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

Email addresses:

Hiroaki Kato: hkato@med.shimane-u.ac.jp
Mitsuhiro Shimizu: shimizum@chem.meisei-u.ac.jp
Takeshi Urano: turano@med.shimane-u.ac.jp

Corresponding author:

Hiroaki Kato
Department of Biochemistry, Shimane University School of Medicine
89-1 Enya-cho, Izumo 693-8501, Japan
Tel: +81-853-20-2127; Fax: +81-853-20-2125
hkato@med.shimane-u.ac.jp

Abstract [350 words]

Background:

Assessing the nucleosome-forming potential of specific DNA sequences is important for understanding complex chromatin organization. Methods for predicting nucleosome positioning include bioinformatics and biophysical 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. Nucleosome maps are constructed using micrococcal nuclease digestion followed by high-throughput sequencing (MNase-seq). However, as MNase has a strong preference for A/T-rich sequences, MNase-seq 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 mice) are currently available. However, these chemical maps have yet to be integrated into publicly available computational methods.

Results:

We developed a Bioconductor package (named nuCpos) to demonstrate the superiority of chemical maps in predicting nucleosome positioning. The accuracy of chemical map–based prediction in rotational settings 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 cells could also be predicted. nuCpos calculates individual histone binding affinity (HBA) scores for given 147-bp sequences to examine their suitability for nucleosome formation. We also established local HBA as a new parameter to predict nucleosome formation, which was calculated for 13 overlapping nucleosomal DNA subsequences. 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.

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 is defined by rotational and translational settings [4, 5]. Rotational settings define the histone-DNA contact surfaces on the DNA string, whereas translational position is the distinct position of the nucleosome (i.e., position of the dyad base) with respect to the genomic coordinate. The positioning of nucleosomes along genomic DNA is critical for various biological events, including transcription, DNA replication and repair, and chromosome segregation [4]. In these events, nucleosome positioning is combinatorially regulated by *trans*-acting factors (e.g., transcription factors and chromatin remodeling factors) and by the inherent suitability of the DNA sequence for nucleosome formation [5-7]. Focusing on the latter requirement, computational methods that assess the nucleosome-forming potential of given DNA sequences have been developed [8].

These methods are classified into two groups: biophysical and bioinformatics [8, 9]. 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, 10, 11]. The use of the 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.

98 In contrast to the biophysical prediction methods, bioinformatics
99 methods are based on maps of *in vivo* nucleosomes determined using
100 high-throughput sequencing techniques. The use of complex, natural
101 nucleosomal sequences may cause difficulty in the interpretation of prediction
102 results; however, it may also lead to unexpected findings that enhance
103 understanding of nucleosome-based regulation.
104 A weakness of currently available bioinformatics methods is that the
105 statistical models are constructed based on *in vivo* nucleosome maps generated
106 using micrococcal nuclease (MNase) digestion followed by high-throughput
107 sequencing (MNase-seq). MNase digests linker DNA that connects NCPs,
108 allowing researchers to obtain nucleosomal DNA that has been protected from
109 enzymatic digestion. In MNase-seq experiments, DNA fragments of
110 mono-nucleosome size (~150 bp) are sequenced and mapped to the reference
111 genome. However, as MNase strongly prefers A/T-rich sequences, unwanted
112 digestion of A/T-rich nucleosomal sequences and less-effective digestion of
113 G/C-rich linkers may occur, leading to variation in nucleosome fragment size [7,
114 12]. Because of this enzymatic bias, precise determination of the genomic
115 coordinates of *in vivo* nucleosomes using MNase-seq (i.e., dyad base calling) is
116 challenging. In addition, although some correction methods for MNase-seq
117 coverage against enzymatic bias have been developed [13-15], no method that
118 corrects the deduced nucleosome position by shifting it to an appropriate
119 rotational position is available. Thus, the resolution of MNase-seq–based maps
120 with respect to the genomic coordinates is expected to be inevitably low.
121 One bioinformatics method considered successful in a thorough
122 comparative study [9] is NuPoP, which is built upon a duration hidden Markov
123 model (dHMM) that considers information regarding both the nucleosome and
124 linker DNA [16]. NuPoP enables the prediction of nucleosome positioning in nine
125 different species by rescaling the base composition of both the nucleosome and
126 linker models for the budding yeast *Saccharomyces cerevisiae* with those of
127 target organisms. The budding yeast models are generated using an
128 MNase-seq–based nucleosome map by considering the central base of each
129 nucleosomal fragment of varying length the dyad base [16]. This may particularly
130 affect construction of the time-dependent statistical model for the nucleosome

131 state, which ideally requires that the dyad bases be determined at bp resolution.
132 The construction of the linker model, which requires extraction of linker
133 sequences as genomic sequences not covered by nucleosomes, may also be
134 affected. Thus, although selecting the MNase-seq–based map was the best
135 practice at the time of the software's development, the prediction results output
136 provided by NuPoP can be biased by both the enzymatic preference of MNase
137 and the difficulty of dyad base calling.

138 A breakthrough in nucleosome mapping occurred with the development
139 of a method for site-directed chemical cleavage of nucleosomal DNA followed by
140 high-throughput sequencing [17, 18]. This technique, called chemical mapping,
141 was first applied to budding yeast and subsequently to the fission yeast
142 *Schizosaccharomyces pombe* and embryonic stem cells of the house mouse,
143 *Mus musculus* [17, 19-22]. In cells expressing histone H4 with an S47C amino
144 acid substitution, nucleosomal DNA is specifically cleaved near the nucleosome
145 center by site-directed chemical cleavage. By examining the cleavage sites,
146 previous studies determined the dyad positions of *in vivo* nucleosomes in the
147 three species at bp resolution [17, 19, 22]. The published maps contained whole
148 and representative sets of identified nucleosomes, which are referred to as
149 "redundant" and "unique" nucleosomes, respectively [17, 19, 22]. In addition to
150 the histone H4-S47C approach, chemical mapping of nucleosomes using a
151 histone H3 mutation (Q85C) determined the precise positions of −1 and +1
152 nucleosomes for protein-coding genes [12]. However, the bp-resolution chemical
153 maps have not yet been implemented in publicly available software for prediction
154 of nucleosome positioning.

155 One goal of nucleosome positioning prediction is *in silico* reproduction
156 of *in vivo* observations, preferably at the locus level. If this goal is achieved,
157 predicting the effects of genetic alterations on nucleosome positioning and
158 prediction-based engineering of nucleosome-forming and -depleted sequences
159 would also be practical. In this regard, budding and fission yeasts would serve
160 as preferable model organisms, because nucleosomes formed on native and
161 modified sequences *in vivo* have been precisely mapped [23-28]. These
162 nucleosomes are suitable for *in silico* reproduction of *in vivo* observations.
163 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 [29-38].

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 with respect to the genomic coordinates 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 generated with H4-S47C–dependent DNA cleavage (see **Methods** and **Table 1**). The core dHMM algorithm implemented in our method (nuCpos) was not altered from the original algorithm developed for the MNase-seq–based method (NuPoP) to clarify the advantages of using the chemical maps.

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. Aiming to evaluate the prediction accuracy with respect to the genomic coordinates, prediction results were assessed by checking the matching status of predicted nucleosomes with those experimentally determined by means of histone H4-S47C–based chemical cleavage [17, 19]. Here, MNase-seq–based maps are inappropriate to use as true-position references due to their lower resolution, as described in the Background section. Ideally, prediction results should be evaluated at 1-bp resolution, so that it will be possible to accurately discern the

197 rotational settings of given nucleosomes. As the resolution of the chemical maps
198 that we used for model construction could be 1-2 bp or lower, the resolution of
199 the prediction results would be expected to be at a similar level. Indeed, the
200 distance between chemically predicted nucleosomes and the nearest *in vivo*
201 nucleosomes was generally 0 bp (located at the same position) or 1 bp (located
202 at the juxtaposed nucleotide positions) (**Figure 1A**, right panel).
203 In order to determine the appropriate resolution for examining
204 prediction results, we calculated the matching rate between predicted and *in vivo*
205 nucleosomes with stepwise widening of the matching window (**Figure 1B**). Here,
206 we defined the matching window as the allowable distance between
207 nucleosomes regarded as matched. In chemical map–based predictions
208 (nuCpos), the matching rate doubled to over 50% when the matching window
209 was widened from 0 bp (1-bp resolution) to 1 bp (2-bp resolution) (**Figure 1B**,
210 red lines). However, the rate did not markedly increase upon further widening of
211 the matching window, suggesting that the chemical models distinguish the
212 histone-facing side of the DNA string.
213 In MNase-seq–based predictions (NuPoP), by contrast, the matching
214 rate increased gradually as the matching window was widened from 0 to 9 bp
215 (**Figure 1B**, blue lines). The matching rate drew a very loose anti-S shape,
216 suggesting that the MNase-seq–based model still recognized the histone-facing
217 side of the DNA string, but only to a small extent. We also noticed that a
218 substantial population of nucleosomes predicted with the MNase-seq–based
219 model was positioned approximately 5 nucleotides away from the *in vivo*
220 nucleosomes (**Figure 1A**, left panel, marked with a blue horizontal bar). These
221 observations suggest that the MNase-seq–based model has difficulty in
222 discriminating rotationally mispredicted nucleosomes. Considering these results,
223 we concluded that prediction results should be evaluated at 2-bp resolution, as
224 further widening of the matching window would increase the misprediction rate.
225 **Figure 1C** presents the evaluation of prediction results against
226 representative “unique” and whole “redundant” nucleosomes [17, 19] at 2-bp
227 resolution. When the budding yeast MNase-seq–based model (NuPoP) was
228 applied to the prediction of nucleosome positioning in the budding yeast genome
229 (Target: Sc, Model: Sc), only 3.2% and 12.6% of the nucleosomes on the most

230 probable (Viterbi) path matched *in vivo* unique (red) and redundant (orange)
231 nucleosomes, respectively. When the fission yeast model was applied to the
232 prediction of nucleosome positioning in the fission yeast genome (Target: Sp,
233 Model: Sp), the MNase-seq–based model scored 2.7% and 11.0%, respectively.
234 In contrast, when the budding yeast chemical map–based model (nuCpos) was
235 used for predictions (Target: Sc, Model: Sc), 23.5% and 50.9% of the
236 nucleosomes on the Viterbi path matched unique and redundant nucleosomes,
237 respectively. For the fission yeast (Target: Sp, Model: Sp), the corresponding
238 scores were 25.5% and 54.2%, respectively. Thus, with the chemical map–
239 based models, over half of the nucleosomes predicted as being on the Viterbi
240 path are present in living cells at exactly the same positions or at the juxtaposed
241 nucleotide positions.
242 In the chemical maps of budding and fission yeasts, *in vivo* unique
243 nucleosomes accounted for 19.6% and 17.8%, respectively, of all redundant
244 nucleosomes [17, 19]. Thus, in the MNase-seq–based predictions, enrichment
245 of unique nucleosomes on the Viterbi paths was only 1.3- and 1.4-fold,
246 respectively. In contrast, the chemical map–based predictions exhibited 2.4- and
247 2.6-fold enrichment, respectively. The higher enrichments suggest that chemical
248 map–based models better distinguish sequences favorable for nucleosome
249 formation [17, 19], compared with MNase-seq–based models.
250 Since the dHMMs were constructed using nucleosome and linker
251 information based solely on the unique nucleosomes (see **Methods**), detection
252 of portion of the redundant nucleosomes (from which unique nucleosomes have
253 been removed) is worthwhile. The nucleosomes in this category accounted for
254 27.4% and 28.7% of the Viterbi nucleosomes in the chemical-map–based
255 predictions for the budding and fission yeasts, respectively, and the proportions
256 were higher than the MNase-seq–based predictions (9.4% and 8.3%,
257 respectively). Giving these results, the chemical map–based models provide
258 better prediction outcomes with respect to the genomic coordinates than the
259 MNase-seq–based models.
260
261 **Prediction of *in vitro* reconstituted nucleosome positions**

262 The evaluations described above suggested that chemical map–based models
263 are superior in terms of the prediction of rotational settings. To confirm this
264 observation, we compared HBA scores [16] calculated along the original 282-bp
265 Widom 601 sequence [10] with the chemical map–based and MNase-seq–
266 based models (**Figure 2**). HBA scores are calculated in the dHMM algorithms in
267 order to examine the suitability of given 147-bp sequences for nucleosome
268 formation [16], and they are expected to predict the rotational setting of
269 nucleosomes if the resolution of the nucleosome map used for model
270 construction is sufficiently high. The original 282-bp Widom 601 sequence
271 contained primer sequences for the SELEX study at both ends and additional
272 sequences flanking the nucleosome-forming 147-bp sequence centered at
273 nucleotide position 154 [10]. When analyzed using the chemical map–based
274 model, discrete HBA peaks with an interval of approximately 10 bp were
275 observed at positions 124, 135, 143, 154, 164, and 175 (asterisks). This
276 periodicity indicated that nearly the same surface of the DNA string was
277 predicted to be in contact with the histone core in these potential translational
278 positions. Among these positions, the one with the highest HBA score
279 corresponds to the translational position of the *in vitro* stable nucleosome
280 (position 154, orange vertical lines). In contrast, the MNase-seq–based HBA
281 score was relatively low at position 154, and the peaks did not exhibit the
282 apparent 10-bp periodicity. Analyses of the mouse mammary tumor virus
283 (MMTV) 3'–long terminal repeat (LTR) sequence and *Xenopus borealis* 5S RNA
284 gene also showed the advantage of the chemical model (**Additional file 1:**
285 **Figure S1**). Collectively, the chemical map–based model reliably predicted the
286 rotational settings of the nucleosomes reconstituted with these suitable
287 sequences.
288
289 **Prediction of *in vivo* nucleosome positions**
290 Next, we calculated HBA scores around −1 and +1 nucleosomes for 5,542
291 protein-coding genes identified by means of histone H3-Q85C–based chemical
292 cleavage [12]. Due to the differences in detection methods, positions of
293 H3-Q85C–based nucleosomes are not always identical to those of the
294 corresponding H4-S47C–based nucleosomes [12]. In this regard, the H3-Q85C–

295 based map differed from the H4-S47C–based map that was used for the
296 statistical model construction. Thus, the H3-Q85C–based nucleosomes would
297 be a good reference for evaluation of prediction accuracy at bp-resolution.
298 As shown in **Figure 3A**, mean HBA scores calculated using both
299 models exhibited a periodicity with a 10-bp interval. However, the amplitude of
300 the chemical map–based HBA scores (red) around the tested nucleosomes was
301 larger than that of the MNase-seq–based HBA scores (blue). This suggested
302 that the chemical model could clearly distinguish the surface of the DNA string
303 that interacts with the histone core even when natural genomic sequences were
304 queried. We found that the chemical map–based HBA scores were highest at
305 the −1 and +1 nucleosome positions (0 bp). Chemical map–based HBA scores
306 for the three neighboring translational positions (10, 20, and 30 bp in distance)
307 relative to the nucleosome-depleted regions (marked with asterisks in **Figure**
308 **3A**) remained relatively high despite the gradual increase in the A/T frequency of
309 the tested sequences (**Figure 3B**). After the A/T frequency reached a plateau
310 (A/T≈0.64), the chemical map–based HBA scores began to decrease. In
311 contrast, MNase-seq–based HBA scores at the −1 and +1 nucleosome positions
312 (0 bp) were already at moderate levels, and they simply decreased as the A/T
313 frequency increased. Thus, the MNase-seq–based model was adversely
314 influenced by the enzymatic bias of MNase, whereas the chemical map–based
315 model was not.
316 The higher accuracy of chemical map–based models at the genomic
317 level prompted us to perform a locus-level evaluation of prediction results, using
318 the budding yeast *TRP1ARS1* minichromosome [25, 27]. On this well-studied,
319 circular, 1,465-bp DNA, three (numbered I to III) and four (IV to VII)
320 nucleosomes are interspersed with two nucleosome-depleted regions or
321 nuclease-hypersensitive regions designated *HSRA* and *HSRB*. *HSRA* includes
322 the DNA replication origin *ARS1*, whereas *HSRB* is located upstream of the
323 *TRP1* marker gene. At a glance, the chemical map– and MNase-seq–based
324 model prediction results appeared very different (**Figure 3C**). Each model
325 suggested eight and seven representative nucleosomes, respectively. Seven of
326 the eight nucleosomes predicted by the chemical map–based model (**Figure 3C**,
327 nuCpos) reasonably matched the positions of *in vivo* nucleosomes determined

328 previously [27]. The only exception was a predicted nucleosome located in
329 *HSRA*, in which functional *ARS1* would be occupied by the ARS-binding factor
330 Abf1 and the origin recognition complex instead of the nucleosome to maintain
331 this circular DNA across generations [39]. This prediction was reasonable, as
332 DNA replication origins are known to be covered by nucleosomes when the
333 subunits of the origin recognition complex or Abf1 are perturbed *in vivo* [39-41].
334 In this native sequence, the chemical map–based HBA scores
335 exhibited a periodicity of an approximately 10-bp interval (**Additional file 1:**
336 **Figure S2**, nuCpos). As the dHMM considers the linker length distribution as a
337 major factor affecting nucleosome-linker state transition and does not allow
338 nucleosomes with shorter DNAs (<147 bp), nucleosome positions tend to be
339 selected in a limited manner. Although there were many high-HBA positions,
340 only a few were clearly selected by the dHMM as high-probability nucleosome
341 dyad positions. Thus, the predicted occupancy should be cautiously considered,
342 as it appears to provide only a rough image of chromatin state. Instead, it should
343 be kept in mind that periodic HBA scores may contribute to redundant
344 nucleosome positioning *in vivo*.
345 In contrast to the chemical map–based model, most of the nucleosome
346 locations predicted by the MNase-seq–based model were not in agreement with
347 their positions *in vivo* (**Figure 3C**, NuPoP); most of the *in vivo* nucleosome
348 centers (inverted triangles in figure) were located in the predicted linkers or
349 nucleosome ends. This result was consistent with the lower genome-wide
350 prediction accuracy (**Figure 1**). Importantly, MNase-seq–based HBA scores for
351 this sequence appeared to mirror the A/T-content over the X-axis (**Figure 3C**,
352 A/T); high-HBA positions were generally A/T-poor, and low-HBA positions were
353 A/T-rich. In addition, the local amplitude of MNase-seq–based HBA scores was
354 apparently lower compared with the chemical map–based model (**Additional**
355 **file 1: Figure S2**). These observations also demonstrated that MNase-seq–
356 based model prediction results are strongly influenced by the A/T preference of
357 MNase. A/T-content exhibited a similar effect in the MNase-seq–based
358 nucleosome positioning prediction around the fission yeast *ura4*$^{+}$ gene
359 (**Additional file 1: Figure S3**) [42]. Given these results, we concluded that the

360 chemical map–based models output better genome-wide and locus-level
361 prediction results.
362
363 **Prediction of effects of genetic alterations on nucleosome positioning**
364 Next, we examined whether the effects of genetic alterations on nucleosome
365 positioning could be predicted. In **α** cells of budding yeast, the **a**-cell-specific
366 gene *BAR1* is repressed by *α2*-repressor–dependent positioning of the
367 nucleosome at its promoter, which masks the gene's *TATA* box [23, 26]. The
368 chemical map–based model predicted the positioning of this repressive
369 nucleosome with high accuracy (**Figure 4**, WT). The HBA score for this
370 translational position (0 bp), which is 82 bp away from the *α2*-operator, was very
371 high (HBA=8.47), suggesting that repression of the *BAR1* gene is assisted by
372 the intrinsic suitability of the promoter DNA for nucleosome formation. When a
373 36-bp sequence consisting of 12 repeats of CTG is inserted at the center of this
374 nucleosome, repression of *BAR1* and nucleosome positioning are not affected *in*
375 *vivo* [26]. Prediction results agreed that this insertion should not cause
376 nucleosome depletion (**Figure 4**, CTG_{12}). Other repeat sequences causing no
377 transcriptional derepression *in vivo* [26] were also predicted to not cause
378 nucleosome depletion (**Figure 4**, Sac_5 and Sac_6). It is also known that insertion
379 of a 30-bp A-stretch or a 10-bp CG (i.e., CpG) repeat sequence inhibits *in vivo*
380 nucleosome formation and causes derepression of *BAR1* [26]. Remarkably, the
381 effect of these genetic alterations on nucleosome positioning was reproduced *in*
382 *silico* with the chemical map–based model (**Figure 4**, A_{30} and CG_5). We noted
383 that the disruptive effect of fragment insertion on nucleosome positioning was
384 more pronounced in predictions than *in vivo* observations. Insertion of a 20-bp
385 A-stretch, which does not cause complete nucleosome depletion *in vivo*, was
386 clearly predicted to inhibit nucleosome formation (**Figure 4**, A_{20}). However, in the
387 case of CG repeat insertion, shortening of the repeat to 8 bp, which satisfies *in*
388 *vivo* nucleosome formation [26], was predicted to preserve the capability to form
389 nucleosomes (**Figure 4**, CG_4). We also confirmed that the inhibition of
390 nucleosome formation by telomeric DNA insertion demonstrated in a previous
391 study [43] was also largely predictable (**Additional file 1: Figure S4**). Thus,

392 overall, the effects of genetic alterations on nucleosome positioning can be
393 predicted with the chemical map–based model.
394
395 **Local HBA for nucleosomal DNA subsegments**
396 The successful increase in prediction accuracy and better discrimination of
397 rotationally favorable nucleosome positions with chemical map–based models
398 suggested that the chemical map–based HBA score for a given 147-bp
399 sequence is a good indicator of the suitability of that sequence for nucleosome
400 formation. Considering that nucleosomal DNA makes contact with histone
401 proteins at each superhelical location (SHL ±0.5-6.5), we defined a novel
402 parameter, designated "local HBA", to examine the suitability of
403 intra-nucleosomal DNA segments for each histone-DNA contact. In order to
404 implement this idea, a 147-bp whole nucleosomal segment was divided into 13
405 overlapping nucleosomal DNA subsegments, designated A through M
406 (**Additional file 1: Figure S5**, see **Methods**). Each 20- or 21-bp segment
407 corresponded to two histone-DNA contact sites, each of which was shared by
408 neighboring segments. For example, segment B corresponded to SHL −5.5 and
409 −4.5, which were shared by segments A and C, respectively. We took this
410 overlapping approach because it remains unclear how surrounding sequences
411 affect histone-DNA contacts. Calculating local HBA scores is conceptually the
412 same as that of HBA calculations, which consider the probability of a whole
413 147-bp sequence to be a nucleosome [16]. In local HBA calculation, for each
414 intra-nucleosomal segment of a given 147-bp DNA that wraps a histone octamer
415 to form a nucleosome, the probability of the selected 20- or 21-bp sequence to
416 be located at that part of the nucleosome and the probability of the same
417 sequence functioning as a linker are calculated. The log likelihood ratio of these
418 probabilities is defined as the local HBA score for that segment. Thus, the sum of
419 local HBA scores for the non-overlapping set of seven 21-bp segments, A, C, E,
420 G, I, K, and M (147 bp in total), would be nearly equal to the HBA score
421 calculated for the same 147-bp sequence.
422 In order to examine whether local HBA scores indicate the suitability of
423 nucleosomal subsegments for histone-DNA contacts, we used the 282-bp
424 Widom 601 sequence, which was used for the above HBA analysis (**Figure 2**),

425 as a test sequence. **Figure 5A** presents a heatmap of the calculated local HBA
426 scores. The topmost track, marked “A”, represents local HBA scores for the A
427 segments of 147-bp sequences centered at the indicated dyad positions. At
428 potential translational positions with high HBA scores (**Figure 5A**, arrows), local
429 HBA scores for some segments (e.g., segments D and E of the 147-bp
430 sequence centered at position 154) were very high, as expected. However, each
431 147-bp sequence also contained segments with relatively low local HBA scores.
432 Interestingly, at translational positions with low HBA scores (those intervened by
433 the high-scoring HBA positions), some segments (e.g., segments F and G at
434 position 157) exhibited very low local HBA scores. This tendency was
435 particularly notable for the central segments E through I. The finding of these
436 low-scoring local HBA segments suggests that sub-sequences that are
437 disfavored in other rotational settings play a role in determining rotational
438 settings.
439 Next, we focused on the 147-bp Widom 601 sequence, of which the
440 dyad base is located at nucleotide position 154 of the original sequence (**Figure**
441 **5A**, red arrow). We chose this sequence because its derivatives are
442 well-characterized in reconstitution studies, as it forms a homogeneous
443 nucleosome, and because the left and right halves exhibit different features
444 [44-48]. In agreement with previous transcription studies [44, 45, 49], local HBA
445 scores for the 147-bp sequence were highest for segments D and E, both of
446 which share SHL −2.5, in which the high-affinity R3 element is located (**Figure**
447 **5B**, red dots). The region covered by segments D and E (SHL −3.5 to −1.5) is
448 also known for its very strong histone-DNA interaction [46]. These data indicate
449 that local HBA scores are useful for evaluating the suitability of nucleosomal
450 subsegments for histone-DNA contacts.
451 When the translational position was shifted along the 282-bp sequence
452 from nucleotide position 154 to surrounding nucleotide positions ranging from
453 150 to 158 (**Figure 5C**, −4 to +4 with respect to 154), the local HBA scores for
454 segments B through I changed dramatically (**Figure 5B**, black dots). However,
455 such shifting resulted in modest changes at segments J through M, with their
456 scores remaining at relatively high levels. These observations suggest that
457 segments J through M in the right-half sequence are generally suitable for

458 nucleosome formation, which may also be true even when the translational
459 position is shifted by several base pairs. In other words, segments J through M
460 do not seem to have intra-nucleosomal disfavored elements that strongly limit
461 the rotational setting. This nature might account for the observed poor crystal
462 development or lower homogeneity [46], as this 'slippery' sequence could disrupt
463 the uniformity of histone-DNA interactions.
464
465 **Local HBA scores for modified nucleosome-forming sequences**
466 The above testing with the sequence previously used in reconstitution studies
467 suggested that calculating local HBA scores would enhance understanding of
468 how DNA sequences characterize the nucleosomes formed *in vitro*. This
469 prompted us to examine whether we could apply local HBA calculations to the
470 evaluation of intracellular nucleosome formation. González et al. reported that
471 sequence-based determinants of nucleosome positioning are dispersed across
472 nucleosomes [42]. One gene that they thoroughly investigated was the fission
473 yeast gene, *ura4*$^{+}$. This gene has six nucleosomes (designated +1 to +6) that
474 overlap the coding sequence, five of which were predictable with the chemical
475 map–based dHMM (**Additional file 1: Figure S3**, nuCpos). Although the
476 remaining nucleosome (+5) was not clearly observed in the predicted occupancy
477 plot, there was one position with a relatively high HBA score that reasonably
478 corresponded to the *in vivo* nucleosome (marked with an asterisk in the figure).
479 A heatmap of local HBA scores for this gene suggested that there is an element
480 disfavored for nucleosome formation near the +5 nucleosome (**Figure 6**, WT,
481 see the blueish slanted line spanning segments A through M around dyad
482 position 600-750). Thus, this element may have caused the lower predicted
483 occupancy of the +5 nucleosome.
484 González et al. generated three modified sequences, namely Dyad,
485 Linker, and Int, for the *ura4*$^{+}$ gene and studied nucleosome positioning on these
486 sequences *in vivo* [42]. In Dyad and Linker, 51-bp sequences centered to each
487 nucleosome dyad and each linker, respectively, were replaced with artificial
488 randomized sequences of the same length. In the sequence Int, two
489 intra-nucleosomal sequences of each nucleosome (positions from −51 to −24
490 and 24 to 51) were replaced. According to their report, the occupancy of the +1

491 nucleosome was markedly decreased in the Dyad sequence, that of the +2
492 nucleosome was slightly increased in the Dyad and Linker sequences, and the
493 position of the +3 nucleosome was shifted in the Int sequence [42]. However,
494 why these changes selectively occur on specific nucleosomes *in vivo* remains
495 unclear due to a lack of appropriate analytical methods.
496 A heatmap of local HBA scores for the Dyad sequence clearly showed
497 that the replacements in this sequence caused a marked decrease in local HBA
498 scores only around the position of the +1 nucleosome (**Figure 6**, Dyad, blue
499 blots not present in the WT map). In contrast, the 51-bp replacements at other
500 nucleosomal centers did not appear to dramatically change their suitability for
501 nucleosome formation. As to the +2 nucleosome, we did not observe any
502 supportive signatures that could explain why the Dyad and Linker replacements
503 slightly increased its occupancy (**Figure 6**, Dyad and Linker). We observed that
504 the Int replacements primarily affected local HBA scores around the position of
505 the +3 nucleosome (**Figure 6**, Int, see the blueish slanted line spanning
506 segments A through M). This suggests that the mutation specifically affects the
507 +3 nucleosome positioning and triggers a shift of it to a more suitable position *in*
508 *vivo*. Giving these results, local HBA score calculation for nucleosomal
509 subsegments is valuable for *in silico* assessment of modified
510 nucleosome-forming sequences.
511
512 **Chemical map–based prediction of nucleosome positioning in mice**
513 The observation that the chemical-map–based models successfully improved
514 the prediction accuracy of nucleosomes in both yeasts suggested that these
515 approaches would be useful for mice. We therefore examined nucleosome
516 positioning in the mouse genome using the MNase-seq–based and chemical
517 map–based models. Matching of the predicted nucleosomes on the Viterbi paths
518 at a 2-bp resolution demonstrated the better performance of the chemical map–
519 based model (**Figure 7A**). In the MNase-seq–based prediction, 1.4% and 11.0%
520 of predicted nucleosomes matched unique and redundant nucleosomes,
521 respectively. In contrast, the chemical map–based prediction scored 10.0% and
522 41.1%, respectively. Thus, the chemical map–based model is more suitable for

523 prediction of nucleosome positioning with respect to the genomic coordinates
524 than the MNase-seq–based model.
525 In order to determine whether the mouse chemical map−based model
526 could better predict the rotational setting of nucleosomes, as shown in the case
527 of yeasts, we calculated the matching rate between predicted and *in vivo*
528 nucleosomes with stepwise widening of the matching window (**Figure 7B**). As
529 expected, the matching rate for MNase-seq–based prediction increased
530 gradually as the matching window widened. In chemical map–based predictions,
531 a dramatic increase in matching rate was observed when the matching window
532 was widened from 0 to 1 bp. However, the matching rate also increased
533 gradually as the matching window was widened further (See **Figures 1B and 7B**
534 for comparison between species). The shapes of the matching rate curves
535 suggested the chemical map–based model still distinguished rotational settings
536 to some extent. However, the mouse chemical model did not appear to be as
537 good at discriminating rotationally mispredicted nucleosomes compared with the
538 yeast models.
539
540 **Discussion:**
541 We developed a chemical map–based computational method to predict
542 nucleosome positioning and assessed the prediction results. Training dHMMs
543 with chemical maps improved the prediction accuracy of nucleosome locations
544 with respect to the genomic coordinates in budding yeast, fission yeast, and the
545 house mouse, *Mus musculus* (**Figures 1 and 7**). A total of 41-54% of the
546 predicted nucleosomes on the Viterbi paths matched those of *in vivo*
547 nucleosomes at 2-bp resolution. Genome- and locus-level evaluations showed
548 that the software successfully predicted the positions of *in vivo* nucleosomes
549 (**Figures 3 and 4, and Additional file 1: Figures S2 and S3**). The *in vivo*
550 nucleosomes generally had high HBA scores at reasonably near genomic
551 positions. Furthermore, we demonstrated that perturbation of nucleosome
552 positioning associated with genetic alterations could also be predicted (**Figures**
553 **4 and 6, and Additional file 1: Figure S4**). Thus, we propose that the software
554 nuCpos can be used for predicting nucleosome positions and also for
555 engineering of nucleosome-forming sequences.

556 As expected, the use of chemical maps in prediction led to strict
557 recognition of rotational settings, which was demonstrated with *in vitro* and *in*
558 *vivo* nucleosome-forming sequences (**Figures 1, 2 and 7**). Indeed, chemical
559 map–based HBA scores for the original Widom 601 sequence indicated that the
560 147-bp sequence centered at nucleotide position 154 was suitable for
561 nucleosome formation (**Figure 2**). The *in vitro* nucleosome positions formed on
562 natural sequences were also predictable at a rotational level (**Additional file 1:**
563 **Figure S1**); however, the HBA amplitude along these sequences was smaller
564 than along the *in vitro*–optimized Widom sequence. In contrast to the chemical
565 map–based models, the MNase-seq–based models largely failed to recognize
566 rotational settings (**Figures 1, 2 and 7**). This may have been due to the A/T
567 preference of MNase and the difficulty of dyad base calling in the construction of
568 MNase-seq–based nucleosome maps [7, 12]. Indeed, the MNase-map–based
569 model simply predicted A/T-rich regions as nucleosome-depleted regions
570 (**Figure 3 and Additional file 1: Figure S3**). Therefore, instead of
571 lower-resolution MNase-seq–based maps, base-pair-resolution chemical maps
572 should be used in the development of bioinformatics methods to obtain better
573 prediction results.
574 In the field of synthetic biology, the MNase-seq–based method NuPoP
575 has contributed to the design of synthetic promoters and terminators [29, 30].
576 However, some studies reported that the histone binding affinity and predicted
577 nucleosome occupancy output by NuPoP do not correlate with synthetic
578 promoter and terminator activity [34, 35]. Differences in the DNA sequences of
579 the tested synthetic elements may of course account for this discrepancy, as
580 discussed elsewhere [34, 35]. Clearly, the A/T frequency does affect the
581 prediction results of NuPoP (**Figure 3 and Additional file 1: Figure S3**). As
582 demonstrated in the Results section, the chemical map–based method is more
583 accurate and not affected by the A/T frequency. Thus, we expect that nuCpos is
584 applicable to prediction-based engineering and more effective for synthesizing
585 functional elements. Similarly, nuCpos can also be used to examine whether
586 DNA sequences of interest are suitable for nucleosome formation, as previously
587 done with NuPoP [37, 38].

588 Dividing 147-bp nucleosomal DNA into subsegments and calculating
589 subsegment local HBA scores revealed unexpected landscapes of the tested
590 nucleosomes. Our study demonstrated that the Widom sequence had segments
591 that are both favored and disfavored for nucleosome formation (**Figure 5A**).
592 Importantly, segments with high local HBA scores in the Widom 601 sequence
593 matched the high-affinity histone-contacting regions (**Figure 5B**). Furthermore,
594 the Widom sequence had segments exhibiting the highest local HBA scores at
595 the *in vitro* rotational settings (**Figure 5B**). In contrast to the left-half sequence,
596 the right-half sequence contained segments exhibiting relatively high local HBA
597 scores that did not change markedly with different rotational settings. Thus,
598 these sequence-specific features may account for the different behavior of each
599 half in *in vitro* experiments [44-48]. Collectively, our data indicate that local HBA
600 scores for given sequences provide insights that enhance understanding of
601 nucleosome features.
602 Locus-level assessments of prediction results suggested that software
603 users should not place undue confidence in the predicted nucleosome
604 occupancy output from chemical map–based models. Indeed, there was a
605 misplaced nucleosome that occupied the replication origin *ARS1* (**Figure 3**). We
606 speculate that these discrepancies were due primarily to the fact that the models
607 do not consider involvement of functional DNA elements or transacting factors
608 that regulate nucleosome formation at specific genomic locations. In other words,
609 chromatin regions highlighted by the differences in nucleosome occupancy
610 between prediction and *in vivo* observations could be the targets of future
611 investigations aiming to uncover region-specific regulatory mechanisms.
612 We demonstrated that the prediction accuracy of the chemical map–
613 based mouse model (41%) was apparently lower than that of yeast models (over
614 50%) (**Figures 1 and 7**). A disadvantage of mouse prediction is that the
615 chemical map probably contains substantial numbers of pseudo-positive
616 nucleosomes, which could not be omitted due to the lower cleavage depth than
617 yeasts [22]. As a consequence, the density of nucleosomes on the mouse
618 genome is one nucleosome per 3 bp, which is about 10 times higher than yeast
619 values (see **Methods**). Thus, we speculate that the accuracy of mouse
620 nucleosome positioning prediction can be increased if other chemical maps of

621 higher quality are produced and used for model construction. In human
622 genomics, the relationship between various phenotypes and genetic variants
623 has been explored [50-53]. However, the effects of genetic variants on
624 nucleosome formation in this regard remain to be studied. Now that our study
625 showed that insertions and replacements that disrupt nucleosomes in yeast cells
626 can be predictable *in silico*, it is possible that nucleosome-disrupting human
627 mutations can be found with the aid of chemical map–based predictions.
628
629 **Conclusions:**
630 In this study, we demonstrated that the accuracy of dHMM-based nucleosome
631 positioning prediction can be substantially increased by using
632 base-pair-resolution nucleosome maps for model construction. Our prediction
633 results suggest that chemical map–based models are useful for predicting
634 nucleosome positioning in wild-type and modified sequences at the locus level.
635 We also demonstrated that strong histone-DNA contacts in a nucleosome and
636 their rotational settings can be predicted. Furthermore, as another advantage of
637 bioinformatics methods, our models indicate that the commonly used Widom
638 sequence contains subnucleosomal segments that are disfavored, or statistically
639 very rare in *in vivo* nucleosomes, at their nucleosomal positions in shifted
640 rotational settings. We expect that our prediction method will provide further
641 insights that will enhance understanding of nucleosome-based epigenetic
642 regulation.
643
644 **Methods:**
645 **Software and data sets**
646 Most analyses were performed in the GNU R environment
647 (https://www.r-project.org, ver. 3.6.1). R packages were obtained from CRAN
648 (https://cran.r-project.org/) and Bioconductor (http://bioconductor.org/). Chemical
649 maps for budding yeast, fission yeast, and house mouse (*Mus musculus*)
650 embryonic stem cells [17, 19, 22] were used for model construction and testing
651 the prediction accuracy. The budding yeast's chemical map in the sacCer2
652 coordinate was lifted over to the sacCer3 coordinate, as described elsewhere
653 [27]. The number of unique and redundant nucleosomes in the budding yeast

654 genome was 67,548 and 344,709, respectively; fission yeast, 75,828 and
655 425,653; mice, 10,677,016 and 850,701,275. Reference genomes of budding
656 yeast (R64-1-1), fission yeast (ASM294v2), and mice (mm9/NCBIM37.67) were
657 used. The original 282-bp Widom 601 sequence [10], the 485-bp MMTV 3'–LTR
658 sequence [54] and the somatic 5S RNA gene of *Xenopus borealis* [55] were
659 used as test sequences.
660
661 **Construction of the nuCpos package**
662 The codes for the construction of parameters used in the nuCpos package (ver.
663 1.2.0) are available online (**https://doi.org/10.5281/zenodo.3362065**). Genomic
664 regions covered with the 147-bp non-redundant (unique) chemically mapped
665 nucleosomes and uncovered were defined as nucleosome and linker regions,
666 respectively. DNA sequences of these regions were used to construct
667 parameters that were transferred to internal Fortran programs. Nucleosomes of
668 which dyads were located within 73 bp of the chromosomal ends were omitted.
669 For construction of the mouse model, hard-masked genomic sequences were
670 used, and nucleosome and linker regions containing N were omitted before
671 parameter construction to avoid potential prediction bias caused by repeat
672 elements. In total, 67,538 nucleosome regions and 50,622 linker regions were
673 obtained for the budding yeast genome (sacCer3); fission yeast (ASM294v2),
674 75,826 nucleosome and 46,557 linker regions; mice (mm9), 4,147,972
675 nucleosome and 2,484,347 linker regions.
676 We developed an R function designated *predNuCpos*, which predicts
677 nucleosome positioning based on a dHMM, as previously proposed by Xi et al.
678 [16]. Like its ancestral function *predNuPoP* in the NuPoP package
679 (https://doi.org/doi:10.18129/B9.bioc.NuPoP, ver. 1.34.0), *predNuCpos* receives
680 a DNA sequence of any length, invokes an internal Fortran program, and
681 outputs the prediction result either in the working directory or in the working
682 environment of R. In *predNuCpos*, construction of the dHMM is based on
683 chemical maps, as described below.
684 Parameters used in the *predNuCpos* function were constructed
685 according to the NuPoP paper [16] using the functionalities of the Biostrings
686 package (https://doi.org/doi:10.18129/B9.bioc.Biostrings, ver. 2.52.0). The

687 parameters were as follows: *freqL*, one-base frequencies for linker regions; *tranL*,
688 *tranL2*, *tranL3*, and *tranL4*, First- to 4th-order transition probabilities for linker
689 regions, respectively; *freqN4*, four-base frequencies at the first four nucleotide
690 positions of nucleosome regions; *tranN4*, time-dependent 4th-order transition
691 probabilities for nucleosome regions; *Pd*, linker length distribution that ranges
692 from 1 to 500 bp. Linker sequences of 7-500 bp in length were used for linker
693 model construction, as described elsewhere [16]. *freqL* and *freqN4* were
694 obtained using the *oligonucleotideFrequency* function of Biostrings; *tranL*,
695 *tranL2*, *tranL3*, *tranL4*, and *tranN4* were obtained using the
696 *oligonucleotideTransitions* function of Biostrings. Moving average smoothing
697 using the *SMA* function of the TTR package
698 (https://cran.r-project.org/package=TTR, ver. 0.23-4) at a 3-bp window was
699 applied to the 4th-order transition probability parameter *tranN4* and to the linker
700 length distribution parameter *Pd*. The parameters used in *predNuCpos* were
701 also used in another function, *mutNuCpos*, which predicts the effect of genetic
702 alterations on nucleosome positioning.
703 Xi et al. proposed the HBA score [16], which is also referred to as the
704 'nucleosome affinity score', as the log likelihood ratio of the probability for a
705 given 147-bp sequence to be a nucleosome versus a linker. According to their
706 definition, the HBA score for the 147-bp region x centering at position i (a_i) on a
707 given genomic sequence is,

708 $$a_i := log\left[\frac{P_N(x_{i-73},\ldots,x_{i+73})}{G_L(x_{i-73},\ldots,x_{i+73}|147)}\right],$$

709 where P_N and G_L represent the probability of observing the 147-bp sequence as
710 a nucleosome or a linker, respectively [16]. The probability of being a
711 nucleosome is calculated by referring to the parameters *freqN4* and *tranN4*,
712 which are derived from nucleosomal DNA sequences. Similarly, calculation of
713 the probability of being a linker is based on linker DNA sequences. As
714 nucleosomal and linker sequences do not overlap in terms of their genomic
715 coordinates, negativity of HBA does not directly mean that the tested sequence
716 is inappropriate for nucleosome formation. The *predNuCpos* function calculates
717 chemical map–based HBA scores along the input sequence and outputs them
718 as raw values as its default behavior. We developed an independent function

719 designated *HBA*, which only calculates the HBA score for a given 147-bp
720 sequence. The *HBA* function uses the abovementioned chemical parameters for
721 *predNuCpos*: *freqL*, *tranL*, *tranL2*, *tranL3*, *tranL4*, *freqN4*, and *tranN4*.
722 We defined 13 overlapping nucleosomal subsegments, A through M,
723 and developed a function designated *localHBA* that calculates "local" HBA
724 scores for each segment. Segment A corresponds to nucleosomal nucleotide
725 positions 1-21; B, 12-31; C, 22-42; D, 33-52; E, 43-63; F, 54-73; G, 64-84; H,
726 75-94; I, 85-105; J, 96-115; K, 106-126; L, 117-136; and M, 127-147. Similar to
727 the calculation of HBA [16], the local HBA score for segment A of the 147-bp
728 potential nucleosomal region x centering at position i (l_i) on a given genomic
729 sequence is calculated as,

730 $$l_i := log\left[\frac{P_N(x_{i-73},\ldots,x_{i-53})}{G_L(x_{i-73},\ldots,x_{i-53}|21)}\right],$$

731 where the probabilities of observing the 21-bp sequence as segment A of a
732 nucleosome and a linker are calculated. Local HBA scores for the other
733 segments are calculated in the same way, except that the considered nucleotide
734 positions are set appropriately. At the implementation level, four-base frequency
735 values for the first four nucleotide positions of each segment were prepared:
736 *freqN4SA* corresponds to nucleosome positions 1-4; *freqN4SB*, 12-15;
737 *freqN4SC*, 22-25; *freqN4SD*, 33-36; *freqN4SE*, 43-46; *freqN4SF*, 54-57;
738 *freqN4SG*, 64-67; *freqN4SH*, 75-78, *freqN4SI*, 85-88; *freqN4SJ*, 96-99;
739 *freqN4SK*, 106-109; *freqN4SL*, 117-120; and *freqN4SM*, 127-130. These
740 four-base frequency values were used to calculate the probability of the segment
741 as that part of nucleosomal DNA as done for HBA calculations [16].
742

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

745 The codes for prediction, evaluation, and figure drawing used in this study are
746 available online (**https://doi.org/10.5281/zenodo.4083950**). Unmasked
747 reference genomes for the budding yeast, fission yeast, and house mouse (*Mus*
748 *musculus*) were used as prediction target sequences. For the mouse
749 experiments, only the 19th chromosome (Chr19) was used for prediction
750 because its length (61,342,430 bp) was sufficient for evaluation, at

751 approximately five-times longer than the total length of yeast chromosomes
752 (budding yeast, 12,071,326 bp; fission yeast, 12,571,820 bp). Mouse Chr19
753 contained 249,210 unique and 19,965,481 redundant nucleosomes that had
754 been determined based on histone H4 S47C-dependent cleavage [22]. The
755 density of redundant nucleosomes on mouse Chr19 was one per 3.07 bp. With
756 this density, one in three randomly located nucleosomes in the chromosome
757 could be counted as a truly predicted nucleosome. Thus, we selected 2,044,747
758 nucleosomes with the highest NCP scores as true redundant nucleosomes for
759 the matching experiment. This selection yielded a density of one nucleosome
760 per 30.0 bp, which was comparable to that of budding yeast and fission yeast
761 (one nucleosome per 35.0 bp and 29.5 bp, respectively).
762 The *predNuPoP* function of NuPoP and the *predNuCpos* function of
763 nuCpos were used for dHMM-based predictions. In order to predict using
764 budding yeast models, the species arguments were set as “7” for *predNuPoP*
765 and “sc” for *predNuCpos*. Similarly, fission yeast (“9” and “sp”) and mouse (“2”
766 and “mm”) models were specified. We defined any nucleosomes on the Viterbi
767 path as predicted nucleosomes. Matching of predicted nucleosomes using *in*
768 *vivo* chemical nucleosomes was done by widening the genomic coordinates of *in*
769 *vivo* nucleosomes to both sides in a step-wise manner from zero to nine base
770 pairs. For instance, at 1-bp resolution, a predicted nucleosome was regarded as
771 “matched” when the genomic coordinate of its dyad was equal to that of the
772 nearest *in vivo* nucleosome. If the dyad of the predicted nucleosome was 1 bp
773 away from that of the nearest *in vivo* nucleosome, it was regarded as “not
774 matched.”
775 Chemical map−based HBA scores were calculated using the
776 *predNuCpos* function or the *HBA* function of nuCpos, which yield the same
777 scores. The species argument was set as “sc” for budding yeast sequences and
778 sequences of *in vitro* reconstitution studies; for the fission yeast sequences, “sp.”
779 The *predNuPoP* function of NuPoP smooths raw HBA scores using a moving
780 average of a 55-bp window before outputting the processed HBA scores. This
781 prevents software users from examining the actual HBA score for each genomic
782 coordinate to determine if the 147-bp sequence of interest is suitable for
783 nucleosome formation. To overcome this problem, non-smoothed MNase-seq–

784 based HBA scores were calculated using in-house functions that utilize the
785 parameters for the *predNuPoP* function of NuPoP. For comparison of *in vivo*
786 nucleosomes, 5,542 sets of −1 and +1 nucleosomes determined by Chereji et al.
787 [12] were used for HBA calculations. The 347-bp sequences around these
788 nucleosomes (from −173 to +173 nucleotide positions with respect to the dyads)
789 were extracted from the budding yeast genome. The HBA score was calculated
790 for each possible 147-bp segment along the sequences. A/T-frequency was
791 calculated using the *letterFrequency* function of Biostrings for each 147-bp
792 sequence. HBA and A/T-frequency scores were averaged at each coordinate
793 with respect to the −1 or +1 nucleosomes. Chemical map−based local HBA
794 scores were calculated using the *localHBA* function of nuCpos. Heatmaps of
795 local HBA scores were drawn using the *levelplot* function of the rasterVis
796 package (https://cran.r-project.org/package=rasterVis, ver. 0.46).
797
798 **List of abbreviations:**
799 dHMM: duration hidden Markov model
800 HBA: histone binding affinity
801 LTR: long terminal repeat
802 MMTV: mouse mammary tumor virus
803 MNase: micrococcal nuclease
804 NCP: nucleosome core particle
805 NDR: nucleosome depleted region
806 SELEX: Systematic Evolution of Ligands by EXponential enrichment
807 SHL: superhelical location
808
809 **Declarations:**
810 **Ethics approval and consent to participate:**
811 Not applicable.
812
813 **Consent for publication:**
814 Not applicable.
815
816 **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.

850 5. Struhl K, Segal E: **Determinants of nucleosome positioning.** *Nat Struct*
851 *Mol Biol* 2013, **20:**267-273.
852 6. Lieleg C, Krietenstein N, Walker M, Korber P: **Nucleosome positioning**
853 **in yeasts: methods, maps, and mechanisms.** *Chromosoma* 2015,
854 **124:**131-151.
855 7. Chereji RV, Clark DJ: **Major Determinants of Nucleosome Positioning.**
856 *Biophys J* 2018, **114:**2279-2289.
857 8. Teif VB: **Nucleosome positioning: resources and tools online.** *Brief*
858 *Bioinform* 2016, **17:**745-757.
859 9. Liu H, Zhang R, Xiong W, Guan J, Zhuang Z, Zhou S: **A comparative**
860 **evaluation on prediction methods of nucleosome positioning.** *Brief*
861 *Bioinform* 2014, **15:**1014-1027.
862 10. Lowary PT, Widom J: **New DNA sequence rules for high affinity**
863 **binding to histone octamer and sequence-directed nucleosome**
864 **positioning.** *J Mol Biol* 1998, **276:**19-42.
865 11. Pisano S, Galati A, Cacchione S: **Telomeric nucleosomes: forgotten**
866 **players at chromosome ends.** *Cell Mol Life Sci* 2008, **65:**3553-3563.
867 12. Chereji RV, Ramachandran S, Bryson TD, Henikoff S: **Precise**
868 **genome-wide mapping of single nucleosomes and linkers in vivo.**
869 *Genome Biol* 2018, **19:**19.
870 13. Chereji RV, Bryson TD, Henikoff S: **Quantitative MNase-seq accurately**
871 **maps nucleosome occupancy levels.** *Genome Biol* 2019, **20:**198.
872 14. West JA, Cook A, Alver BH, Stadtfeld M, Deaton AM, Hochedlinger K,
873 Park PJ, Tolstorukov MY, Kingston RE: **Nucleosomal occupancy**
874 **changes locally over key regulatory regions during cell**
875 **differentiation and reprogramming.** *Nat Commun* 2014, **5:**4719.
876 15. Mieczkowski J, Cook A, Bowman SK, Mueller B, Alver BH, Kundu S,
877 Deaton AM, Urban JA, Larschan E, Park PJ, et al: **MNase titration**
878 **reveals differences between nucleosome occupancy and chromatin**
879 **accessibility.** *Nat Commun* 2016, **7:**11485.
880 16. Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang JP:
881 **Predicting nucleosome positioning using a duration Hidden Markov**
882 **Model.** *BMC Bioinformatics* 2010, **11:**346.

883 17. Brogaard K, Xi L, Wang JP, Widom J: **A map of nucleosome positions**
884 **in yeast at base-pair resolution.** *Nature* 2012, **486:**496-501.
885 18. Brogaard KR, Xi L, Wang JP, Widom J: **A chemical approach to**
886 **mapping nucleosomes at base pair resolution in yeast.** *Methods*
887 *Enzymol* 2012, **513:**315-334.
888 19. Moyle-Heyrman G, Zaichuk T, Xi L, Zhang Q, Uhlenbeck OC, Holmgren
889 R, Widom J, Wang JP: **Chemical map of Schizosaccharomyces**
890 **pombe reveals species-specific features in nucleosome positioning.**
891 *Proc Natl Acad Sci U S A* 2013, **110:**20158-20163.
892 20. Henikoff S, Ramachandran S, Krassovsky K, Bryson TD, Codomo CA,
893 Brogaard K, Widom J, Wang JP, Henikoff JG: **The budding yeast**
894 **Centromere DNA Element II wraps a stable Cse4 hemisome in either**
895 **orientation in vivo.** *Elife* 2014, **3:**e01861.
896 21. Ramachandran S, Zentner GE, Henikoff S: **Asymmetric nucleosomes**
897 **flank promoters in the budding yeast genome.** *Genome Res* 2015,
898 **25:**381-390.
899 22. Voong LN, Xi L, Sebeson AC, Xiong B, Wang JP, Wang X: **Insights into**
900 **Nucleosome Organization in Mouse Embryonic Stem Cells through**
901 **Chemical Mapping.** *Cell* 2016, **167:**1555-1570 e1515.
902 23. Shimizu M, Roth SY, Szent-Gyorgyi C, Simpson RT: **Nucleosomes are**
903 **positioned with base pair precision adjacent to the alpha 2 operator**
904 **in Saccharomyces cerevisiae.** *EMBO J* 1991, **10:**3033-3041.
905 24. Roth SY, Dean A, Simpson RT: **Yeast alpha 2 repressor positions**
906 **nucleosomes in TRP1/ARS1 chromatin.** *Mol Cell Biol* 1990,
907 **10:**2247-2260.
908 25. Thoma F, Bergman LW, Simpson RT: **Nuclease digestion of circular**
909 **TRP1ARS1 chromatin reveals positioned nucleosomes separated by**
910 **nuclease-sensitive regions.** *J Mol Biol* 1984, **177:**715-733.
911 26. Morohashi N, Yamamoto Y, Kuwana S, Morita W, Shindo H, Mitchell AP,
912 Shimizu M: **Effect of sequence-directed nucleosome disruption on**
913 **cell-type-specific repression by alpha2/Mcm1 in the yeast genome.**
914 *Eukaryot Cell* 2006, **5:**1925-1933.

915 27. Fuse T, Katsumata K, Morohoshi K, Mukai Y, Ichikawa Y, Kurumizaka H,
916 Yanagida A, Urano T, Kato H, Shimizu M: **Parallel mapping with**
917 **site-directed hydroxyl radicals and micrococcal nuclease reveals**
918 **structural features of positioned nucleosomes in vivo.** *PLoS One*
919 2017, **12:**e0186974.
920 28. Fuse T, Yanagida A, Shimizu M: **The Yeast Minichromosome System**
921 **Consisting of Highly Positioned Nucleosomes in Vivo.** *Biol Pharm*
922 *Bull* 2019, **42:**289-294.
923 29. Curran KA, Crook NC, Karim AS, Gupta A, Wagman AM, Alper HS:
924 **Design of synthetic yeast promoters via tuning of nucleosome**
925 **architecture.** *Nat Commun* 2014, **5:**4002.
926 30. Morse NJ, Gopal MR, Wagner JM, Alper HS: **Yeast Terminator**
927 **Function Can Be Modulated and Designed on the Basis of**
928 **Predictions of Nucleosome Occupancy.** *ACS Synth Biol* 2017,
929 **6:**2086-2095.
930 31. Wight A, Yang D, Ioshikhes I, Makrigiannis AP: **Nucleosome Presence**
931 **at AML-1 Binding Sites Inversely Correlates with Ly49 Expression:**
932 **Revelations from an Informatics Analysis of Nucleosomes and**
933 **Immune Cell Transcription Factors.** *PLoS Comput Biol* 2016,
934 **12:**e1004894.
935 32. Creamer KM, Job G, Shanker S, Neale GA, Lin YC, Bartholomew B,
936 Partridge JF: **The Mi-2 homolog Mit1 actively positions nucleosomes**
937 **within heterochromatin to suppress transcription.** *Mol Cell Biol* 2014,
938 **34:**2046-2061.
939 33. Small EC, Xi L, Wang JP, Widom J, Licht JD: **Single-cell nucleosome**
940 **mapping reveals the molecular basis of gene expression**
941 **heterogeneity.** *Proc Natl Acad Sci U S A* 2014, **111:**E2462-2471.
942 34. Portela RM, Vogl T, Kniely C, Fischer JE, Oliveira R, Glieder A: **Synthetic**
943 **Core Promoters as Universal Parts for Fine-Tuning Expression in**
944 **Different Yeast Species.** *ACS Synth Biol* 2017, **6:**471-484.
945 35. MacPherson M, Saka Y: **Short Synthetic Terminators for Assembly of**
946 **Transcription Units in Vitro and Stable Chromosomal Integration in**
947 **Yeast S. cerevisiae.** *ACS Synth Biol* 2017, **6:**130-138.

948 36. Mahmoudi T: **The BAF complex and HIV latency.** *Transcription* 2012,
949 **3:**171-176.
950 37. Rafati H, Parra M, Hakre S, Moshkin Y, Verdin E, Mahmoudi T:
951 **Repressive LTR nucleosome positioning by the BAF complex is**
952 **required for HIV latency.** *PLoS Biol* 2011, **9:**e1001206.
953 38. Gimeno-Valiente F, Riffo-Campos AL, Vallet-Sanchez A, Siscar-Lewin S,
954 Gambardella V, Tarazona N, Cervantes A, Franco L, Castillo J,
955 Lopez-Rodas G: **ZNF518B gene up-regulation promotes**
956 **dissemination of tumour cells and is governed by epigenetic**
957 **mechanisms in colorectal cancer.** *Sci Rep* 2019, **9:**9339.
958 39. Lipford JR, Bell SP: **Nucleosomes positioned by ORC facilitate the**
959 **initiation of DNA replication.** *Mol Cell* 2001, **7:**21-30.
960 40. Eaton ML, Galani K, Kang S, Bell SP, MacAlpine DM: **Conserved**
961 **nucleosome positioning defines replication origins.** *Genes Dev* 2010,
962 **24:**748-753.
963 41. Rodriguez J, Lee L, Lynch B, Tsukiyama T: **Nucleosome occupancy as**
964 **a novel chromatin parameter for replication origin functions.**
965 *Genome Res* 2017, **27:**269-277.
966 42. Gonzalez S, Garcia A, Vazquez E, Serrano R, Sanchez M, Quintales L,
967 Antequera F: **Nucleosomal signatures impose nucleosome**
968 **positioning in coding and noncoding sequences in the genome.**
969 *Genome Res* 2016, **26:**1532-1543.
970 43. Ichikawa Y, Morohashi N, Nishimura Y, Kurumizaka H, Shimizu M:
971 **Telomeric repeats act as nucleosome-disfavouring sequences in**
972 **vivo.** *Nucleic Acids Res* 2014, **42:**1541-1552.
973 44. Kulaeva OI, Gaykalova DA, Pestov NA, Golovastov VV, Vassylyev DG,
974 Artsimovitch I, Studitsky VM: **Mechanism of chromatin remodeling and**
975 **recovery during passage of RNA polymerase II.** *Nat Struct Mol Biol*
976 2009, **16:**1272-1278.
977 45. Gaykalova DA, Kulaeva OI, Volokh O, Shaytan AK, Hsieh FK,
978 Kirpichnikov MP, Sokolova OS, Studitsky VM: **Structural analysis of**
979 **nucleosomal barrier to transcription.** *Proc Natl Acad Sci U S A* 2015,
980 **112:**E5787-5795.

981 46. Chua EY, Vasudevan D, Davey GE, Wu B, Davey CA: **The mechanics**
982 **behind DNA sequence-dependent properties of the nucleosome.**
983 *Nucleic Acids Res* 2012, **40:**6338-6352.
984 47. Vasudevan D, Chua EYD, Davey CA: **Crystal structures of**
985 **nucleosome core particles containing the '601' strong positioning**
986 **sequence.** *J Mol Biol* 2010, **403:**1-10.
987 48. Tsunaka Y, Fujiwara Y, Oyama T, Hirose S, Morikawa K: **Integrated**
988 **molecular mechanism directing nucleosome reorganization by**
989 **human FACT.** *Genes Dev* 2016, **30:**673-686.
990 49. Bondarenko VA, Steele LM, Ujvari A, Gaykalova DA, Kulaeva OI,
991 Polikanov YS, Luse DS, Studitsky VM: **Nucleosomes can form a polar**
992 **barrier to transcript elongation by RNA polymerase II.** *Mol Cell* 2006,
993 **24:**469-479.
994 50. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, Hess GT,
995 Zappala Z, Strober BJ, Scott AJ, et al: **The impact of rare variation on**
996 **gene expression across tissues.** *Nature* 2017, **550:**239-243.
997 51. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G,
998 Statistical Methods groups-Analysis Working G, Enhancing Gg, Fund
999 NIHC, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, et al: **Genetic effects on**
1000 **gene expression across human tissues.** *Nature* 2017, **550:**204-213.
1001 52. Chakravorty S, Hegde M: **Inferring the effect of genomic variation in**
1002 **the new era of genomics.** *Hum Mutat* 2018, **39:**756-773.
1003 53. Chakravorty S, Hegde M: **Gene and Variant Annotation for Mendelian**
1004 **Disorders in the Era of Advanced Sequencing Technologies.** *Annu*
1005 *Rev Genomics Hum Genet* 2017, **18:**229-256.
1006 54. Flaus A, Richmond TJ: **Positioning and stability of nucleosomes on**
1007 **MMTV 3'LTR sequences.** *J Mol Biol* 1998, **275:**427-441.
1008 55. Ura K, Hayes JJ, Wolffe AP: **A positive role for nucleosome mobility in**
1009 **the transcriptional activity of chromatin templates: restriction by**
1010 **linker histones.** *EMBO J* 1995, **14:**3752-3765.
1011
1012
1013 **Figure legends:**

1014 **Figure 1. Comparison of prediction accuracy between chemical map–**
1015 **based and MNase-seq–based models for yeasts.** (**A**) Distance between
1016 predicted nucleosomes and nearest *in vivo* nucleosomes. Fraction of predicted
1017 nucleosomes on the Viterbi path located at a particular distance from the nearest
1018 *in vivo* nucleosomes is plotted. Species of the target genome and for model
1019 construction were both *S. cerevisiae*. Magnified view is presented inside each
1020 plot. Arrows point to the fractions of predicted nucleosomes that are 1-bp apart
1021 from *in vivo* nucleosomes. (**B**) Matching rate between predicted and in vivo
1022 nucleosomes with variable matching windows. The species of genomic
1023 sequences used for prediction (Target) and the species of the chemical map
1024 used for model construction (Model) are indicated. "Sc" stands for *S. cerevisiae*,
1025 whereas "Sp" stands for *S. pombe*. Indicated R packages were used for
1026 predictions. (**C**) Matching of nucleosomes on the Viterbi paths with *in vivo* unique
1027 and redundant nucleosomes at 2-bp resolution. Note that the redundant
1028 nucleosome datasets contain both unique nucleosomes (red), which are
1029 representative, and non-representative nucleosomes (orange).
1030
1031 **Figure 2. HBA scores along the Widom 601 sequence.** HBA scores along the
1032 282-bp original Widom sequence were calculated using chemical map–based
1033 and MNase-seq–based budding yeast models. The scores were normalized by
1034 subtracting the mean value from the raw value on a per sequence basis. The
1035 HBA score for a given 147-bp nucleosome sequence was assigned to its dyad
1036 position. Orange vertical lines indicate the dyad positions of *in vitro* reconstituted
1037 nucleosome: nucleotide position 154. Asterisks indicate high-scoring HBA
1038 positions around the *in vitro* positioning site.
1039
1040 **Figure 3. Comparison between chemical map–based and MNase-seq–**
1041 **based models on selected *in vivo* regions.** (**A and B**) Chemical map–based
1042 and MNase-seq–based HBA scores (**A**) and A/T nucleotide frequency (**B**) along
1043 the sequences containing budding yeast −1 and +1 nucleosomes for
1044 protein-coding genes. Gray vertical lines indicate the dyad positions of
1045 respective nucleosomes. Nucleosomes of which the dyad base is located at 0 bp
1046 are shown as ovals above the plots. Asterisks indicate positions with relatively

1047 high HBA scores. NDR: nucleosome-depleted region. (**C**) Prediction results for
1048 the budding yeast *TRP1ARS1* mini-chromosome. Schematic representation of *in*
1049 *vivo* nucleosome positioning is shown above the plots. Nucleosomes labeled I
1050 through VII are indicated as ovals. Note that this sequence is circularized *in vivo*
1051 by being linked at the *Eco*RI sites. The top two panels show the prediction
1052 results output by nuCpos, whereas the next two panels show NuPoP results.
1053 The upper panel in each set shows predicted occupancy of the nucleosome
1054 (Occup., gray polygons) and probabilities of the tested 147-bp sequences for
1055 being in the nucleosome state (P-dyad, blue vertical lines). The lower panels
1056 show HBA values for the tested 147-bp sequences calculated using the
1057 indicated models. The very bottom panel shows the A/T-frequencies for the
1058 tested 147-bp sequences. Horizontal lines at the bottom of each plot indicate a
1059 unit of the circular *TRP1ARS1* mini-chromosome (colored in gray, nucleotide
1060 positions 0 to 1,464), the coding region of *TRP1* (red, 115-879), and the *B3*
1061 (black, 763-788), *B2* (purple, 810-820), *B1* (green, 847-859), and *ACS* (blue,
1062 869-879) elements of *ARS1*. Inverted triangles indicate the histone H4
1063 S47C-dependent cleavage centers determined by indirect end-labeling; orange
1064 for mini-chromosome and light blue for genomic experiments [27].
1065
1066 **Figure 4. Predicting the effects of repeat insertion on nucleosome**
1067 **positioning.** Chemical map–based HBA scores and predicted nucleosomal
1068 occupancy were calculated for original and modified *BAR1* promoter sequences.
1069 The sequences were centered at the dyad position of the repressive
1070 nucleosome, in which the indicated repeat sequences were inserted. Horizontal
1071 lines at the bottom of each plot indicate the *α2*-operator (purple), the *TATA*
1072 (blue) element, the coding region of *BAR1* (orange), and the insert (red). Note:
1073 the sequence insertion caused a shift of the left of the three predicted
1074 nucleosomes to the right to cover the *α2*-operator in the occupancy plots.
1075 Overconfidence should not be placed in this shift, as it may simply reflect that the
1076 dHMM tends to follow the linker length distribution, which generally does not
1077 allow longer linkers.
1078

1079 **Figure 5. Local HBA analysis of the Widom 601 sequence.** (**A**) Chemical
1080 map–based local HBA scores along the 282-bp Widom 601 sequence. The
1081 scores for each tested 147-bp sequence were heatmapped at the center position
1082 of that sequence. The heatmap scale indicates the local HBA score. Black
1083 arrows indicate high-HBA-score positions (124, 135, 143, 164, and 175). Red
1084 arrow indicates the dyad position of the *in vitro* reconstituted nucleosome (154).
1085 The region evaluated in (**B**) is indicated as a horizontal gray bar under the X-axis.
1086 (**B**) Local HBA scores at (red) and around (black) the *in vitro* translational
1087 position for each segment are plotted in columns. The scores for the M segment
1088 are magnified to the left to show the indices on the X-axis. The relative positions
1089 −4, 0, and +4 correspond to nucleotide positions 150, 154, and 158 of the
1090 282-bp sequence. The location of the R3 element is indicated at the top. (**C**)
1091 Schematic representation of shifting of translational positions in (**B**). Nucleotide
1092 positions relative to the dyad of the central translated position (i=154) are
1093 indicated. Only three (−4, 0, and +4) of the nine positions evaluated in (**B**) are
1094 shown.
1095
1096 **Figure 6. Local HBA analysis of modified nucleosome-forming sequences.**
1097 Chemical map–based local HBA scores along the original and modified fission
1098 yeast *ura4*$^{+}$ gene. The scores for each tested 147-bp sequence were
1099 heatmapped at the center position of that sequence. The heatmap scale
1100 indicates the local HBA score. Black arrows indicate the nucleosome centers
1101 (nucleotide positions +17, +164, +332, +470, +629, and +806) determined by
1102 MNase-seq [42].
1103
1104 **Figure 7. Comparison of prediction accuracy between chemical map–**
1105 **based and MNase-seq–based models for mice.** (**A**) Matching of nucleosomes
1106 on the Viterbi paths with *in vivo* unique and redundant nucleosomes. (**B**)
1107 Matching rate between predicted and *in vivo* nucleosomes with variable
1108 matching windows. The sequence of mouse Chr19 was used for prediction.
1109
1110 **Table 1. Functions of the software packages used in this study.**

NuPoP [16]	nuCpos	Description
*predNuPoP**	*predNuCpos***	Prediction of nucleosome positioning based on dHMMs
readNuPoP		Loading of prediction results
plotNuPoP		Plotting of prediction results
	HBA	Calculation of HBA scores
	localHBA	Calculation of local HBA scores
	mutNuCpos	Prediction of the effects of genetic alterations on nucleosome positioning

*MNase-seq–based. HBA scores are smoothed before outputting the prediction results. **Chemical map–based. Smoothing of HBA scores is optional.

Supplementary Figure Legends:

Additional file 1: Figure S1. HBA scores along *in vitro* nucleosome-forming sequences. HBA scores along nucleosome-forming sequences were calculated using chemical map–based and MNase-seq–based budding yeast models. Sequences analyzed were the 3'-LTR of MMTV (**A**) and the *Xenopus borealis* 5S rDNA dinucleosome-forming sequence (**B**). 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 139 and 335 (**A**) and 106 and 303 (**B**). 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.

Additional file 1: Figure S2. 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.

1134 **Additional file 1: Figure S3. Prediction results for the fission yeast *ura4*$^{+}$**
1135 **gene.** Schematic representation of *in vivo* nucleosome positioning is shown
1136 above the plots. Nucleosomes numbered +1 to +6 are indicated as ovals. The
1137 top two panels show the prediction results output by nuCpos, whereas the next
1138 two panels show NuPoP results. The upper panel in each set shows predicted
1139 occupancy of nucleosomes (Occup., gray polygons) and probability that the
1140 tested 147-bp sequences are in the nucleosome state (P-dyad, blue vertical
1141 lines). Lower panels show HBA values for the tested 147-bp sequences
1142 calculated using the indicated models. The very bottom panel shows the
1143 A/T-frequency for the tested 147-bp sequences. Horizontal lines at the bottom of
1144 each plot indicate the 5'- and 3'-untranslated regions (colored in gray, nucleotide
1145 positions −151 to −1 and +795 to +986, respectively) and the protein-coding
1146 region (red, 0 to +794) of the *ura4*$^{+}$ gene. Inverted triangles indicate the
1147 nucleosome centers determined by MNase-seq [42].
1148
1149 **Additional file 1: Figure S4. Prediction of the effects of repeat insertion in**
1150 **the fourth nucleosome of the *TALS* minichromosome.** Chemical map–based
1151 HBA scores and predicted nucleosomal occupancy were calculated for original
1152 and modified *TALS* sequences [43]. The sequences were centered at the dyad
1153 position of the fourth nucleosome (0 bp), in which the indicated repeat
1154 sequences were inserted. Note: the position for the fourth nucleosome was
1155 predicted to shift to the right. Horizontal lines at the bottom of each plot indicate
1156 the *α2*-operator (purple) and the insert (red). hTEL stands for human telomeric
1157 repeat (5'-TTAGGG-3'); SI-A, sequence isomer-A (5'-TGTAGG-3'); SI-B,
1158 sequence isomer-B (5'-TGTGAG-3'). When telomeric DNA fragments are
1159 inserted at the center of the fourth nucleosome, shorter fragments (hTEL2 and
1160 hTEL4) were predicted to not affect nucleosome positioning. In contrast, longer
1161 fragments (hTEL12 and hTEL29) were predicted to severely compromise
1162 nucleosome formation. In addition, sequence isomers of the telomeric repeat
1163 (SI-A6, SI-A12, SI-B6, and SI-B12) were predicted to not inhibit nucleosome
1164 formation. These prediction results agreed with previous *in vivo* observations
1165 [43]. The hTEL6 insertion, which causes nucleosome depletion *in vivo*,

1166 appeared not to affect nucleosome formation in the prediction, suggesting a
1167 limitation of the prediction.
1168
1169 **Additional file 1: Figure S5. Thirteen overlapping nucleosomal DNA**
1170 **subsegments for local HBA calculation.** Segments A through M are colored in
1171 orange on the nucleosome structure 1AOI [1]. The structures were drawn using
1172 MacPyMOL (v1.8.4.1). The base pair length (red) and corresponding
1173 superhelical locations (black) of each segment are indicated. The color of the
1174 right-half segments (H-M) is faded because they are behind the front helix.
1175
1176

Figure 1

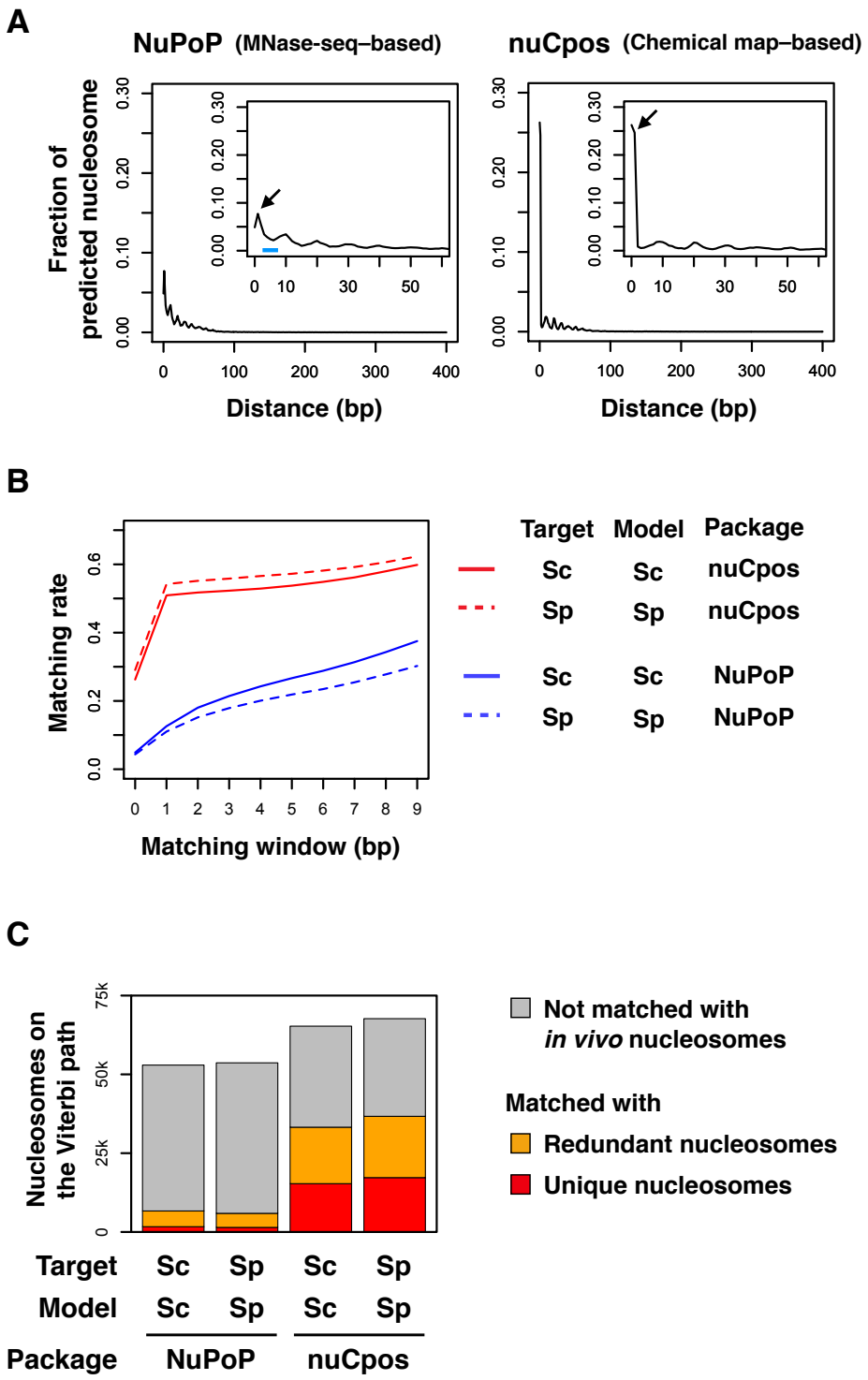

A
NuPoP (MNase-seq–based)
nuCpos (Chemical map–based)
Fraction of predicted nucleosome
Distance (bp)
B
Matching rate
Matching window (bp)
Target Model Package
Sc Sc nuCpos
Sp Sp nuCpos
Sc Sc NuPoP
Sp Sp NuPoP
C
Nucleosomes on the Viterbi path
Not matched with in vivo nucleosomes
Matched with
Redundant nucleosomes
Unique nucleosomes
Target Sc Sp Sc Sp
Model Sc Sp Sc Sp
Package NuPoP nuCpos

Figure 2

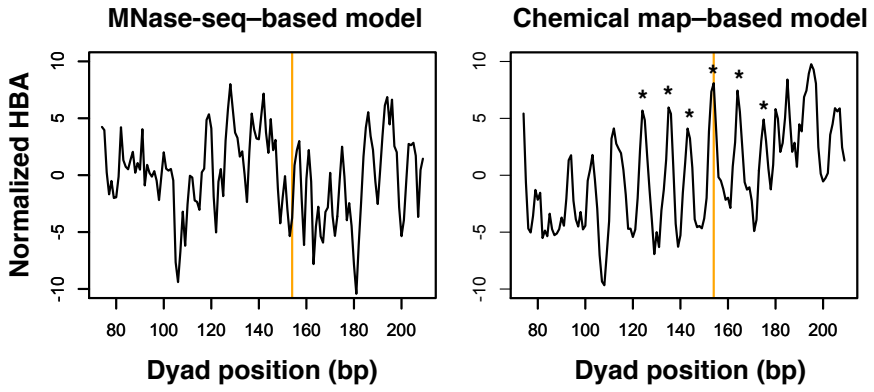

MNase-seq–based model
Chemical map–based model
Normalized HBA
10
5
0
-5
-10
80
100
120
140
160
180
200
Dyad position (bp)
Dyad position (bp)

Figure 3

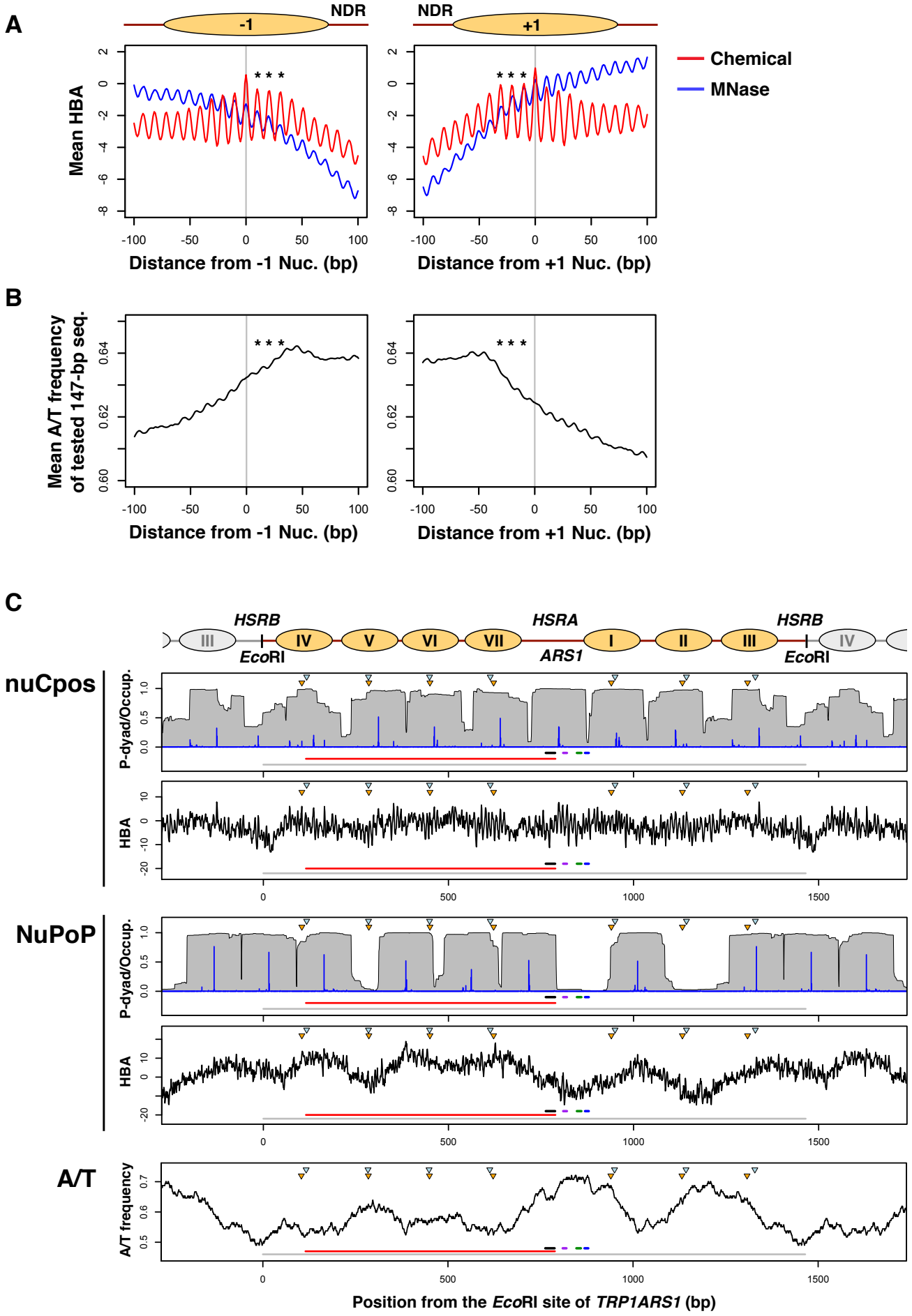

A
NDR
-1
NDR
+1
Mean HBA
Chemical
MNase
Distance from -1 Nuc. (bp)
Distance from +1 Nuc. (bp)
B
Mean A/T frequency of tested 147-bp seq.
Distance from -1 Nuc. (bp)
Distance from +1 Nuc. (bp)
C
HSRB
HSRA
HSRB
III
IV
V
VI
VII
I
II
III
IV
EcoRI
ARS1
EcoRI
nuCpos
P-dyad/Occup.
HBA
NuPoP
P-dyad/Occup.
HBA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

Figure 4

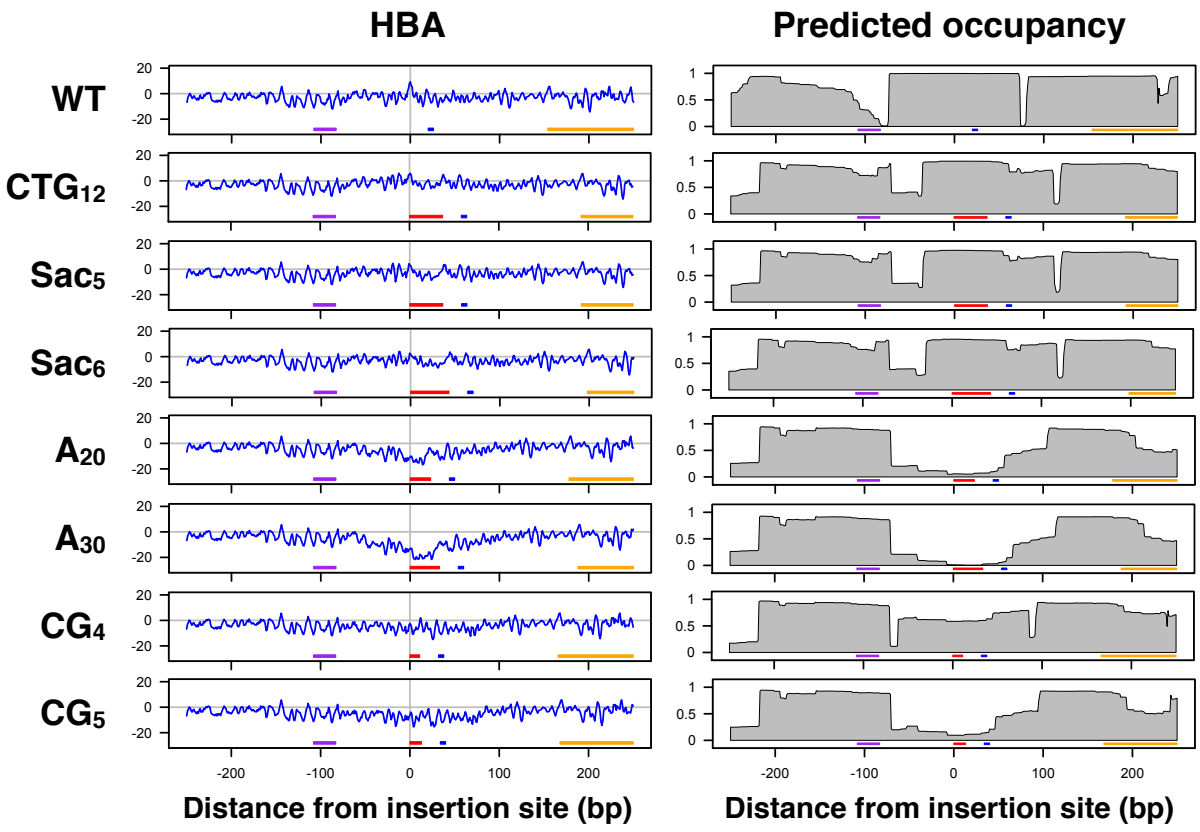

HBA
Predicted occupancy
WT
CTG12
Sac5
Sac6
A20
A30
CG4
CG5
20
0
-20
1
0.5
0
-200
-100
0
100
200
Distance from insertion site (bp)
Distance from insertion site (bp)

Figure 5

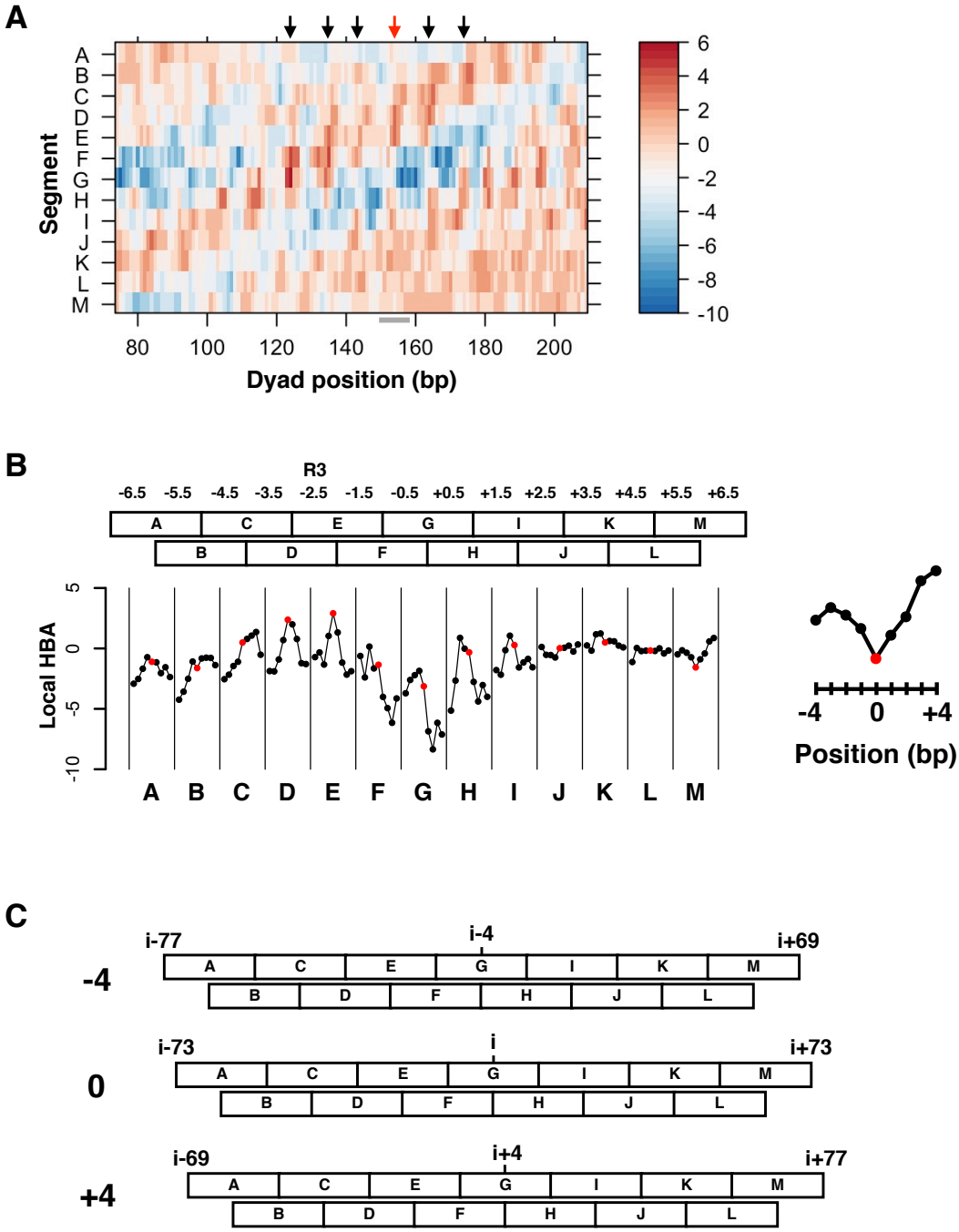

A
Segment
A
B
C
D
E
F
G
H
I
J
K
L
M
80
100
120
140
160
180
200
Dyad position (bp)
6
4
2
0
-2
-4
-6
-8
-10
B
R3
-6.5 -5.5 -4.5 -3.5 -2.5 -1.5 -0.5 +0.5 +1.5 +2.5 +3.5 +4.5 +5.5 +6.5
A C E G I K M
B D F H J L
Local HBA
5
0
-5
-10
A B C D E F G H I J K L M
-4 0 +4
Position (bp)
C
-4
i-77
i-4
i+69
A C E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

Figure 6

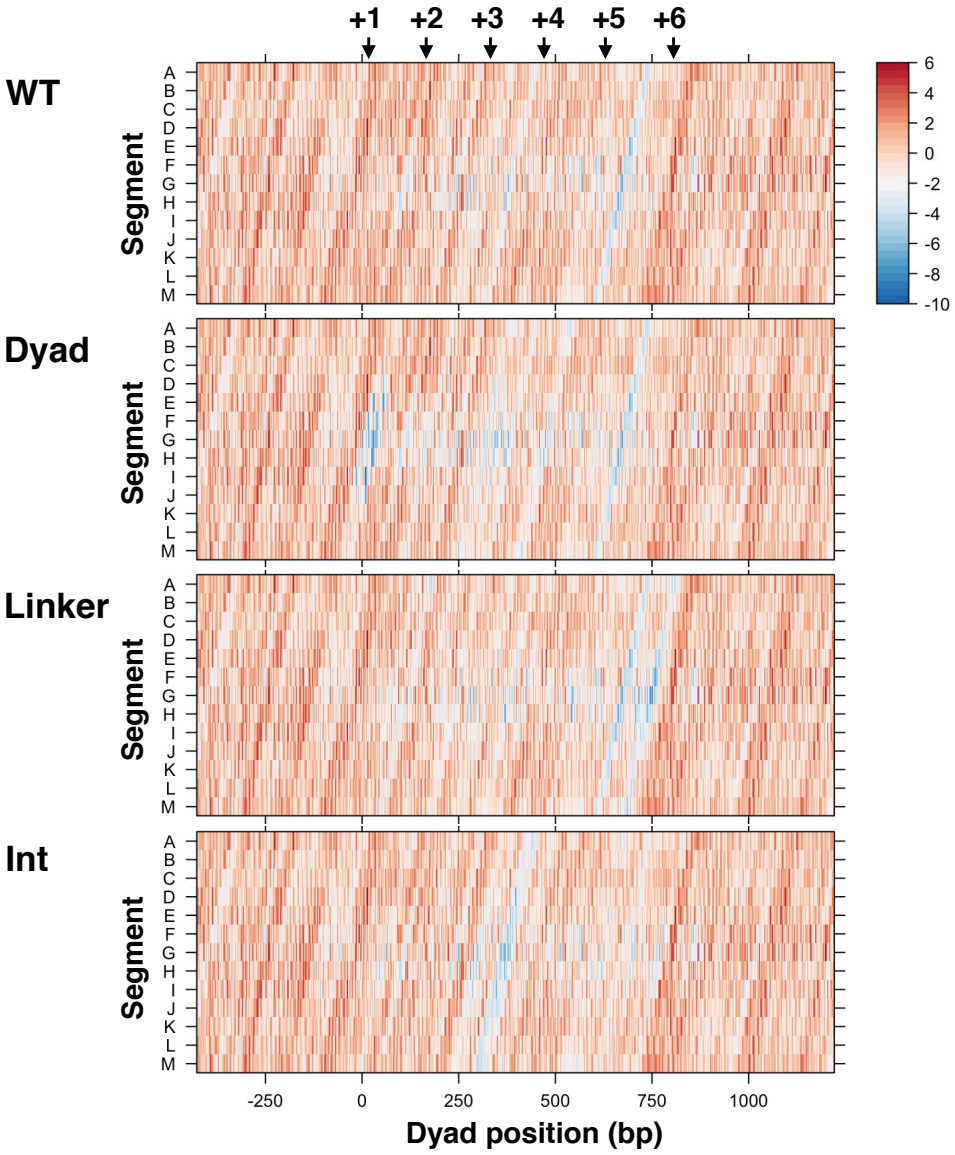
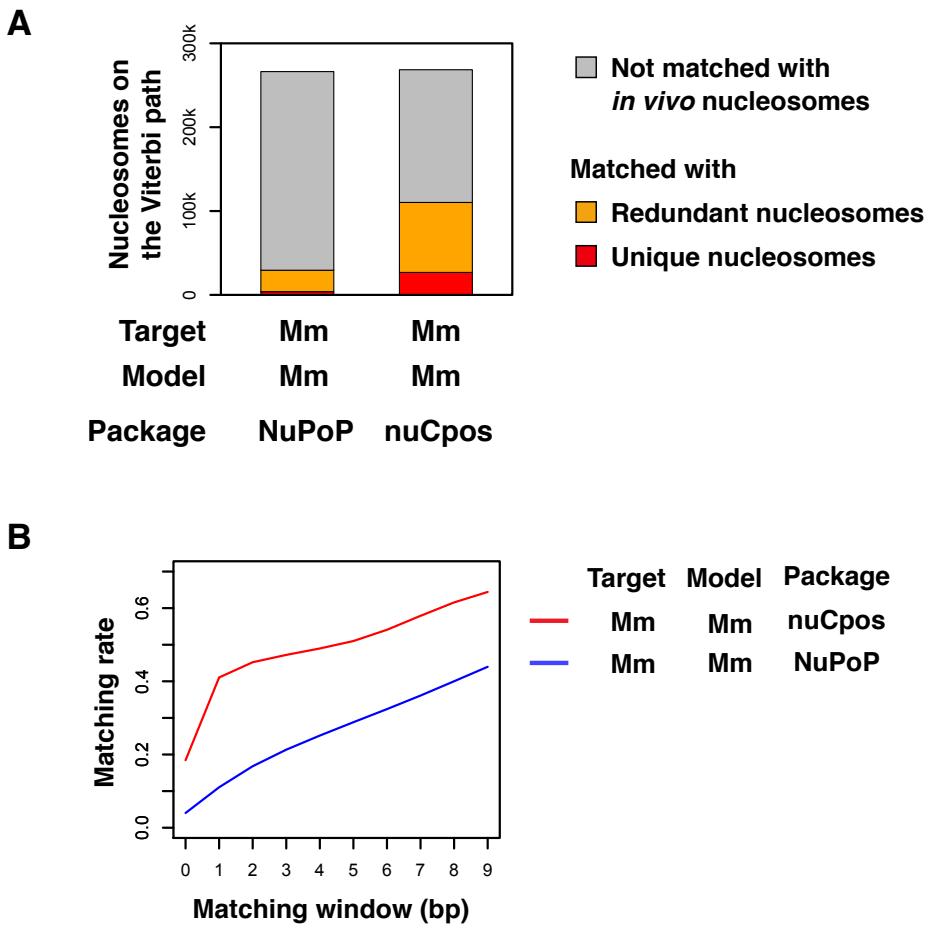

+1
+2
+3
+4
+5
+6
WT
Dyad
Linker
Int
Segment
A
B
C
D
E
F
G
H
I
J
K
L
M
6
4
2
0
-2
-4
-6
-8
-10
-250
0
250
500
750
1000
Dyad position (bp)

Figure 7


A
Nucleosomes on the Viterbi path
0
100k
200k
300k
Target
Model
Package
Mm
Mm
NuPoP
Mm
Mm
nuCpos
Not matched with *in vivo* nucleosomes
Matched with
Redundant nucleosomes
Unique nucleosomes
B
Matching rate
0.0
0.2
0.4
0.6
0 1 2 3 4 5 6 7 8 9
Matching window (bp)
Target Model Package
Mm Mm nuCpos
Mm Mm NuPoP