

## LETTERS

# The DNA-encoded nucleosome organization of a eukaryotic genome

Noam Kaplan<sup>1\*</sup>, Irene K. Moore<sup>3\*</sup>, Yvonne Fondufe-Mittendorf<sup>3</sup>, Andrea J. Gossett<sup>4</sup>, Desiree Tillo<sup>5</sup>, Yair Field<sup>1</sup>, Emily M. LeProust<sup>6</sup>, Timothy R. Hughes<sup>5,7,8</sup>, Jason D. Lieb<sup>4</sup>, Jonathan Widom<sup>3</sup> & Eran Segal<sup>1,2</sup>

Nucleosome organization is critical for gene regulation<sup>1</sup>. In living cells this organization is determined by multiple factors, including the action of chromatin remodellers<sup>2</sup>, competition with site-specific DNA-binding proteins<sup>3</sup>, and the DNA sequence preferences of the nucleosomes themselves<sup>4–8</sup>. However, it has been difficult to estimate the relative importance of each of these mechanisms *in vivo*<sup>7,9–11</sup>, because *in vivo* nucleosome maps reflect the combined action of all influencing factors. Here we determine the importance of nucleosome DNA sequence preferences experimentally by measuring the genome-wide occupancy of nucleosomes assembled on purified yeast genomic DNA. The resulting map, in which nucleosome occupancy is governed only by the intrinsic sequence preferences of nucleosomes, is similar to *in vivo* nucleosome maps generated in three different growth conditions. *In vitro*, nucleosome depletion is evident at many transcription factor binding sites and around gene start and end sites, indicating that nucleosome depletion at these sites *in vivo* is partly encoded in the genome. We confirm these results with a micrococcal nuclease-independent experiment that measures the relative affinity of nucleosomes for ~40,000 double-stranded 150-base-pair oligonucleotides. Using our *in vitro* data, we devise a computational model of nucleosome sequence preferences that is significantly correlated with *in vivo* nucleosome occupancy in *Caenorhabditis elegans*. Our results indicate that the intrinsic DNA sequence preferences of nucleosomes have a central role in determining the organization of nucleosomes *in vivo*.

We sought to establish the extent to which the DNA sequence determines nucleosome organization in living cells. Our strategy, previously used by others for two yeast promoters<sup>12</sup>, was to compare *in vivo* nucleosome organization with that obtained by an *in vitro* assembly procedure using only purified nucleosomes and purified DNA. To obtain a genome-wide map of nucleosome occupancy governed solely by nucleosome sequence preferences, we purified chicken erythrocyte histone octamers and assembled them on purified yeast genomic DNA by salt gradient dialysis<sup>13</sup>. We then isolated mononucleosomes by micrococcal nuclease digestion, and used parallel sequencing to determine nucleosome positions. We performed two independent experiments, resulting in ~10,000,000 DNA sequence reads that map uniquely to the yeast genome. For comparison to *in vivo* nucleosome positions, we isolated mononucleosomes from living cells<sup>5,7,9,10</sup>, and obtained ~25,000,000 sequence reads from 6 independent experiments. For each map, we determined the average nucleosome occupancy at every base pair, calculated as the log-ratio between the number of reads that cover that base

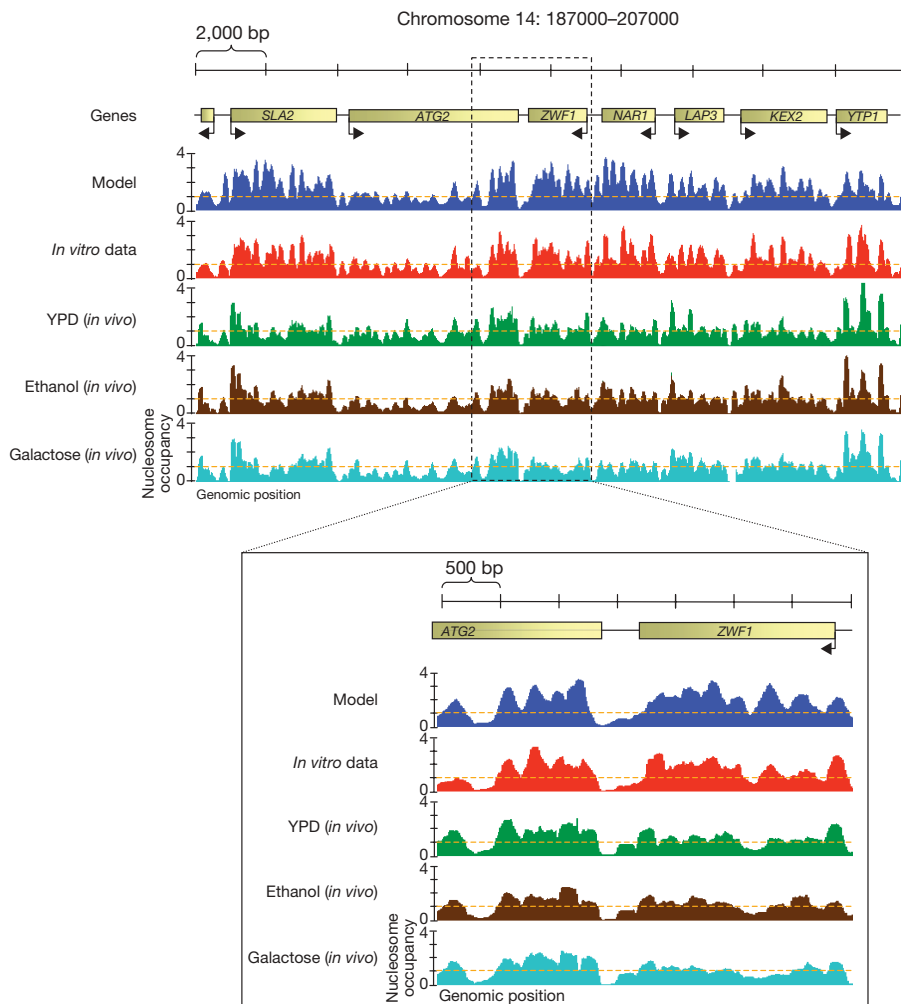
pair and the genome-wide average coverage per base pair (see Methods).

The nucleosome organizations of the *in vitro* and *in vivo* maps are notably similar, although not identical (Fig. 1), with a correlation of 0.74 between the nucleosome occupancy per base pair (Fig. 2a). On the scale of individual nucleosomes, the *in vitro* data separate regions that are enriched in nucleosomes *in vivo* from regions depleted of nucleosomes with high accuracy (Supplementary Fig. 1). Similarly, we found a significant correspondence between the positions of stable nucleosomes in the two maps (Supplementary Fig. 2). This high degree of similarity between the maps indicates that nucleosome sequence preferences have a dominant role in determining *in vivo* nucleosome organization.

The correlation between the maps is not uniform across the genome. We found a higher correlation between the maps at non-promoter intergenic regions located at ends of convergently transcribed genes (0.83) and a lower correlation at promoter (0.69) and coding (0.69) regions. In addition, the depletion level *in vivo* relative to that measured *in vitro* at coding regions increases with the expression level of the associated genes (Fig. 2b). These results indicate that transcription factors, chromatin regulators and active transcription influence the resulting nucleosome organization *in vivo*.

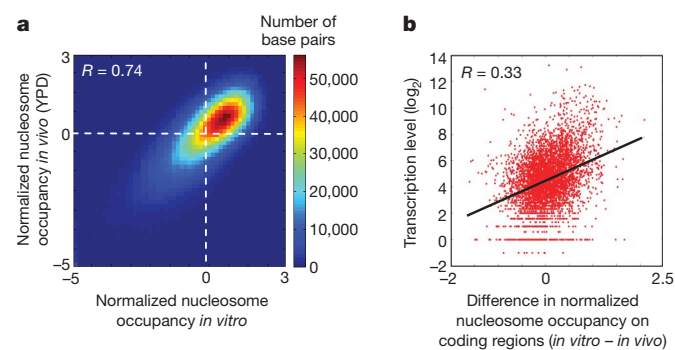
Because the nucleosome organization *in vitro* is determined only by the DNA sequence, we asked whether we could derive rules that are predictive of nucleosome positioning and occupancy. For each of the 1,024 sequences of length 5 base pairs, we computed the average nucleosome occupancy of that sequence across all of its instances in the genome. We found a near perfect agreement (correlation of 0.98) between the average occupancy of these 5-base-pair sequences *in vivo* and *in vitro* (Fig. 3a). Many 5-base-pair sequences showed strong preferences for nucleosome-enriched or nucleosome-depleted regions. For example, AAAAA has the lowest average nucleosome occupancy both *in vivo* and *in vitro*, consistent with the reduced nucleosome affinity that poly(dA-dT) sequences have *in vitro*<sup>14</sup>, and with the nucleosome depletion observed over poly(dA-dT) sequences *in vivo*<sup>9,15</sup>. Consistent with previous reports<sup>4,5,11,16</sup>, we also found clear ~10-bp periodicities of dinucleotides along the nucleosome length, both *in vitro* and *in vivo* (Fig. 3b, c). Notably, the dynamic range of these periodicities is greater *in vitro*, suggesting that the fraction of nucleosomes positioned by these periodic motifs *in vitro* is greater than that *in vivo*. This difference may be due to the action of chromatin remodellers and transcription factors *in vivo*, which may cause nucleosomes to deviate from the locations dictated by the nucleosome sequence preferences. The higher

<sup>1</sup>Department of Computer Science and Applied Mathematics, <sup>2</sup>Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel. <sup>3</sup>Department of Biochemistry, Molecular Biology, and Cell Biology, Northwestern University, 2153 Sheridan Road, Evanston, Illinois 60208, USA. <sup>4</sup>Department of Biology, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. <sup>5</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. <sup>6</sup>Agilent Technologies Inc., Genomics—LSSU, 5301 Stevens Creek Boulevard, MS 3L/MT Santa Clara, California 95051, USA. <sup>7</sup>Terrence Donnelly Centre for Cellular & Biomolecular Research, <sup>8</sup>Banting and Best Department of Medical Research, 160 College Street, Toronto, Ontario M5S 3E1, Canada. \*These authors contributed equally to this work.



**Figure 1 | The intrinsic DNA-encoded nucleosome organization at a typical genomic region.** Shown are the four different maps of nucleosome occupancy measured in this study for a typical 20,000-bp-long genomic region: the *in vitro* map, which reflects only the intrinsic nucleosome sequence preferences, and *in vivo* yeast maps for three different growth conditions (YPD, ethanol and galactose). Each track plots the measured nucleosome occupancy per base pair, computed by summing all of the nucleosome reads obtained in that experiment, and dividing that number by the average number of reads per base pair across the genome. The line of  $y = 1$  thus represents the genome-wide average and is shown as a dashed orange line. The average nucleosome occupancy predictions from our model are shown in blue.

concentration of nucleosomes *in vivo* relative to the concentration used to create our *in vitro* map may also contribute to this difference, because higher nucleosome concentrations generally increase the



**Figure 2 | *In vitro* and *in vivo* maps are highly similar.** **a**, Shown is a density dot plot comparison of the normalized nucleosome occupancy per base pair in the *in vitro* ( $x$  axis) and *in vivo* ( $y$  axis) maps (see Methods). Values above zero indicate nucleosome enrichment relative to the genome-wide average. The colour of each point represents the number of base pairs that map to that point in the graph. The Pearson correlation between the maps is indicated. **b**, Nucleosome depletion *in vivo* relative to *in vitro* over coding regions increases with the expression level of associated genes. Shown is a dot plot comparison between the expression level of every yeast gene (measured in ref. 26) and the difference between the average normalized nucleosome occupancy of the coding region of that gene in the *in vitro* map compared with the *in vivo* map (that is, higher values indicate larger nucleosome depletion *in vivo* relative to *in vitro*). The Pearson correlation of the dot plot is indicated.

contribution of non-specific binding, thus diminishing the contribution of the  $\sim 10$ -bp sequence periodicities. Nevertheless, the conservation of the  $\sim 10$ -bp dinucleotide periodicities and the near-identity of 5-base-pair nucleosome occupancies demonstrate that nucleosomes have clear sequence preferences that are highly similar *in vitro* and *in vivo*.

To test whether general sequence-based rules can be derived from our *in vitro* data and be used to predict nucleosome occupancy *in vivo*, we constructed a simple probabilistic model based on both the global preferences over sequences of length 5 and the position-dependent dinucleotide preferences<sup>5,17</sup>, which scores the nucleosome formation potential of every 147-bp sequence. Importantly, this model is learned only from *in vitro* nucleosome data, and therefore represents only nucleosome sequence preferences, whereas previous models<sup>5–8,18</sup>, which were learned from *in vivo* data, may also capture sequence preferences of other factors<sup>7</sup>, as well as indirect effects due to chromatin remodelling activities. We tested the model in a cross-validation scheme in which the nucleosome occupancy of each chromosome was predicted using a model that was constructed from the data from all other chromosomes. Our model has high correlations of 0.89 and 0.75 with the *in vitro* and *in vivo* maps, respectively (Fig. 3d, e), and separates nucleosome-enriched regions from nucleosome-depleted regions (Supplementary Fig. 3), indicating that the model successfully identified general predictive rules for the sequence preferences of nucleosomes.

If nucleosome sequence preferences are important in other eukaryotes, then our model should also be predictive of their *in vivo* nucleosome organization. Indeed, we found a good (0.60) correlation between the nucleosome occupancy per base pair predicted by our

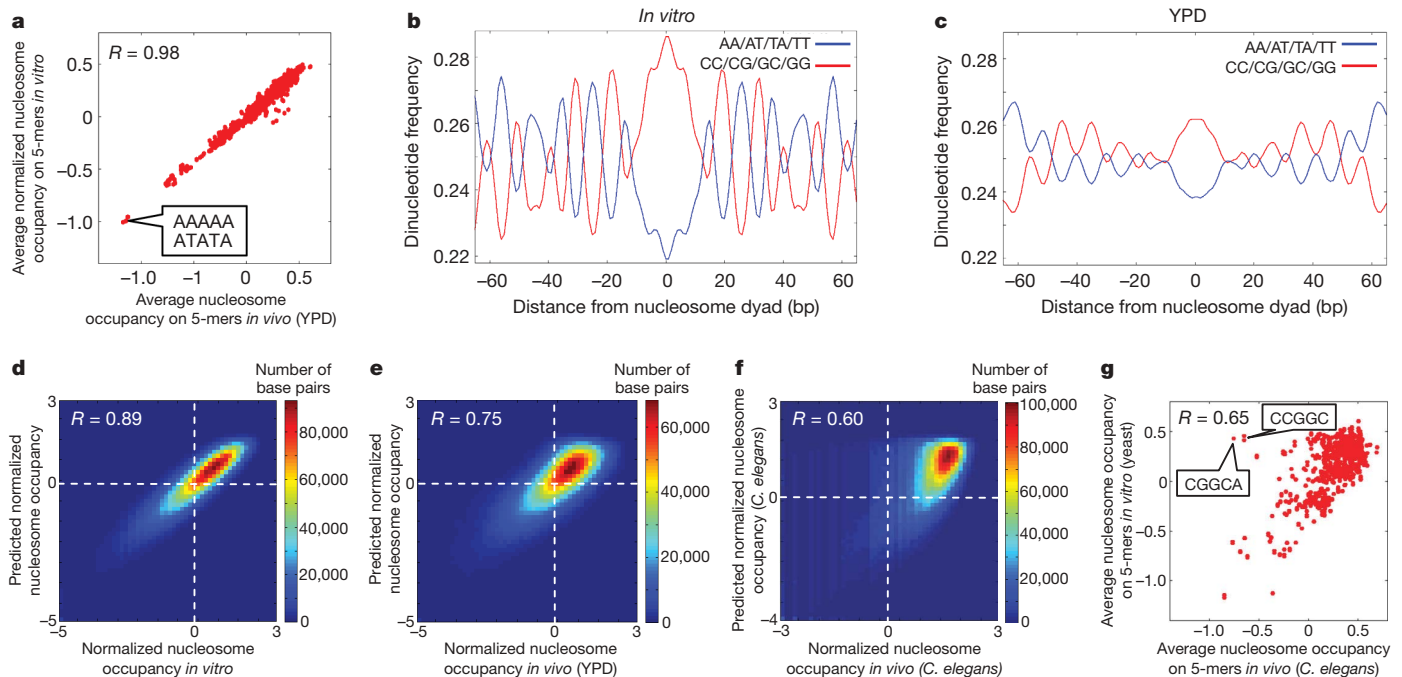
*in vitro* yeast-based model and that of *in vivo* nucleosome occupancy in *C. elegans*<sup>19</sup> (Fig. 3f). Moreover, our model classifies nucleosome-enriched regions from nucleosome-depleted regions in *C. elegans* with high accuracy (Supplementary Fig. 4), and the 5-base-pair sequence preferences of the *C. elegans in vivo* map agree well with those of the yeast *in vitro* map (Fig. 3g). The poorer classification performance in comparison with yeast may indicate that factors other than the DNA sequence preferences make a greater contribution to nucleosome organization in more complex eukaryotes. Alternatively, the poorer performance may indicate that distinct sequence types are present in *C. elegans* for which our yeast *in vitro* data do not provide statistics. Nonetheless, our model is significantly correlated with the *in vivo* nucleosome organization across *C. elegans*.

We next compared the DNA-encoded nucleosome organization of the *in vitro* map with nucleosome organization under growth conditions that cause substantial transcriptional changes relative to log-phase growth in rich medium (that is glucose). In addition to our map obtained from yeast cells grown in rich medium, we also measured the nucleosome organization of yeast cells grown separately in galactose, and in ethanol, and found that the overall nucleosome occupancy is very similar between all three *in vivo* maps, although localized differences are apparent (Fig. 1 and Supplementary Fig. 5). All three *in vivo* maps are highly correlated with the *in vitro* map and show the sequence characteristics seen *in vitro* (Supplementary Fig. 6). These results imply that intrinsic sequence preferences of nucleosomes have a dominant role in determining nucleosome

organization in several growth conditions, with local, condition-specific changes superimposed.

To address concerns regarding biases that may be caused by the sequence specificity of micrococcal nuclease<sup>20</sup> and possible biases in parallel sequencing, we performed a different kind of *in vitro* experiment that measures the relative nucleosome affinity of ~40,000 double-stranded 150-bp oligonucleotides without the use of micrococcal nuclease or parallel sequencing. The resulting 5-base-pair nucleosome sequence preferences are in excellent agreement with those discovered in the genome-wide *in vitro* reconstitution (correlation of 0.83), and there is a good correlation (0.51) between the measured oligonucleotide affinities and those predicted by the model constructed from the genome-wide *in vitro* map (Supplementary Fig. 7). These results are wholly independent of either micrococcal nuclease or parallel sequencing, and thus confirm that the sequence specificities derived from our previous experiments were caused by intrinsic nucleosome preferences, rather than being an artefact of our experimental approach.

Previous studies identified nucleosome depletion around transcription start and stop sites<sup>5–7,9–11</sup>. However, because these studies were based on *in vivo* data, it was not possible to determine which mechanism accounted for the observed patterns. The *in vitro* and *in vivo* maps show highly similar stereotypic nucleosome depletion at translation end sites, indicating that this depletion is largely encoded by nucleosome sequence preferences (Fig. 4b and Supplementary Fig. 8). The two maps also show stereotypic nucleosome depletion



**Figure 3 | The *in vitro* sequence preferences of nucleosomes are highly similar to those of nucleosome-bound sequences *in vivo* and are predictive of nucleosome occupancy in *C. elegans*.** **a**, Comparison of genome-wide relative nucleosome occupancy of nucleosomes over sequences of length 5. For the *in vitro* and *in vivo* maps of nucleosome occupancy, we separately computed the average normalized nucleosome occupancy of each of the 1,024 sequences of length 5, across all of its instances in the genome. Shown is a comparison between the distributions of these 5-base-pair sequences in both maps. Also shown is the Pearson correlation between these distributions. **b**, Position-dependent sequence preferences of nucleosomes in the *in vitro* map. We aligned the individual nucleosome reads in the *in vitro* nucleosome collection. Shown is the fraction (3-bp moving average) of AA/AT/TT/TA and CC/CG/GC/GG dinucleotides at each position of the alignment. **c**, Same as **b**, for the *in vivo* map. **d**, Shown is a density dot plot comparison between the normalized nucleosome occupancy per base pair in the *in vitro* map (*x* axis) and the normalized nucleosome occupancy per base

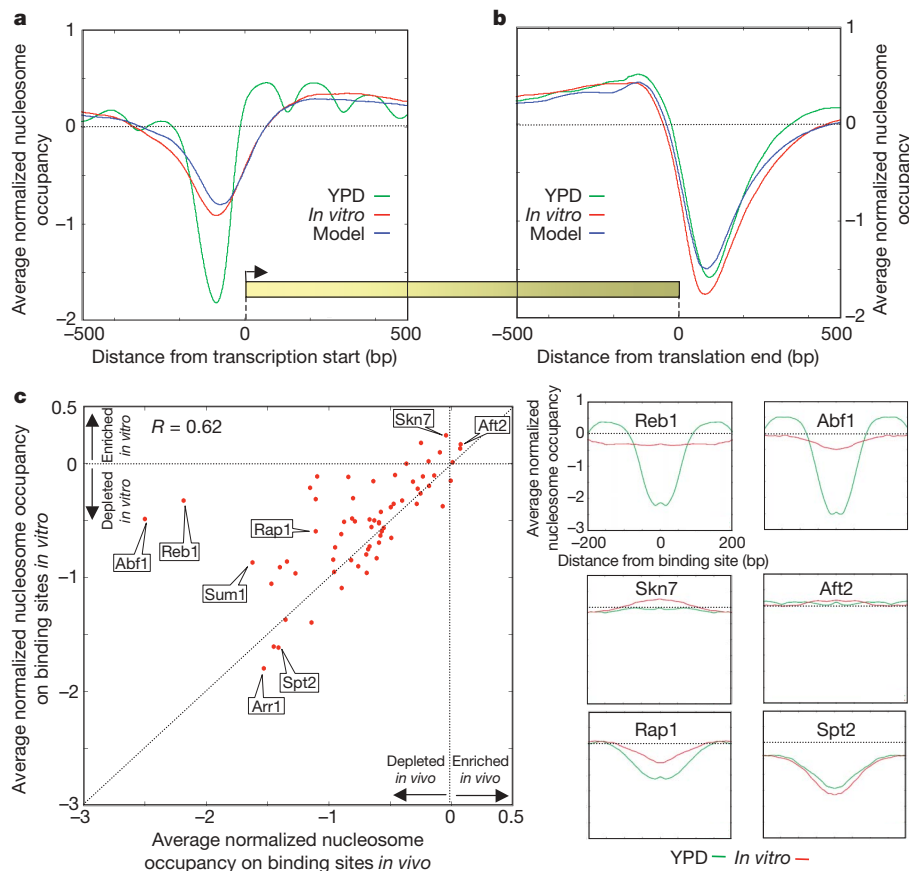
pair predicted by our cross-validated computational model of nucleosome sequence preferences (*y* axis). Values above zero indicate nucleosome enrichment relative to the genome-wide average. The colour of each point represents the number of base pairs that map to that point in the graph. The Pearson correlation between the maps is indicated. **e**, Same as **d**, comparing our model predictions to the *in vivo* map. **f**, *In vitro* nucleosome sequence preferences on yeast genomic DNA are predictive of the *in vivo* nucleosome organization in *C. elegans*. Same as **d**, comparing our model predictions and the *in vivo* nucleosome occupancy map of *C. elegans* on chromosome 2 (ref. 19). **g**, Comparison of yeast nucleosome sequence preferences *in vitro* and those of *C. elegans in vivo*. For each of the maps we separately computed the average normalized nucleosome occupancy of every possible sequence of length 5. For *C. elegans*, we performed these computations on chromosome 2. Shown is a comparison of these 5-base-pair sequence distributions between the yeast *in vitro* map and the *in vivo* map of *C. elegans*, along with the Pearson correlation between these distributions.

at transcription start sites, indicating that this depletion is also partly encoded by nucleosome sequence preferences (Fig. 4a). However, the level of depletion around transcription start sites *in vitro* is smaller than *in vivo*, indicating that transcription factors, chromatin remodellers, the transcription initiation machinery and other mechanisms also contribute to the depletion. Another difference around start sites (Fig. 4a) is the longer-range ordering of nucleosomes into coding regions observed in the *in vivo* map. This may be partly explained by the higher nucleosome concentration *in vivo* (see Methods), which causes increased ordering by statistical positioning<sup>21</sup>.

Nucleosome depletion has also been reported around transcription factor binding sites *in vivo*<sup>7,9,11</sup>. In both our *in vivo* and *in vitro* maps, nucleosome depletion was observed, on average, around the chromatin-immunoprecipitation-determined binding sites<sup>22</sup> of most transcription factors. The maps also agree on the degree of depletion around binding sites (correlation of 0.62; Fig. 4c and Supplementary Fig. 9). These results indicate that nucleosome depletion around regulatory factor binding sites is partly encoded in the genome's intrinsic nucleosome organization, and that this intrinsic organization may facilitate transcription initiation and assist in directing transcription factors to their appropriate sites in the genome<sup>5,23</sup>.

Binding sites for the yeast factors Abf1 and Reb1 show the largest deviation from the above agreement, with sites for both factors being more depleted *in vivo* than *in vitro*. Notably, both factors are highly abundant and influence chromatin structure<sup>24,25</sup>, indicating that the depletion around their sites *in vivo* may be attributable to their own action. The large nucleosome depletion over these factor sites resulted in them being major components of a nucleosome occupancy model created from *in vivo* data<sup>7</sup>. However, because the model constructed here is based solely on *in vitro* data, in which these sites have only moderate depletion, they are not major components of our model.

In summary, we find marked similarities between the nucleosome organization governed only by the DNA sequence preferences of nucleosomes, and the organization of nucleosomes *in vivo* measured under different growth conditions. This result indicates that nucleosome sequence preferences are important determinants of nucleosome organizations *in vivo*. Our analysis indicates that genomes may use their intrinsically encoded nucleosome organization to facilitate functions such as transcription factor binding and transcription. Despite the overall similarity between the *in vitro* and *in vivo* maps, there are differences, consistent with previous studies showing that



**Figure 4 | The intrinsic nucleosome organization over transcripts and transcription factor binding sites.** **a**, For the *in vitro* and *in vivo* nucleosome occupancy maps, and for our model, shown is the normalized nucleosome occupancy per base pair around the transcription start site, averaged across all yeast genes. The long-range ordering of nucleosome occupancy which is present in the *in vivo* maps but not in the *in vitro* map may be partly explained by the lower nucleosome concentration in which the *in vitro* experiment was carried out (see Methods), because higher nucleosome concentration *in vivo* is predicted to cause long-range ordering of nucleosome arrays<sup>21</sup>. **b**, Same as **a**, but around translation end sites of genes (translation end was chosen because transcription end sites are poorly annotated). The depletion around gene ends may be due to the presence of termination signals<sup>27</sup>, which disfavour nucleosome formation *in vitro* (Supplementary Fig. 8). The fact that these signals tend to occur in a specific

orientation with respect to the direction of transcription<sup>27</sup> is consistent with a function in transcript processing, but does not exclude the possibility that one or more of these motifs functions primarily to disfavour nucleosomes. **c**, Comparison of the nucleosome occupancy over transcription factor binding sites between the *in vitro* and the YPD *in vivo* maps. For each transcription factor with at least 50 functional binding sites<sup>22</sup>, we computed, separately for the *in vivo* and *in vitro* maps, the average normalized nucleosome occupancy over its binding sites. Shown is a comparison of these nucleosome occupancies per factor, between the *in vivo* and *in vitro* maps, along with the Pearson correlation between them. For six factors taken from different regions of the plot, we also show the average normalized nucleosome occupancy around those factors' binding sites, for both the *in vitro* and the *in vivo* maps.

factors other than nucleosome sequence preferences contribute to nucleosome organization *in vivo*. Future studies will focus on understanding how nucleosomes are remodelled locally and the function of such remodelling in transcriptional regulation.

## METHODS SUMMARY

*In vivo* maps from yeast nucleosome DNAs were prepared from log-phase cells grown in rich medium (YPD, six independent replicates) as described previously<sup>5</sup>, as well as from cells grown in YP media supplemented with 2% galactose (three replicates) or 2.8% ethanol (four replicates) instead of glucose. The resulting DNAs were subjected to Illumina sequencing-by-synthesis. For the *in vitro* map, histone octamer was purified from chicken erythrocytes, assembled on purified yeast genomic DNA by salt gradient dialysis<sup>13</sup>, digested with micrococcal nuclease and subjected to Illumina sequencing (two independent replicates). The resulting *in vitro* map has a lower concentration of nucleosomes along the DNA than obtained *in vivo*. This technical limitation was necessitated by our finding that reconstitutions at the *in vivo* stoichiometry on long genomic DNA resulted in insoluble chromatin that was inaccessible to micrococcal nuclease. We mapped the resulting reads to the genome and removed reads that mapped to multiple genomic locations. We extended the nucleosome reads of each experiment to the average nucleosome length in that experiment (always between 140–170 bp). For each map, we then calculated the normalized nucleosome occupancy at every base pair as the log-ratio between the number of reads that cover that base pair and the average number of reads per base pair across the genome. We then set the genomic mean in each sample to zero by subtracting the genome-wide mean from every base pair. The independent replicates for each experiment type were in excellent agreement, so we averaged the replicates within each type. The resulting tracks are termed normalized nucleosome occupancy throughout the manuscript. The detailed formulation of our sequence-based model for nucleosome positioning is given in the Methods and is similar to that described in ref. 17, except that it was learned using only the *in vitro* data. For our data, results and model, see <http://genie.weizmann.ac.il/pubs/nucleosomes08/>, and GEO accession number GSE13622.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Received 2 October; accepted 26 November 2008.**

**Published online 17 December 2008.**

- Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285–294 (1999).
- Vignali, M., Hassan, A. H., Neely, K. E. & Workman, J. L. ATP-dependent chromatin-remodeling complexes. *Mol. Cell. Biol.* **20**, 1899–1910 (2000).
- Korber, P., Luckenbach, T., Blaschke, D. & Horz, W. Evidence for histone eviction in trans upon induction of the yeast PHO5 promoter. *Mol. Cell. Biol.* **24**, 10965–10974 (2004).
- Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675 (1986).
- Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
- Ioshikhes, I. P., Albert, I., Zanton, S. J. & Pugh, B. F. Nucleosome positions predicted through comparative genomics. *Nature Genet.* **38**, 1210–1215 (2006).
- Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet.* **39**, 1235–1244 (2007).
- Yuan, G. C. & Liu, J. S. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.* **4**, e13 (2008).
- Yuan, G. C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309**, 626–630 (2005).
- Shivaswamy, S. *et al.* Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* **6**, e65 (2008).

- Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446**, 572–576 (2007).
- Sekinger, E. A., Moqtaderi, Z. & Struhl, K. Intrinsic histone–DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell* **18**, 735–748 (2005).
- Thastrom, A., Bingham, L. M. & Widom, J. Nucleosomal locations of dominant DNA sequence motifs for histone–DNA interactions and nucleosome positioning. *J. Mol. Biol.* **338**, 695–709 (2004).
- Anderson, J. D. & Widom, J. Poly(dA–dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Mol. Cell. Biol.* **21**, 3830–3839 (2001).
- Iyer, V. & Struhl, K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* **14**, 2570–2579 (1995).
- Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M. & Trifonov, E. N. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.* **262**, 129–139 (1996).
- Field, Y. *et al.* Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.* **4**, e1000216 (2008).
- Peckham, H. E. *et al.* Nucleosome positioning signals in genomic DNA. *Genome Res.* **17**, 1170–1177 (2007).
- 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).
- Horz, W. & Altenburger, W. Sequence specific cleavage of DNA by micrococcal nuclease. *Nucleic Acids Res.* **9**, 2643–2658 (1981).
- Kornberg, R. D. & Stryer, L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.* **16**, 6677–6690 (1988).
- MacIsaac, K. D. *et al.* An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7**, 113 (2006).
- Liu, X., Lee, C. K., Granek, J. A., Clarke, N. D. & Lieb, J. D. Whole-genome comparison of Leu3 binding *in vitro* and *in vivo* reveals the importance of nucleosome occupancy in target site selection. *Genome Res.* **16**, 1517–1528 (2006).
- Lascaris, R. F., Groot, E., Hoen, P. B., Mager, W. H. & Planta, R. J. Different roles for Abf1p and a T-rich promoter element in nucleosome organization of the yeast *RPS28A* gene. *Nucleic Acids Res.* **28**, 1390–1396 (2000).
- Raisner, R. M. *et al.* Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. *Cell* **123**, 233–248 (2005).
- Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
- Graber, J. H., Cantor, C. R., Mohr, S. C. & Smith, T. F. Genomic detection of new yeast pre-mRNA 3'-end-processing signals. *Nucleic Acids Res.* **27**, 888–894 (1999).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank H. Kelkar for bioinformatics support in handling the sequencing files, P. Mieczkowski and J. McPherson for Illumina sequencing (library preparation and instrument operation), and the members of our respective laboratories for technical assistance, discussions and comments on the manuscript. The UNC sequencing facility is funded by the Lineberger Comprehensive Cancer Center and University Cancer Research Funds. J.W. acknowledges the use of instruments at Northwestern University's Keck Biophysics Facility. This work was supported by a grant from CIHR to T.R.H. and C. Nislow, a grant from the NIH to J.D.L., a grant from the NIH to J.W., and grants from the European Research Council (ERC) and NIH to E.S. D.T. holds an NSERC postgraduate scholarship. N.K. is a Clore scholar. E.S. is the incumbent of the Soretta and Henry Shapiro career development chair.

**Author Contributions** This work was performed jointly in the Hughes, Lieb, Widom and Segal laboratories and with substantial contributions from each.

**Author Information** This has been submitted to GEO under accession number GSE13622. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to J.W. ([j-widom@northwestern.edu](mailto:j-widom@northwestern.edu)) or E.S. ([eran.segal@weizmann.ac.il](mailto:eran.segal@weizmann.ac.il)).

## METHODS

**Parallel sequencing of yeast nucleosomes *in vivo*.** Four samples of yeast nucleosome DNAs (*in vivo* nucleosomes) were prepared from log-phase cells grown in rich medium (YPD) as described previously<sup>5</sup>. Additionally, samples were prepared from cells grown in YP media supplemented with 2% galactose (two samples) or 2.8% ethanol (two samples) instead of glucose. Five additional samples (two in YPD, one in galactose and two in ethanol) were prepared from cells that were formaldehyde-crosslinked before isolating nucleosomes. Crosslinking was achieved with 1% formaldehyde for 15 min at 30 °C and stopped with 125 mM glycine. The cells were then washed, made into spheroplasts with lyticase as above, treated with micrococcal nuclease (MNase) as described previously<sup>12</sup>, and nucleosomal DNAs isolated as described<sup>10</sup>. All of the resulting DNAs were subjected to Illumina/Solexa sequencing-by-synthesis. The results from crosslinked or uncrosslinked cells for a given growth medium were very similar; therefore, we averaged the replicates within each growth medium (six YPD, four YPetOH, three YPGal).

**Parallel sequencing of nucleosomes reconstituted *in vitro* on yeast genomic DNA.** *S. cerevisiae* genomic DNA was purified from strain YLC8 (*MATa ura3Δ leu2Δ his3Δ met15Δ*) using standard methods with additional steps to remove contaminating RNA (see Supplementary Information). Histone octamer was purified from chicken erythrocytes using salt extraction and hydroxyapatite column chromatography, as described previously<sup>28</sup>. Genomic DNA was reconstituted into nucleosomes under selective pressure for nucleosome-favouring sequences by salt gradient dialysis<sup>13</sup>, using 40 μg histone octamer plus 100 μg DNA in a 200 μl volume. This results in a lower concentration of nucleosomes along the DNA than obtained *in vivo*. This technical limitation was necessitated by our finding that reconstitutions at the *in vivo* stoichiometry on this long genomic DNA resulted in insoluble chromatin that was inaccessible to micrococcal nuclease. Two independent reconstitutions were carried out. The resulting nucleosomes were biochemically isolated by MNase digestion, using  $6 \times 10^{-3}$  units MNase (Sigma Chemical Company) per 10 μg competitively reconstituted DNA, in 10 mM Tris, pH 8.0, and 1 mM CaCl<sub>2</sub> for 5 min at 37 °C. DNA was extracted, and protected fragments of length ~147 bp were isolated by polyacrylamide gel electrophoresis, and then extracted from the gel. Both samples were independently subjected to Illumina sequencing at the UNC high-throughput sequencing facility. We note that, because the chicken histones represent a mixture of post-translationally modified and unmodified histones as well as variant histones, there is a possibility that these modifications or variants influence the measured nucleosome positions.

**Mapping and post-processing of parallel sequencing reads.** We mapped reads resulting from sequencing experiments to the yeast genome, allowing at most one mismatch and no gaps. To estimate the mean DNA fragment length in each experiment, we superimposed the nucleosome reads of one strand and examined the distribution of nucleosome reads of the opposite strand. As expected, this distribution showed a strong peak at ~140–170 bp for all experiments, with slight variations between experiments. We used the maximum of the peak as an estimation of the mean DNA fragment length and extended all nucleosome reads to this length. We defined repetitive regions as regions that were matched by a read that mapped to more than one place in the genome. We excluded repetitive regions and their 150-bp vicinity from our analyses. To obtain genomic nucleosome occupancy tracks we summed for each position all reads covering it. We noticed that in some replicates from each of the three *in vivo* conditions (YPD, YPetOH and YPGal), but not in either of the *in vitro* replicates, a small fraction of the mapped base pairs (in all cases less than 0.5%) were covered by an aberrantly large number of reads (more than ten times the median genomic base pair coverage) that may potentially result from sequencing biases. For these regions, we set the coverage of these base pairs to equal ten times the median genomic base-pair coverage. Next, for each map, we calculated the average nucleosome occupancy at every base pair by dividing the number of reads that cover that base pair by the average number of reads per base pair across the genome. Finally, for each replicate we took the log of the ratio at every base pair and set the genomic mean to zero by subtracting the mean from every base pair. Thus, base pairs for which the occupancy value is below 0 showed a relative depletion of nucleosomes in comparison with the genome-wide average, whereas base pairs with values above 0 have a relative enrichment. The independent replicates for each experiment type were in excellent agreement, so we averaged the replicates within each type to create four nucleosome occupancy maps: three *in vivo* (YPD, YPetOH, YPGal) and one *in vitro*. These final tracks are termed the normalized nucleosome occupancy. The data in Fig. 1 are presented without the log transformation.

**Analysis of nucleosomes reconstituted on oligonucleotides.** Two concerns with our experimental approach were the known sequence specificity of micrococcal nuclease<sup>20</sup> and the possibilities of biases introduced by the parallel

DNA-sequencing strategy. We addressed both of these concerns simultaneously with an independent and different kind of experiment that obtains large-scale measurements of *in vitro* nucleosome sequence preferences without the use of micrococcal nuclease or parallel sequencing. In brief, we created a pool of ~40,000 double-stranded oligonucleotides of length 150 bp, each flanked by common priming sites<sup>29</sup>, and combined the pool with limiting amounts of chicken histones to form nucleosomes. DNAs in reconstituted nucleosomes were separated from unincorporated DNAs by native gel electrophoresis, without the use of any micrococcal nuclease. We then extracted from the gel the DNA that had successfully competed to form nucleosomes, and used both parallel sequencing and microarrays to compare the nucleosomal DNAs to DNAs in the initial pool. For each sequence, we calculated the log-ratio between the reconstituted fraction and the initial pool as a measure of the nucleosome affinity of that sequence. Thus, a log-ratio above zero indicates a higher than average affinity to form nucleosomes, whereas a value below zero indicates a lower than average affinity. We note that this experiment is inherently noisy, as shown by a large variation in the number of sequence reads in the unselected reference sample, which may be due to the use of two rounds of PCR in the procedure.

Specifically, we created 197-mer oligonucleotides (containing 150-mer unique sequences flanked by common priming sites) using a modification of the microarray-based method described previously<sup>29</sup>. Sequences were designed to represent tiled yeast genomic sequences (22,236 oligonucleotides), and a survey of randomly generated sequences selected to represent a wide variety of DNA sequence characteristics (nucleotide and dinucleotide composition, periodic features and DNA structural parameters<sup>7</sup>). DNA obtained from the array was PCR-amplified and assembled into nucleosomes under selective pressure for nucleosome-favouring sequences by salt gradient dialysis<sup>13</sup>, using 6 μg histone octamer and 12 μg DNA in a 200-μl volume. The resulting nucleosomes were separated away from the remaining naked DNA by native polyacrylamide gel electrophoresis (5% (w/v) polyacrylamide) in 1/3× TBE buffer (TBE is 90 mM Tris, 90 mM boric acid, 2 mM EDTA, pH 8.3). The nucleosome-containing band was excised from the gel, and its DNA extracted by crushing and soaking into 0.3 M NaOAc, 0.5 M NH<sub>4</sub>OAc, 0.1 mM EDTA and 0.1% (w/v) SDS, followed by ethanol precipitation. For Illumina analysis, the pools before and after nucleosome assembly were re-amplified and NruI sites present in the primers were cleaved before sequencing from the ends. In our analyses, we used oligonucleotides that were sequenced at least once and at most 500 times in each experiment. For microarray analyses, the pools before and after assembly were re-amplified with fluor-labelled primers, and hybridized to an Agilent array designed to detect the individual sequences represented in the pools. Additional details are available in the Supplementary Information.

To remove any direct biases resulting from the primers in the computation of preferences over sequences of length 5 from these data (Supplementary Fig. 7), we omitted the primers altogether in the computation of 5-base-pair preferences (that is, all 5-base-pair sequences were extracted only from the variable oligonucleotides without the primers). This left open the possibility of indirect biases in preferences of sequences of length 5 resulting from the primers; for example, if one primer sequence attracted nucleosomes and if there were systematic biases in compositions of sequences of length 5 near that end of the designed oligonucleotides. However, half (~20,000) of the oligonucleotides were selected from the yeast genome, and the other half were designed in several unrelated ways (see Supplementary Methods) and do not have such biases. The two approaches are in strong agreement regarding these preferences over sequences of length 5 base pairs, diminishing concerns regarding biases caused by the primers (the correlations between the preferences computed from the first approach and the second approach are 0.81 and 0.83 for reading out the measurements of the second approach by parallel sequencing or microarrays, respectively; see Supplementary Fig. 7).

**Data sets.** The yeast genome sequence (May 2006 build) and gene and chromosome annotations were obtained from SGD<sup>30</sup>. Yeast transcription start sites were compiled from refs 31–33: for each gene, the transcription start site was taken as that with the most sequence reads from refs 31 and 32, or from ref. 33 when no sequencing data was available. Functional transcription factor DNA binding sites in yeast, defined as sites that are bound by their cognate transcription factor were obtained from refs 22 and 34. Expression levels in yeast were taken from ref. 26. The *in vivo* map of nucleosome organization in *C. elegans* was taken from ref. 19; specifically, we used both the ‘adjusted nucleosome coverage’ and ‘raw nucleosome coverage’, applying the same normalization method to the latter as we did to our yeast nucleosome maps<sup>19</sup>.

**Classifying nucleosome-enriched regions from nucleosome-depleted regions.** To test whether the two maps agree on the scale of individual nucleosomes, we took the target map, and from it defined nucleosome-enriched regions as maximal consecutive regions longer than 50 bp for which minimum occupancy at every base pair is above some threshold,  $t_c$ , and nucleosome-depleted regions as

maximal consecutive regions longer than 50 bp for which maximum occupancy at every base pair is below some threshold,  $t_d$ . We chose 50 bp as a compromise between partitioning as much of the genome as possible into nucleosome-enriched and nucleosome-depleted regions while still deriving each partition from enough data to reduce noise that may arise when regions are too short (a choice of 25 bp yielded equivalent results). At various thresholds, we then scored each of the resulting nucleosome-enriched and nucleosome-depleted regions by its average occupancy in the predicting map. Note that at the threshold  $t_c = t_d = 0$ , the evaluation is done across the entire genome (typically ~97%, because regions <50 bp are excluded), and the median length of the nucleosome-enriched regions is typically 150–190 bp. If the predicting map were fully predictive of the target map, then the predicting map occupancy score of every 'enriched' region in the target map would be higher than that of every 'depleted' region in the target map. A standard quantification of this predictive power is the receiver operating characteristic curve, whose area under the curve is 1 for perfect performance and 0.5 for random guessing.

**Sequence-based model for nucleosome positioning.** We use the *in vitro* map to represent the sequence preferences of nucleosomes by devising a probabilistic model, similar in formulation to that proposed in ref. 17, that assigns a score to every 147-bp (nucleosome-length) sequence. Our model is based on both the position-dependent and global preferences of sequences of length 5 that we characterized previously (Fig. 3). The first component,  $P_N$ , represents the distribution over dinucleotides at each position along the nucleosome length, and thus captures the periodic signal of dinucleotides along the nucleosome. The second component,  $P_L$ , represents the position-independent distribution of nucleosomes over sequences of length 5, and thus captures sequences that are generally favoured or disfavoured by nucleosomes regardless of their detailed position within the nucleosome. We chose to represent this component using sequences of length 5, because this is the highest order  $k$ -mer for which our data have sufficient statistics to robustly estimate each of the associated parameters. The final score that our model assigns to a 147-bp sequence  $S$  is then given by the log-ratio of these two model components:

$$\begin{aligned} \text{Score}(S) &= \log \frac{P_N(S)}{P_L(S)} \\ &= \log \frac{P_{N,1}(S[1]) \prod_{i=2}^{147} P_{N,i}(S[i]|S[i-1])}{P_L(S[1]) \prod_{i=2}^{147} P_L(S[i]|S[\max(1, i-4)], \dots, S[i-1])} \end{aligned} \quad (1)$$

where  $P_{N,i}$  is the  $i$ -th component of the dinucleotide model component and represents the conditional probability distribution over nucleotides at position  $i$  given the nucleotide that appeared at position  $(i-1)$ , and  $P_L$  is the position-independent component of the second component of our model ( $P_L$ ). Note that  $P_{N,1}$  is represented by a mononucleotide model over the nucleotide at the first position.

We now describe in detail how each of the two components of our model was derived. To estimate the position-specific dinucleotide component of our model,  $P_N$ , we took all of the nucleosome reads (a total of over 10,000,000) from our collection of nucleosome-bound sequences, and first aligned these sequences about their centre. Each sequence was added twice to the alignment, once in its original form and once in its reverse complement form, to account for the two-fold symmetry in the nucleosome structure<sup>35</sup>. With each position  $i$ , we then associated a dinucleotide distribution,  $P_{N,i}$ , which we estimated from the combined dinucleotide counts at alignment positions  $[i-2, i-1]$ ,  $[i-1, i]$  and  $[i, i+1]$  (the two end positions of the nucleosome are averaged with fewer positions). Combining the dinucleotides at the two neighbouring positions smoothes the resulting dinucleotide distribution at each position with a 3-bp moving average, and is motivated by the experimental evidence that small  $\pm 1$ -bp changes in spacing of key nucleosome DNA sequence motifs can occur with a relatively small cost to the free energy of histone–DNA interactions<sup>36</sup>. To remove sequence composition biases from this component, we normalized the distribution over each position separately, by dividing the final probability of every dinucleotide at each position by the average probability of that dinucleotide across all positions, and finally normalized the resulting dinucleotide weights to a probability distribution. We used this estimation procedure in the 127 central positions of the nucleosome, and we forced a uniform distribution over the 10 remaining positions at each end of the nucleosome profile. This was done to avoid biases in nucleotide distributions that may arise from the sequence specificity of the micrococcal nuclease used to isolate the nucleosome, because this way we do not include statistics that are taken from the cut site of the

nuclease. Note that our above construction produces a reverse complement symmetrical distribution; that is, the probability of a sequence and its reverse complement are equal by construction.

The position-independent component of our model,  $P_L$ , the purpose of which is to represent sequences that are generally favoured or disfavoured regardless of their position within the nucleosome, assigns a score to each 147-bp sequence, as the product of a position-independent Markov model,  $P_i$ , of order 4. Thus,  $P_L$  defines a probability distribution over every one of the 1,024 possible sequences of length 5. We chose to model the distribution over sequences of length 5, because this is the highest order for which our data still provides sufficient statistics to robustly estimate the value of each of the 1,024 parameters. Given a map of nucleosome positions, the probability of each sequence of length 5 is equal to the reciprocal of the average nucleosome occupancy of that sequence across all instances of the 5-base-pair-long sequence in the map, where this reciprocal average occupancy is then scaled to a probability by dividing it by the sum of all such reciprocal occupancies across all 5-base-pair-long sequences (we use the reciprocal of the average occupancy so that higher probabilities correspond to disfavoured sequences). Thus, this Markov model,  $P_i$ , includes contributions both from sequences that are disfavoured by nucleosomes and from sequences that are favoured by nucleosomes, because it models the distribution over all sequences of length 5, with the disfavoured sequences having a relatively high probability and the favoured sequences having a relatively low probability.

The above probabilistic model assigns a nucleosome formation score to each sequence of (nucleosome-length) 147 bp. We then followed the approach of ref. 5, and used the scores of this model to compute the genome-wide distribution over nucleosome positions, taking into account steric hindrance constraints between neighbouring nucleosomes. These computations can be done efficiently using dynamic programming, and their output is a track of the average nucleosome occupancy per base pair, that is, the probability that each base pair in the genome is covered by any nucleosome.

In developing our model, we also separately evaluated the performance of the  $P_N$  and  $P_L$  components, and found that the performance of the  $P_L$  component alone is nearly the same as that of the full model (per-base-pair nucleosome occupancy correlation of 0.876 between a  $P_L$ -only model and the *in vitro* map, compared with 0.880 for the full model using both  $P_L$  and  $P_N$ ), whereas the performance of the  $P_N$  component alone is highly predictive but slightly worse than the full model (correlation of 0.820). Thus, for practical purposes of generating predictions, using only the  $P_L$  component may suffice. Nevertheless, the  $P_N$  component captures real periodic aspects of nucleosome sequence preferences that have been reported by several other groups<sup>4,6,37</sup>, and that we now find in the genome-wide *in vitro* nucleosome map in Fig. 3b. For these reasons, we retain the  $P_N$  component as part of the model.

For the synthetic oligonucleotide analysis (Supplementary Fig. 7), we used the nucleosome formation potential of equation (1). In all other analyses we converted this genome-wide nucleosome formation potential into genome-wide nucleosome occupancy per base pair, as described in ref. 5. We selected the concentration and temperature parameters ( $\tau$  and  $\beta$ , respectively) in a cross-validation manner to be  $\tau = 0.03$ ,  $\beta = 1$  in yeast and  $\tau = 0.1$ ,  $\beta = 1$  in *C. elegans*. Finally, to appropriately compare the model with log-transformed data, we applied to the predictions of the model the same log normalization that was applied to the data.

28. Feng, H. P., Scherl, D. S. & Widom, J. Lifetime of the histone octamer studied by continuous-flow quasielastic light scattering: test of a model for nucleosome transcription. *Biochemistry* **32**, 7824–7831 (1993).
29. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4**, 931–936 (2007).
30. Cherry, J. M. *et al.* SGD: *Saccharomyces* genome database. *Nucleic Acids Res.* **26**, 73–79 (1998).
31. Zhang, Z. & Dietrich, F. S. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5' SAGE. *Nucleic Acids Res.* **33**, 2838–2851 (2005).
32. Miura, F. *et al.* A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc. Natl Acad. Sci. USA* **103**, 17846–17851 (2006).
33. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA* **103**, 5320–5325 (2006).
34. Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
35. Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
36. Ong, M. S., Richmond, T. J. & Davey, C. A. DNA stretching and extreme kinking in the nucleosome core. *J. Mol. Biol.* **368**, 1067–1074 (2007).
37. Bolshoy, A., Ioshikhes, I. & Trifonov, E. N. Applicability of the multiple alignment algorithm for detection of weak patterns: periodically distributed DNA pattern as a study case. *Comput. Appl. Biosci.* **12**, 383–389 (1996).