Genome-Wide Prediction of Nucleosome Occupancy in Maize Reveals Plant Chromatin Structural Features at Genes and Other Elements at Multiple Scales\textsuperscript{[W][OA]}

Justin A. Fincher, Daniel L. Vera, Diana D. Hughes, Karen M. McGinnis, Jonathan H. Dennis, and Hank W. Bass\textsuperscript{*}

Department of Biological Science, Florida State University, Tallahassee, Florida 32306–4295 (J.A.F., D.L.V., D.D.H., K.M.M., J.H.D., H.W.B.); and Department of Computer Science, Florida State University, Tallahassee, Florida 32306–4530 (J.A.F.)

The nucleosome is a fundamental structural and functional chromatin unit that affects nearly all DNA-templated events in eukaryotic genomes. It is also a biochemical substrate for higher order, cis-acting gene expression codes and the monomeric structural unit for chromatin packaging at multiple scales. To predict the nucleosome landscape of a model plant genome, we used a support vector machine computational algorithm trained on human chromatin to predict the nucleosome occupancy likelihood (NOL) across the maize (Zea mays) genome. Experimentally validated NOL scores provide a novel genomic annotation that highlights gene structures, repetitive elements, and chromosome-scale domains likely to reflect regional gene density. We established a new genome browser (http://www.genomaize.org) for viewing support vector machine-based NOL scores. This annotation provides sequence-based comprehensive coverage across the entire genome, including repetitive genomic regions typically excluded from experimental genomics data. We find that transposable elements often displayed family-specific NOL profiles that included distinct regions, especially near their termini, predicted to have strong affinities for nucleosomes. We examined transcription start site consensus NOL plots for maize gene sets and discovered that most maize genes display a typical +1 nucleosome positioning signal just downstream of the start site but not upstream. This overall lack of a −1 nucleosome positioning signal was also predicted by our method for Arabidopsis (Arabidopsis thaliana) genes and verified by additional analysis of previously published Arabidopsis MNase-Seq data, revealing a general feature of plant promoters. Our study advances plant chromatin research by defining the potential contribution of the DNA sequence to observed nucleosome positioning and provides an invariant baseline annotation against which other genomic data can be compared.

Eukaryotic genomes consist of DNA that is packaged together with histone proteins into chromatin. The basic subunit of chromatin, the nucleosome, is composed of approximately 150 bp of DNA wrapped 1.65 times around a histone octamer. This octamer is composed of two copies each of the four histone proteins H2A, H2B, H3, and H4 (for review, see Kornberg and Lorch, 1999) and has been described at atomic resolution (Luger et al., 1997). Nucleosomes can be positioned along a DNA molecule to form a 10-nm fiber described as “beads-on-a-string,” a structure that is visible by electron microscopy (Finch and Klug, 1976; Finch et al., 1977). Models for higher order chromatin organization are rapidly advancing, and views on how to measure and interpret chromatin fiber structural data are varied and controversial (Lieberman-Aiden et al., 2009; Li and Reinberg, 2011; Mirny, 2011; Nishino et al., 2012; Schlick et al., 2012).

The distribution of nucleosomes is controlled by a combination of factors including chromatin regulatory complexes and features intrinsic to the DNA sequence. Steady-state nucleosome distribution affects DNA-binding interactions required for nuclear processes such as transcription, replication, recombination, repair, and transposition (Jiang and Pugh, 2009). The nonrandom distribution of nucleosomes along the DNA strand is influenced by the cis-acting physiochemical properties of the DNA sequences that encode local relative affinities for nucleosome binding (Drew and Travers, 1985; for review, see Fincher and Dennis, 2011). In addition, trans-acting factors, such as those involved in histone modifications or chromatin remodeling, also affect nucleosome positioning (Kingston and Narlikar, 1999; Kouzarides, 2007; Zhang et al., 2011). Typical locations of covalent histone modifications near genes are well documented (Jiang and Pugh, 2009), as are the roles that ATP-dependent chromatin remodelers play in gene regulation. How these histone modifications and nucleosome distributions might interact to affect chromatin

\textsuperscript{1} This work was supported by the National Science Foundation (grant no. IOS-1025954) and the American Heart Association Greater Southeast Affiliate (grant no. 12POST12070101 to J.A.F.).

\textsuperscript{*} Corresponding author; e-mail bass@bio.fsu.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Hank W. Bass (bass@bio.fsu.edu).

\textsuperscript{[W]} The online version of this article contains Web-only data.

\textsuperscript{[OA]} Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.113.216432
function remains largely unknown. Knowledge of nucleosome occupancy, therefore, provides a critically important point of reference for understanding the role of nucleosomes in genome regulation.

Empirical data on nucleosome distribution have been valuable in determining the chromatin landscape of eukaryotic genomes. Micrococcal nuclease (MNase) is a preferentially internucleosomal cleavage reagent widely used to delineate nucleosome-protected from nucleosome-free regions of DNA in preparations of fixed nuclei. The isolated approximately 150-bp DNA fragments from MNase-digested nuclei represent nucleosome footprints that can be quantified and mapped onto genomic sequences by microarray or sequencing technologies (for review, see Spetman et al., 2011).

High-resolution maps of nucleosome occupancy in budding yeast (Saccharomyces cerevisiae) reveal the basic chromatin features associated with regulatory elements and transcription start sites (TSSs; Yuan et al., 2005). From these and related studies, canonical chromatin landscapes of nucleosome positioning around transcription initiation sites are emerging. For example, several nucleosome-mapping studies established the existence of nucleosome-free regions, an important observation confirming earlier models of nucleosome phasing such as the statistical positioning of nucleosomes (Kornberg, 1981; Gupta et al., 2008; Yuan and Liu, 2008; Zhang et al., 2011).

Maps of nucleosome distribution have led to the development of computational models that predict nucleosome occupancy from the DNA sequence. General properties of nucleosome-forming and nucleosome-inhibitory sequences were described nearly 30 years ago (Drew and Travers, 1985). Since then, sophisticated computational tools have been used to devise rules that predict the intrinsic nucleosome occupancy likelihood (NOL) for any DNA sequence. Seminal work by Segal et al. (2006) proposed a computational model for the sequence-based prediction of nucleosome positioning in S. cerevisiae that used dinucleotide distributions collected from a training set of nucleosome-forming sequences. This work verified the importance of DNA sequence in determining nucleosome position and revealed the utility of combined genomics and computational approaches for chromatin research. A complementary computational model was proposed (Peckham et al., 2007; Gupta et al., 2008) that is discriminative, rather than generative, and uses only sequences that show the strongest or weakest affinity for forming nucleosomes. In these studies, DNA sequences protected from MNase digestion were queried with DNA microarrays (Dennis et al., 2007; Ozsolak et al., 2007), and the probe sequences with the highest and lowest nucleosome occupancy signals were used to train support vector machine (SVM) classifiers that can be applied to any genomic sequence (Gupta et al., 2008). A comparative assessment of available nucleosome occupancy prediction algorithms revealed that the SVM trained on human chromatin worked well on related species with relatively large, complex genomes (Tanaka and Nakai, 2009). Although several opisthokont genomes have been relatively well characterized by these approaches, very little is known about the genome-wide nucleosome landscape of members of the plant kingdom.

Here, we describe the use of the human SVM model trained on human chromatin (Gupta et al., 2008) as a predictor of nucleosome occupancy in maize (Zea mays). Given the high degree of conservation of histone protein sequences and the conserved structural role of the canonical histone octamer, we expected that the human SVM algorithm would produce features for NOL in maize similar to those in humans, including, for example, specific signatures for nucleosome positioning around TSSs. We provide a comprehensive genome-wide NOL profile for the maize genome, describe our results from analyzing these predictions, and introduce a new maize genome bioinformatic resource, Genomaize (http://www.genomaize.org), to facilitate structural and functional genomics of maize.

RESULTS

To examine chromatin structure at the level of nucleosome occupancy, we applied a previously validated SVM model (Gupta et al., 2008) to maize genomic DNA. This model calculates a NOL score to any 50-bp sequence on the basis of empirical nucleosome occupancy data (A375 model; Gupta et al., 2008). We initially determined how well this human-trained SVM would perform on a well-characterized plant gene, maize Suc Synthase1 (also known as Shrunken1). The resulting NOL profile was remarkably similar to those of human genes in that it displayed prominent NOL peaks near the TSS and within exons (data not shown). We next extended this NOL prediction to the entire maize genome and analyzed the resulting data in relation to genes, repetitive elements, and whole chromosomes.

Genome-Wide Maize NOL Score Distributions

We carried out a genome-wide NOL prediction for the approximately 3 billion-bp maize genome (unmasked B73 RefGen_v2; Schnable et al., 2009). Negative NOL scores predict underrepresentation in nucleosome-protected DNA fractions, whereas positive NOL scores predict overrepresentation in nucleosome-protected DNA fractions. We first analyzed the basic statistical properties of NOL scores using a 25-bp step size. The NOL score frequency distribution histogram for the maize genome is shown in Figure 1A, along with that of human. Minimum, maximum, median, and upper and lower 1% and 5% values are indicated on the histograms and tabulated for these and other sequence data sets for comparison (Fig. 1; Table I; Supplemental Fig. S1).

The NOL values for the human genome exhibit a nonnormal distribution and range from −5.2 to +9.0, with 50% of all scores falling between −1.47 and −0.12, a range of 1.35. In humans, the top 5% of NOL scores

null
are above 0.86 and the bottom 5% are below –1.99. The maize genome (Fig. 1B; Table I) showed a NOL score distribution similar to that of humans, but with a markedly larger proportion of high values. For example, the top 5% of maize NOL scores were higher than the top 1% of human NOL scores. The NOL values for maize ranged from –5.6 to +10.0, and 50% of all scores fell between –1.08 and +0.91, a range of 1.99. We also compared the distribution of NOL scores for the entire maize genome with that of maize chromosome 9 alone using a 1-bp step size; the two showed nearly identical distributions (Table I; Supplemental Fig. S1, compare B and C).

To compare these values with those from randomized sequences with the same base composition, we plotted the NOL score distribution for a randomized maize chromosome 9 sequence (Supplemental Fig. S1D). The resulting distribution was dramatically different, resembling a normal distribution with less extreme values at each end. Notably, 10% of maize chromosome 9 sequences, but only 1% of scores from randomized chromosome 9, exceeded a score of 2.4. Similarly, 9% of the maize chromosome 9 sequences, but only 1% of the randomized maize chromosome 9 sequences, fell below –1.62.

We also considered nonnuclear genomes as another type of control comparison. These genomes lack eukaryotic nucleosome organizing information, yet they contain biological sequence complexity, with evolved genes and regulatory elements. The NOL scores for maize mitochondrial and chloroplast genomes showed a very slight leftward skew but, like the randomized maize chromosome 9, lacked a significant number of extreme values (Table I; Supplemental Fig. S1, E and F). Similarly, the NOL values for the Escherichia coli genome (Supplemental Fig. S1G) resembled those of

![Figure 1](https://example.com/figure1.png)

Figure 1. Histograms, box plots, and percentile scores of NOL for two genomes. The entire human (hg19) and maize (B73 RefGen_v2) genomes were generated with a sliding 50-bp window and 25-bp step size. The box plots for genic regions, exons, and introns for maize are shown as an inset to B.

<table>
<thead>
<tr>
<th>Organism/Region</th>
<th>Region¹</th>
<th>Total bp</th>
<th>Minimum</th>
<th>0.01</th>
<th>0.05</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.95</th>
<th>0.99</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Whole genome</td>
<td>3,160</td>
<td>–5.19</td>
<td>–2.30</td>
<td>–1.99</td>
<td>–1.47</td>
<td>–0.95</td>
<td>–0.12</td>
<td>0.86</td>
<td>2.07</td>
<td>9.01</td>
</tr>
<tr>
<td>Maize</td>
<td>Whole genome</td>
<td>2.1</td>
<td>–5.63</td>
<td>–2.15</td>
<td>–1.82</td>
<td>–1.08</td>
<td>–0.25</td>
<td>0.91</td>
<td>2.93</td>
<td>4.26</td>
<td>9.96</td>
</tr>
<tr>
<td>Maize chromosome 9</td>
<td>Chromosome 9</td>
<td>159</td>
<td>–5.63</td>
<td>–2.15</td>
<td>–1.82</td>
<td>–1.08</td>
<td>–0.25</td>
<td>0.93</td>
<td>2.94</td>
<td>4.26</td>
<td>9.96</td>
</tr>
<tr>
<td>Maize chromosome 9</td>
<td>Randomized</td>
<td>159</td>
<td>–2.95</td>
<td>–1.62</td>
<td>–1.19</td>
<td>–0.47</td>
<td>0.10</td>
<td>0.71</td>
<td>1.67</td>
<td>2.38</td>
<td>6.17</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>Whole genome</td>
<td>0.57</td>
<td>–2.83</td>
<td>–1.92</td>
<td>–1.55</td>
<td>–0.92</td>
<td>–0.37</td>
<td>0.25</td>
<td>1.32</td>
<td>2.23</td>
<td>5.73</td>
</tr>
<tr>
<td>Chloroplast</td>
<td>Whole genome</td>
<td>0.14</td>
<td>–2.70</td>
<td>–2.12</td>
<td>–1.83</td>
<td>–1.23</td>
<td>–0.67</td>
<td>0.00</td>
<td>1.12</td>
<td>1.95</td>
<td>4.20</td>
</tr>
<tr>
<td>E. coli</td>
<td>Whole genome</td>
<td>4.6</td>
<td>–3.19</td>
<td>–1.68</td>
<td>–1.05</td>
<td>0.04</td>
<td>0.77</td>
<td>1.49</td>
<td>2.53</td>
<td>3.27</td>
<td>5.79</td>
</tr>
</tbody>
</table>

¹DNA sequence sources are described in “Materials and Methods.”
the randomized maize chromosome 9 distribution. These data suggest that the human SVM run on the maize nuclear genome detected significantly more sequences predicted to be nucleosome bound or nucleosome free than would be expected at random or for non-chromatin-associated genomes.

NOL Plots Highlight Gene Structures and Are Validated by Empirical Measurements

We next investigated the NOL scores around the TSSs of a few select genes (presumed nonmutant alleles), as shown in Figure 2. The NOL plots and gene models are shown for Alcohol dehydrogenase 1 (Adh1), a gene with multiple transcript isoforms, Sad1-Unc84-like 5 (Sun5), a low-expressed gene limited to pollen expression (Murphy et al., 2010), Rolled leaf 1 (Rld1), a developmental regulatory gene, and Multidrug resistance-associated protein 3 (Mrp3), a gene encoding an ATP-binding cassette domain 2 of a multidrug resistance-associated protein. Each gene exhibits a unique NOL signature, with a tendency for higher positive scores in the exonic regions (Fig. 2B). We also note that inspection of many maize genes reveals a lower NOL score in the regions upstream of the TSS. Interestingly, in some cases, relatively high positive NOL scores just upstream of the designated canonical transcript (Fig. 2, arrows) overlap with EST or alternative transcript models in which upstream promoters may function (e.g. Adh1 in Fig. 2A).

In order to confirm these predictions with empirical measurements, we carried out DNA microarray hybridization experiments to map nucleosome occupancy using MNase protection assays for a group of 400 genes, including those shown in Figure 2. The empirical nucleosome occupancy data for nuclei from two sources, ear shoot and seedling, are coplotted with the prediction and gene models. We found that they show good general agreement across the regions for which we designed probe coverage (canonical TSS ± 1,500 bp). On a probe-by-probe basis (Fig. 2, scatterplots), the correlations observed (r of 0.45–0.71 for these four examples) validate the model and support the claim that the human SVM algorithm performs well on the maize genome. In fact, when comparing the global correlation of all probes on our 400-gene array with a similar-sized data set for human genes, the maize data (r = 0.63) were more highly correlated with the NOL predictions than the corresponding human data (r = 0.59), as summarized in Figure 3, A to C. When the correlations between the predicted and measured values were analyzed gene by gene and binned by r value increments of 0.10, we found that the most frequent class for maize was r = 0.6 to 0.7 and that for human was r = 0.5 to 0.6. This observation is in good agreement with the recent report from Labonne et al. (2013) in which measured and predicted nucleosome occupancy were determined across different tissues and genotypes. Together, these results demonstrate that the human SVM provides reliable and informative estimates of NOL in maize, adding a new informational dimension for the annotation of the maize genome.

A Novel Plant Promoter Landscape Revealed by the Analysis of Genome-Wide Nucleosome Data

To determine if there was a general pattern of NOL scores for maize genes, we aligned the NOL plots at the TSS for several large gene model sets and averaged their scores (Fig. 3D). For each of the gene sets, we observed a striking pattern of NOL scores that reached a peak value a few hundred bases downstream of the annotated TSS. The larger gene sets (working gene set [WGS] and putative unigene transcripts [PUTI]) are known to include a substantial number of pseudogenes, which may be related to the difference in strength of this +1 nucleosome signal. The NOL peak signal in the TSS-average plot is centered around the +200 region in maize. Surprisingly, the maize NOL plots did not show a strongly positioned −1 nucleosome signal, as one might expect from prior studies of metazoan and fungal species (for review, see Jiang and Pugh, 2009). Instead, it appeared that maize genes may essentially lack a −1 nucleosome signal.

We wanted to know if this feature, the lack of a −1 nucleosome signal, was peculiar to maize or possibly a general feature of plants. To address this question, we calculated and coplotted NOL-TSS consensus plots for maize, human, and Arabidopsis (Arabidopsis thaliana) genes (Fig. 3E). Whereas the human genes showed a prominent peak for a −1 nucleosome, neither maize nor Arabidopsis showed a comparable signal. Given the unexpected nature of this pattern, we sought experimental validation on a larger scale. For this, we plotted the microarray-based nucleosome occupancy data across 387 TSSs (Fig. 3F) and the next-generation sequencing-based nucleosome occupancy calculated for 21,899 Arabidopsis TSSs (Fig. 3G) calculated from Chudavarapu et al. (2010). In both cases, one eudicot and one monocot, the lack of a −1 nucleosome predicted from the plant NOL plots was also observed via genome-wide nuclease protection assays. From this, we concluded that the pattern predicted for maize promoters (Fig. 3D) was typical for a plant and not peculiar to maize.

Repetitive Elements Exhibit Unique NOL Profiles

The maize genome is known for its large, diverse, and dynamic repetitive element population (Bennetzen and Hake, 2009; Schnable et al., 2009). Collectively, repetitive elements constitute the majority of the genome and, therefore, the majority of chromatin- and nucleosome-binding sequences. Importantly, recognition and interest are emerging in their contribution to gene regulation and genome dynamics and evolution (Feschotte et al., 2002; Wang and Dooner, 2006; Baucom et al., 2009; Devos, 2010). Despite their significance,
Figure 2. Experimental validation of NOL predictions for maize genomic DNA. NOL-predicted and empirically measured nucleosome occupancy are shown for four maize genes. To the right are correlation values and scatterplots of empirically measured nucleosome occupancy scores versus the associated NOL scores for the probe sequences. A, Adh1 (GRMZM2G442658). B, Sun5 (AC194341.4_FGT003). C, Rld1 (GRMZM2G109987). D, Mrpa3 (GRMZM2G111903).
Figure 3. Correlations between predicted and measured nucleosome occupancy in both human and maize and occupancy profiles at TSS. A, The correlations for 386 maize TSSs were determined and plotted as a frequency histogram with bin sizes of 0.1 (black bars). For comparison, a similar data set was analyzed using 411 human TSSs (gray bars). B and C, Scatterplots of microarray versus NOL scores for all microarray probes shown as smoothed gray-scale kernel density representations for maize (B) and human (C). The correlation r values for maize and human are indicated in the plots. D, Average NOL scores aligned to the TSSs of maize genes for four different sets of gene annotations: canonical filtered gene sets (FGS), filtered gene sets, WGS, and PUTs. E, Average NOL score at the TSSs of annotated human (hg19; solid line), maize (B73 RefGenv2; dashed line), and Arabidopsis (TAIR 10; dotted-dashed line) genes. F, Average measured nucleosome occupancy at the TSSs of genes in maize seedlings (solid line) and ear shoots (dashed line). G, Average measured nucleosome occupancy at TSSs in Arabidopsis. In D to G, TSS positions and directions are indicated with vertical lines and arrows, respectively.
these sequences are often excluded from genome-wide analyses, such as microarray and high-throughput sequencing assays, because of the inability to map sequences to unique positions in the genome. The ability of the SVM to predict nucleosome occupancy from contiguous DNA sequences of any level of copy number provides an unprecedented view of the maize chromatin landscape. Given the prevalence and importance of repetitive sequences in both structural and functional genomics in maize, we produced NOL plots for some of the most common DNA transposable elements and retroelements using exemplar or published sequences, as shown in Figure 4.

The DNA transposable elements Activator (Ac), Suppressor-mutator (Spm), Mutator (Mu), and a representative mutator-like element, Trap, all exhibit very high NOL scores at their 5' ends, and Ac, Mu, and Trap, but not Spm, exhibit high scores at their 3' ends (Fig. 4, A–D). Three of these four DNA elements had at least one region with NOL scores in the top 1% range, above a value of 4.26 (Fig. 4, dashed lines indicate the top 1%). We also plotted the NOL scores for 10 common long terminal repeat (LTR) elements, including six gypsy-like (Fig. 4, E–J) and four copia-like (Fig. 4, K–N) retroelements. A diversity of patterns was observed for these, both in the LTRs and within the bodies of the elements. The NOL scores for the LTRs ranged from extremely high for Xilon to very high for Opie2, Huck1, and Prem2α to moderately high for CRM and Stonor to relatively low for Zeon1 and ZmCopia1. Given that the LTRs are transcribed, the presence of TSS-like spikes in NOL scores (e.g. Xilon and Opie2) is not surprising, but Zeon1 and ZmCopia1 are interesting exceptions. Interestingly, no general pattern or overall trend could be discerned for the retroelements as a group, indicating that different retroelements have evolved distinct nucleosome affinity profiles.

**Bacterial Artificial Chromosome-Scale NOL Plots Highlight Intraspecies Variation around the bz-uce2 Region**

We next examined the NOL plots in the bronze (bz) region on chromosome 9, taking advantage of previously characterized comparative genomic sequences in this region (Wang and Dooner, 2006). The raw and smoothed NOL plots for the region delineated by bz and uce2 on maize chromosome 9 are shown in Figure 5 along with the locations of known or predicted genes and LTR retroelements. Most of the protein-coding genes showed the expected feature of relatively high NOL scores, especially near the 5' ends of the genes (Fig. 5B, arrows). As observed with the analysis of NOL scores around the TSS of select genes (Fig. 2), the 5' ends of individual genes (stl1, bz, str1, rpl35a, and znf; Fig. 5, A and B) possess relatively high NOL scores, often exceeding the top 1% of scores for the entire maize genome. In addition, prominent NOL score peaks appeared in the LTR regions of some, but not all, retroelements (Fig. 5B, arrowheads). For example, the Xilon2, Tekay, and Grande1 LTRs showed very high peaks (Fig. 5B). Together, the genic TSS regions and the retroelement LTRs dominate the NOL landscape with discrete peaks that are readily visualized at this approximately 100-kb viewing scale.

Maize as a species is well known for its extraordinary genotypic diversity, even among breeding lines belonging to the same subspecies, Z. mays mays. To examine NOL variation in relation to genotypic variation, we plotted the NOL scores for the same bz-uce2 region in two other inbred lines, McC (Fig. 5C) and A188 (Fig. 5D). Despite the marked variation in intragenic genome sequences, the NOL profiles for the genes themselves remained largely the same, as might be expected on the basis of genic DNA sequence conservation. The NOL plots for these two genotypes were conspicuously differentiated by the signals associated with their various retroelement compositions. One notable retroelement family, Huck, displays high NOL scores across the entire element, not just within the LTRs (compare Grande1 in Fig. 5B with Huck1a and Huck1b in Fig. 5C).

**NOL Plots Provide Continuous Genome-Wide Annotation across All Maize Chromosomes**

To examine the predicted nucleosome occupancy genome wide at the whole-chromosome scale, we generated NOL plots, displayed in Figure 6, for all 10 chromosomes. We used fixed x and y axes to allow for direct comparisons of features across all 10 plots. The copotted gene density heat maps reveal that the NOL scores at this scale are not explained simply by gene density, but a general trend is apparent of higher NOL scores toward the ends of the chromosomes, where gene density is generally higher. One region of chromosome 6 (near 11 Mb) exceeds the upper limit we chose for the y axis (Fig. 6, arrow). This region is relatively gene poor, and the assembly has a high frequency of gaps, which we believe may contribute to NOL score distortion. We also noticed that the NOL scores around individual centromeres varied, without an obvious pattern for all centromeres viewed at this scale. The functional significance of the features revealed at this level remains largely untested, such that the relevant or apparent features at this scale may become more clear as structural-genomic data are added in the future.

In exploring the NOL plots at different scales (from 1,000 to 100,000,000 bp) and using differing degrees of data smoothing, we noticed that the features across any given segment of DNA seemed variable, especially with regard to smoothing. The most basic form of the NOL score data are single values for 50-bp windows, plotted at 1-bp step size. But for viewing larger segments of DNA, smoothing aids visualizing signal over noise (Fig. 5, compare A with B). To explore more systematically the effect of smoothing, we plotted the...
Figure 4. NOL plots for notable repeat elements in the maize genome. Three classes of elements are shown: DNA transposable elements, gypsy LTR retroelements, and copia LTR retroelements. The dotted lines indicate the 99% cutoff for all maize NOL scores. On the x axis, major tick marks indicate 1 kb, medium tick marks indicate 500 bp, and minor tick marks indicate 100 bp. LTRs are marked by black triangles.
NOL scores for chromosome 9 with increasing degrees of LOESS smoothing from 0.001 to 0.33, as shown in Supplemental Figure S2. In this comparison, we marked three regions (Supplemental Fig. S2, squares, circles, and triangles) to illustrate the effect smoothing can have on particular loci or the relative NOL scores across pairs of loci. They change their relative positions along the y axis (NOL score) in different ways depending on the degree of smoothing. For the whole-chromosome view, a LOESS smoothing value of 0.01 to 0.02 was found to provide a suitable value for whole-chromosome inspection. At the highest level of smoothing (LOESS value of 0.33; Fig. 6), a clear pattern emerges of peak-trough-peak, likely reflecting the overall

Figure 5. BAC-scale analysis of NOL data reveals features highlighting genes and repetitive elements. NOL plots for selected regions of BAC clones in three maize genotypes are shown. Genes are annotated with black pentagons; the point of each pentagon indicates the directionality of the gene. Notable repetitive elements are indicated by gray rectangles, and black triangles indicate the LTRs of these elements. Arrows indicate peaks at the 5′ ends of genes, and arrowheads indicate notable LTR signals. A, NOL scores generated for the BAC in the B73 genotype (GenBank no. AF448416). B, The same locus with the NOL scores smoothed by LOESS. C and D, NOL scores smoothed with the same algorithm for the McC (GenBank no. 391808; C) and A188 (GenBank no. DQ493650; D) genotypes.
Figure 6. NOL plots for each of the 10 maize chromosomes. Each plot is smoothed by LOESS. Beneath the NOL scores are gene-density heat maps (numbers of unique TSSs per 1 Mb). The centromere is marked by “C” directly above the heat map for
gene density of the chromosome, known to be higher in the arms and lower around the centromere, most of which are metacentric or submetacentric in maize (Schnable et al., 2009).

**Annotation of the Maize Genome on a New Public Genome Browser, Genomaize**

We produced a publicly available B73 genome browser, Genomaize (http://www.genomaize.org), based on the University of California, Santa Cruz Genome Browser platform (Kent et al., 2002). This resource is intended to enhance the genomic data visualization of nucleosome occupancy information in the context of available genomic annotations, empirical or predicted. The genome-wide NOL data are provided as data tracks in two public browsers, Genomaize (“Nucleosome Occupancy Likelihood” track within “Nucleosomes”) and MaizeGDB (“Nucleosome Occupancy Likelihood (FSU HsA375, ZmB73rg2)” track). These browsers allow for the visualization, navigation, and comparison of NOL data in parallel with many other genomic features. A particularly useful feature of the UCSC genome browser platform is the ability to render data very quickly, allowing rapid navigation of data sets at all scales. Sample screen shots are shown in Figure 7 for the bz-uce2 region on chromosome 9. Major genetic elements, such as genes and retroelements, are highlighted with this new annotation. In addition to the NOL plots, the empirical data analyzed in Figures 2 and 3 are available online, providing a new genomics resource in the form of NOL plots that display a fixed, distinctive, and continuous maize genome annotation. This information is expected to inform investigative research that bears on questions of chromatin architecture for the maize genome.

**DISCUSSION**

Prediction of nucleosome occupancy in maize with a 1-bp sliding window across the maize genome produced a population of 3 billion 50-bp segments, each with an associated NOL score. The previously published and experimentally verified algorithm (SVM model A375 from Gupta et al., 2008) was trained on data from human nucleosome-mapping assays and was shown here to perform equally well (Fig. 3, B and C) on a plant species with a similarly large, complex genome, maize. The population of maize NOL scores had a median value of −0.25 and nonrandom distribution with a distinct positive skew. This skew is also characteristic for human NOL scores (Fig. 1A) but is not seen for NOL scores of noneukaryotic genomes such as maize mitochondria, maize plastid, or E. coli (Table I; Supplemental Fig. S1, E–G) or for a randomized set of sequences with the same complexity and base composition as maize chromosome 9 (Table I; Supplemental Fig. S1D). This rightward skew, representing a large fraction of relatively strong nucleosome-positioning signals, is likely the result of evolutionary selection within genes and repetitive elements (Figs. 2, 4, and 5).

Our results are consistent with contributions from both a genomic code for nucleosome positioning and statistical positioning of nucleosomes. The genomic code for nucleosome positioning (Segal et al., 2006) may play a significant role in maize because of the nonrandom prevalence of high (nucleosome-enriched) and low (nucleosome-depleted) NOL scores. These results indicate that maize nucleosome positioning may be strongly influenced by boundary features, which could then affect statistical positioning (Kornberg, 1981) to set up a gene-specific or repeat element-specific nucleosomal landscape, similar to what has been described for yeast (Zhang et al., 2011). By identifying regions of DNA with intrinsic tendencies favoring nucleosome-binding versus nucleosome-depleted DNA, our study will guide comparative and genetic analyses of chromatin structure and dynamics in relation to genome structure and function.

At individual genes, we find that NOL scores frequently show peaks near their 5’ ends (Figs. 2 and 3) that exceed the top one percentile value of the whole genome (Figs. 1 and 2; Table I). These high gene-associated scores are quite remarkable considering the fact that the population of NOL scores for maize is predominantly over 80% nongenic. Our findings are also consistent with the general patterns seen in yeast and humans, where the +1 nucleosome is often the most strongly positioned nucleosome of the gene (Yuan et al., 2005; Barski et al., 2007; Mavrich et al., 2008a, 2008b; Jiang and Pugh, 2009). Because NOL plots clearly highlight transcription initiation sites of maize genes, they may potentially complement other means of improving gene models and identifying putative TSSs in regions lacking EST evidence.

Many human genes exhibit a nucleosome landscape of a peak-trough-peak near the TSS, composed of an upstream or −1 nucleosome and a downstream or +1 nucleosome, with a nucleosome-free region in between (Schones et al., 2008; Valouev et al., 2011). The −1 nucleosome is often the target of covalent histone modifications associated with gene activation or repression (for review, see Jiang and Pugh, 2009). This −1 nucleosome has received attention because of its tendency to be modified, replaced, or differentially sensitive to fixation and nuclease digestion conditions (Barski...
et al., 2007; Schones et al., 2008; Henikoff et al., 2009, 2011; Jiang and Pugh 2009, Weiner et al., 2010; Xi et al., 2011).

Although some maize genes display a NOL signal representing a −1 nucleosome (e.g. ZmSun5 in Fig. 2B), a surprising observation in this study was the lack of such a predicted or measured nucleosome in this −1 region for many maize and Arabidopsis genes (Fig. 3E). It is important to consider the effect of technical contributions to the possibility that a −1 nucleosome may be present but undermeasured. For instance, the data from Arabidopsis were derived from unfixed nuclei (Chodavarapu et al., 2010). Even so, the use of unfixed nuclei was shown to cause either negligible (Henikoff et al., 2011) or partial (Xi et al., 2011) reduction in measured occupancy at the −1 position.

**Figure 7.** Visualization of genomic features, including NOL scores, in Genomaize (http://www.genomaize.org). Shown are tracks for bins, assembly gaps, Munich Information Center for Protein Sequences (MIPS) repeats, UniformMu insertions, gene models, and NOL scores. Genomic coordinates are indicated at the top. A, A region of the maize genome similar to that shown in Figure 2A. B and C, Enlarged regions marked above the genomic coordinates in A. Directionalities of the gene transcription are indicated by gray arrows.
when compared directly with that of fixed nuclei. A stronger effect may have to do with the degree of MNase digestion used in nucleosome occupancy assays (Weiner et al., 2010; Henikoff et al., 2011; for review, see Zentner and Henikoff, 2012). In our studies, with relatively high levels of MNase digestion, we cannot exclude the possibility that hypersensitive nucleosomes, like those described for yeast (Weiner et al., 2010), also exist around some TSS regions in plants. Relatedly, our findings are compatible with DNase I-hypersensitive site mapping in rice (Oryza sativa), in which DNase I-hypersensitive sites were prevalent just upstream of the TSS (Zhang et al., 2012). Our findings in maize and Arabidopsis are in general agreement for both predicted and measured promoter architecture. Given that maize and Arabidopsis diverged about 150 million years ago, this study has important implications for the recognition of a plant-wide trend in which the proximal promoter region lacks a strong DNA-based nucleosome occupancy signal. To further define the nature and generality of the observations reported here, it will be important to investigate plant chromatin structure under various defined experimental conditions as well as to develop new SVM models trained on plant chromatin data.

Maize repetitive elements constitute the majority of chromatin in the maize genome. We found that different families of transposable elements are predicted to have unique NOL profiles, even across families belonging to the same class. For example, two of the six gypsy-family retroelements, Opie2 (Fig. 4K) and Xilon (Fig. 4I), had very high NOL scores in their LTR regions. Given the biological significance of the sequence-driven affinity of nucleosome-forming potential, retroelement-specific differences in nucleosome occupancy may reflect functional differences with important evolutionary consequences. Retrotransposons employ an obligatory transcription for their transposition and, therefore, provide a potentially informative alternative system for defining the role of chromatin structure in transcriptional control and, possibly, the integration of mobile elements.

NOL plots provide an informative view of the predicted chromatin architecture at bacterial artificial chromosome (BAC)-size scales. Specifically, NOL plots highlight the locations of genes, LTRs, and other repetitive elements and, therefore, may also guide the discovery of unannotated transcription units such as rarely expressed genes, noncoding RNAs, and repetitive elements. In addition, we found that the human-trained SVM could highlight these features across diverse maize lines with genome structure variation, allowing the investigation of how genome structure variation affects the chromatin landscape at multiple scales. Finally, recombination is a major aspect of plant breeding and crop improvement and is also being recognized as a chromatin-mediated process (Wang and Dooner, 2006; Schnable et al., 2009). Therefore, NOL plots should help to inform investigations into how chromatin structure might interact with the distribution or regulation of meiotic recombination.

When NOL scores are visualized and interpreted at larger scales (Fig. 6; Supplemental Fig. S2), the scale of view, which can span 6 orders of magnitude, is an important consideration, especially when patterns such as NOL score peaks and valleys are observed (Supplemental Fig. S2C) across a linear representation of the genome. Which features, at different scales, reflect functional biological or structural units of chromatin remains an open question. At the whole-chromosome scale (150–350 Mb), the gene-rich areas near the ends of the chromosomes show a relatively high NOL score per unit area (Fig. 6). This distribution reflects the general organization of maize chromosomes as metacentric, with a tendency for gene-dense regions toward telomeres. Similar patterns in which the arms are different from the central, centromeric regions can be observed in other genomic data, such as epigenetic marks and repetitive sequence distribution (Baucom et al., 2009; Schnable et al., 2009; Eichten et al., 2011).

In considering the overall rationale for this study, an important fact is that the NOL scores are assignments based on primary DNA sequence, whereas chromatin structure is three dimensional and dynamic. As a genome annotation, the NOL plots represent an invariant baseline of intrinsic features derived directly from the DNA sequence. These plots, on Genomaize or other browsers, provide a platform to facilitate a comparative analysis of nucleosome structure with other genomic information such as epigenetic marks in relation to development or in response to stimuli (Wang et al., 2009; Waters et al., 2011; Zhang et al., 2012). The NOL plots are expected to serve, therefore, as a valuable new tool for plant biology, enhancing future investigations defining the functional relationships between DNA sequence, nucleosome occupancy, and epigenetic landscapes.

MATERIALS AND METHODS

SVM A375

All NOL scores were generated with the SVM model (described by Gupta et al., 2008) derived from empirical nucleosome occupancy data from a human melanoma cell line, A375. This SVM classifier uses a model file (A375-SVM) and a python script (Gupta et al., 2008) that scores 50-mer DNA sequences and assigns them a single numerical value. This NOL score indicates a nucleosome-forming (positive value) or nucleosome-inhibitory (negative value) potential for the sequence. For sequences longer than 50 bp, a sliding 50-bp window is used with l-bp steps. To smooth NOL score data, a LOESS smoothing algorithm was used as implemented in the R statistical programming language. The parameter “span” was used to determine the level of smoothing calculated.

Genome and Gene Model Sequences

Genomic sequences were obtained from MaizeSequence.org (http://www.maizesequence.org/) for nuclear, mitochondrial, and plastid genomes (B73 RefGen_v2/AGPv2, downloaded in January 2012 from ftp://ftp.maizesequence.org/release-8b/assembly/). Maize (Ziz mays) chromosome 9 was randomized by sequential randomization with a python script. Human genomic sequences were downloaded from the UCSC human genome browser (hg19; downloaded in January 2012 from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/). Escherichia coli genomic sequences were from the National Center for Biotechnology Information (K12-MG1655; downloaded in January 2012 from ftp://ftp.ncbi.nlm.nih.gov/
Fisher et al.

Nucleosome Occupancy from Oligonucleotide Microarrays
Maize nuclei were isolated and digested as follows. Ear shoots of 3 to 5 cm were harvested from field-grown B73 plants (Mission Road Research Facility in Tallahassee), flash frozen in liquid nitrogen, and stored at −80°C. One gram of frozen ear shoots was ground under liquid nitrogen in a mortar and pestle and fixed at room temperature in 20 ml of buffer A (50 mM sodium borat, 15 mM PIPES, pH 6.8, 80 mM KCl, 20 mM EDTA, 0.5 mM EGTa, 0.15 mM spermine, 0.5 mM spermidine, and 1 mM dithiothreitol) with 1% formaldehyde. After 10 min, the cross linking was stopped with 125 mM Gly, and the tissue was pelleted by centrifugation at 1,000g for 10 min and then resuspended in buffer A with 1% Triton X-100. After incubation at 4°C for 15 min, nuclei were filtered through two layers of Miracloth, washed once with buffer A, and resuspended in 1 ml of MNase digestion buffer (50 mM Tris, pH 7.5, 12.5% glycerol, 25 mM KCl, 4 mM MgCl2, and 4 mM CaCl2). A total of 100 μL of nuclei suspension was treated with 0.2 units of MNase (Worthington) for 5 min at 37°C, and the digestion was stopped with 10 μl MNase. Human cells were digested as follows. lIERT-RPE1 cells (Coriell) were grown to 70% confluency in Dulbecco’s Modified Eagle Medium: Nutrient Mixture F-12 with 10% fetal bovine serum. Cells were fixed in growth medium with 1% formaldehyde at room temperature for 10 min, and the cross linking was stopped with 125 mM Gly. Cells were washed twice with phosphate-buffered saline, permeabilized for 10 min in MNase digestion buffer with 0.1% Triton X-100, digested with 40 units ml−1 MNase digestion buffer with 0.1% Triton X-100 for 5 min at room temperature, and the digestion was stopped with 10 μl MNase. Digested maize nuclei and human cells were digested as well in 1% formaldehyde at room temperature for 10 min, and the cross linking was stopped with 125 mM Gly. Cells were washed twice with phosphate-buffered saline, permeabilized for 2 min in MNase digestion buffer with 0.1% Triton X-100, digested with 40 units ml−1 MNase digestion buffer with 0.1% Triton X-100 for 5 min at room temperature, and the digestion was stopped with 10 μl MNase. Digested maize nuclei and human cells were digested as well in 1% formaldehyde at room temperature for 10 min, and the cross linking was stopped with 125 mM Gly. Cells were washed twice with phosphate-buffered saline, permeabilized for 2 min in MNase digestion buffer with 0.1% Triton X-100, digested with 40 units ml−1 MNase digestion buffer with 0.1% Triton X-100 for 5 min at room temperature, and the digestion was stopped with 10 μl MNase. Digested maize nuclei and human cells were digested as well in 1% formaldehyde at room temperature for 10 min, and the cross linking was stopped with 125 mM Gly. Cells were washed twice with phosphate-buffered saline, and permeabilized for 2 min in MNase digestion buffer with 0.1% Triton X-100, digested with 40 units ml−1 MNase digestion buffer with 0.1% Triton X-100 for 5 min at room temperature, and the digestion was stopped with 10 μl MNase.

Analysis of Arabidopsis Data
MNase-Seq data obtained from Chodavarapu et al. (2010); Gene Expression Omnibus no. GSE21673. Reads were filtered to only contain reads in which 90% of the bases had a Sanger quality score of 25 with FASTX-Toolkit and aligned to the TAIR 10 genome assembly with Bowtie2 using default parameters. Only uniquely aligning reads were kept, and base-wise read densities were calculated using bedtools software. To obtain TSS-centered averages of nucleosome occupancy scores, base-wise read densities of TAIR 10 genes were aligned by the TSS and averaged.

For Arabidopsis plots, data from Chodavarapu et al. (2010) was first mapped from TAIR 7 to TAIR 10 build of the genome. TAIR 10 gene models were then used for the generation of mean nucleosome occupancy profiles at regions surrounding TSSs (refer to TAIR 10 gene models). For plotting, read counts were limited to a maximum value of 40 to exclude skew caused by extremely large read counts mapping to a few regions. Maize NOL and microarray data were plotted in a publicly accessible genome browser installation (genomazoo.org) using the UCSC genome browser platform (Kent et al., 2002) currently loaded with the B73 RefGenV2 genome assembly.

Repetitive Sequences
Sequences for DNA transposable elements were downloaded from GenBank (AC, AF353378.1; MuDR, M67987.1; and SPM Sadler, M25427). The DNA sequence for the Trap element was provided by D. Lisch (University of California, Berkeley [for review, see Lisch, 2002]). Retroelement sequences were selected from Meyers et al. (2001) to represent common, abundant elements from TE REPbase (http://www.ncbi.nlm.nih.gov/repbase/) for sequence exemplars. The entire element was constructed by concatenation of the LTR sequence on either end of the internal sequence. We did so for Cinful1 (AF484161), CRM (AY129008), Grandi1 (X97604), Huck11 (TE REPbase “HUCK11-I_ZM”), Xil1 (AF484161), Zoon1 (U11059), Opid1 (U68408), Premz2 (AF904447), Stonor (TE REPbase “STONOR_F”), and ZmCop1 (AF215823).

BAC Sequences, Gene Sets, and Gene Models
The NOL plots for BAC sequences were produced with annotated sequence files from GenBank for maize inbred lines B73 (GenBank no. AF448416; start, 27,982; end, 106,186), McC (GenBank no. AF391808; start, 110,111; end, 225,742), and A188 (GenBank no. DQ493630; start, 1; end, 54,821). Maize gene annotations were downloaded from MaizeSequence.org (AGPv2-5b; downloaded in January 2012 from http://ftp.maizegenome.org/release-5b/), and human annotations were downloaded from the UCSC Genome Browser MySQL database (Table hg19.mhgFullMm).

Supplemental Data
The following materials are available in the online version of this article.
Supplemental Figure S1. NOL score distributions in control genomes.
Supplemental Figure S2. Different smoothing levels for maize chromosome 9.

ACKNOWLEDGMENTS
We thank Anne Thistle and members of the H.W. Bass, J.H. Dennis, and K.M. McGinnis laboratories for critique of the manuscript. We also acknowledge contributions from anonymous reviewers for suggested experiments and improvements associated with extending and testing the predictions reported. Received February 14, 2013; accepted April 8, 2013; published April 9, 2013.

LITERATURE CITED

Received February 14, 2013; accepted April 8, 2013; published April 9, 2013.