Supporting Information

Auerbach et al. 10.1073/pnas.0905443106

SI Text

Supplementary Materials and Methods

Growth of Cells for ChIP-Seq, Sono-Seq, MNase Digestion, Naked DNA and qPCR. For RNA Polymerase II ChIP-Seq, normal IgG ChIP-Seq and Sono-Seq, HeLa S3 cells were grown in suspension in Joklik's modified MEM, supplemented with 10% FBS to a density of 6 × 10^5 cells/mL. Cells were fixed with 1% formaldehyde at room temperature for 10 min and the fixation was terminated by the addition of glycine to a final concentration of 125 mM. The cells were washed in Dulbecco's PBS (Invitrogen), snap-frozen as cell pellets in liquid nitrogen, and provided to us by the National Cell Culture Center (Biovest International Inc.). For the preparation of the MNase-treated cells and naked DNA samples, HeLa S3 cells were grown in SMEM (Invitrogen), supplemented with glutamine, 10% FBS (Invitrogen), and antibiotics (penicillin-streptomycin) and harvested without cross-linking at a density of 5 × 10^5 cells/mL. For qPCR, HeLa cells were grown in MEM supplemented with 10% FBS (Atlanta Biologicals), 100 units/mL penicillin and 100 μg/mL streptomycin. A total of 10^7 cells were trypsinized, fixed in 10 mL MEM supplemented with 1% formaldehyde for 10 min at room temperature and quenched by the addition of glycine to a final concentration of 125 mM.

Construction and Sequencing of Illumina Libraries. DNA samples were run through Qiagen MinElute PCR columns, eluted with 15 μL of Qiagen buffer EB, and size-selected on 2% agarose E-gels (Invitrogen). Band-isolated fragments were gel-purified by using a Qiagen gel extraction kit. Libraries were prepared according to DNA Sample Kit instructions (Illumina Part# 0801–0303) but substituting kit enzymes with those available from other suppliers. Briefly, DNA was end-repaired and phosphorylated with the End-It kit from Epicentre (Cat# ER0720). The blunt, phosphorylated ends were treated with Klenow fragment (3′ to 5′ exo minus; NEB, Cat# M0212s) and dATP to yield a protruding 3′-'A' base for ligation of Illumina's adapters, which have a single 'T' base overhang at the 3′ end. After adapter ligation (LigaFast, Promega Cat#M8221) DNA was PCR-amplified with Illumina genomic DNA primers 1.1 and 2.1 for 15 cycles by using a program of (*i*) 30 s at 98 °C, (*ii*) 15 cycles of 10 s at 98 °C, 30 s at 65 °C, 30 s at 72 °C, and (*iii*) a 5-min extension at 72 °C. For the Sono-Seq DNA with large inserts of 350–800 bp, the extension time for the 15 cycles was increased to 1 min. The final libraries were band-isolated from an agarose gel to remove residual primers and adapters. Library concentrations and A_{260}/A_{280} ratios were determined by UV-Vis spectrometry on a NanoDrop ND-1000 spectrophotometer (NanoDrop). Purified library DNA was captured on an Illumina flowcell for cluster generation and sequenced on an Illumina Genome Analyzer II following the manufacturer's protocols.

Preparation of DNA for qPCR. Formaldehyde-fixed HeLa cell pellets were resuspended in cell lysis buffer [25 mM Hepes (pH 7.9), 1.5 mM $MgCl_2$, 10 mM KCl, 1 mM DTT, 0.1% Nonidet P-40] supplemented with EDTA-free protease inhibitor mixture (Roche) and 0.5 mM PMSF at a concentration of 10^7 cell equivalents/mL and incubated on ice for 10 min. After centrifugation, the crude nuclear pellet was resuspended in nuclear lysis buffer [50 mM Hepes (pH 7.9), 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS] supplemented with EDTA-free protease inhibitor mixture and 0.5 mM PMSF at a concentration of 10^7 cell equivalents/mL. Chromatin was sheared at 4 °C, 10 × 30 s at 30+ s intervals on a Branson Microtip Sonifier 450 set at constant duty and an output level of 4. After centrifugation for 10 min at 16,000 × *g*, chromatin was sonicated for an additional 0 to 5 min at 10 s intervals in 0.5 mL aliquots by using a cup horn on a Misonix sonicator 4000 set at level 6. The sonicated chromatin was treated with RNase A (Invitrogen) for 10 min at room temperature and decross-linked by boiling for 10 min. After an additional centrifugation for 10 min at 16,000 × *g*, DNA was extracted with phenol:chloroform:isoamyl alcohol (25:24:1) (Ambion) and purified through Qiagen PCR purification columns (Qiagen). DNA was resolved by agarose gel electrophoresis and 100–500-bp and 1,000–6,000-bp-sized DNA bands were excised and purified again through Qiagen PCR columns. The concentration and purity of the DNA was measured by A_{260} and A_{280} UV-Vis spectrometry on a NanoDrop ND-1000 spectrophotometer.

Quantitation by Real-Time PCR. For chromatin size-selection experiments samples were assayed by quantitative PCR (qPCR) to assess the enrichment of genomic regions in either the 100–500-bp or 1,000–6,000-bp chromatin samples. PCRs contained 2 μL DNA template, 3 μL of 3.3 mM primer pairs, and 5 μL of 2X EvaGreen reaction mix (FluoProbes). qPCR was performed on an Applied Biosystems 7500 Fast unit by using a 10-min soak at 95 °C, followed by 40 cycles of 5 s at 95 °C, 5 s at 55 °C and 20 s at 72 °C. Ct values were determined at threshold of 0.01. For each amplification product, the relative enrichment in the 100–500-bp sample versus the 1,000–6,000-bp sample was determined by using the formula relative enrichment = $a*1.9^{\mathrm{Ct(1{,}000–6{,}000bp)}-\mathrm{Ct(I100–500bp)}}$, where *a* was the constant associated with the ratio of the DNA concentration of the 100–500-bp and 1,000–6,000-bp samples and Ct is the threshold cycle.

Creation of ChIP-Seq and Reference DNA Sample Aggregation Plots. Uniquely mapped reads were extracted from the corresponding standard Eland output files for each factor/reference DNA type, signal maps created, and aggregation plots generated. A Python script was then used to create a signal map file in sgr format by using a sliding window approach. The size of the sliding window used for each dataset is shown in Table S1. For each list of features, coordinates were obtained and converted to build hg18 of the human genome, when necessary, by using UCSC's Liftover tool. A Perl script was then used to perform the aggregation. This script divides the region immediately upstream and downstream of a feature's start site into several bins. For each bin, all reads present for each nucleotide within the bin are summed and the average signal for the bin calculated. The bin scores corresponding to the same relative position for each feature are then averaged to produce a mean signal for each bin upstream and downstream of a feature's start position. These signals are finally normalized to the sum of the averages for the first four and last four bins for each feature type to produce the final ChIP–Seq aggregated signal values. For all factors/reference types in this analysis, we chose to use a total of 46 nonoverlapping bins (23 on each side of a feature's start position) of length 90 base pairs for all features other than CpG islands. Because of the varied lengths of CpG islands, we could not apply a standard bin size across all islands. Instead, we partitioned each CpG island into 35 equal-sized bins and aggregated over each fraction. We then extended this method to include additional bins of the same size regions flanking each CpG island.

Calculating Percent Feature Composition for Sono-Seq DNA and Pol II DNA. From the ranked lists described in the manuscript, all enriched regions comprised of fewer than 20 tags or possessing a fold-enrichment <5 were discarded. A subset of the peaks was then analyzed in a stepwise fashion by intersecting enriched regions against promoter regions of expressed Ensembl genes and against promoter regions of nonexpressed genes. Remaining enriched regions were deemed to lie outside of promoter regions and classified as "Other." For this analysis, promoter regions are again defined as within ±2.5 kb of a transcription start site (TSS). Data were added to the subset in 5% increments (i.e., iteration 1 would intersect enriched regions above the fifth percentile, iteration 2 above the tenth percentile, etc) until 100% of enriched regions were analyzed.

Creation of the Pol II and Sono-Seq Rank-Order Plot. To create the rank-order plot, enriched regions are ranked by tag count (minimum: 20) and enrichment factor (minimum: 5), and q-value (maximum: 0.05). This subset of enriched regions is deemed to be "high-quality enriched regions." For Pol II and Sono-Seq, data are added in a stepwise-fashion at 5% increments in decreasing order of enrichment to create an analysis set. During each iteration, enriched regions contained in the analysis set are intersected against 5′ ends of Ensembl genes. This process is repeated until all enriched regions are included in the analysis set. Regions intersecting promoter regions of Ensembl genes are further subdivided based on whether they overlap promoters of expressed or nonexpressed Ensembl genes. Enriched regions lying distal (±2.5 kb) to the TSS of an Ensembl gene are classified as "other."

Supplementary Results

Sono-Seq DNA Signals Show Little Increase over CTCF Regions Distal to Promoters. Because Sono-Seq DNA regions are often associated with promoter regions, particularly those of expressed genes, we examined whether the Sono-Seq signal is depleted over regions of closed chromatin or insulators. CCCTC-binding factor (CTCF) plays many roles in the human genome including behaving as a chromatin barrier, binding insulator elements to restrict transcriptional enhancers from activating unrelated promoters, and acting as an anchor for positioning neighboring nucleosomes (1). We analyzed the association of CTCF sites distal to promoters by removing sites found within ±2.5 kb of 5′ ends of known genes from a list of 127,172 CTCF sites obtained from Barski et al. (2). ChIP signals were then aggregated by using a random sample of 100,000 sites from the remaining 119,940 distal sites. We find that the Pol II signal is elevated over both proximal and distal CTCF sites, as well as Sono-Seq and MNase-digested DNA signals to a lesser degree (Fig. S9).

Highly Transcribed Regions Are Sonication-Sensitive Whereas Centromeric Repeats Are Sonication-Resistant. As a complementary approach for examining how different genomic regions are affected by the size of DNA fragments in sonicated samples, we performed qPCR analysis. DNA from sonicated chromatin was electrophoretically separated into small (100–500 bp) and large (1,000–6,000 bp) DNA fragments, and the amounts of DNA for various genomic regions were determined by qPCR analysis (Fig. S8). The results are presented as the small:large ratio, where a value of 1.0 indicates equimolar representation in the two samples. Several Pol II promoter regions are strongly overrepresented among the small DNA fragments (ratios ranging between 5–20). In contrast, the corresponding coding regions as well as two nonannotated, transcriptionally inactive regions of chromosome 21 are comparably represented in both the small and large DNA samples. Two of four Pol III genes, as well as the 18S and 28S regions of the ribosomal DNA genes, are also highly overrepresented among the small DNA fragments (ratios between 20–30). Interestingly, whereas a telomeric region is equally represented in the two samples, a centromeric region is extremely underrepresented among the small DNA fragments (ratio of 0.03). Thus, sonication of cross-linked chromatin samples occurs in a highly nonrandom fashion (variation among genomic regions occurs over a 200-fold range), with preferential fragmentation occurring at promoters and in highly transcribed regions and strong resistance to fragmentation occurring near centromeric repeats.

1. Fu Y, Sinha M, Peterson CL, Weng Z (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across- the human genome. *PLoS Genet* 4:e1000138.
2. Barski A, et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129:823–837.
3. Morin R, et al. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* 45:81–94.
4. Rozowsky JS, et al. (2007) The DART classification of unannotated transcription within the ENCODE regions: Associating transcription with known and novel loci. *Genome Res* 17:732–745.
5. Affymetrix and Cold Spring Harbor Laboratory ENCODE Transcriptome Projects (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* 457:1028–1032.
6. Cuddapah S, et al. (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* 19:24–32.
7. Heintzman ND, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459:108–112.

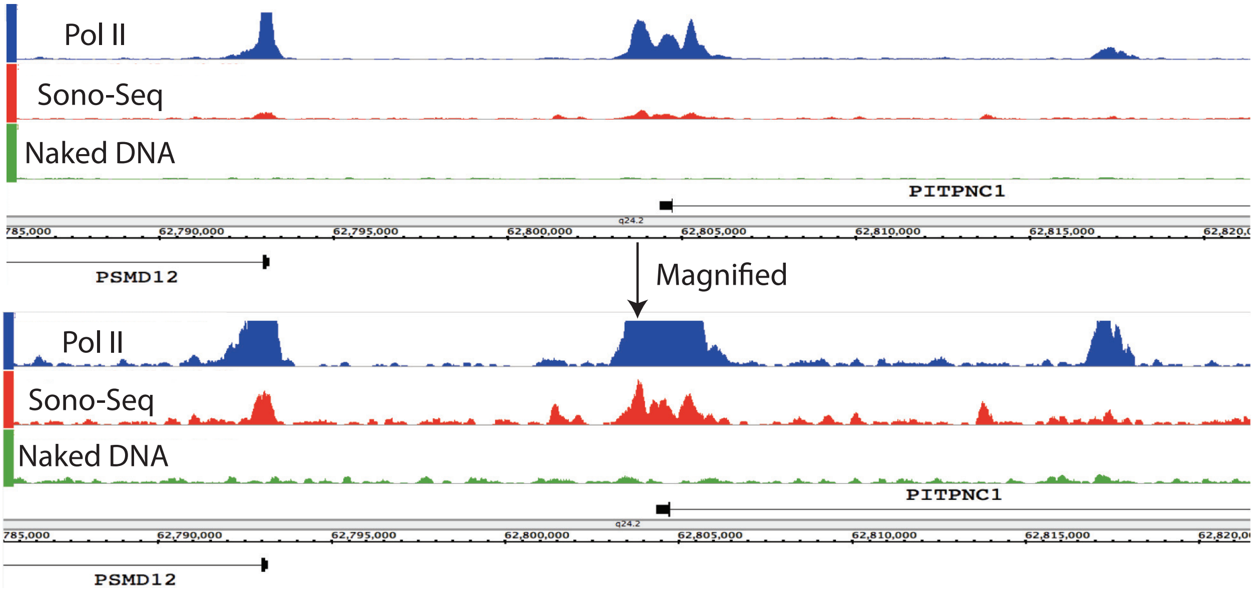

Fig. S1. Signal map showing Pol II ChIP DNA, Sono-Seq DNA from HeLa S3 cells and naked DNA. Signal maps are created with the IGB Browser (Affymetrix), and tracks are scaled based on the number of uniquely mapped reads obtained for each sample type. The magnified view shows the same region but uniformly alters the scale for all tracks to show additional peak detail. This figure shows signal levels between positions 62,755,000 and 62,821,500 of chromosome 17. Both *PSMD12* and *PITPNC1* are expressed in HeLa S3 based on RNA-Seq data (3).

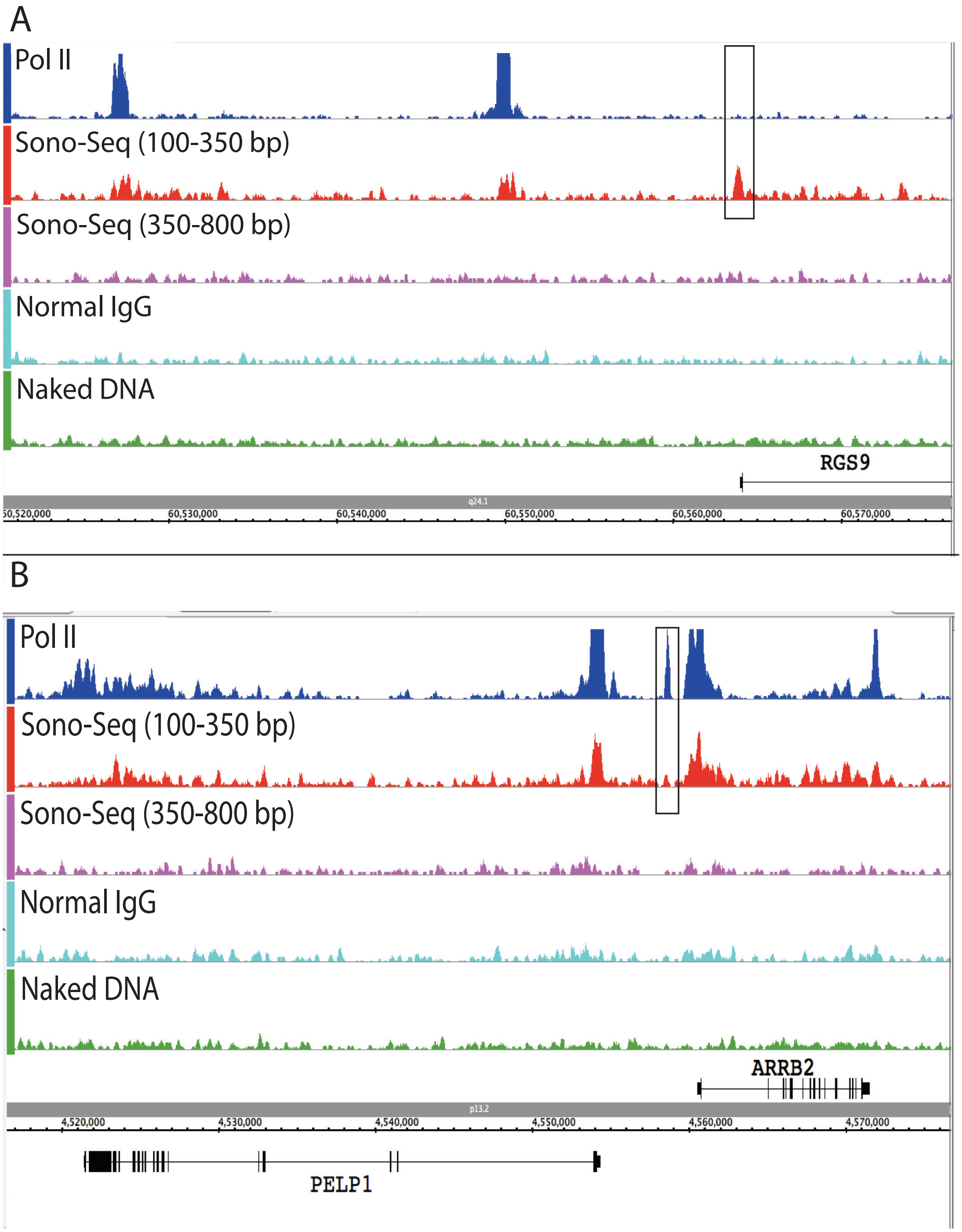

Fig. S2. Signal map showing Pol II ChIP DNA, Sono-Seq DNA (small and large fragment sizes), normal IgG, and naked DNA. All signals are in HeLa S3 cells. (*A*) Signal map showing Pol II ChIP DNA, Sono-Seq DNA (small and large fragment sizes), normal IgG, and naked DNA. All signals are in HeLa S3 cells. This figure shows signal levels between positions 60,520,000 and 60,576,000 of chromosome 17. Signal maps are created with the IGB Browser (Affymetrix), and tracks are scaled based on the number of uniquely mapped reads obtained for each sample type. *RGS9* is not expressed in HeLa S3 based on RNA-Seq data (3). The boxed region shows a large Sono-Seq peak in the absence of a corresponding peak in Pol II. (*B*) This figure shows signal levels between positions 4,517,000 and 4,576,000 of chromosome 17. Signal maps are created with the IGB Browser (Affymetrix), and tracks are scaled based on the number of uniquely mapped reads obtained for each sample type. Both *PELP1* and *ARRB2* are expressed in HeLa S3 based on RNA-Seq data (3). The boxed region illustrates a region where an auxiliary Pol II peak is observed without an accompanying Sono-Seq peak near the *ARRB2* promoter region, although other large Sono-Seq and Pol II peaks are located at this loci.

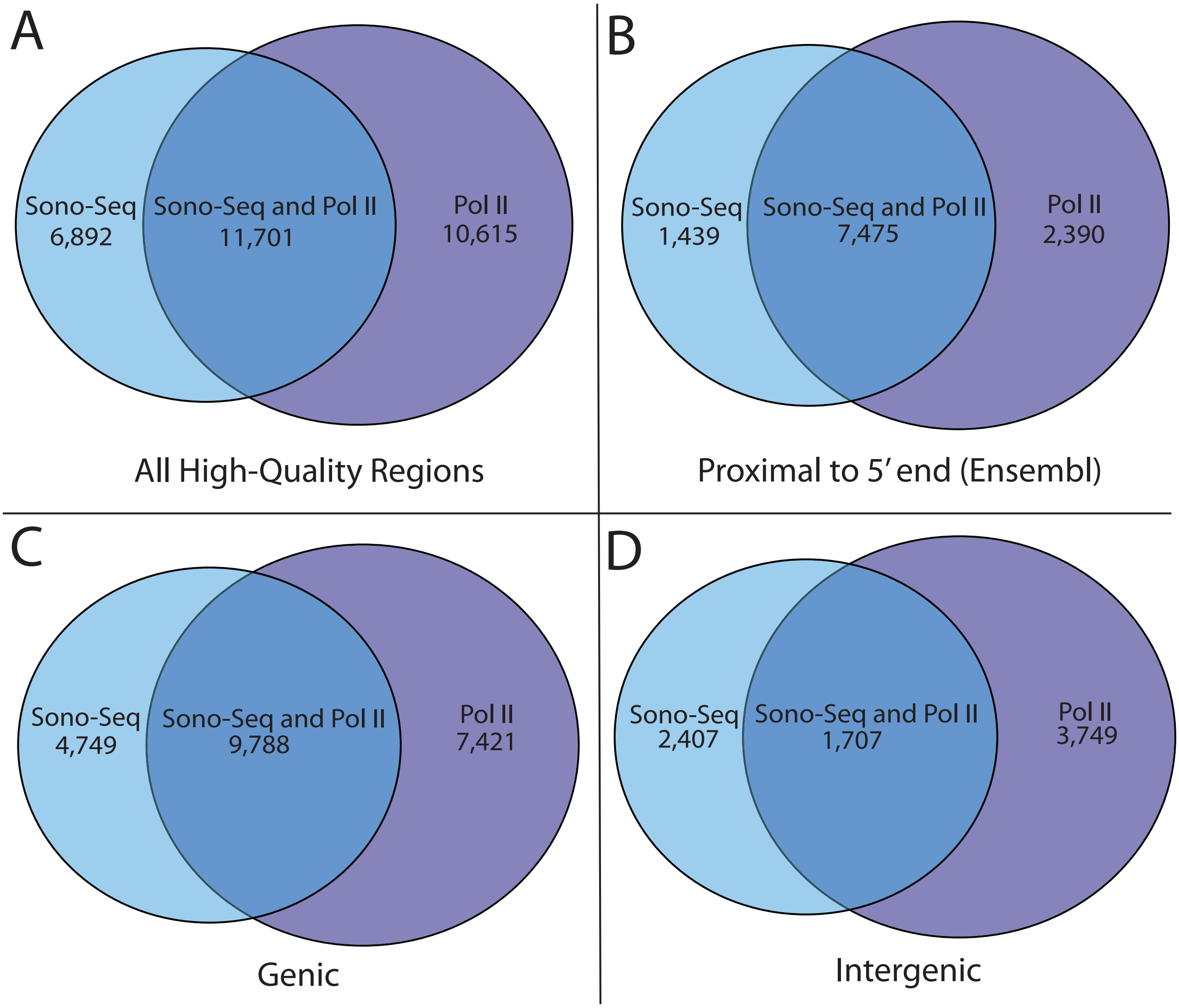

Fig. S3. Venn diagrams showing the number of Sono-Seq DNA and Pol II ChIP DNA regions, in which both datasets were collected from HeLa S3 cells. Diagrams show intersections of all highly enriched regions of Pol II and Sono-Seq DNA in the entire genome (*A*), proximal (within ±2.5 kb) to an Ensembl gene TSS (*B*), within or proximal to Ensembl genes (*C*), and distal to Ensembl genes (*D*). Intersections were performed by using the Active Region Comparer, which merged all peaks occurring within 500 bp before performing the intersections (4). Results are different from one-way intersections used in the paper, as Sono-Seq and Pol II hits do not necessarily exhibit a one-to-one relationship.

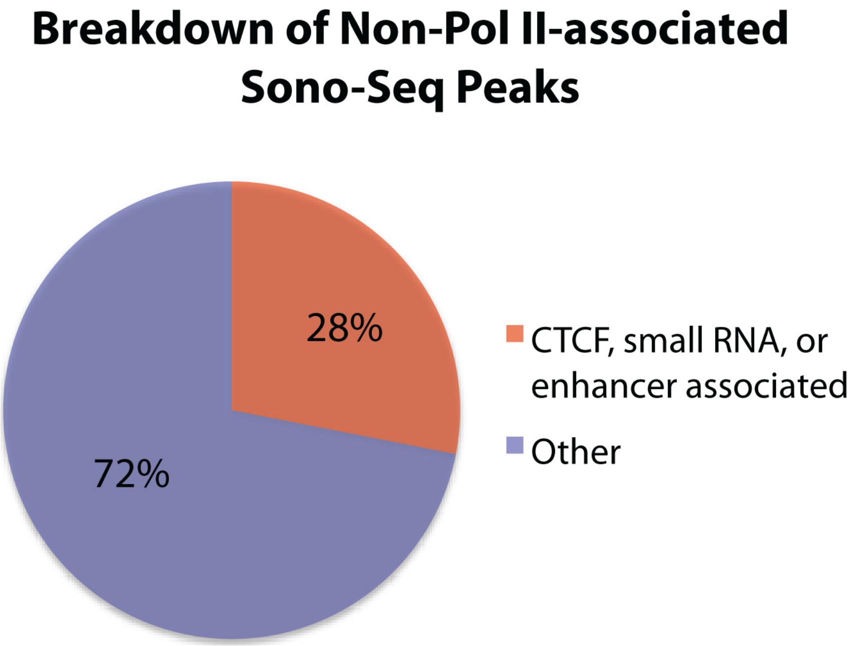

Fig. S4. Breakdown of non-Pol II-associated Sono-Seq peaks with respect to small RNAs, CTCF sites, and enhancers, where all three of these intersecting sets were HeLa-derived (5–7). Sono-Seq regions located >1 kb from a Pol II peak were intersected against CTCF and enhancer sites (within 200 bp) and small RNAs (within 2 kb). Twenty-eight percent of the non-Pol II-associated peaks fall within one of these categories.

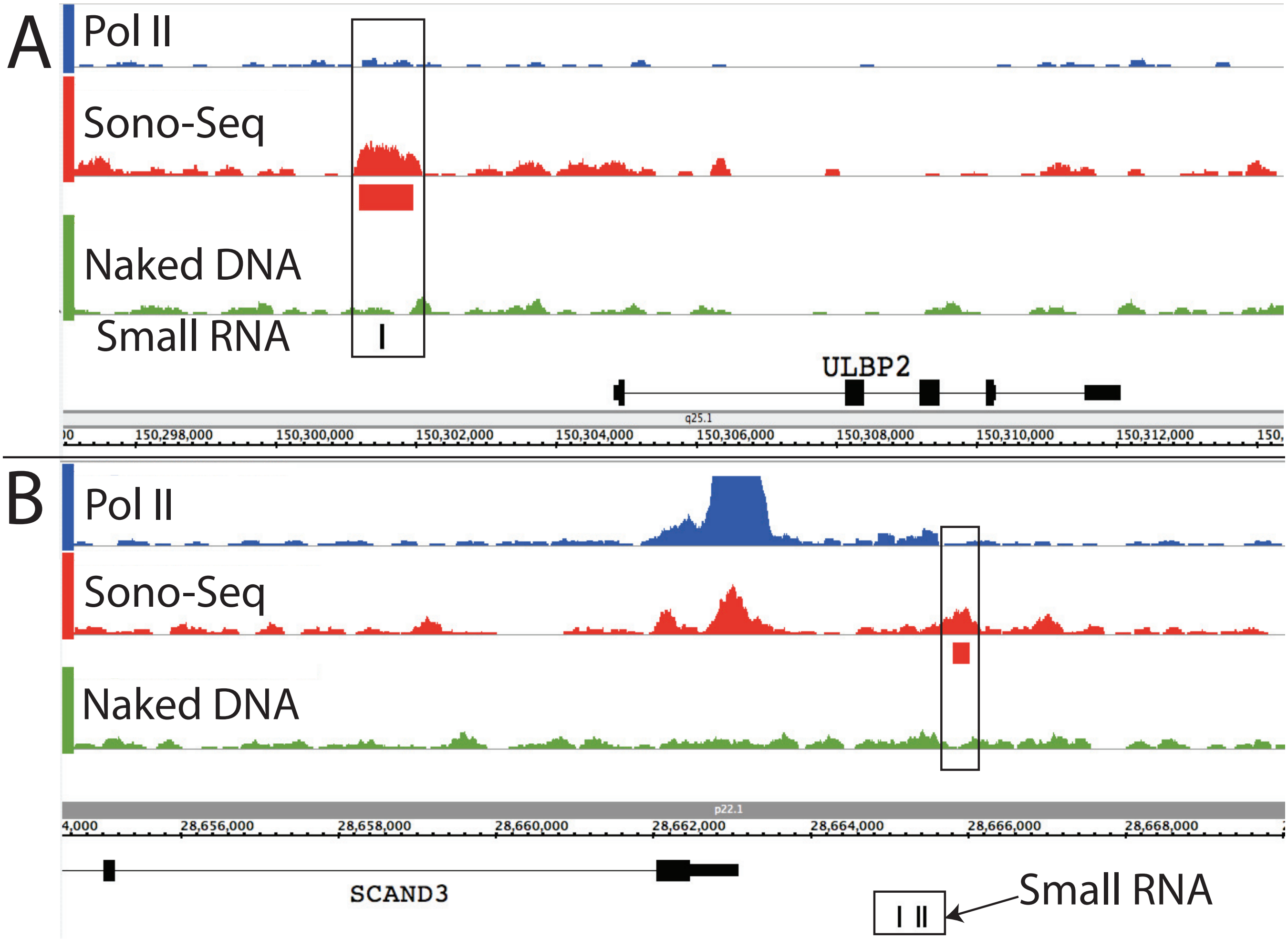

Fig. S5. Signal map showing Pol II ChIP DNA, Sono-Seq DNA (small fragment sizes) and naked DNA relative to several small (<200 nucleotides) RNAs from HeLa cells (5). Signal tracks are scaled based on the number of uniquely mapped reads obtained for each sample type. Boxed regions in *A* and *B* show Sono-Seq peaks in the absence of corresponding RNA Pol II peaks and where several small RNAs (small black rectangles near gene annotations) are within 1 kb of the Sono-Seq peaks. Neither *SCAND3* nor *ULBP2* are expressed in HeLa S3 cells based on RNA-Seq data (3). Regions shown are from 28,654,000–28,670,000 on chromosome 6 for *SCAND3* and from 150,297,200–150,314,000 on chromosome 6 for *ULBP2*. Signal maps are created with the IGB Browser (Affymetrix).

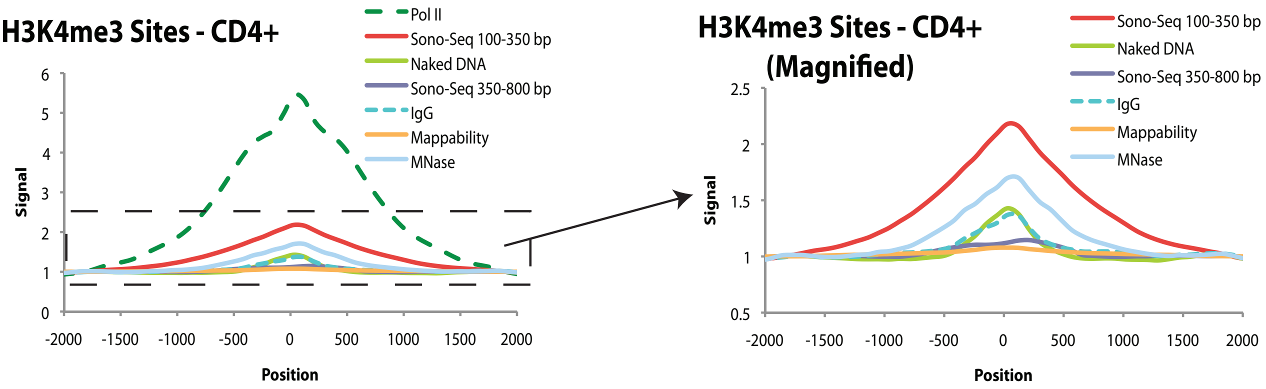

Fig. S6. Aggregation plot depicting average ChIP signal across a random sample of 100,000 H3K4me3 sites identified in CD4$^+$ cells. Frame *B* is a magnified view of the region enclosed by the dotted box in *A*. In *B*, Pol II is removed and the scale is altered to allow for better comparison between reference sample types. Vertical axis units are consistent between all plots. Horizontal axis units are given in terms of nucleotides from the feature start site.

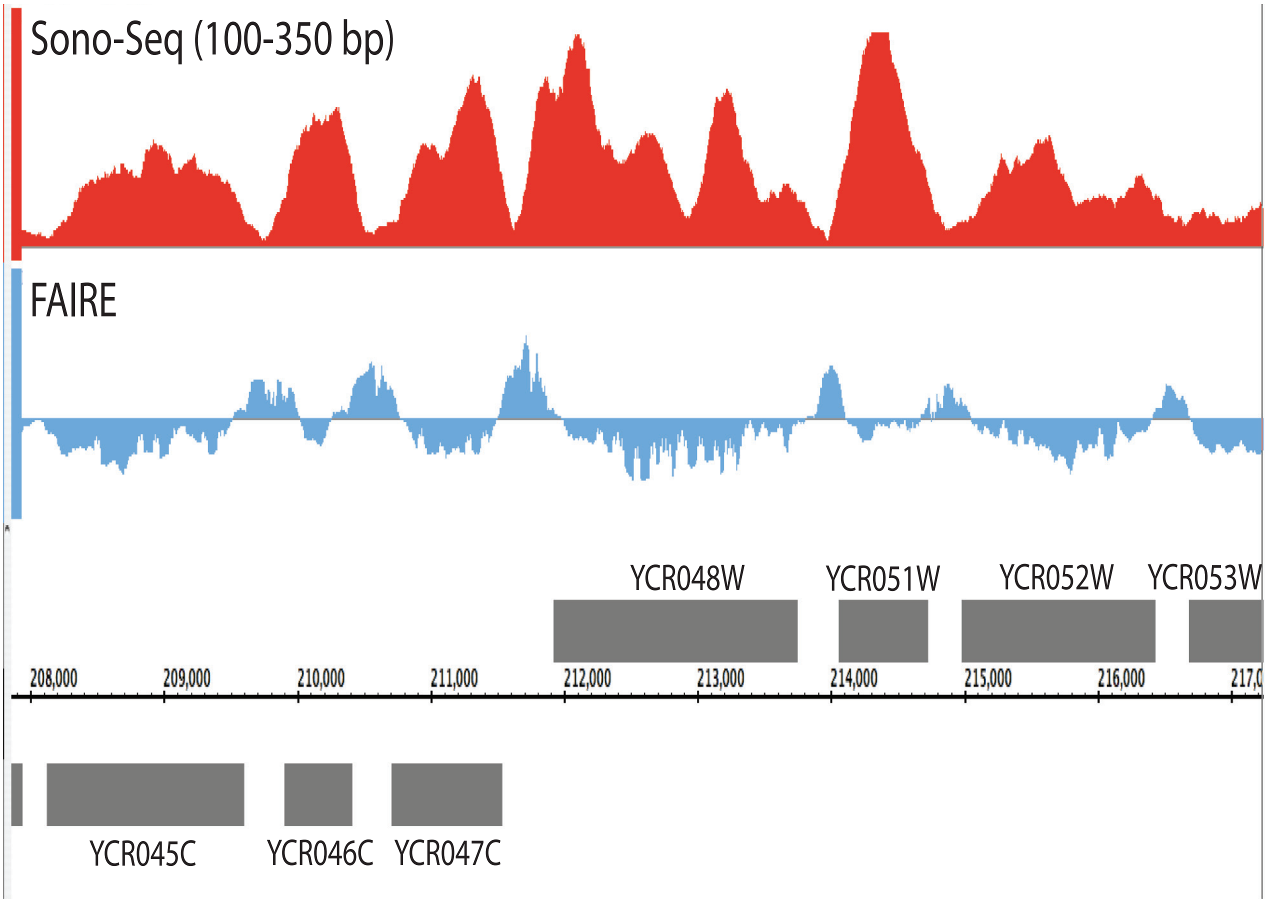

Fig. S7. Signal map showing Sono-Seq DNA and FAIRE signals in *Saccharomyces cerevisiae*. For FAIRE, signal levels above the axis are enriched, whereas levels beneath the axis are depressed. Regions enriched in Sono-Seq appear anticorrelated with FAIRE signal. This figure shows signal levels between positions 207,900–217,200 of *S. cerevisiae* chromosome 3. Signal maps are created with the IGB Browser (Affymetrix).

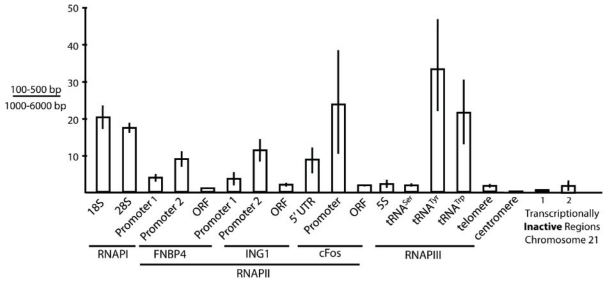

Fig. S8. Sonication efficiency varies greatly among genomic regions. DNA samples derived from 100–500 bp, and 1,000–6,000-bp chromatin fragments were analyzed by qPCR and normalized to the concentration of the input DNA. The amount of DNA in the 100–500 bp sample was 2.8 ± 0.5 higher than in the 1,000–6,000-bp sample.

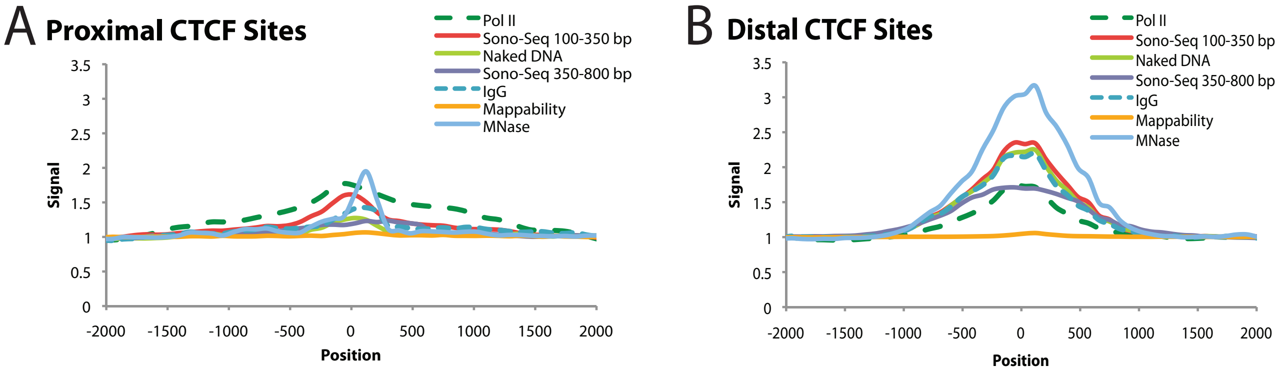

Fig. S9. Aggregation plots depicting average ChIP signal across 7,232 proximal and 100,000 distal CTCF sites. The same units and conventions as those from Fig. 3 in the main text are used for the axes.

Table S1. Sliding window sizes used to generate signal maps

Factor/reference sample type	Signal map (SGR) window size
Pol II	200
Sono-Seq (100–350 bp, HeLa S3)	200
Sono-Seq (350–800 bp, HeLa S3	575
Naked DNA	200
IgG	200
MNase	200