Independent and complementary methods for large-scale structural analysis of mammalian chromatin

Jonathan H. Dennis,[1,6] Hua-Ying Fan,[1,6] Sheila M. Reynolds,[2,6] Guocheng Yuan,[3] James C. Meldrim,[4] Daniel J. Richter,[4] Daniel G. Peterson,[5] Oliver J. Rando,[3] William S. Noble,[2] and Robert E. Kingston[1,7]

[1]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA; [2]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA; [3]Bauer Center for Genomics Research, Harvard University, Cambridge, Massachusetts 02138, USA; [4]The Broad Institute, Cambridge, Massachusetts 02141, USA; [5]Department of Plant and Soil Sciences, Mississippi State University, Mississippi State, Mississippi 39762, USA

The fundamental building block of chromatin, the nucleosome, occupies 150 bp of DNA in a spaced arrangement that is a primary determinant in regulation of the genome. The nucleosomal organization of some regions of the human genome has been described, but mapping of these regions has been limited to a few kilobases. We have explored two independent and complementary methods for the high-throughput analysis of mammalian chromatin structure. Through adaptations to a protocol used to map yeast chromatin structure, we determined sites of nucleosomal protection over large regions of the mammalian genome using a tiling microarray. By modifying classical primer extension methods, we localized specific internucleosomally cleaved mammalian genomic sequences using a capillary electrophoresis sequencer in a manner that allows high-throughput nucleotide-resolution characterization of nucleosome protection patterns. We developed algorithms for the automated and unbiased analysis of the resulting data, a necessary step toward large-scale analysis. We validated these assays using the known positions of nucleosomes on the mouse mammary tumor virus LTR, and additionally, we characterized the previously unreported chromatin structure of the *LCMT2* gene. These results demonstrate the effectiveness of the combined methods for reliable analysis of mammalian chromatin structure in a high-throughput manner.

[Supplemental material is available online at www.genome.org.]

In eukaryotic cells, DNA is efficiently and compactly organized into chromatin consisting of nucleosomal units of 150 bases of DNA wrapped 1.65 times around a histone octamer (Kornberg and Lorch 1999). Chromatin is the substrate for virtually all nuclear events: transcription, replication, recombination, and repair (Kornberg and Lorch 2002). Chromatin condenses and decondenses in response to different molecular cues, and the spatiotemporal specificity of nuclear processes appears to be well-coordinated with this dynamic nature of nucleosomal organization and genomic structure (Lu et al. 1994; Wallrath et al. 1994; Anderson and Widom 2000). Nucleosome spacing and positioning are generally accepted to be a major determinant of chromatin structure (Kingston and Narlikar 1999). No means thus far have been able to test this model in mammals, however, as there are no data on the position and spacing of nucleosomes over a large and varied genomic area. Our goal was to develop high-throughput, cost-effective, reliable, and robust methods for the analysis of nucleosome protection over broad areas of the human genome.

Recently a protocol has been described in which a tiling microarray of nearly 500 kb of the *Saccharomyces cerevisiae* genome was probed with mononucleosomal DNA (Yuan et al. 2005). In addition to identifying the translational position of a majority of the nucleosomes, conventional patterns of nucleosome deposition and density were described for Pol II promoters (Yuan et al. 2005). We reasoned that we could adapt this technique to mammalian genomes by creating a gene-enriched mononucleosomal library with which to interrogate a custom human genome-tiling array. The maximum resolution of this technique is directly related to the length and spacing of the oligonucleotides on the array. To corroborate the results of the tiling microarray in a high-resolution manner, we adapted the ligation-mediated polymerase chain reaction (LM-PCR) for analysis on a capillary electrophoresis sequencer. This allows changes in chromatin cleavage sensitivity at single nucleotide resolution. LM-PCR is unrivaled as the most sensitive technique to map cleavage sites at the nucleotide level in genomic DNA, and thus is an ideal complement to the mononucleosomal hybridization experiment.

Translational positioning of nucleosomes has been documented as a feature of several loci in mammalian genomes (Simpson et al. 1993). This positioning can result from the effects of regulatory factors binding to chromatin as well as features intrinsic to the DNA sequence itself (Fragoso et al. 1995). The mouse mammary tumor virus long terminal repeat (MMTV-LTR) has served as a powerful tool in the elucidation of the coordination between translational positioning and transcriptional status. MMTV-LTR is organized into six nucleosomes (Richard-Foy and

[6]These authors contributed equally to this work.
[7]Corresponding author.
E-mail kingston@frodo.mgh.harvard.edu; fax (617) 643-2119.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.5636607.
Freely available online through the *Genome Research* Open Access option.

17:928–939 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

Hager 1987; Truss et al. 1995; Belikov et al. 2000). We chose a cell line with a stable incorporation of the MMTV-LTR as the proof-of-principle case for the above technologies (Wilson et al. 2002). In addition, we characterized the previously undescribed nucleosome protection pattern of the promoter region of *LCMT2*, an important regulator of cell cycle (De Baere et al. 1999).

Here we describe the adaptation of these two nucleosome mapping methodologies for the automated, high-throughput, long-range analysis of mammalian chromatin structure (Fig. 1). This combination should pioneer a robust, cost-effective, automatable, interpretable, and quantifiable procedure for the long-range mapping of nucleosomes. The adaptations used to establish these protocols include preparation and enrichment of internucleosomally cleaved, primer-extendable template, generation of independent and complementary high throughput readouts, and development of software to align and analyze data from different experiments. We believe that a systematic, automated mapping of chromatin architecture over regions of the genome orders of magnitude greater than those examined previously will provide a useful tool for the identification of regulatory elements and the formulation of hypotheses concerning the regulation of higher-order genomic structure.

Results

Development of two parallel platforms for the description of chromatin structure

Preparation of MNase-cleaved genomic template

Analysis of chromatin structure can be accomplished using enzymes such as micrococcal nuclease (MNase) that cleave chromatin at regions between nucleosomes (Telford and Stewart 1989a,b). To develop technologies for broad-scale mapping of mammalian chromatin, we needed a cleavage procedure that maintained faithful structure. We pre-extracted cells with phosphate-buffered saline modified to contain 300 mM NaCl and mild detergents to obtain bona fide nucleosomal material. In order to preserve the native state of the chromatin through the MNase cutting reaction, we then fixed the chromatin with formaldehyde (Kornberg et al. 1989; Fragoso and Hager 1997).

Two distinct nucleosomal DNA populations were needed for the two different protocols described below. The protocol that uses tiling microarrays was performed by interrogating these arrays with protected mononucleosomal DNA fragments. As the readout for the array hybridization experiment relies on the absence of hybridization to the probes corresponding to internucleosomal regions, inclusion of any nucleosomal ladder fragments larger than mononucleosomes might result in an increase in noise and decrease in signal. In contrast, the LM-PCR primer extension protocol requires longer DNA fragments as template because each primer extension read can cover two or three nucleosomes. Ideally, this DNA population should be at least twice as long as the maximum read-length desired to maximize information from singly cut molecules of DNA. Thus, because a capillary sequencer is maximally capable of a read-length of up to 1000 nt, the population of DNA fragments used in the primer extension experiments was chosen to range between 150 and 3000 bases in length.

We titrated MNase digestion reactions so that the same samples could be used for the purification of mononucleosomal DNA for the array hybridization experiment and for preparation of longer templates for the LM-PCR primer extension. A typical result of in-

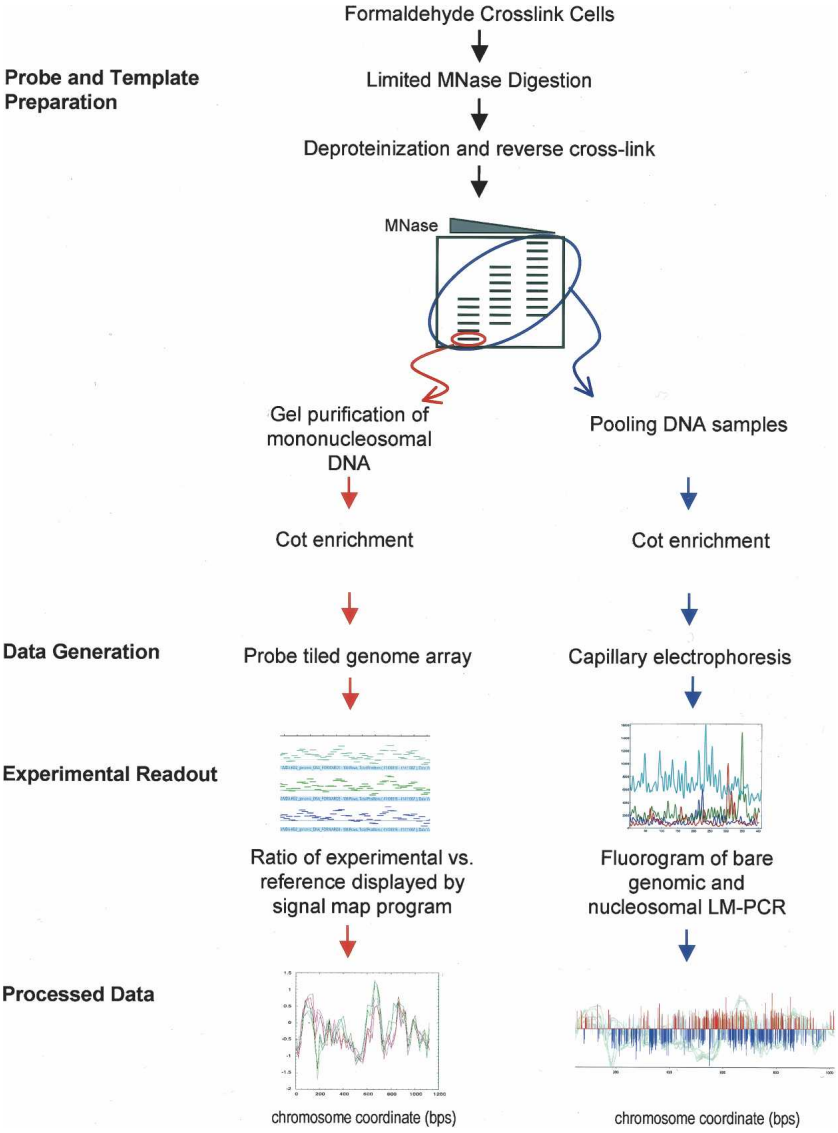

Figure 1. Schematic representation of mononucleosomal array and primer extension capillary electrophoresis procedures.

ternucleosomal cleavage reactions carried out on pre-extracted, formaldehyde-cross-linked nuclei is shown in Figure 2. The starting material for both protocols can be generated simultaneously. This allowed us to directly compare the ability of each protocol to characterize nucleosome structures.

For the purpose of mononucleosomal purification, DNA ladders that consisted of one to seven nucleosomes were used. We chose this population size in order to reduce the number of nicks occurring within the bounds of the nucleosome, as it has been reported that DNA digested completely to mononucleosomes contains nicks in the nucleosomally protected region (Fragoso and Hager 1997). The nucleosomal ladder from this condition was isolated using preparative electrophoresis, and the mononucleosomal band was excised and gel-purified.

The LM-PCR experiment template was prepared by combining all lanes of the titration such that the population of nucleosomal fragments contained one to 15 nucleosomes (~150–3000 bp). We reasoned that the combination of the differing ranges of the MNase digestion would give more consistent length reads in the primer extension experiments, thereby facilitating the comparison of biological replicates.

Because there is a slight sequence preference to digestion of DNA by MNase, a bare genomic DNA sample was always included in addition to the digested chromatin (Fig. 2). As with the internucleosomally cleaved samples, the bare genomic DNA was cleaved with increasing amounts of MNase, and the reactions were combined to give a bare genomic library of ~100–3000 bp.

Cot enrichment from total human genomic template

We tested our two protocols on the MMTV promoter in a cell line with a low copy number of integrated MMTV, and on the single-copy *LCMT2* gene (Lu et al. 1994; Wallrath et al. 1994). One way to create a more reliable template for mapping cleavage sites on single-copy genes in mammals is to increase the unique (and genic) component and reduce the repetitive component of the genome, thus making the template more similar to that prepared from lower eukaryotes. We used a DNA reassociation kinetics technique, Cot enrichment, to increase the complexity of our sample and thereby achieve this goal (Britten and Kohne 1968; Peterson et al. 2002a,b).

In Cot-based enrichment, total genomic DNA is heat-denatured and allowed to reassociate to a Cot value at which a majority of the repetitive component reassociates, but the single- and low-copy component remains single-stranded. (The Cot value is the product of the molar concentration of nucleotides, reassociation time, and a factor based on the cation concentration of the buffer.) Hydroxyapatite chromatography is used to separate double-stranded, repetitive DNA from single-stranded, low-copy DNA. The single-stranded fraction contains the gene-enriched template (Peterson et al. 2002a,b). We performed Cot enrichment by reassociating denatured human DNA to a Cot value of 3000 M·sec and collecting the single-stranded eluent. The Cot value of 3000 M·sec was chosen because it is near the predicted $Cot_{1/2}$ value for the single-copy component of the human genome.

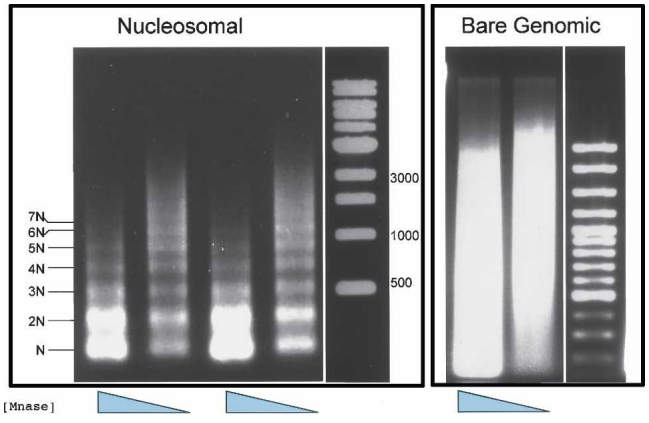

Figure 2. Micrococcal nuclease cleavage of nucleosomal and bare genomic DNA. Nuclei from MDA-kb2 cells and chromosomal DNA were digested with micrococcal nuclease as described in Methods. DNAs were isolated, and 2 µg was loaded in each lane of a 1.2% agarose gel.

To test the efficacy of our Cot enrichment, we used real-time quantitative PCR (RT-Q-PCR) (Fig. 3). We used either 5 or 50 ng of both nucleosomal and bare genomic DNA cleaved with varying concentrations of MNase as template in two RT-Q-PCR experiments. The threshold cycle values (C_T) were compared for both the MMTV-LTR and the *LCMT2* gene. When 5 ng of template was amplified using primers specific to the MMTV-LTR, the total genomic samples required between 7.7 and 10.0 more cycles to reach the C_T than their Cot-enriched counterparts. Likewise, when 50 ng of template was amplified with the same primers, the total genomic samples required 5.5–7.7 more cycles to reach the C_T. This pattern, which reflects the degree of enrichment, also held true for the *LCMT2* gene in which total genomic samples required between 6.4 and 11 more cycles than their Cot-enriched counterparts to reach the C_T when 5 ng of template was amplified, and 6.8–8.4 more cycles when 50 ng of template was amplified. These results indicate a consistent and significant enrichment of these single-copy genes in the human genome using Cot hybridization. These enriched DNA samples served as the material for both mononucleosomal hybridization to the tiling array and LM-PCR primer extension experiments.

Hybridization of Cot-enriched mononuclosomes to tiling microarray

We interrogated a tiling microarray using our gel-purified, Cot-filtered mononucleosomal DNA fragments. In this proof-of-principle experiment we chose to design probes that spanned two genomic regions: MMTV-LTR and *LCMT2*. We first tested the MMTV-LTR, which is widely regarded as the best characterized example of translational positioning of nucleosomes in mammals. This sequence, spanning 1.2 kb, is organized into six nucleosomes (Richard-Foy and Hager 1987; Truss et al. 1995; Belikov et al. 2000). We also probed the *LCMT2* gene, an important regulator of the cell cycle. We designed our tiling microarray of overlapping 50-mer oligonucleotides spaced 20 bases apart. Each gene was spotted in triplicate on both forward and reverse strands to give redundant and overlapping data sets. Gel-purified, Cot-enriched DNA was hybridized to a DNA tiling array by Nimblegen. The bare and nucleosomal DNA preparations were fluorescently labeled with Cy5 and Cy3, respectively, and hybridized to the tiling array. The raw data are displayed as a $\log_2$ ratio plot [mononucleosomal DNA signal (Cy3)/genomic DNA signal (Cy5)] spanning the tiling array for MMTV and *LCMT2* (Fig. 4A,C). Regions protected from MNase digestion are expected to result in peaks, while regions more accessible to MNase digestion will result in valleys in the $\log_2$ ratio plot. Clear peaks and valleys were seen when both MMTV and *LCMT2* were used probed the tiling microarray.

The $\log_2$-transformed raw data were normalized using the quantile-quantile method (Bolstad et al. 2003) to rescale the dynamic ranges of all replicates to be identical. The bare genomic ($\log_2$ Cy5) values were then subtracted from the mononucleosomal ($\log_2$ Cy3) values. The signals (log ratios) of three replicates on both forward and reverse strands were then plotted (Fig. 4B,D).

There is a high level of consistency of the signal between the replicates on each strand. The Pearson correlation between replicates on an individual strand is >0.9 for each dye channel. Because we tiled both strands, we could also compare patterns between strands, which would be expected to be similar as nucleosomes organize double-stranded DNA. The agreement between the probes on the forward and reverse complementary strands was striking, with a Pearson correlation >0.8.

The MNase-accessibility pattern at the MMTV-LTR obtained from probing MNase-resistant mononucleosomal DNA to our tiling microarray agrees with the published data on the translational positioning of nucleosomes in this region. The MMTV-LTR region spotted onto our tiling array covers an area in which we expect to see five of the six positioned nucleosomes. Based on the published positions (Richard-Foy and Hager 1987; Truss et al. 1995; Belikov et al. 2000), we expected nucleosomal protection between bases 75 and 265 for nucleosome B, 265 and 460 for nucleosome C, 460–670 for nucleosome D, 670–840 for nucleosome E, and 840–1030 for nucleosome F. Within the resolution of the indirect end-labeling experiment on which these protections are based, the microarray data are concordant with the previously published data on this promoter. We do see a larger, less distinct, protection in the area correlating with nucleosome C; this might reflect multiple positions for this nucleosome in this particular cell line and this particular integration event of the MMTV construct.

We next examined the MNase-accessibility pattern of a previously uncharacterized region, the *LCMT2* promoter. The 2-kb region centering the transcription start site of *LCMT2* displayed a high signal-to-noise ratio. Seven discrete peaks of protection as well as an eighth area of more diffuse protection were easily identified in this region. The seven peaks ranged from ~100 to 200 bases of protection, suggesting that they might result from protection by translationally positioned nucleosomes. The eighth more diffuse peak appears to occupy ~450 bases. This noncanonical protection might be the result of one or more nucleosomes occupying several translational frames, or protection of this region by a protein complex not removed during the pre-extraction of the nuclei. The results demonstrate the potential to use this microarray-based technique to map chromatin structure of uncharacterized mammalian genomic loci.

LM-PCR amplification and labeling of Cot-enriched template

We next sought to recapitulate the reproducible pattern of nucleosomal protection in the microarray experiment using a protocol based on LM-PCR primer extension (Mueller and Wold 1989). In contrast to the microarray protocol that locates protected regions, this protocol maps DNA cleavage sites introduced by MNase, and thus offers a completely independent method to measure cleavage of a sample by MNase using a methodology that can be adapted to high throughput. We reasoned that two independent high-throughput protocols would allow us to validate cleavage patterns.

Briefly, the LM-PCR allows the analysis of fragments extended from a gene-specific primer to multiple MNase cleavage points. In this procedure, a universal linker is ligated to the blunt end of a population of fragments extended from a primer to a specific region of the genome. Subsequent PCRs using nested gene-specific primers and the known universal linker are used to amplify and detect the MNase-cleaved fragments. The first step in this process was to use kinase to phosphorylate the Cot-enriched DNA sample, because cleavage by MNase does not leave an intact 5′-phosphate, necessary for ligation. The second step was to primer extend gene-specific primers to a blunt end suitable for ligation of a universal primer. The resulting gene-specific blunt ends were ligated with a universal linker. This DNA population was then subjected to an exponential PCR using the universal linker and a nested gene-specific primer. This PCR product was then labeled in a linear PCR using a third nested 6-FAM fluorescently labeled primer. These labeled products were loaded onto a capillary

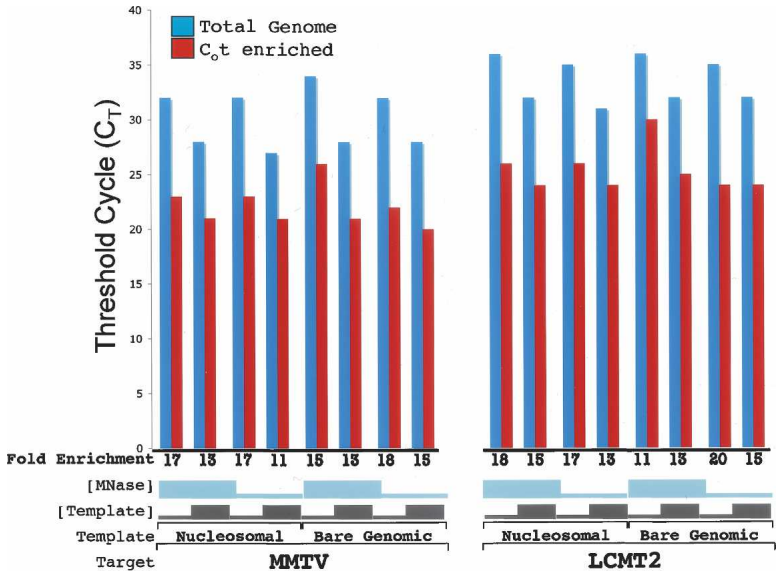

Figure 3. Quantitative PCR of total genomic and Cot-enriched DNA samples. Quantitative PCR was used to confirm enrichment of MMTV and *LCMT2* target DNA. Nuclei from MDA-kb2 cells (nucleosomal) and chromosomal (bare genomic) DNA were digested with two concentrations of micrococcal nuclease as described in Methods. DNAs were isolated, melted, and allowed to rehybridize to a Cot value of ~3332.2 M·sec. The slowly reassociating single-stranded component was purified by hydroxyapatite chromatography as described in Methods. Five or 50 ng of this Cot-enriched DNA (red columns) was compared with 5 or 50 ng of the non-Cot-enriched starting material (blue columns) in a quantitative PCR experiment assessing the enrichment of two loci, MMTV and *LCMT2*. The fold enrichment between the total genomic sample and the Cot-enriched sample was calculated using an amplification efficiency of 1.84. The value shown is C_T, the fractional cycle number at which the amount of amplified copies reaches a fixed threshold.

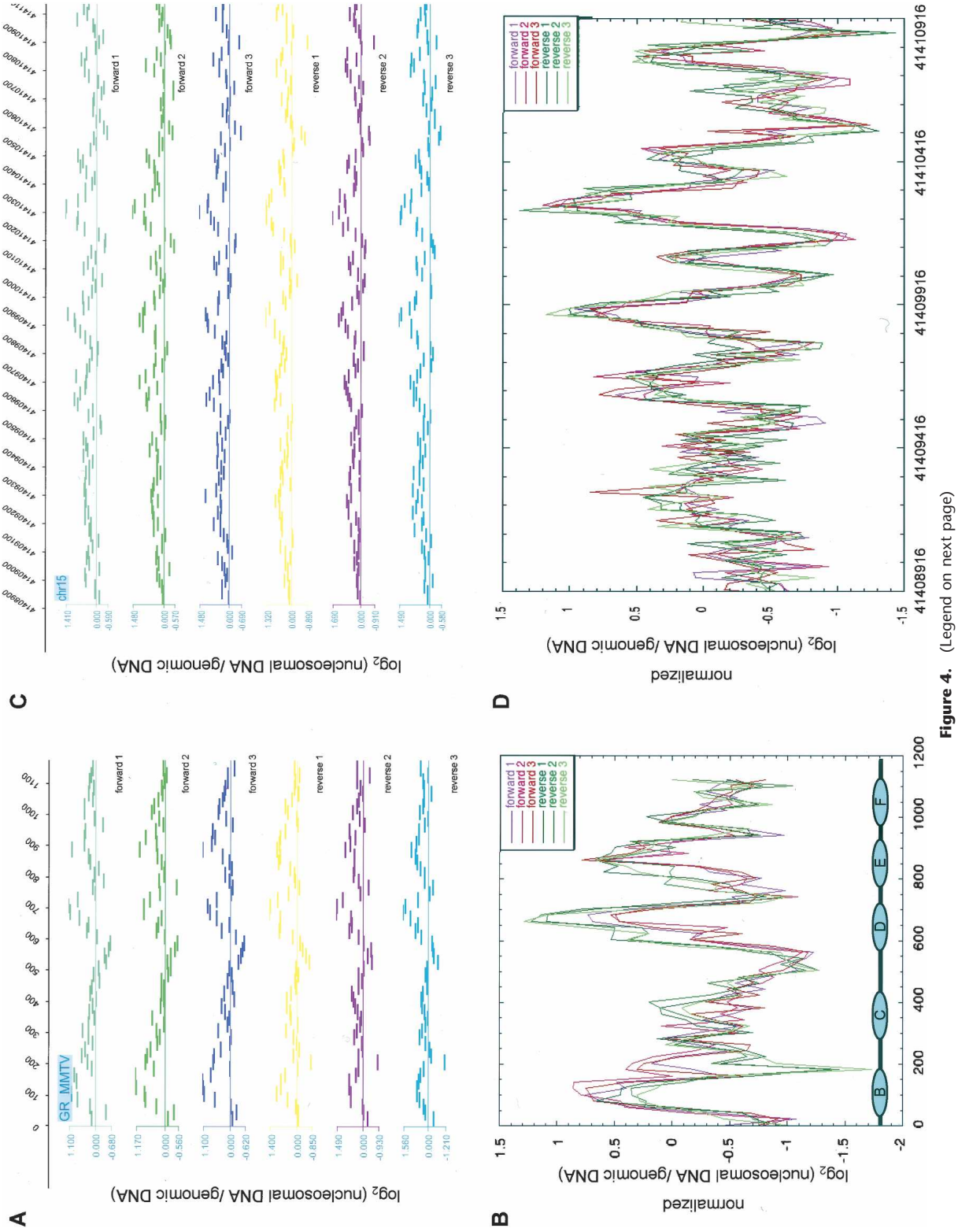

Figure 4. (Legend on next page)

electrophoresis sequencer. The binary output file from the capillary sequencer contains the data from the 6-FAM fluorescently labeled PCR products, and a set of LIZ fluorescently labeled molecular weight markers (MWMs) that were run simultaneously through the capillary. These unprocessed data maintained all the original information regarding peak height. Because peak height is a reflection of the amount of MNase cleavage, we designed our analysis to interpret the peak height at each nucleotide as an indicator of the amount of cleavage occurring at that position.

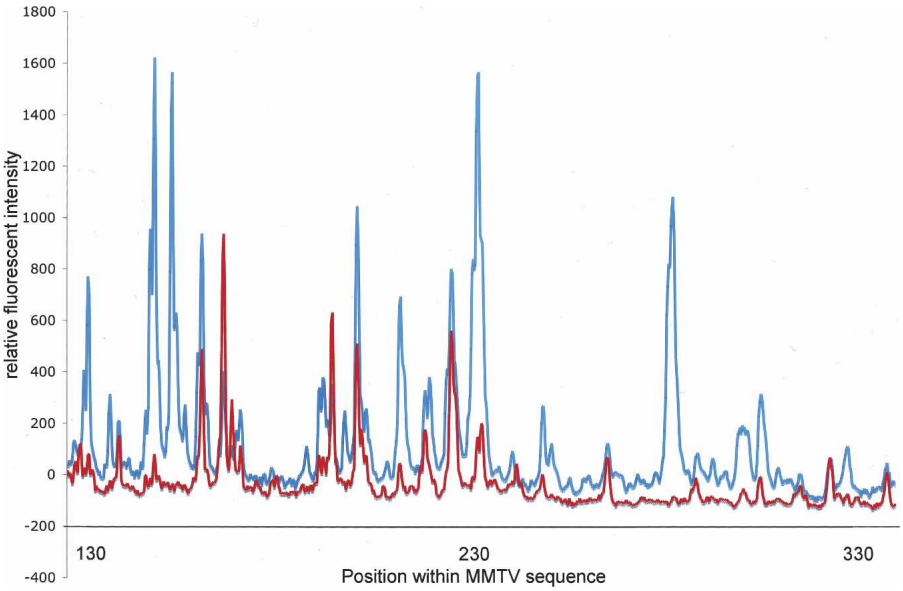

Figure 5. Example of an aligned primer extension capillary electrophoresis fluorogram trace. Raw fluorescence data from an ABI3730xl sequencer were aligned to ABI GeneScan500 molecular weight markers. Bare genomic DNA is in red, and internucleosomally cleaved DNA is in blue.

Analysis of capillary electrophoresis primer extension data

MMTV–LTR

We used primer extension to analyze cleavage sites in the MMTV-LTR. The entire data set consists of three biological replicates each for nucleosomal (N1, N2, N3) and bare-genomic (B1, B2, B3) DNA samples. Each sample was amplified through two separate LM-PCR steps (1, 2), and then each of these was run twice on the capillary electrophoresis sequencer (a, b). In addition, for the MMTV analysis, we used six different overlapping primers: three on the forward strand and three on the reverse strand. For each primer, we therefore have 24 samples at each base position. At this stage, we performed a log transform of the raw data, which resulted in data that are approximately Gaussian-distributed. Reproducibility within all N and B samples was extremely high, with a Pearson correlation coefficient >0.9 in all cases, and a median value of 0.97.

A key issue with this protocol is the development of software tools to allow automated interpretation of the data sets. Manual alignment and measurement of peak heights are not practical if this technology is to be used in a high-throughput manner. We developed a systematic and automated protocol for the unbiased measurement of the relative amount of cleavage at each nucleotide. The first step in the analysis involves aligning the data track to the DNA sequence by using the positions of the MWMs as guides. This alignment is done using a piecewise linear interpolation between adjacent MWMs. The smallest marker corresponds to 50 bp from the 5′-end of the primer, and the largest marker is at 510 bp. The output of the alignment is a uniformly sampled fluorogram from 50 bp to 510 bp, with four samples per base (Fig. 5).

These aligned data were then analyzed using the classical statistical method known as analysis of variance (ANOVA) (Sahai and Agell 2000). Our ANOVA used four factors to account for different sources of variance in the data: the sample type, the LM-PCR, the capillary electrophoresis run, and the sample index (which corresponds to the day on which the analyses were run). The ANOVA as used here does not consider interaction effects. A preliminary ANOVA that did include interaction terms showed that, as expected, these effects are minimal for our data set. The ANOVA results give us a *P*-value for each factor at each base position, indicating whether that factor is the cause of a difference at that base position. The single factor causing the most statistically significant differences was, as expected, the sample type (N vs. B). The results of this analysis are shown in Figure 6. The red and blue lines show the significance ($\log_{10}$ *P*-value) at each base. Note that a *P*-value is always <1, and therefore the log will always be negative. In these plots, the height of the red or blue bars is equal to the magnitude |log(*P*-value)|. The choice of color and whether the bar extends up or down from the $Y = 0$ midline is based on whether B > N (red, above midline) or N > B (blue, below midline). The black line shows the average $\log_2$(intensity ratio of the bare genomic signal vs. the nucleosomal signal).

In order to combine the results from the six primers, the primers were ranked according to which yielded the most significant peaks. When piecing together the six individual traces shown in the six panels in Figure 6A, wherever there was overlap, the highest-ranked single trace was used. We reasoned that the number of peaks associated with a given primer set should be an indicator of the quality of the data associated with that primer set, and that selecting the higher quality data should therefore give better results than simple averaging. This approach allowed us to use a defined parameter to assess the quality of data and to remove from analysis regions of lower quality. Based on this

Figure 4. Hybridization of Cot-enriched mononucleosomal DNA fragments to a tiling microarray. Gel-purified Cot-enriched mononucleosomal DNA was labeled with Cy3, and sonicated Cot-enriched bare genomic DNA was labeled with Cy5; both were hybridized to a tiling microarray containing (*A*,*B*) the MMTV LTR and (*C*,*D*) the promoter region of the *LCMT2* gene. Probes were 50 bases long and spaced 20 bases apart. Each probe was spotted in triplicate on both the forward and reverse strands. Replicate probe data from MMTV (*A*) and *LCMT2* (*C*) are shown as the $\log_2$ ratio of mononucleosomal DNA (Cy3) to bare genomic DNA (Cy5). (*B*,*D*) Each probe from the six replicate data sets (three from the forward strand and three from the reverse strand) was log-transformed, normalized, and plotted as log ratios for MMTV (*B*) and *LCMT2* (*D*).

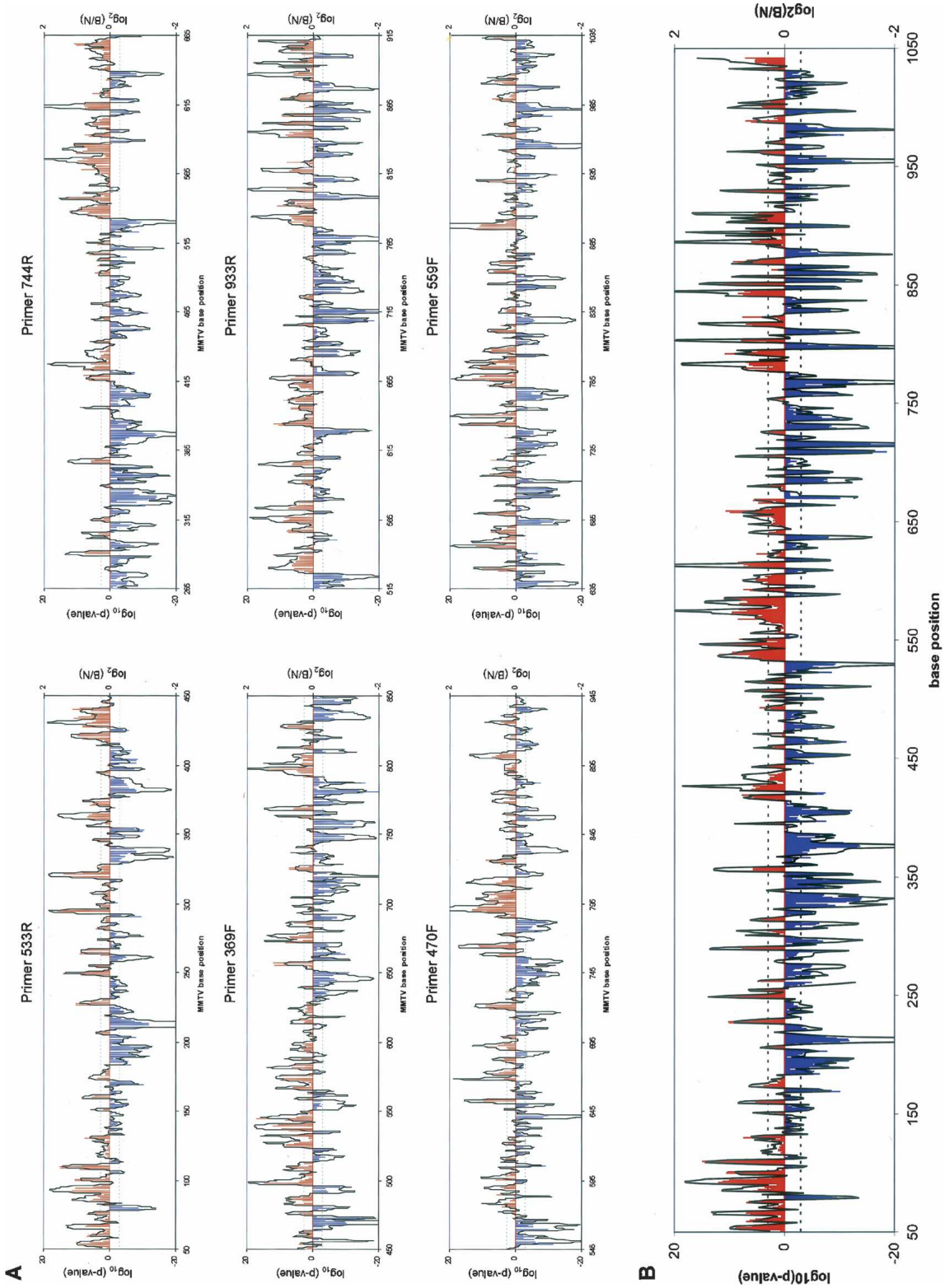

Figure 6. (Legend on next page)

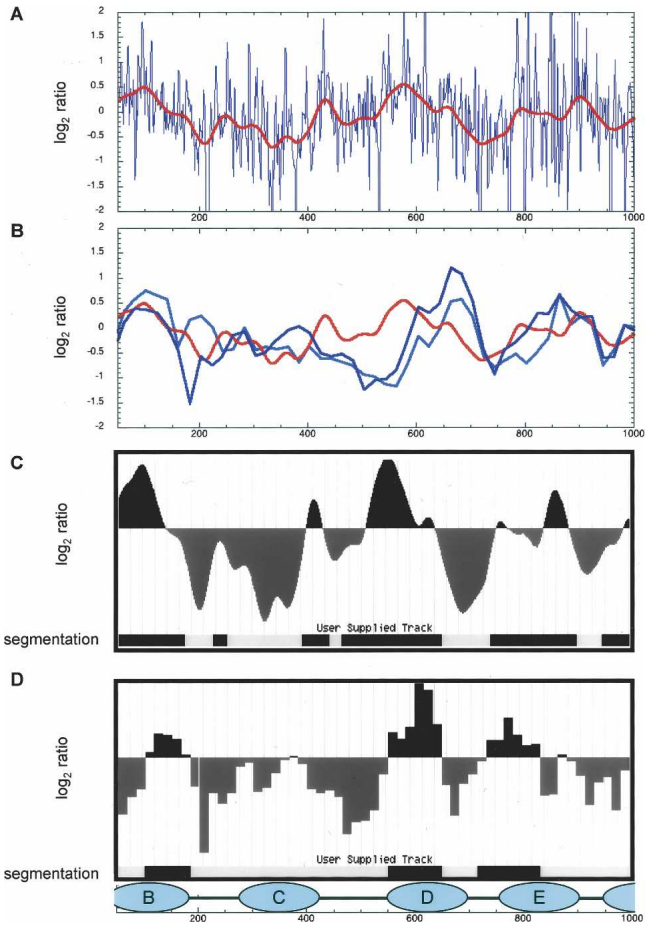

Figure 7. Smoothed $\log_2$(intensity ratio [B/N]) for MMTV-LTR primer extension and comparison with microarray data. (*A*) The information provided by each primer, as shown in Figure 6, was combined to produce the summary graph shown here. Data points in blue show the $\log_2$(intensity ratio [B/N]) of the mean intensities for each trace. The red line is a smoothed curve from a Gaussian window length of 75 and a standard deviation of 15 bases. (*B*) The smoothed primer extension curve in red is overlaid on the microarray data, which is shown as $\log_2$(intensity ratio[N/B]). Note that the primer extension curves show the ratio as [B/N], while the microarray curves are shown as the ratio [N/B]; by plotting the data this way, a "high" region implies protection from cleavage in both cases. The light blue trace and the dark blue trace represent the average of the three replicate forward probes and the three replicate reverse probes, respectively. (*C*, *D*) These data were further segmented into two states—"accessible," shown in black, and "inaccessible," shown in white—using a HMM. A HMM was trained on the $\log_2$intensity ratio signal for the smoothed primer extension data (*C*) or the microarray reverse strand data (*D*) for the purposes of comparison. The data traces in *C* are identical to those shown *B*. The bar below each data trace shows the segmentation of the data.

ranking, the three bottom-strand primers were preferred. This analysis across MMTV-LTR is shown in Figure 6B.

The $\log_{10}$(*P*-value) and $\log_2$(intensity ratio) curves show a strong high-frequency variation. In order to get a sense of the slower underlying variation that is related to the chromatin accessibility, we smoothed these curves using a Gaussian window length of 75 bases and standard deviation of 15 bases. These smoothed curves are shown in Figures 7 and 8. They show a strong degree of similarity to the microarray data for the same regions of DNA.

In order to further segment these data into two states, which can be interpreted as "accessible" and "inaccessible" to MNase, we used a two-state, first-order hidden Markov model (HMM) trained with a single Gaussian at each state. The parameters of the model were learned in an unsupervised fashion using the expectation maximization (EM) algorithm. We started EM 100 times from random initial parameters and selected the learned parameters that best explain the data (i.e., that yield the largest posterior probability). Using the trained model, we then segmented the data using the Viterbi algorithm (Figs. 7C, 8C).

Inspection of this output reveals multiple clusters of significant points both above and below the central N = B line. We found that the clustering of points with significant *P*-values where N > B correlated with proposed MNase accessibility, including the region between nucleosomes B and C: 175–225 bp, and the region between nucleosomes D and E: 675–750 bp. Additionally, we found that clusters of points with significant *P*-values where B > N correlated with regions of nucleosomal protection, most notably from 50–180, identifying nucleosome B, 500–675 bp, identifying protection by nucleosome D, and 750–900 bp, identifying protection by nucleosome E. Interestingly, less clustering of points was seen to correlate with protection by nucleosome C; MNase was able to cut within this predicted area of protection, indicating that this nucleosome has a probabilistic distribution at multiple locations between 200 and 500 bp.

LCMT2

The *LCMT2* data set consisted of the same N and B samples, and used one primer set to analyze the 5′-end of this promoter. Data from this primer were handled in a manner identical to that of the MMTV-LTR. Inspection of this data set suggests that the first ~250 bases from the primer are MNase-accessible followed by an ~75-base region of protection where the trace ends (Fig. 8).

Discussion

Characterization of how chromatin organization affects DNA availability is central to our understanding of how regulatory factors such as proteins and RNA gain access to specific areas of the genome. There are few reports describing accessibility to specific areas of mammalian genomes in cultured cells, and these studies are generally limited to areas much smaller than the transcription of a gene. Here we present two independent, high-throughput methods for the analysis of chromatin structure in mammalian cells, a microarray-based method and a primer extension-based method. Each of these techniques can be readily adapted to explore structure over large regions.

Figure 6. Primer extension data for the MMTV-LTR from each primer and their combined result. (*A*) The PCR extensions from each primer were analyzed to determine locations of significant difference between the bare genomic and nucleosomal DNA (B and N, respectively). The 5′-start position of each primer is given as well as the strand it primes from (F, forward; R, reverse). In each graph, the red and blue lines indicate the locations and $\log_{10}$(*P*-value) of significant differences between the two, while the black line shows the $\log_2$(intensity ratio) of the mean intensities for each trace. The dotted lines represent a *P*-value of 0.001, and have been included solely as a reference point for the reader. Above midline indicates B > N, below midline indicates N > B. (*B*) Combined information from all six primers to reveal MNase accessibility across the MMTV-LTR.

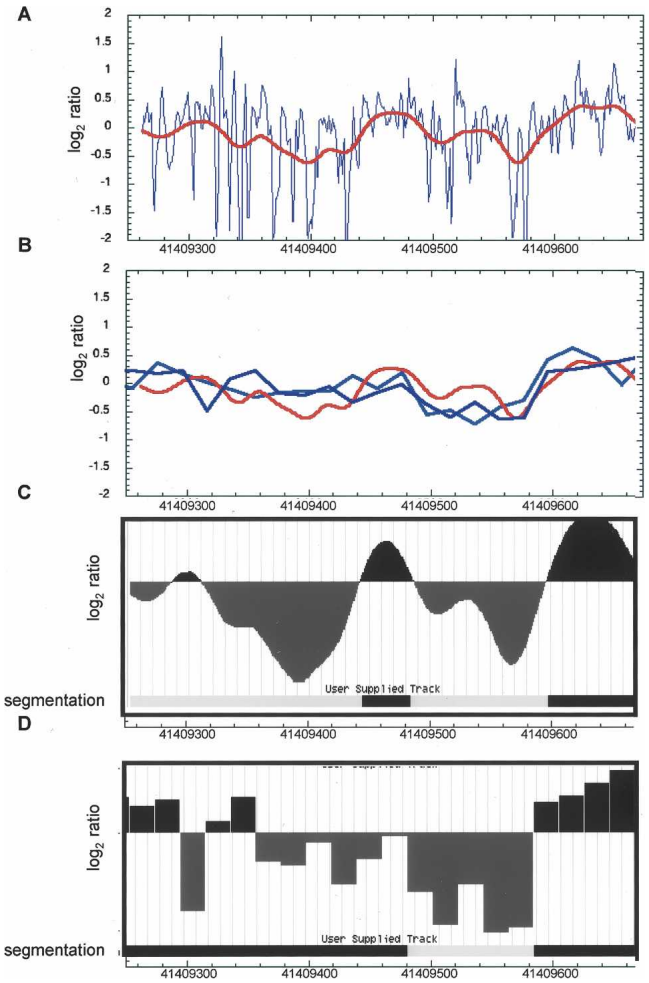

Figure 8. Smoothed $\log_2$(intensity ratio [B/N]) for *LCMT2* primer extension and comparison with microarray data. (*A*) Data points in blue show the $\log_2$(intensity ratio [B/N]) of the mean intensities for each trace. The red line is a smoothed curve from a Gaussian window length of 75 and a standard deviation of 15 bases. (*B*) The smoothed primer extension curve in red (scale modified from that shown in panel *A* to show detail) is overlaid on the microarray data, which is shown as $\log_2$(intensity ratio[N/B]). Note that the primer extension curves show the ratio as [B/N], while the microarray curves are shown as the ratio [N/B]; by plotting the data this way, a "high" region implies protection from cleavage in both cases. The light blue trace and the dark blue trace represent the average of the three replicate forward probes and the three replicate reverse probes, respectively. (*C*, *D*) These data were further segmented into two states—"accessible," shown in black, and "inaccessible," shown in white—using a HMM. A HMM was trained on the $\log_2$intensity ratio signal for the smoothed primer extension data (*C*) or the microarray reverse strand data (*D*) for the purposes of comparison. The data traces in *C* are identical to those shown in *B*. The bar below each data trace shows the segmentation of the data.

Comparison of the microarray and primer extension data

The microarray technique and the primer extension technique each give information regarding chromatin structure. Broadly speaking, the microarray protocol measures nucleosome occupancy, while the primer extension protocol measures nuclease accessibility. These two parameters should be closely related, as nucleosome occupancy is defined, both historically and in the work presented above, by regions resistant to MNase cleavage. If the protocols we describe were accurate, then we anticipated that there should be concordance between the two methods. The two technologies would not be expected to be completely concordant, because, for example, the microarray protocol will not detect situations where there is a nucleosomal site that happens to be sensitive to MNase cleavage, whereas the primer extension protocol will detect such a site. Both protocols should reflect true nucleosome position, however, as both protocols should detect internucleosomal regions with high amounts of MNase cleavage. We compared the cleavage patterns produced by both technologies on the MMTV and *LCMT2* promoters (Figs. 7, 8).

There are examples of the synergy between the data resulting from the two techniques. For example, in the region between 750 and 850 in the MMTV sequence, the microarray data from the replicate experiments on the forward and reverse probes are slightly divergent, making it difficult to infer what the state of the chromatin is at this nucleosome boundary. The primer extension data for this region, which are consistent between two overlapping primer sets representing both strands, show a region of protection framed by areas of accessibility on either side. This pattern suggests that the nucleosome here might occupy two positions, or that there may be DNA access within the bounds of the histone octamer. This situation holds true for the region between 600 and 700, as well. In this region we see a local minimum in the microarray data, which occurs consistently in all six traces. The primer extension data at this minimum indicate that DNA in this area was exposed, and available for cleavage by MNase: another example of DNA access within the bounds of the histone octamer.

The primer extension data, which have single-nucleotide resolution, should allow us to visualize subnucleosomal cleavages, a level of detail not possible in the microarray experiment. An example of this is seen with previously described nuclease accessibility within the MMTV-LTR that are not coincident with internucleosomal regions (Belikov et al. 2000). These cleavages are found within nucleosome B, where there are multiple sites for DNA-binding proteins, as well as within the bounds of the area protected by nucleosome D. We observe a region of accessibility between bases 175 and 225 in nucleosome B using the primer extension data; accessibility in this region is less distinct in the microarray data. The HMM segmentation recognizes this increased accessibility in the primer extension experiment, but not in the microarray experiment (Fig. 7). Similarly, accessibility within the bounds of nucleosome D, bases 450–500, is suggested by both methods but is more pronounced in the primer extension experiment. Again, this cleavage is recognized in the HMM segmentation exclusively in the primer extension experiment. These examples indicate that the primer extension technology can augment the microarray analysis in detecting cleavage events that are not caused by canonical nucleosome structure.

Each method has differing levels of resolution. The microarray offers the simultaneous interrogation of hundreds of kilobases of sequence, with a resolution of ~50 bases. The primer extension method is approximately an order of magnitude more limited in range (analysis of tens of kilobases is reasonable) but has a commensurate increase in resolution, giving the precise nucleotide position of DNA accessibility. The synthesis of information from these technologies has the potential to provide a very clear picture of chromatin accessibility throughout large stretches of the mammalian genome. The patterns of chromatin accessibility revealed by these technologies reflect protection of the DNA by proteins or protein complexes, including nucleosomes. Interpretation of this information does not require that

we call actual nucleosome position, as understanding the precise location of changes in chromatin structural changes is useful even without understanding the precise alterations to nucleosome location that give rise to those structural changes. Data from these experiments can lead to tentative assignment of nucleosome position, however, that can be verified by further experiments (i.e., indirect end labeling, restriction endonuclease accessibility, as well as in vitro biochemical experiments). Additionally, experiments described herein involving cell lines differentiated along different paths are currently underway and should yield valuable insight into how chromatin structure play a role in transcriptional memory and maintenance of cellular function.

The potential to take advantage of these methods is within the scope of most laboratories. Core sequencing facilities and core microarray facilities are becoming more prevalent in research departments; additionally, commercial offerings for each of these services are numerous. One well-designed chip can simultaneously interrogate up to 200 kb, while one 384-well plate can cover 10 kb in an overlapping and redundant manner using the primer extension methodology. A microarray is generally probed with 10 µg of DNA, and the 384-well plate described above could be completed with 100 µg of DNA. The number of cells this work requires is thus within the growth capacity of any lab doing tissue culture.

Reliability and sensitivity of the two methods

The two methods probe opposite aspects of the accessibility phenomenon. The signal from the hybridization of mononucleosomal DNA to the tiling microarray comes from the regions of DNA that are protected from cleavage. The signal from the primer extension experiment is the result of the extension of chromatin DNA up to its MNase-accessible region. Both experiments give highly concordant results. Not only does each method suggest the five predicted nucleosomal locations within the MMTV-LTR region investigated, but both technologies have very similar chromatin accessibility patterns (Fig. 7). Additionally, in the case of the *LCMT2* gene, when the output from each of the methods is compared, the similarity is striking (Fig. 8).

While the two data sets are highly concordant, the differences between the microarray and the primer extension experiment could possibly be the result of slight differences in the representative genomic populations. The set of mononucleosomal DNA fragments may represent a particular population of rapidly digestible positions on the cross-linked chromatin. Additionally, aspects of the LM-PCR, including biases in the ligation of the universal linker and the exponential nature of the amplification step, could be responsible for the over- or under-representation of particular MNase-cleaved fragments. In spite of these technical issues, the signal from each of these methods is able to indicate the accessibility of the chromatin in a reproducible and redundant manner.

High–throughput analysis for high–throughput methods

The high-throughput nature of a technology is frequently limited by the ability to analyze the data as much as by generating the data. We have developed the necessary tools to rapidly understand the quality and nature of the data from each method. Furthermore, we have implemented a two-state HMM to distill the fine information of each method into a binary call of accessible or inaccessible. This tool will be valuable in the archiving and display of this information as the catalog of chromosomal regions characterized by these methods grows.

Improvements to these methods can be made. The signal-to-noise ratio and data analysis can be improved in the microarray analysis. Signal quality and read length can be improved in the primer extension experiments. The availability of high-throughput methods that are complementary and can be applied to large regions of mammalian genomes makes an unbiased analysis of chromatin structural changes over entire genetic loci (e.g., the HOX clusters) feasible. These technologies allow chromatin structure to be analyzed on the same genomic scale that ChIP analysis can be reliably performed. Correlating changes in chromatin structure with changes in the binding of regulatory factors, or with changes in histone modification, will allow a more thorough analysis of chromatin structural changes during regulatory events (Fan et al. 2004). This will facilitate the formation of hypotheses concerning direct and indirect mechanistic interfaces between histone covalent modification, binding events, and structural transitions.

Methods

Cell growth and nuclei purification

MDA-kb2 cells (ATCC number: CRL-2713) were grown in Leibovitz's L-15 medium. Ninety-five percent confluent MDA-kb2 cells were pre-extracted with PBS supplemented with 150 mM NaCl, 0.2% Tween 20, and 0.2% Triton X-100 for 5 min at room temperature. The cells were then crossed-linked by adding formaldehyde to final concentration 1% and incubated for 10 min at room temperature. Glycine was then added to 125 mM to quench formaldehyde. Fixed cells were collected by scraping followed by centrifugation at 1000*g* for 10 min.

Cell pellets were resuspended in sucrose buffer (0.3 M sucrose, 2 mM MgOAc, 3 mM $CaCl_2$, 1% Triton X-100, and 10 mM HEPES at pH 7.8), and Dounce-homogenized with a loose-fitting pestle. The homogenate was then diluted 1:1 with GB buffer (25% glycerol, 5 mM $MgAc_2$, 0.1 mM EDTA, and 10 mM HEPES at pH 7.8). To isolate the nuclei, the resulting solution was layered on an equal volume of GB buffer, and spun at 1000*g* for 15 min.

MNase cleavage and mononucleosomal purification

Fixed nuclei and total bare genomic DNA were digested with titrated amounts of MNase (Sigma). MNase-digested nuclei were then treated with proteinase K, and the cross-links were reversed by overnight incubation at 65°C. To isolate mononucleosomal DNA for the tiling DNA microarray, the nucleosomal DNA ladder was resolved on 1% low melt agarose, and the mononucleosomal DNA band was then cut out and purified using the QIAGEN Gel Purification System (QIAGEN).

Cot enrichment and quantitative PCR

To enrich for single-copy genes, DNA at a concentration of 2.5 mg/mL in 180 mM sodium phosphate (pH 6.8) was boiled for 10 min, and DNA was allowed to anneal at 55°C overnight (approximate Cot calculation: 0.04 M nucleotides × 50,000 sec × 1.5 buffer factor = Cot 3000). DNA grade hydroxyapatite (HAP) (Bio-Rad) resuspended in 180 mM phosphate buffer (pH 6.8) was then added to the DNA, and the HAP-DNA mixture was incubated at 55°C. After 1 h, HAP was removed by spinning the entire reaction through a Costar SpinX column (Fisher). The flow-through contained Cot-enriched DNA. Phosphate buffer was removed by buffer exchanged into TE using a MicroCon MWCO 10K spin

column (Amicon). The Cot-enriched DNA was then NaOAc/EtOH-precipitated and resuspended in 0.1× TE.

DNA from total genome or Cot enrichment was quantified by full-scale OD in TE containing 0.1M NaOH to correct for the hypochromic effect of double-stranded DNA and allow for the accurate comparison of DNA concentrations between the total genomic sample and the single-stranded Cot-enriched sample. PCRs were set up using either 5.0 or 50.0 ng of DNA as template. Both bare genomic samples and nucleosomal samples cleaved with titrated amounts of micrococcal nuclease were used a template for primers to the MMTV-LTR or the *LCMT2* gene (MMTV forward, 5′-GGAAAACCTTTCCCCAAAAG-3′, and MMTV reverse, 5′-TGGGATAGGTGGGTCACAAT-3′ giving a 187-bp product; and *LCMT2* forward, 5′-TAGTCTGCGCTCTCAAAGCA-3′, and *LCMT2* reverse 5′-TGGCTTCGACTCGCTCTATT-3′ giving a 194-bp product). PCRs (50 µL) were set up using 250 fmol of primer per microliter using Bio-Rad SYBR Green SuperMix (Bio-Rad). Reactions were cycled 50 times and analyzed on the Bio-Rad i-Cycler.

LM-PCR

DNA phosphorylation reaction: 100 µg of Cot-enriched DNA was prepared in a 250-µL reaction volume with 25 µL of 10× T4 DNA ligase buffer (NEB). The volume was brought up to 230 µL with ddH_2O, and 20 µL of T4 polynucleotide kinase (10 U/µL) was added. The reaction was incubated for 1 h at 37°C. The reaction was stopped by incubation for 10 min at 75°C. The phosphorylated DNA was stored at −20°C.

Primer extension: 5.0 µg of phosphorylated DNA was brought up to 40 µL in 1× ThermoPol Buffer (NEB) containing 3 pmol of Primer1, and incubated for 10 min at 95°C then for 30 min at T_mPrimer1 − 2°C. Ten microliters of 1× ThermoPol Buffer containing 1 mM dNTPs and 1.0 U of Deep $vent_R$ Exo-DNA polymerase (NEB) was added. The solution was incubated for 10 min at T_mPrimer1 − 2°C, and then for 10 min at 76°C. This reaction was then placed on ice.

Linker ligation: 50 µL of freshly prepared 1× T4 DNA ligase buffer (NEB) containing 0.3 pmol/µL was added to the primer extension reaction, followed by 50 µL of T4 DNA ligase at a concentration of 1 U/µL. The ligation reaction was incubated overnight at 16°C, and then NaOAc/EtOH-precipitated.

PCR amplification: The precipitated pellet was resuspended in 50 µL of 0.1× TE containing 0.2 µM Primer2. Fifty microliters of 2× PCR MasterMix (Promega) was then added to the reaction. The reaction was amplified as follows: 4 min at 94°C, then 25 cycles of 30 sec at 94°C, 2 min at T_mPrimer2 − 2°C, and 4 min at 72°C; the reaction was finished with a 5-min incubation at 72°C.

Labeling reaction: The amplification reaction was used immediately in a 6-FAM labeling reaction. Five microliters of 0.02 µM 6-FAM-labeled Primer3 in 0.1× TE was added to10 µL of the labeling reaction. Five microliters of 2× PCR MasterMix (Promega) was then added to the reaction. The labeling reaction was cycled as follows: 2 min at 94°C, then five cycles of 30 sec at 94°C, 2 min at T_mPrimer3 − 2°C, and 4 min at 72°C; the reaction was finished with a 5-min incubation at 72°C. The labeling reaction was NaOAc/EtOH-precipitated, washed three times with 70% EtOH, and dried. Primer sequences for all reactions are available upon request.

Microarray hybridization and processing

Tiling genomic DNA microarrays were custom designed (NimbleGen Systems, Inc.). The 50-mer probes were selected every 20 bases with no repeat masking, from both forward and reverse strands. Three replicates for each strand were spotted on the array. Mononucleosomal DNA and genomic DNA were labeled with Cy3 and Cy5, respectively, and hybridized to the array by the manufacturer.

ABI3730xl capillary electrophoresis

Immediately before capillary electrophoresis, LM-PCR sample pellets were resolubilized in 5 µL of HiDi Formamide (ABI) containing 0.1 µL of Genescan 500LIZ molecular weight markers (ABI), heated to 94°C, and quick chilled on ice. An additional 20 µL of 0.03% molten agarose in ddH_2O was added to the sample, and the sample was run using the sequencing parameters on an ABI3730xl sequencer. Three alterations to the standard sequencing protocol were made: the electrokinetic injection voltage was increased to 2.0 kV, the electrokinetic injection time was increased to 60 sec, and the GeneMapper G5 Dyeset (compatible with 6-FAM and LIZ labels) was used.

Acknowledgments

We thank D. Altshuler for advice throughout the course of this research, T. Gillis for skillful assistance with customizing the ABI3730xl sequencer for these experiments, and C. Cotsapas and members of the Kingston Laboratory for a critical reading of this manuscript. This work was supported by NHGRI/ENCODE grant HG003141 (to R.E.K.), NHGRI/ENCODE grant HG003161 and NHGRI grant GM071923 (to W.S.N.), NSF grant DBI-0421717 (to D.G.P.), American Cancer Society grant PF-03-042-01-GMC (to J.H.D.), NIH/NCI award CA-093660 (to H.-Y.F.), and an NDSEG fellowship (to S.M.R.).

References

Anderson, J.D. and Widom, J. 2000. Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. *J. Mol. Biol.* **296:** 979–987.

Belikov, S., Gelius, B., Almouzni, G., and Wrange, O. 2000. Hormone activation induces nucleosome positioning in vivo. *EMBO J.* **19:** 1023–1033.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19:** 185–193.

Britten, R.J. and Kohne, D.E. 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161:** 529–540.

De Baere, I., Derua, R., Janssens, V., Van Hoof, C., Waelkens, E., Merlevede, W., and Goris, J. 1999. Purification of porcine brain protein phosphatase 2A leucine carboxyl methyltransferase and cloning of the human homologue. *Biochemistry* **38:** 16539–16547.

Fan, H.Y., Narlikar, G.J., and Kingston, R.E. 2004. Noncovalent modification of chromatin: Different remodeled products with different ATPase domains. *Cold Spring Harb. Symp. Quant. Biol.* **69:** 183–192.

Fragoso, G. and Hager, G.L. 1997. Analysis of in vivo nucleosome positions by determination of nucleosome-linker boundaries in crosslinked chromatin. *Methods* **11:** 246–252.

Fragoso, G., John, S., Roberts, M.S., and Hager, G.L. 1995. Nucleosome positioning on the MMTV LTR results from the frequency-biased occupancy of multiple frames. *Genes & Dev.* **9:** 1933–1947.

Kingston, R.E. and Narlikar, G.J. 1999. ATP-dependent remodeling and acetylation as regulators of chromatin fluidity. *Genes & Dev.* **13:** 2339–2352.

Kornberg, R.D. and Lorch, Y. 1999. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98:** 285–294.

Kornberg, R.D. and Lorch, Y. 2002. Chromatin and transcription: Where do we go from here. *Curr. Opin. Genet. Dev.* **12:** 249–251.

Kornberg, R.D., LaPointe, J.W., and Lorch, Y. 1989. Preparation of nucleosomes and chromatin. *Methods Enzymol.* **170:** 3–14.

Lu, Q., Wallrath, L.L., and Elgin, S.C. 1994. Nucleosome positioning and gene regulation. *J. Cell. Biochem.* **55:** 83–92.

Mueller, P.R. and Wold, B. 1989. In vivo footprinting of a muscle specific enhancer by ligation mediated PCR. *Science* **246:** 780–786.

Peterson, D.G., Schulze, S.R., Sciara, E.B., Lee, S.A., Bowers, J.E., Nagel, A., Jiang, N., Tibbitts, D.C., Wessler, S.R., and Paterson, A.H. 2002a. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res*. **12:** 795–807.

Peterson, D.G., Wessler, S.R., and Paterson, A.H. 2002b. Efficient capture of unique sequences from eukaryotic genomes. *Trends Genet.* **18:** 547–550.

Richard-Foy, H. and Hager, G.L. 1987. Sequence-specific positioning of nucleosomes over the steroid-inducible MMTV promoter. *EMBO J.* **6:** 2321–2328.

Sahai, H. and Agell, M.I. 2000. *The analysis of variance*. Birkhauser, Boston.

Simpson, R.T., Roth, S.Y., Morse, R.H., Patterton, H.G., Cooper, J.P., Murphy, M., Kladde, M.P., and Shimizu, M. 1993. Nucleosome positioning and transcription. *Cold Spring Harb. Symp. Quant. Biol.* **58:** 237–245.

Telford, D.J. and Stewart, B.W. 1989a. Characteristics of chromatin release during digestion of nuclei with micrococcal nuclease: Preferential solubilization of nascent RNA at low enzyme concentration. *Int. J. Biochem.* **21:** 1235–1240.

Telford, D.J. and Stewart, B.W. 1989b. Micrococcal nuclease: Its specificity and use for chromatin analysis. *Int. J. Biochem.* **21:** 127–137.

Truss, M., Bartsch, J., Schelbert, A., Hache, R.J., and Beato, M. 1995. Hormone induces binding of receptors and transcription factors to a rearranged nucleosome on the MMTV promoter in vivo. *EMBO J.* **14:** 1737–1751.

Wallrath, L.L., Lu, Q., Granok, H., and Elgin, S.C. 1994. Architectural variations of inducible eukaryotic promoters: Preset and remodeling chromatin structures. *Bioessays* **16:** 165–170.

Wilson, V.S., Bobseine, K., Lambright, C.R., and Gray Jr., L.E. 2002. A novel cell line, MDA-kb2, that stably expresses an androgen- and glucocorticoid-responsive reporter for the detection of hormone receptor agonists and antagonists. *Toxicol. Sci.* **66:** 69–81.

Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., and Rando, O.J. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309:** 626–630.

Received June 13, 2006; accepted in revised form November 9, 2006.