Title:

Chemical map–based prediction of nucleosome positioning using the Bioconductor package nuCpos

Authors:

Hiroaki Kato[1], Mitsuhiro Shimizu[2] and Takeshi Urano[1].

Institutional addresses:

1. Department of Biochemistry, Shimane University School of Medicine, Izumo, Shimane 693-8501, Japan
2. Department of Chemistry, Graduate School of Science and Engineering, Program in Chemistry and Life Science, School of Science and Engineering, Meisei University, Hino, Tokyo 191-8506, Japan

Corresponding author:

Hiroaki Kato
Department of Biochemistry, Shimane University School of Medicine
89-1 Enya-cho, Izumo 693-8501, Japan
Tel: +81-853-20-2126; Fax: +81-853-20-2125
hkato@med.shimane-u.ac.jp

Abstract

Background:

Methods for predicting nucleosome positioning include bioinformatics, biophysical, and combined approaches. An advantage of bioinformatics methods, which are based on *in vivo* nucleosome maps, is the use of natural sequences that may contain previously unknown elements involved in nucleosome positioning *in vivo*. The accuracy of such prediction attempts reflects the genomic coordinate resolution of the nucleosome maps applied. Thus, nucleosome maps constructed using *Micrococcus* nuclease (MNase), which has a strong preference for A/T-rich sequences, may not be appropriate for this purpose. In addition to MNase seq–based maps, base pair–resolution chemical maps of *in vivo* nucleosomes from three different species (budding and

fission yeasts, and mouse) are currently available. However, these maps have yet to be integrated into publicly available computational methods.

Results:

We developed a Bioconductor package (named nuCpos) that uses chemical maps to train duration hidden Markov models (dHMMs) to predict *nuc*leosome *pos*itioning. The accuracy of chemical map–based prediction was higher than that of the previously developed MNase seq–based approach. With our method, predicted nucleosome occupancy reasonably matched *in vivo* observations and was not affected by A/T nucleotide frequency. Effects of genetic alterations on nucleosome positioning that had been observed in living yeast could also be predicted. In addition to dHMM-based prediction, nuCpos can calculate individual histone binding affinity (HBA) scores for given 147-bp sequences to examine their suitability for nucleosome formation. Local HBA scores for 13 overlapping nucleosomal DNA subsegments can also be calculated. HBA and local HBA scores for various sequences agreed well with previous *in vitro* and *in vivo* studies. Furthermore, our results suggest that nucleosomal subsegments that are disfavored in different rotational settings contribute to the defined positioning of nucleosomes.

Conclusions:

Our results demonstrate that chemical map–based statistical models are beneficial for studying nucleosomal DNA features. Studies employing nuCpos software can enhance understanding of chromatin regulation and the interpretation of genetic alterations and facilitate the design of artificial sequences. nuCpos is freely available from Bioconductor under the GPL-2 license.

Keywords:

nucleosome, prediction of nucleosome positions, dHMM, HBA, local HBA, translational and rotational settings, nucleosomal DNA subsegments

Background:

The nucleosome is a structure in which a ~147-base pair (bp) double-stranded DNA forms a left-handed super-helix to wrap a histone octamer consisting of two copies of histones H2A, H2B, H3, and H4 [1-3]. The positioning of nucleosomes along genomic DNA sequences affects various biological events, including transcription, DNA replication and repair, and chromosome segregation [4]. The importance of nucleosome positioning has prompted the development of computational methods to predict nucleosome positions along a given DNA sequence [5]. These methods are classified into three groups: biophysical, bioinformatics, and combined [5, 6]. Most biophysical methods are based on the structures of reconstituted nucleosome core particles (NCPs) such as NCP147, which contains a palindromic 147-bp sequence with intrinsic positional stability [3, 7, 8]. The use of palindromic sequence optimized for homogeneous nucleosome reconstitution and the highest resolution of the structure contributed greatly to the development of prediction algorithms. However, relying on the representative structure itself could also be a potential weakness, as there could be unrecognized natural intra-nucleosomal DNA elements that mediate specific physiologic functions in cells.

Studies using reconstituted nucleosomes other than NCP147 have provided many insights into nucleosome organization. In particular, studies using derivatives of the Widom 601 sequence [7] demonstrated the importance of intra-nucleosomal DNA elements from a physiologic context. When a nucleosome with this sequence is placed before elongating RNA polymerase II, the nucleosome acts as a polar barrier to transcription [9]. In this scenario, the polymerase coming from the right side of the sequence pauses in a manner dependent on an element called R3, which is found at the superhelical location (SHL) −2.5 to the left and strongly contacts the histone core [10, 11]. Structural studies demonstrated that a palindromic sequence consisting of two copies of the left-half of the Widom sequence yields a high-resolution X-ray structure, whereas the nucleosome with a right-half palindrome is less stable and yields poor crystals [12, 13]. Efficient reconstitution with the left-half sequence was confirmed in another study [14]. Indeed, the positioning power of the left-half sequence is exploited in reconstitution studies that require nucleosomes at targeted positions [15, 16]. However, why the left- and right-half sequences

behave differently and how the left-half sequence determines the rotational and translational settings remain to be fully elucidated.

In contrast to the biophysical prediction methods, bioinformatics methods are based on maps of *in vivo* nucleosomes determined using high-throughput sequencing techniques. *In vivo* nucleosomes contain various DNA elements that should not necessarily be appropriate for nucleosome formation (e.g., protein-coding sequences and recognition motifs for transcription factors). In addition, various types of *trans*-acting factors (e.g., transcription factors, histone chaperones, and chromatin remodeling factors) are known to regulate nucleosome positioning in cells, as reviewed elsewhere [17-19]. Furthermore, similarity in the usage of dinucleotides at each superhelical turn in the nucleosome may help nucleosomes position redundantly along the genomic coordinates with a periodicity of approximately 10 bp [20-23]. Indeed, *in vivo*, most nucleosomes exhibit redundant positions on the genome [24-26]. The use of complex, natural nucleosomal sequences may cause difficulty in the interpretation of prediction results; however, use of these sequences may also lead to unexpected findings that enhance understanding of nucleosome-based regulation.

A weakness of currently available bioinformatics methods is that the statistical models are constructed based on *in vivo* nucleosome maps generated using *Micrococcus* nuclease (MNase) digestion followed by high-throughput sequencing (MNase-seq). MNase digests linker DNA that connects NCPs, allowing researchers to obtain nucleosomal DNA that has been protected from enzymatic digestion. In MNase-seq experiments, DNA fragments of mono-nucleosome size (~150 bp) are sequenced and mapped to the reference genome. However, as MNase strongly prefers A/T-rich sequences, unwanted digestion of A/T-rich nucleosomal sequences and less-effective digestion of G/C-rich linkers may occur, leading to variation in nucleosome fragment size [19, 27]. Because of this enzymatic bias, precise determination of the genomic coordinates of *in vivo* nucleosomes using MNase-seq (i.e., dyad base calling) is challenging.

One bioinformatics method considered successful in a thorough comparative study [6] is NuPoP, which is built upon a duration hidden Markov

model (dHMM) that considers information regarding both the nucleosome and linker DNA [28]. NuPoP enables the prediction of nucleosome positioning in nine different species by rescaling the base composition of both the nucleosome and linker models for the budding yeast *Saccharomyces cerevisiae* with those of target organisms. The budding yeast models are generated using an MNase-seq–based nucleosome map by considering the central base of each nucleosomal fragment of varying length the dyad base [28]. This may particularly affect construction of the time-dependent statistical model for the nucleosome state, which ideally requires that the dyad bases be determined at bp resolution. The construction of the linker model, which requires extraction of linker sequences as genomic sequences not covered by nucleosomes, may also be affected. Thus, although selecting the MNase-seq–based map was the best practice at the time of the software's development, the prediction results output provided by NuPoP can be biased by both the enzymatic preference of MNase and the difficulty of dyad base calling.

In order to examine the suitability of a given 147-bp sequence for nucleosome formation using the NuPoP algorithm, a histone binding affinity (HBA) score, also referred to as a 'nucleosome affinity score', is calculated [28]. According to the authors' definition, this score should ideally predict the rotational setting of nucleosomes along the genome. However, this capability would only be expected when the *in vivo* nucleosome map used for construction of the statistical model is at bp resolution. This means that the low coordinate resolution of the MNase-seq–based map could cause ambiguity in the prediction. Indeed, at the implementation level, NuPoP smooths raw HBA scores using a moving average of a 55-bp window and standardizes the scores on a per sequence basis before outputting the processed HBA scores. This prevents software users from examining the actual HBA score for each genomic coordinate to see if the 147-bp sequence of interest is suitable for nucleosome formation.

A breakthrough in nucleosome mapping occurred with the development of a method for site-directed chemical cleavage of nucleosomal DNA followed by high-throughput sequencing [24, 29]. This technique, called chemical mapping, was first applied to budding yeast and then to the fission yeast

Schizosaccharomyces pombe and embryonic stem cells of the house mouse, *Mus musculus* [24-26, 30, 31]. In cells expressing histone H4 with an S47C amino acid substitution located near the nucleosome center, nucleosomal DNA is specifically cleaved. By examining the cleavage sites, previous studies determined the dyad positions of *in vivo* nucleosomes in the three species at bp resolution [24-26]. In addition to the histone H4-S47C approach, chemical mapping of nucleosomes using a histone H3 mutation (Q85C) determined the precise positions of −1 and +1 nucleosomes for protein-coding genes [27]. Molye-Hyerman and colleagues demonstrated that dHMMs trained with chemical maps can roughly predict the phasing pattern of nucleosome occupancy in budding and fission yeasts [25]. For an unknown reason, they used a uniform distribution as the linker length distribution in the calculation. In order to predict the relative positions of nucleosomes using the dHMM as it was originally proposed, a non-uniform, experimentally obtained distribution of linker lengths should be used. Most importantly, the bp-resolution chemical maps have not yet been implemented in publicly available software for prediction of nucleosome positioning.

One goal of nucleosome positioning prediction is *in silico* reproduction of *in vivo* observations, preferably at the locus level. If this goal is achieved, predicting the effects of genetic alterations on nucleosome positioning and prediction-based engineering of nucleosome-forming and -depleted sequences would also be practical. In this regard, budding and fission yeasts would serve as preferable model organisms, because nucleosomes formed on native and modified sequences *in vivo* have been precisely mapped [32-37]. These nucleosomes are suitable for *in silico* reproduction of *in vivo* observations. Importantly, for the currently available bioinformatics method NuPoP, the reliability of locus-level nucleosome prediction has seldom been demonstrated, despite its wide use in previous studies [38-47].

In this study, we developed a nucleosome positioning prediction method that utilizes available chemical maps of three species for construction of dHMMs. We show that the prediction accuracy of chemical map–based models is higher than that of MNase-seq–based models. The chemical map–based models appear to recognize appropriate sides of DNA strings as histone-DNA

contact surfaces. The effects of genetic alterations on nucleosome positioning previously observed in living yeast cells were reasonably reproduced *in silico*, suggesting that this software is useful for chromatin engineering and interpretations of mutation effects. We found that the prediction accuracy for mice was lower than that for yeasts, which should be acknowledged by software users. Our software, nuCpos, is available at the Bioconductor website (https://doi.org/doi:10.18129/B9.bioc.nuCpos).

Results:

Chemical map–based prediction of nucleosome positioning in yeasts

In order to determine whether the bp-resolution chemical maps are suitable for predicting nucleosome positioning, we trained dHMMs with the nucleosome maps determined with H4-S47C–dependent DNA cleavage (see **Methods**). Linker length distributions derived from the chemical maps, which were implemented in our method (nuCpos), exhibited a 10-bp periodicity, as previously reported [24-27, 48] (**Additional file 1: Figure S1**). In contrast, linker length distributions of the MNase-seq–based method, NuPoP, did not exhibit such periodicity, reflecting the difficulty of mapping nucleosomes at bp resolution using MNase digestion.

First, we predicted nucleosome positioning in the genomes of budding and fission yeasts (denoted ‘Sc’ and ‘Sp’, respectively, in **Figure 1**) with models that were constructed with *in vivo* nucleosome maps of target species (hereafter referred to as “matched-species prediction”). Based on the prediction results, we drew receiver operating characteristic (ROC) curves (**Figure 1A**, upper panels) and calculated area under the curve (AUC) scores (**Additional file 1: Table S1**). Here, only the predicted nucleosomes for which the dyad bases matched those of *in vivo* nucleosomes at a resolution of 1 bp were considered to be correctly predicted (see **Methods**). The prediction performance of the MNase-seq–based model for unique (representative) budding yeast nucleosomes as determined by AUC was 0.532, and it was 0.526 for redundant nucleosomes (**Figure 1A**, Target: Sc, Model: Sc, blue). Prediction for the fission yeast genome resulted in similar results (**Figure 1A**, Target: Sp, Model: Sp, blue). In contrast, better prediction performance was observed with the chemical map–based model

(**Figure 1A**, upper panels, red). In both yeasts, AUC scores for unique nucleosomes (budding yeast, 0.751; fission yeast, 0.743) were more than 0.07 points higher than those for redundant nucleosomes (**Additional file 1: Table S1**). These results suggest that the chemical map–based models better predict the suitability of a given sequence for nucleosome formation, as unique nucleosomes are known to have sequences preferred for nucleosome formation [24, 25].

Prediction performance was also evaluated using cross-species models; nucleosome positioning in the target genomes was predicted using models constructed with nucleosome maps of different species. Using the budding yeast chemical model for prediction of nucleosome positioning in the fission yeast genome (**Figure 1A**, Target: Sp, Model: Sc, red) resulted in lower AUC scores for unique and redundant nucleosomes (0.667 and 0.626, respectively) in comparison with the matched-species scores (**Additional file 1: Table S1**). A reciprocal prediction with the budding yeast genome and the fission yeast chemical model resulted in similar scores (**Figure 1A**, Target: Sc, Model: Sp, red). These results indicate that the species of the nucleosome map from which the model is constructed must be matched to the prediction target species for optimal prediction results. For MNase-seq–based predictions, matched-species and cross-species predictions performed similarly (**Additional file 1: Table S1 and Figure 1A**, compare blue lines in upper and bottom panels). Interestingly, these MNase-seq–based scores were much lower than those obtained from chemical map–based cross-species predictions. These results thus highlight the usefulness of chemical map–based models and indicate that nucleosome-forming sequences exhibit similar properties in both types of yeast, as previously reported [25].

Ideally, prediction results should be evaluated at a 1-bp resolution, as done above, because if nucleosomes can be predicted at such a high resolution, it might be possible to discern the rotational settings of given nucleosomes. As the resolution of the chemical maps that we used for model construction could be 1-2 bp or lower, the resolution of the prediction results would be expected to be at a similar level. Indeed, the distance between chemically predicted nucleosomes and the nearest *in vivo* nucleosomes was generally 0 bp (located

at the same position) or 1 bp (located at the juxtaposed nucleotide positions) (**Figure 1B**, left panel). In order to determine the appropriate resolution for examining prediction results, we calculated the matching rate between predicted and *in vivo* nucleosomes with stepwise widening of the matching window (**Additional file 1: Figure S2**). Here, we defined the matching window as the allowable distance between nucleosomes regarded as matched. In chemical map–based predictions (nuCpos), the matching rate doubled to over 50% when the matching window was widened from 0 bp (1-bp resolution) to 1 bp (2-bp resolution) (**Additional file 1: Figure S2**, red lines). However, the rate did not markedly increase upon further widening of the matching window, suggesting that the chemical models distinguish the histone-facing side of the DNA string.

In MNase-seq–based predictions (NuPoP), by contrast, the matching rate increased gradually as the matching window was widened from 0 to 9 bp (**Additional file 1: Figure S2**, blue lines). The matching rate drew a very loose anti-S shape, suggesting that the MNase-seq–based model still recognized the histone-facing side of the DNA string, but only to a small extent. We also noticed that a substantial population of nucleosomes predicted with the MNase-seq–based model was positioned approximately 5 nucleotides away from the *in vivo* nucleosomes (**Figure 1B**, right panel, marked with a blue horizontal bar). These observations suggest that the MNase-seq–based model has difficulty in discriminating rotationally mispredicted nucleosomes. Considering these results, we concluded that prediction results should be evaluated at 2-bp resolution, as further widening of the matching window would increase the misprediction rate.

The left panel of **Figure 1C** presents the evaluation of matched-species predictions at 2-bp resolution. When the budding yeast MNase-seq–based model (NuPoP) was applied to the prediction of nucleosome positioning in the budding yeast genome (Target: Sc, Model: Sc), only 3.2% and 12.6% of the nucleosomes on the most probable (Viterbi) path matched *in vivo* unique (red) and redundant (orange) nucleosomes, respectively. When the fission yeast model was applied to the prediction of nucleosome positioning in the fission yeast genome (Target: Sp, Model: Sp), the MNase-seq–based model scored 2.7% and 11.0%, respectively. In contrast, when the budding yeast chemical

map–based model (nuCpos) was used for predictions (Target: Sc, Model: Sc), 23.5% and 50.9% of the nucleosomes on the Viterbi path matched unique and redundant nucleosomes, respectively. For the fission yeast (Target: Sp, Model: Sp), the corresponding scores were 25.5% and 54.2%, respectively. Thus, with the chemical map–based models, over half of the nucleosomes predicted as being on the Viterbi path are present in living cells at exactly the same positions or at the juxtaposed nucleotide positions. In the chemical maps of budding and fission yeasts, *in vivo* unique nucleosomes accounted for 19.6% and 17.8%, respectively, of all redundant nucleosomes [24, 25]. Thus, in the MNase-seq–based predictions, enrichment of unique nucleosomes on the Viterbi paths was only 1.3- and 1.4-fold, respectively. In contrast, the chemical map–based predictions exhibited 2.4- and 2.6-fold enrichment, respectively. The higher enrichments suggest that chemical map–based models better distinguish sequences favorable for nucleosome formation, compared with MNase-seq–based models [24, 25].

The right panel of **Figure 1C** presents the evaluation of cross-species predictions at 2-bp resolution. When the fission yeast MNase-seq–based model (NuPoP) was used to predict nucleosome positioning in the budding yeast genome (Target: Sc, Model: Sp), 12.7% of the nucleosomes on the Viterbi path matched *in vivo* redundant nucleosomes. Prediction in the fission yeast genome with the budding yeast model (Target: Sp, Model: Sc) scored 10.8%. When the chemical map–based model (nuCpos) was used for cross-species predictions, greater proportions of predicted nucleosomes matched *in vivo* redundant nucleosomes: 36.7% (Target: Sc, Model: Sp) and 37.1% (Target: Sp, Model: Sc). Thus, although the prediction accuracy tends to decrease, cross-species predictions using chemical map–based models appear to produce better results than MNase-seq–based models.

Prediction of *in vitro* reconstituted nucleosome positions

The evaluations described above suggested that chemical models perform better in the prediction of rotational settings. To confirm this observation, we compared HBA scores calculated along the original 282-bp Widom 601 sequence [7] with the chemical map–based and MNase-seq–based models

(**Figure 2A**). As mentioned in the Introduction section, non-smoothed HBA scores are expected to predict the rotational setting of nucleosomes if the resolution of the nucleosome map used for model construction is sufficiently high. The original 282-bp Widom 601 sequence contained primer sequences for the SELEX study at both ends and additional sequences flanking the nucleosome-forming 147-bp sequence centered at nucleotide position 154 [7]. When analyzed using the chemical map–based model, discrete HBA peaks with an interval of approximately 10 bp were observed at positions 124, 135, 143, 154, 164, and 175 (asterisks). This periodicity indicated that nearly the same surface of the DNA string was predicted to be in contact with the histone core in these potential translational positions. Among these positions, the one with the highest HBA score corresponds to the translational position of the *in vitro* stable nucleosome (position 154, orange vertical lines). In contrast, the MNase-seq–based HBA score was relatively low at position 154, and the peaks did not exhibit the apparent 10-bp periodicity. Thus, the chemical map–based model performed better for the prediction of rotational settings, at least when the Widom 601 sequence was tested.

We also calculated HBA scores along the 485-bp mouse mammary tumor virus (MMTV) 3'–long terminal repeat sequence (**Figure 2B**). The dyad bases of two nucleosomes, designated B and A, were located at nucleotide positions 139 and 335 (−127 and +70 relative to the transcriptional start site), respectively [49]. At these positions, chemical map–based HBA scores were the highest within the 50-bp-wide surrounding regions. For nucleosome B, alternative translational positions with an interval of approximately 10 bp were suggested (asterisks), probably due to the higher dinucleotide periodicity of nucleosome B compared with nucleosome A [49]. Note that the periodicity of the chemical map–based HBA peaks on this natural sequence was not as clear as seen with the *in vitro*–optimized Widom sequence (**Figure 2A**). We also noticed that the MNase-seq–based HBA score was highest at position 339, which was quite near nucleosome-forming position 335. However, the 4-base difference suggested that the MNase-seq–based model primarily recognized the opposite side of the DNA string as the histone-contacting surface.

Figure 2C shows the results of HBA calculations for a dinucleosomal DNA fragment derived from the somatic 5S RNA gene of *Xenopus borealis* [50]. As one of the two nucleosomes occupies a 150-bp region spanning nucleotide positions 30 to 179 (−70 to +79 with respect to the +1 nucleotide position of the 5S RNA gene), the dyad base should be around nucleotide positions 103 to 106 [50]. In this tandemly joined duplex sequence, the dyad base of the second nucleosome should be located around positions 300 to 303. The chemical map–based model revealed potential nucleosomes with their dyad bases at positions 105 to 107 and 302 to 304, with relatively high HBA scores. Around these potential dyad positions, 10-bp periodicity of the chemical map–based HBA was observed (asterisks). Selection of *in vitro* nucleosome-forming positions from these candidates could be made by linker histones [50, 51]. In contrast to the prediction with the chemical map–based model, the MNase-seq–based model did not appear to successfully recognize appropriate rotational settings, as in the case of the MMTV sequence.

Prediction of *in vivo* nucleosome positions

Next, we calculated HBA scores around budding yeast −1 and +1 nucleosomes (n=5,542 each) mapped by means of histone H3-Q85C–based chemical cleavage [27]. As shown in **Figure 3A**, mean HBA scores calculated using both models exhibited a periodicity with a 10-bp interval. However, the amplitude of the chemical map–based HBA scores (red) around the tested nucleosomes was larger than that of the MNase-seq–based HBA scores (blue). This suggested that the chemical model could clearly distinguish the surface of the DNA string that interacts with the histone core even when natural genomic sequences were queried. We found that the chemical map–based HBA scores were highest at the −1 and +1 nucleosome positions (0 bp). Chemical map–based HBA scores for the three neighboring translational positions (10, 20, and 30 bp in distance) relative to the nucleosome-depleted regions (marked with asterisks in **Figure 3A**) remained relatively high despite the gradual increase in the A/T frequency of the tested sequences (**Figure 3B**). After the A/T frequency reached a plateau (A/T≈0.64), the chemical map–based HBA scores began to decrease. In contrast, MNase-seq–based HBA scores at the −1 and +1 nucleosome positions

(0 bp) were already at moderate levels, and they simply decreased as the A/T frequency increased. Thus, the MNase-seq–based model was adversely influenced by the enzymatic bias of MNase, whereas the chemical map–based model was not.

The higher accuracy of chemical map–based models at the genomic level prompted us to perform a locus-level evaluation of prediction results, using the budding yeast *TRP1ARS1* minichromosome [34, 36]. On this well-studied, circular, 1,465-bp DNA, three (numbered I to III) and four (IV to VII) nucleosomes are interspersed with two nucleosome-depleted regions or nuclease-hypersensitive regions designated *HSRA* and *HSRB*. *HSRA* includes the DNA replication origin *ARS1*, whereas *HSRB* is located upstream of the *TRP1* marker gene. At a glance, the chemical map– and MNase-seq–based model prediction results appeared very different (**Figure 3C**). Each model suggested eight and seven representative nucleosomes, respectively. Seven of the eight nucleosomes predicted by the chemical map–based model (**Figure 3C**, nuCpos) reasonably matched the positions of *in vivo* nucleosomes determined previously [36]. The only exception was a predicted nucleosome located in *HSRA*, in which functional *ARS1* would be occupied by the ARS-binding factor Abf1 and the origin recognition complex instead of the nucleosome to maintain this circular DNA across generations [52]. This prediction was reasonable, as DNA replication origins are known to be covered by nucleosomes when the subunits of the origin recognition complex or Abf1 are perturbed *in vivo* [52-54].

In this native sequence, the chemical map–based HBA scores exhibited a periodicity of an approximately 10-bp interval (**Figure 3C**, nuCpos, and **Additional file 1: Figure S3**). Some nucleotide positions with high HBA scores were selected by the chemical map–based dHMM as representative nucleosome positions. As the dHMM considers the linker length distribution as a major factor affecting nucleosome-linker state transition and does not allow nucleosomes with shorter DNAs (<147 bp), nucleosome positions tend to be selected in a limited manner. Although there were many high-HBA positions, only a few were clearly selected by the dHMM. Thus, although the predicted occupancy provides a rough image of chromatin state, it should be cautiously

considered. Instead, it should be kept in mind that periodic HBA scores may contribute to redundant nucleosome positioning *in vivo*.

In contrast to the chemical map–based model, most of the nucleosomes predicted by the MNase-seq–based model were mislocated (**Figure 3C**, NuPoP); most of the *in vivo* nucleosome centers (inverted triangles in figure) were located in predicted linkers or nucleosome ends. This tendency to mislocate nucleosomes may explain the lower genome-wide prediction accuracy (**Figure 1**). Importantly, MNase-seq–based HBA scores for this sequence appeared to mirror the A/T-content over the X-axis (**Figure 3C**, A/T); high-HBA positions were generally A/T-poor, and low-HBA positions were A/T-rich. In addition, the local amplitude of MNase-seq–based HBA scores was apparently lower compared with the chemical map–based model (**Additional file 1: Figure S3**). These observations also demonstrated that MNase-seq–based model prediction results are influenced by the A/T preference of MNase. A/T-content exhibited a similar effect in the MNase-seq–based nucleosome positioning prediction around the fission yeast *ura4*$^{+}$ gene (**Additional file 1: Figure S4**) [55]. Given these results, we concluded that the chemical map–based models output better genome-wide and locus-level prediction results.

Prediction of effects of genetic alterations on nucleosome positioning

Next, we examined whether the effects of genetic alterations on nucleosome positioning could be predicted. In α cells of budding yeast, expression of the a-specific gene *BAR1* is repressed by *α2*-operator–dependent positioning of the nucleosome at its promoter, which masks the gene's *TATA* box [32, 35]. The chemical map–based model perfectly predicted the positioning of this repressive nucleosome (**Figure 4**, WT). The HBA score for this translational position (0 bp), which is 82 bp away from the *α2*-operator, was very high (HBA=8.47), suggesting that repression of the *BAR1* gene is assisted by the intrinsic suitability of the promoter DNA for nucleosome formation. When a 36-bp sequence consisting of 12 repeats of CTG is inserted at the center of this nucleosome, repression of *BAR1* and nucleosome positioning are not affected *in vivo* [35]. Prediction results agreed that this insertion should not cause nucleosome depletion (**Figure 4**, $[CTG]_{12}$). Other repeat sequences causing no

transcriptional derepression *in vivo* [35] were also predicted to not cause nucleosome depletion (**Figure 4**, Sac_5 and Sac_6). It is also known that insertion of a 30-bp A-stretch or a 10-bp CG (i.e., CpG) repeat sequence inhibits *in vivo* nucleosome formation and causes derepression of *BAR1* [35]. Interestingly, the effect of these genetic alterations on nucleosome positioning was reproduced *in silico* with the chemical map–based model (**Figure 4**, A_{30} and CG_5). We noted that the disruptive effect of fragment insertion on nucleosome positioning was more pronounced in predictions than *in vivo* observations. Insertion of a 20-bp A-stretch, which does not cause complete nucleosome depletion *in vivo*, was clearly predicted to inhibit nucleosome formation (**Figure 4**, A_{20}). However, in the case of CG repeat insertion, shortening of the repeat to 8 bp, which satisfies *in vivo* nucleosome formation [35], was predicted to preserve the capability to form nucleosomes (**Figure 4**, CG_4). Thus, despite some exceptions, the effects of genetic alterations on nucleosome positioning can be predicted with the chemical map–based model.

We also confirmed that the inhibition of nucleosome formation by telomeric DNA insertion demonstrated in a previous study [56] was also largely predictable (**Additional file 1: Figure S5**). In budding yeast cells, the fourth nucleosome of a derivative of the *TRP1ARS1* minichromosome called *TALS* is formed in a manner dependent on the *α2*-operator [33, 37]. Note that the chemical map–based model predicted the location of this nucleosome at a shifted position that unexpectedly would cover the *α2*-operator (**Additional file 1: Figure S5**, WT); the central of the three predicted nucleosomes should have been centered at nucleotide position 0. The shift may be a consequence of the relatively low HBA score for the *in vivo* translational position (position 0), where the nucleosomal DNA contains an artificial sequence of bacterial origin [33, 37]. When telomeric DNA fragments are inserted at the center of the fourth nucleosome, shorter fragments (hTEL2 and hTEL4) were predicted to not affect nucleosome positioning. In contrast, longer fragments (hTEL12 and hTEL29) were predicted to severely compromise nucleosome formation. These prediction results agreed with previous *in vivo* observations [56]. In addition, sequence isomers of the telomeric repeat (SI-A6, SI-A12, SI-B6, and SI-B12) were predicted to not inhibit nucleosome formation, which also agreed with the *in vivo*

experiments [56]. The hTEL6 insertion, which causes nucleosome depletion *in vivo*, appeared not to affect nucleosome formation in the prediction, suggesting a limitation of the prediction. Nonetheless, these results suggest that the chemical map–based models are useful for predicting the effects of genetic alterations on nucleosome positioning.

Local HBA for nucleosomal DNA subsegments

The successful increase in prediction accuracy and better recognition of rotational settings with chemical map–based models suggested that the chemical map–based HBA score for a given 147-bp sequence is a good indicator of the suitability of that sequence for nucleosome formation. However, considering that nucleosomal DNA makes contact with histone proteins at each superhelical location (SHL ±0.5-6.5), we thought that the suitability of intra-nucleosomal DNA segments for each histone-DNA contact could also be calculated as “local” HBA scores. In order to implement this idea, we divided a 147-bp whole nucleosomal segment into 13 overlapping nucleosomal DNA subsegments, designated A through M (**Additional file 1: Figure S6**, See **Methods**). Each 20- or 21-bp segment corresponded to two histone-DNA contact sites, each of which was shared by neighboring segments. For example, segment B corresponded to SHL −5.5 and −4.5, which were shared by segments A and C, respectively. We took this overlapping approach because it remains unclear how surrounding sequences affect histone-DNA contacts. Calculating local HBA scores is conceptually the same as that of HBA calculations, which consider the probability of a whole 147-bp sequence to be a nucleosome [28]. In local HBA calculation, for each intra-nucleosomal segment of a given 147 bp, the probability of the selected 20- or 21-bp sequence to be located at that part of the nucleosome and the probability of the same sequence functioning as a linker are calculated. The log likelihood ratio of these probabilities is defined as the local HBA score for that segment. Thus, the sum of local HBA scores for the non-overlapping set of seven 21-bp segments, A, C, E, G, I, K, and M (147 bp in total), would be nearly equal to the HBA score calculated for the same 147-bp sequence.

In order to examine whether local HBA scores indicate the suitability of nucleosomal subsegments for histone-DNA contacts, we used the 282-bp Widom 601 sequence, which was used for the above HBA analysis (**Figure 2A**), as a test sequence. At potential translational positions with high HBA scores, local HBA scores for some segments were very high, as expected (**Figure 5A**, arrows). However, each 147-bp sequence also contained segments with relatively low local HBA scores. Interestingly, at translational positions with low HBA scores (those intervened by the high-scoring HBA positions), some segments exhibited very low local HBA scores (blue blots in the heatmap). This tendency was particularly notable for the central segments E through I. The finding of these low-scoring local HBA segments suggests that sub-sequences that are disfavored in other rotational settings play a role in determining rotational settings.

Next, we focused on the 147-bp Widom 601 sequence, of which the dyad base is located at nucleotide position 154 of the original sequence (**Figure 5A**, red arrow). We chose this sequence because its derivatives are often used in reconstitution studies, as it forms a homogeneous nucleosome, and because the left and right halves exhibit different features. In agreement with the transcription studies mentioned in the Background section [10, 11], local HBA scores for the 147-bp sequence were highest for segments D and E, both of which share SHL −2.5, in which the high-affinity R3 element is located (**Figure 5B**, red dots). The region covered by segments D and E (SHL −1.5 to −3.5) are known to exhibit high permanganate reactivity, which is indicative of DNA distortion resulting from very strong histone-DNA interaction at these segments [12]. These data indicate that local HBA scores are useful for evaluating the suitability of nucleosomal subsegments for histone-DNA contacts.

The overall image of the local HBA for the 147-bp Widom 601 sequence did not change with variations in the terminal sequences for biochemical handling (**Additional file 1: Figure S7**, 601LR, red dots) [12]. In the palindromic sequences made of each half, the local HBA landscapes were conserved compared with that of the original sequence, as expected by their sequence identity (**Additional file 1: Figure S7**, 601L and 601R, red dots). Indeed, the left-half palindrome (601L) had four segments with very-high HBA

scores (segments D, E, I, and J) and four other segments with relatively low scores (A, B, L, and M). The local HBA score for segment G increased from a low to moderate level following palindromization. In contrast to the left half, the local HBA scores for the right-half palindrome (601R) were moderate for segments A through F and H through M and relatively low for segment G. The absence of segments with very-high local HBA scores in the right-half palindrome may contribute to the lower salt stability of the reconstituted nucleosome [12].

When the translational position was shifted along the 282-bp sequence from nucleotide position 154 to surrounding nucleotide positions ranging from 150 to 158 (**Figure 5C**, −4 to +4 with respect to 154), the local HBA scores for segments B through I changed dramatically (**Figure 5B**, black dots). However, such shifting resulted in modest changes at segments J through M, with their scores remaining at relatively high levels. These observations suggest that segments J through M in the right-half sequence are generally suitable for nucleosome formation, which may also be true even when the translational position is shifted by several base pairs. In other words, segments J through M do not seem to have intra-nucleosomal disfavored elements that strongly limit the rotational setting. Reflecting the property of the non-palindromic sequence, palindromization of the left half yielded six segments (D-F and H-J) that exhibited the highest local HBA scores at the *in vitro* translational position (**Additional file 1: Figure S7**, 601L, black dots). In addition, the shifting score patterns for segments C and K suggest that shifting to either side may reduce the sum score. Thus, the rotational setting of this sequence appears to be determined through overall suitability for histone-DNA contacts. For the right-half palindrome, only the central segments with low to moderate local HBA scores (E through I) could contribute to the rotational setting. This nature might account for the observed poor crystal development or lower homogeneity [12], as this ‘slippery’ sequence could disrupt the uniformity of histone-DNA interactions.

As a comparison, we analyzed the palindromic sequence of NCP147 (**Additional file 1: Figure S7**, NCP147). In this sequence, the local HBA scores for outer segments A through C and K through M were high, whereas the inner part (segments D through J) scored lower (red dots). In contrast to the Widom

derivatives, local HBA scores for most of the segments were not high at the *in vitro* translational position. However, each segment appeared to contribute to the determination of rotational setting. For example, while the score for segment B decreased when the translational position was shifted to the left side, that for segment C decreased when shifted to the opposite side. The same was true for segments K and L. Similarly, local HBA scores for segments F and H decreased when the translational position was shifted inward. These observations suggest that being a palindrome itself helps generate homogeneous nucleosomes by preventing alternative positioning.

Local HBA scores for modified nucleosome-forming sequences

The above testing with the sequences previously used in reconstitution studies suggested that calculating local HBA scores would enhance understanding of how DNA sequences characterize the nucleosomes formed *in vitro*. This prompted us to examine whether we could apply local HBA calculations to the evaluation of intracellular nucleosome formation. González et al. reported that sequence-based determinants of nucleosome positioning are dispersed across nucleosomes [55]. One gene that they thoroughly investigated was the fission yeast gene, *ura4*$^{+}$. This gene has six representative nucleosomes (designated +1 to +6) that overlap the coding sequence, five of which were predictable with the chemical map–based dHMM (**Additional file 1: Figure S4**, nuCpos). Although the remaining nucleosome (+5) was not clearly observed in the predicted occupancy plot, there was one position with a relatively high HBA score that reasonably corresponded to the *in vivo* nucleosome (marked with an asterisk in the figure). A heatmap of local HBA scores for this gene suggested that there is an element disfavored for nucleosome formation near the +5 nucleosome (**Figure 6A**, WT, see the blueish slanted line spanning segments A through M around position 700). Thus, this element may have caused the lower predicted occupancy of the +5 nucleosome.

González et al. also generated three modified sequences, namely Dyad, Linker, and Int, for the *ura4*$^{+}$ gene and studied nucleosome positioning on these sequences *in vivo* [55]. In Dyad and Linker, 51-bp sequences centered to each nucleosome dyad and each linker, respectively, were replaced with

artificial randomized sequences of the same length. In the sequence Int, two intra-nucleosomal sequences of each nucleosome (positions from −51 to −24 and 24 to 51) were replaced. According to their report, the occupancy of the +1 nucleosome was markedly decreased in the Dyad sequence, that of the +2 nucleosome was slightly increased in the Dyad and Linker sequences, and the position of the +3 nucleosome was shifted in the Int sequence [55]. However, why these changes selectively occur on specific nucleosomes *in vivo* remains unclear due to a lack of appropriate analytical methods.

A heatmap of local HBA scores for the Dyad sequence clearly showed that the replacements in this sequence caused a marked decrease in local HBA scores only around the position of the +1 nucleosome (**Figure 6A**, Dyad, blue lines not present in the WT map). The heatmap also suggested that the replaced segment in the +1 nucleosome exerts a strong effect only on the central nucleosomal subsegments (E through I). Indeed, at the *in vivo* translational position, the Dyad replacement caused marked decreases in HBA scores (**Figure 6B**, +1 Nuc., left, compare gray and green lines), and the effect was limited to the central segments (**Figure 6B**, +1 Nuc., right, asterisks). In contrast, the 51-bp replacements at other nucleosomal centers did not appear to dramatically change their suitability for nucleosome formation. As to the +2 nucleosome, we did not observe any supportive signatures that could explain why the Dyad and Linker replacements slightly increased its occupancy (**Figure 6A**, Dyad and Linker, and **Figure 6B**, +2 Nuc.). We observed that the Int replacements primarily affected local HBA scores around the position of the +3 nucleosome (**Figure 6A**, Int, see the blueish slanted line spanning segments A through M). The Int replacement caused decreases in neighboring HBA scores, and the effect was greatest at segments C and D (asterisks), where local HBA scores were among the highest (**Figure 6B**, +3 Nuc., compare gray and black lines). This suggests that the mutation triggers a shift of the nucleosome to a more suitable position *in vivo*. Giving these results, local HBA score calculation for nucleosomal subsegments is useful for *in silico* assessment of modified nucleosome-forming sequences.

Chemical map–based prediction of nucleosome positioning in mice

The improved prediction accuracy of the chemical map–based models in both yeasts suggested that the prediction accuracy in mice would also be improved. A big difference between the genomes of yeasts and vertebrates is that the latter have CG modifications. We observed that the CG dinucleotide frequency in mouse nucleosomal DNA sequences was much lower than the frequencies of other SS dinucleotides (each S stands for a C or G nucleotide) (**Additional file 1: Figures S8A and S8B**). Even in yeast nucleosomes, CG was the least favored dinucleotide, as previously discovered in budding yeast H2A.Z-containing nucleosomes [23]. This suggests that CG could be targeted by modification-mediated regulation of nucleosome-forming suitability. After normalization with the genomic frequency, mouse nucleosomes did not appear to exhibit a strong bias in CG frequency overall (**Figure 7A**). In detail, when compared with the two yeasts, the normalized CG frequency in mouse nucleosomes was relatively low at SHL 0 (**Figure 7B**, marked with a triangle) and high at the inner parts of some histone-DNA contact sites (asterisks), as well as some parts of outer regions (gray horizontal lines). This suggests that CG dinucleotides can affect the formation of murine nucleosomes at these intra-nucleosomal regions.

We predicted nucleosome positioning in the mouse genome using the MNase-seq–based and chemical map–based models. Based on the prediction results, we drew ROC curves and calculated AUC scores (**Figure 7C**). The prediction performance of the MNase-seq–based model for the unique nucleosomes measured by AUC was 0.511, and it was 0.509 for the redundant nucleosomes. As expected, the chemical map–based model performed better for both unique (AUC=0.711) and redundant (AUC=0.646) nucleosomes. Matching of the predicted nucleosomes on the Viterbi paths at a 2-bp resolution also demonstrated the better performance of the chemical map–based model (**Figure 7D**). In the MNase-seq–based prediction, 1.4% and 11.0% of predicted nucleosomes matched unique and redundant nucleosomes, respectively. In contrast, the chemical map–based prediction scored 10.0% and 41.1%, respectively. Thus, although the accuracy was not as high as in the case of yeasts (**Figure 1C**, scored over 50%), the chemical map–based model predicted

the suitability of given sequences for nucleosome formation in mice better than the MNase-seq–based model.

In order to determine whether the mouse chemical map–based model could better predict the rotational setting of nucleosomes, as shown in the case of yeasts, we calculated the matching rate between predicted and *in vivo* nucleosomes with stepwise widening of the matching window (**Additional file 1: Figure S9**). As expected, the matching rate for MNase-seq–based prediction increased gradually as the matching window widened. In chemical map–based predictions, a dramatic increase in matching rate was observed when the matching window was widened from 0 to 1 bp. However, unexpectedly, the matching rate also increased gradually as the matching window was widened further (See **Additional file 1: Figures S2 and S9** for comparison between species). The shapes of the matching rate curves suggested the chemical map–based model still distinguished rotational settings to some extent. However, the mouse chemical model did not appear to be as good at discriminating rotationally mispredicted nucleosomes compared with the yeast models.

Discussion:

We developed a chemical map–based computational method for predicting nucleosome positioning and assessed the prediction results. Training dHMMs with chemical maps improved prediction accuracy in budding yeast, fission yeast, and the house mouse, *Mus musculus* (**Figures 1 and 7, and Additional file 1: Figures S2 and S9 and Table S1**). 41-54% of the predicted nucleosomes on the Viterbi paths matched those of *in vivo* nucleosomes at 2-bp resolution. Genome- and locus-level evaluations showed that the software successfully predicted the positions of representative *in vivo* nucleosomes (**Figures 3 and 4, and Additional file 1: Figures S3 and S4**). The *in vivo* nucleosomes that were not predicted as being representative had high HBA scores at reasonably near genomic positions. Furthermore, we demonstrated that perturbation of nucleosome positioning associated with genetic alterations could also be predicted (**Figures 4 and 6, and Additional file 1: Figure S5**). Thus, we

propose that the software nuCpos can be used for predicting nucleosome positioning and also for engineering of nucleosome-forming sequences.

As expected, the use of chemical maps in prediction led to strict recognition of rotational settings, which was demonstrated with *in vitro* and *in vivo* nucleosome-forming sequences (**Figure 2 and Additional file 1: Figures S2 and S9**). Indeed, chemical map–based HBA scores for the original Widom 601 sequence indicated that the 147-bp sequence centered at nucleotide position 154 was suitable for nucleosome formation (**Figure 2A**). The *in vitro* nucleosome positions formed on natural sequences were also predictable at a rotational level (**Figures 2B and C**); however, the HBA amplitude along these sequences was smaller than along the *in vitro*–optimized Widom sequence. In contrast to the chemical map–based models, the MNase-seq–based models largely failed to recognize rotational settings (**Figure 2 and Additional file 1: Figures S2 and S9**). This may have been due to the A/T preference of MNase and the difficulty of dyad base calling in the construction of MNase-seq–based nucleosome maps [19, 27]. Indeed, the MNase-map–based model simply predicted A/T-rich regions as nucleosome-depleted regions (**Figure 3 and Additional file 1: Figure S4**). Therefore, instead of lower-resolution MNase-seq–based maps, base-pair-resolution chemical maps should be used in the development of bioinformatics methods to obtain better prediction results.

In the field of synthetic biology, the MNase-seq–based method NuPoP has contributed to the design of synthetic promoters and terminators [38, 39]. However, some studies reported that the histone binding affinity and predicted nucleosome occupancy output by NuPoP do not correlate with synthetic promoter and terminator activity [43, 44]. Differences in the DNA sequences of the tested synthetic elements may of course account for this discrepancy, as discussed elsewhere [43, 44]. Clearly, the A/T frequency does affect the prediction results of NuPoP (**Figure 3 and Additional file 1: Figure S4**). As demonstrated in the Results section, the chemical map–based method is more accurate and not affected by the A/T frequency. Thus, we expect that nuCpos is applicable to prediction-based engineering and more effective for synthesizing functional elements. Similarly, nuCpos can also be used to examine whether

DNA sequences of interest are suitable for nucleosome formation, as previously done with NuPoP [46, 47].

Dividing 147-bp nucleosomal DNA into subsegments and calculating subsegment local HBA scores revealed unexpected landscapes of the tested nucleosomes. Our study revealed that the Widom sequence had segments that are both favored and disfavored for nucleosome formation (**Figure 5A**). Importantly, segments with high local HBA scores in the Widom 601–derived sequences matched the high-affinity histone-contacting regions (**Figure 5B and Additional file 1: Figure S7**). Furthermore, the Widom derivatives had segments exhibiting the highest local HBA scores at the *in vitro* rotational settings, which was further emphasized by palindromization with the left-half sequence (**Additional file 1: Figure S7**). In contrast to the left-half sequence, the right-half sequence contained segments exhibiting relatively high local HBA scores that did not change markedly with different rotational settings. Thus, these sequence-specific features may account for the different behavior of the Widom derivatives in *in vitro* experiments. For the NCP147 sequence, it appeared that palindromization itself repressed the shifting of rotational settings (**Additional file 1: Figure S7**). Collectively, our data indicate that local HBA scores for given sequences provide insights that enhance understanding of nucleosome features.

Locus-level assessments of prediction results suggested that software users should not place undue confidence in the predicted nucleosome occupancy output from chemical map–based models. Indeed, there was a misplaced nucleosome that occupied the replication origin *ARS1* (**Figure 3**) and a shifted nucleosome that covered the *α2*-operator (**Additional file 1: Figure S5**). Also, the +5 nucleosome of *ura4*$^{+}$ was not predicted as a representative nucleosome (**Additional file 1: Figure S4**). We speculate that these discrepancies were due primarily to the fact that the models do not consider involvement of functional DNA elements or transacting factors that regulate nucleosome formation at specific genomic locations. In other words, chromatin regions highlighted by the differences in nucleosome occupancy between prediction and *in vivo* observations could be the targets of future investigations aiming to uncover region-specific regulatory mechanisms. From another point of

786 view, software users should keep in mind that the dHMMs consider linker length
787 distribution but do not consider partial unwrapping of nucleosomal DNA in
788 calculating probable paths of nucleosome and linker states. Instead, our
789 observations suggest that HBA scores indicate redundant positioning of
790 nucleosomes in the cells; HBA scores are simply calculated for a given 147-bp
791 sequence, independently of surrounding sequences.
792 In the three species with available chemical nucleosome maps, mouse
793 genomic DNA undergoes covalent modification of cytosine in CG dinucleotides,
794 whereas that of budding or fission yeast does not. There is a negative correlation
795 between CG frequency and nucleosome occupancy in mammalian cells [57]. In
796 addition, hypermethylation of cytosine is observed when the chromatin of the
797 mouse *Pou5f1* gene, which encodes the pluripotency factor Oct4, is changed to
798 a closed state, suggesting a link between DNA methylation and nucleosome
799 positioning [58-60]. Furthermore, DNA methylation is correlated with
800 nucleosome occupancy in humans [61]. Indeed, the nucleosome structure itself
801 has been suggested as necessary for the maintenance of DNA methylation [62,
802 63]. In addition, many histone variants exist in mice, which might cause
803 variability in the preferred sequence for nucleosome formation [64-66]. Taking
804 this species- or mammalian-specific evidence into account, for nucleosome
805 positioning prediction in mice, statistical models should be better constructed
806 from the nucleosome map of mice, not by rescaling yeast models, if applicable.
807 We demonstrated that the prediction accuracy of the chemical
808 map–based mouse model was apparently lower than that of yeast models
809 (**Figures 2 and 7**). A disadvantage of mouse prediction is that the chemical map
810 probably contains substantial numbers of pseudo nucleosomes [26], which could
811 not be omitted due to the lower cleavage depth than yeasts. As a consequence,
812 the density of nucleosomes on the mouse genome is one redundant
813 nucleosome per 3 bp, which is about 10 times higher than yeast values (see
814 **Methods**). Thus, we speculate that the accuracy of mouse nucleosome
815 positioning prediction can be increased if other chemical maps of higher quality
816 are produced and used for model construction. In human genomics, the
817 relationship between various phenotypes and genetic variants has been
818 explored [67-70]. However, the effects of genetic variants on nucleosome

formation in this regard remain to be studied. Now that our study showed that insertions and replacements that disrupt nucleosomes in yeast cells can be predictable *in silico*, it is possible that nucleosome-disrupting human mutations can be found with the aid of chemical map–based predictions. We showed that cross-species chemical map–based models exhibit relatively high prediction accuracy compared with MNase-seq–based models (**Figure 1**). Therefore, before generating high-quality chemical maps for species other than the budding and fission yeasts, the yeast models can alternatively be applied for the prediction of nucleosome positioning in those species.

Conclusions:

In this study, we demonstrated that the accuracy of dHMM-based nucleosome positioning prediction can be substantially increased by using base-pair-resolution nucleosome maps for model construction. Our prediction results suggest that chemical map–based models are useful for predicting nucleosome positioning in wild-type and modified sequences at the locus level. We also demonstrated that strong histone-DNA contacts in a nucleosome and their rotational settings can be predicted. Furthermore, as another advantage of bioinformatics methods, our models indicate that the commonly used Widom sequence contains subnucleosomal segments that are disfavored, or statistically very rare, in *in vivo* nucleosomes at their nucleosomal positions in shifted rotational settings. We expect that our prediction method will provide further insights that will enhance understanding of nucleosome-based epigenetic regulation.

Methods:

Software and data sets

Most analyses were performed in the GNU R environment (https://www.r-project.org, ver. 3.6.1). R packages were obtained from CRAN (https://cran.r-project.org/) and Bioconductor (http://bioconductor.org/). Chemical maps for budding yeast, fission yeast, and house mouse (*Mus musculus*) embryonic stem cells [24-26] were used for model construction and testing the prediction accuracy. The budding yeast's chemical map in the sacCer2

coordinate was lifted over to the sacCer3 coordinate, as described elsewhere [36]. The number of unique and redundant nucleosomes in the budding yeast genome was 67,548 and 344,709, respectively; fission yeast, 75,828 and 425,653; mice, 10,677,016 and 850,701,275. Reference genomes of budding yeast (R64-1-1), fission yeast (ASM294v2), and mice (mm9/NCBIM37.67) were used.

Construction of the nuCpos package

The codes for the construction of parameters used in the nuCpos package (ver. 1.2.0) are available online (**https://doi.org/10.5281/zenodo.3362065**). Genomic regions covered with the 147-bp non-redundant (unique) chemically mapped nucleosomes and uncovered were defined as nucleosome and linker regions, respectively. DNA sequences of these regions were used to construct parameters that were transferred to internal Fortran programs. Nucleosomes of which dyads were located within 73 bp of the chromosomal ends were omitted. For construction of the mouse model, hard-masked genomic sequences were used, and nucleosome and linker regions containing N were omitted before parameter construction to avoid potential prediction bias caused by repeat elements. In total, 67,538 nucleosome regions and 50,622 linker regions were obtained for the budding yeast genome (sacCer3); fission yeast (ASM294v2), 75,826 nucleosome and 46,557 linker regions; mice (mm9), 4,147,972 nucleosome and 2,484,347 linker regions.

We developed an R function designated *predNuCpos*, which predicts nucleosome positioning based on a dHMM, as previously proposed by Xi et al. [28]. Like its ancestral function *predNuPoP* in the NuPoP package (https://doi.org/doi:10.18129/B9.bioc.NuPoP, ver. 1.34.0), *predNuCpos* receives a DNA sequence of any length, invokes an internal Fortran program, and outputs the prediction result either in the working directory or in the working environment of R. In *predNuCpos*, construction of the dHMM is based on chemical maps, as described below.

Parameters used in the *predNuCpos* function were constructed according to the NuPoP paper [28] using the functionalities of the Biostrings package (https://doi.org/doi:10.18129/B9.bioc.Biostrings, ver. 2.52.0). The

parameters were as follows: *freqL*, one-base frequencies for linker regions; *tranL*, *tranL2*, *tranL3*, and *tranL4*, 1st- to 4th-order transition probabilities for linker regions, respectively; *freqN4*, four-base frequencies at the first four nucleotide positions of nucleosome regions; *tranN4*, time-dependent 4th-order transition probabilities for nucleosome regions; *Pd*, linker length distribution that ranges from 1 to 500 bp. Linker sequences of 7-500 bp in length were used for linker model construction, as described elsewhere [28]. *freqL* and *freqN4* were obtained using the *oligonucleotideFrequency* function of Biostrings; *tranL*, *tranL2*, *tranL3*, *tranL4*, and *tranN4* were obtained using the *oligonucleotideTransitions* function of Biostrings. Moving average smoothing using the *SMA* function of the TTR package (https://cran.r-project.org/package=TTR, ver. 0.23-4) at a 3-bp window was applied to the 4th-order transition probability parameter *tranN4* and to the linker length distribution parameter *Pd*. The parameters used in *predNuCpos* were also used in another function *mutNuCpos*, which predicts the effect of genetic alterations on nucleosome positioning.

Xi et al. proposed the HBA score as the log likelihood ratio of the probability for a given 147-bp sequence to be a nucleosome versus a linker [28]. According to their definition, the HBA score for the 147-bp region x centering at position i (a_i) on a given genomic sequence is,

$$a_i := log\left[\frac{P_N(x_{i-73},\ldots,x_{i+73})}{G_L(x_{i-73},\ldots,x_{i+73}|147)}\right],$$

where P_N and G_L represent the probability of observing the 147-bp sequence as a nucleosome or a linker, respectively [28]. The probability of being a nucleosome is calculated by referring to the parameters *freqN4* and *tranN4*, which are derived from nucleosomal DNA sequences. Similarly, calculation of the probability of being a linker is based on linker DNA sequences. As nucleosomal and linker sequences do not overlap in terms of their genomic coordinates, negativity of HBA does not directly mean that the tested sequence is inappropriate for nucleosome formation. The *predNuCpos* function calculates chemical map–based HBA scores along the input sequence and outputs them as raw values as its default behavior. We developed an independent function designated *HBA*, which only calculates the HBA score for a given 147-bp

sequence. The *HBA* function uses the abovementioned chemical parameters for *predNuCpos*: *freqL*, *tranL*, *tranL2*, *tranL3*, *tranL4*, *freqN4*, and *tranN4*.

We defined 13 overlapping nucleosomal subsegments, A through M, and developed a function designated *localHBA* that calculates "local" HBA scores for each segment. Segment A corresponds to nucleosomal nucleotide positions 1-21; B, 12-31; C, 22-42; D, 33-52; E, 43-63; F, 54-73; G, 64-84; H, 75-94; I, 85-105; J, 96-115; K, 106-126; L, 117-136; and M, 127-147. Similar to the calculation of HBA [28], the local HBA score for segment A of the 147-bp potential nucleosomal region x centering at position i (l_i) on a given genomic sequence is calculated as,

$$l_i := log\left[\frac{P_N(x_{i-73},\ldots,x_{i-53})}{G_L(x_{i-73},\ldots,x_{i-53}|21)}\right],$$

where P_N and G_L represent the probability of observing the 21-bp sequence as segment A of a nucleosome or a linker, respectively. Local HBA scores for the other segments are calculated in the same way, except that the considered nucleotide positions are set appropriately. At the implementation level, four-base frequency values for the first four nucleotide positions of each segment were prepared: *freqN4SA* corresponds to nucleosome positions 1-4; *freqN4SB*, 12-15; *freqN4SC*, 22-25; *freqN4SD*, 33-36; *freqN4SE*, 43-46; *freqN4SF*, 54-57; *freqN4SG*, 64-67; *freqN4SH*, 75-78, *freqN4SI*, 85-88; *freqN4SJ*, 96-99; *freqN4SK*, 106-109; *freqN4SL*, 117-120; and *freqN4SM*, 127-130. These four-base frequency values were used to calculate the probability of the segment as that part of nucleosomal DNA as done for HBA calculations [28]. The parameter *freqN4SA* was identical to *freqN4*, which was used for HBA calculations.

Prediction of nucleosome positioning with nuCpos and NuPoP and evaluation of the prediction results

The codes for prediction, evaluation, and figure drawing used in this study are available online (**https://doi.org/10.5281/zenodo.3523573**). Unmasked reference genomes for the budding yeast, fission yeast, and house mouse (*Mus musculus*) were used as prediction target sequences. For the mouse experiments, only the 19th chromosome (Chr19) was used for prediction

because its length (61,342,430 bp) was sufficient for evaluation, at approximately five-times longer than the total length of yeast chromosomes (budding yeast, 12,071,326 bp; fission yeast, 12,571,820 bp). Mouse Chr19 contained 249,210 unique and 19,965,481 redundant nucleosomes that had been determined based on histone H4 S47C-dependent cleavage [26]. The density of redundant nucleosomes on mouse Chr19 was one per 3.07 bp. With this density, one in three randomly located nucleosomes in the chromosome could be counted as a truly predicted nucleosome. Thus, we selected 2,044,747 nucleosomes with the highest NCP scores as true redundant nucleosomes for the matching experiment. This selection yielded a density of one nucleosome per 30.0 bp, which was comparable to that of budding yeast and fission yeast (one nucleosome per 35.0 bp and 29.5 bp, respectively).

The *predNuPoP* function of NuPoP and the *predNuCpos* function of nuCpos were used for dHMM-based predictions. In order to predict using budding yeast models, the species arguments were set as “7” for *predNuPoP* and “sc” for *predNuCpos*. Similarly, fission yeast (“9” and “sp”) and mouse (“2” and “mm”) models were specified. We defined any nucleosomes with *P*-values greater than zero (*P*>0) as predicted nucleosomes. Matching of predicted nucleosomes using *in vivo* chemical nucleosomes was done by widening the genomic coordinates of *in vivo* nucleosomes to both sides in a step-wise manner from zero to nine base pairs. For instance, at 1-bp resolution, a predicted nucleosome was regarded as “matched” when the genomic coordinate of its dyad was equal to that of the nearest *in vivo* nucleosome. If the dyad of the predicted nucleosome was 1 bp away from that of the nearest *in vivo* nucleosome, it was regarded as “not matched.” At 2-bp resolution, a predicted nucleosome was regarded as “matched” when the distance from its dyad to that of the nearest *in vivo* nucleosome was <2 bp. For drawing ROC curves and calculation of AUC, a matrix of two columns was generated for each prediction result. In the first column, *P*-values of nucleosomes that were *P*>0 were stored. In the second column, the matching status for each nucleosome according to the matching experiments at 1-bp resolution was stored. After processing using the *prediction* function of the ROCR package (https://cran.r-project.org/package=ROCR, ver. 1.0-7), ROC curves were drawn

using the *plot* function of R. The AUC was calculated using the *performance* function of ROCR. The matching rate of the nucleosomes on the Viterbi paths, which were predicted by *predNuPoP* and *predNuCpos*, was calculated using in-house scripts and plotted using the *barplot* function of R.

Dinucleotide frequencies for nucleosomal DNA and unmasked genomic DNA were calculated using the *oligonucleotideFrequency* function of Biostrings. For normalization of dinucleotide frequencies, the nucleosomal frequency was divided by the genomic frequency. Heatmaps of 16 dinucleotide frequencies along the nucleosome were drawn using the *levelplot* function of the rasterVis package (https://cran.r-project.org/package=rasterVis, ver. 0.46). The *plot* function was used for plotting the frequencies of CG and other SS dinucleotides.

Chemical map–based HBA scores were calculated using the *predNuCpos* function or the *HBA* function of nuCpos, which yield the same scores. The species argument was set as “sc” for budding yeast sequences and sequences of *in vitro* reconstitution studies; for the fission yeast sequences, “sp.” Non-smoothed MNase-seq–based HBA scores were calculated using in-house functions that utilize the parameters for the *predNuPoP* function of NuPoP. For comparison of *in vivo* nucleosomes, 5,542 sets of −1 and +1 nucleosomes determined by Chereji et al. [27] were used for HBA calculations. The 347-bp sequences around these nucleosomes (from −173 to +173 nucleotide positions with respect to the dyads) were extracted from the budding yeast genome. The HBA score was calculated for each possible 147-bp segment along the sequences. A/T-frequency was calculated using the *letterFrequency* function of Biostrings for each 147-bp sequence. HBA and A/T-frequency scores were averaged at each coordinate with respect to the −1 or +1 nucleosomes.

Chemical map–based local HBA scores were calculated using the *localHBA* function of nuCpos. Heatmaps of local HBA scores were drawn using the *levelplot* function of rasterVis.

List of abbreviations:

AUC: area under the curve

CG: cytosine-guanine dinucleotide
dHMM: duration hidden Markov model
HBA: histone binding affinity
LTR: long terminal repeat
MMTV: mouse mammary tumor virus
MNase: *Micrococcus* nuclease
NCP: nucleosome core particle
NDR: nucleosome depleted region
ROC: receiver operating characteristic
SELEX: Systematic Evolution of Ligands by EXponential enrichment
SHL: superhelical location

Declarations:

Ethics approval and consent to participate:

Not applicable.

Consent for publication:

Not applicable.

Availability of data and materials:

The datasets generated and analyzed during the current study are available online (see **Methods** for the URLs).

Competing interests:

The authors declare that they have no competing interests.

Funding:

This study was supported by JSPS KAKENHI Grant Numbers JP18K05556 and JP19H05264.

Authors' contributions:

H.K. designed the study, created the software, performed the experiments, and was a major contributor in writing the manuscript. T.U. and M.S. were involved in

study design, data validation, and manuscript writing. All authors read and approved the final manuscript.

Acknowledgments:

The authors thank Dr. K. Ura for providing the dinucleosome-forming sequence and Dr. Y. Ichikawa for confirming the DNA sequences. The authors also thank Drs. H. Kimura, H. Kurumizaka, J. Nakayama, and Y. Ohkawa for their helpful comments and encouragement.

References:

1. 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:**251-260.
2. Luger K, Richmond TJ: **DNA binding within the nucleosome core.** *Curr Opin Struct Biol* 1998, **8:**33-40.
3. Richmond TJ, Davey CA: **The structure of DNA in the nucleosome core.** *Nature* 2003, **423:**145-150.
4. Lai WKM, Pugh BF: **Understanding nucleosome dynamics and their links to gene expression and DNA replication.** *Nat Rev Mol Cell Biol* 2017, **18:**548-562.
5. Teif VB: **Nucleosome positioning: resources and tools online.** *Brief Bioinform* 2016, **17:**745-757.
6. Liu H, Zhang R, Xiong W, Guan J, Zhuang Z, Zhou S: **A comparative evaluation on prediction methods of nucleosome positioning.** *Brief Bioinform* 2014, **15:**1014-1027.
7. Lowary PT, Widom J: **New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning.** *J Mol Biol* 1998, **276:**19-42.
8. Pisano S, Galati A, Cacchione S: **Telomeric nucleosomes: forgotten players at chromosome ends.** *Cell Mol Life Sci* 2008, **65:**3553-3563.
9. Bondarenko VA, Steele LM, Ujvari A, Gaykalova DA, Kulaeva OI, Polikanov YS, Luse DS, Studitsky VM: **Nucleosomes can form a polar**

barrier to transcript elongation by RNA polymerase II. *Mol Cell* 2006, **24:**469-479.

10. Kulaeva OI, Gaykalova DA, Pestov NA, Golovastov VV, Vassylyev DG, Artsimovitch I, Studitsky VM: **Mechanism of chromatin remodeling and recovery during passage of RNA polymerase II.** *Nat Struct Mol Biol* 2009, **16:**1272-1278.
11. Gaykalova DA, Kulaeva OI, Volokh O, Shaytan AK, Hsieh FK, Kirpichnikov MP, Sokolova OS, Studitsky VM: **Structural analysis of nucleosomal barrier to transcription.** *Proc Natl Acad Sci U S A* 2015, **112:**E5787-5795.
12. Chua EY, Vasudevan D, Davey GE, Wu B, Davey CA: **The mechanics behind DNA sequence-dependent properties of the nucleosome.** *Nucleic Acids Res* 2012, **40:**6338-6352.
13. Vasudevan D, Chua EYD, Davey CA: **Crystal structures of nucleosome core particles containing the '601' strong positioning sequence.** *J Mol Biol* 2010, **403:**1-10.
14. Tsunaka Y, Fujiwara Y, Oyama T, Hirose S, Morikawa K: **Integrated molecular mechanism directing nucleosome reorganization by human FACT.** *Genes Dev* 2016, **30:**673-686.
15. Kujirai T, Ehara H, Fujino Y, Shirouzu M, Sekine SI, Kurumizaka H: **Structural basis of the nucleosome transition during RNA polymerase II passage.** *Science* 2018, **362:**595-598.
16. Ehara H, Kujirai T, Fujino Y, Shirouzu M, Kurumizaka H, Sekine SI: **Structural insight into nucleosome transcription by RNA polymerase II with elongation factors.** *Science* 2019, **363:**744-747.
17. Struhl K, Segal E: **Determinants of nucleosome positioning.** *Nat Struct Mol Biol* 2013, **20:**267-273.
18. Lieleg C, Krietenstein N, Walker M, Korber P: **Nucleosome positioning in yeasts: methods, maps, and mechanisms.** *Chromosoma* 2015, **124:**131-151.
19. Chereji RV, Clark DJ: **Major Determinants of Nucleosome Positioning.** *Biophys J* 2018, **114:**2279-2289.

20. Lowman H, Bina M: **Correlation between dinucleotide periodicities and nucleosome positioning on mouse satellite DNA.** *Biopolymers* 1990, **30:**861-876.
21. Satchwell SC, Drew HR, Travers AA: **Sequence periodicities in chicken nucleosome core DNA.** *J Mol Biol* 1986, **191:**659-675.
22. Drew HR, Travers AA: **DNA bending and its relation to nucleosome positioning.** *J Mol Biol* 1985, **186:**773-790.
23. Ioshikhes I, Hosid S, Pugh BF: **Variety of genomic DNA patterns for nucleosome positioning.** *Genome Res* 2011, **21:**1863-1871.
24. Brogaard K, Xi L, Wang JP, Widom J: **A map of nucleosome positions in yeast at base-pair resolution.** *Nature* 2012, **486:**496-501.
25. Moyle-Heyrman G, Zaichuk T, Xi L, Zhang Q, Uhlenbeck OC, Holmgren R, Widom J, Wang JP: **Chemical map of Schizosaccharomyces pombe reveals species-specific features in nucleosome positioning.** *Proc Natl Acad Sci U S A* 2013, **110:**20158-20163.
26. Voong LN, Xi L, Sebeson AC, Xiong B, Wang JP, Wang X: **Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping.** *Cell* 2016, **167:**1555-1570 e1515.
27. Chereji RV, Ramachandran S, Bryson TD, Henikoff S: **Precise genome-wide mapping of single nucleosomes and linkers in vivo.** *Genome Biol* 2018, **19:**19.
28. Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang JP: **Predicting nucleosome positioning using a duration Hidden Markov Model.** *BMC Bioinformatics* 2010, **11:**346.
29. Brogaard KR, Xi L, Wang JP, Widom J: **A chemical approach to mapping nucleosomes at base pair resolution in yeast.** *Methods Enzymol* 2012, **513:**315-334.
30. Henikoff S, Ramachandran S, Krassovsky K, Bryson TD, Codomo CA, Brogaard K, Widom J, Wang JP, Henikoff JG: **The budding yeast Centromere DNA Element II wraps a stable Cse4 hemisome in either orientation in vivo.** *Elife* 2014, **3:**e01861.

31. Ramachandran S, Zentner GE, Henikoff S: **Asymmetric nucleosomes flank promoters in the budding yeast genome.** *Genome Res* 2015, **25:**381-390.
32. Shimizu M, Roth SY, Szent-Gyorgyi C, Simpson RT: **Nucleosomes are positioned with base pair precision adjacent to the alpha 2 operator in Saccharomyces cerevisiae.** *EMBO J* 1991, **10:**3033-3041.
33. Roth SY, Dean A, Simpson RT: **Yeast alpha 2 repressor positions nucleosomes in TRP1/ARS1 chromatin.** *Mol Cell Biol* 1990, **10:**2247-2260.
34. Thoma F, Bergman LW, Simpson RT: **Nuclease digestion of circular TRP1ARS1 chromatin reveals positioned nucleosomes separated by nuclease-sensitive regions.** *J Mol Biol* 1984, **177:**715-733.
35. Morohashi N, Yamamoto Y, Kuwana S, Morita W, Shindo H, Mitchell AP, Shimizu M: **Effect of sequence-directed nucleosome disruption on cell-type-specific repression by alpha2/Mcm1 in the yeast genome.** *Eukaryot Cell* 2006, **5:**1925-1933.
36. Fuse T, Katsumata K, Morohoshi K, Mukai Y, Ichikawa Y, Kurumizaka H, Yanagida A, Urano T, Kato H, Shimizu M: **Parallel mapping with site-directed hydroxyl radicals and micrococcal nuclease reveals structural features of positioned nucleosomes in vivo.** *PLoS One* 2017, **12:**e0186974.
37. Fuse T, Yanagida A, Shimizu M: **The Yeast Minichromosome System Consisting of Highly Positioned Nucleosomes in Vivo.** *Biol Pharm Bull* 2019, **42:**289-294.
38. Curran KA, Crook NC, Karim AS, Gupta A, Wagman AM, Alper HS: **Design of synthetic yeast promoters via tuning of nucleosome architecture.** *Nat Commun* 2014, **5:**4002.
39. Morse NJ, Gopal MR, Wagner JM, Alper HS: **Yeast Terminator Function Can Be Modulated and Designed on the Basis of Predictions of Nucleosome Occupancy.** *ACS Synth Biol* 2017, **6:**2086-2095.
40. Wight A, Yang D, Ioshikhes I, Makrigiannis AP: **Nucleosome Presence at AML-1 Binding Sites Inversely Correlates with Ly49 Expression:**

Revelations from an Informatics Analysis of Nucleosomes and Immune Cell Transcription Factors. *PLoS Comput Biol* 2016, **12:**e1004894.

41. Creamer KM, Job G, Shanker S, Neale GA, Lin YC, Bartholomew B, Partridge JF: **The Mi-2 homolog Mit1 actively positions nucleosomes within heterochromatin to suppress transcription.** *Mol Cell Biol* 2014, **34:**2046-2061.

42. Small EC, Xi L, Wang JP, Widom J, Licht JD: **Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity.** *Proc Natl Acad Sci U S A* 2014, **111:**E2462-2471.

43. Portela RM, Vogl T, Kniely C, Fischer JE, Oliveira R, Glieder A: **Synthetic Core Promoters as Universal Parts for Fine-Tuning Expression in Different Yeast Species.** *ACS Synth Biol* 2017, **6:**471-484.

44. MacPherson M, Saka Y: **Short Synthetic Terminators for Assembly of Transcription Units in Vitro and Stable Chromosomal Integration in Yeast S. cerevisiae.** *ACS Synth Biol* 2017, **6:**130-138.

45. Mahmoudi T: **The BAF complex and HIV latency.** *Transcription* 2012, **3:**171-176.

46. Rafati H, Parra M, Hakre S, Moshkin Y, Verdin E, Mahmoudi T: **Repressive LTR nucleosome positioning by the BAF complex is required for HIV latency.** *PLoS Biol* 2011, **9:**e1001206.

47. Gimeno-Valiente F, Riffo-Campos AL, Vallet-Sanchez A, Siscar-Lewin S, Gambardella V, Tarazona N, Cervantes A, Franco L, Castillo J, Lopez-Rodas G: **ZNF518B gene up-regulation promotes dissemination of tumour cells and is governed by epigenetic mechanisms in colorectal cancer.** *Sci Rep* 2019, **9:**9339.

48. Wang JP, Fondufe-Mittendorf Y, Xi L, Tsai GF, Segal E, Widom J: **Preferentially quantized linker DNA lengths in Saccharomyces cerevisiae.** *PLoS Comput Biol* 2008, **4:**e1000175.

49. Flaus A, Richmond TJ: **Positioning and stability of nucleosomes on MMTV 3'LTR sequences.** *J Mol Biol* 1998, **275:**427-441.

50. Ura K, Hayes JJ, Wolffe AP: **A positive role for nucleosome mobility in the transcriptional activity of chromatin templates: restriction by linker histones.** *EMBO J* 1995, **14:**3752-3765.
51. Ura K, Nightingale K, Wolffe AP: **Differential association of HMG1 and linker histones B4 and H1 with dinucleosomal DNA: structural transitions and transcriptional repression.** *EMBO J* 1996, **15:**4959-4969.
52. Lipford JR, Bell SP: **Nucleosomes positioned by ORC facilitate the initiation of DNA replication.** *Mol Cell* 2001, **7:**21-30.
53. Eaton ML, Galani K, Kang S, Bell SP, MacAlpine DM: **Conserved nucleosome positioning defines replication origins.** *Genes Dev* 2010, **24:**748-753.
54. Rodriguez J, Lee L, Lynch B, Tsukiyama T: **Nucleosome occupancy as a novel chromatin parameter for replication origin functions.** *Genome Res* 2017, **27:**269-277.
55. Gonzalez S, Garcia A, Vazquez E, Serrano R, Sanchez M, Quintales L, Antequera F: **Nucleosomal signatures impose nucleosome positioning in coding and noncoding sequences in the genome.** *Genome Res* 2016, **26:**1532-1543.
56. Ichikawa Y, Morohashi N, Nishimura Y, Kurumizaka H, Shimizu M: **Telomeric repeats act as nucleosome-disfavouring sequences in vivo.** *Nucleic Acids Res* 2014, **42:**1541-1552.
57. Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I, Andrau JC: **CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters.** *Genome Res* 2012, **22:**2399-2408.
58. Hattori N, Nishino K, Ko YG, Hattori N, Ohgane J, Tanaka S, Shiota K: **Epigenetic control of mouse Oct-4 gene expression in embryonic stem cells and trophoblast stem cells.** *J Biol Chem* 2004, **279:**17063-17069.
59. Osorno R, Tsakiridis A, Wong F, Cambray N, Economou C, Wilkie R, Blin G, Scotting PJ, Chambers I, Wilson V: **The developmental dismantling**

of pluripotency is reversed by ectopic Oct4 expression. *Development* 2012, **139:**2288-2298.

60. West JA, Cook A, Alver BH, Stadtfeld M, Deaton AM, Hochedlinger K, Park PJ, Tolstorukov MY, Kingston RE: **Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming.** *Nat Commun* 2014, **5:**4719.

61. Collings CK, Anderson JN: **Links between DNA methylation and nucleosome occupancy in the human genome.** *Epigenetics Chromatin* 2017, **10:**18.

62. Ishiyama S, Nishiyama A, Saeki Y, Moritsugu K, Morimoto D, Yamaguchi L, Arai N, Matsumura R, Kawakami T, Mishima Y, et al: **Structure of the Dnmt1 Reader Module Complexed with a Unique Two-Mono-Ubiquitin Mark on Histone H3 Reveals the Basis for DNA Methylation Maintenance.** *Mol Cell* 2017, **68:**350-360 e357.

63. Li T, Wang L, Du Y, Xie S, Yang X, Lian F, Zhou Z, Qian C: **Structural and mechanistic insights into UHRF1-mediated DNMT1 activation in the maintenance DNA methylation.** *Nucleic Acids Res* 2018, **46:**3218-3231.

64. Maehara K, Harada A, Sato Y, Matsumoto M, Nakayama KI, Kimura H, Ohkawa Y: **Tissue-specific expression of histone H3 variants diversified after species separation.** *Epigenetics Chromatin* 2015, **8:**35.

65. Mendiratta S, Gatto A, Almouzni G: **Histone supply: Multitiered regulation ensures chromatin dynamics throughout the cell cycle.** *J Cell Biol* 2019, **218:**39-54.

66. Harada A, Maehara K, Ono Y, Taguchi H, Yoshioka K, Kitajima Y, Xie Y, Sato Y, Iwasaki T, Nogami J, et al: **Histone H3.3 sub-variant H3mm7 is required for normal skeletal muscle regeneration.** *Nat Commun* 2018, **9:**1400.

67. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, Hess GT, Zappala Z, Strober BJ, Scott AJ, et al: **The impact of rare variation on gene expression across tissues.** *Nature* 2017, **550:**239-243.

68. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, et al: **Genetic effects on gene expression across human tissues.** *Nature* 2017, **550:**204-213.
69. Chakravorty S, Hegde M: **Inferring the effect of genomic variation in the new era of genomics.** *Hum Mutat* 2018, **39:**756-773.
70. Chakravorty S, Hegde M: **Gene and Variant Annotation for Mendelian Disorders in the Era of Advanced Sequencing Technologies.** *Annu Rev Genomics Hum Genet* 2017, **18:**229-256.

Figure legends:

Figure 1. Comparison of prediction accuracy between chemical map–based and MNase-seq-based models for yeasts. (**A**) ROC curves of prediction results with matched-species models (upper panels) and cross-species models (lower panels). Orange lines represent lines of slope 1 and y-intercept 0. The species of genomic sequences used for prediction (Target) and the species used for model construction (Model) are indicated. “Sc” stands for *S. cerevisiae*, whereas “Sp” stands for *S. pombe*. Chemical map–based (nuCpos) and MNase-seq–based (NuPoP) models were used for predictions. Prediction performance for representative (Unique) and redundant (Redundant) nucleosomes are presented. Matching window was set to 0 bp. (**B**) Distance between predicted nucleosomes and nearest *in vivo* nucleosomes. Probabilities of Viterbi nucleosomes located at a particular distance from the nearest *in vivo* redundant nucleosomes are plotted. Species of the target genome and for model construction were both *S. cerevisiae* (matched-species models). Magnified view is presented inside each plot. Arrows point to the probability of predicted nucleosomes that are 1-bp apart from *in vivo* nucleosomes. (**C**) Matching of nucleosomes on the Viterbi paths with *in vivo* unique and redundant nucleosomes. Note that the redundant nucleosome datasets contain both unique nucleosomes, which are representative (colored red in the bar graphs), and non-representative nucleosomes (orange).

Figure 2. HBA scores along *in vitro* nucleosome-forming sequences. (**A-C**) HBA scores along nucleosome-forming sequences were calculated using chemical map–based and MNase-seq–based budding yeast models. Sequences analyzed were the 282-bp original Widom sequence (**A**), the 3'-LTR of MMTV (**B**), and the *Xenopus borealis* 5S rDNA dinucleosome-forming sequence (**C**). The scores were normalized by subtracting the mean value from the raw value on a per sequence basis. The HBA score for a given 147-bp nucleosome sequence was assigned to its dyad position. Orange vertical lines indicate the dyad positions of *in vitro* reconstituted nucleosomes: nucleotide positions 154 (**A**), 139 and 335 (**B**), and 106 and 303 (**C**). Asterisks indicate high-scoring HBA positions around the *in vitro* positioning sites. Note that for the 5S rDNA sequence, two identical sequences are joined at position 230, causing a difference in HBA scores around the left sides of the two *in vitro* nucleosome positions.

Figure 3. Comparison between chemical map–based and MNase-seq–based models on selected *in vivo* regions. (**A and B**) Chemical map–based and MNase-seq–based HBA scores (**A**) and A/T nucleotide frequency (**B**) along the sequences containing budding yeast −1 and +1 nucleosomes for protein-coding genes. Gray vertical lines indicate the dyad positions of respective nucleosomes. Asterisks indicate positions with relatively high HBA scores. NDR: nucleosome-depleted region. (**C**) Prediction results for the budding yeast *TRP1ARS1* mini-chromosome. Schematic representation of *in vivo* nucleosome positioning is shown above the plots. Nucleosomes labeled I through VII are indicated as ovals. Note that this sequence is circularized *in vivo* by being linked at the *Eco*RI sites. The top two panels show the prediction results output by nuCpos, whereas the next two panels show NuPoP results. The upper panel in each set shows predicted occupancy of the nucleosome (Occup., gray polygons) and probabilities of the tested 147-bp sequences for being in the nucleosome state (P-dyad, blue vertical lines). The lower panels show HBA values for the tested 147-bp sequences calculated using the indicated models. The very bottom panel shows the A/T-frequencies for the tested 147-bp sequences. Horizontal lines at the bottom of each plot indicate a

unit of the circular *TRP1ARS1* mini-chromosome (colored in gray, nucleotide positions 0 to 1,464), the coding region of *TRP1* (red, 115-879), and the *B3* (black, 763-788), *B2* (purple, 810-820), *B1* (green, 847-859), and *ACS* (blue, 869-879) elements of *ARS1*. Inverted triangles indicate the histone H4 S47C-dependent cleavage centers determined by indirect end-labeling; orange for mini-chromosome and light blue for genomic experiments [36].

Figure 4. Predicting the effects of repeat insertion on nucleosome positioning. Chemical map–based HBA scores and predicted nucleosomal occupancy were calculated for original and modified *BAR1* promoter sequences. The sequences were centered at the dyad position of the repressive nucleosome, in which the indicated repeat sequences were inserted. Horizontal lines at the bottom of each plot indicate the *α2*-operator (purple), the *TATA* (blue) element, the coding region of *BAR1* (orange), and the insert (red). Note: the sequence insertion caused a shift of the left of the three predicted nucleosomes to the right to cover the *α2*-operator in the occupancy plots. Overconfidence should not be placed in this shift, as it may simply reflect that the dHMM tends to follow the linker length distribution, which generally does not allow longer linkers.

Figure 5. Local HBA analysis of the Widom 601 sequence. (**A**) Chemical map–based local HBA scores along the 282-bp Widom 601 sequence. The scores for each tested 147-bp sequence were heatmapped at the center position of that sequence. The heatmap scale indicates the local HBA score. Black arrows indicate high-HBA-score positions (124, 135, 143, 164, and 175). Red arrow indicates the dyad position of the *in vitro* reconstituted nucleosome (154). The region evaluated in (**B**) is indicated as a horizontal gray bar under the X-axis. (**B**) Local HBA scores at (red) and around (black) the *in vitro* translational position for each segment are plotted in columns. The scores for the M segment are magnified to the left to show the indices on the X-axis. The relative positions −4, 0, and +4 correspond to nucleotide positions 150, 154, and 158 of the 282-bp sequence. The location of the R3 element is indicated at the top. (**C**) Schematic representation of shifting of translational positions in (**B**). Nucleotide

positions relative to the dyad of the central translated position (i=154) are indicated. Only three (−4, 0, and +4) of the nine positions evaluated in (**B**) are shown.

Figure 6. Local HBA analysis of modified nucleosome-forming sequences. (**A**) Chemical map–based local HBA scores along the original and modified fission yeast *ura4*$^{+}$ gene. The scores for each tested 147-bp sequence were heatmapped at the center position of that sequence. The heatmap scale indicates the local HBA score. Black arrows indicate the nucleosome centers (nucleotide positions +17, +164, +332, +470, +629, and +806) determined by MNase-seq [55]. (**B**) Chemical map–based HBA and local HBA scores for +1 to +3 nucleosomes of the *ura4*$^{+}$ gene. Relative dyad positions with respect to the *in vivo* positions are indicated for HBA plots. Local HBA scores for the *in vivo* positions are shown. Asterisks indicate substantially affected segments.

Figure 7. Comparison of prediction accuracy between chemical map–based and MNase-seq–based models for mice. (**A**) Normalized frequencies of dinucleotides starting at indicated nucleosomal positions. The heatmap scale indicates the normalized dinucleotide frequency. (**B**) Normalized CG frequency across the nucleosome. Triangles, stars, and horizontal bars indicate positions and regions with mouse-specific characteristics. Gray vertical lines indicate the positions of dyad (solid) and histone-DNA contact sites (SHL ±0.5-6.5). (**C**) ROC curves of prediction results with matched-species models. “Mm” stands for *M. musculus*. Prediction performance for representative (Unique) and redundant (Redundant) nucleosomes are presented. (**D**) Matching of nucleosomes on the Viterbi paths with *in vivo* unique and redundant nucleosomes.

Supplementary Figure Legends:

Additional file 1: Figure S1. Linker length distributions utilized in the construction of chemical map–based and MNase-seq–based models. The linker length distributions of three species implemented in each package are

plotted. A magnified view ranging from 0 to 100 bp in linker length is presented inside each plot.

Additional file 1: Table S1. Area under the curve scores for the prediction results shown in Figure 1A.

Additional file 1: Figure S2. Matching rate between predicted and *in vivo* nucleosomes with variable matching windows. The species of genomic sequences used for prediction (Target) and the species used for model construction (Model) are indicated. “Sc” stands for *S. cerevisiae*, whereas “Sp” stands for *S. pombe*. Chemical map–based (nuCpos) and MNase-seq–based (NuPoP) models were used for predictions. Matching rates for representative (Unique) and redundant (Redundant) nucleosomes are presented.

Additional file 1: Figure S3. Magnified view of prediction results for *TRP1ARS1*. Prediction results for the budding yeast *TRP1ARS1* mini-chromosome output by nuCpos and NuPoP (**Figure 3C**) were magnified, being centered at nucleosome II.

Additional file 1: Figure S4. Prediction results for the fission yeast *ura4*$^{+}$ gene. Schematic representation of *in vivo* nucleosome positioning is shown above the plots. Nucleosomes numbered +1 to +6 are indicated as ovals. The top two panels show the prediction results output by nuCpos, whereas the next two panels show NuPoP results. The upper panel in each set shows predicted occupancy of nucleosomes (Occup., gray polygons) and probability that the tested 147-bp sequences are in the nucleosome state (P-dyad, blue vertical lines). Lower panels show HBA values for the tested 147-bp sequences calculated using the indicated models. The very bottom panel shows the A/T-frequency for the tested 147-bp sequences. Horizontal lines at the bottom of each plot indicate the 5’- and 3’-untranslated regions (colored in gray, nucleotide positions −151 to −1 and +795 to +986, respectively) and the protein-coding region (red, 0 to +794) of the *ura4*$^{+}$ gene. Inverted triangles indicate the nucleosome centers determined by MNase-seq [55].

Additional file 1: Figure S5. Prediction of the effects of repeat insertion in the fourth nucleosome of the *TALS* minichromosome. Chemical map–based HBA scores and predicted nucleosomal occupancy were calculated for original and modified *TALS* sequences [56]. The sequences were centered at the dyad position of the fourth nucleosome (0 bp), in which the indicated repeat sequences were inserted. Note: the position for the fourth nucleosome was predicted to shift to the right. Horizontal lines at the bottom of each plot indicate the *α2*-operator (purple) and the insert (red). hTEL stands for human telomeric repeat (5'-TTAGGG-3'); SI-A, sequence isomer-A (5'-TGTAGG-3'); SI-B, sequence isomer-B (5'-TGTGAG-3').

Additional file 1: Figure S6. Thirteen overlapping nucleosomal DNA subsegments for local HBA calculation. Segments A through M are colored in orange on the nucleosome structure 1AOI [1]. The structures were drawn using MacPyMOL (v1.8.4.1). The base pair length (red) and corresponding superhelical locations (black) of each segment are indicated. The color of the right-half segments (H-M) is faded because they are behind the front helix.

Additional file 1: Figure S7. Local HBA analysis of the *in vitro* reconstitution sequences. Local HBA scores at (red) and around (black) the *in vitro* nucleosome position for each segment are plotted in columns. Due to the lack of flanking sequences, shifting the translational position to the left and right sides inhibits local HBA calculations for terminal segments. Names of reconstituted nucleosomes are indicated to the left. Note that the 601LR sequence is not completely identical to the original 147-bp sequence centering at nucleotide position 154 (**Figure 5B**). The sequences of 601LR, 601L, and 601R differ at their termini for biochemical handling. For 601LR, the four original terminal bases (5'-CTGG-3' and 5'-CTGT-3') have been replaced with three-base sequences (5'-ATC-3' and 5'-GAT-3'; i.e., blunt ends created by *Eco*RV digestion) [12, 13]. Here, the nucleotide base A was added to both ends of the 145-bp sequence to lengthen it to 147-bp for local HBA calculations.

Additional file 1: Figure S8. Dinucleotide frequency across chemical unique nucleosomes. (**A**) Non-normalized frequencies of dinucleotides starting at indicated nucleosome positions. The heatmap scale indicates the dinucleotide frequency (log10). Note: AA-TT, AC-GT, AG-CT, CA-TG, CC-GG, and GA-TC dinucleotide pairs exhibit rotational symmetry to each other, whereas AT, CG, GC, and TA dinucleotides exhibit point symmetry by themselves. AA-TT, AC-GT, AG-CT, and TA dinucleotides are highly enriched at SHL ±0.5. (**B**) SS dinucleotide frequency. Gray vertical lines indicate the positions of dyad (solid) and histone-DNA contact sites (SHL ±0.5-6.5). Mean frequency of each dinucleotide in the unique nucleosomes is indicated to the right and shown as an orange horizontal bar in the plot.

Additional file 1: Figure S9. Matching rate analysis for the mouse model. Matching rate between predicted and *in vivo* nucleosomes with variable matching window is shown. The sequence of mouse Chr19 was used for prediction with the mouse chemical map–based (nuCpos) and MNase-seq–based (NuPoP) models.

Figure 1

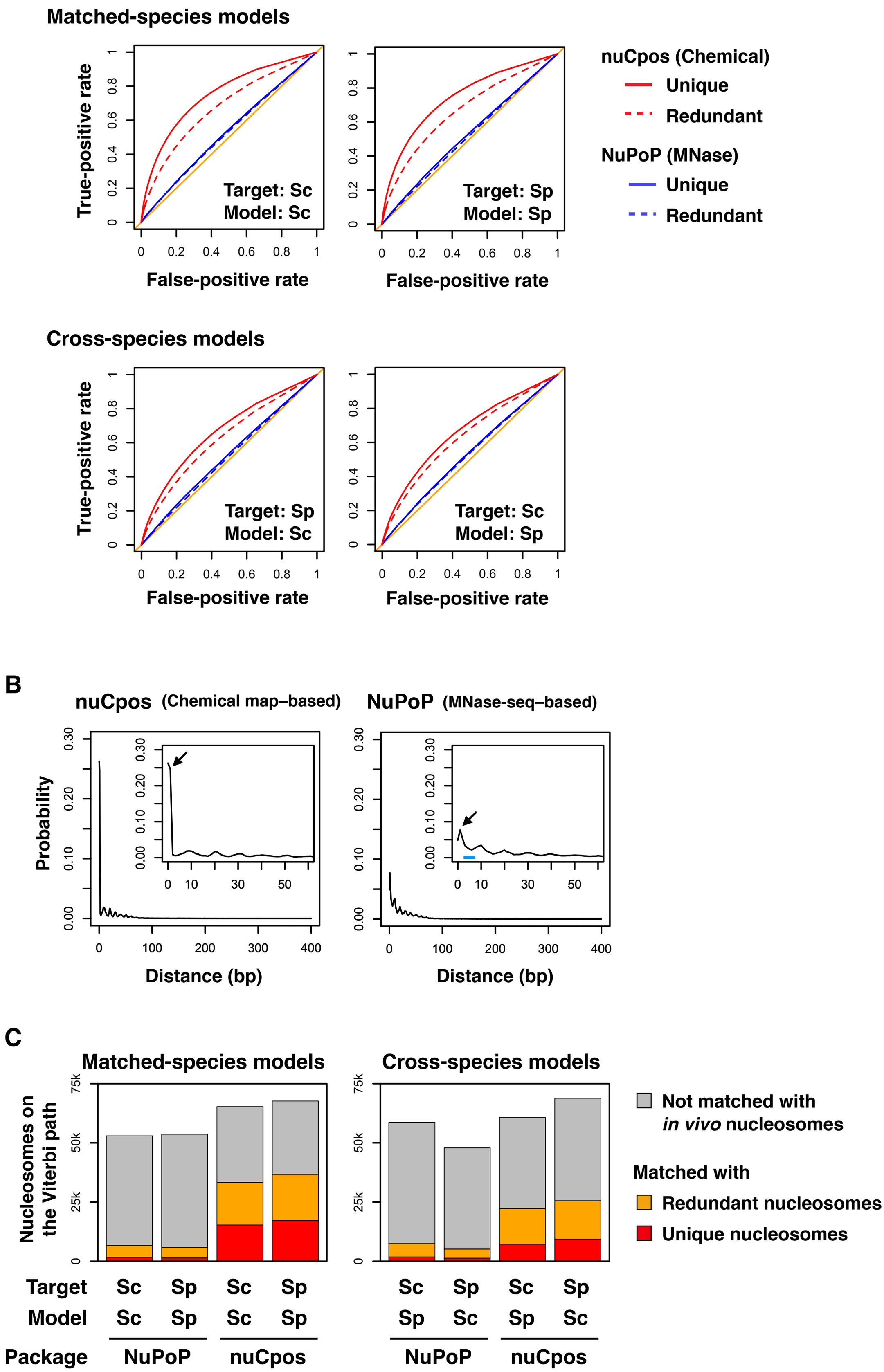

A
Matched-species models
True-positive rate
False-positive rate
Target: Sc
Model: Sc
Target: Sp
Model: Sp
nuCpos (Chemical)
Unique
Redundant
NuPoP (MNase)
Unique
Redundant
Cross-species models
Target: Sp
Model: Sc
Target: Sc
Model: Sp
B
nuCpos (Chemical map–based)
NuPoP (MNase-seq–based)
Probability
Distance (bp)
C
Matched-species models
Cross-species models
Nucleosomes on the Viterbi path
Not matched with in vivo nucleosomes
Matched with
Redundant nucleosomes
Unique nucleosomes
Target Sc Sp Sc Sp
Model Sc Sp Sc Sp
Package NuPoP nuCpos
Sc Sp Sc Sp
Sp Sc Sp Sc
NuPoP nuCpos

Figure 2

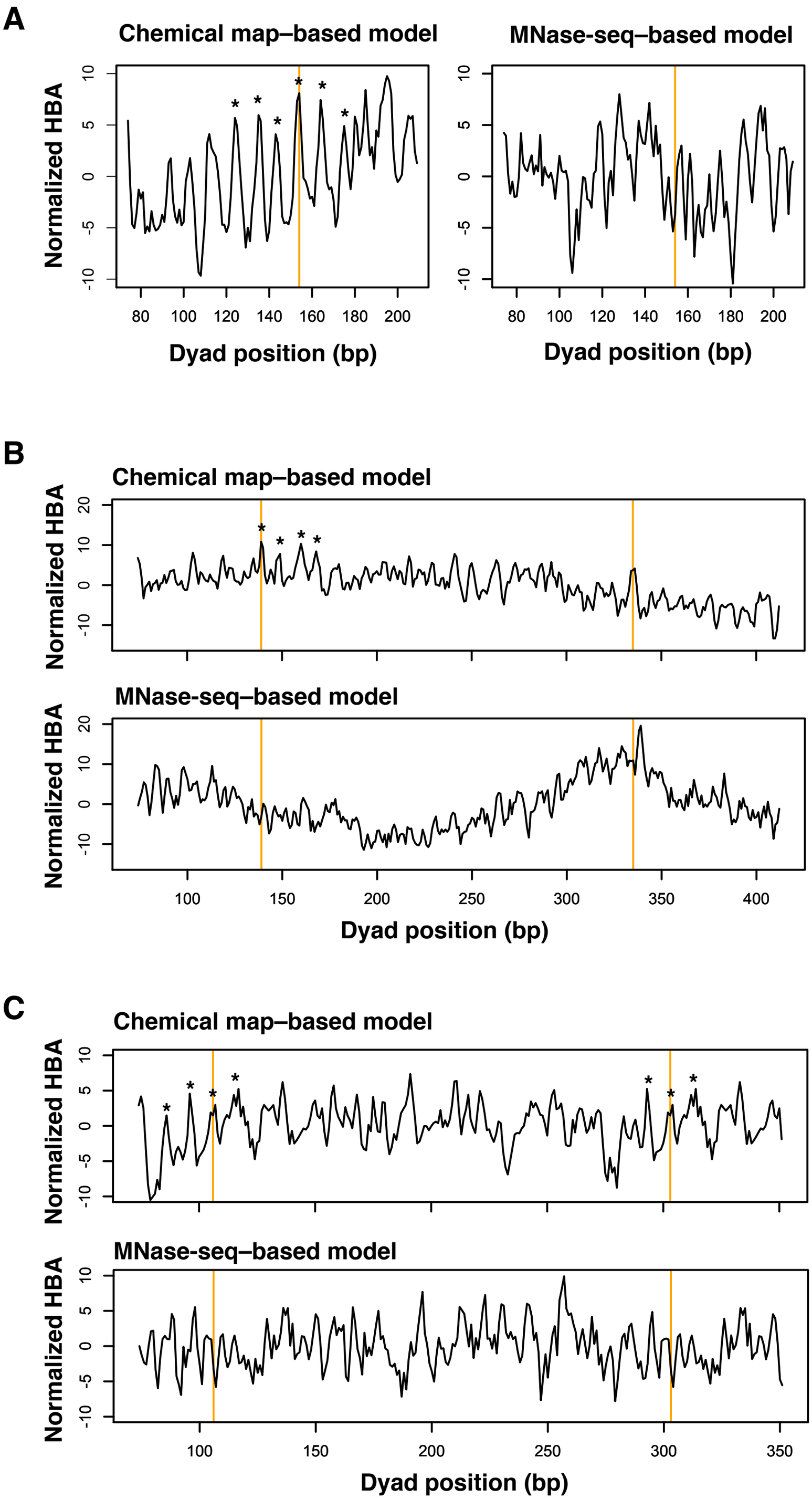

A
Chemical map–based model
MNase-seq–based model
Normalized HBA
Dyad position (bp)
B
Chemical map–based model
MNase-seq–based model
Normalized HBA
Dyad position (bp)
C
Chemical map–based model
MNase-seq–based model
Normalized HBA
Dyad position (bp)

Figure 3

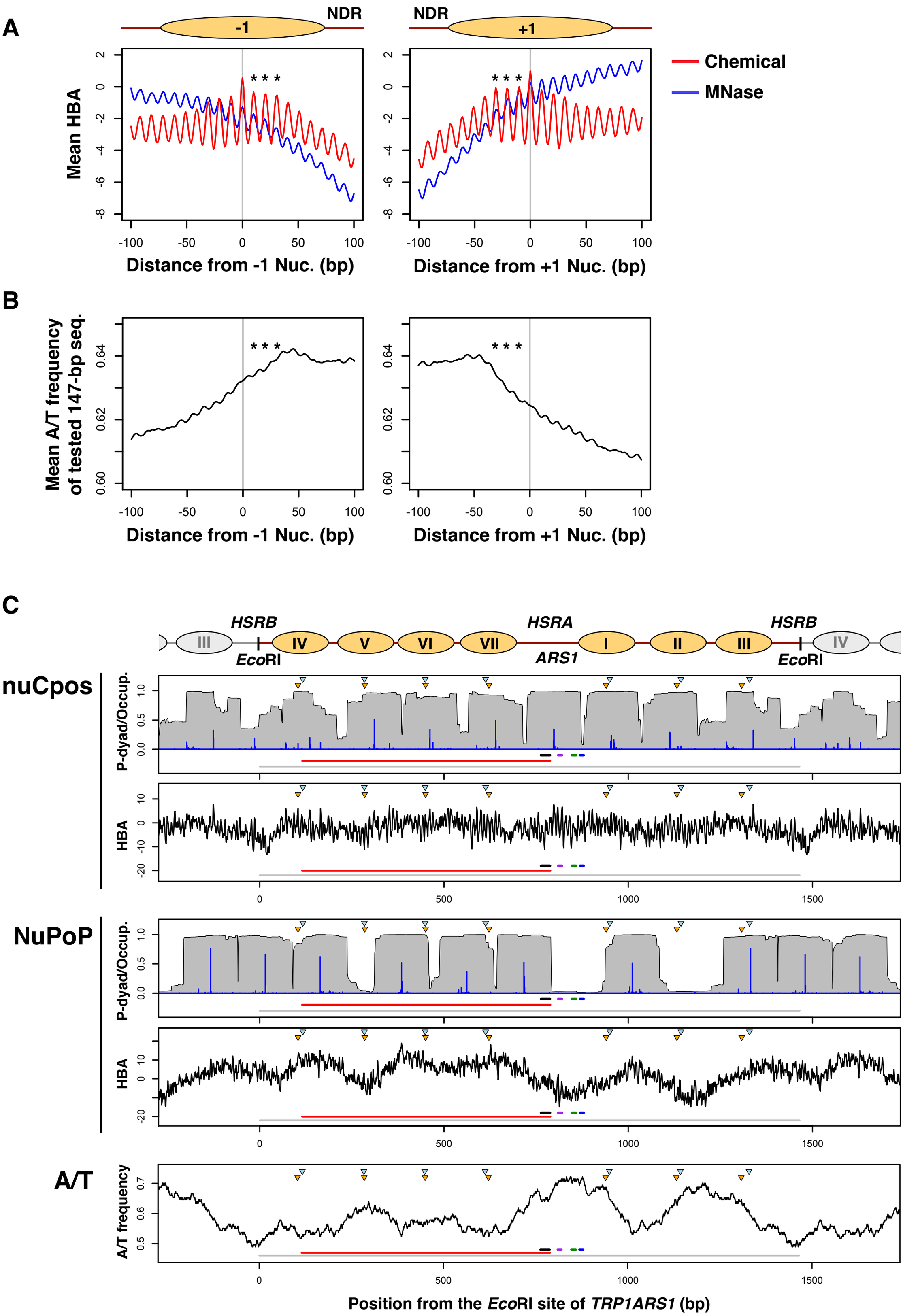

A
NDR
-1
NDR
+1
Mean HBA
Chemical
MNase
Distance from -1 Nuc. (bp)
Distance from +1 Nuc. (bp)
B
Mean A/T frequency of tested 147-bp seq.
Distance from -1 Nuc. (bp)
Distance from +1 Nuc. (bp)
C
HSRB
HSRA
HSRB
EcoRI
ARS1
EcoRI
nuCpos
NuPoP
A/T
P-dyad/Occup.
HBA
A/T frequency
Position from the EcoRI site of TRP1ARS1 (bp)

Figure 4

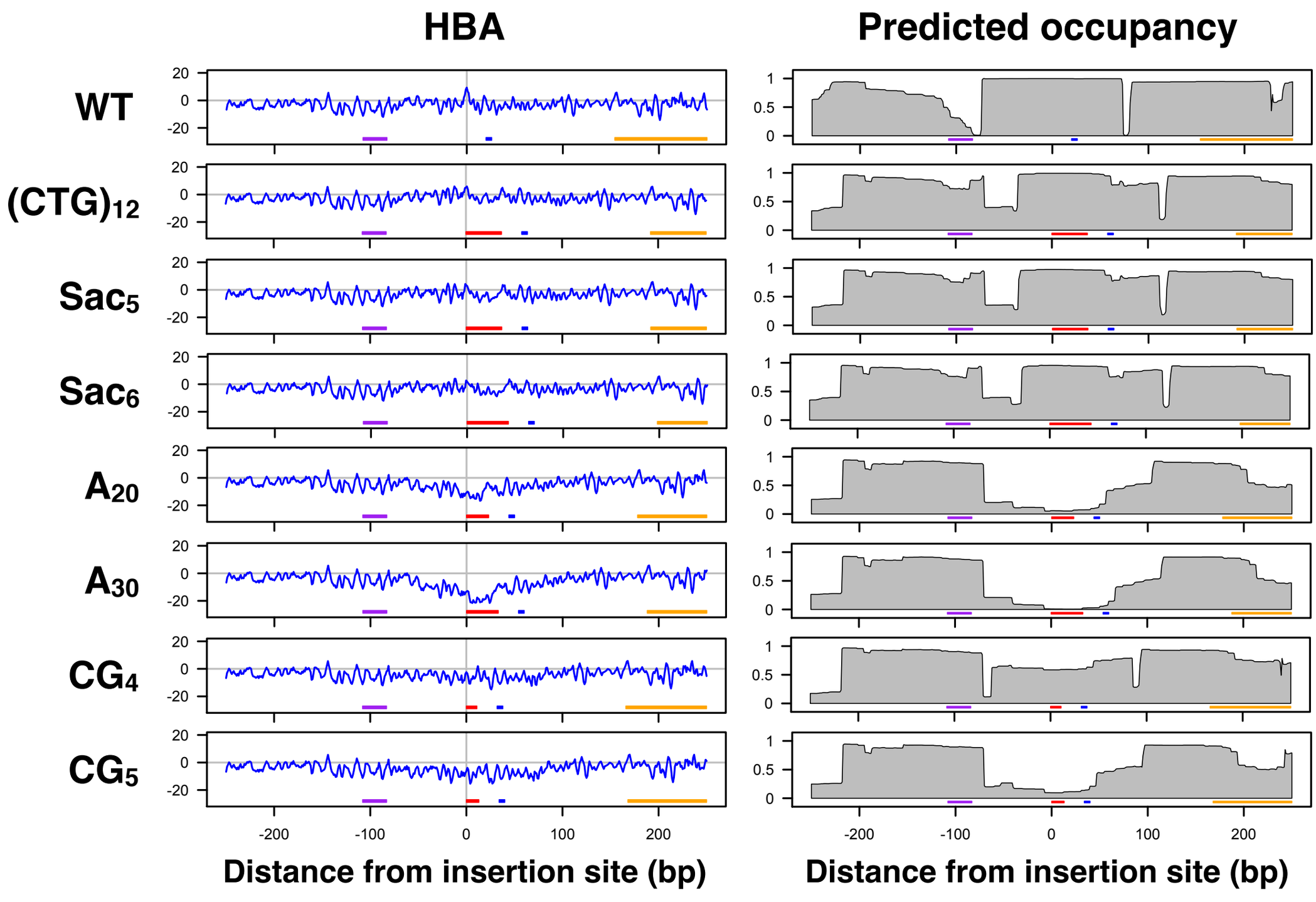

HBA
Predicted occupancy
WT
(CTG)12
Sac5
Sac6
A20
A30
CG4
CG5
20
0
-20
1
0.5
0
-200
-100
0
100
200
Distance from insertion site (bp)
Distance from insertion site (bp)

Figure 5

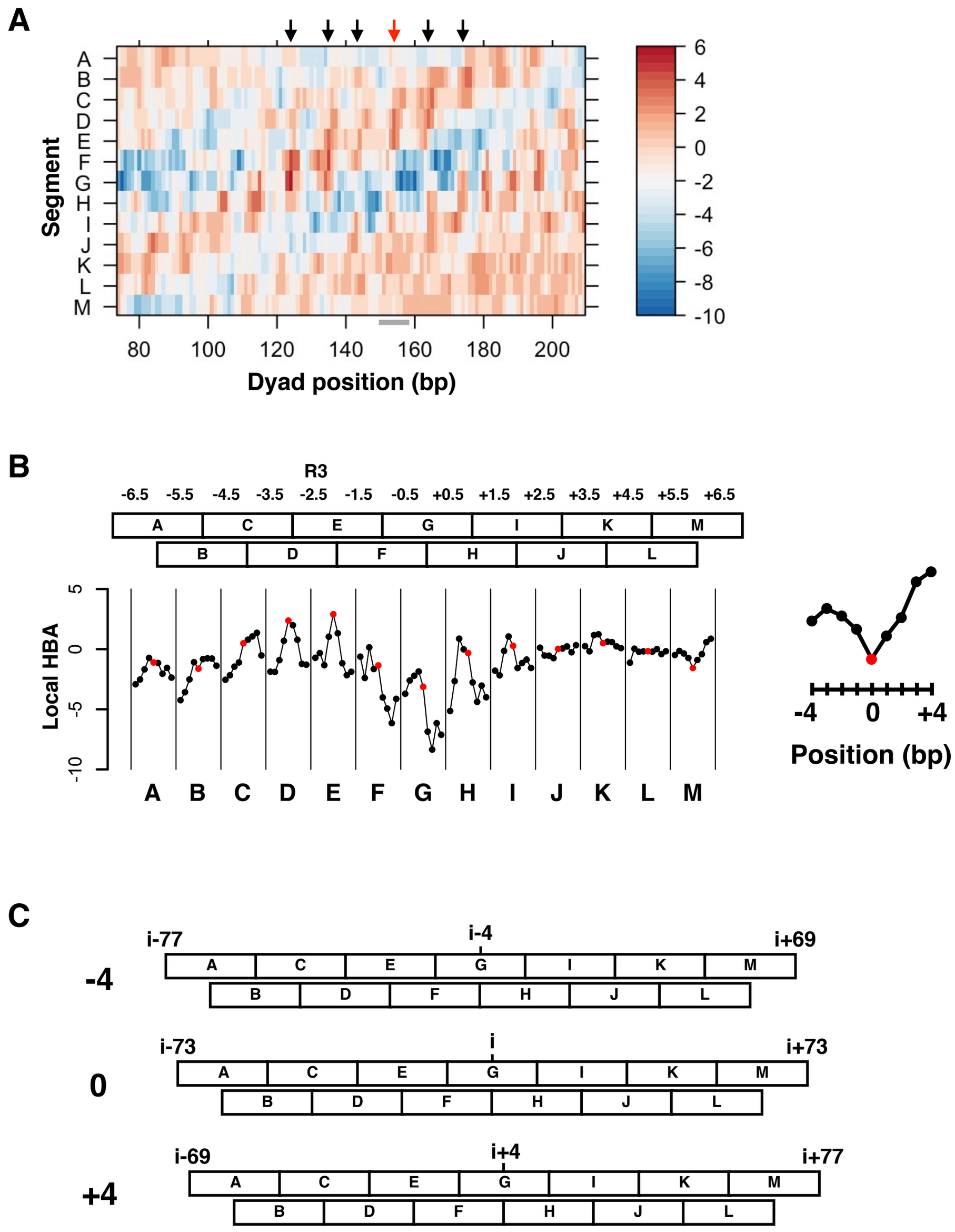
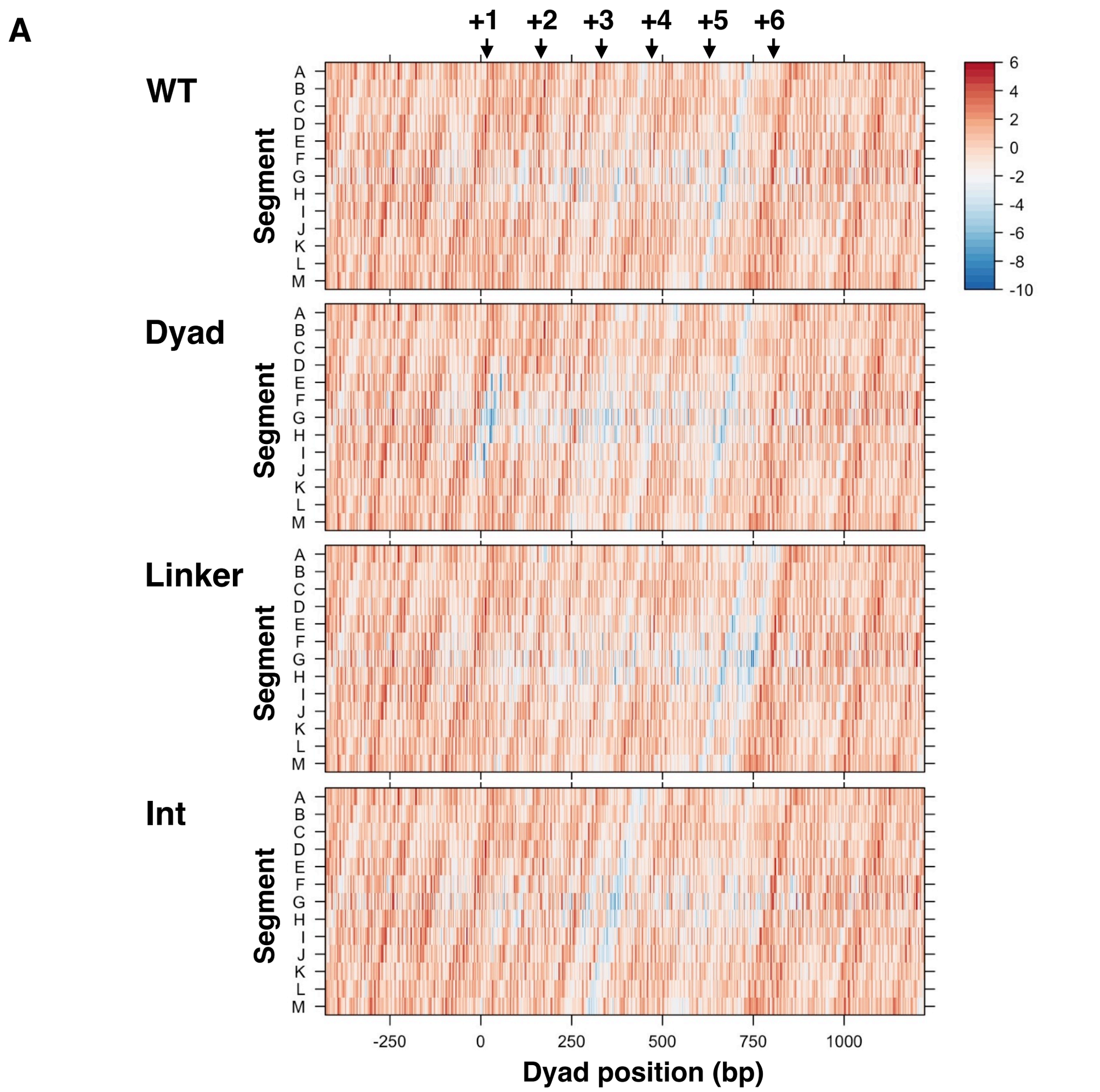

A
Segment
A
B
C
D
E
F
G
H
I
J
K
L
M
80
100
120
140
160
180
200
Dyad position (bp)
6
4
2
0
-2
-4
-6
-8
-10
B
R3
-6.5 -5.5 -4.5 -3.5 -2.5 -1.5 -0.5 +0.5 +1.5 +2.5 +3.5 +4.5 +5.5 +6.5
A C E G I K M
B D F H J L
Local HBA
5
0
-5
-10
A B C D E F G H I J K L M
-4
0
+4
Position (bp)
C
i-77
i-4
i+69
-4
A C E G I K M
B D F H J L
i-73
i
i+73
0
A C E G I K M
B D F H J L
i-69
i+4
i+77
+4
A C E G I K M
B D F H J L

Figure 6


A
+1
+2
+3
+4
+5
+6
WT
Dyad
Linker
Int
Segment
A B C D E F G H I J K L M
6 4 2 0 -2 -4 -6 -8 -10
-250 0 250 500 750 1000
Dyad position (bp)

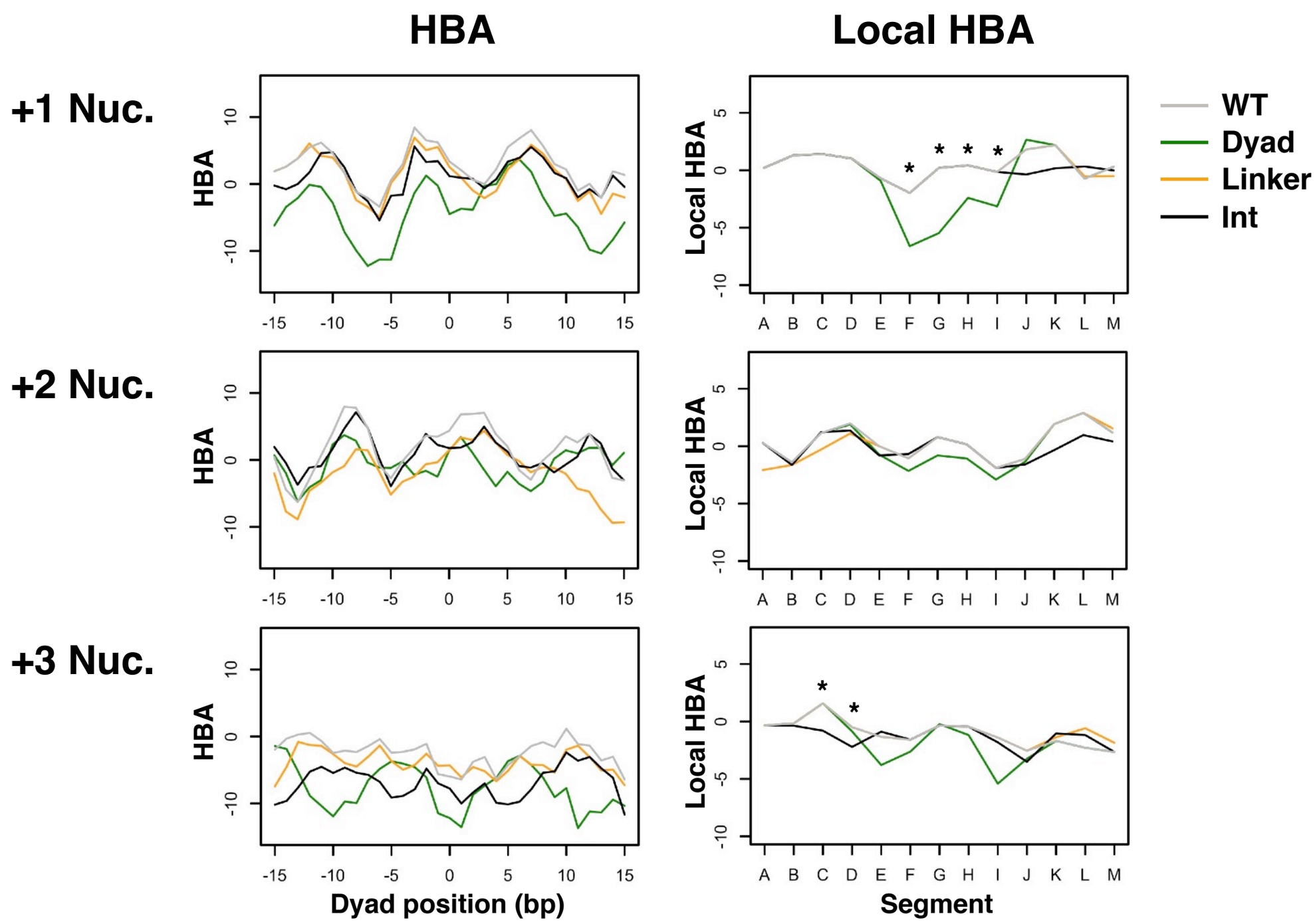
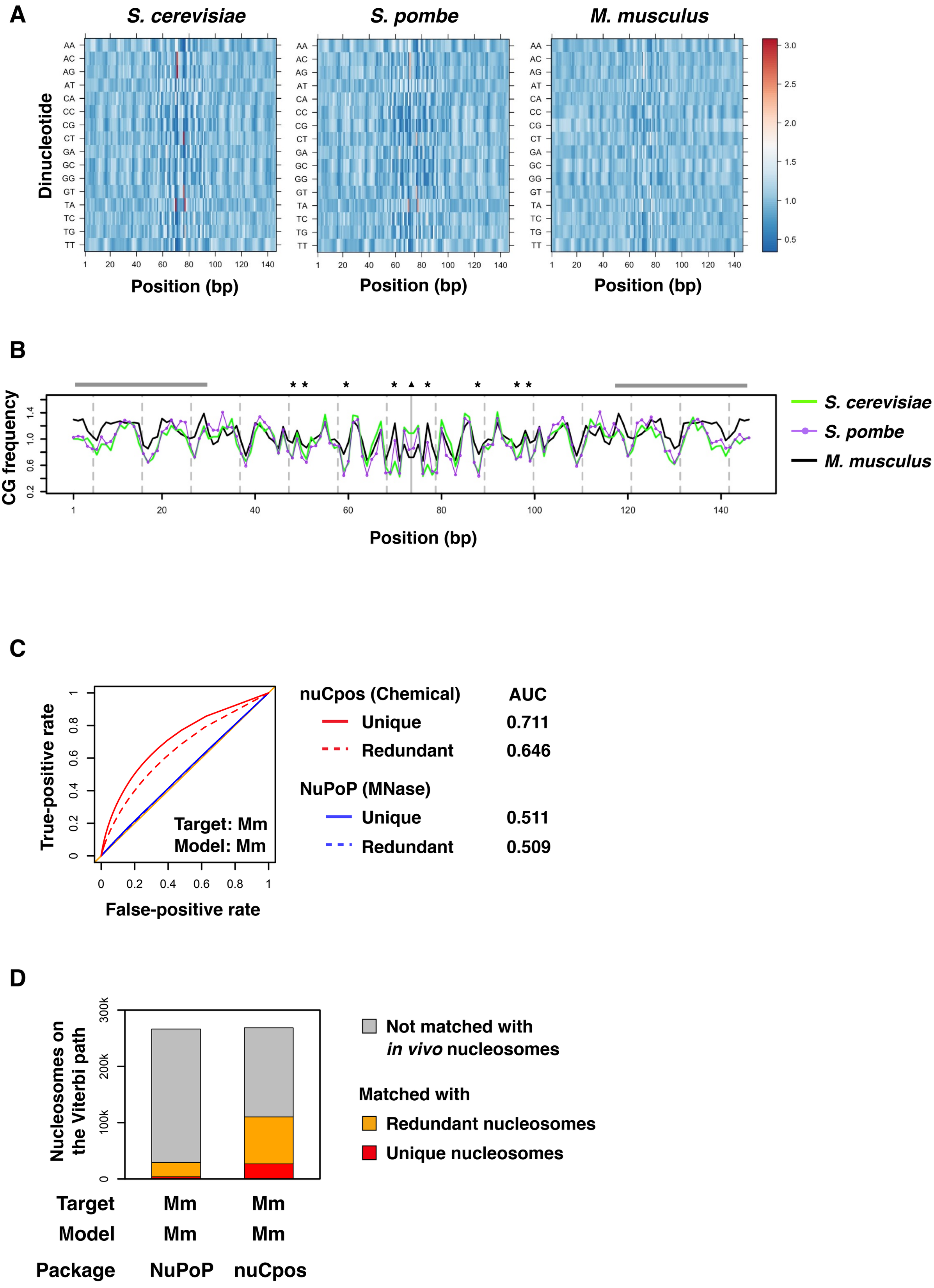

B
HBA
Local HBA
+1 Nuc.
+2 Nuc.
+3 Nuc.
HBA
Local HBA
-15 -10 -5 0 5 10 15
A B C D E F G H I J K L M
WT
Dyad
Linker
Int
Dyad position (bp)
Segment

Figure 7


A
S. cerevisiae
S. pombe
M. musculus
Dinucleotide
AA
AC
AG
AT
CA
CC
CG
CT
GA
GC
GG
GT
TA
TC
TG
TT
Position (bp)
3.0
2.5
2.0
1.5
1.0
0.5
B
Normalized CG frequency
Position (bp)
S. cerevisiae
S. pombe
M. musculus
C
True-positive rate
False-positive rate
Target: Mm
Model: Mm
nuCpos (Chemical)
AUC
Unique
0.711
Redundant
0.646
NuPoP (MNase)
Unique
0.511
Redundant
0.509
D
Nucleosomes on the Viterbi path
Not matched with in vivo nucleosomes
Matched with
Redundant nucleosomes
Unique nucleosomes
Target
Mm
Mm
Model
Mm
Mm
Package
NuPoP
nuCpos