Determinants of nucleosome organization in primary human cells

Anton Valouev1, Steven M. Johnson2, Scott D. Boyd1, Cheryl L. Smith1, Andrew Z. Fire3,3 & Arend Sidow1,3

Nucleosomes are the basic packaging units of chromatin, modulating accessibility of regulatory proteins to DNA and thus influencing eukaryotic gene regulation. Elaborate chromatin remodelling mechanisms have evolved that govern nucleosome organization at promoters, regulatory elements, and other functional regions in the genome. Analyses of chromatin landscape have uncovered a variety of mechanisms, including DNA sequence preferences, that can influence nucleosome positions. To identify major determinants of nucleosome organization in the human genome, we used deep sequencing to map nucleosome positions in three primary human cell types and in vitro. A majority of the genome showed substantial flexibility of nucleosome positions, whereas a small fraction showed reproducibly positioned nucleosomes. Certain sites that position in vitro can anchor the formation of nucleosomal arrays that have cell type-specific spacing in vivo. Our results unveil an interplay of sequence-based nucleosome preferences and non-nucleosomal factors in determining nucleosome organization within mammalian cells.

Previous studies in model organisms as well as initial analyses in human cells have identified fundamental aspects of nucleosome organization. Here we focus on the dynamic relationships between sequence-based nucleosome preferences and chromatin regulatory function in primary human cells. We mapped tissue-specific and sequence-based nucleosome preferences and chromatin regulatory mechanisms have evolved that govern nucleosome organization at eukaryotic gene regulation. Elaborate chromatin remodelling mechanisms have evolved that allow high sequence coverage throughout the genome. Analyses of chromatin landscape have uncovered a variety of mechanisms, including DNA sequence preferences, that can influence nucleosome positions. To identify major determinants of nucleosome organization in the human genome, we used deep sequencing to map nucleosome positions in three primary human cell types and in vitro. A majority of the genome showed substantial flexibility of nucleosome positions, whereas a small fraction showed reproducibly positioned nucleosomes. Certain sites that position in vitro can anchor the formation of nucleosomal arrays that have cell type-specific spacing in vivo. Our results unveil an interplay of sequence-based nucleosome preferences and non-nucleosomal factors in determining nucleosome organization within mammalian cells.

We first focused on global patterns of nucleosome positioning and spacing by calculating fragment distograms and phasograms. Distograms (histograms of distances between mapped reads’ start positions aligning in opposing orientation, Supplementary Fig. 3a) reveal the average core fragment size as a peak if there are many sites in the genome that contain consistently positioned nucleosomes. A positioning signal that is strongly amplified by conditioning the analysis on sites with three or more read starts (reflecting a positioning preference; 3-pile subset), is present not only in vivo (Fig. 1a), but also in vitro (Fig. 1b), demonstrating that many genomic sites bear intrinsic, sequence-driven, positioning signals. Phasograms (histograms of distances between mapped reads’ start positions aligning in the same orientation, Supplementary Fig. 3b) reveal consistent spacing of positioned nucleosomes by exhibiting a wave-like pattern with a period that represents genome-average internucleosome spacing. In granulocytes, the wave peaks are 193 bp apart (Fig. 1c, adjusted $P$-value $<10^{-15}$), which, given a core fragment length of 147 bp, indicates an internucleosome linker length of 46 bp. By contrast, the phasograms of both types of T cells have spacing that is wider by 10 bp (Fig. 1d), equivalent to a 56 bp average linker length. These results are consistent with classical observations of varying nucleosome phases in different cell types. Linker length differences have been tied to differences in linker histone gene expression, which we found to be 2.4 times higher in T cells compared to granulocytes (84 reads per kilobase of mature transcript per million mapped reads (RPKM) vs 35 RPKM). The in vitro phasogram (Fig. 1e) reveals no detectable stereotypic spacing of positioned nucleosomes, demonstrating a lack of intrinsic phasing among DNA-encoded nucleosome positioning sites.

Using a positioning stringency metric (Methods; Supplementary Fig. 4) that quantifies the fraction of defined nucleosome positions within a given segment, we calculated the fraction of the genome that is occupied by preferentially positioned nucleosomes at different stringency thresholds. The maximum number of sites at which some positioning preference can be detected statistically is 120 million, covering just over 20% of the genome (Supplementary Fig. 5) at the low stringency of 23%. Thus, the majority of nucleosome positioning preferences is weak, and nucleosomes across the majority of the human genome are not preferentially positioned, either by sequence or by cellular function.

Next we focused on how transcription and chromatin functions affect nucleosome organization regionally. For each cell type, we generated deep RNA-seq data and binned genes into groups according to their expression levels. The average spacing of nucleosomes was greatest within silent genes (CD4+ T cells, 206 bp, Fig. 2a) and decreased by as much as 11 bp as the expression levels went up ($t$-statistic $P$-value $=6.5 \times 10^{-35}$). This suggests that transcription-induced cycles of nucleosome eviction and reoccupation cause denser packing of nucleosomes and slight reduction in nucleosome occupancy (Supplementary Fig. 6). On the basis of this result, we hypothesized that higher-order chromatin organization as implied by specific

1Department of Pathology, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, California 94305, USA. 2Department of Microbiology and Molecular Biology, Brigham Young University, 757 WIDB, Provo, Utah 84602-5253, USA. 3Department of Genetics, Stanford University School of Medicine, Poste Drive, Stanford, California 94305, USA.

©2011 Macmillan Publishers Limited. All rights reserved
 chromatin modifications might be associated with specific spacing patterns. Using previously published ChIP-seq data, we identified regions of enrichment for histone modifications that are found within heterochromatin (H3K27me3, H3K9me3)16, gene-body euchromatin (H3K27me3, H3K9me3)16, gene-body euchromatin

Figure 1 | Global parameters of cell-specific nucleosome phasing and positioning in human. a, In vivo granulocyte distogram (calculation explained in Supplementary Fig. 2a). x-axis represents the range of observed distances. y-axis represents frequencies of observed distances within 1-pile (blue) and 3-pile (red) subsets. 1-pile subset represents the entire data set, 3-pile subset represents a subset of sites containing three or more coincident read starts. b, Distogram of the in vitro reconstituted nucleosomes showing 1-pile and 3-pile subsets as in (a). c, In vivo granulocyte phasogram (calculation explained in Supplementary Fig. 3b), x-axis shows the range of observed phases. y-axis shows frequencies of corresponding phases. Phasograms of 1-pile, 3-pile and 5-pile subsets are plotted. Inset, linear fit to the positions of the phase peaks within 3-pile subsets (Slope = 193 bp). d, Phasograms of blood cell types. Inset, linear fits in CD4+ T cells (203 bp) and granulocytes (193 bp). e, Phasograms of 1-pile, 3-pile and 5-pile subsets in the in vitro data.

Figure 2 | Transcription and chromatin modification-dependent nucleosome spacing. a, Nucleosome spacing as a function of transcriptional activity. x-axis represents gene expression values binned according to RPKM values. Internucleosome spacing is plotted along the y-axis. Dashed lines represent genome-wide average spacing for each cell type. b, Nucleosome spacing within genomic regions marked by specific histone marks in CD4+ T cells. Bar height plots estimated nucleosome spacing for each histone modification. Bar colours differentiate chromatin types (euchromatin vs heterochromatin).

(hH4K20me1, H3K27me1)16, or euchromatin associated with promoters and enhancers (H3K4me1, H3K27ac, H3K36ac)17, and estimated spacing of nucleosomes for each of these epigenetic domains. We found that active promoter-associated domains contained the shortest spacing of 178–187 bp, followed by a larger spacing of 190–195 bp within the body of active genes, whereas heterochromatin spacing was largest at 205 bp (Fig. 2b). These results reveal striking heterogeneity in nucleosome organization across the genome that depends on global cellular identity, metabolic state, regional regulatory state, and local gene activity.

To characterize DNA signals responsible for consistent positioning of nucleosomes, we identified 0.3 million sites occupied in vitro by nucleosomes at high stringency (>0.5; Methods). The region occupied by the centre of the nucleosome (dyad) exhibits a significant increase in G/C usage (Poisson P-value < 10^-9; Fig. 3a) Flanking regions increase in A/T usage as the positioning strength increases (Fig. 3b). A subset of in vitro positioned nucleosomes (stringency >0.5) which are also strongly positioned in vivo (stringency >0.4) revealed increased A/T usage within the flanks (Fig. 3c) compared to in vitro-only positioning sites (Fig. 3a), which underscores the importance of flanking repelling elements for positioning in vivo. We term such elements with strong G/C cores and A/T flanks ‘container sites’ to emphasize the proposed positioning mechanism (Fig. 3d). This positioning signal is different from a 10-bp dinucleotide periodicity observed in populations of nucleosome core segments isolated from a variety of species8,9,10 and proposed to contribute to precise positioning and/or rotational setting of DNA on nucleosomes11 on a fine scale (Supplementary Fig. 7). G/C-rich signals are known to promote nucleosome occupancy10,20, whereas AA-rich sequences repel nucleosomes11, and our data demonstrate that precise arrangement of a core-length attractive segment flanked by repelling sequences can produce a strongly positioned nucleosome (Fig. 3d).
Dyad frequencies around container sites (Fig. 3c) show a strong peak of enrichment in vivo, confirming that DNA positions nucleosomes in vivo over these sites. Additionally, wave-like patterns emanate from these sites in vivo (but not in vitro), reflecting the nucleation of phased arrays by positioned cellular nucleosomes. Viewing these results in light of the nucleosome barrier model, which proposes that nucleosomes are packed into positioned and phased arrays against a chromatin barrier, we conclude that sequence-positioned nucleosome can initiate propagation of adjacent stereotypically positioned nucleosomes. Importantly, wave periods around container sites are shorter in granulocytes than in T cells, allowing tissue-specific variation in linker length (Fig. 1d) to alter placement of nucleosomes over distances of as much as 1 kilobase from an initial container site. Functional consequences of such rearrangements may include global shifts in regulatory properties that could contribute to distinct transcription factor accessibility profiles in different cell types.

The cellular environment can drive nucleosomes to sequences not intrinsically favourable to being occupied, as is evident in a genome-wide comparison of observed nucleosome coverage of all possible tetranucleotides between the granulocyte and the in vitro data (Fig. 4a). In vitro, nucleosome occupancy is strongly associated with AT/GC content, but this preference is abolished in vivo; the exception are C/G rich tetramers that contain CpG dinucleotides, which show a 30% reduction in apparent nucleosome occupancy despite having high core coverage in vitro. Consistent with this, CpG islands are fivefold depleted for observed nucleosome coverage in vivo (Fig. 4b). No such decrease is observed in the in vitro data set.

The decreased nucleosome occupancy of promoters could be due to promoter-related functions of mammalian CpG islands, similar to promoter-associated nucleosome-free regions observed in flies and yeast, which do not have CpG islands. We therefore analysed transcription-dependent nucleosome packaging around promoters. As in other organisms, promoters of active genes have a nucleosome-free region (NFR) of about 150 bp overlapping the transcriptional start site and arrays of well-positioned and phased nucleosomes that radiate from the NFR (Fig. 4c). A notable reduction in apparent nucleosome occupancy extends up to 1 kb into the gene body. We also observed consistent nucleosome coordinates in an independent data set of H3K4me3-bearing nucleosomes (Fig. 4d). Comparison of the nucleosome data (Fig. 4d) with binding patterns of RNA polymerase II (Fig. 4d) around active promoters indicates that phasing of positioned nucleosomes can be explained by packing of nucleosomes against Pol II stalled at the promoter, with Pol II potentially acting as the ‘barrier’. The set of inactive promoters, by contrast, exhibits neither a pronounced depletion of nucleosomes, nor a positioning and phasing signal (Fig. 4c). The transition of an inactive promoter to an active one is therefore likely to involve eviction of nucleosomes, coupled with positioning and phasing of nucleosomes neighbouring RNA Pol II (Fig. 4e). These results indicate that CpG-rich segments in mammalian promoters override intrinsic signals of high nucleosome affinity (Supplementary Fig. 8) to become active; this would be in contrast to fly and yeast, where AT-rich promoters may comprise intrinsic sequence signals that are particularly prone to nucleosome eviction.

To explore how regulatory factors interact with sequence signals to influence nucleosome organization outside of promoters, we focused on binding sites of the NRSF/REST repressor protein and the insulator protein CTCF. NRSF and CTCF sites are flanked by arrays of positioned nucleosomes (Fig. 4g and Supplementary Fig. 9), consistent with barrier-driven packing previously reported for CTCF. Both proteins occupy additional linker space, with NRSF taking up an extra...
37 bp and CTCF 74 bp. In agreement with sequence-based predictions, both CTCF and NRSF sites intrinsically encode high nucleosome occupancy as can be seen from the in vivo data (Fig. 4f and Supplementary Fig. 9), but this signal is overridden by occlusion of these sites from associating with nucleosomes. Additionally, phasing of nucleosomes around these regulatory sites is more compact in granulocytes compared to T cells (Supplementary Fig. 9), again exemplifying the importance of cellular parameters for placement of nucleosomes.

Our genome-wide, deep sequence data of nucleosome positions facilitated an initial characterization of the determinants of nucleosome organization in primary human cells. Spacing of nucleosomes differs between cell types and between distinct epigenetic domains in the same cell type, and is influenced by transcriptional activity. We confirm positioning preferences in regulatory elements such as promoters and chromatin regulator binding sites, but find that the majority of the human genome exhibits little if any detectable positioning. The influence of sequence on positioning of nucleosomes in vivo is modest but detectable. Despite DNA sequence being a potent driver of nucleosome organization at certain sites, the cellular environment often overrides sequence signals and can drive nucleosomes to occupy intrinsically unfavourable DNA elements or evict nucleosomes from intrinsically favourable sites. We find evidence for the barrier model of nucleosome distribution around the top 1,000 NRSF sites in vivo and in vitro. Distances from the NRSF binding sites are plotted along the x-axis, y-axis represents the normalized frequency of nucleosome dyads. Blue ovals depict hypothetical nucleosome positions. NRSF binding site is shown by grey lines pointing to nucleosome centres. Schematic depiction of nucleosome organization around promoters of repressed and active genes. Promoters of repressed genes do not have a well-defined nucleosome organization, whereas promoters of active genes have a nucleosome-free region (NFR, blue), RNA Pol II (orange) localized at the NFR boundary, and positioned nucleosomes (red) radiating from the NFR. Height of the ovals represents nucleosome frequency (inferred from c). Nucleosome distribution around promoters of active genes have a nucleosome-free region (NFR, blue), RNA Pol II (orange) localized at the NFR boundary, and positioned nucleosomes (red) radiating from the NFR. Height of the ovals represents nucleosome frequency (inferred from c). Nucleosome distribution around the top 1,000 NRSF sites in vivo and in vitro.

Methods Summary
Neutrophil granulocytes, CD4+ T cells, and CD8+ T cells were isolated from donor blood using Histopaque density gradients and Ig-coupled beads against blood cell surface markers (pan T and CD4+ microbeads, Miltenyi Biotec). Nucleosome cores

Figure 4 | Influence of gene regulatory function on nucleosome positioning. a, Comparison of sequence preferences of nucleosomes in vivo and in vitro. Normalized nucleosome core coverage in vivo (granulocytes) for a given sequence 4-mer is plotted along the x-axis. In vitro core coverage is plotted along the y-axis. Each data point on the plot represents one of the 256 possible 4-mers (coloured according to their G/C content). The diagonal line depicts the positions in the plot for which sequence-based preferences of nucleosomes would be the same in vivo and in vitro. b, Nucleosome core coverage over CpG islands in vivo and in vitro. x-axis represents coordinates within CpG islands (0–100%) and flanking upstream of the transcriptional start sites (TSS) (left) and downstream of the TSS (right). Normalized frequencies of nucleosome cores in vivo (upper plot) and in vitro (lower plot) are plotted along the y-axis. c, In vivo CD4+ T-cell nucleosome organization around promoters. x-axis represents distance from the TSS (blue arrow). Normalized frequencies of nucleosome dyads are plotted along the y-axis. Nucleosome arrangements within four gene groups are shown (not expressed 0–0.1 RPKM, low expressed 0.1–1 RPKM, moderately expressed 1–8 RPKM, highly expressed > 8 RPKM). Pie chart depicts distribution of RPKM values across gene groups. d, RNA Pol II binding signal within highly expressed genes (orange curve) and H3K4me3-marked nucleosome dyad frequency (green curve) within highly expressed genes (>8 RPKM). Nucleosomes show consistent positions, indicated by grey lines pointing to nucleosome centres. Schematic depiction of nucleosome organization around promoters of repressed and active genes. Promoters of repressed genes do not have a well-defined nucleosome organization, whereas promoters of active genes have a nucleosome-free region (NFR, blue), RNA Pol II (orange) localized at the NFR boundary, and positioned nucleosomes (red) radiating from the NFR. Height of the ovals represents nucleosome frequency (inferred from c). E, Nucleosome distribution around the top 1,000 NRSF sites in vivo and in vitro. Distances from the NRSF binding sites are plotted along the x-axis, y-axis represents the normalized frequency of nucleosome dyads. Blue ovals depict hypothetical nucleosome positions. NRSF binding site is shown by the green rectangle.
were prepared as described previously; cells were snap-frozen and crushed to release chromatin, followed by micrococcal nuclease treatment. *In vitro* nucleosomes were prepared by combining human genomic DNA with recombinantly-derived histone octamers at an average ratio of 1 octamer per 850 bp. Unbound DNA was then digested using micrococcal nuclease. After digestion, reactions were stopped with EDTA, samples were treated with proteinase K, and nucleosome-bound DNA was extracted with phenol-chloroform and precipitated with ethanol (Supplementary methods). Purified DNA was size-selected (120–180 bp) on agarose to obtain mononucleosome cores, followed by sequencing library construction. RNA was isolated by homogenizing purified cells in TRIzol, poly-A RNA was purified using a Qiagen Oligotex kit and RNA-seq libraries were constructed using a SOLiD Whole Transcription Analysis kit. All sequence data was obtained using the SOLiD 35 bp protocol and aligned using the SOLiD pipeline against the human hg18 reference genome. Downstream analyses were all conducted using custom scripts (Methods).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 August 2010; accepted 18 March 2011.

Published online 22 May 2011.

1. Mellor, J. The dynamics of chromatin remodeling at promoters. *Mol. Cell* **19**, 147–157 (2005).

2. Radman-Livaja, M. & Rando, O. J. Nucleosome positioning: how is it established, and why does it matter? *Dev. Biol.* **359**, 258–266 (2010).

3. Kaplan, N. et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009).

4. Berstein, B. E., Liu, C. L., Humphrey, E. L., Perlstein, E. O. & Schreiber, S. L. Global genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309**, 626–630 (2000).

5. Johnson, S. M., Tan, F. J., Mccullough, H. L., Riordan, D. P. & Fire, A. Z. Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res.* **16**, 1505–1516 (2006).

6. Valouev, A. et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).

7. Schones, D. E. et al. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).

8. Trifonov, E. N. & Sussman, J. L. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl Acad. Sci. USA* **77**, 3816–3820 (1980).

9. Kornberg, R. D. Structure of chromatin. *Ann. Rev. Biochem.* **46**, 931–954 (1977).

10. Widom, J. A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells. *Proc. Natl Acad. Sci. USA* **89**, 1095–1099 (1992).

11. Schlegel, R. A., Hase, K. R., Litwack, A. H. & Phelps, B. M. Nucleosome repeat lengths in the definitive erythroid series of the adult chicken. *Biochim. Biophys. Acta* **606**, 316–330 (1980).

12. Fan, Y. et al. Histone H1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation. *Cell* **29**, 1199–1212 (2005).

13. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).

14. Valouev, A. et al. Genome-wide analysis of transcription factor binding sites based on ChiP-Seq data. *Nature Methods* **5**, 829–834 (2008).

15. Barski, A. et al. High-resolution profiling of histone modifications in the human genome. *Cell* **129**, 823–837 (2007).

16. Wang, Z. et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genet.* **40**, 897–903 (2008).

17. Schwalke, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675 (1986).

18. Segal, E. et al. A genomic code for nucleosome positioning. *Nature* **422**, 772–778 (2006).

19. Hughes, A. & Rando, O. J. Chromatin ‘programming’ by sequence – is there more to the nucleosome code than %GC? *J. Biol.* **8**, 96 (2009).

20. Tillo, D. et al. High nucleosome occupancy is encoded at human regulatory sequences. *PLoS ONE* **5**, e12199 (2010).

21. Mavrich, T. N. et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18**, 1073–1085 (2008).

22. Mavrich, T. N. et al. Nucleosome organization in the Drosophila genome. *Nature* **453**, 358–362 (2008).

23. Lee, W. et al. A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet.* **39**, 1235–1244 (2007).

24. Gu, S. G. & Fire, A. Partitioning the C. elegans genome by nucleosome modification, occupancy, and positioning. *Chromosoma* **119**, 73–87 (2010).

25. Sasaki, S. et al. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science* **323**, 401–404 (2009).

26. Zhang, Y. et al. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nature Struct. Mol. Biol.* **16**, 847–852 (2009).

27. Field, Y. et al. Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nature Genet.* **41**, 438–445 (2009).

28. Chudakov, S. et al. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* **19**, 24–32 (2009).

29. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4**, e1000138 (2008).

30. Albert, I. et al. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* **446**, 572–576 (2007).

31. Wellinger, R. E. & Thoma, F. Nucleosome structure and positioning modulate nucleotide excision repair in the non-transcribed strand of an active gene. *EMBO J.* **16**, 5046–5056 (1997).

32. Scha, K. et al. Distributed probing of chromatin structure in vivo reveals pervasive chromatin accessibility for expressed and non-expressed genes during tissue differentiation in *C. elegans*. *BMC Genomics* **11**, 465 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Stanford Genetics/Pathology Sequencing Initiative. We thank G. Narlikar for help with *in vitro* experiments, Life Technologies, especially J. Briggs, for help with generating sequencing data, P. Lacroix for help with sequence alignment, S. Galil for valuable discussions, L. Gracey for critical reading of the manuscript, and members of the Sidow and Fire labs for valuable feedback and discussions. Work in the Fire lab was partially supported by NIGMS (R01GM37706). A.V. was partially supported by an ENCODE subcontract to A.S. (NHGRI U01HG004695). S.M.J. was partially supported by the Stanford Genome Training program (NHGRI T32HG00044).

Author Contributions A.V., S.M.J., A.S. and A.Z.F. designed the experiments. S.M.J., A.V., C.L.S., and S.O.B. performed the experiments. A.V. designed and carried out analyses with input from A.S., A.Z.F. and S.M.J.; A.V.; A.S. and A.Z.F. wrote the manuscript.

Author Information All sequence data were submitted to Sequence Read Archive (accession number GSE25133). Sites containing strongly positioned in vitro nucleosomes are available as a supplementary data file. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.S. (arend@stanford.edu) or A.Z.F. (alfre@stanford.edu).
 DNA and 0.80 have contributed an end-based signal at shorter fragment sizes. Lack of end-
Extraction Kit (Qiagen). DNA fragment lengths several-fold larger than nucleo-
100 V for 1 h. A smear of fragments with lengths from 850–2,000 bp (the bulk of
CD4 isolate mononucleosome core DNA from human cells, neutrophil granulocytes,
micrococcal nuclease (Roche) resuspended at 1 U
10 mM Hepes at pH 7.9 (final concentrations) and digesting with 20 units of
EDTA, followed by ethanol precipitation. The digested DNA was run on a 2.5% agarose
gel and the smear of DNA fragments from 135–225 bp was excised from the gel and purified using a QIAquick Gel Extraction Kit (Qiagen) as noted above.
End repair, linker ligation and library amplification. The ends of isolated
mononucleosome core DNAs (granulocytes, CD4 lymphocytes and CD8 lymph-
hytes), in vitro core DNAs and genomics control DNAs were processed by treating 0.3–0.5 μg of the DNA samples with T4 polynucleotide kinase (New England Biolabs) at 37 °C for 2.5 h followed by ethanol precipitation and sub-
sequent treatment with T4 DNA polymerase (New England Biolabs) in the presence
doNPs for 15 min at 12 °C. After purification using either a QIAquick Gel Extraction Kit as described above or a QIAquick PCR Purification Kit (Qiagen), linkering of previously annealed duplexes AF-SJ-47 (5′-OH-CCACTACGGCT CCGTTTCTCCTCTTCTGTGG-3′)/AF-SJ-48 (5′-P-ATCAC CGACTGCCCATAGAAGGAAAGGGGAGGTTAGTTT-3′) and AF-
SJ-49 (5′-OH-CTGGGCTCGCGCCCTATCCTCTTCTC-3′)/AF-SJ-50 (5′-P-AGAG AATGAGGAACCCGGGGCAGTT-3′) to the samples was accomplished with
T4 DNA ligase during a 6.5-h room-temperature incubation. The ligation reactions were
separated on a 2% agarose gel, and the relevant band isolated as described above. Amplification of the linked libraries was accomplished with 8 (granulocyte mononucleosome library), 10 (CD4 lymphocytes, CD8 lymphocytes and genomics control libraries) or 12 (in vitro library) cycles of polymerase chain reaction (PCR) using primers AF-SJ-47 (SOLiD P1 primer) and AF-SJ-49 (SOLiD P2 primer) with subsequent separation and purification using a 2% agarose gel and the QIAquick Gel Extraction Kit as described above. The number of cycles used in the PCR amplification were monitored and described as in ref. 25.
RNA-seq library preparation. Cells were homogenized in TRIzol using an 18G needle, followed by total RNA extraction using phenol-chloroform-isooamy alcohol. Poly-A RNA was isolated from total RNA using a Qiagen OligoTect kit according to the manufacturer’s instructions. The RNA-seq SOLiD sequencing library was built from 100 ng of poly-A RNA according to the manufacturer’s instructions (SOLiD whole transcriptome analysis kit).
DNA sequencing and mapping. Both nucleosome fragment and RNA-seq libraries were sequenced using the SOLiD DNA sequencing platform to produce 35 bp reads. All sequence data was mapped using SOLiD software pipeline against the human hg18 assembly using the first 25 bp from each read. This was done to maximize the number of the reference-mapped reads, as the higher error rate in read positions 26–35 of that version of the SOLiD chemistry prevented a substantial fraction of reads from mapping to the genome. For the genome-wide analysis we retained only unambiguously mapped reads.
Genome coverage by nucleosome cores was calculated as: core coverage = (number of mapped reads) × (147)/(genome size)

mRNA sequencing and data analysis. RNA-seq libraries were sequenced on the SOLiD platform to produce 35 bp reads and then the first 25 bp of each read were mapped to hg18 using the SOLiD mapping pipeline which resulted between 77 and 99 million mapped reads for each cell type. RPKM values were calculated as in ref. 14, with a modification that adjusted for transcript length, which was calculated according to the formula L = L – 50 × (E – 1), where L is the actual transcript length, and E is the number of exons in the gene. This modification is needed because of the lack of mappings across splice junctions.

Mathematical notations. Start counts: S_i \{ j \} (j) represent counts of 5′ coordinates of reads that map in + or – orientation at the j-th position of the reference strands. For example, if read maps to the interval [x,y) on the plus strand, then its 5′ coordinate is x, if it maps to – strand, then it’s y – 1. Indicator functions: I(condition) = 1 if condition is satisfied, 0 otherwise.

Nucleosome positioning stringency metric: nucleosome positioning stringency metric quantifies the fraction of nucleosomes covering a given position that are ‘well positioned’. The stringency at position j of the genome is calculated according to the formula:

S(j,w=30) = \sum_{i=-15}^{w+15} \frac{D(i,w=30)}{2w+1}

where D(i,w) is a kernel-smoothed dyad count calculated according to the formula:

D(i,w) = \sum_{j \in \mathbb{Z}} K(i-j, w) \delta(j)

©2011 Macmillan Publishers Limited. All rights reserved
where \( L \) is the size of a given chromosome, and \( K(u,w) \) is a smoothing kernel function of the form:

\[
K(u,w) = \left(1 - \frac{u}{w}\right)^2 I(|u| < w),
\]

and

\[
\int_{-\infty}^{\infty} (1 - u^2)^2 du = 1 / 1.09,
\]

and \( d(j) \) represents the number of dyads that occurs at the position \( j \):

\[
d(j) = s_j (j - 1/2) + s_{j - 1}(j + 1/2).
\]

Here \( L \) is the average library size (1 = 153 for in vivo data sets, 147 for in vitro data set). The core size is inferred from the 3-pile diagram peak position in the range of 100–200 bp.

The numerator of the stringency formula represents a kernel-smoothed count of nucleosome centres (dyads) at position \( i \) in the genome, whereas the denominator represents the count of nucleosome centres that infringe on the nucleosome centred at that position, which is inferred by integration of the dyad density estimate over an area of nucleosome infringement. The stringency is constructed in such a way that it would achieve a maximum of 1 if all nucleosomes were perfectly centred at that position (Supplementary Fig. 4). If two alternative, mutually exclusive, equally frequent nucleosome positions are observed in the data, then the stringency would be 0.5 or 50% for each alternative site (illustrated in Supplementary Fig. 4).

Application of the Kernel Density Estimation allowed obtaining smooth estimates of the stringency, which was useful for detection of nucleosome centres and robustly estimating the degree of positioning. We experimented with other smooth kernels and obtained highly consistent results. In principle, the kernel choice should not affect the results substantially as long as there is sufficient nucleosome core coverage (which follows from the convergence property of Kernel Density Estimation).

The kernel bandwidth \( w \) is an important parameter of the stringency formula and provides a means to control the smoothness of the stringency profile. Larger values of \( w \) provide higher smoothing but result in less accurate estimates of positioning centres, which is acceptable in cases of low core coverage. On the other hand, lower values of \( w \) result in less smoothing but more accurate estimation of the positioning centres, which is desirable in cases when nucleosome core coverage is high. We decided to use \( w = 30 \) in our calculation as it provided a sufficient amount smoothing across all of our data sets without sacrificing the sharpness of the positioning estimate.

Nucleosome positioning stringency was used for calculation of the fraction of the genome containing preferentially positioned nucleosomes (Supplementary Fig. 5). Positioned nucleosomes used in the container site analysis (Fig. 3a–c) were identified with the positioning stringency metric (as shown) and additional filters on nucleosome occupancy (in vivo occupancy > 30) to improve the statistical confidence of the positioning estimates.

**Nucleosome dyad coordinates.** Nucleosome dyads were inferred from 5′ coordinates of reads by shifting them by half the average nucleosome core size towards the 3′ end. The average nucleosome core size was estimated by a maximum value of the 3-pile diagram in a size range of 100–200 bp.

**Rotational positioning analysis.** We examined oligonucleotide preferences of rotational positioning of nucleosomes, which is associated with 10-bp patterning of short \( k \)-mers within nucleosome cores\(^{18,31}\). Plotting the frequencies of dyads around specific oligomers within the genome showed that the strongest patterning was exhibited by C-polymers (CC,CCC) with an exact helical period of 10.15 bp (Supplementary Fig. 7a, \( P \)-value < 2 \( \times \) 10\(^{-15} \)), indicating that they are important for rotational positioning. In vivo, such rotational preferences are much less pronounced (Supplementary Fig. 7b), indicating that cellular factors or conditions often override the sequence-encoded rotational settings.

**Characterization MNase cleavage patterns.** MNase is known to have sequence preferences that can affect both individual and bulk analyses of chromatin structure. Previous studies comparing MNase with alternative probes in model systems, both at specific loci (for example, ref. 32) and genome wide (for example, ref. 33), support the correspondence between the patterns of nucleosomes inferred from MNase digestion of chromatin and the in vivo chromatin landscapes. Nonetheless, it remained important to characterize the patterns of MNase activity in our data.

We investigated the extent of cleavage bias by MNase by examining sequence preferences within the cleavage sites, which correspond to 5′ end read positions in our data (Supplementary Figure 7a–c). Consistent with previous observations, MNase exhibits a pronounced but imperfect tendency to cleave at A or T nucleotides in naked DNA (Supplementary Fig. 11a). This same bias is detectable but, importantly, weaker when nucleosomes occupy the DNA, both in vivo and in vitro (1-pile subsets, top row b–c). Sites of more frequent cleavage (3-pile and 5-pile subsets, middle and bottom rows) revealed preferences that were virtually indistinguishable from the single-site preference.

The fact that the cleavage bias does not extend beyond 1–2 base pairs suggests that our analyses of nucleosome positioning preferences, which have substantially less than single-base resolution, should be robust to biases introduced by the MNase digestion. A case in point is the above-discussed rotational positioning analysis, whose resolution is on the order of 10 bp and which involves oligonucleotides that do not resemble the MNase cleavage site (Fig. 7a).

To investigate whether the sequence-driven nucleosome positioning element identified by the in vitro reconstitution experiment (Fig. 3) was a result of particularly pronounced MNase digestion bias within specific sites, we examined nucleotide preferences of nucleosome fragments overlapping sites of medium (>0.5) and high (>0.7) positioning stringency (Supplementary Fig. 11f, g). Preferences within these sites are identical to genome-wide preferences, ruling out the possibility that their positioning is an artefact of MNase digestion. In addition, we observe wave-like patterns in vivo around the CTCF site (Fig. 3e) consistent with previous observations that our analyses of nucleosome positioning preferences, which have substantially less than single-base resolution, should be robust to biases introduced by the MNase digestion. A case in point is the above-discussed rotational positioning analysis, whose resolution is on the order of 10 bp and which involves oligonucleotides that do not resemble the MNase cleavage site (Fig. 7a).

The lack of systematic differences in cleavage bias in our experimental data sets, in conjunction with the fact that naked DNA is affected most by the cleavage bias, suggests that our conclusions are robust to the use of MNase.

**Analyses of independent data sets.** We conducted additional analyses on independent data not generated by us to address any lingering concerns about biases or reproducibility. First, we sought to confirm independently that MNase cuts the linker DNA separating nucleosomes. In our data, CTCF sites (Supplementary Fig. 9) are surrounded by arrays of highly positioned and phased nucleosomes extending at least 1 kb in each direction. We investigated the frequency of cleavage by DNase I, a nuclease with preferences different from those of MNase, around CTCF sites within lymphoblastoid cell lines, using publicly available data from the ENCODE project. In agreement with our MNase results, we observed strongly phased peaks in the DNase I ENCODE data that align with linker DNA sites in our nucleosome data (Supplementary Fig. 12).

The estimates of spacing between nucleosomes as depicted in Fig. 1d are consistent between the two types of T cells we analysed. To assess whether these estimates were also reproducible by a different approach, we turned to a published data set that was generated for a different purpose, and by different means. Ref. 8 compared nucleosome distribution between resting and activated CD4+ T cells using MNase treatment of the cellular chromatin. We analysed spacing of nucleosomes in their data and obtained a highly concordant estimate of 202 and 203 bp (Supplementary Fig. 13) which is in agreement with the 203 bp spacing we see in our data (Fig. 1d).

34. Luger, K., Rechsteiner, T. J. & Richmond, T. J. Preparation of nucleosome core particle from recombinant histones. Methods Enzymol. 304, 3–19 (1999).