

DNase-seq predicts regions of rotational nucleosome stability across diverse human cell types

Deborah R. Winter,^{1,2} Lingyun Song,¹ Sayan Mukherjee,^{1,3} Terrence S. Furey,^{4,6} and Gregory E. Crawford^{1,5,6}

¹Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina 27708, USA; ²Computational Biology & Bioinformatics, Duke University, Durham, North Carolina 27708, USA; ³Departments of Statistical Science, Computer Science, and Mathematics, Duke University, Durham, North Carolina 27708, USA; ⁴Department of Genetics, Department of Biology, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA; ⁵Department of Pediatrics, Division of Medical Genetics, Duke University, Durham, North Carolina 27708, USA

DNase-seq is primarily used to identify nucleosome-depleted DNase I hypersensitive (DHS) sites genome-wide that correspond to active regulatory elements. However, ~40 yr ago it was demonstrated that DNase I also digests with a ~10-bp periodicity around nucleosomes matching the exposure of the DNA minor groove as it wraps around histones. Here, we use DNase-seq data from 49 samples representing diverse cell types to reveal this digestion pattern at individual loci and predict genomic locations where nucleosome rotational positioning, the orientation of DNA with respect to the histone surface, is stably maintained. We call these regions DNase I annotated regions of nucleosome stability (DARNS). Compared to MNase-seq experiments, we show DARNS correspond well to annotated nucleosomes. Interestingly, many DARNS are positioned over only one side of annotated nucleosomes, suggesting that the periodic digestion pattern attenuates over the nucleosome dyad. DARNS reproduce the arrangement of nucleosomes around transcription start sites and are depleted at ubiquitous DHS sites. We also generated DARNS from multiple lymphoblast cell line (LCL) samples. We found that LCL DARNS were enriched at DHS sites present in most of the original 49 samples but absent in LCLs, while multi-cell-type DARNS were enriched at LCL-specific DHS sites. This indicates that variably open DHS sites are often occupied by rotationally stable nucleosomes in cell types where the DHS site is closed. DARNS provide additional information about precise DNA orientation within individual nucleosomes not available from other nucleosome positioning assays and contribute to understanding the role of chromatin in gene regulation.

[Supplemental material is available for this article.]

Most of the human genome, like other eukaryotes, is concentrated in nucleosomes, which consist of ~147 bases of DNA wrapped ~1.7 times around a histone octamer and connected by DNA linkers of variable length (Richmond and Davey 2003; Wang et al. 2008). As the basic unit of packaging in chromatin, nucleosomes incorporate the majority of bases in the genome (Felsenfeld and Groudine 2003; Valouev et al. 2011). Nucleosomes play a role in regulating gene transcription by controlling the accessibility of transcription factor (TF) binding sites at key locations (Albert et al. 2007).

The most common experimental method for genome-wide *in vivo* mapping of nucleosomes within the cell is to treat nuclei with micrococcal nuclease (MNase-seq) followed by high-throughput sequencing (Schones et al. 2008; Chodavarapu et al. 2010; Valouev et al. 2011). MNase preferentially digests within linkers, resulting in mononucleosome fragments that can be used to infer the dyad or center of well-positioned nucleosomes that maintain their exact coordinates across a population of cells (Noll 1974; Pugh 2010). Current studies have used MNase or chemical modifications to improve the resolution of “translational” positioning, which is de-

finer as the location of the histone core along the DNA (Valouev et al. 2011; Brogaard et al. 2012). However, it is difficult to determine the “rotational” positioning—the orientation of the DNA major and minor groove relative to the histone surface—from these data. Moreover, analysis of MNase-seq results may overlook “fuzzy” nucleosomes whose positions fluctuate across a population of cells, making it difficult to determine the alignment of the dyad. But even fuzzy nucleosomes may maintain rotational setting preferences as they undergo translational shifts of ~10 bp (Albert et al. 2007; Gaffney et al. 2012) that preserve the relationship with underlying dinucleotide periodicities (Trifonov and Sussman 1980; Satchwell et al. 1986; Segal et al. 2006). Nucleosomes that are crucial in regulating gene expression may shift at this period to maintain orientation as they allow TF access in response to cell conditions (Hu et al. 2011). In addition, the conservation of rotational positioning strikes a balance between the limited contribution of periodic sequence preferences (Satchwell et al. 1986; Segal et al. 2006; Kaplan et al. 2009) and the statistical positioning theory, which posits that most nucleosomes are uniformly packed between nucleosome boundaries formed by promoters, TF binding sites, and nucleosome-occluding sequences (Mavrich et al. 2008; Zhang et al. 2009). In the human genome, a previous study estimated that only 20% of nucleosomes are well-positioned translationally (Valouev et al. 2011), although the remaining nucleosomes may conserve their rotational setting and reflect additional regulatory potential.

Corresponding authors

E-mail tsfurey@email.unc.edu

E-mail greg.crawford@duke.edu

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/10.1101/gr.150482.112>.

Here, we present a new, complementary method that uses DNase I digestion data to predict regions of nucleosome rotational stability. DNase I preferentially digests DNA in nucleosome-depleted open chromatin regions and is generally used to locate all types of active regulatory elements (Wu et al. 1979; Gross and Garrard 1988). When coupled with high-throughput sequencing (DNase-seq), DNase I hypersensitive (DHS) sites are identified by the enrichment in DNase-seq reads. Each DNase-seq read indicates a digestion site at a specific locus with single-base resolution (Gross and Garrard 1988; Boyle et al. 2008). Early studies documented that DNase I-digested mononucleosomes resulted in DNA fragments that spaced ~ 10 bp apart using gel electrophoresis (Noll 1974). By plotting the distribution of distances between pairs of DNase-seq reads mapping outside of DHS sites, we previously observed an oscillation at ~ 10.4 bp (Boyle et al. 2008). Since the DNA helix completes one full turn in the same period (Wang 1979), this is consistent with the preference of DNase I to digest in the minor groove and its periodic exposure as DNA wraps around the nucleosome (Noll 1974; Cousins et al. 2004). Thus, the presence of this DNase I digestion pattern appears to be correlated with the position of nucleosomes.

In this study, we created a model that predicts individual DNase-annotated regions of nucleosome stability (DARNS) of varying length that are consistently occupied by nucleosomes with preserved rotational settings. Whereas previous studies demonstrated the 10.4-bp DNase I nucleosome pattern as a cumulative signal (Noll 1974; Boyle et al. 2008; Gaffney et al. 2012), we further this work by revealing that this period can be detected at individual loci and used to reverse engineer the location of DARNS. We determined that the rotational stability was highly conserved across 49 distinct DNase-seq data sets from diverse cell types. Unexpectedly, when we evaluated DARNS against other nucleosome annotations, we found that many were positioned on only one side of the nucleosome without crossing the dyad, where the periodic digestion pattern appears to be attenuated. By using recently available DNase-seq data for lymphoblastoid cell lines from multiple individuals (Degner et al. 2012), we also annotated DARNS in a single cell type. We found evidence of dynamic nucleosome positioning across cell-type-specific DHS sites compared with the multi-cell-type DARNS. Our results provide the first genome-wide annotation of the orientation of nucleosomes and reveal high-resolution features of nucleosomes that have precise rotational settings.

Results

Periodic DNase I digestion pattern is conserved across cell types

The 10.4-bp spacing of aligned DNase-seq reads was previously shown in genome-wide aggregate plots within a single-cell-type DNase-seq experiment (Boyle et al. 2008) but has been difficult to identify at individual loci due to the lack of sufficient read coverage. To increase overall read coverage, we integrated DNase-seq results from multiple cell types, isolated from various tissues in different individuals grown in nonidentical conditions (for list of cell types, see Methods) that individually demonstrated the 10.4-bp periodicity between DNase I digestion sites (data not shown). To verify that this periodicity was conserved between experiments, we plotted the pairwise spacing between reads from different cell types. The ~ 10 -bp oscillation in the plots was maintained for reads originating in human umbilical vein endothelial cell line (HUVEC) compared with reads on the same strand from the GM12878 lymphoblast cell line (LCL) (Fig. 1A); similar results

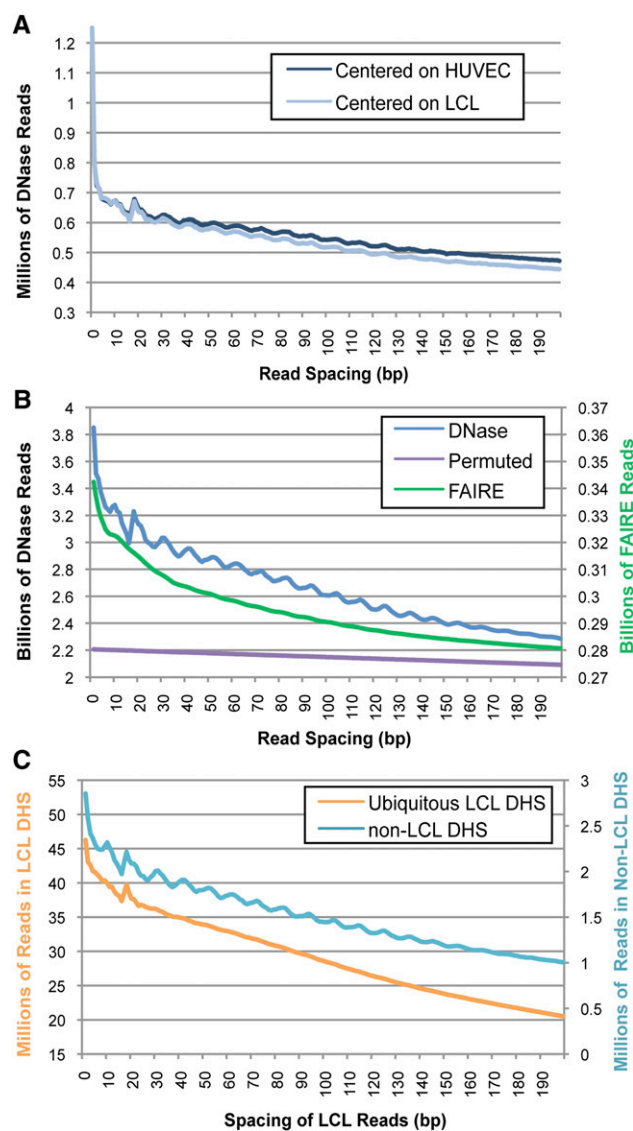


Figure 1. Characteristics of DNase-seq 10-bp periodicity pattern. (A) Distances between reads from two different cell types (human umbilical vein endothelial cell line [HUVEC] and GM12878 lymphoblastoid cell line [LCL]) are plotted and show the 10-bp periodic spacing preference. We show lines both displaying distances from HUVEC reads to LCL reads and vice versa. (B) Similar read spacing plots showing 10-bp period is present in combined data from multiple cell types in DNase-seq data but absent in FAIRE-seq (Formaldehyde-Assisted Identification of Regulatory Elements) or permuted DNase data. (C) Read spacing plot showing 10-bp period in LCL DNase-seq reads overlapping non-LCL DNase I hypersensitive (DHS) sites but not in LCL-specific DHS sites.

were obtained from all pairs of cell types tested (data not shown). Thus, the 10.4-bp spacing in digestion sites is generally in phase across cell types, suggesting that rotational positioning is widely conserved.

We proceeded to combine DNase-seq data from 49 samples representing 43 distinct cell types generated by our group at Duke as part of the ENCODE project (for combination algorithm, see Methods) (McDaniell et al. 2010; Song et al. 2011; The ENCODE Project Consortium 2012; Thurman et al. 2012). We compiled only the uniquely aligned reads from each of the DNase-seq experiments totaling ~ 1.5 billion data points on either strand. The dis-

tribution of distances between pairs of reads in the combined data set exhibited the expected spacing on both the negative (Fig. 1B) and positive (data not shown) strands with an estimated period of 10.45 ± 0.21 . When we sequenced a library of naked genomic DNA that was digested with DNase I, we found that the read spacing from these data does not exhibit periodicity (Supplemental Fig. S1A). In addition, we found that positive-strand digestion tended to be ~ 3 bp offset from the negative-strand digestion (Supplemental Fig. S1B). This is consistent with our previous results reported in one cell type (Boyle et al. 2008) and is likely associated with the 3' overhang previously documented at DNase I cut sites (Sollner-Webb et al. 1978, Cousins et al. 2004). We also noted that, unlike similar periodic profiles seen in MNase-seq studies that demonstrated a substantial increase in signal at ~ 150 bp (Valouev et al. 2011), the periodic spacing of DNase I reads attenuated as the distance between reads increased with no significant rise at the length of a nucleosome. These results support periodic DNase I digestion along nucleosome-bound DNA, as opposed to MNase digestion primarily within the linker.

To ensure these results were not due to chance, we generated a pair of negative control data sets. First, we permuted the base positions of the combined DNase-seq data within 1000-bp windows. This effectively disrupted local spacing patterns while maintaining local distributions of read counts. For this permuted data set, no obvious periodicity was observed, but rather a more uniform distribution of read pair distances was found (Fig. 1B). Second, to confirm that the 10-bp period of digestion sites was unique to DNase-seq digestion and not an inherent property of open chromatin assays, we performed a similar analysis on FAIRE-seq (Formaldehyde-Assisted Identification of Regulatory Elements) data, which similarly employs high-throughput sequencing to identify open chromatin (Giresi et al. 2007). Since FAIRE-seq uses random sonication to fractionate the genome, we expected that the positions of aligned reads from the ends of these fragments would be unrelated to fine nucleosome positioning. We combined FAIRE-seq data generated from 19 samples shared with the DNase-seq data as part of the ENCODE project, totaling ~ 830 million data points on each strand (Song et al. 2011; M. Schaner, J.M. Simon, K.A. Showers, Z. Zhang, P.G. Giresi, L. Song, D. London, T.S. Furey, G.E. Crawford, J.D. Lieb, unpubl.). When the distances between pairs of FAIRE-seq reads were plotted as above, no oscillation signal was evident (Fig. 1B). Hence, the 10-bp spacing of reads appears unique to the specificity of digestion sites of the DNase I enzyme.

Periodic DNase I digestion patterns are absent in nucleosome-depleted DHS regions

In our previous study, we showed that the 10.4-bp period in DNase I digestion disappeared when only DNase-seq reads originating from DHS sites in that cell type were included (Boyle et al. 2008). Similarly, we separately plotted the spacing between pairs of reads that mapped within DHS sites identified in the cell type from which the data were generated, and reads that mapped outside of these DHS sites; we observed the oscillation pattern only for the latter set of non-DHS site reads (Supplemental Fig. S1C).

We performed a similar analysis but focused on reads from only the seven LCLs that aligned to DHS sites identified in any cell type. We compared spacing between reads that mapped within DHS sites found in all LCLs (ubiquitous LCL DHS) and those that mapped within DHS sites found in other cell types but not LCLs (non-LCL DHS). We found that the characteristic oscillation was evident in the spacing of reads from non-LCL DHS sites, where we

would expect nucleosome in LCLs, but was absent for reads in ubiquitous LCL DHS sites that would be nucleosome-depleted (Fig. 1C). This suggests that in cell types where a DHS site is closed, the occupying nucleosome appears to establish a position that contains a precise rotational setting, and further supports the connection of periodic digestion with the dynamic positioning of nucleosomes.

Local DNase I digestion site spacing is predominantly 10.4 bp in length

By examining the combined DNase-seq data, we find that some genomic locations with sufficiently high read density clearly display the 10.4-bp digestion pattern on both strands, with the positive-strand digestion peaks 3 bp downstream from the negative-strand digestion peaks (Fig. 2A). This was not observed in either FAIRE-seq or the permuted data sets. Although nucleosomes likely incorporate the vast majority of the genome, the 10.4-bp period in preferred digestion sites is not readily detectable over much of the genome. This inconsistency is possibly due to noise in the data, the imbalanced coverage of reads in the DNase-seq data, or periodic patterns that are not maintained across cell types.

To better detect recurring periods of any length in the DNase I digestion patterns, we used Fourier transforms to decompose the DNase-seq data in small segments of the genome into component frequencies. For each overlapping 1000-bp window sliding along chromosome 1, we calculated the dominant period, indicating the most prevalent pattern in the window. Based on the Fourier analysis results, we estimate the average period of digestion around the nucleosome is between 10.3 and 10.4 bp (Fig. 2B; Supplemental Fig. S2A). As a negative control, we similarly analyzed the permuted (Fig. 2C) and FAIRE-seq data (Supplemental Fig. S2B) and observed that the most prevalent periods were very small (< 3 bp). This is consistent with the lack of oscillation described previously (Fig. 1B). Hence, the 10.4-bp period is recoverable at single loci in DNase-seq data but not in other data sets.

Identifying DNase I digestion patterns at individual loci

To systematically detect the 10.4-bp digestion pattern genome-wide given variation in signal and read depth, we first created a 91-bp pattern from the observed genome-wide DNase-seq read spacing that represented the expected digestion around the nucleosome (see Methods) (Fig. 1B, blue DNase line; for pictorial representation, see Supplemental Fig. 3A). Then, we calculated the Pearson correlation coefficient between this pattern and the distribution of the combined DNase-seq reads in 91-bp windows centered on each base of the genome, independently testing both strands (Fig. 3A). High positive correlations, anticipated once per complete DNA turn, indicated a good match to the expected pattern in nucleosomes, while high negative correlations occurred whenever the pattern was out of phase. Compared with FAIRE-seq and permuted data sets, the distribution of DNase-seq correlation scores was enriched at the extreme values ($P < 1^{-10}$) (Fig. 3B), indicating that these correlations may be used to predict stable rotational nucleosome positioning at base-pair resolution.

To identify bases with high correlation, we first defined local maxima or peaks in the correlation scores and plotted the distance between pairs of peaks from opposite strands (for definition of peaks, see Methods) (Fig. 3C). From this plot, we noted that correlation peaks (1) occurred with ~ 10.4 -bp spacing; (2) were 2–3 bp offset between positive and negative strands, consistent with

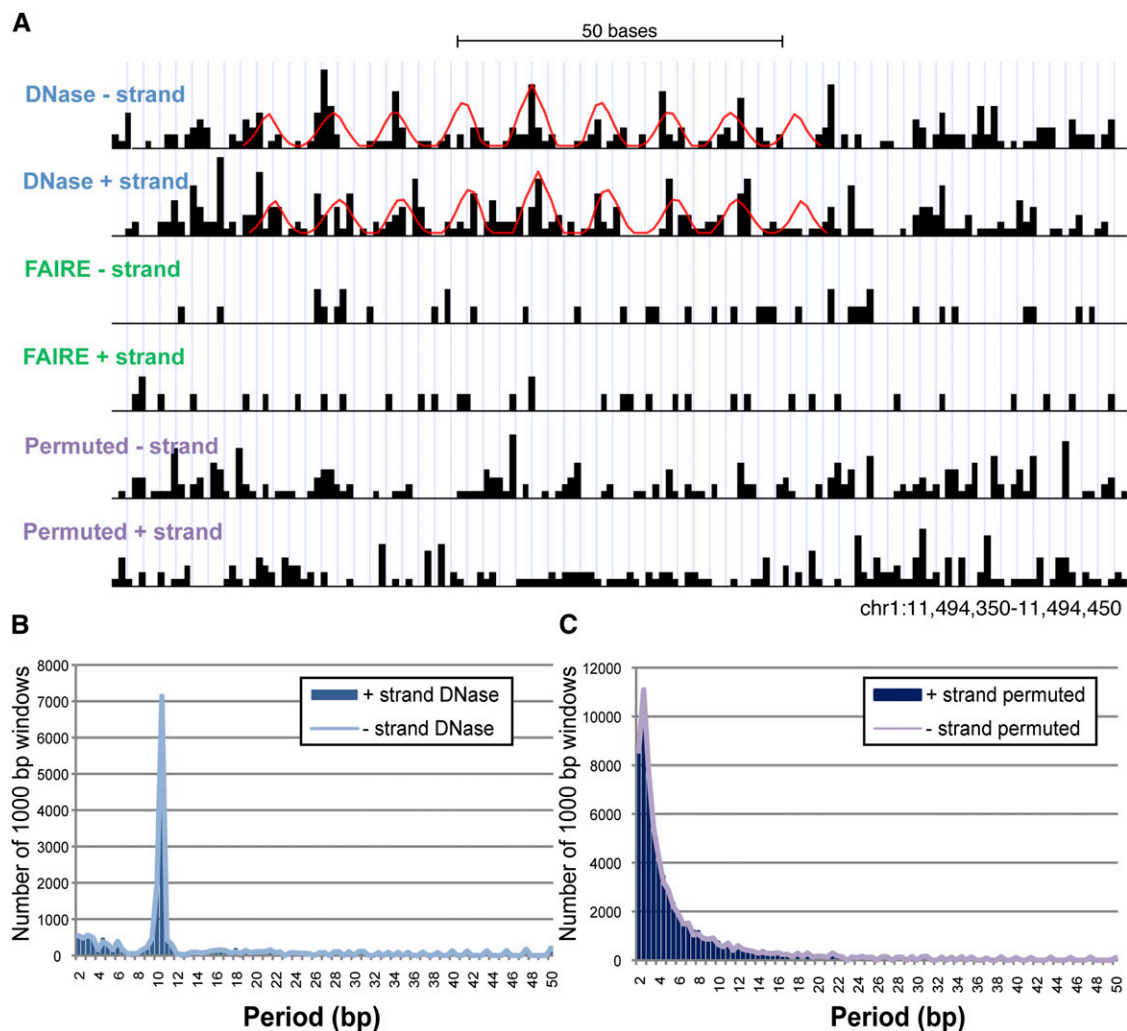


Figure 2. The 10-bp periodic digestion can be detected at individual genomic loci. (A) The combined read counts at individual bases shows the preference for a 10-bp spacing between reads in DNase-seq data, but not in FAIRE-seq data or in permuted DNase-seq data. The 91-bp nucleosome pattern used to identify DARNs (see Supplemental Fig. S3A) is superimposed on DNase-seq data. (B) Fourier analysis on 1000-bp windows from chromosome 1 reveals a prominent 10-bp period within DNase-seq data. (C) The same Fourier analysis on permuted data does not reveal a meaningful period.

reported 3' overhang of DNase I digestion; and (3) tended to be closely spaced, indicating multiple matches of the 91-bp pattern across whole nucleosomes. These features were not present in the permuted data set (Fig. 3C, purple line). As described below, we exploited these characteristics in creating a model to predict regions of nucleosome rotational stability.

Predicting regions of nucleosome rotational stability

To predict individual regions that display rotational nucleosome stability, we devised a hidden Markov model (HMM) to identify sequences of DNA with average spacing of 10.4 bp between correlation peaks on the same strand and where the positive-strand peaks trailed the negative-strand peaks by ~ 3 bp. The HMM transitioned between a background digestion state and a nucleosome meta-state composed of a cycle of states that generated the observed spacing between positive and negative correlation peaks associated with DNase I digestion of the nucleosome (for a more complete description, see Methods; for HMM diagram, see Supple-

mental Fig. S3B). Figure 4A illustrates a representative region depicting the combined DNase-seq data, correlations scores, correlation peaks, and regions assigned by the HMM to the nucleosome meta-state.

We labeled the “nucleosome” regions called by the HMM as DARNs because they represent predictions of DNA sequences covered by consistently positioned, rotationally stable nucleosomes without necessarily defining the exact boundaries of each nucleosome. We annotated ~ 14 million DARNs covering 890 million bases (30.77%) of the genome. Across all DARNs, the mean distance between adjacent correlation peaks on the same strand is 10.364 with a standard deviation of ~ 0.8 . While some DARNs cover only a portion of the nucleosome, others are much larger than the average size of a nucleosome and may reflect nucleotides over which one or more nucleosomes maintain their DNA orientation as their translational position fluctuates. The DARNs ranged in length from 15–1282 bp with a median of 50 bp. Of the three DARNs that were >1 kb, two mapped near centromeres; this is consistent with previous studies recording high, stable nucleosome occupancy in peri-

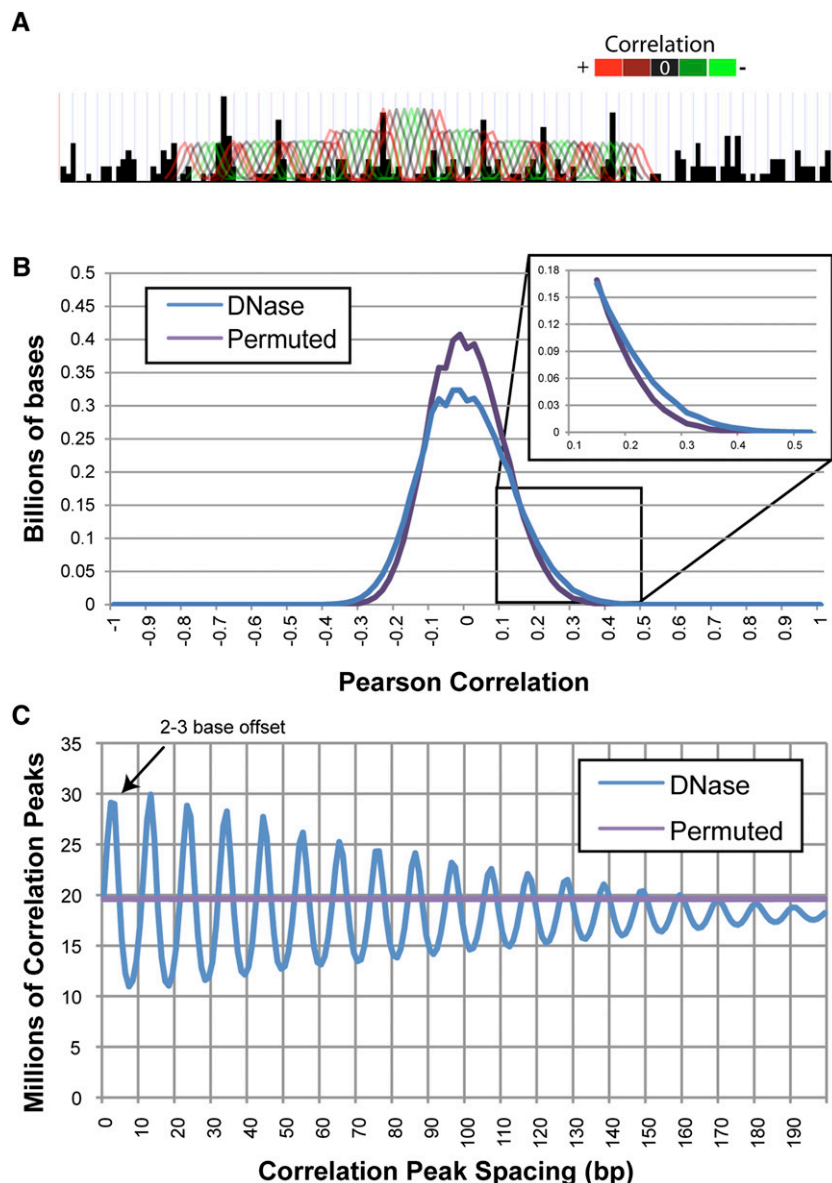


Figure 3. Correlating DNase-seq data to expected nucleosome pattern. (A) The expected 91-bp pattern of DNase I digestion around the nucleosome (Supplemental Fig. S3A) was correlated with the DNase-seq data at each base across the genome. (B) The distribution of correlation scores shows that DNase has more positive and negative values at the extremes compared with the permuted data set. (C) The distances between correlation peaks on the plus and minus strands in DNase-seq data and the permuted DNase-seq are plotted. Note that the 3-bp offset of the ~10 bp between correlation peaks on opposite strands is only detected in the DNase data.

centromeric regions (Chodavarapu et al. 2010, Gaffney et al. 2012). The distance between pairs of DARNs tended to be multiples of 10.4 bp (Fig. 4B). This may indicate that either smaller DARNs are from the same nucleosome but with spurious correlation peaks between them that deviate from the HMM model or that phasing occurs between individual nucleosomes (Valouev et al. 2011; Gaffney et al. 2012). The latter may reflect the positions of nucleosomes in higher-order structures.

As a control, we also used our HMM to annotate regions using the FAIRE-seq and permuted DNase-seq data sets. These control sets each predicted ~12 million regions covering ~540 million (18.6%) and ~520 million (17.9%) of bases in the genome, respec-

tively. We found these control regions did not significantly overlap DARNs defined using DNase I data (Table 1). We compared these annotations to the extended nucleosome boundaries of well-positioned dyads inferred from *in vitro* MNase-seq experiments (Valouev et al. 2011). We found that DNase-defined DARNs showed a significant overlap (~39 million, or 39% of 99 million bases covered by *in vitro* nucleosomes; $P = 0.025$), but a similar overlap was not present in either FAIRE-seq or permuted data sets (Table 1). We further found that annotated control regions were significantly different from DARNs when comparing inherent features of these annotations, including length, read coverage, correlation scores, spacing between annotations, and HMM posterior probability ($P < 10^{-10}$; for complete list of 16 features, see Supplemental Material). We used these differences to estimate the false-discovery rate (FDR) by developing a multivariate linear regression classifier based on these features that distinguished DNase-defined DARNs from control data annotations. This provided a confidence score for each DARNs and estimated an FDR for DARNs of 8.14% (for details, see Methods).

To further investigate the relationship between DARNs and annotated nucleosome positions, we plotted the *in vitro* dyad positions, as well as *in vivo* nucleosome occupancy (Stanford Nucleosome track, UCSC Genome Browser), relative to DARNs. We first noted that the ~600,000 *in vitro* dyads showed strong correspondence with nucleosome occupancy signal from *in vivo* MNase-seq supporting their accuracy (see Methods) (Supplemental Fig. S4A). We found that *in vitro* dyads were prevalent around the midpoints of DARNs but were not around regions called from the permuted data (Fig. 4C). Interestingly, the dyad signal was not centered at DARNs midpoints, but rather shows two distinct regions of significant enrichment ~60 bp upstream and downstream ($P < 0.05$). These dual crests, along with the 50-bp median DARNs length, suggest that DARNs were preferentially positioned on either side of the dyad within the *in vitro* predicted nucleosomes but did not generally overlap the dyad. A similar profile of dual enrichment peaks was detected when comparing *in vivo* MNase-seq data from the GM12878 LCL ($P < 0.05$) (Fig. 4D) and the K562 erythroleukemia cell line (data not shown). The lack of DARNs overlapping the exact center of nucleosomes may signify a disruption or attenuation of the digestion periodicity around or across the dyad, analogous to the previously reported reduction in 10-bp periodic dinucleotide signal at the dyad (Albert et al. 2007; Ioshikhes et al. 2011). We propose that digestion of the

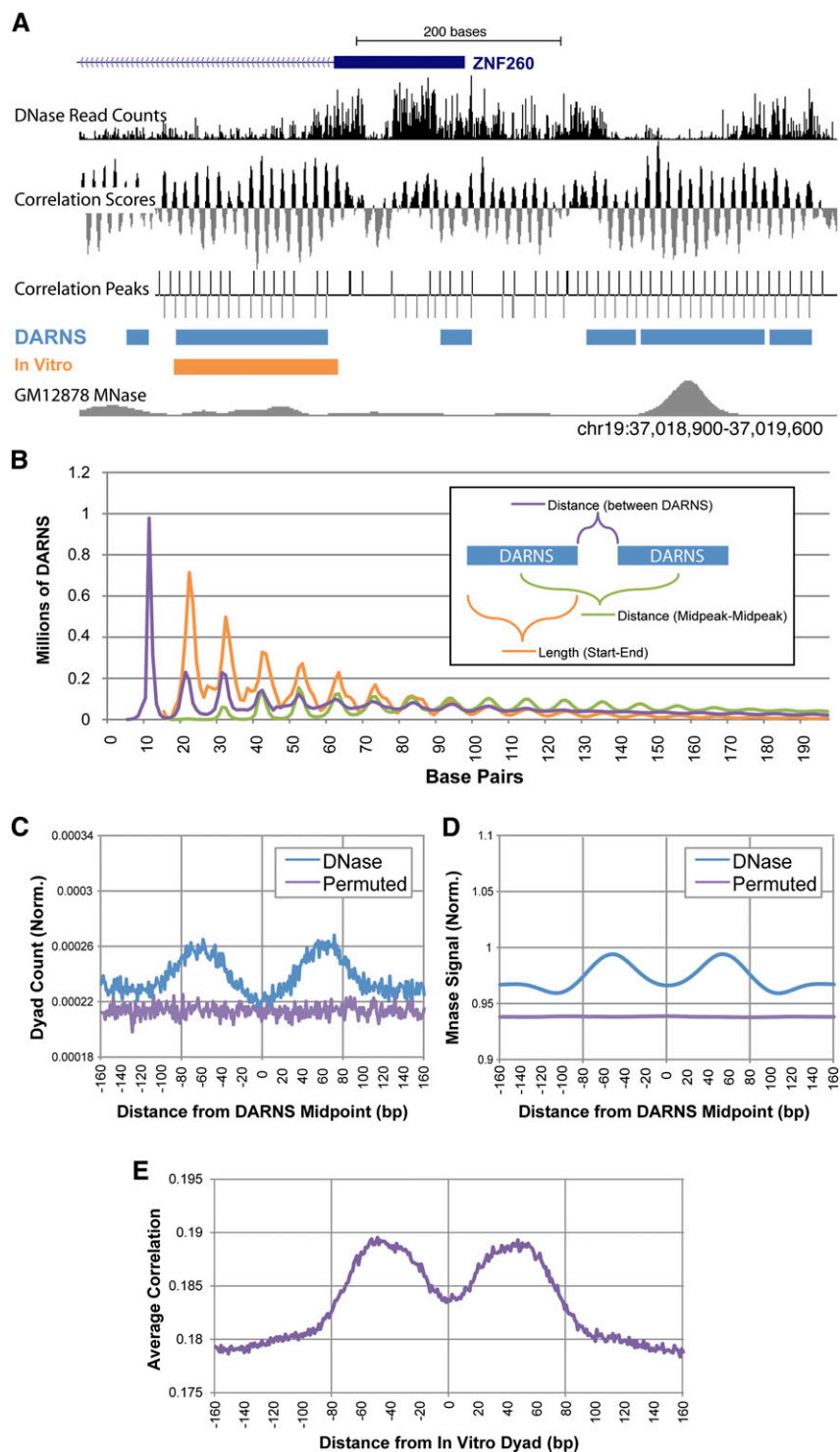



Figure 4. Hidden Markov model (HMM) identified DNase annotated regions of nucleosome stability (DARNS). (A) UCSC Browser screenshot of DNase-seq read counts, correlation scores, correlation peaks, and HMM identified DARNS (light blue boxes). In vitro positioned nucleosomes (orange box) (Valouev et al. 2011) and the in vivo GM12878 MNase-seq show a similar overlap with DARNS. (B) Length of and distance between DARNS display an ~ 10 -bp pattern. (C) Distribution of in vitro dyads shows twin crests of significant enrichment ($\sim [-80:-45]$ and $[45:80]$) around DARNS midpoint. (D) Distribution of in vivo GM12878 MNase signal shows twin crests of significant enrichment ($\sim [-80:-25]$ and $[25:80]$) around DARNS midpoint. (E) Distribution of average correlation at peaks from both strands shows a dip in scores near in vitro dyads.

DNA at the dyad is not as constrained as flanking DNA that is more directly associated with the histone core. This is supported by the high levels of DNase-seq reads around the in vitro dyads (Supplemental Fig. S5A) combined with their low average correlation scores (Fig. 4E). The coordinates of DARNS and correlation peaks within DARNS are available at <http://fureylab.web.unc.edu/datasets/darns/> and in Supplemental Material.

Properties of DARNS covering nucleosome halves

To further investigate the bimodal signal seen in Figure 4, C and D, we compiled subsets of DARNS covering the upstream or 5' half and the downstream 3' half of the nucleosome based on their relative position to the nearest in vitro dyad and labeled according to the orientation of the positive strand in the genome assembly sequence. We created these separate sets to ensure that this analysis was not due to strand alignment biases or artefacts. This resulted in $\sim 475,000$ 5' end and $498,000$ 3' end DARNS, with $\sim 269,000$ corresponding to opposite sides of the same nucleosome. With a total of $\sim 600,000$ annotated dyads, the number of dyads with a pair of DARNS covering both halves is significantly less than that expected by random chance ($P < 10^{-10}$); this suggests that our model was less likely to annotate both sides of the same nucleosome and may indicate varying exposure of nucleosome halves or a difference in stable histone contacts across the nucleosome. When we independently compared these 5' and 3' end DARNS to in vivo MNase-seq data, we detected enrichment in MNase-seq signal on only one side of the DARNS, further supporting that we correctly assigned the DARNS locations relative to the dyad (Supplemental Fig. S4B).

For each of the 5' and 3' end DARNS, we plotted the distribution of aligned positive- and negative-strand DNase-seq read counts, as well as MNase-seq data to show relative locations of the nucleosome and linker (for 5', see Fig. 5A; for 3', see Fig. 5B). In both plots, the region of the nucleosome covered by the DARNS demonstrated the strongest periodicity relative to the linker and the opposite side of the nucleosome. We observed a decrease in overall read counts in the presumed linker region, indicating reduced DNase I digestion between nucleosomes. A similar plot of DNase-seq read counts relative to in vitro dyads does not display

Table 1. Overlap of HMM nucleosome calls


		In vitro Nucleosomes	DUKE DNase	Permuted DNase	FAIRE	UC DNase
	# of Sites	616,857	14,368,288	12,069,380	11,773,074	12,882,169
	Coverage (bp)	98,687,672	891,534,284	518,081,818	537,821,557	693,421,171
	Coverage (%)	3.41%	30.77%	17.88%	18.56%	23.93%
In vitro	Coverage (bp)		38,675,758	19,118,288	20,097,313	31,586,925
	Overlap of Row (%)		39.19%	19.37%	20.36%	32.01%
	[P-Value]		[0.025]	[0.56]	[0.54]	[0.010]
DUKE DNase	Coverage (bp)	38,675,758		174,951,169	182,401,882	311,897,617
	Overlap of Row (%)	4.34%		19.62%	20.46%	34.98%
	[P-Value]	[0.025]		[0.65]	[0.59]	[<1e-10]
Permuted DNase	Coverage (bp)	19,118,288	174,951,169		105,530,733	140,257,510
	Overlap of Row (%)	3.69%	33.77%		20.37%	27.07%
	[P-Value]	[0.56]	[0.65]		[0.61]	[0.29]
FAIRE	Coverage (bp)	20,097,313	182,401,882	105,530,733		135,696,098
	Overlap of Row (%)	3.74%	33.91%	19.62%		25.23%
	[P-Value]	[0.54]	[0.59]	[0.61]		[0.93]
UC DNase	Coverage (bp)	31,586,925	311,897,617	140,257,510	135,696,098	
	Overlap of Row (%)	4.56%	44.98%	20.23%	19.57%	
	[P-Value]	[0.010]	[<1e-10]	[0.29]	[0.93]	

Data sets that show significant overlap ($P < 0.05$) are shaded light gray.

periodicity, further supporting that the midpoints of DARNs are not dyads (Supplemental Fig. S5A).

To test whether this pattern was due to a bias in DNase I digestion at the nucleotide level, we analyzed these regions in DNase-seq data from naked DNA extracted from the K562 cell line. We did detect a periodicity in digestion at these same sites that matched what we observed in fully constituted nucleosomal DNA, but at a greatly reduced amplitude (Supplemental Fig. S6) suggesting that this was not the predominant cause of this digestion pattern.

We also investigated dinucleotide base frequencies in these subsets of DARNs. Periodic weak ($W = A/T$) dinucleotides have been associated with well-positioned nucleosomes and tend to be out of phase with strong ($S = G/C$) dinucleotides (Satchwell et al. 1986; Segal et al. 2006). We noted that in both 5' end DARNs (Fig. 5C) and 3' end DARNs (Fig. 5D), SS dinucleotide levels were highest in DNA incorporated into the nucleosome and in phase with DARNs peaks, while WW dinucleotide levels were highest in the linker and out of phase with DARNs peaks. This is consistent with previous reports of nucleosomes with GC-rich cores and AT-rich flanks (Valouev et al. 2011) as well as SS dinucleotides aligning to exposed positions where the minor groove faces outward from the histone. A similar trend was evident in an equivalent plot relative to the in vitro dyads (Supplemental Fig. S5B). Not surprisingly, the periodicity in each plot was again strongest over the region of the nucleosome covered by the DARNs, supporting the relationship between rotational stability and dinucleotide pe-

riodicity. To determine whether the periodicity could be recovered at the in vitro dyads, we aligned them by the nearest negative-strand correlation peak. The periodicity in dinucleotide frequency was indeed visible and fairly symmetrical across the surrounding region (Supplemental Fig. S5C). Taken together, these results support the periodic features of DNA as it is incorporated into the nucleosome. We also note that as expected, all of these results are symmetric for the 5' and 3' end DARNs when adjusting for orientation to the dyad.

Comparison to DARNs in a single cell type and their relationship to genomic features

DNase-seq data (~1.26 billion reads) from 70 HapMap Yoruba LCLs was recently generated by the Pritchard and Gilad laboratories at the University of Chicago (UC) using the same DNase-seq protocol (Degner et al. 2012). To demonstrate the reproducibility and enable analysis in a single cell type, we annotated DARNs using these data. The UC data produced ~13 million DARNs covering ~690 million (23.9%) bases of the genome (Table 1). The UC data showed the same patterns as the Duke DNase-seq data with respect to aligned read spacing (Supplemental Fig. S7A), Fourier analysis (Supplemental Fig. S7B), and distribution of correlation scores compared with permuted DNase-seq data ($P < 10^{-10}$) (Supplemental Fig S7C). Likewise, the UC DARNs significantly overlapped in vitro dyads (31.6 million bases, or 32% of the in vitro

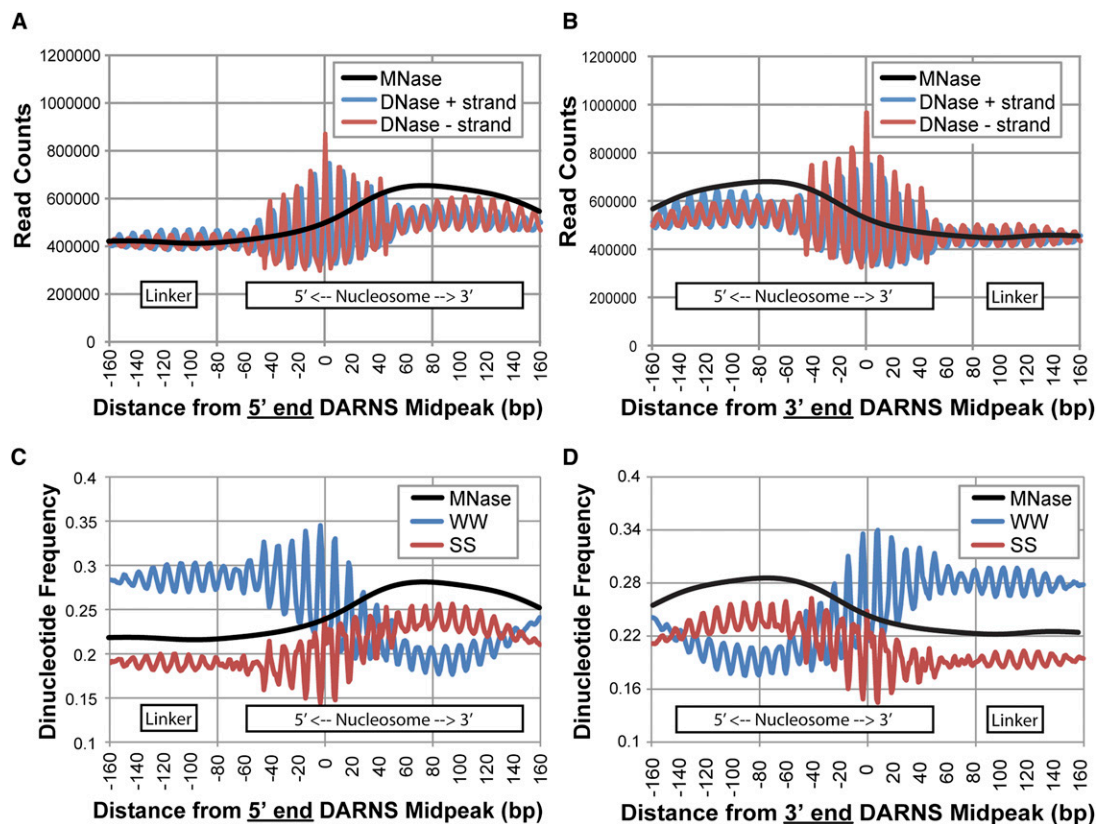


Figure 5. Properties of DARNs that map to the 5' end or 3' end of nucleosomes. In vitro dyads (Valouev et al. 2011) were used to distinguish DARNs that mapped to either the 5' end (A,C) or the 3' end (B,D) of the nucleosome (for details, see Methods). (A) DNase reads aligned by centermost negative-strand correlation peak (midpeak) of DARNs that map to the 5' end of the nucleosome exhibit a greater oscillation pattern and DNase signal on the 5' end of the nucleosome compared to either the linker or the 3' end of the nucleosome. In vivo LCL MNase-seq signal is transposed on top (black line) to help designate the locations of the nucleosome and the linker (indicated at bottom). (B) Same as A, but for DARNs that map to the 3' end of the nucleosome. (C,D) Same plot as A and B, but showing dinucleotide (W = A/T, S = C/G) frequency. Note that the central peak in the SS dinucleotide signal aligns to the midPeak of the DARNs, suggesting that they occur when the minor groove is exposed.

nucleosomes; $P < 0.010$), as well as with the positions of Duke DARNs called from the 49 samples (312 million bases, or 35% of Duke DARNs; $P < 10^{-10}$) (Table 1, bottom row). Therefore, the DNase I digestion pattern is a reproducible feature of this DNase-seq protocol and leads to consistent DARNs annotations from independent DNase-seq experiments.

Promoters often contain a nucleosome-free region (NFR) at the transcription start site (TSS) with a -1 nucleosome and a prominent $+1$ nucleosome followed by decreasingly well-positioned nucleosomes over the gene body (Jiang and Pugh 2009). We found a significant depletion in Duke (Fig. 6A) and UC (Fig. 6B) DARNs at TSSs compared with intergenic regions ($P < 10^{-10}$), but a 20% enrichment around the promoter area compared with background levels ($P < 10^{-10}$), particularly downstream. Similarly, we found reduced DARNs density immediately downstream from the transcription termination site (TTS) (Supplemental Fig. S8A). Gene deserts had reduced levels of DARNs despite the fact that in aggregate the 10.4-bp period was still present (see Methods) (Supplemental Fig. S8A,C). Thus, our data were in agreement with the established profile of nucleosomes around genes.

CpG islands, often found at gene promoters, occur in various sizes and have been shown to occlude nucleosome binding (Fenuil et al. 2012). When we plot the distribution of DARNs around CpG islands categorized by varying length thresholds, we find that the area depleted of DARNs increases with the size of the CpG islands

(Supplemental Fig. S8B), supporting the interference of nucleosome formation by these regions. Moreover, we find that DNase-seq reads mapping to CpG islands do not exhibit the periodicity associated with nucleosomes (Supplemental Fig. S8C).

Active DHS sites are nucleosome depleted by definition, and DHS has been shown to correlate with the strength of nearby nucleosome positioning (Gaffney et al. 2012). Both Duke (Fig. 6C) and UC (Fig. 6D) DARNs demonstrated a strong depletion in ubiquitous DHS sites detected in all cell types included in this study compared with random intergenic sites ($P < 10^{-10}$). Ubiquitous LCL DHS sites, found in all seven LCL samples from the Duke data, also demonstrated DARNs depletion for both sets ($P < 10^{-10}$). These features indicate that DARNs reproduce the depletion of nucleosomes at DHS sites that are active in all relevant cell types.

We used the UC DARNs from the LCLs to explore their profile at variably open DHS sites present in only LCLs compared with the non-LCL cell types in the Duke data. To do this, we divided the cumulative set of DARNs into three subsets: (1) 6.7 million Duke-specific DARNs covering 315 million bases (10.9% of the genome); (2) 5 million UC-specific DARNs covering 207 million bases (7.2%); and (3) 9 million DARNs annotated by both covering 312 million bases (10.8%). We considered the Duke-specific DARNs to be largely ubiquitous because they were only found by the diverse cell type data and considered the UC-specific DARNs to be cell-type-

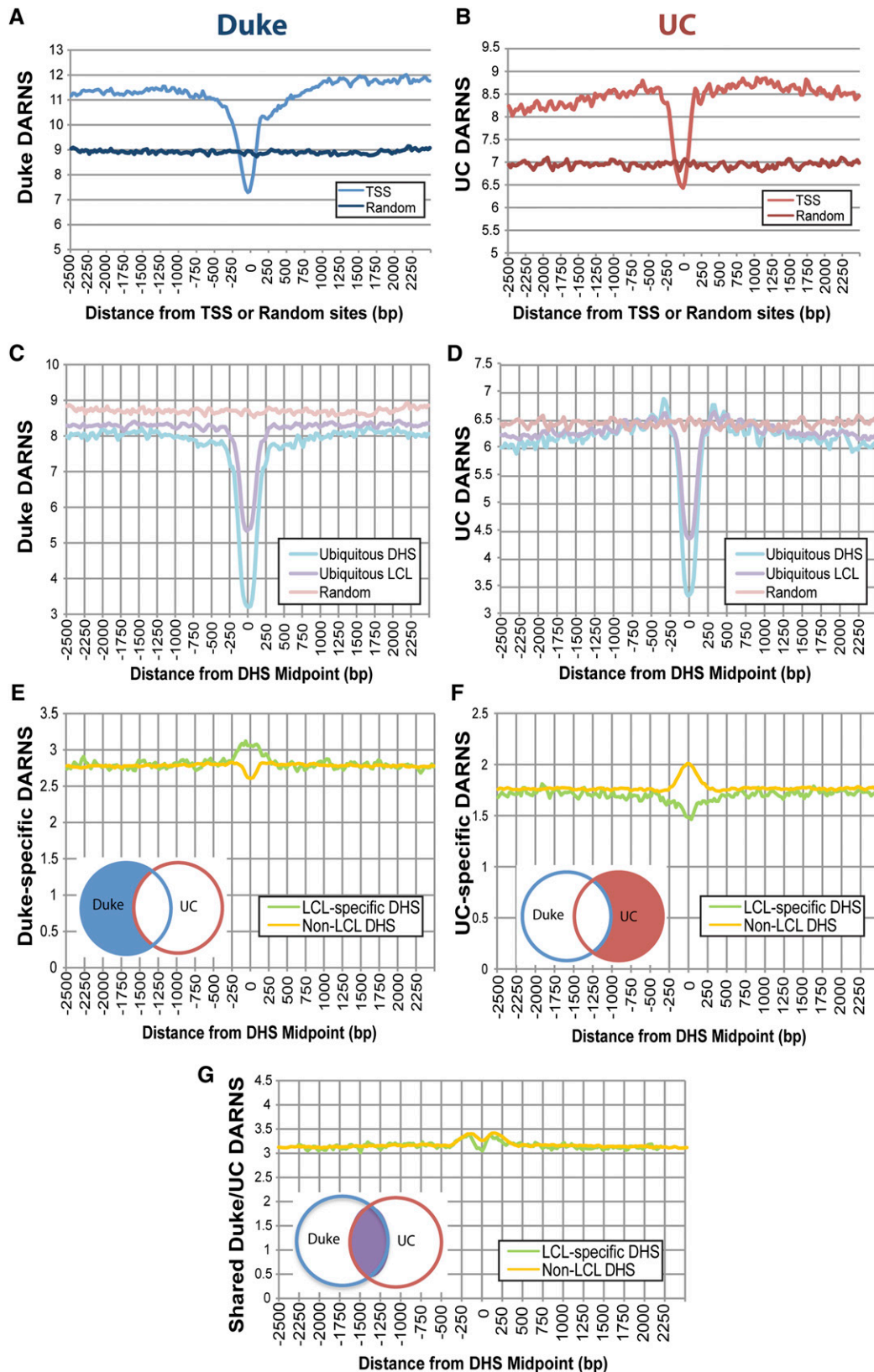


Figure 6. Properties of DARNs around transcription start sites (TSSs) and DHS sites. (Y-axis) The normalized density of DARNs. DARNs identified from Duke (*left*) and the University of Chicago (UC; *right*) (Degner et al. 2012). DNase-seq data are depleted around TSS (A,B) and enriched in surrounding areas (particularly downstream toward the gene body) relative to random intergenic coordinates. (C,D) DARNs (*left*, Duke; *right*, UC) are most depleted around ubiquitous DHS sites present in all cell types (ubiquitous DHS), as well as ubiquitous DHS sites identified in all LCL cell lines (ubiquitous LCL). DARNs were identified as Duke-specific (E), UC-specific (F), and shared by both data sets (G). (E) Duke-specific DARNs, but depleted around non-LCL DHS sites (yellow) and enriched around LCL-specific DHS (green). (F) In contrast, UC-specific DARNs are enriched for non-LCL DHS sites and depleted around LCL-specific DHS sites. (G) DARNs shared by both data sets show similar patterns around these DHS sites.

specific because they were only found by the LCL data. We found a Duke-specific enrichment ($P < 10^{-10}$) (Fig. 6E) and UC-specific depletion ($P < 10^{-10}$) (Fig. 6F) of DARNs over the middle of DHS sites present in LCLs but not in other cell types (LCL-specific DHS). In contrast, DHS sites present only in non-LCL cell types (non-LCL DHS) showed the reverse trend with a Duke-specific depletion ($P < 10^{-10}$) and UC-specific enrichment ($P < 10^{-10}$) of DARNs. The shared DARNs showed similar profiles for both sets of DHS sites (Fig. 6G). Likewise, when we analyzed DHS sites detected in a single non-LCL cell type (HUVEC), Duke-specific DARNs and shared DARNs were depleted ($P < 10^{-10}$) (Supplemental Fig. S9). This indicates that DARNs common to Duke and UC data were among the strongest and most-conserved across all cell types, while cell-type-specific DARNs were depleted from DHS sites present within that cell type and enriched in cell types where the DHS site was absent. Therefore, this suggests that many inactive regulatory regions are occupied by rotationally stable nucleosomes.

Discussion

In this study, we provide evidence that the rotational positioning of nucleosomes at individual loci can be annotated from genome-wide DNase I digestion patterns derived from DNase-seq data and that many regions of the genome maintain nucleosome rotational stability across diverse cell types. In addition to enabling us to explore ubiquitously stable nucleosomes, we found evidence of cell-type-specific nucleosome positioning by comparing to a second, independent DNase-seq data set from a single cell type. With future higher throughput sequencing, it will be possible to use our model on single-cell-type DNase-seq experiments to further explore cell type specificity. We note that using DNase-seq in this manner signifies a distinct approach from MNase-seq in uncovering properties of and annotating nucleosome positions. Although DNase-seq will likely not supplant MNase-seq for identifying translational nucleosome positions, it can provide a complementary view of stable nucleosomes and provide unique high-resolution information on rotational positioning.

DARNs provide spatial information regarding the orientation of DNA in a nucleosome. Therefore, they can be used to give context to or align other features of DNA such as nucleotide frequencies and regulatory motifs. This was illustrated by aligning dinucleotide periodicities around *in vitro* dyads (Supplemental Fig. 5C). Since DNase I is known to digest in the minor groove of DNA, positions corresponding to correlation peaks in DARNs will indicate where the minor grooves are likely facing away from the histone surface (Noll 1974; Cousins et al. 2004). A comparable connection was shown between the rotational positioning of nucleosomes and DNA methylation in the minor groove (Chodavarapu et al. 2010).

Many DARNs appear to cover only one-half of annotated nucleosomes. We hypothesize that the periodic pattern that forms the basis of our model is weaker at the dyad, preventing DARNs from extending across to the opposite half. Since a nucleosome is composed of DNA wrapped around two histone tetramers, we suggest the periodic constraint imposed on DNase digestion is relaxed as it transitions from one histone tetramer to the next. This may be related to the loss in conservation of the dinucleotide periodicity at the dyad (Albert et al. 2007; Ioshikhes et al. 2011). Additionally, it has been proposed that pairs of nucleosomes incorporated into higher-order structures, like 30-nm fibers, are asymmetrically protected from DNase digestion (Staynov 2000), which may contribute to DARNs mapping to only one side of the nucleosome but not extending across the dyad.

The positioning of ubiquitous and cell-type-specific DARNs may help our understanding of cell-specific gene regulation. Our results suggested that although variably open DHS sites were strongly depleted of DARNs when active, DARNs reappear in cell types where the DHS sites are closed. For regions that contain a well-positioned nucleosome, our ability to determine the orientation of the major and minor grooves relative to the histone surface may allow us to better understand how TFs initially access these nucleosome bound *cis*-regulatory elements in response to changing cell conditions. For example, a binding site that maps to DNA in a nucleosome can be exposed on cue by chromatin remodelers that evict the nucleosome or shift the rotational settings (Jiang and Pugh 2009). Several models have been proposed for estimating TF binding while taking into account predicted nucleosome occupancy (Narlikar et al. 2007; Raveh-Sadka et al. 2009); these may benefit from incorporating relevant information from *in vivo* DARNs.

Finally, DARNs may be used as a starting point for investigating higher-order nucleosome structures *in vivo*. There are two proposed structures for how linear arrays of nucleosomes form 30-nm fibers (Tremethick 2007). These hierarchical organizations of nucleosomes are likely to form inaccessible regions that may result in a recognizable pattern of DNase I digestion (Staynov 2000): This may be reflected as higher-order patterns in our data. Nucleosomes are also further compacted into chromosomal structures like centromeres, telomeres, and heterochromatin (van Holde and Zlatanova 1995). DNase-seq data may be able to contribute to our understanding of how nucleosomes are incorporated into these chromosome architectures, which will help to elucidate how the genome is spatially organized.

Methods

DNase-seq and FAIRE-seq data

The Duke DNase-seq data is a composite of the results from the following 49 samples (seven LCLs and 42 unique cell types) with replicates: GM12891, GM12892, GM12878, GM19238, GM19239, GM19240, GM18507, A549, Chorion, CLL, D721, E_myoblast, FB0167P, FB8470, Fibroblasts_park, FSHD_myoblast, H1_ES, H54, H9_ES, HeLaS3, HeLaS3_IFNA, Hepatocytes, HepG2, HMEC, HPDE6, Huh7_5, Huh7, HUVEC, iPS, K562, LHSR, LHSR_induced, LnCAP, LnCAP_induced, MCF7, Melanocyte, Myoblast, Myometrial, Myotube, NHEK, Osteoblast, Pancreatic_islets_dedif, Pancreatic_islets, PATu8988T, SM_SFM, Stellate, T47, TE, Trophoblast. The UNC Chapel Hill FAIRE-seq data is a composite of the results from 19 samples (a subset of the 49) with replicates from the Lieb laboratory (Giresi et al. 2007): GM12891, GM12892, GM12878, GM19238, GM18507, A549, D721, H1_ES, H54, HeLaS3, HeLaS3_IFNA, HepG2, HUVEC, K562, LHSR, LHSR_induced, NHEK, Pancreatic_islets, Trophoblast. More information on these cell types can be found at genome.ucsc.edu/ENCODE/cellTypes.html. The UC DNase-seq data originated from LCLs from 70 individuals with replicates (Degner et al. 2012).

The Naked DNA data set was generated by purifying total genomic DNA from unfixed K562 cells using phenol extraction and ethanol precipitation. This naked DNA was treated with DNase I in the same manner as the above cell types and used to generate a DNase-seq library.

All Duke DNase-seq and UNC FAIRE-seq data were aligned to the hg19 Human Genome Assembly. The raw UC DNase-seq data were initially aligned to hg18, but DARNs are given in hg19 (see Nucleosome Annotations and Comparisons). DNase-seq and FAIRE-

seq were generated as part of the ENCODE Project and are available at the UCSC Genome Browser (<http://genome.ucsc.edu/ENCODE>) or at the NCBI Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>) under accession numbers GSE32970 (DNase-seq) and GSE35239 (FAIRE-seq).

Combining multiple cell types

Sequence files for individual cell types consisted of reads combined across all replicates. In our merged data, we counted, for each base, the number of cell types with an aligned read starting at that base, considering each strand separately. The value at each base on each strand was used as the representative aligned sequence read count for that base. This was done for each of the Duke DNase-seq, UC DNase-seq, and FAIRE-seq data sets.

Fourier transform analysis

The *fft* (fast Fourier transform) function in MATLAB was used to calculate the Fourier transform for 1000-bp sliding windows (100-bp overlap). The dominant frequency in each window was set to the maximum of all frequencies greater than 0.001. These values were inverted to calculate the period and used to create the histogram of the dominant periods across all windows.

Pattern of DNase I digestion around the nucleosome and correlation peaks

The frequencies of distances between DNase I digestion sites (y -axis) were extracted for the distances 31–76 bases (x -axis) from the plot shown in Figure 1B. This subset was selected to avoid the artifact at ~20 bp resulting from sequence length. Moreover, this pattern is large enough to capture the DNase digestion period but small enough to match multiple times within typical nucleosome regions. We “flattened” these values by setting minima points to zero and rescaling the remaining values accordingly and then mirroring this pattern to the left to create a symmetric pattern of DNase I digestion around the nucleosome (Supplemental Fig. S3A). For each strand, we calculated the correlation between the nucleosome pattern and the read counts in 91-bp sliding windows across the genome and assigned the score to the center base. We estimated the significance of the difference in variance between the distributions of DNase and permuted correlations using an *F*-test. Correlation peaks were defined as local maximums that were bounded by negative correlation values.

HMM for determining DARNs

Genomic positions were labeled as corresponding to a positive correlation peak (1), a negative correlation peak (−1), or neither/both (0), with each strand being considered separately. The resulting strings of labels were input into an HMM that consisted of 14 states with transitions as depicted in Supplemental Figure S3B. In the single background state, the emissions probabilities for the base labels (−1, 0, 1) were empirically derived from random genomic regions. The remaining “nucleosome” states represented the expected cycle between positive and negative correlation peaks within the nucleosome; the path through the states as well as the associated emission and transition probabilities were derived from observed frequencies and distances between peaks. A region with the desired pattern, which we refer to as a DARNs, starts and ends in the nucleosome state representing either the negative (−1) or positive (1) peak. Then, the number of zero or “no peak” states between correlation peak states reflects the observed probable spacing. The *hmmviterbi* function in MATLAB was used to determine the op-

timal path through the HMM for each chromosome. The Mid-Peak was set to the middle most negative peak within each DARNs. We trained a regression classifier to estimate the FDR as described in the Supplemental Material. Coordinates, confidence scores, and correlation peaks for DARNs are available at <http://fureylab.web.unc.edu/datasets/darns/> and in Supplemental Material.

Nucleosome annotations and comparisons

The program *featureBits*, part of the UCSC toolbox (Kent et al. 2002), was used to calculate base overlaps between genomic regions annotated as DARNs, control regions, and extended *in vitro* dyads. *P*-values for the percentage of bases in the overlaps were calculated using hypergeometric tests. The locations of *in vitro* dyads inferred from well-positioned nucleosomes were based on MNase-seq experiments performed in the Sidow laboratory on naked DNA combined with histones (Valouev et al. 2011). To determine overlap, we generated *in vitro* nucleosome boundaries by extending 80bp in both directions from the dyad. The program *liftOver*, also part of the UCSC toolbox, was used to translate the positions of both the *in vitro* nucleosomes and the UC DARNs from the hg18 to hg19 assembly (Kent et al. 2002). We removed sites that corresponded to “blacklisted” regions as designated by the ENCODE project and available within the UCSC Genome Browser, Mappability annotation track (<http://genome.ucsc.edu/>; A. Kundaje and E. Birney, unpubl.). The regions of significant enrichment in the *in vitro* dyad profile (Fig. 4C) scored $P < 0.05$ after Bonferroni correction of the Poisson distribution derived from the permuted data. The MNase-seq data sets for the GM12878 LCL and the K562 leukemia cell line were produced by the Snyder laboratory and are also available within the UCSC Genome Browser, Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU (<http://genome.ucsc.edu/>; M. Snyder, D. Raha, S. Johnson, E. Winters, A. Sidow, Z. Weng, C. Smith, P. Lacroute, P. Cayting, A. Kundaje, unpubl.). These data were downloaded as a processed file with normalized nucleosome occupancy scores for each base. The regions of enrichment in the GM12878 MNase-seq occupancy signal profile (Fig. 4D) were significant at $P < 0.05$ after Bonferroni correction based on the Gaussian distribution derived from the permuted data.

Distribution of DARNs around genomic features

For comparison, we chose ~20,000 random intergenic sites from large alignable regions (>2500-bp excluding repetitive regions and alignment gaps) across the genome. We determined the locations of TSS, TTS, gene deserts (>100 kb without gene), and CpG islands using annotations from the UCSC Genome Browser downloaded for hg19 (<http://genome.ucsc.edu/>). DHS sites are annotated according to the method described previously (Song et al. 2011), and the sets are defined as in Supplemental Table S1. To investigate the distribution of DARNs around genomic features, we summed the number of instances in which DARNs overlapped each base around the reference point (TSS, TTS, or midpoint) of the feature (DHS site, CpG Island, gene desert, or random intergenic) and binned into 25-bp windows. Then, we normalized for the number of sites in the set of genomic features (Fig. 6; Supplemental Fig. S8A,B) or the number of bases covered by each type of DARNs (Supplemental Fig. S9). For the UC DARNs, we used the translated hg19 assembly positions. *P*-values were calculated using the χ^2 statistic by comparing the number of DARNs covering the window of the TSS or DHS midpoint with the number covering random intergenic sites. Fold enrichment of promoter regions was calculated similarly by comparing to the average random level.

Acknowledgments

We thank the Lieb laboratory at UNC Chapel Hill for access to FAIRE-seq data, the Pritchard laboratory at the University of Chicago for access to and help with their DNase-seq data, and the Sidow laboratory at Stanford University for access to and help with their *in vitro* dyad nucleosome data. This work was supported by a Canadian NSERC PGS-D graduate fellowship (D.R.W.), NIH Grant U54-HG004563 (T.S.F., G.E.C.), and the University of North Carolina Cancer Research Fund (T.S.F.).

References

- Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF. 2007. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446**: 572–576.
- Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**: 311–322.
- Brogaard K, Xi L, Wang P, Widom J. 2012. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**: 496–501.
- Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, et al. 2010. Relationship between nucleosome positioning and DNA methylation. *Nature* **466**: 388–392.
- Cousins DJ, Islam SA, Sanderson MR, Proykova YG, Crane-Robinson C, Staynov DZ. 2004. Redefinition of the cleavage sites of DNase I on the nucleosome core particle. *JMB* **335**: 1199–1211.
- Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, Leon SD, Michelini K, Lewellen N, Crawford GE, et al. 2012. DNASE1 sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**: 390–394.
- The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.
- Felsenfeld G, Groudine M. 2003. Controlling the double helix. *Nature* **421**: 448–453.
- Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I, et al. 2012. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res* **22**: 2399–2408.
- Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K, Widom J, Gilad Y, Pritchard JK. 2012. Controls of nucleosome positioning in the human genome. *PLoS Genet* **8**: e1003036.
- Giresi PG, Kim J, McDaniel RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**: 877–885.
- Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57**: 159–197.
- Hu G, Schones DE, Cui K, Ybarra R, Northrup D, Tang Q, Gattinoni L, Restifo NP, Huang S, Zhao K. 2011. Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res* **21**: 1650–1658.
- Ioshikhes I, Hosid S, Pugh BF. 2011. Variety of genomic DNA patterns for nucleosome positioning. *Genome Res* **21**: 1863–1871.
- Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: Advances through genomics. *Nat Rev Genet* **10**: 161–172.
- Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**: 362–366.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **6**: 996–1006.
- Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* **18**: 1073–1083.
- McDaniel R, Lee B-K, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, et al. 2010. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**: 235–239.
- Narlikar L, Gordan R, Hartemink AJ. 2007. Nucleosome occupancy information improves *de novo* motif discovery. *Res Comp Mol Biol* **4453**: 107–121.
- Noll M. 1974. Internal structure of the chromatin subunit. *Nucleic Acids Res* **1**: 1573–1578.
- Pugh BF. 2010. A preoccupied position on nucleosomes. *Nat Struct Mol Biol* **17**: 923.
- Raveh-Sadka T, Levo M, Segal E. 2009. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res* **19**: 1480–1496.
- Richmond TJ, Davey CA. 2003. The structure of DNA in the nucleosome core. *Nature* **423**: 145–150.
- Satchwell SC, Drew HR, Travers AA. 1986. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* **191**: 659–675.
- Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**: 887–898.
- Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang J-PZ, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* **442**: 772–778.
- Sollner-Webb B, Melchior W Jr, Felsenfeld G. 1978. DNase I, DNase II, and staphylococcal nuclease cut at different, yet symmetrically located, sites in the nucleosome core. *Cell* **14**: 611–627.
- Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee B-K, Sheffield NC, Gräf S, Huss M, Keefe D, et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* **21**: 1757–1767.
- Staynov DZ. 2000. DNase I digestion reveals alternating asymmetrical protection of the nucleosome by the higher order chromatin structure. *Nucleic Acids Res* **28**: 3092–3099.
- Thurman RE, Rynes E, Humbert R, Vierstra J, Maurana MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–78.
- Tremethick DJ. 2007. Higher-order structures of chromatin: The elusive 30 nm fiber. *Cell* **128**: 651–654.
- Trifonov EN, Sussman JL. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc Natl Acad Sci* **77**: 3816–3820.
- Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. 2011. Determinants of nucleosome organization in primary human cells. *Nature* **474**: 516–520.
- van Holde K, Zlatanova J. 1995. Chromatin higher order structure: Chasing a mirage? *J Bio Chem* **270**: 8373–8376.
- Wang JC. 1979. Helical repeat of DNA in solution. *Proc Natl Acad Sci* **76**: 200–203.
- Wang J-P, Fondufe-Mittendorf Y, Xi L, Tsai G-F, Segal E, Widom J. 2008. Preferentially quantized linker DNA lengths in *Saccharomyces cerevisiae*. *PLoS Comput Biol* **4**: e1000175.
- Wu C, Bingham PM, Livak KJ, Holmgren R, Elgin SCR. 1979. The chromatin structure of specific genes: 1. Evidence for higher order domain of defined DNA sequence. *Cell* **16**: 797–806.
- Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, Liu XS, Struhl K. 2009. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions *in vivo*. *Nat Struct Mol Biol* **16**: 847–852.

Received December 18, 2012; accepted in revised form April 30, 2013.