LETTER

doi:10.1038/nature10002

Determinants of nucleosome organization in primary human cells

Anton Valouev[1], Steven M. Johnson[2], Scott D. Boyd[1], Cheryl L. Smith[1], Andrew Z. Fire[1,3] & Arend Sidow[1,3]

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[1]. Analyses of chromatin landscape have uncovered a variety of mechanisms, including DNA sequence preferences, that can influence nucleosome positions[2–4]. 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[3–7] as well as initial analyses in human cells[8] 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 DNA-encoded nucleosome organization across granulocytes and two types of T cells ($CD4^+$ and $CD8^+$) isolated from the blood of a single human donor, by isolating cellular chromatin and treating it with micrococcal nuclease (MNase) followed by deep sequencing of the resulting nucleosome-protected fragments (Methods, Supplementary Fig. 1). To provide sufficient depth for both local and global analyses, we used high-throughput SOLiD technology, generating 584, 342 and 343 million mapped reads for granulocytes, $CD4^+$ and $CD8^+$ T cells, respectively. These are equivalent to 16–28× genome coverage by 147 bp nucleosome footprints (cores; see Methods). The depth of sequence was critical for our subsequent analysis: although shallower coverage can illuminate features of nucleosome positions through statistical analysis (for example, refs 6, 8), any definitive map and thus comparison of static and dynamic positioning requires high sequence coverage throughout the genome.

To provide complementary data on purely sequence-driven nucleosome positioning in the absence of cellular influences, we reconstituted genomic DNA *in vitro* with recombinantly derived histone octamers to produce *in vitro* nucleosomes (Methods, Supplementary Fig. 2), and generated over 669 million mapped reads, representing 32× core coverage of the genome. To identify primary nucleosome positioning sites in DNA, the reconstitution was performed under conditions of DNA excess (see methods). We also generated a control data set of 321 million mapped reads from MNase-digested naked DNA. In the population of granulocytes (our deepest *in vivo* data set), over 99.5% of the mappable genome is engaged by nucleosomes (Methods), and 50 percent of nucleosome-depleted bases occur in regions shorter than 160 bp.

We first focused on global patterns of nucleosome positioning and spacing by calculating fragment distograms and phasograms[6,7,9]. 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 $R^2 = 1$, 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[10,11]. Linker length differences have been tied to differences in linker histone gene expression[12,13], 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)[14] 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 $= 6.5 \times 10^{-34}$). 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

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

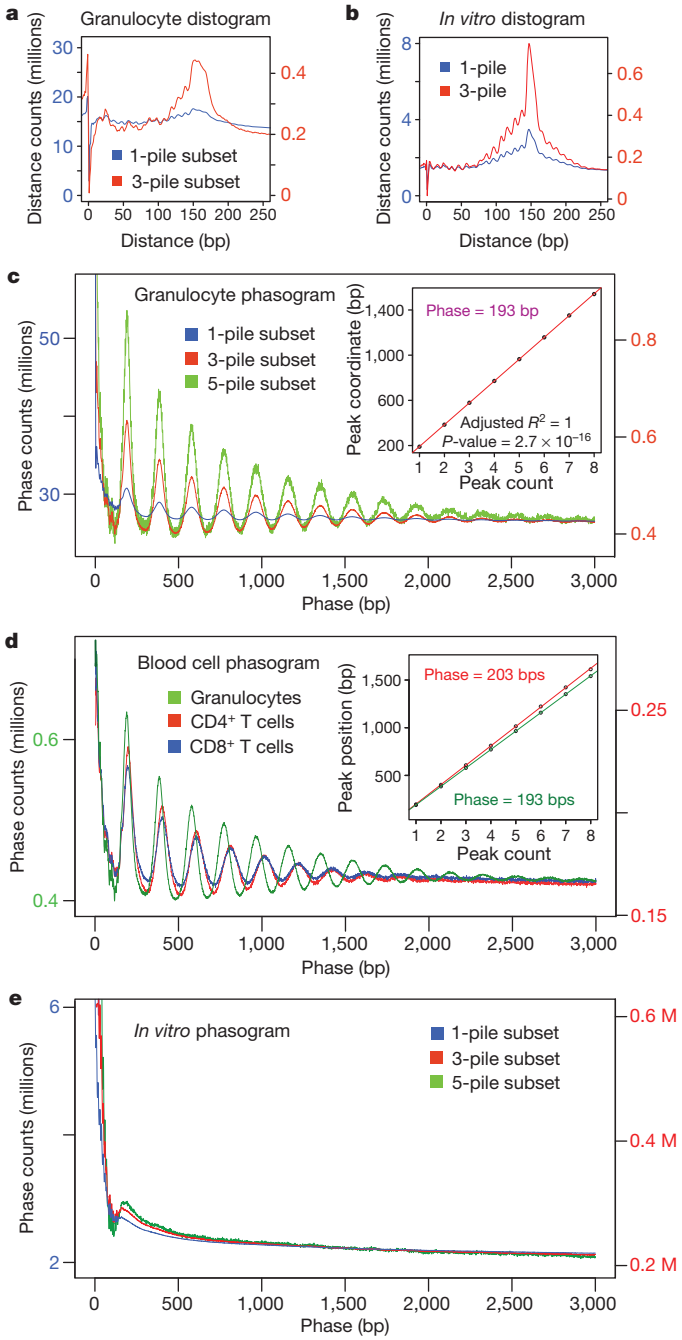


Figure 1 | Global parameters of cell-specific nucleosome phasing and positioning in human. **a**, *In vivo* granulocyte distogram (calculation explained in Supplementary Fig. 3a). *x*-axis represents the range of recorded distances. *y*-axis represents frequencies of observed distances within 1-pile (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 recorded 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.

chromatin modifications might be associated with specific spacing patterns. Using previously published ChIP-seq data, we identified regions of enrichment[15] for histone modifications that are found within heterochromatin (H3K27me3, H3K9me3)[16], gene-body euchromatin

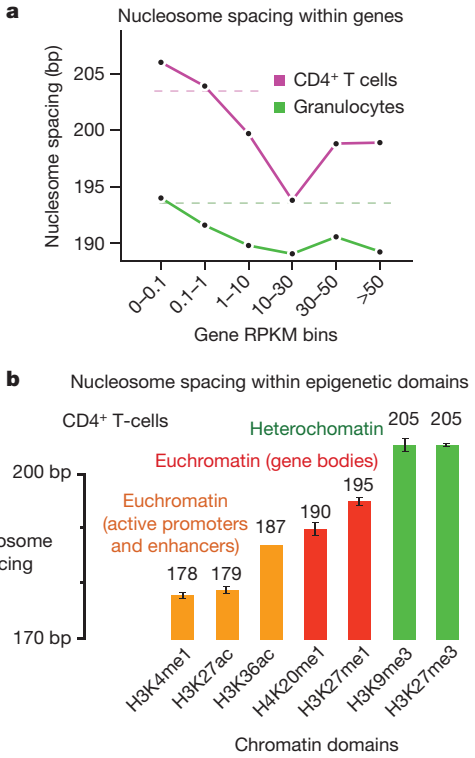


Figure 2 | Transcription and chromatin modification-dependent nucleosome spacing. **a**, Nucleosome spacing as a function of transcriptional activity. *x*- axis represents gene expression values binned according to RPKM values. Internucleosome spacing is plotted along the *y*-axis. Dashed lines represent 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).

(H4K20me1, 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 $< 10^{-100}$; 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 species[18,19] and proposed to contribute to precise positioning and/or rotational setting of DNA on nucleosomes[19] on a fine scale (Supplementary Fig. 7). G/C-rich signals are known to promote nucleosome occupancy[20,21], whereas AA-rich sequences repel nucleosomes[4], 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).

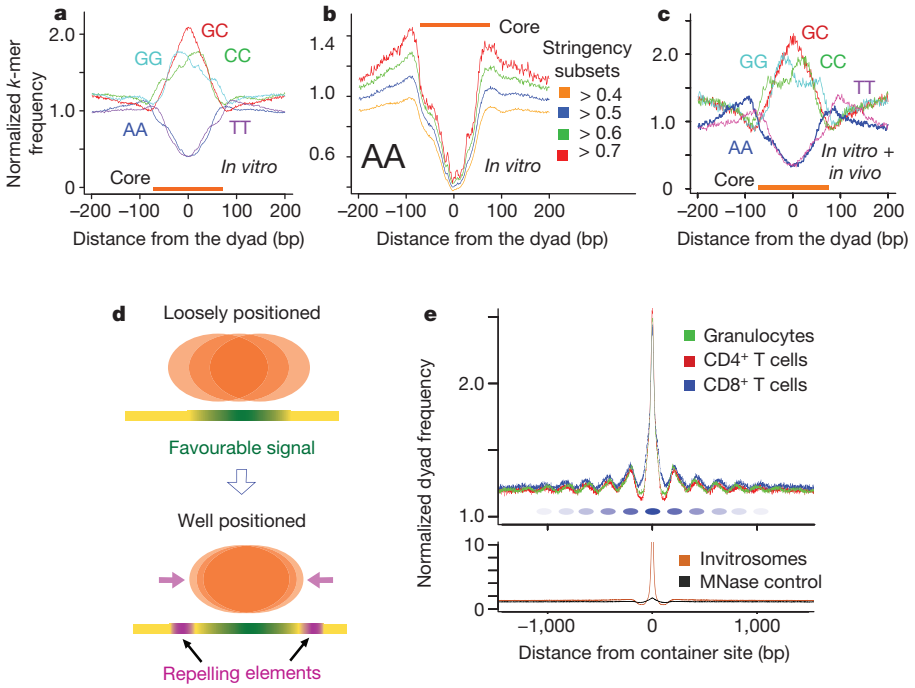


Figure 3 | Sequence signals that drive nucleosome positioning. **a**, Sequence signals within sites containing moderately positioned *in vitro* nucleosomes (stringency > 0.5). Distance from the positioned dyad to a given dinucleotide is plotted along the *x*-axis; *y*-axis represents frequency of a given *k*-mer divided by its genome-wide expectation. The 147-bp footprint of a nucleosome is indicated by an orange band. **b**, Changes in AA dinucleotide usage with increasing positioning stringency. *x* and *y* axes same as in (**a**). Curves of AA usage within the sites of increasingly positioned dyads are shown (stringency cutoffs of 0.4, 0.5, 0.6, 0.7). **c**, Sequence signals within sites containing *in-vitro*-positioned nucleosomes (stringency > 0.5) that also have high *in vivo* stringency (stringency > 0.4). *x* and *y* axes same as in (**a**). **d**, Schematic depiction of the container site positioning mechanism. The C/G-rich core area (green) favours occupancy, but does not precisely position the nucleosome (top). Adding flanking A/T-rich repelling elements (purple, bottom) restricts the position of the nucleosome. **e**, Nucleosome organization around container sites *in vivo* and *in vitro*. *x*-axis represents distances from the dyads to container sites (based on 300,000 container sites). Frequencies of nucleosome dyads around those sites are plotted along the *y*-axis. The upper plot shows distribution of *in vivo* dyads across $CD4^+$ cells, $CD8^+$ cells and granulocytes. The ovals depict hypothetical nucleosome positions across the site with colour intensities reflecting their positioning strength. The lower plot shows distribution of dyads *in vitro* and in MNase control.

Dyad frequencies around container sites (Fig. 3e) 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[22], 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 might 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[23] and yeast[5], which do not have CpG islands. We therefore analysed transcription-dependent nucleosome packaging around promoters. As in other organisms[23–27], 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[16] (Fig. 4d). Comparison of the nucleosome data (Fig. 4d) with binding patterns of RNA polymerase II[16] (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[28].

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[15] and the insulator protein CTCF. NRSF and CTCF sites are flanked by arrays of positioned nucleosomes (Fig. 4f and Supplementary Fig. 9), consistent with barrier-driven packing previously reported for CTCF[29,30]. Both proteins occupy additional linker space, with NRSF taking up an extra

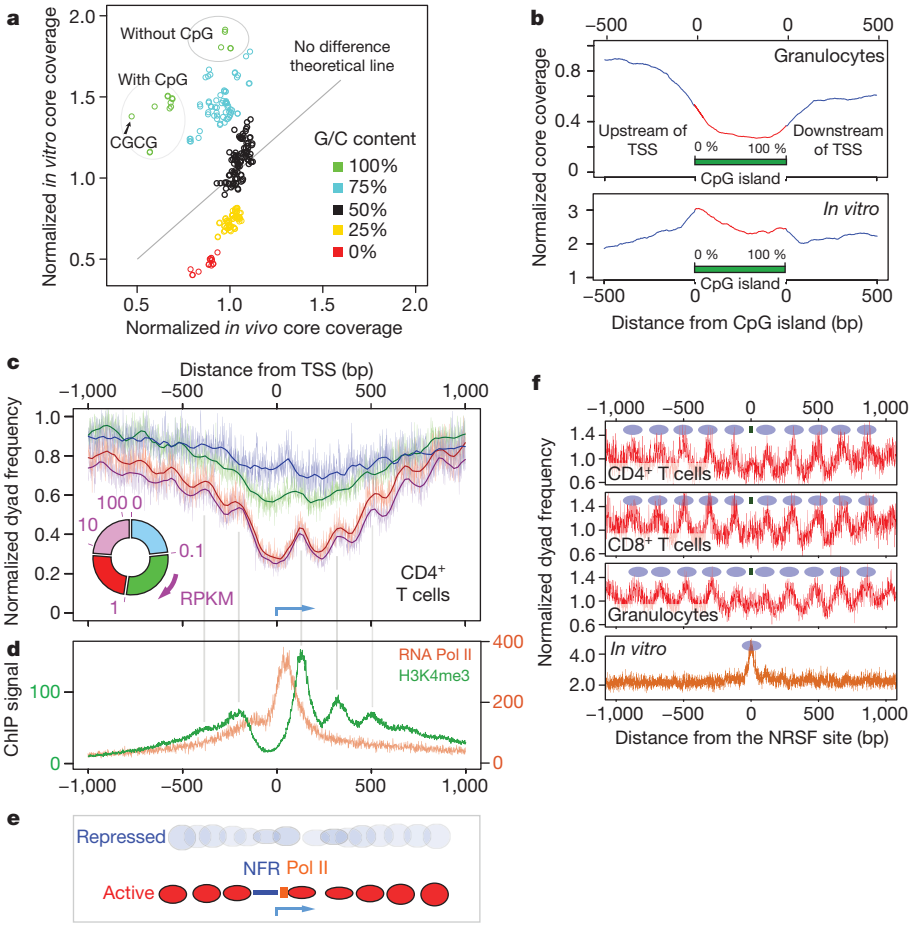


Figure 4 | Influence of gene regulatory function on nucleosome positioning. **a**, Comparison of sequence preferences of nucleosomes *in vivo* and *in vitro*. Normalized nucleosome core coverage *in vivo* (granulocytes) for a given sequence 4-mer is plotted along the *x*-axis. *In vitro* core coverage 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. **e**, 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**). **f**, 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.

37 bp and CTCF 74 bp. In agreement with sequence-based predictions[21], both CTCF and NRSF sites intrinsically encode high nucleosome occupancy as can be seen from the *in vitro* data (Fig. 4f and Supplementary Fig. 9), but this signal is overridden *in vivo* 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 for nucleosome organization, and that barriers can be nucleosomes (positioned by container sites), RNA polymerase II (stalled at the promoter), or sequence-specific regulatory factors. Our nucleosome maps should be useful for investigating how nucleosome organization affects gene regulation and vice versa, as well as for pinpointing the mechanisms driving regional heterogeneity of nucleosome spacing.

METHODS SUMMARY

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

were prepared as described previously[7]; 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 Transcriptome 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; corrected 23 June 2011 (see full-text HTML version for details).

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.* **339,** 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 nucleosome occupancy in yeast. *Genome Biol.* **5,** R62 (2004).
5. Yuan, G.-C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309,** 626–630 (2005).
6. 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).
7. 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).
8. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–898 (2008).
9. Trifonov, E. N. & Sussman, J. L. The pitch of chromatin DNA is reflected in it its nucleotide sequence. *Proc. Natl Acad. Sci. USA* **77,** 3816–3820 (1980).
10. Kornberg, R. D. Structure of chromatin. *Ann. Rev. Biochem.* **46,** 931–954 (1977).
11. 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).
12. Schlegel, R. A., Haye, 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).
13. 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).
14. 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).
15. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods* **5,** 829–834 (2008).
16. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
17. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genet.* **40,** 897–903 (2008).
18. Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191,** 659–675 (1986).
19. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442,** 772–778 (2006).
20. Hughes, A. & Rando, O. J. Chromatin 'programming' by sequence - is there more to the nucleosome code than %GC? *J. Biol.* **8,** 96 (2009).
21. Tillo, D. *et al.* High nucleosome occupancy is encoded at human regulatory sequences. *PLoS ONE* **5,** e9129 (2010).
22. Mavrich, T. N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18,** 1073–1083 (2008).
23. Mavrich, T. N. *et al.* Nucleosome organization in the *Drosophila* genome. *Nature* **453,** 358–362 (2008).
24. Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet.* **39,** 1235–1244 (2007).
25. Gu, S. G. & Fire, A. Partitioning the *C. elegans* genome by nucleosome modification, occupancy, and positioning. *Chromosoma* **119,** 73–87 (2010).
26. Sasaki, S. *et al.* Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science* **323,** 401–404 (2009).
27. 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).
28. Field, Y. *et al.* Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nature Genet.* **41,** 438–445 (2009).
29. Chuddapah, 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).
30. 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).
31. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446,** 572–576 (2007).
32. 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).
33. Sha, 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. Lacroute for help with sequence alignment, S. Galli 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.D.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. (afire@stanford.edu).

METHODS

Cell purification. Blood samples were obtained from the Stanford Blood Center. Samples were screened for any medical history of malignancy or signs of infectious disease, and tested for serologic evidence of viral infections to ensure that samples came from healthy donors. The Stanford Blood Center procedures used for the cells in this study are the same as those used for transfusion of patients and are routinely inspected by the FDA, the American Association of Blood Banks, and the College of American Pathologists. The blood for the experiments was processed immediately upon donation to avoid any change in quality as a result of sample storage.

Buffy coat (36 ml) from a blood donor was diluted in PBS to a total volume of 200 ml. The cells were layered on a Histopaque gradient with densities 1.119 and 1.077 g ml^{-1} according to manufacturer's instructions (Sigma HISTOPAQUE-1119 and 1077) and separated by centrifugation to yield granulocytes and mononuclear fractions. T cells were isolated from mononuclear cells using a Pan T isolation Kit (Miltenyi Biotec), followed by separation into CD4$^+$ and CD8$^+$ fractions using CD4$^+$ microbeads (Miltenyi Biotec).

Isolation of mononucleosome core DNA fragments from human cells. To isolate mononucleosome core DNA from human cells, neutrophil granulocytes, CD4$^+$ lymphocytes and CD8$^+$ lymphocytes were flash-frozen in liquid nitrogen in 0.34 M sucrose Buffer A and ground, digested on different days, and isolated as described in ref 7. By carrying out an MNase digestion in a short time frame (12 min at 16 °C) following grinding of the samples, we minimize the potential for nucleosome mobility. To maximize uniformity of representation, we use an extraction protocol after MNase digestion that does not rely on solubility of the individual core particles; this resulted in recovery of the bulk of input DNA as a mono-nucleosome band (Supplementary Fig. 10), limiting the degree to which the protocol might select for specific (for example, accessible) chromosomal regions. The mean nucleosome core length obtained for analysis (153 nucleotides) indicates an average overhang of 3 nucleotides on each side of individual cores (147 bp + 2 × 3 bp = 153 bp). Subsequent analyses assign nucleosome positions accounting for this mean overhang and making use of the ability to define location based on interpolation between values calculated from plus-oriented and minus-oriented reads (see below).

Preparation of *in vitro* nucleosomes. Naked genomic DNA isolated from neutrophil granulocytes from our *in vivo* studies was sheared by sonication using a Covaris sonicator and separated on a 1% UltraPure Agarose (Invitrogen) gel run at 100 V for 1 h. A smear of fragments with lengths from 850–2,000 bp (the bulk of the sheared DNA) was isolated and extracted from the gel using the QIAquick Gel Extraction Kit (Qiagen). DNA fragment lengths several-fold larger than nucleosome cores were chosen for this analysis to minimize any end-effects that could have contributed an end-based signal at shorter fragment sizes. Lack of end-preference in the reconstitutions was then confirmed under the conditions of these assays using a series of defined restriction fragments as templates for assembly (S.M.J. and A.F., results not shown).

The ends of the sheared DNA fragments were repaired as described below and then were assembled with recombinant *Xenopus* histones into nucleosomes as described previously[34] at a 1.1:1 molar ratio of DNA to histone octamer such that on average one nucleosome would occupy 850 bp of DNA. Specifically, 4.9 μg of DNA and 0.80 μg of octamer were reconstituted in a total volume of 200 μl.

The ref. 34 conditions (in which DNA was not limiting) were used for our analysis in order to focus specifically on primary sequence effects on nucleosome position. We note that two recent studies in yeast use somewhat different conditions, with a higher ratio of nucleosomes to DNA[3,27]. Assays at high nucleosome:DNA ratio provide a composite readout reflecting both (1) primary preferences of nucleosomes (caused by sequence signals within the nucleosome-bound DNA) and (2) secondary effects due to steric hindrance as a result of dense packing of nucleosomes. Although such data are certainly valuable in modelling chromosome dynamics, the goals of our study (definition of individual sequence elements that can initiate positioning) were best served with the lower nucleosome:DNA assay conditions[34].

Isolation of *in vitro* nucleosome core DNA fragments. *In vitro* nucleosome core DNAs were isolated by diluting 70 μl of the reconstituted *in vitro* nucleosome into a total volume of 200 μl containing 5 mM $MgCl_2$, 5 mM $CaCl_2$, 70 mM KCl and 10 mM Hepes at pH 7.9 (final concentrations) and digesting with 20 units of micrococcal nuclease (Roche) resuspended at 1 U μl^{-1} for 15 min at room temperature. The digestion was stopped by adding an equal volume of 3% SDS, 100 mM EDTA and 50 mM Tris. Octamer proteins were removed by treating with one-tenth volume proteinase K (20 mg ml^{-1} in TE at pH 7.4) for 30 min at 50 °C followed by phenol/chloroform and chloroform extractions and ethanol precipitation. This procedure was repeated twice to process the entire *in vitro* sample, and then *in vitro* DNA cores were isolated on a 2% UltraPure Agarose (Invitrogen) gel run at 100 V for 1 h followed by DNA extraction from the gel using a QIAquick Gel Extraction Kit (Qiagen) following the standard protocol with the exception of allowing the isolated gel sample to incubate in Buffer QG at room temperature until dissolved.

Genomic MNase digest control library preparation. For control libraries, genomic DNA (20 μg) from human neutrophil granulocytes in 0.34 M sucrose Buffer A with 1× BSA (New England Biolabs) and 1 mM $CaCl_2$ was digested with 200 units of micrococcal nuclease (Roche) (0.4 U μl^{-1} final concentration) in a total volume of 500 μl for 10 min at 23 °C. The digestion was stopped by addition of 10 μl 0.5 M 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$^+$ lymphocytes), *in vitro* core DNAs and genomic 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 subsequent treatment with T4 DNA polymerase (New England Biolabs) in the presence of dNTPs 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-CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT-3′)/AF-SJ-48 (5′-P-ATCACCGACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTGGTT-3′) and AF-SJ-49 (5′-OH-CTGCCCCGGGTTCCTCATTCTCT-3′)/AF-SJ-50 (5′-P-AGAGAATGAGGAACCCGGGGCAGTT-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 linkered libraries was accomplished with 8 (granulocyte mononucleosome library), 10 (CD4$^+$ lymphocytes, CD8$^+$ lymphocytes and genomic control libraries) or 12 (*in vitro* library) cycles of polymerase chair 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 selected as described in ref. 25.

RNA-seq library preparation. Cells were homogenized in TRIzol using an 18G needle, followed by total RNA extraction using phenol-chloroform-isoamyl alcohol. Poly-A RNA was isolated from total RNA using a Qiagen Oligotex 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 $L' = L - 50 \times (E - 1)$, 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: $S_{+/-}(j)$ 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 + 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 i of the genome is calculated according to the formula:

$$S(i, w=30) = \frac{D(i, w=30)}{\sum_{i-150}^{i+150} \frac{1.09}{w} D(j, w=30)},$$

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

$$D(i,w) = \sum_{j=0}^{L} K(i-j,w) \cdot d(j),$$

where L is the size of a given chromosome, and $K(u,w)$ is a smoothing kernel function of the form:

$$K(u,w) = (1-(u/w)^2)^3 I\{|u| < w\},$$

and

$$\int_{-1}^{1} (1-u^2)^3 du = 1/1.09,$$

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

$$d(j) = s_+(j-l/2) + s_-(j+l/2).$$

Here l is the average library size ($l = 153$ for *in vivo* data sets, 147 for *in vitro* data set). The core size is inferred from the 3-pile distogram 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 vitro* 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 cores size was estimated by a maximum value of the 3-pile distogram 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 $< 2 \times 10^{-16}$), 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–e). 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–e). 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 (Supplementary 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 these sites (Fig. 3e) consistent with existence of a chromatin barrier in the form of a well-positioned nucleosome.

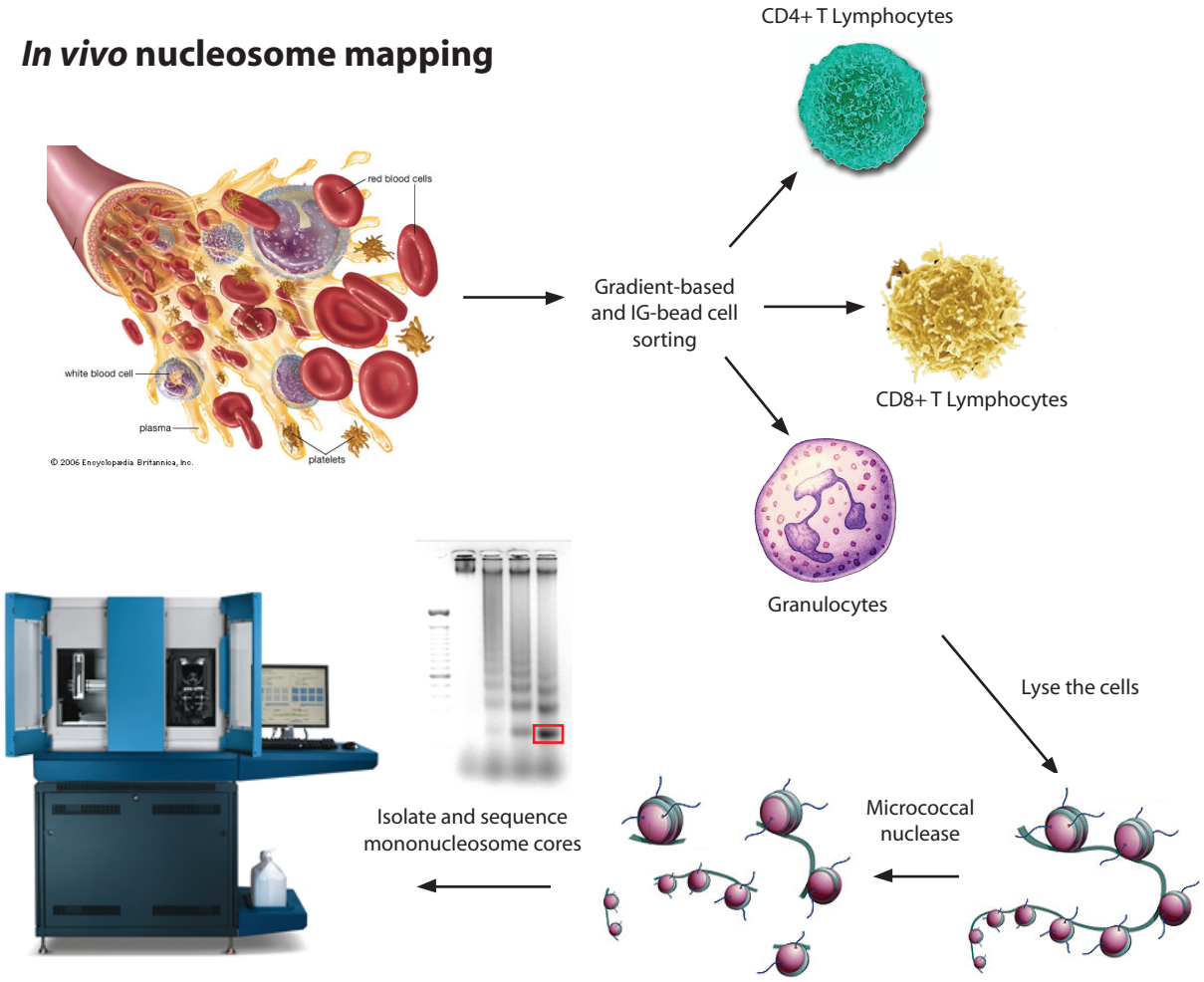
The lack of systematic differences in cleavage bias in our experimental data sets, in conjunction with the fact that naked DNA is affected most by the cleavage bias, suggests that our conclusions are robust to the use of MNase.

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).

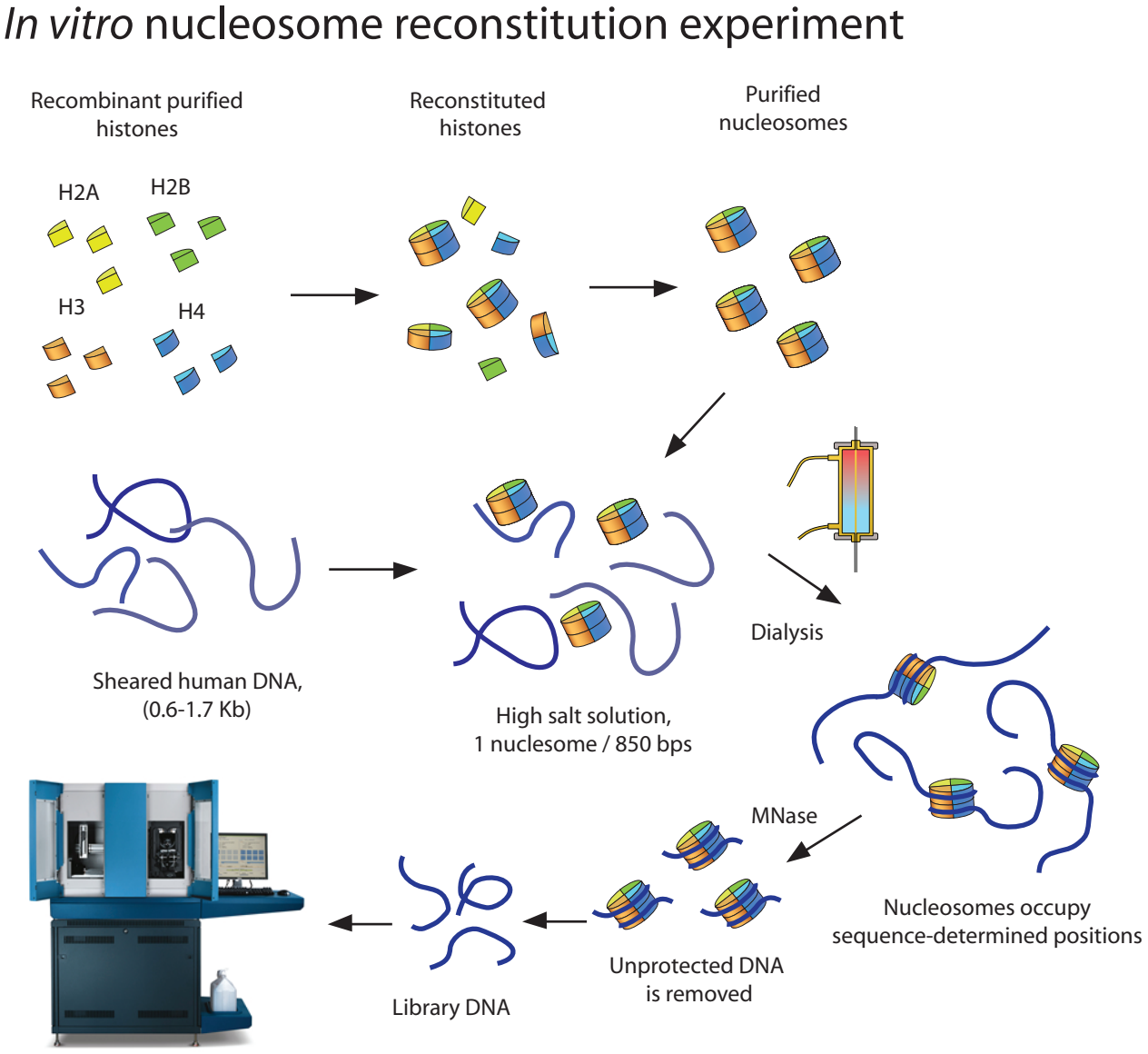
The estimates of spacing between nucleosomes as depicted in Fig. 1d are consistent between the two types of T cells we analysed. To ask whether these estimates were also reproducible by a different approach, we turned to a published data set that was generated for a different purpose, and by different 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).

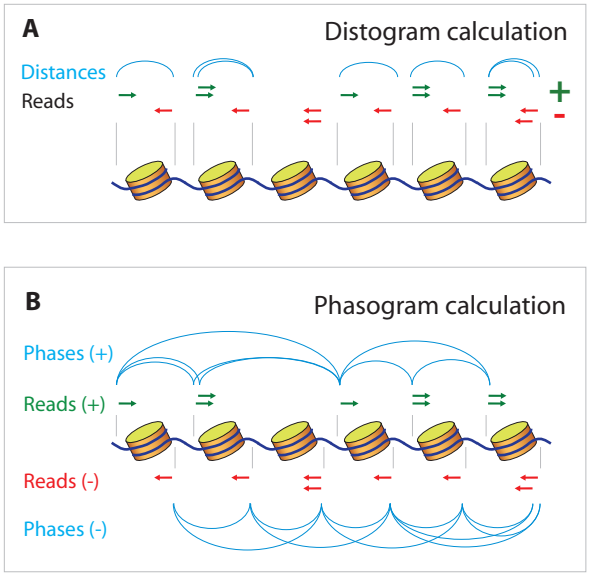
doi:10.1038/nature10002



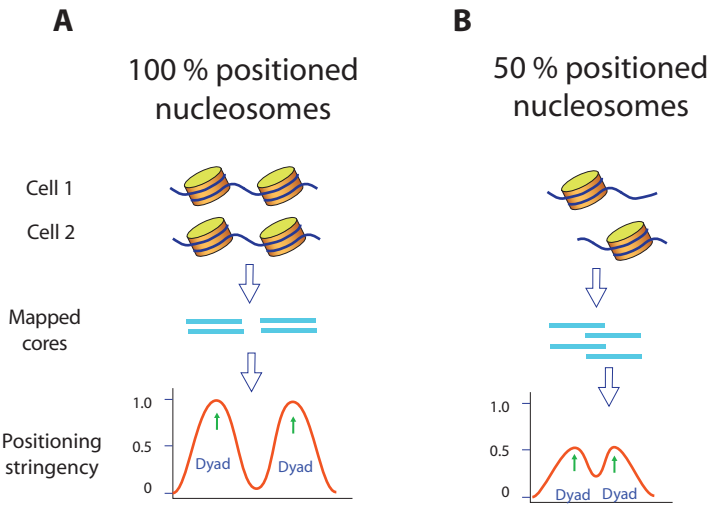
Supplementary Figure 1. Schematic depiction of *in vivo* nucleosome mapping experiment. Blood cells were isolated from a human donor blood and sorted into populations representing CD4+ T-cells, CD8+ T-cells and granulocytes. Nuclear chromatin was released by crushing the cells, followed by Micrococcal nuclease treatment. Mononucleosome fraction was isolated by gel electrophoresis and sequenced to high depth using SOLiD platform.



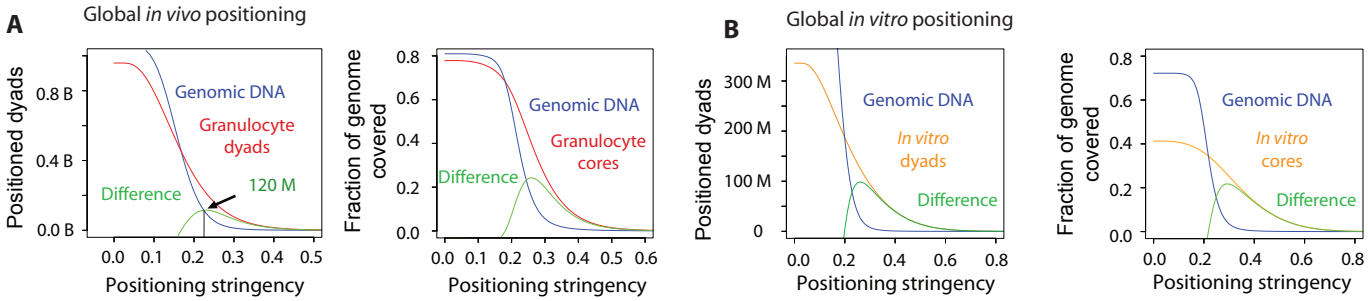
Supplementary Figure 2. Schematic representation of *in vitro* reconstitution experiment. Recombinant histones were assembled to produce the histone octamer particles. Human genomic DNA was sheared to a range of 0.6-1.7 Kb and combined with octamers at a ratio of one octamer per 850 bps of DNA. The salt was gradually dialyzed away and unbound DNA was removed by Micrococcal nuclease treatment. Nucleosome-bound DNA was purified and sequenced on the SOLiD platform.



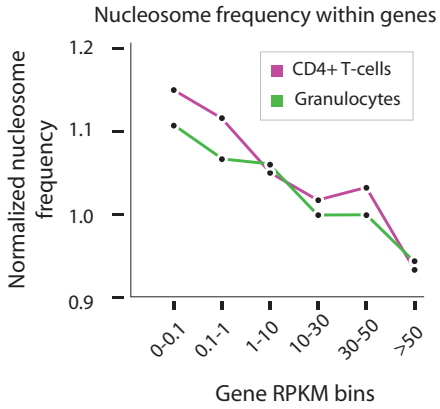
Supplementary Figure 3. Distograms and phasograms. (A) Schematic depiction of the distogram calculation. Blue arcs represent recorded distances between nucleosome reads that map on opposite strands. Distance frequencies are represented as a histogram (distogram, see Fig. 1A-B of the main text). Distograms are used to reveal the existence of consistently positioned nucleosomes in the main data. **(B)** Schematic depiction of the phasogram calculation. Blue arcs represent recorded phases between the nucleosome reads mapping on the same strand of the reference genome. Phase frequencies are represented as a histogram (phasogram, see Fig 1C-D). Phasograms are used to reveal the existance of consistently spaced nucleosomes forming regular arrays.



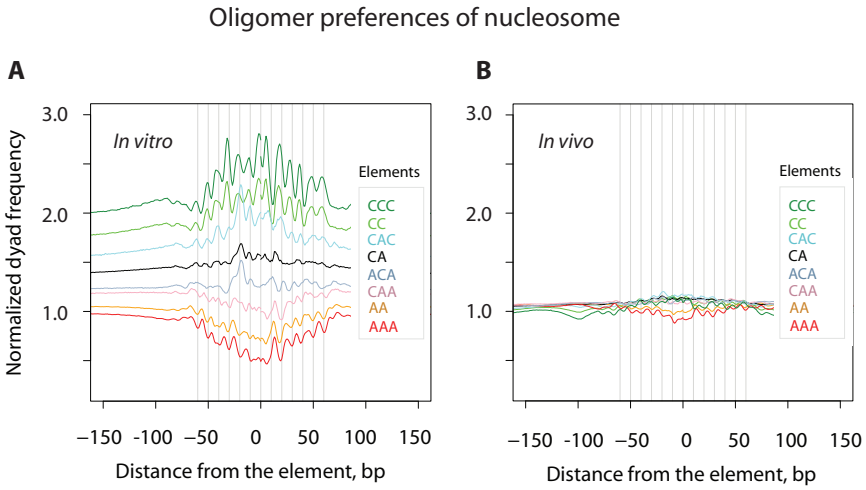
Supplementary Figure 4. Schematic depiction of the nucleosome positioning stringency metric. At the sites containing perfrectly positioning nucleosomes (panel **A**) the stringency values are 1.0 (100% positioning), and at the sites containing two mutually exclusive nucleosome positions which are utilized with 50% frequency across cells (panel **B**), the stringency values are 0.5 (50% positioning frequency at each of the two sites). Nucleosome dyad positions are identified as the local maxima of the stringency profile (green arrows).



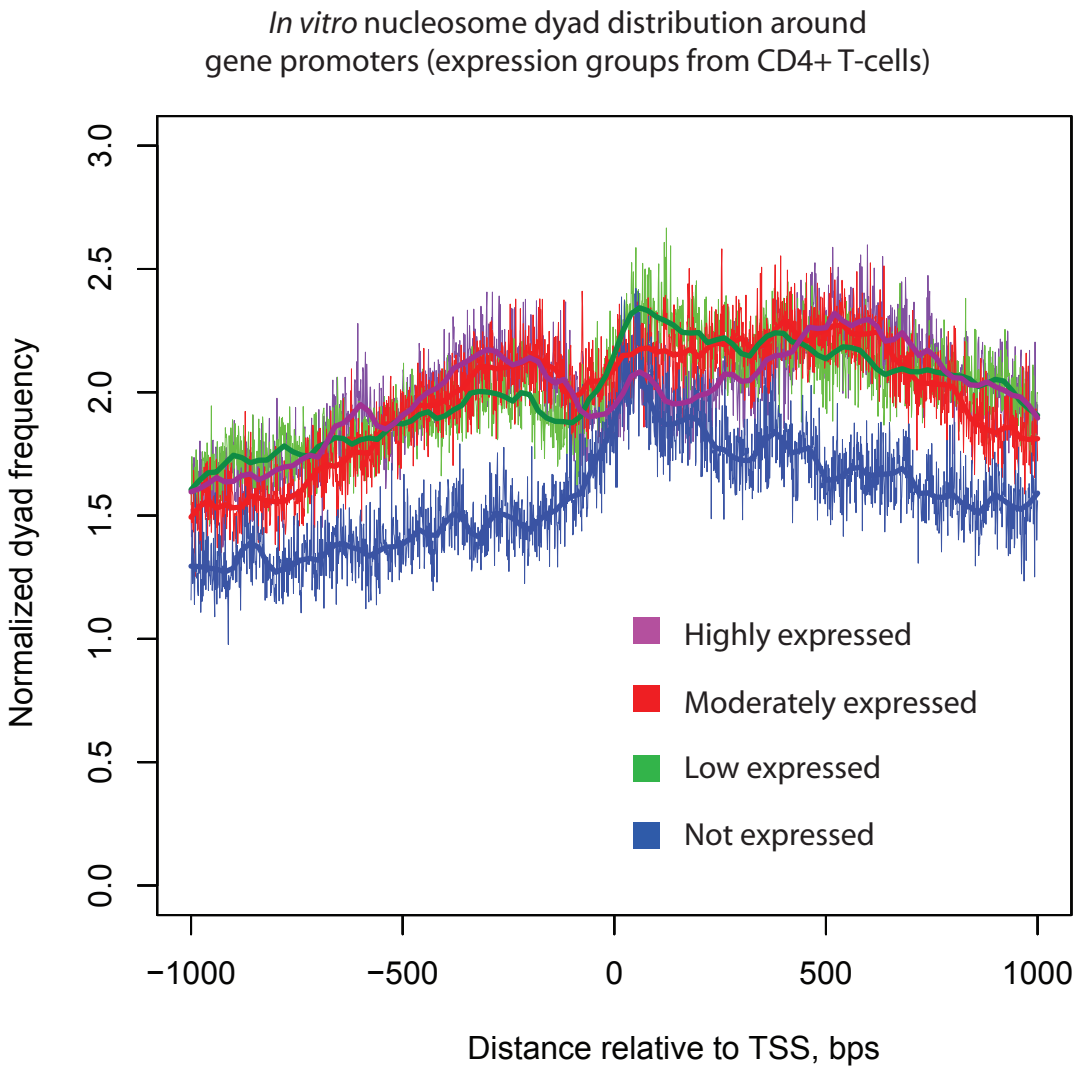
Supplementary Figure 5. Genome-wide positioning of nucleosomes. (A) Global *in vivo* nucleosome positioning of granulocytes. In both panels, X axis represents a range of positioning stringency cutoffs. In the left panel, Y axis represents the number of positioned dyads at a given positioning stringency cutoff. The red curve represents granulocyte data, the blue curve represents genomic DNA control matched to the number of granulocyte reads, the green curve represents the difference curve that provides the number of statistically positioned dyads at a given stringency cutoff. In the right panel, Y axis represents the fraction of the genome covered by 147 bp nucleosome cores centered at the dyad positions exceeding a given stringency. The red curve represents granulocyte nucleosome data, the blue curve represents genomic control matching the granulocyte data read number, and the green curve represents the difference between granulocytes and control curves and gives the fraction of the genome covered by statistically positioned nucleosomes. **(B)** Global *in vitro* nucleosome positioning. The data are plotted as in **(A)** using *in vitro* data and control matching the read number of the *in vitro* data set.



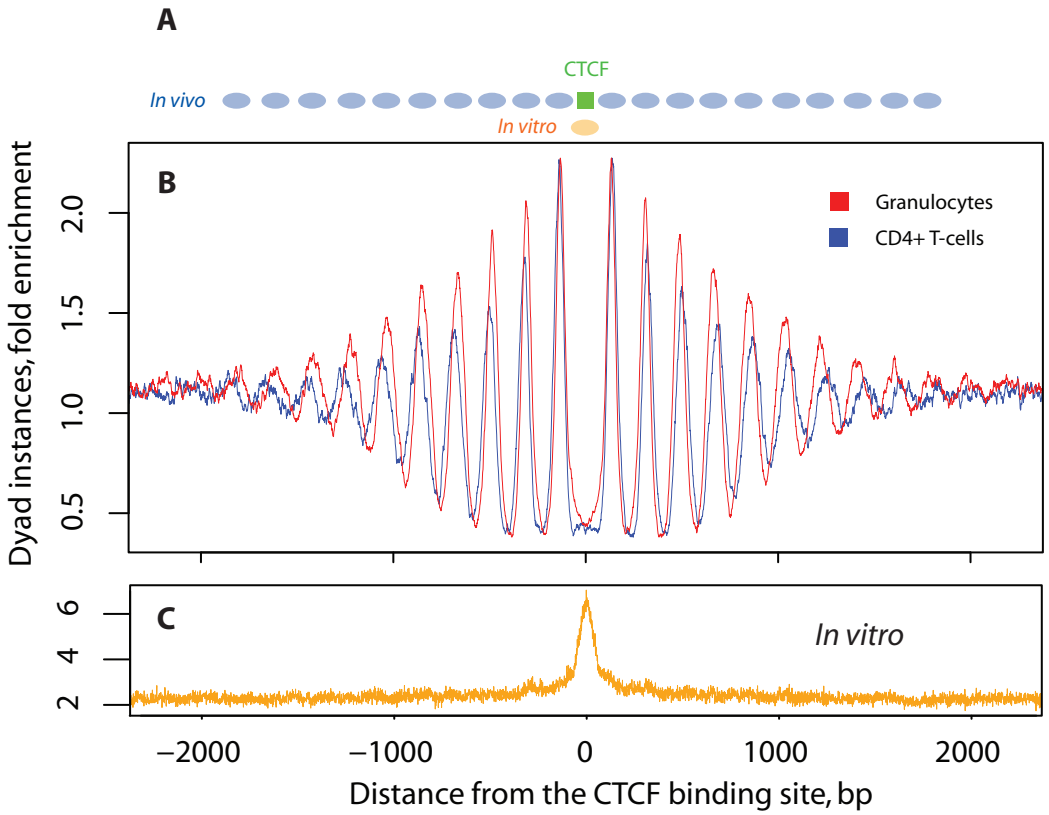
Supplementary Figure 6. Association between transcriptional levels and measured nucleosome occupancy. X axis represents gene expression values binned according to their RPKM values. Y axis represents normalized frequencies of observed nucleosome coverage within the regions occupied by genes in each bin.



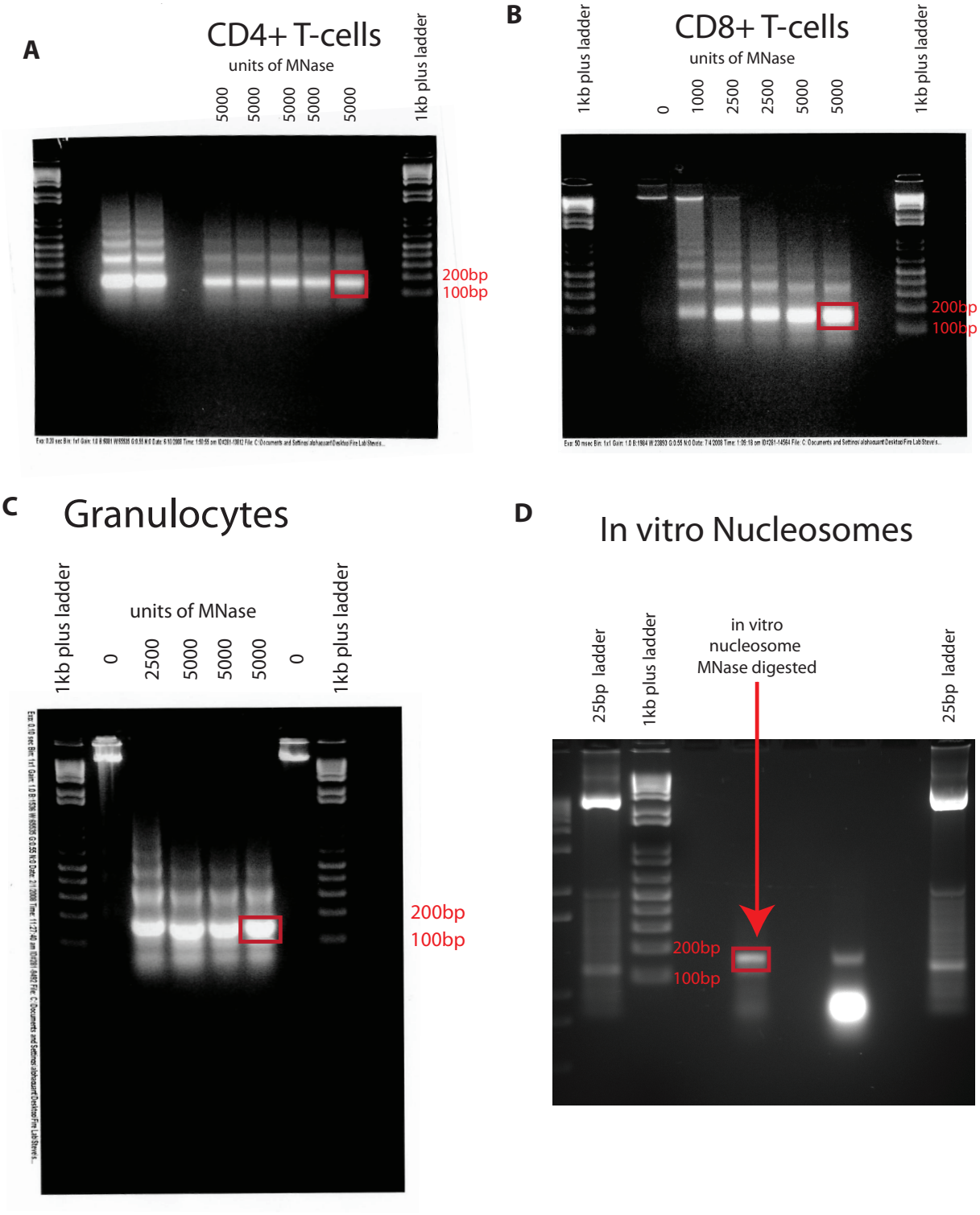
Supplementary Figure 7. **(A)** Signatures of rotational positioning of *in vitro* nucleosomes. Shown are preferences relative to most dimers and trimers composed of Cs and As. X axis represents a distance from a given oligomer to a dyad inferred from mapped sequence reads. Y axis represents the frequency of dyads at a given distance normalized to the expected frequency. 10bp-spaced peaks represent helical rotational preferences of oligomers relative to nucleosome surface. **(B)** Signatures of rotational positioning of *in vivo* granulocyte nucleosomes against the same panel of oligomers.



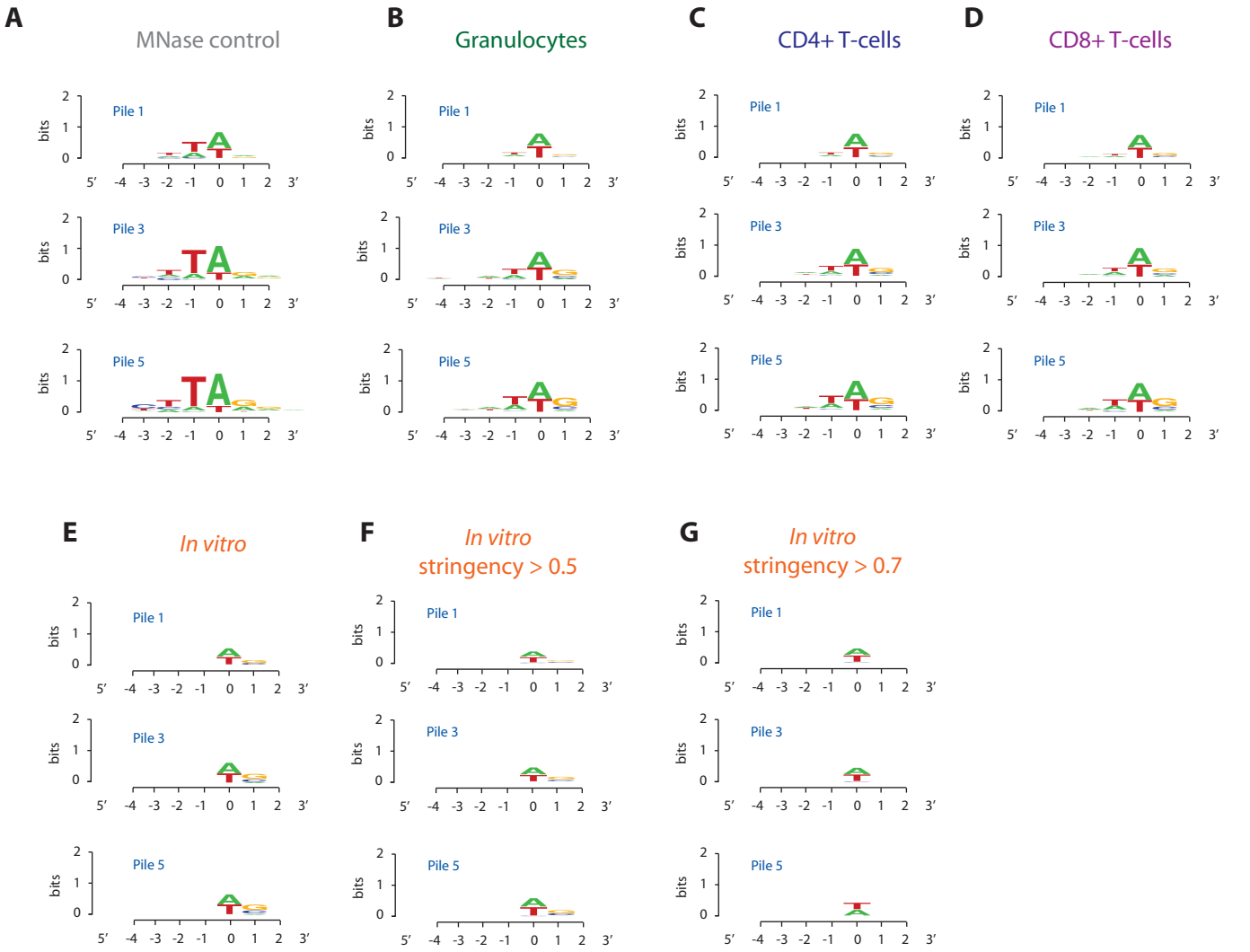
Supplementary Figure 8. Sequence-encoded nucleosome organization around TSS. Plotted are frequencies of *in vitro* nucleosome dyads around promoters of genes binned according to their expression levels in CD4+ T-cells. X axis represents the distances relative to the TSS (left of zero is away from the gene). Y-axis represents frequencies of nucleosome dyads normalized to the genome-wide average. Each of the 4 gene bins is represented by a line of a corresponding color displayed in the legend.



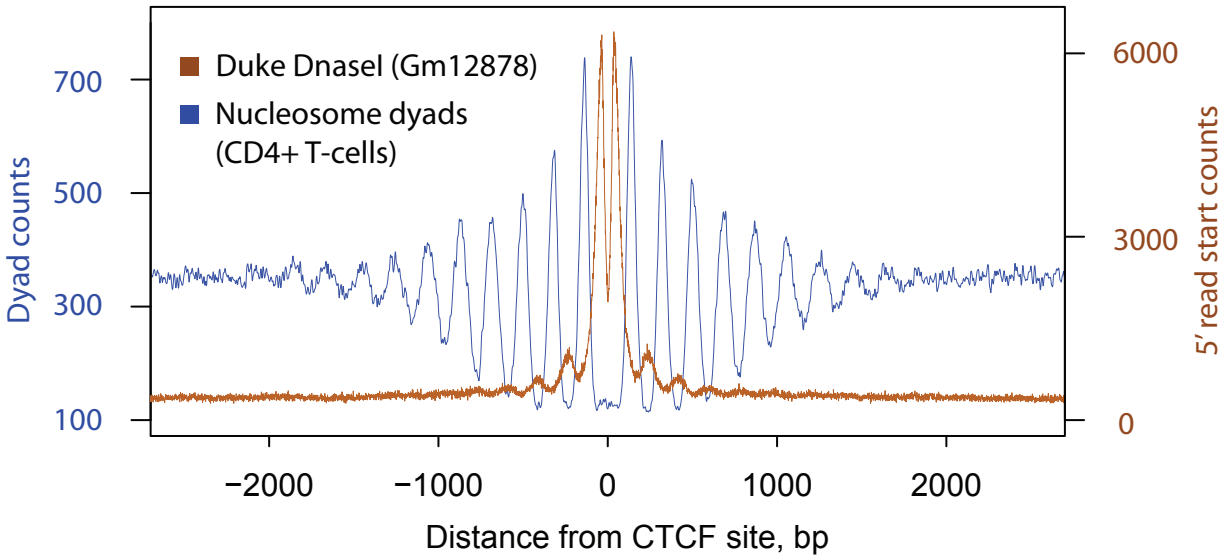
Supplementary Figure 9. Nucleosome organization around CTCF binding sites. (**A**) Schematic depiction of nucleosome organization inferred from the data. The blue ovals represent *in vivo* nucleosome positions, the green square represents binding of CTCF protein which is flanked by two well-positioned nucleosoomes. The orange oval represents preferred position of nucleosomes *in vitro*. (**B**) Dyad frequencies around CTCF binding site. Binidng sites were aligned so that position 0 represents coordinate of CTCF binding inferred from CTCF data in CD4+ T-cells. X-axis represents 4 Kbp window around CTCF binding site, Y-axis represents normlized frequencies of dyads across the regions. The red curve represents smoothed frequency of nucleosome dyads from granulocytes, the blue curve represents smoothed nucleosome dyad frequency in CD4+ T-cells. (**C**) Dyad frequencies in the *in vitro* reconstitution data around CTCF binding sites.



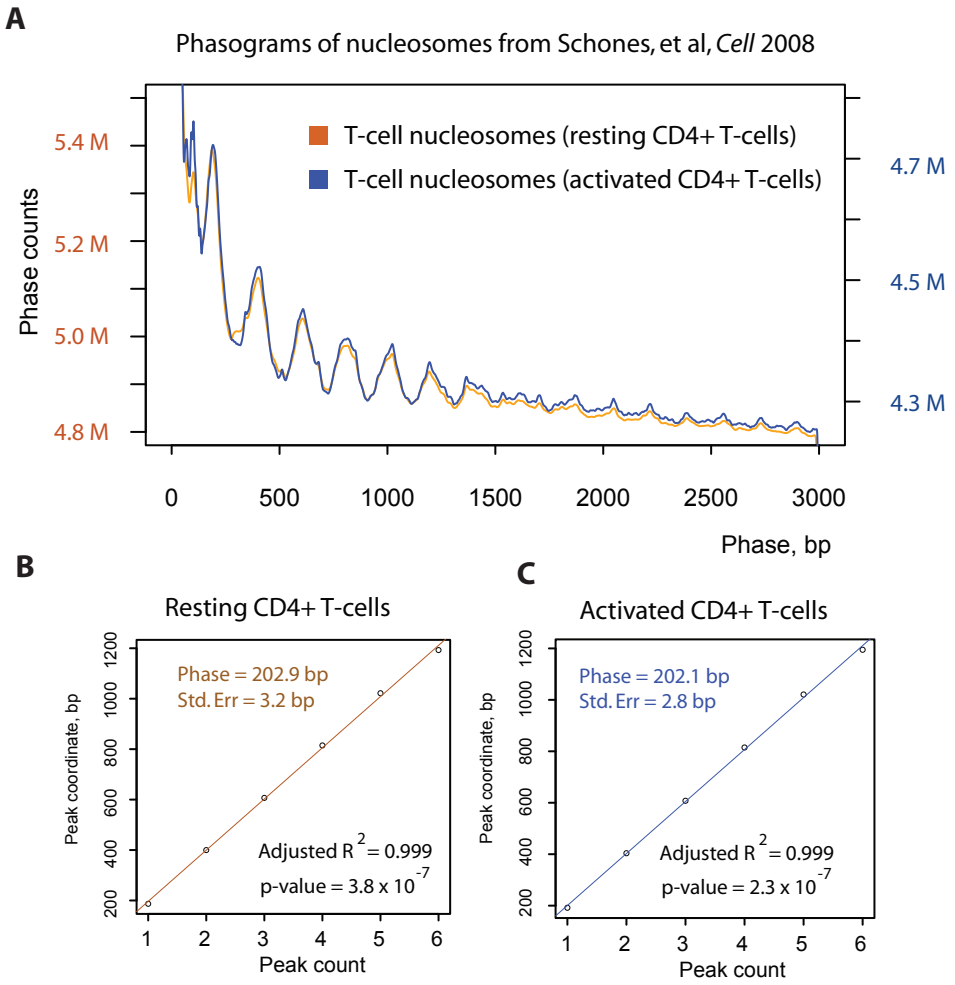
Supplementary Figures 10. Isolation of nucleosome-bound DNA. Agarose gels of nucleosome-bound DNA after micrococcal-treatment in CD4+ T-cells (**A**), CD8+ T-cells (**B**), Granulocytes (**C**), and in vitro reconstituted nucleosomes (**D**). Bands isolated for sequencing are marked by red rectangles.



Supplementary Figure 11. Micrococcal nuclease sequence bias anaysis. Shown are Weblogos (Crooks et al 2004) across sites cleaved by micrococcal nuclease in the control data (A), *in vivo* nucleosome data (B-D), and *in vitro* nucleosome data (E-G). We examined sites containing nucleosomes of increasing positioning strength (Pile1, sites with 1 or more read starts on the same strand; Pile3, sites with 3 or more read starts; Pile5, sites with 5 or more reads starts). For each subset, we aligned start positions and plotted nucleotide frequency at corresponding sites, with 0 representing the first sequenced base of the fragments. For the sites containing positioned *in vitro* nucleosomes (stringency > 0.5 and > 0.7), we plotted nucleotide frequencies from overlapping nucleosome fragments.



Supplementary Figure 12. Chromatin structure around CTCF sites. We plotted Dnase I cutting frequency (brown) and dyad frequencies (blue) around CTCF binding sites. Dnase I cleavage frequency is represented by plotting frequency of 5′ ends from Dnase I sequence reads using Duke Dnase-seq protocol (Song and Crawford, 2010) in the lymphoblastoid cell line. Peaks of Dnase I are in strong counter-phase with dyads, representing cleavage sites localizing within the nucleosome linker DNA. In addition, a strong peak of Dnase I can be seen between the CTCF binding site and the first wel-positioned nuclesome.



Supplementary Figure 13. Nucleosome spacing in resting and activated T-cells. (**A**) Phasograms of nucleosomes in resting and activated T-cells (Schones et al, 2008). Nucleosome spacing was estimated using a linear fit to peak positions in the corresponding phasograms. (**B**) Spacing was estimated to be 202.9 bps in resting T-cells, and (**C**) 202.1 bps in activated T-cells. These results provide independent replication of phasing estimates in CD4+ and CD8+ T cells (Fig. 1D).

Dynamic transcriptional events in embryonic stem cells mediated by the super elongation complex (SEC)

Chengqi Lin, Alexander S. Garrett, Bony De Kumar, et al.

Genes Dev. 2011 25: 1486-1498
Access the most recent version at doi:10.1101/gad.2059211

Dynamic transcriptional events in embryonic stem cells mediated by the super elongation complex (SEC)

Chengqi Lin,[1] Alexander S. Garrett,[1] Bony De Kumar,[1] Edwin R. Smith,[1] Madelaine Gogol,[1] Christopher Seidel,[1] Robb Krumlauf,[1,2] and Ali Shilatifard[1,3]

[1]Stowers Institute for Medical Research, Kansas City, Missouri 64110, USA; [2]Department of Anatomy and Cell Biology, University of Kansas Medical Center, Kansas City, Kansas 66160, USA

Transcriptional regulation of developmentally controlled genes is at the heart of differentiation and organogenesis. In this study, we performed global genomic analyses in murine embryonic stem (ES) cells and in human cells in response to activation signals. We identified an essential role for the ELL (eleven–nineteen lysine-rich leukemia gene)/P-TEFb (positive transcription elongation factor)-containing super elongation complex (SEC) in the regulation of gene expression, including several genes bearing paused RNA polymerase II (Pol II). Paused Pol II has been proposed to be associated with loci that respond rapidly to environmental stimuli. However, our studies in ES cells also identified a requirement for SEC at genes without paused Pol II, which also respond dynamically to differentiation signals. Our findings suggest that SEC is a major class of active P-TEFb-containing complexes required for transcriptional activation in response to environmental cues such as differentiation signals.

[*Keywords*: transcription elongation; immediate early genes; ELL2; AFF4]

Supplemental material is available for this article.

Received April 14, 2011; revised version accepted June 10, 2011.

Transcriptional regulation by RNA polymerase II (Pol II) is a multifaceted process requiring the concerted action of a large number of factors for the steadfast synthesis of full-length messenger RNA (mRNA) (Workman and Kingston 1998; Shilatifard et al. 2003; Sims et al. 2004; Bres et al. 2008; Boettiger and Levine 2009). Transcription by Pol II is divided into four stages: initiation, promoter clearance, elongation, and termination. The initiation stage of transcription requires nucleosomal remodeling around the enhancer and promoter regions followed by the recognition of the promoter elements by the basal transcription machinery and Pol II. Once the basal factors and Pol II are recruited to the promoter elements, the catalysis of the first phosphodiester bond marks the initiation of transcription (Shilatifard 1998; Sims et al. 2004). For many years, it was considered that transcription initiation was the rate-limiting step to the transcription process as a whole. However, a large number of studies demonstrated that the elongation stage of transcription regulated by a number of factors is essential for productive transcription (Shilatifard et al. 2003; Sims et al. 2004; Levine 2011). In support of a vital role for the elongation stage of transcription in development, it has been demonstrated that the perturbation of this stage of transcription or the factors involved in this process results in the pathogenesis of human diseases, including cancer (Shilatifard et al. 2003; Mohan et al. 2010).

In addition to the control of the productive elongation stage of transcription by Pol II elongation factors, many developmentally regulated genes in mammalian cells are marked by stalled or paused Pol II at their proximal-promoter regions (Muse et al. 2007; Zeitlinger et al. 2007; Core et al. 2008; Boettiger and Levine 2009; Rahl et al. 2010; Levine 2011). Such polymerases have already been initiated and are awaiting proper developmental signals to enter the processive stage of transcription elongation (Rougvie and Lis 1988). Some studies have suggested that marking such developmentally regulated genes by paused Pol II could enhance their ability to be induced rapidly in a robust manner (Nechaev and Adelman 2008). Other studies, however, have proposed that the presence of paused Pol II at developmentally regulated genes allows for a synchronous induction of the same set of genes in distinct cell populations at the appropriate stage of development (Boettiger and Levine 2009).

Multiple factors have been identified to achieve proper promoter clearance and the processive elongation stage of transcription during development. These factors include Elongin A, DSIF (DRB sensitivity-inducing factor), NELF

[3]**Corresponding author.**
E-mail ASH@Stowers.org.
Article is online at http://www.genesdev.org/cgi/doi/10.1101/gad.2059211.

(negative elongation factor), P-TEFb (positive transcription elongation factor), and ELL (eleven–nineteen lysine-rich leukemia gene) (Jones and Peterlin 1994; Shilatifard et al. 2003; Sims et al. 2004; Peterlin and Price 2006; Bres et al. 2008; Levine 2011). Both Elongin A and DSIF are capable of increasing the catalytic rate of the productive transcription by Pol II; however, in addition to its role in this process, DSIF also works with NELF to regulate Pol II arrest (Yamaguchi et al. 1999; Shilatifard et al. 2003; Cheng and Price 2007). Such arrested Pol IIs are released by the Cdk9 kinase activity of P-TEFb, which phosphorylates the C-terminal domain (CTD) of Pol II and many of the other transcription factors, signaling the release of the stalled Pol II into productive transcription (Jones and Peterlin 1994; Fuda et al. 2009). ELL was purified based on its catalytic properties to increase the V_{max} rate of transcription elongation by Pol II (Shilatifard et al. 1996; Shilatifard 1998). Translocations of ELL involving the mixed-lineage leukemia gene *MLL* are associated with the pathogenesis of childhood leukemia and misregulation of developmental genes (Thirman et al. 1994). In addition to ELL, a large number of genes with very little sequence or obvious functional similarities are found in translocations with MLL in leukemia (Mohan et al. 2010).

In support of the hypothesis that the elongation stage of transcription by Pol II has an essential role in development and cancer pathogenesis, ELL and several other MLL translocation partners were biochemically isolated as part of the super elongation complex (SEC) that contains P-TEFb (Lin et al. 2010). SEC has also been shown to play a role in regulating the elongation stage of transcription controlled by HIV-1 Tat (He et al. 2010; Sobhian et al. 2010). These studies suggest that MLL translocations function by regulating the elongation stage of transcription on developmentally regulated genes, such as the *HOX* loci, through the association of MLL chimeras with P-TEFb within the ELL-containing SEC. This association of SEC with MLL through chromosomal translocations can result in the premature release of paused Pol II at developmental loci (Mohan et al. 2010; Smith et al. 2011).

P-TEFb participates in a variety of complexes, both active and inactive (Bres et al. 2008; He and Zhou 2011). Both Brd4 and Myc-containing P-TEFb complexes have been considered to be major regulators of transcription elongation (Zhou and Yik 2006; Zippo et al. 2009; Donner et al. 2010; Rahl et al. 2010). To investigate to what degree SEC functions genome-wide in transcription elongation control, we performed chromatin immunoprecipitation (ChIP) and sequencing (ChIP-seq) studies in both mouse embryonic stem (ES) cells in response to retinoic acid (RA) induction and human HCT-116 cells in response to serum stimulation. Our studies in mouse ES cells identified gene targets for SEC, many of which are developmental regulators with paused Pol II that were rapidly induced to high, but relatively uniform, levels. Our studies in human HCT-116 cells found that SEC is also a major regulator of immediate early genes induced by growth factors.

Together, these findings suggest that the presence of paused Pol II at promoter-proximal regions and recruitment of SEC upon activation may represent a major cellular mechanism for rapid and uniform induction of gene expression upon exposure to key developmental signals. Intriguingly, our global genomic studies in ES cells also identified a requirement for SEC at *Cyp26a1*, a gene that does not bear paused Pol II at its promoter-proximal region, yet responds dynamically to RA in an even more rapid manner than other genes that have paused Pol II at their promoter-proximal regions. Our findings suggest that the recruitment of SEC allows genes to respond in a rapid and dynamic manner to developmental signals in different cell types in mammals, and that SEC is involved in transcriptional induction that is both dependent on and independent of the presence of paused Pol II.

Results

The SEC is enriched on highly transcribed genes in ES cells

We previously identified the SEC as a transcription elongation complex containing the known elongation factors ELL1–3, EAF1–2, P-TEFb, and several other translocation partners of MLL found in leukemia (Fig. 1A), including AFF1, AFF4, AF9, and ENL (Lin et al. 2010). SEC was previously shown to be required for a normal cellular function, the induction of the *HSP70* gene upon stress, and the misregulation of transcription of the *HOXA9* and *HOXA10* genes by MLL chimeras (Lin et al. 2010). SEC was later found to be required for Tat-mediated HIV transactivation (He et al. 2010; Sobhian et al. 2010). Our previous studies demonstrated a close relationship between AFF4 and ELL2; AFF4 was central for the formation of SEC, and the RNAi-mediated knockdown of AFF4 led to the destabilization of the ELL2 protein (He et al. 2010; Lin et al. 2010). To investigate a possible role of SEC in the control of developmental genes poised for activation in early development, we developed antibodies to SEC components (Supplemental Fig. S1; Lin et al. 2010) and performed a genome-wide occupancy analysis of the SEC components in mouse ES cells using ChIP-seq of AFF4, ELL2, Cdk9, and RNA Pol II. These SEC components co-occupy many of the same genes, including highly expressed housekeeping genes such as the histone genes (Supplemental Fig. S2); however, SEC is only found at a subset of highly transcribed genes (Fig. 1B,C,E). SEC components are enriched at the transcription start site (TSS) regions of these genes and within the gene body similar to the Pol II distribution (Fig. 1D; Supplemental Fig. S2). The co-occupancy of the AFF4 and ELL2 components of SEC correlates with a high level of expression of genes in mouse ES cells (Fig. 1E) suggesting that SEC is frequently associated with highly transcribed regions.

Paused Pol II and recruitment of SEC at the developmentally regulated Hox *clusters in ES cells*

The SEC was discovered based on the purification of several of the MLL chimeras that are commonly found in

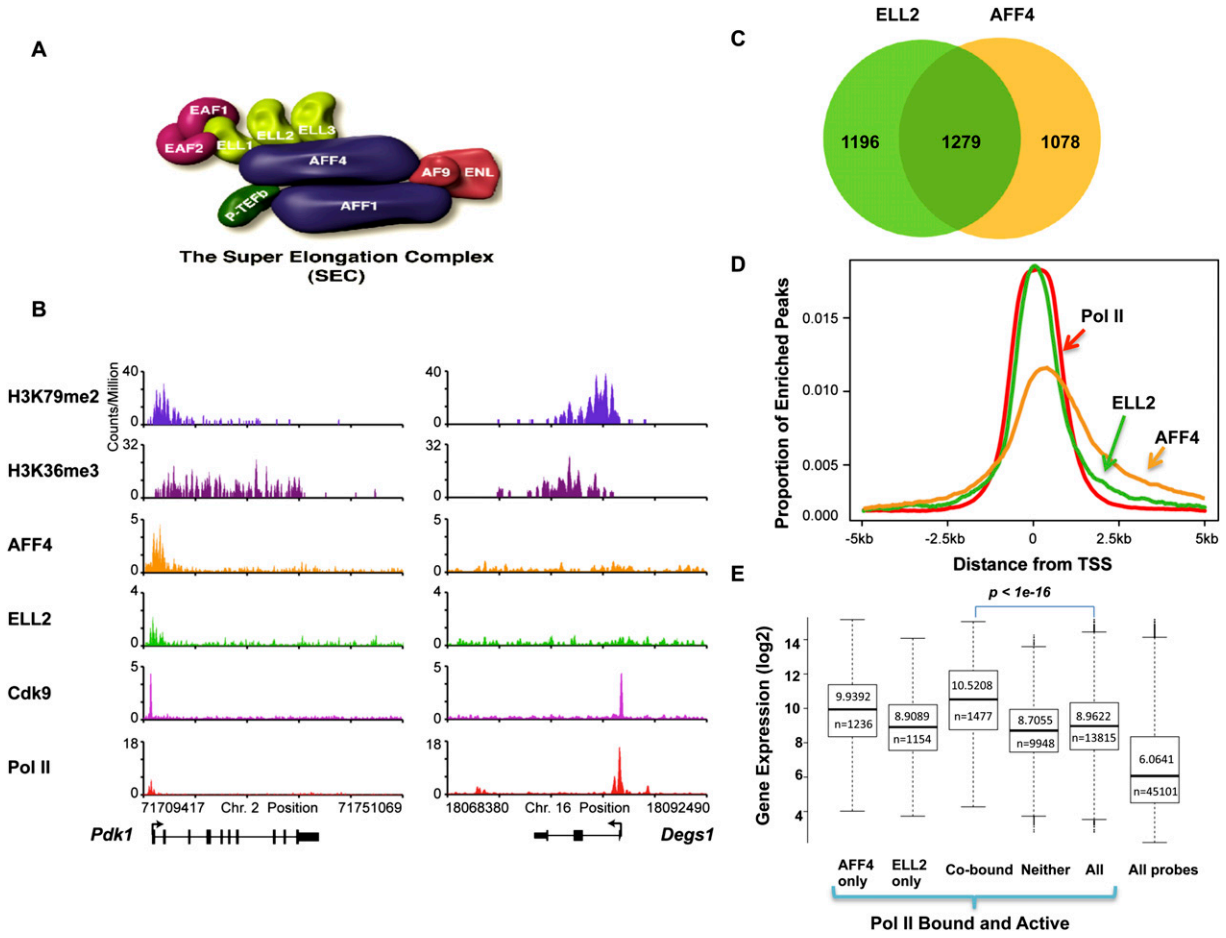


Figure 1. Global occupancy of the SEC subunits in mouse ES cells. (*A*) Schematic representation of the SEC. The SEC is a P-TEFb-containing complex that contains various combinations of four types of proteins: ELL1–3; EAF1–2; AFF1 and AFF4; and AF9 and ENL. P-TEFb itself consists of Cdk9 and CycT1/2 and is best characterized as an RNA Pol II CTD kinase. (*B*) Genome-wide analysis of SEC components AFF4, ELL2, and Cdk9 by ChIP-seq in ES cells found SEC enriched at a subset of actively transcribed genes. Shown are two genes with high levels of expression in ES cells. (*Left* panel) The *Pdk1* gene is occupied by the SEC subunits ELL2 and AFF4. (*Right* panel) The *Degs1* gene does not have significant levels of the SEC components AFF4 and ELL2. H3K36 trimethylation (H3K36me3) and H3K79 dimethylation (H3K79me2) data from Marson et al. (2008) are shown as markers of actively transcribed genes. (*C*) Venn diagram analysis of AFF4- and ELL2-occupied genes. Around 50% of AFF4-enriched genes are also occupied by ELL2, demonstrating that in mouse ES cells, these two proteins share a similar global occupancy. (*D*) Histogram of the genome-wide occupancy of AFF4, ELL2, and Pol II. The canonical TSS of each gene in the genome was used to measure the distance to the nearest bound region, which is plotted if falling within 5 kb of the TSS. This analysis shows that SEC components are enriched over the TSS, similar to Pol II occupancy. (*E*) AFF4 and ELL2 co-occupy highly transcribed genes. The dark lines in the box plots and the number *above* the line indicate the median level of expression for the gene subset indicated. The number *below* the line indicates the number of Affymetrix probe sets that correspond to the gene subset. Probe sets for ELL2 and AFF4 cobound genes show significantly higher expression compared with all Pol II-bound and active genes ($P < 1 \times 10^{-16}$ by Wilcoxon two-sample rank sum test). The gene subset containing neither AFF4 nor ELL2 also shows some highly expressed genes. Genes were called active if they were determined present on the array by the MAS5 algorithm.

MLL-rearranged leukemias (Lin et al. 2010). It is not currently known why SEC components are so frequently found to be translocation partners with MLL. One possibility is that the genes misregulated by MLL chimeras, such as the *HOX* genes, are normal SEC targets. In these leukemias, MLL target genes become misregulated when SEC is recruited inappropriately and prematurely activates transcription by releasing paused Pol II (Lin et al. 2010; Mohan et al. 2010; Smith et al. 2011). Many developmentally regulated genes in flies and mammals have paused Pol II at the TSS before their activation during development (Muse et al. 2007; Zeitlinger et al. 2007). In mammalian stem cells, these genes are characterized by a bivalent mark of both H3K4 and H3K27 trimethylation on the same gene (Bernstein et al. 2006). Looking within the *Hox* clusters in ES cells, we found bivalent marks co-occurring with Pol II at the TSS at four of the *Hoxa* cluster genes (*Hoxa1*, *Hoxa3*, *Hoxa4*, and *Hoxa7*), but not at the promoters of the *Hoxb* genes (Fig. 2).

The regulation of gene transcription at the level of paused Pol II and its controlled release have been best

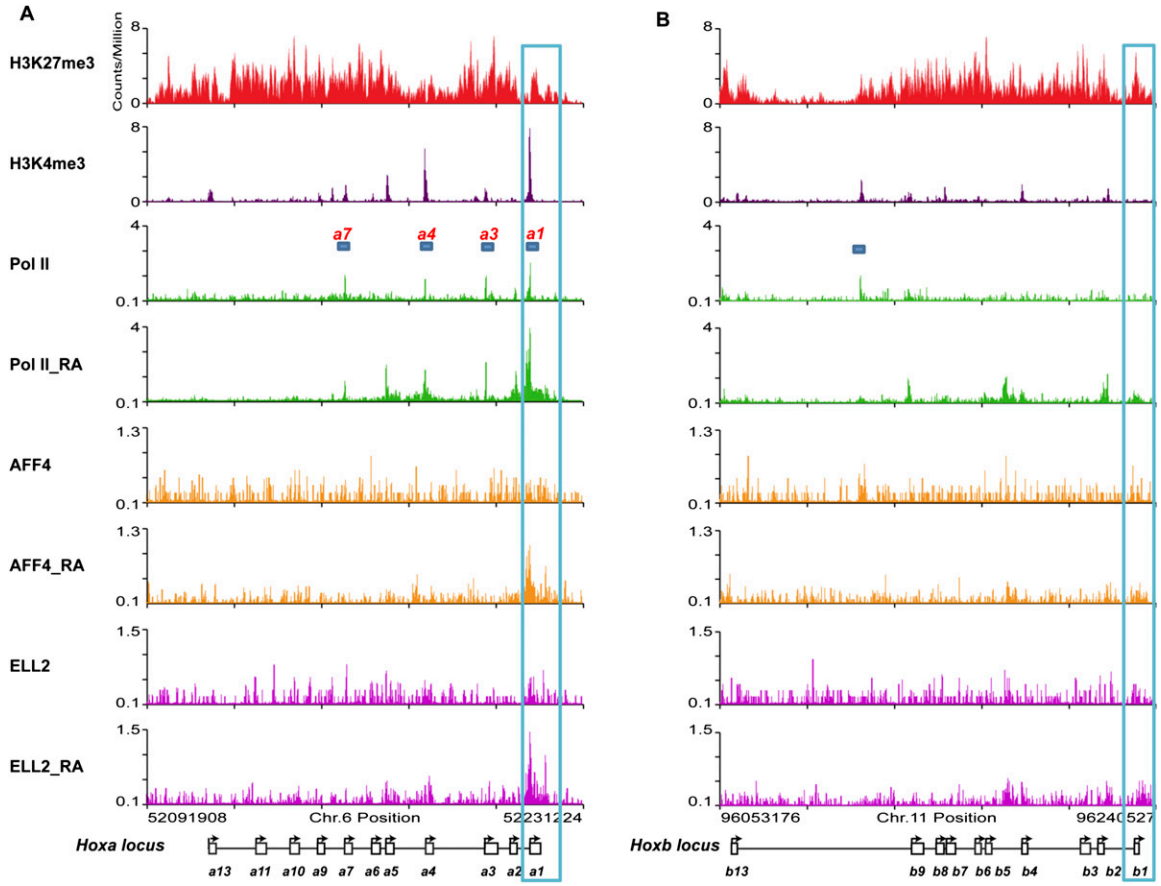


Figure 2. The *Hoxa1* promoter is preloaded with Pol II and recruits SEC after RA treatment in ES cells. (*A*) Bivalent marks, paused Pol II, and SEC recruitment to the *Hoxa* cluster. In ES cells, the whole *Hoxa* cluster is highly enriched for H3K27me3, and also contains H3K4me3 at the promoters of a subset of genes, including *Hoxa1*, *Hoxa3*, *Hoxa4*, and *Hoxa7*. These regions are preloaded with Pol II (bars indicate regions that have both a bivalent mark and Pol II). (*B*) Bivalent marks and paused Pol II are both largely absent from the *Hoxb* genes, which do not recruit SEC after 6 h of RA treatment. While H3K27me3 marks the whole cluster of *Hoxb* genes, only *Hoxb4*, *Hoxb7*, and *Hoxb9* contain H3K4me3 at their promoters and can be considered bivalent. There is no significant Pol II detected on the promoters of the *Hoxb* genes in ES cells. The bar marks a peak of significant Pol II that does not correspond to a known gene feature. Before RA treatment, there is no detectable AFF4 and ELL2 signal on the *Hoxa* or *Hoxb* cluster genes. Both AFF4 and ELL2 are recruited to the *Hoxa1*, but not the *Hoxb1*, gene promoter after exposure to RA for 6 h. Blue boxes highlight the *Hoxa1* and *Hoxb1* genes. Expanded views of the *Hoxa1* and *Hoxb1* regions are shown in Supplemental Figure S4.

studied at the heat-shock genes such as *HSP70*, as well as in the control of HIV transcription, and both processes require SEC (He et al. 2010; Lin et al. 2010; Sobhian et al. 2010). Genes with paused Pol II, such as *HSP70*, are transcriptionally engaged with pausing 30–40 nucleotides downstream from the TSS (Core et al. 2008; Nechaev et al. 2010). These genes contain basal transcriptional machinery at their promoters and have a form of Pol II phosphorylated on Ser5, but not Ser2, of the CTD, and the Pol II is associated with DSIF/NELF (Nechaev and Adelman 2011). By all of these criteria, *Hoxa1*, but not *Hoxb1*, is occupied and engaged by paused Pol II (Supplemental Fig. S3).

This led us to ask whether SEC was differentially recruited to *Hoxa1* and not *Hoxb1* upon induction by RA treatment. We performed AFF4, ELL2, and Pol II ChIP-seq after 6 h of RA treatment of mouse ES cells. Pol II was recruited to both *Hoxa1* and *Hoxb1* promoters after exposure to RA for 6 h (Fig. 2; Supplemental Fig. S4). Interestingly, AFF4 and ELL2 were only recruited to *Hoxa1*, and not *Hoxb1*, by 6 h of RA treatment (Fig. 2; Supplemental Fig. S4). However, we cannot rule out the possibility that SEC was not detected at *Hoxb1* due to lower levels of Pol II and a concomitant decrease in SEC that falls below our detection level. Our genome-wide analyses suggest that our ability to detect SEC occupancy on a gene is not strictly dependent on levels of Pol II or transcription levels (Fig. 1B,E).

SEC is required for the rapid induction of Hoxa1

Promoter-proximal paused Pol II has been proposed to allow for a more rapid induction of genes upon differentiation cues (Nechaev and Adelman 2008). Therefore, we assayed the induction kinetics of *Hoxa* and *Hoxb* cluster genes by RT-qPCR after RA treatment from 2–24 h (Fig. 3A,B). We found that *Hoxa1* and *Hoxb1* were the first

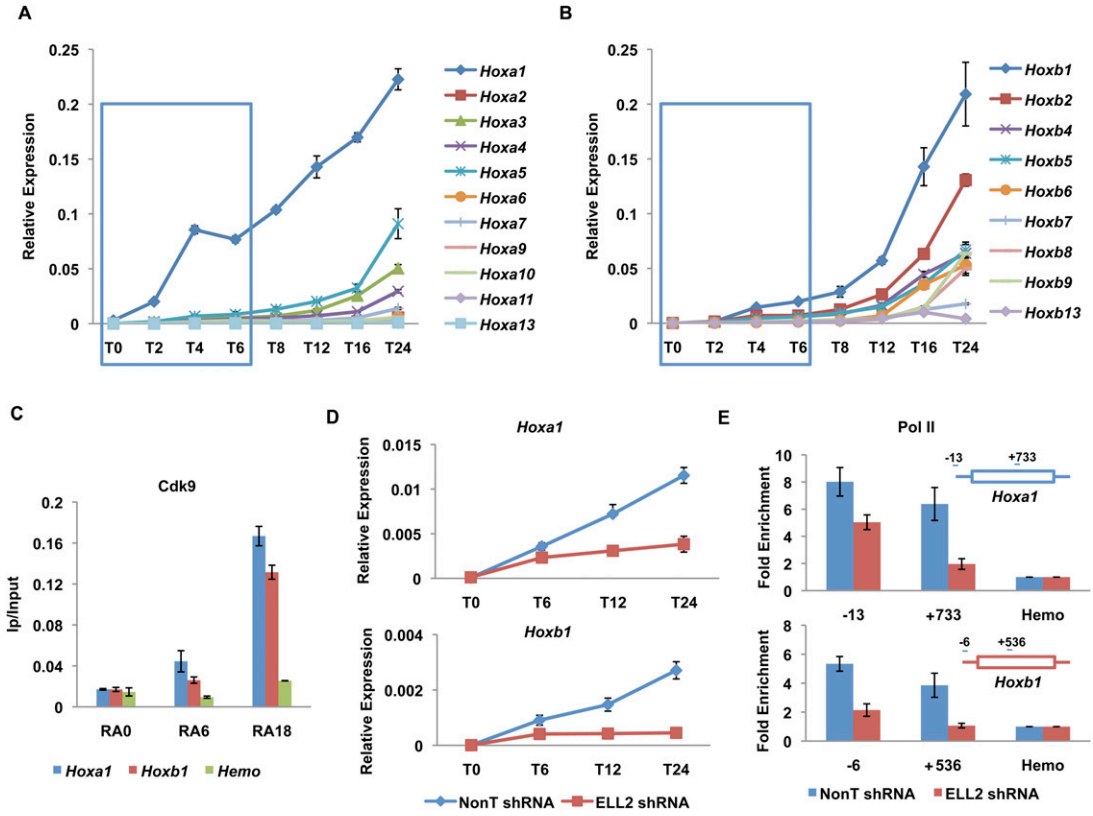


Figure 3. SEC is required for the rapid induction of the *Hoxa1* gene. (*A*,*B*) RT-qPCR analysis of *Hoxa* and *Hoxb* cluster genes upon RA treatment. ES cells were treated with RA for different time points as indicated. Total RNAs were extracted from these cells and then subjected to RT-qPCR analysis using an Applied Biosystems' custom TaqMan array card. *Hoxa1* was the first *Hox* gene to be induced by RA. Compared with *Hoxa1*, the induction of *Hoxb1* was much slower within the first 6 h of RA treatment. The blue boxes indicate the first three RA induction time points. (*C*) Cdk9 is recruited to both the *Hoxa1* and *Hoxb1* gene promoters. Cdk9 ChIP was performed to measure its enrichment on *Hoxa1* and *Hoxb1* after RA treatment. A hemoglobin gene, *Hba* (Hemo), serves as a nontranscribed control gene. (*D*) ELL2 RNAi inhibits the induction of *Hoxa1* and *Hoxb1* by RA. shRNA targeting ELL2 or nontargeting shRNA (NonT) was introduced by lentiviral infection for 3 d before RA treatment. (*E*) Knockdown of ELL2 reduces Pol II occupancy at *Hoxa1* and *Hoxb1* after 6 h of RA treatment. Pol II occupancy was assayed by ChIP at the start site of transcription and in the ORF of *Hoxa1* and *Hoxb1* in RA-induced cells. Pol II is reduced in the ORF of both *Hoxa1* and *Hoxb1*, and *Hoxb1* also shows dramatically reduced levels of Pol II at its promoter after ELL2 RNAi. The *Hoxa1* promoter, but not the *Hoxb1* promoter, has prebound Pol II before RA treatment (see Fig. 2; Supplemental Fig. S4). Error bars represent the standard deviation.

genes rapidly induced within their respective clusters, followed more slowly by other members of the clusters, in general agreement with the collinearity of expression that occurs during normal embryonic development (Duboule and Dolle 1989; Graham et al. 1989; McGinnis and Krumlauf 1992). The Hoxa1 and Hoxb1 paralogs functionally synergize in regulating the hindbrain pattern formation and cranial nerve patterning (Gavalas et al. 2001). During normal mouse development, *Hoxa1* is the first *Hox* gene expressed in neural tissue directly induced by RA through a 3′ RA response element (RARE). It is closely followed by RA-mediated induction of *Hoxb1* through a similar 3′ RARE. Hoxa1 also participates in the proper activation of *Hoxb1* by binding to *Hoxb1*'s 5′ autoregulatory element (*ARE*), and Hoxb1 further stimulates transcription of its own gene (Popperl et al. 1995; Studer et al. 1998). Indeed, when looking within the first 6-h window of RA treatment of mouse ES cells, we observe that *Hoxa1* is induced more rapidly than *Hoxb1*, mirroring their normal kinetics of induction in neural development (Fig. 3A,B, blue boxes). The more rapid induction of the *Hoxa1* locus compared with *Hoxb1* could result from the presence of paused Pol II before RA treatment.

Since SEC was only detected at the *Hoxa1* promoter, and not at the *Hoxb1* promoter, we asked whether the Pol II CTD kinase Cdk9 was also differentially localized to these two genes at early induction time points. Direct comparisons of Cdk9 at *Hoxa1* and *Hoxb1* show that Cdk9 is recruited to both genes as early as 6 h and has increased occupancy at 18 h (Fig. 3C). When ES cells are induced with RA for various time points, in the presence or absence of the Cdk9 inhibitor flavopiridol (Chao and Price 2001), the induction of both *Hoxa1* and *Hoxb1* are diminished (Supplemental Fig. S5). This indicates that Cdk9 is required for the activation of both genes, even though the kinetics of their induction differ. This indicates that the recruitment of P-TEFb within SEC,

specifically to *Hoxa1*, functions in its rapid induction. In support of this statement, ELL2 RNAi also reduces the induction of *Hoxa1* (Fig. 3D; Supplemental Fig. S5). ELL2 is also required for the induction of *Hoxb1*; however, this observation could be explained by the requirement of the Hoxa1 protein for the full induction of *Hoxb1* (Studer et al. 1998). Accordingly, in the absence of ELL2 (ELL2 RNAi), we also observe the loss of Pol II in the body of the *Hoxa1* gene with no significant change or slight reduction in occupancy of Pol II at the *Hoxa1* promoter (Fig. 3E, top panel). Furthermore, since *Hoxb1* expression requires Hoxa1 activity and lacks prior paused Pol II in ES cells, in the absence of ELL2, we detect a loss in Pol II occupancy both at the promoter and in the body of the *Hoxb1* locus (Fig. 3E, bottom panel). Therefore, *Hoxa1* is likely to be a direct target of SEC, and *Hoxb1* is likely to be an indirect target of SEC. In summary, given the fact that *Hoxa1*, and not *Hoxb1*, possesses paused Pol II and recruits SEC upon a differentiation signal, we hypothesized that the recruitment of SEC to genes bearing paused Pol II is associated with rapid induction.

SEC is required for the induction of other rapidly induced genes in ES cells bearing paused Pol II

Using genome-wide approaches, we asked whether there were other genes that were regulated similarly to *Hoxa1*. We performed gene expression analyses of ES cells treated for 2–6 h with RA using Affymetrix expression arrays with probes representing ~30,000 genes (Fig. 4A). Sorting the gene expression data by fold expression over time showed that only a small number of genes demonstrated rapid and sustained induction over this time frame in a manner similar to *Hoxa1* (Fig. 4A,B). We found that 37 genes were rapidly induced at least twofold at 2, 4, and 6 h post-induction (Fig. 4A,B). Among these genes was *Hoxb1*, which our RT-qPCR data had shown was not as rapidly induced as *Hoxa1* (Fig. 3A,B). We therefore

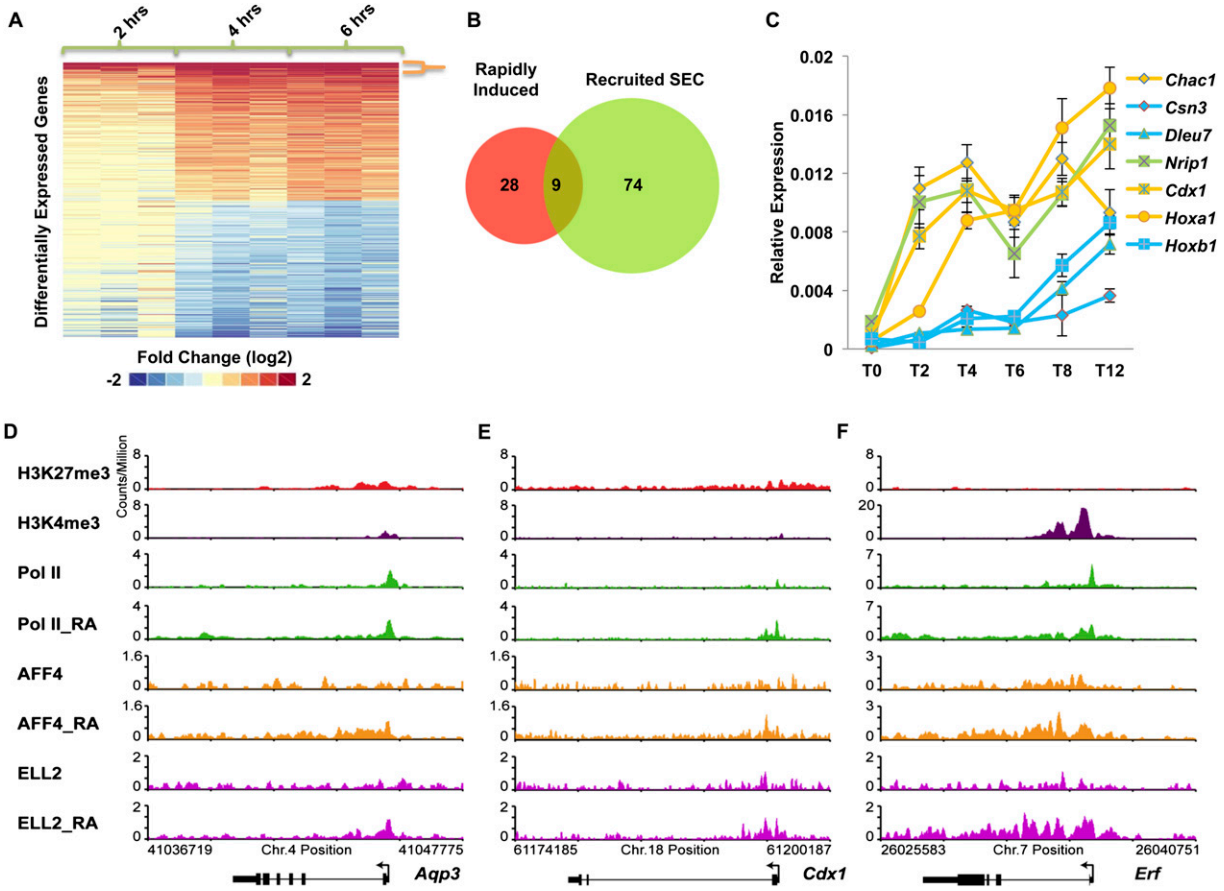


Figure 4. SEC regulates the rapid induction of RA signaling. (*A*, *left* panel) Microarray analyses of RA induction of ES cells as a function of time (2, 4, and 6 h) in biological triplicate. Differentially expressed probes (twofold or more) at the sixth hour post-induction compared with no induction are shown. Thirty-seven genes were induced twofold or more at each of the 2-, 4-, and 6-h time points (demarcated by the orange bracket). (*B*) Of the 37 induced genes, nine of them recruited SEC (ELL2 and AFF4). Newly recruited SEC genes are cobound at 6 h post-induction and are not cobound before induction. (*C*) RT-qPCR analysis of some of the induced genes identified from the microarray analysis. ES cells were treated with RA for the indicated time points, 0 h (T0), 2 h (T2), 4 h (T4), 6 h (T6), 8 h (T8), and 12 h (T12). Genes that recruit SEC are shown in yellow and genes that do not recruit SEC are shown in blue. *Nrip1*, which does not recruit SEC but is rapidly induced, is shown in green. Error bars represent the standard deviation. (*D–F*) Examples of ChIP-seq data showing SEC recruitment to RA-induced genes. Shown are *Aqp3*, *Cdx1*, and *Erf*, three of the nine genes from *B*.

performed RT-qPCR analyses of other genes from the top of this list to confirm their patterns of induction (Fig. 4C). These RT-qPCR studies demonstrated that two of the genes identified by microarray, *Dleu7* and *Csn3* (Fig. 4C, blue), behaved similarly to *Hoxb1* and were not as rapidly induced as *Hoxa1*, while two others showed the kinetics of rapid induction similar to *Hoxa1* (Fig. 4C, yellow).

Many of the induced genes that recruited SEC had bivalent histone marks and paused Pol II prior to RA induction. Genome browser track files for some examples are shown in Figure 4, D–F. Regardless of whether genes are rapidly or more slowly induced, Cdk9 was recruited and required for their induction (Supplemental Fig. S7). This analysis indicates that several genes that recruit SEC with Cdk9 respond more rapidly and uniformly to developmental signals than genes recruiting Cdk9 without SEC. However, the existence of genes like *Nrip1*, which is induced with similar kinetics to *Hoxa1* (Fig. 4C, green) but does not recruit SEC, suggests that while SEC is a major form of the Cdk9 complexes recruited to genes for rapid gene activation, other pathways to rapid gene activation are also possible.

We tested for the presence of another P-TEFb interactor, Brd4, on these genes and demonstrated that although Brd4 is recruited to these loci upon RA induction (Fig. 5A), its reduction by RNAi has very little to do with their activation by RA, except for the *Aqp3* gene (Fig. 5B,C). This observation suggests that although Brd4 is also recruited to those SEC target sites, it might not play a major role for their activation (Fig. 5A–C). Perhaps, as in the case of HIV-1 transcriptional regulation, Brd4 has

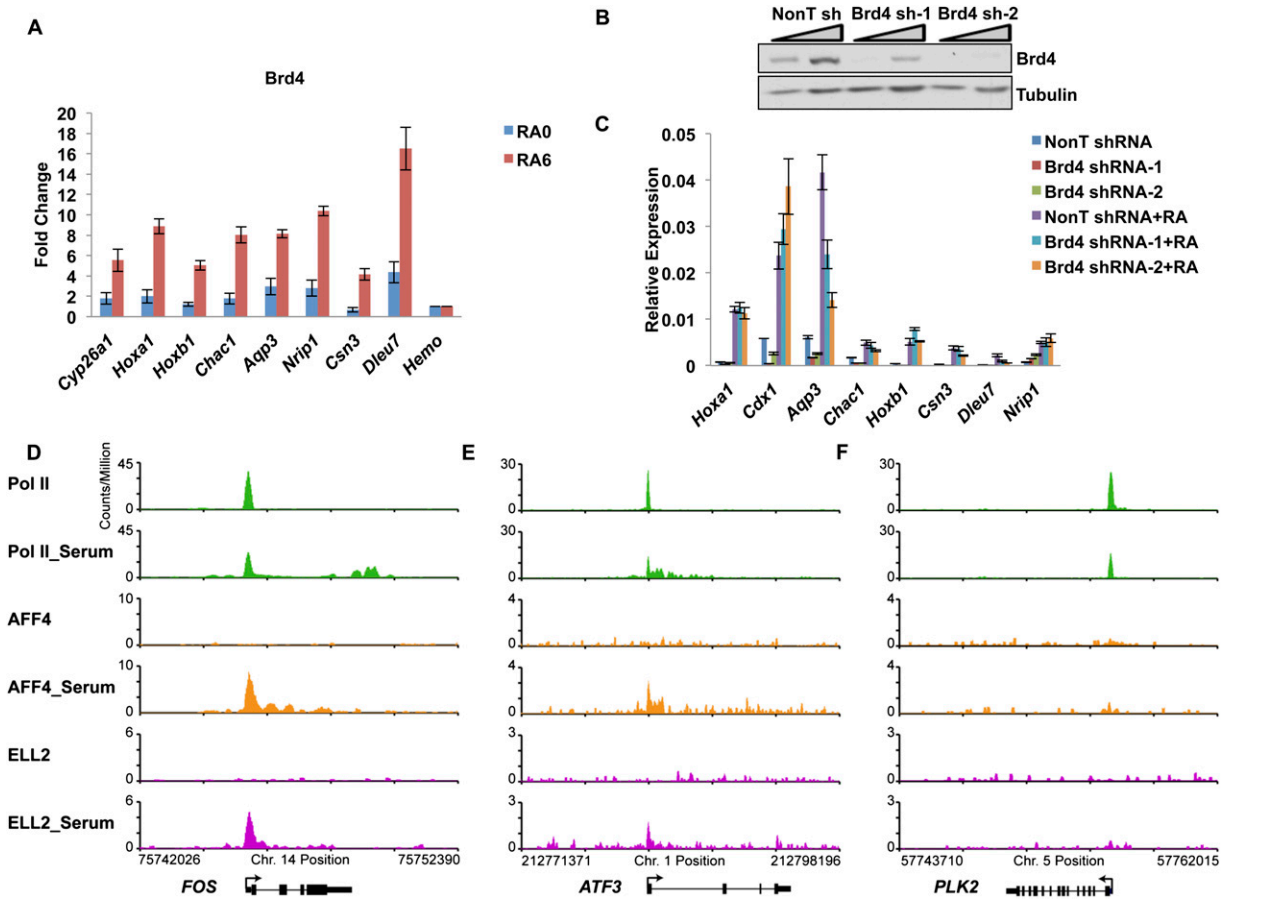


Figure 5. Brd4 is broadly present, but not broadly required, for RA induction of genes. (*A*) ChIP of Brd4 at RA-6-induced genes. Brd4 levels significantly increase at all RA-6-induced genes tested. The *Hba* gene serves as a nontranscribed control gene. (*B*) shRNA-mediated knockdown of Brd4. Two different shRNA constructs targeting Brd4 and a nontargeting shRNA (NonT) were introduced by lentiviral infection for 3 d before RA treatment. Brd4 levels were significantly reduced by Western analysis. Triangles indicate titrations of cell extracts. Tubulin serves as a loading control. (*C*) Induction of genes with RA is not broadly affected by Brd4 knockdown. Several genes, identified in Figure 4 as rapidly induced, were assayed for expression levels before and after RA treatment. Only *Aqp3* showed a significant decrease in its induction. Error bars represent the standard deviation. (*D–F*) SEC is recruited to the immediate early genes in HCT-116 cells after serum stimulation, genes previously identified as regulated by Brd4-containing P-TEFb complexes. HCT-116 cells were starved for 40 h before the 30-min add-back of serum. (*D–F*) Genome browser track files of three serum-induced genes, which were previously shown to be rapidly induced after serum stimulation (Donner et al. 2010). SEC is recruited to *ATF3* and *FOS* concomitant with release of paused Pol II into the gene body. *PLK2 is* shown for comparison as a gene induced by serum that does not recruit SEC.

a role in maintaining basal levels of transcription, but not in the activation of these genes (Yang et al. 2005).

SEC is also required for the rapid transcriptional induction of many growth-related immediate early genes in human cells

Given the small number of RA-induced genes in the mouse ES system, we sought another system to determine to what degree SEC regulates rapid transcriptional responses to environmental signals. Therefore, we investigated the role of SEC in the induction of genes in response to serum in human cells (Fig. 5D–F; Supplemental Fig. S8). The immediate early genes induced by growth factors are some of the best-characterized genes regulated at the level of the release of paused Pol II (Simone et al. 2001; Kong et al. 2005). We performed ChIP-seq of SEC and Pol II in HCT-116 cells before and after serum stimulation. SEC components are also enriched at the TSS in HCT-116 cells, consistent with their distribution in ES cells (Fig. 1E, Supplemental Fig. S8A). SEC was newly recruited to 55 genes within 30 min of serum stimulation (Supplemental Fig. S8B). Similar to what we observed in ES cells (Fig. 1E), genes bound by AFF4 and ELL2 showed higher levels of expression than those that lacked SEC (Supplemental Fig. S8C,D). Previous gene expression analysis of serum-inducible genes in HCT-116 cells identified 29 genes that were up-regulated twofold or more within 30 min of serum stimulation (Donner et al. 2010), 12 of which recruited both AFF4 and ELL2. We also performed RNA-seq analysis in these cells in the presence and absence of serum stimulation and identified 66 genes, which were induced above twofold, including 26 out of the 29 genes identified by Donner et al. (2010). To more precisely characterize the induction kinetics of these genes, we performed RT-qPCR on 17 serum-induced genes at different times after serum stimulation (Supplemental Fig. S8E). As we had seen with RA induction, serum-responsive genes were induced at varying rates, with SEC recruitment frequently occurring on the most rapidly induced genes (Supplemental Fig. S8D,E). Thus, SEC appears to be one of the major factors in the rapid release of paused Pol II in response to developmental and environmental stimuli.

Dynamic transcriptional induction requiring SEC without the presence of paused Pol II

To date, published studies indicate that paused Pol II functions in the rapid and robust induction of many developmentally regulated genes (Rougvie and Lis 1988; Muse et al. 2007; Zeitlinger et al. 2007; Nechaev and Adelman 2008; Boettiger and Levine 2009). However, our genome-wide expression and ChIP-seq data identified one gene that is extremely rapidly induced by RA: the *Cyp26a1* gene (Fig. 6). *Cyp26a1* encodes a cytochrome P450 that metabolizes RA (Duester 2008). The *Cyp26a1* gene bears several RAREs in its promoter and is known to be one of the most rapidly induced genes after exposure to RA (Alexander et al. 2009). Loss of *Cyp26a1* is toxic to development in mice, but this toxicity can be rescued by the loss of RA receptor γ (RARγ) (Abu-Abed et al. 2001; Sakai et al. 2001). While the *Cyp26a1* gene appears to have high levels of H3K27 trimethylation, it contains very low levels of H3K4 trimethylation compared with *Hoxa1* (see Figs. 2A,B, 6A). Also, this gene lacks paused Pol II in the untreated ES cells (Fig. 6A). After RA addition, Pol II and SEC are recruited to *Cyp26a1* by 6 h post-induction (Fig. 6A). In mouse ES cells, *Cyp26a1* is more rapidly induced when compared with *Hoxa1* and *Hoxb1* (Figs. 4C, 6B). Knockdown of ELL2 by shRNA treatment causes a reduction in *Cyp26a1* activation and also affects the recruitment of Pol II in its promoter and gene body (Fig. 6C,D), while flavopiridol completely eliminates *Cyp26a1* induction, indicating that this gene requires Cdk9 for its rapid induction by RA treatment (Supplemental Fig. S9). Furthermore, reduction of the Brd4 level by RNAi did not significantly affect *Cyp26a1* induction, suggesting that it is the SEC version of P-TEFb that regulates this gene. The dynamic induction of *Cyp26a1* without pre-existing paused Pol II suggests that there are other mechanisms for rapid induction of transcription during early development, which involves SEC.

Discussion

Our study reports that genes recruiting SEC in human and mouse cells are capable of rapidly responding in a dynamic manner to developmental cues. Although the presence of paused Pol II has been associated with rapid transcriptional induction, the response to some of the genes we identified in our study does not require the presence of paused Pol II. These findings are supported by the following observations. (1) *Hoxa1*, but not *Hoxb1*, contains paused Pol II on its promoter-proximal region and, upon developmental cues, SEC is recruited to *Hoxa1*, but not *Hoxb1*. (2) *Hoxa1* is induced more rapidly than *Hoxb1* after RA induction in ES cells. (3) Genome-wide analyses in ES cells identified a set of rapidly induced genes that contain paused Pol II, some of which also recruit SEC, and transcriptionally respond in a relatively uniform manner to a differentiation signal. (4) SEC is recruited to some of the immediate early genes upon serum stimulation in human cells, suggesting that SEC is one of the regulators of rapidly induced genes in different cellular contexts. (5) One rapidly induced gene, *Cyp26a1*, does not possess paused Pol II prior to induction, but still requires SEC for its rapid and dynamic activation. Therefore, while rapid induction is a key function of SEC, the function of the paused Pol II state is not simply to mediate rapid induction. It is likely that paused Pol II has some additional role(s) in developmentally controlled gene expression, as genes containing paused Pol II at their promoter-proximal regions transcriptionally respond in a well-regulated and uniform manner upon induction.

Genome-wide occupancy of SEC components

The SEC was identified through the purification of several of the MLL chimeric proteins that mediate leukemogenesis in mixed-lineage leukemias (Lin et al. 2010). SEC contains two classes of elongation factors, ELL

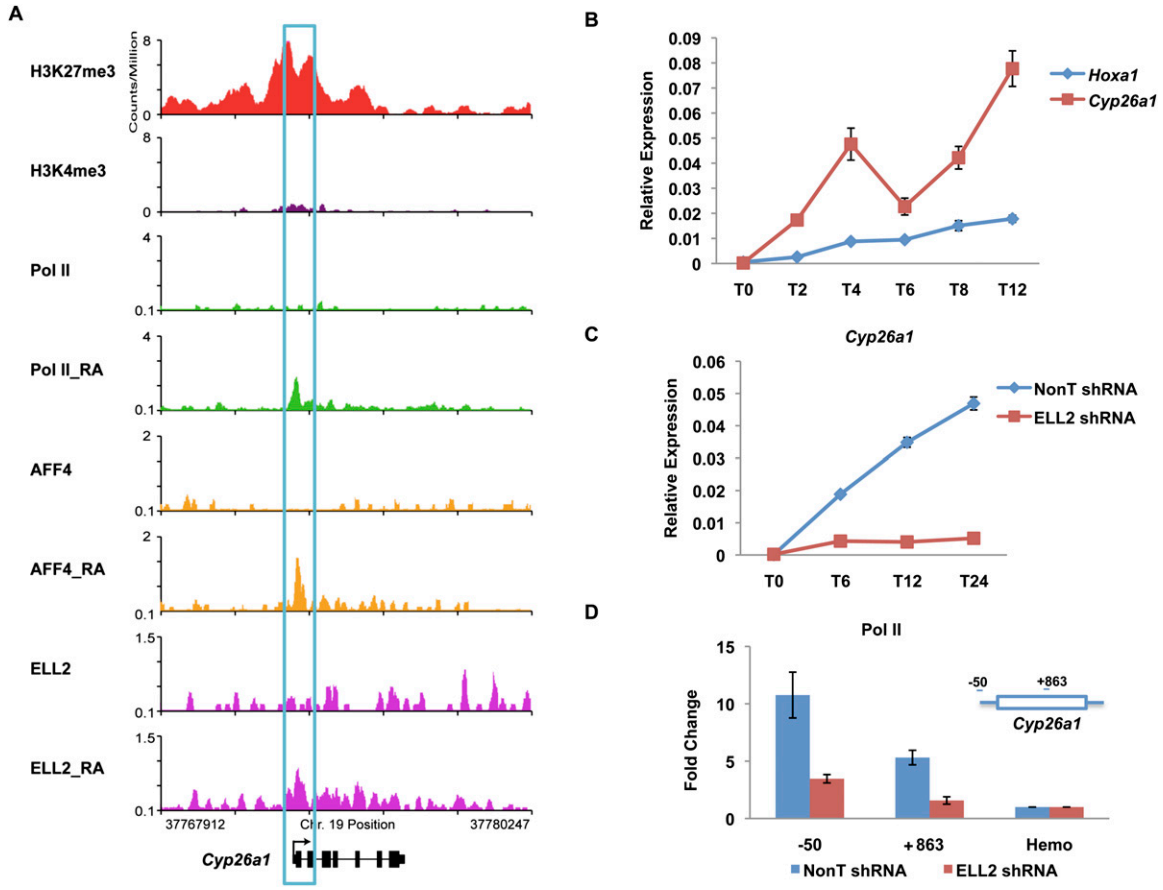


Figure 6. The rapid induction of *Cyp26a1* does not involve preloaded Pol II. (*A*) Pol II, H3K4me3, and H3K27me3 occupancy analysis of the *Cyp26a1* gene before RA induction. Before RA treatment, the *Cyp26a1* promoter is significantly enriched for H3K27me3, with lower levels of H3K4me3. However, there is no detectable Pol II on the promoter. AFF4, ELL2, and Pol II are newly recruited to the *Cyp26a1* gene promoter upon RA treatment. (*B*) RT-qPCR analysis of *Cyp26a1* mRNA levels upon RA treatment. ES cells were treated with RA for the indicated time points, 0 h (T0), 2 h (T2), 4 h (T4), 6 h (T6), 8 h (T8), and 12 h (T12). Total RNAs were extracted from these treated cell samples and then subjected to RT-qPCR analysis. (*C*) ELL2 RNAi inhibits the induction of *Cyp26a1* by RA. shRNA targeting ELL2 or a nontargeting shRNA (NonT) was introduced by lentiviral infection for 3 d before RA treatment. (*D*) Knockdown of ELL2 reduces Pol II occupancy at *Cyp26a1* after 24 h RA treatment. The *Hba* gene serves as a nontranscribed control gene. Error bars represent the standard deviation.

(represented by ELL1–3) and P-TEFb, which consists of Cdk9 and CycT1 or CycT2, which were isolated through biochemical approaches for the discovery of transcription elongation factors (Marshall and Price 1995; Shilatifard et al. 1996). The fact that ELL is a translocation partner of MLL and associates with other translocation partners of MLL such as AFF1, AFF4, ENL, and AF9 indicates that SEC function is important for the misregulation of key developmental genes like the *Hox* loci that are part of the leukemogenic process. Our genome-wide approaches sought to discover the normal developmental role of SEC using mouse ES cells before and after induction of differentiation.

SEC is recruited to genes with paused Pol II for rapid and coordinated transcriptional induction

Many developmentally regulated genes are marked by the presence of bivalent histone marks, the methylation of H3K4 and H3K27, DSIF/NELF, and paused Pol II at the TSS (Bernstein et al. 2006; Stock et al. 2007; Rahl et al. 2010). Since P-TEFb complexes such as the SEC are proposed to release paused Pol II via phosphorylation of the CTD and other general factors within the transcription complex, we asked whether SEC is recruited to these genes after induction of differentiation. We first focused on the *Hox* loci, because misregulation of *Hox* transcription is strongly implicated in leukemogenesis by MLL chimeras. The mammalian *Hox* genes exist in four clusters and are expressed in a collinear manner during development, with the most anterior *Hox* gene (e.g., *Hoxa1* in mammals) being expressed first from its location at the 3′ end of the cluster, and a more posteriorly expressed *Hox* gene (e.g., *Hoxa13*) expressed later during development from its location at the 5′ end of the cluster. Although a large number of developmentally regulated genes contain bivalent marks and paused Pol II at their promoters, we found that only a subset of *Hox* genes

followed this pattern. For example, four genes within the *Hoxa* cluster bear bivalent histone marks and paused Pol II at their promoters, but no *Hoxb* genes are marked in this manner in the ES cells (Fig. 2). Importantly, after induction of differentiation, *Hoxa1* was induced more rapidly than its paralog, *Hoxb1* (Fig. 3). Furthermore, the SEC was specifically recruited to *Hoxa1*, and not *Hoxb1*, suggesting that SEC releases paused Pol II for rapid induction of transcription during development. This mechanism helps to explain the more rapid induction and regulatory roles of *Hoxa1* compared with *Hoxb1* in early neural development (Alexander et al. 2009).

In order to extend these findings beyond the *Hox* loci, we undertook genome-wide analyses of expression and SEC occupancy before and after induction of differentiation with RA. We found additional examples of rapidly induced genes bearing paused Pol II at their promoter-proximal region that also recruited SEC, and many of these were among the most rapidly induced (Fig. 4). These findings were shown to be more general by studying the recruitment of SEC to the immediate early genes in HCT-116 cells after serum induction. The immediate early genes, including the *FOS*, *JUN*, and *EGR* families, are well characterized as genes containing paused Pol II and exhibiting rapid induction kinetics (Galbraith and Espinosa 2011). As with RA-induced genes, SEC was recruited to many of the most rapidly induced genes after serum stimulation (Fig. 5).

Dynamic and rapid transcriptional induction does not always require paused Pol II

Most of the rapidly induced genes contained paused Pol II in the undifferentiated ES cells. However, we were able to identify *Cyp26a1* as a gene that was even more rapidly induced without having prior Pol II occupancy, however, in an SEC-dependent manner. By all molecular characteristics, *Cyp26a1* in the undifferentiated condition appears to be relatively repressed in ES cells, as it bears H3K27 trimethylation, only modest levels of H3K4 methylation, and no detectable paused Pol II on its promoter when compared with other paused Pol II-regulated genes, such as *Hoxa1* (cf. Figs. 2 and 6).

Cyp26a1 encodes a cytochrome, P450, that metabolizes RA and is essential for development. One function of Cyp26a1 is to restrict the response to RA to the appropriate regions of the embryo. Not only is *Cyp26a1* extremely rapidly induced compared with other early RA response genes, but it is induced to much higher levels than the other very early RA-induced genes containing paused Pol II (Figs. 4, 6, 7).

The function of paused Pol II in modulating transcription of developmentally regulated genes

Our genome-wide analyses of RA-induced gene transcription and SEC recruitment identified three classes of genes, two of which require SEC for induction (Fig. 7). One class, which includes *Hoxb1*, lacks paused Pol II and does not recruit SEC upon induction (Fig. 7A). A second class, which includes *Hoxa1*, contains paused Pol II, recruits SEC, and is induced more rapidly than the first class (Fig. 7B). A third class, exemplified by *Cyp26a1*, recruits SEC, is induced just as rapidly as the second class, but to a greater extent than *Hoxa1*, yet lacks paused Pol II at its promoter-proximal region before induction, and requires SEC (Fig. 7C).

The *HSP70*, *FOS*, *JUN*, and *EGR* families of genes are well studied and rapidly induced, and contain paused Pol II in the unstimulated state, leading to the paradigm that rapid induction is the primary function of paused Pol II (Nechaev and Adelman 2008; Donner et al. 2010). However, paused Pol II is not present on *Cyp26a1* before its rapid induction to high levels of transcription, which suggests that paused Pol II is not a prerequisite for rapid induction, but rather facilitates coordinated and controlled induction. Studies in *Drosophila* have shown that developmentally regulated genes that have paused Pol II are activated in a synchronous manner, while developmentally regulated genes that lack paused Pol II have a more stochastic pattern of induction during development (Boettiger and Levine 2009; Levine 2011). Having preloaded Pol II and general transcription factors (GTFs) reduces the number of steps for productive transcription, and thus could result in a more equivalent and uniform way to induce gene expression. Genes such as *Cyp26a1*, while being required for proper development and being induced rapidly to high levels, may not need to be as precisely regulated at the earliest time points of induction.

Overall, our studies demonstrate that SEC is involved in many of the rapid and dynamic inductions of gene expression responses to developmental and environmental cues. P-TEFb was identified >15 years ago (Marshall and Price 1995) and was soon shown to be required for HIV transactivation (Mancebo et al. 1997; Zhu et al. 1997; Wei et al. 1998). Although the majority of P-TEFb is in the inactive Hexim1 complex, P-TEFb has also been shown to associate with a variety of factors that could help recruit it to chromatin in an active form (Bres et al. 2008; He and Zhou 2011). In this study, we show that the recently discovered SEC version of P-TEFb is a major regulator of rapidly induced genes in development. However, our genome-wide analyses indicate that not all rapidly activated genes require the SEC components investigated in this study. How the different P-TEFb complexes are recruited to regulate distinct sets of genes will be an important area of future investigations.

Materials and methods

ES cell culture

Mouse ES cells (KH2 and V6.5) were cultured under mouse ES complete medium on irradiated mouse embryonic fibroblasts (MEFs). For the shRNA knockdown analysis, viral supernatants were collected and concentrated after 48–72 h of packaging in 293T cells. ES cells were infected with concentrated lentiviral particles with polybrene at the concentration of 4 μg/mL. After 3 d of infection, the ES cells were treated with RA for 6 h and grown one passage off feeders before harvesting for RT-qPCR analysis. For the flavopiridol (Sigma catalog no. F3680) treatment, ES cells were cultured in feeder-free medium and plated at

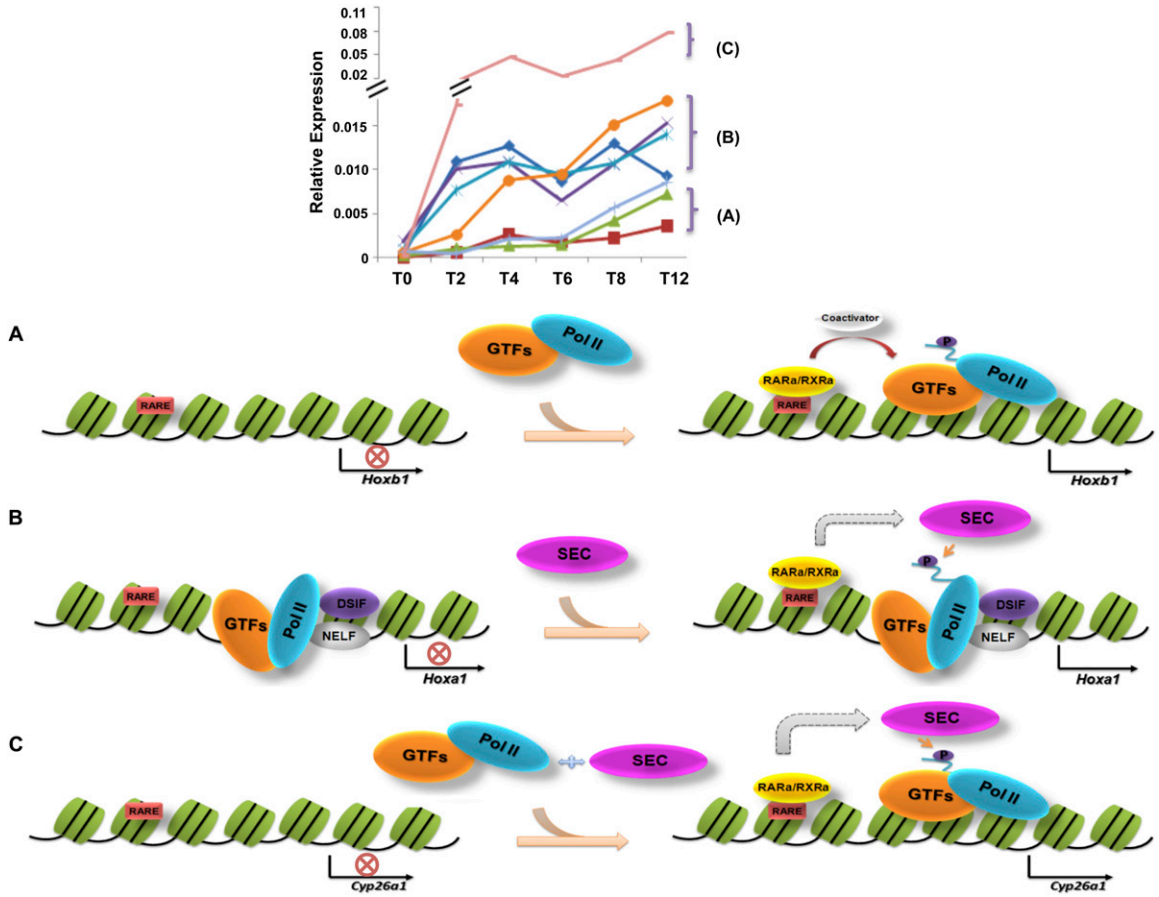


Figure 7. Diverse mechanisms for rapid activation of genes during development. The *top* panel shows that rapidly activated genes can be further subdivided into distinct categories, A–C. (*A*) The *Hoxb1* gene newly recruits Pol II and GTFs in a classical gene activation mechanism, where RAR/RXR binds in the presence of RA and, with the help of coactivators, recruits GTFs and Pol II. (*B*) Paused Pol II, with DSIF/NELF, is present at the TSS of developmentally regulated genes, such as *Hoxa1*. In the presence of RA, RAR/RXR recruits SEC to stimulate transcription elongation through phosphorylation of the DSIF/NELF and the Pol II CTD. (*C*) *Cyp26a1*, a developmentally regulated gene that lacks paused Pol II, is induced by RA in a SEC-dependent manner. All of the same factors are present after RA treatment as seen at *Hoxa1*, but *Cyp26a1* is induced to higher levels, suggesting that paused Pol II may serve to help regulate activation to equivalent levels.

a density of 4×10^5 cells per well in a six-well plate. For the ChIP-seq analysis, cells were grown under feeder-free medium ESGRO (Millipore).

HCT-116 cell culture, serum stimulation, and RNAi treatment

HCT-116 cells were grown in McCoy's 5A medium supplemented with 10% FBS. For serum stimulation, cells were first starved by washing cells twice in PBS, then culturing for 40 h in McCoy's 5A without serum. Cells were then either left untreated or treated with serum for 30 min before harvesting.

Antibodies

Anti-RNA Pol II antibodies were purchased from Covance (8WG16) and from Santa Cruz Biotechnologies (N-20), and H3K4 trimethylation (H3K4me3), AFF4, and ELL2 antibodies were generated in our laboratory as described previously (Lin et al. 2010). A fragment of human CDK9 (amino acids 204–372) was expressed as His tag fusion protein in pET-16b, purified on NTA-agarose according to Qiagen's protocol, and sent to Pocono Rabbit Farm and Laboratory for immunization into rabbits.

ChIP and gene expression analysis

ChIP was performed according to previously described protocols (Wang et al. 2009). RNA for microarray analysis was isolated from ES cells grown on feeder cells. After induction, cells were washed in PBS, trypsinized, and replated onto a fresh plate for 30 min. Unbound cells were harvested and total RNA was extracted with Trizol. For RT-qPCR, RNA was isolated with the RNeasy (Qiagen) kit, treated with DNase I, and repurified with RNeasy. *Hox* expression assays were purchased from Applied Biosystems in a custom TaqMan array card. Five reference controls were experimentally validated using geNorm. *Gapdh* and *Tbp* were found to be stably expressed and were used for normalization. Relative expression levels were determined using the comparative cycle threshold method. ChIP-seq and other data analyses can be found in the Supplemental Material. ChIP-seq and expression data have been deposited at the Gene Expression Omnibus (GEO) under the accession number GSE30268.

Acknowledgments

We thank Tari Parmely and the Stowers Institute Tissue Culture Facility for assistance with mouse ES cells and other cell culture needs, and the Molecular Biology Facility for Illumina sequencing and help with qPCR analysis. We are also grateful to Laura Shilatifard for editorial assistance. We thank Joaquin Espinosa for providing HCT-116 cells. This work was performed to fulfill, in part, requirements for C.L.'s and B.D.K.'s PhD thesis research as students registered with the Open University. R.K. and A.S. are supported in part by funds provided by the Stowers Institute for Medical Research. This study was also supported by funds provided by the Alex's Lemonade Stand Foundation and funds from the National Institute of Health (R01CA150265) to A.S.

References

Abu-Abed S, Dolle P, Metzger D, Beckett B, Chambon P, Petkovich M. 2001. The retinoic acid-metabolizing enzyme, CYP26A1, is essential for normal hindbrain patterning, vertebral identity, and development of posterior structures. *Genes Dev* **15:** 226–240.

Alexander T, Nolte C, Krumlauf R. 2009. Hox genes and segmentation of the hindbrain and axial skeleton. *Annu Rev Cell Dev Biol* **25:** 431–456.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315–326.

Boettiger AN, Levine M. 2009. Synchronous and stochastic patterns of gene activation in the *Drosophila* embryo. *Science* **325:** 471–473.

Bres V, Yoh SM, Jones KA. 2008. The multi-tasking P-TEFb complex. *Curr Opin Cell Biol* **20:** 334–340.

Chao SH, Price DH. 2001. Flavopiridol inactivates P-TEFb and blocks most RNA polymerase II transcription in vivo. *J Biol Chem* **276:** 31793–31799.

Cheng B, Price DH. 2007. Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. *J Biol Chem* **282:** 21901–21912.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848.

Donner AJ, Ebmeier CC, Taatjes DJ, Espinosa JM. 2010. CDK8 is a positive regulator of transcriptional elongation within the serum response network. *Nat Struct Mol Biol* **17:** 194–201.

Duboule D, Dolle P. 1989. The structural and functional organization of the murine HOX gene family resembles that of *Drosophila* homeotic genes. *EMBO J* **8:** 1497–1505.

Duester G. 2008. Retinoic acid synthesis and signaling during early organogenesis. *Cell* **134:** 921–931.

Fuda NJ, Ardehali MB, Lis JT. 2009. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461:** 186–192.

Galbraith MD, Espinosa JM. 2011. Lessons on transcriptional control from the serum response network. *Curr Opin Genet Dev* **21:** 160–166.

Gavalas A, Trainor P, Ariza-McNaughton L, Krumlauf R. 2001. Synergy between Hoxa1 and Hoxb1: the relationship between arch patterning and the generation of cranial neural crest. *Development* **128:** 3017–3027.

Graham A, Papalopulu N, Krumlauf R. 1989. The murine and *Drosophila* homeobox gene complexes have common features of organization and expression. *Cell* **57:** 367–378.

He N, Zhou Q. 2011. New insights into the control of HIV-1 transcription: when Tat meets the 7SK snRNP and super elongation complex (SEC). *J Neuroimmune Pharmacol* **6:** 260–268.

He N, Liu M, Hsu J, Xue Y, Chou S, Burlingame A, Krogan NJ, Alber T, Zhou Q. 2010. HIV-1 Tat and host AFF4 recruit two transcription elongation factors into a bifunctional complex for coordinated activation of HIV-1 transcription. *Mol Cell* **38:** 428–438.

Jones KA, Peterlin BM. 1994. Control of RNA initiation and elongation at the HIV-1 promoter. *Annu Rev Biochem* **63:** 717–743.

Kong SE, Banks CA, Shilatifard A, Conaway JW, Conaway RC. 2005. ELL-associated factors 1 and 2 are positive regulators of RNA polymerase II elongation factor ELL. *Proc Natl Acad Sci* **102:** 10094–10098.

Levine M. 2011. Paused RNA polymerase II as a developmental checkpoint. *Cell* **145:** 502–511.

Lin C, Smith ER, Takahashi H, Lai KC, Martin-Brown S, Florens L, Washburn MP, Conaway JW, Conaway RC, Shilatifard A. 2010. AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia. *Mol Cell* **37:** 429–437.

Mancebo HS, Lee G, Flygare J, Tomassini J, Luu P, Zhu Y, Peng J, Blau C, Hazuda D, Price D, et al. 1997. P-TEFb kinase is required for HIV Tat transcriptional activation in vivo and in vitro. *Genes Dev* **11:** 2633–2644.

Marshall NF, Price DH. 1995. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J Biol Chem* **270:** 12335–12338.

Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J, et al. 2008. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134:** 521–533.

McGinnis W, Krumlauf R. 1992. Homeobox genes and axial patterning. *Cell* **68:** 283–302.

Mohan M, Lin C, Guest E, Shilatifard A. 2010. Licensed to elongate: a molecular mechanism for MLL-based leukaemogenesis. *Nat Rev Cancer* **10:** 721–728.

Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, Grissom SF, Zeitlinger J, Adelman K. 2007. RNA polymerase is poised for activation across the genome. *Nat Genet* **39:** 1507–1511.

Nechaev S, Adelman K. 2008. Promoter-proximal Pol II: when stalling speeds things up. *Cell Cycle* **7:** 1539–1544.

Nechaev S, Adelman K. 2011. Pol II waiting in the starting gates: regulating the transition from transcription initiation into productive elongation. *Biochim Biophys Acta* **1809:** 34–45.

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K. 2010. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327:** 335–338.

Peterlin BM, Price DH. 2006. Controlling the elongation phase of transcription with P-TEFb. *Mol Cell* **23:** 297–305.

Popperl H, Bienz M, Studer M, Chan SK, Aparicio S, Brenner S, Mann RS, Krumlauf R. 1995. Segmental expression of Hoxb-1 is controlled by a highly conserved autoregulatory loop dependent upon exd/pbx. *Cell* **81:** 1031–1042.

Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA. 2010. c-Myc regulates transcriptional pause release. *Cell* **141:** 432–445.

Rougvie AE, Lis JT. 1988. The RNA polymerase II molecule at the 5′ end of the uninduced hsp70 gene of *D. melanogaster* is transcriptionally engaged. *Cell* **54:** 795–804.

Sakai Y, Meno C, Fujii H, Nishino J, Shiratori H, Saijoh Y, Rossant J, Hamada H. 2001. The retinoic acid-inactivating

enzyme CYP26 is essential for establishing an uneven distribution of retinoic acid along the anterio-posterior axis within the mouse embryo. *Genes Dev* **15:** 213–225.

Shilatifard A. 1998. Factors regulating the transcriptional elongation activity of RNA polymerase II. *FASEB J* **12:** 1437–1446.

Shilatifard A, Lane WS, Jackson KW, Conaway RC, Conaway JW. 1996. An RNA polymerase II elongation factor encoded by the human ELL gene. *Science* **271:** 1873–1876.

Shilatifard A, Conaway RC, Conaway JW. 2003. The RNA polymerase II elongation complex. *Annu Rev Biochem* **72:** 693–715.

Simone F, Polak PE, Kaberlein JJ, Luo RT, Levitan DA, Thirman MJ. 2001. EAF1, a novel ELL-associated factor that is delocalized by expression of the MLL–ELL fusion protein. *Blood* **98:** 201–209.

Sims RJ 3rd, Belotserkovskaya R, Reinberg D. 2004. Elongation by RNA polymerase II: the short and long of it. *Genes Dev* **18:** 2437–2468.

Smith E, Lin C, Shilatifard A. 2011. The super elongation complex (SEC) and MLL in development and disease. *Genes Dev* **25:** 661–672.

Sobhian B, Laguette N, Yatim A, Nakamura M, Levy Y, Kiernan R, Benkirane M. 2010. HIV-1 Tat assembles a multifunctional transcription elongation complex and stably associates with the 7SK snRNP. *Mol Cell* **38:** 439–451.

Stock JK, Giadrossi S, Casanova M, Brookes E, Vidal M, Koseki H, Brockdorff N, Fisher AG, Pombo A. 2007. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. *Nat Cell Biol* **9:** 1428–1435.

Studer M, Gavalas A, Marshall H, Ariza-McNaughton L, Rijli FM, Chambon P, Krumlauf R. 1998. Genetic interactions between Hoxa1 and Hoxb1 reveal new roles in regulation of early hindbrain patterning. *Development* **125:** 1025–1036.

Thirman MJ, Levitan DA, Kobayashi H, Simon MC, Rowley JD. 1994. Cloning of ELL, a gene that fuses to MLL in a t(11;19)(q23;p13.1) in acute myeloid leukemia. *Proc Natl Acad Sci* **91:** 12110–12114.

Wang P, Lin C, Smith ER, Guo H, Sanderson BW, Wu M, Gogol M, Alexander T, Seidel C, Wiedemann LM, et al. 2009. Global analysis of H3K4 methylation defines MLL family member targets and points to a role for MLL1-mediated H3K4 methylation in the regulation of transcriptional initiation by RNA polymerase II. *Mol Cell Biol* **29:** 6074–6085.

Wei P, Garber ME, Fang SM, Fischer WH, Jones KA. 1998. A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* **92:** 451–462.

Workman JL, Kingston RE. 1998. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu Rev Biochem* **67:** 545–579.

Yamaguchi Y, Takagi T, Wada T, Yano K, Furuya A, Sugimoto S, Hasegawa J, Handa H. 1999. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97:** 41–51.

Yang Z, Yik JH, Chen R, He N, Jang MK, Ozato K, Zhou Q. 2005. Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Mol Cell* **19:** 535–545.

Zeitlinger J, Stark A, Kellis M, Hong JW, Nechaev S, Adelman K, Levine M, Young RA. 2007. RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nat Genet* **39:** 1512–1516.

Zhou Q, Yik JH. 2006. The Yin and Yang of P-TEFb regulation: implications for human immunodeficiency virus gene expression and global control of cell growth and differentiation. *Microbiol Mol Biol Rev* **70:** 646–659.

Zhu Y, Pe'ery T, Peng J, Ramanathan Y, Marshall N, Marshall T, Amendt B, Mathews MB, Price DH. 1997. Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. *Genes Dev* **11:** 2622–2632.

Zippo A, Serafini R, Rocchigiani M, Pennacchini S, Krepelova A, Oliviero S. 2009. Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation. *Cell* **138:** 1122–1136.