doi:10.1038/nature04979

ARTICLES

A genomic code for nucleosome positioning

Eran Segal[1], Yvonne Fondufe-Mittendorf[2], Lingyi Chen[2], AnnChristine Thåström[2], Yair Field[1], Irene K. Moore[2], Ji-Ping Z. Wang[3] & Jonathan Widom[2]

Eukaryotic genomes are packaged into nucleosome particles that occlude the DNA from interacting with most DNA binding proteins. Nucleosomes have higher affinity for particular DNA sequences, reflecting the ability of the sequence to bend sharply, as required by the nucleosome structure. However, it is not known whether these sequence preferences have a significant influence on nucleosome position *in vivo*, and thus regulate the access of other proteins to DNA. Here we isolated nucleosome-bound sequences at high resolution from yeast and used these sequences in a new computational approach to construct and validate experimentally a nucleosome–DNA interaction model, and to predict the genome-wide organization of nucleosomes. Our results demonstrate that genomes encode an intrinsic nucleosome organization and that this intrinsic organization can explain ~50% of the *in vivo* nucleosome positions. This nucleosome positioning code may facilitate specific chromosome functions including transcription factor binding, transcription initiation, and even remodelling of the nucleosomes themselves.

Eukaryotic genomic DNA exists as highly compacted nucleosome arrays called chromatin. Each nucleosome contains a 147-base-pair (bp) stretch of DNA, which is sharply bent and tightly wrapped around a histone protein octamer[1]. This sharp bending occurs at every DNA helical repeat (~10 bp), when the major groove of the DNA faces inwards towards the histone octamer, and again ~5 bp away, with opposite direction, when the major groove faces outward. Bends of each direction are facilitated by specific dinucleotides[2,3]. Neighbouring nucleosomes are separated from each other by 10–50-bp-long stretches of unwrapped linker DNA[4]; thus, 75–90% of genomic DNA is wrapped in nucleosomes. Access to DNA wrapped in a nucleosome is occluded[1] for polymerase, regulatory, repair and recombination complexes, yet nucleosomes also recruit other proteins through interactions with their histone tail domains[5]. Thus, the detailed locations of nucleosomes along the DNA may have important inhibitory or facilitatory roles[6,7] in regulating gene expression.

DNA sequences differ greatly in their ability to bend sharply[2,3,8]. Consequently, the ability of the histone octamer to wrap differing DNA sequences into nucleosomes is highly dependent on the specific DNA sequence[9,10]. *In vitro* studies show this range of affinities to be 1,000-fold or greater[11]. Thus, nucleosomes have substantial DNA sequence preferences. A key question is whether genomes use these sequence preferences to control the distribution of nucleosomes *in vivo* in a way that strongly impacts on the ability of DNA binding proteins to access particular binding sites. By controlling binding site accessibility in this way, genomes could, for example, target the binding of transcription factors towards appropriate sites and away from irrelevant, non-functional sites[9].

One view is that the sequence preferences of nucleosomes might not be meaningful. Nucleosome positions might be regulated in cells in *trans* by the abundant[12] ATP-dependent nucleosome remodelling complexes[13], which might over-ride the sequence preferences of nucleosomes and move them to new locations whenever needed. Another view, however, is that remodelling factors do not themselves determine the destinations of the nucleosomes that they mobilize. Rather, the remodelling complexes may allow nucleosomes to sample alternative positions rapidly, resulting in a thermodynamic equilibrium between the nucleosomes and the site-specific DNA binding proteins that compete with nucleosomes for occupancy along the genome. In this view, nucleosome positions are regulated in *cis* by their intrinsic sequence preferences, which would then have significant regulatory roles. In this *cis* regulation model, we expect the genome to encode a nucleosome organization, intrinsic to the DNA sequence alone, comprising sequences with both low and high affinity for nucleosomes. Many of the high-affinity sequences should then be occupied by nucleosomes *in vivo*. Moreover, the detailed distribution of nucleosome positions encoded by the genome should significantly influence chromosome functions genome-wide.

Here we report the results of a combined experimental and computational approach to detect the DNA sequence preferences of nucleosomes and the intrinsic nucleosome organization of the genome that these preferences dictate. Our findings demonstrate that eukaryotic genomes use a nucleosome positioning code, and link the resulting nucleosome positions to specific chromosome functions.

Validating a nucleosome–DNA interaction model

To construct a model for nucleosome–DNA interactions in yeast (Fig. 1a), we used a genome-wide assay to isolate DNA regions that were stably wrapped in nucleosomes. Our experimental method maps nucleosomes on the yeast genome with greater accuracy than previous approaches, resulting in a set of 199 mononucleosome DNA sequences of length 142–152 bp (Supplementary Fig. 1). We used this collection of sequences to construct a probabilistic model that represents the DNA sequence preferences of yeast nucleosomes (Supplementary Fig. 2). Our approach resembles that used for representing the binding specificities of transcription factors from a collection of known sites, but with two main distinctions: first, in contrast to the mononucleotide probability distributions used for

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel. [2]Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, 2153 Sheridan Road, Evanston, Illinois 60208, USA. [3]Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, Illinois 60208, USA.

transcription factors, we use dinucleotide probability distributions (as dinucleotides are the simplest sequence elements to capture the sequence-dependent mechanics of DNA bending[14] that are essential for histone–DNA association[3]); second, when constructing the model we represent the two-fold symmetry axis of the nucleosome structure[1] by including the reverse complement of each sequence in the nucleosome collection. More sophisticated nucleosome–DNA interaction models based on mixture models[15] or on the expectation-maximization algorithm[16] yielded equivalent results.

As expected for a nucleosome–DNA interaction model, the resulting model exhibits distinctive sequence motifs that recur periodically at the DNA helical repeat and are known to facilitate the sharp bending of DNA around the nucleosome[3]. These include ~10-bp periodic AA/TT/TA dinucleotides that oscillate in phase with each other (Fig. 1b) and out of phase with ~10-bp periodic GC dinucleotides. Moreover, the same periodicities and phase relationships were derived independently from a collection of 177 natural nucleosomes from chicken[2], and they arose again in three independent *in vitro* experiments that selected for stable nucleosomes. These *in vitro* selection experiments include one on chemically synthesized random DNA[17], one on mouse genomic DNA[18], and a new experiment that we performed on yeast genomic DNA (see Methods). The similarities among these independently derived nucleosome patterns are striking and quantitatively significant (Fig. 1b and Supplementary Figs 3–5), for example, $P < 10^{-50}$ for yeast–chicken *in vivo* similarity.

We experimentally validated the importance of these periodic sequence motifs for nucleosome–DNA interactions *in vitro*. Improving the agreement of a sequence with these motifs increased its binding affinity to the nucleosome, whereas changing the periodicity or deleting the key motifs decreased that affinity (Fig. 1c–e and Supplementary Fig. 6). In addition, these periodic motifs did not arise in alignments of randomly chosen regions in the yeast or chicken genomes (Supplementary Fig. 7). Together, these results establish that the distinctive motifs in our model represent DNA sequence preferences of nucleosomes (Fig. 1f).

If genomes use these sequence preferences, then high-affinity sequences should be prevalent in the genome. Indeed, we found that intergenic and coding regions in the yeast genome contain many more high-affinity DNA sequences than expected by chance ($P < 10^{-200}$ for both intergenic and coding regions; Supplementary Fig. 8), and that scores at positions separated by 10 bp are strongly correlated (Supplementary Fig. 9). Together with the distinctive features of the yeast *in vivo* nucleosome collection, these results show that sequence motifs for positioning nucleosomes are abundantly encoded in the yeast genome and that nucleosomes occupy these sequences *in vivo*.

Predicting nucleosome organization in genomic DNA sequence

We next sought to understand how the encoded nucleosome preferences integrate to specify the intrinsic genome-wide positioning of nucleosomes. This task is non-trivial because encoded nucleosome positions are correlated through steric hindrance. We designed a thermodynamic model that defines an apparent free energy for every organization of nucleosomes on the DNA, taking steric hindrance and competition between nucleosomes into account (see Methods).

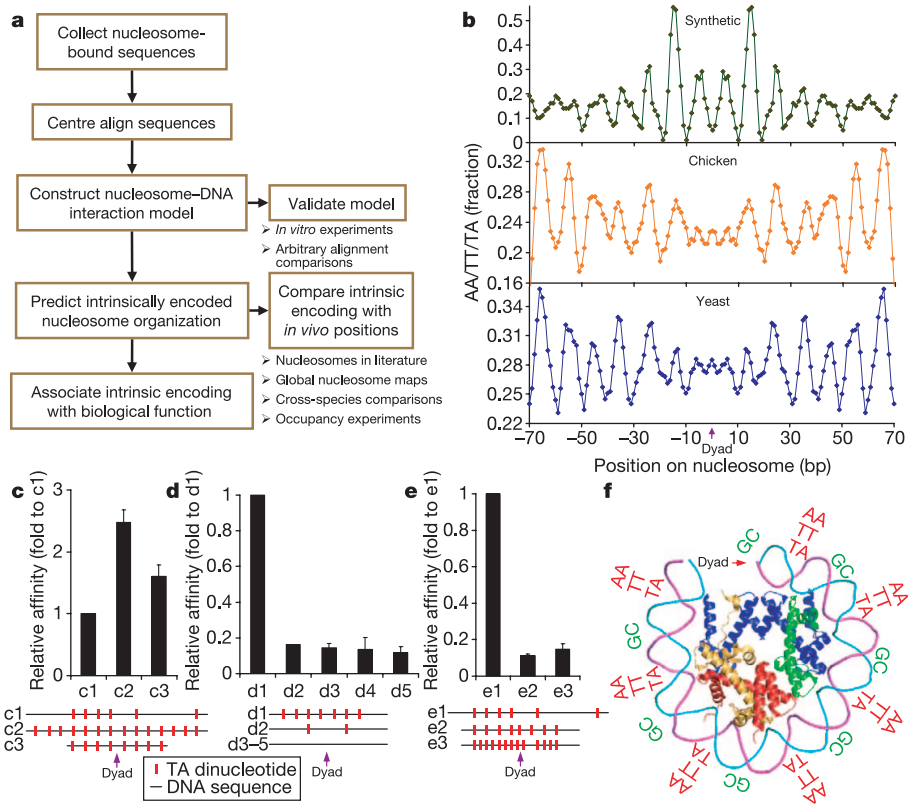

Figure 1 | Probabilistic nucleosome–DNA interaction model. **a**, Flow chart illustrating our approach. **b**, Fraction (3-bp moving average) of AA/TT/TA dinucleotides at each position of centre-aligned yeast, chicken[2] or random chemically synthesized[17] nucleosome-bound DNA sequences, showing ~10-bp periodicity of these dinucleotides. **c–e**, *In vitro* experiments. Positions of the key AA/TT/TA dinucleotides on the tested sequences are indicated. Error bars are s.e.m. **c**, Nucleosome binding affinities of sequences c2 and c3 (ref. 44), which include additional dinucleotide motifs at key positions, relative to the affinity of c1. **d**, Sequences d2–d5 have dinucleotide motifs removed from key positions in e1. **e**, Sequences e2 and e3 have disrupted spacing between the key dinucleotide motifs. **f**, Key dinucleotides inferred from the alignments are shown relative to the three-dimensional structure of one-half of the symmetric nucleosome.

A dynamic programming method[19] evaluated efficiently all sterically allowed organizations, yielding both the probability that each base pair is occupied by any nucleosome (average nucleosome occupancy) and the genomic locations of the sites at which nucleosomes have a high probability of starting (stably positioned nucleosomes).

The resulting intrinsic nucleosome organization differs qualitatively at different genomic locations. In some cases, several mutually exclusive organizations dominate (Supplementary Fig. 10a, b); in others, a single organization dominates (Supplementary Fig. 10c); and yet in others no particular organization dominates (Supplementary Fig. 10d). Comparing these diverse intrinsic organizations to known transcription factor binding sites[20] reveals the potential regulatory role of nucleosomes: nucleosomes may have a strong affinity to occupy transcription factor binding sites (rendering them inaccessible) in some genomic locations (Supplementary Fig. 10a), but a weak affinity to occupy sites (thereby increasing their accessibility) in other locations (Supplementary Fig. 10b).

Predicted nucleosome organization reflects *in vivo* data

By comparing actual *in vivo* nucleosome positions to our predicted or experimentally measured intrinsically encoded positions, we can test whether *in vivo* positions are dictated by the genomic sequence. To this end, we used five different approaches. First, we measured the distance between our predicted stable nucleosome positions (stability probability $\geq$0.2; see Methods) and 99 experimentally mapped nucleosome positions at 11 loci[21–28] (Supplementary Fig. 11). There is some disagreement between different experimental measurements of nucleosome positions (Fig. 2b and Supplementary Fig. 12), hence discrepancies between our predictions and literature reports are attributable to inaccuracies both in our model and in the literature. Even so, six loci showed substantial correspondence (Fig. 2 and Supplementary Figs 13–22). Overall, 54% of our predicted stable nucleosomes were within 35 bp of the literature positions, significantly more than the 39 $\pm$ 1% expected by chance ($P < 10^{-16}$).

Second, we compared our predictions to three genome-wide measurements of nucleosome positions at low[29,30] or higher[31] resolution. Our model showed significant correspondence to these experiments, predicting lower occupancy at nucleosome-depleted (low nucleosome abundance) coding or intergenic regions[29,30] (Supplementary Figs 23–25; 68% of 57 depleted coding regions and 76% of 294 depleted intergenic regions had predicted low occupancy compared with 30% ($P < 10^{-6}$) and 56% ($P < 10^{-9}$), respectively, expected by chance). The model also showed strong correspondence with the higher resolution nucleosome map[31]: 45% of our predicted stable nucleosomes were within 35 bp of experimentally determined nucleosome positions[31] compared with 32 $\pm$ 1% expected by chance, $P < 10^{-15}$ (Supplementary Figs 26 and 27). Notably, our predictions also match closely the stereotyped chromatin organization at Pol II promoters as revealed by the higher resolution nucleosome map[31], and the most stable nucleosome predicted by our model at promoters is located precisely (within 8 bp) where stable nucleosomes containing the histone variant H2A.Z are located *in vivo*[32] (Fig. 5a).

Third, we compared the yeast model predictions to those of a model constructed independently using only nucleosome-bound sequences from chicken. The predictions of the chicken model when applied to the yeast genome correlated strongly with those of the yeast model (Supplementary Fig. 28) and with the genome-wide experimental measurements of nucleosome occupancy at yeast coding and intergenic regions[29–31]: 35% of 57 depleted coding regions and 72% of 294 depleted intergenic regions had predicted low occupancy compared with 4% ($P < 10^{-4}$) and 53% ($P < 10^{-8}$) expected by chance.

Fourth, we carried out a new selection for nucleosome formation on yeast genomic DNA *in vitro*. This experiment directly reveals intrinsically encoded, individual high-affinity nucleosome positions. These *in vitro* nucleosome locations overlap significantly with our *in vivo* yeast nucleosome collection: 32% of 339 selected *in vitro* nucleosomes overlapping the *in vivo* bound sequences compared with 5% ($P < 10^{-5}$) expected by chance. The *in vitro* selected nucleosomes are particularly enriched in intergenic regions that have a high predicted nucleosome occupancy, compared with random genomic locations and to locations immediately upstream or downstream of the selected nucleosomes ($P < 10^{-3}$; Fig. 3c and Supplementary Figs 29 and 30).

Finally, we experimentally tested whether our highest occupancy predictions are highly occupied by nucleosomes *in vivo*, by measuring

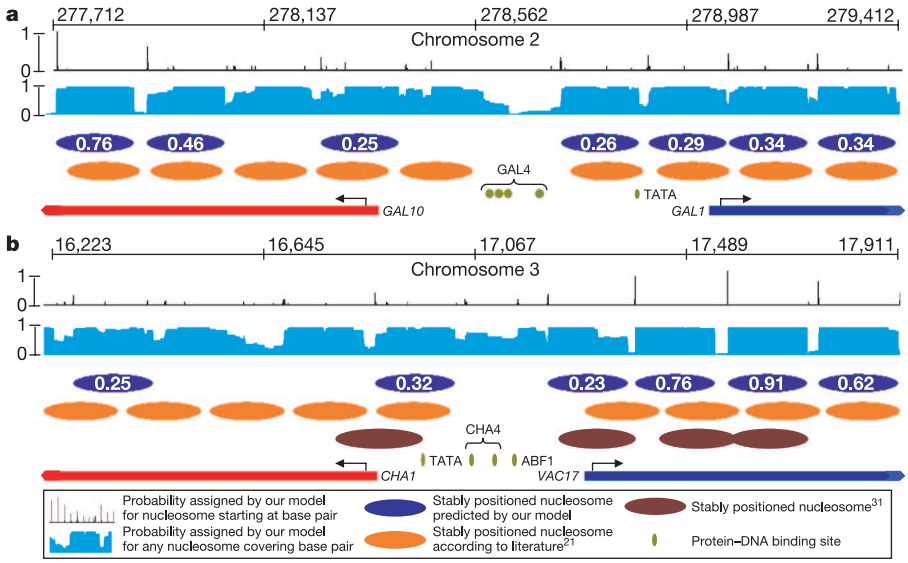

Figure 2 | Genome-wide prediction of intrinsic nucleosome organization and comparison to literature-reported, experimentally identified[31] nucleosome positions. **a**, Detailed view of the *GAL1-10* locus, with literature-reported nucleosome positions[21] (orange ovals). Black trace, probability of a nucleosome starting at each base pair; blue ovals, high probability nucleosomes predicted from our model (probability is indicated); light-blue trace, average occupancy by any nucleosome at each base pair; red and blue bars, protein-coding regions; green ovals, conserved and bound DNA-binding sites[20]. **b**, Same as in **a**, but for the *CHA1* locus[24]; brown ovals, nucleosomes reported from other experiments[31]. The discrepancies between the two sets of literature-reported nucleosome positions highlight the uncertainty in such measurements.

their *in vivo* nucleosome occupancies and comparing them to the occupancies at three nucleosome sites flanking the *GAL1-10* and *PHO5* promoters for which the nucleosome positions are known. Five of the eight predictions tested yielded *in vivo* occupancies comparable to or greater than those of the known nucleosome positions (Fig. 3a), indicating that ~60% of the intrinsically high-occupancy nucleosome sites on the DNA sequence are strongly occupied *in vivo*. In 10 out of 11 cases, these predicted nucleosome positions also had higher occupancy than regions 73 bp (one-half the length of a nucleosome) upstream or downstream from the predicted position (Fig. 3b and Supplementary Figs 31 and 32).

Taken together, these results show that ~50% of the *in vivo* nucleosome organization can be explained solely by the sequence preferences of nucleosomes. Moreover, these results indicate that the nucleosome depletions observed at coding and intergenic regions[29–31] are attributable in part to unstable nucleosomes (that is, positions on the DNA sequence that nucleosomes have a low probability of occupying) encoded in these regions.

Global features of intrinsic nucleosome organization in yeast

We next studied global properties of the intrinsic nucleosome organization in yeast. First, we examined the predicted stability of all 11,802,267 possible genome-wide nucleosome positions; 15,777 were highly stable (stability probability ≥0.5), significantly more than the 10,940 ± 339 ($P < 10^{-20}$) expected by chance. This result may indicate the existence of many genomic locations that encode highly stable nucleosomes, together covering 20% of the genome.

Second, we asked whether individual nucleosomes are organized into higher-ordered nucleosome arrays. The distribution of pairwise distances between positions of the highly stable nucleosomes revealed significant correlations persisting over at least six adjacent nucleosomes, with an average nucleosome repeat length of 177 bp (Fig. 3d). We found similar strong correlations when considering the average nucleosome occupancy predictions (Supplementary Fig. 33). We conclude that the yeast genome not only encodes the preferred positions of individual nucleosomes, but also directly encodes higher structural levels of chromatin organization.

Nucleosome organization varies by type of genomic region

We next asked whether the genome's intrinsic encoding of nucleosome occupancy varies across different types of chromosomal regions, including centromeres, telomeres, intergenic and coding regions, and specific gene classes (Fig. 4a and Supplementary Fig. 34). Indeed, several types of regions had markedly high or low predicted occupancy. The highest predicted occupancy was over centromeres, indicating that centromere function requires enhanced stability of histone–DNA interactions that are encoded in the genomic sequence.

One might think that genomes would facilitate high gene expression levels by encoding unstable nucleosomes over highly expressed genes. Consistent with this expectation, the highly expressed ribosomal RNA and transfer RNA genes stood out as having markedly low predicted nucleosome occupancy.

In contrast to the ubiquitously expressed tRNAs, many other genes vary their expression between high and low levels in different conditions. However, as the genome sequence is static, it cannot simultaneously encode a nucleosome organization that would facilitate both high and low expression levels. Ribosomal proteins are one such example. Our model predicts high nucleosome occupancy encoded over these genes. Thus, the genome sequence does not facilitate the nucleosome depletion[29] and high expression of ribosomal proteins observed during normal growth, which therefore must be governed by other factors. Instead, the genome facilitates the rapid nucleosome reassembly[29] and strong repression of these genes observed under stress[33,34]. These results show how the genome's

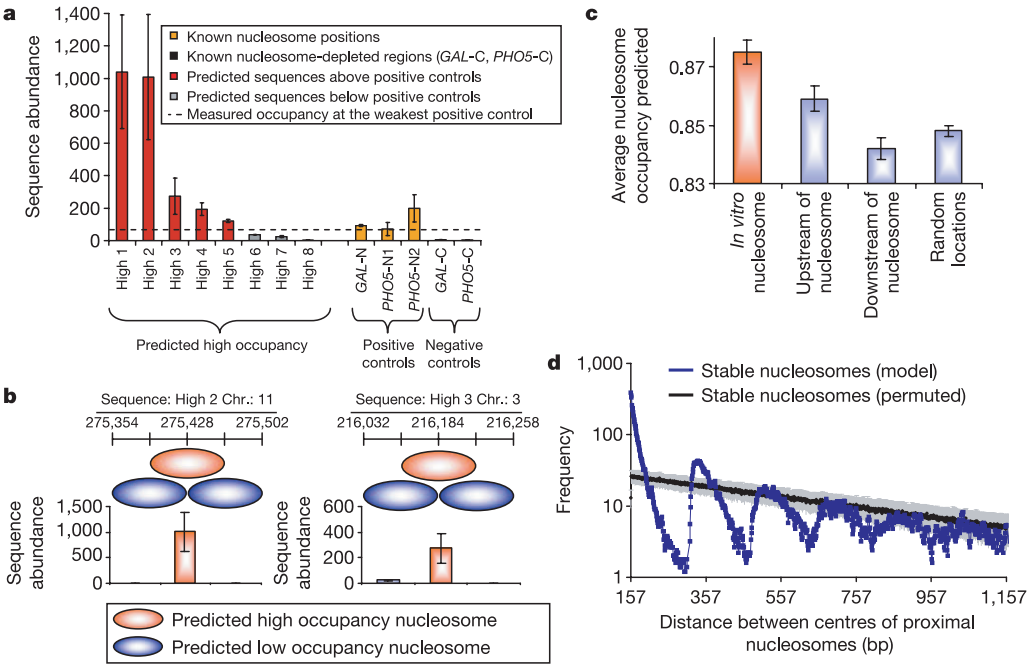

Figure 3 | Higher-order features of intrinsic nucleosome organization and comparison with *in vivo* occupancy experiments. **a**, Experimentally measured nucleosome occupancy *in vivo* for eight high-occupancy predictions, compared with high- and low-occupied locations in the *GAL1-10* and *PHO5* promoters. Error bars are s.d. **b**, *In vivo* nucleosome occupancy measured at predicted low-occupancy regions that are one-half nucleosome distance upstream and downstream (light blue) from the high-occupancy (orange) predictions of **a**. See Supplementary Fig. 31 for additional measurements. Results of **a** and **b** were consistent when normalized for the sequence specificity of micrococcalnuclease (Supplementary Fig. 32). Error bars are s.d. **c**, Predicted nucleosome occupancy in intergenic regions for nucleosomes obtained from an *in vitro* selection experiment (orange) compared with predicted nucleosome occupancy in immediately upstream or downstream locations, or to random genomic locations (light blue). Error bars are s.e.m. **d**, Number of all pairs of proximal stable nucleosomes per centre-to-centre nucleosome distance, compared to the mean (black) and standard deviation (grey) in 100 permutations. Blue, yeast model (stability probability ≥0.5).

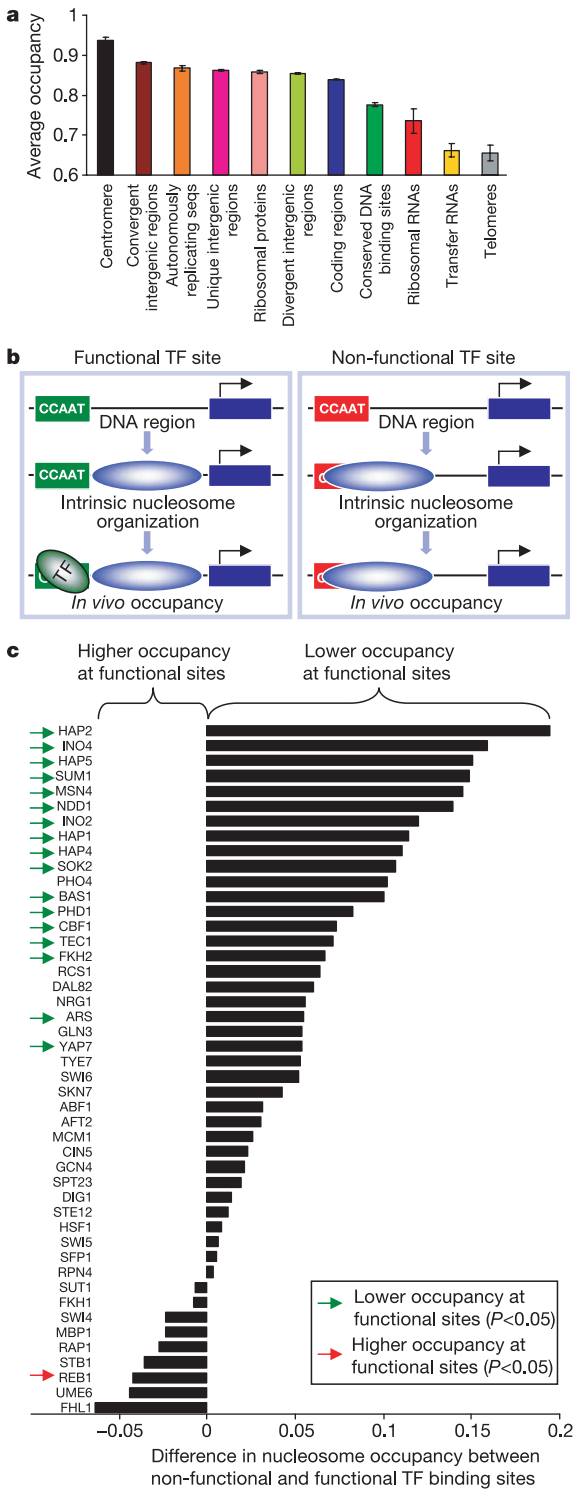

Figure 4 | Intrinsic nucleosome occupancy varies with genomic location type and is low at functional transcription factor binding sites. **a**, Average occupancies and standard errors for different types of genomic regions. **b**, Schematic illustrating how the intrinsic nucleosome organization may facilitate binding of transcription factors (TF) at functional sites, while disfavouring binding at identical non-functional sites that occur by chance. **c**, Difference in predicted nucleosome occupancy between non-functional and functional transcription factor binding sites (absolute occupancy levels are shown in Supplementary Fig. 36). Green arrows, 17 factors having significantly lower nucleosome occupancy at functional sites compared with non-functional sites[20]; red arrow, 1 factor having significantly higher nucleosome occupancy at non-functional sites compared with functional sites.

statically encoded nucleosome organization may contribute to the dynamic process of gene regulation.

Nucleosomes facilitate their own remodelling

We tested whether the variation of nucleosome occupancy that we observed at different types of chromosomal region also extended to

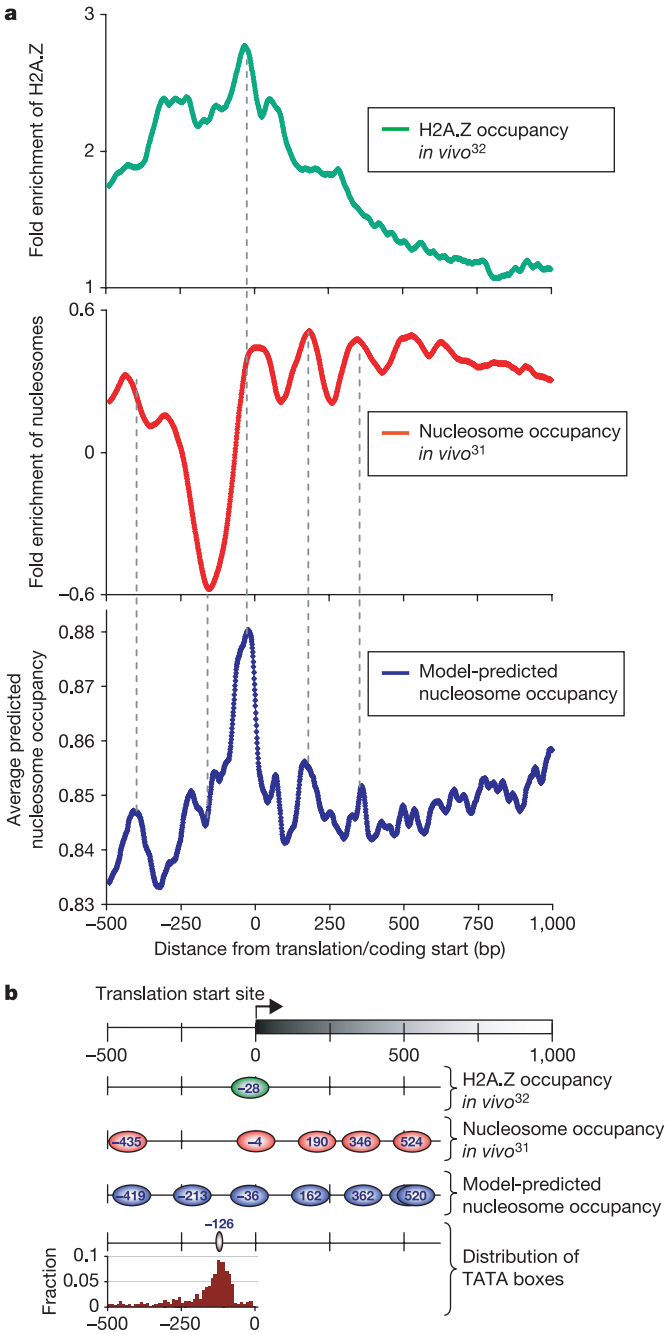

Figure 5 | Genomes encode unstable nucleosomes at transcriptional start sites. **a**, Average across all yeast genes of the *in vivo* occupancy of nucleosomes containing the histone variant H2A.Z[32] (green) or of canonical nucleosomes[31] (red), compared to the nucleosome occupancy predicted by our yeast nucleosome model (blue); all versus distance from the translation open reading frame (ORF) start site. The ORF-proximal peak of our model is statistically significant (Supplementary Fig. 37). **b**, The most probable nucleosome organization, based on **a**. Each nucleosome (ovals; labels represent nucleosome centres) is centred to a corresponding peak in **a**. Bottom graph shows distribution of TATA boxes[42] relative to ORF start sites; brown oval is median TATA box location.

other sets of functionally related genes. We collected 1,949 different sets of yeast genes from a functional gene annotation database[35] and from a wide range of genomic studies[20,36–40], and found that indeed many gene sets showed a significant association with either high or low predicted nucleosome occupancy (Supplementary Fig. 35). Notably, of all gene sets tested, the most significant association predicted low occupancy at regions bound by the chromatin remodelling complex RSC[40] ($P < 10^{-34}$). This implies that genomes facilitate their own chromatin remodelling by encoding intrinsically low nucleosome occupancy at sites destined for remodelling.

Low nucleosome occupancy encoded at functional binding sites

For any given transcription factor, some of its canonical target sites in the genome are occupied by a nucleosome, whereas others are not. Many of the unoccupied sites are thought to occur at random and to be functionally irrelevant[20,41], but the mechanism by which they are kept unoccupied is not known. An intriguing hypothesis is that genomes use their intrinsic nucleosome organization for this task by encoding stable nucleosomes over non-functional sites, thereby decreasing their accessibility to transcription factors (Fig. 4b). We tested this hypothesis by examining our predictions at binding sites for 46 transcription factors. Notably, for 17 (37%) transcription factors the predicted nucleosome occupancy at their functional and conserved DNA binding sites[20] was significantly lower compared with predicted occupancy at their other canonical (but presumed non-functional) sites (Fig. 4c). Only one (2%) factor exhibited significantly higher predicted occupancy at its functional binding sites. These results illustrate how the intrinsic nucleosome organization may help in directing transcription factors towards the appropriate subset of their target sites while excluding them from irrelevant sites.

Low nucleosome occupancy encoded at transcription start sites

Recent nucleosome maps indicate that nucleosomes are depleted from transcriptional start sites[31] (TSSs), but the mechanism for this depletion is not known. For two promoter regions, this depletion was shown experimentally to be intrinsically encoded in the DNA sequence[9]. We asked whether this intrinsically encoded depletion occurs globally by examining the encoded nucleosome organization at all TSSs in yeast (Fig. 5a). We found that the most probable location for TATA elements[42] places them in areas of the genomic sequence that remain unoccupied by nucleosomes; that is, just outside a stably positioned nucleosome (Fig. 5b). Strikingly, the location of the stably positioned nucleosome is conserved across all fungal species (Supplementary Fig. 38). We obtained all of the above results independently, applying both the chicken and yeast models to the yeast genomes. Together, these results may indicate that eukaryotic genomes direct the transcriptional machinery to functional sites by encoding unstable nucleosomes over these elements, thereby enhancing their accessibility.

Conclusions and prospects

Our results establish that nucleosome organization is encoded in eukaryotic genomes. This newly characterized genetic information occurs chromosome-wide, explains ~50% of the *in vivo* nucleosome organization, and may facilitate specific chromosome functions. The consistency between the predictions on the yeast genome using models derived independently from information concerning only yeast or chicken nucleosomes implies that the genomic signals for nucleosome positioning are strong.

Despite its successes, our approach has several limitations and represents only a first step towards understanding the DNA preferences of nucleosomes and the biological implications. First, additional experiments are needed to derive a more accurate nucleosome–DNA interaction model. Second, our representation of nucleosome–nucleosome interactions derived from a thermodynamic model does not yet account for favourable interactions[43], or for the steric hindrance constraints implied by the three-dimensional nucleosome structure. Finally, we examined the intrinsic nucleosome organization without regard for the collection of DNA binding proteins that influence nucleosome positioning by competing for DNA occupancy. At equilibrium, this competition would depend on the concentrations and sequence specificities of both the DNA binding proteins and nucleosomes. The DNA binding proteins have high binding specificity but are present at low concentrations, whereas the nucleosomes have lower binding specificity but are present at high concentrations, covering 75–90% of the DNA. Thus, both are expected to make important contributions to the outcome (Supplementary Figs 39 and 40).

Overall, our results establish that genomes encode the positioning and stability of nucleosomes in regions that are critical for gene regulation and for other specific chromosome functions, and establish that this nucleosome positioning code can be successfully decoded. The genome-wide predictions of nucleosome occupancy and stability that we generated should facilitate the understanding of specific natural gene regulatory phenomena, such as the mechanism by which transcription factors bind preferentially to appropriate sites in promoters rather than to the excess of irrelevant sites in the genome. Our approach may also be useful for improving the performance of engineered transgenes. Our model and results provide a concrete framework for quantitatively integrating chromatin structure into models of gene regulation, and thus represent an essential step towards the goal of developing a quantitative, predictive understanding of transcriptional regulation in all eukaryotes.

METHODS

See Supplementary Information for a more detailed description of the methods.

Molecular biology methods. Mononucleosomes were extracted from log-phase yeast (*Saccharomyces cerevisiae*) cells using standard methods. The DNA was extracted, and protected fragments of length ~147 bp were cloned and sequenced. An *in vitro* selection for nucleosome formation on the yeast genome was performed using purified yeast genomic DNA and substoichiometric purified histone octamer by salt gradient dialysis[44]. The resulting chromatin was treated as for the *in vivo* selection. *In vitro* affinity measurements for core histone $H3_2H4_2$ tetramers were performed as described[44]. *In vivo* nucleosome occupancies were measured as described[9].

Probabilistic nucleosome–DNA interaction model. Given a collection of nucleosome DNA sequences, we aligned all sequences and their reverse complements about their centres, and associated a dinucleotide distribution with each position i, estimated from the combined dinucleotide counts at three neighbouring positions, such that the probability assigned by the model to a 147-bp sequence S is:

$$P(S) = P_1(S_1)\prod_{i=2}^{147} P_i(S_i|S_{i-1})$$

Thermodynamic model for predicting nucleosome positions genome-wide. We used the above probabilistic nucleosome–DNA model within a statistical mechanics framework to compute the nucleosome organization intrinsic to the genomic DNA sequence. We took the partition function to be all 'legal configurations' of nucleosomes on a sequence S, where a legal configuration specifies start positions for a set of non-overlapping 147-bp nucleosomes on S, thus respecting steric hindrance effects between nucleosomes. Using our probabilistic model and an apparent nucleosome concentration parameter, we assigned a statistical weight to each configuration and used the Boltzmann distribution to compute the probability of every configuration. A dynamic programming method[19] was used to efficiently compute the probability that each base pair of S starts a nucleosome or is occupied by a nucleosome.

Additional methods and URLs. For our data, model and genome-wide occupancy predictions, see http://genie.weizmann.ac.il/pubs/nucleosomes06. Our results are also viewable in Genomica (http://Genomica.weizmann.ac.il).

Received 16 March; accepted 14 June 2006.
Published online 19 July 2006.

1. Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
2. Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675 (1986).

3. Widom, J. Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.* **34,** 269–324 (2001).
4. van Holde, K. E. *Chromatin* (Springer, New York, 1989).
5. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293,** 1074–1080 (2001).
6. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98,** 285–294 (1999).
7. Wyrick, J. J. *et al.* Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature* **402,** 418–421 (1999).
8. Trifonov, E. N. Sequence-dependent deformational anisotropy of chromatin DNA. *Nucleic Acids Res.* **8,** 4041–4053 (1980).
9. 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).
10. 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).
11. Thåström, A. *et al.* Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J. Mol. Biol.* **288,** 213–229 (1999).
12. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425,** 737–741 (2003).
13. Cairns, B. R. Chromatin remodeling complexes: strength in diversity, precision through specialization. *Curr. Opin. Genet. Dev.* **15,** 185–190 (2005).
14. Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl Acad. Sci. USA* **95,** 11163–11168 (1998).
15. Wang, J. P. & Widom, J. Improved alignment of nucleosome DNA sequences using a mixture model. *Nucleic Acids Res.* **33,** 6743–6755 (2005).
16. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39,** 1–39 (1977).
17. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* **276,** 19–42 (1998).
18. Widlund, H. R. *et al.* Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.* **267,** 807–817 (1997).
19. Rabiner, L. R. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE* **77,** 257–286 (1989).
20. Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431,** 99–104 (2004).
21. Li, S. & Smerdon, M. J. Nucleosome structure and repair of *N*-methylpurines in the *GAL1–10* genes of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **277,** 44651–44659 (2002).
22. Weiss, K. & Simpson, R. T. High-resolution structural analysis of chromatin at specific loci: *Saccharomyces cerevisiae* silent mating type locus *HMLα*. *Mol. Cell. Biol.* **18,** 5392–5403 (1998).
23. Weiss, K. & Simpson, R. T. Cell type-specific chromatin organization of the region that governs directionality of yeast mating type switching. *EMBO J.* **16,** 4352–4360 (1997).
24. Moreira, J. M. & Holmberg, S. Nucleosome structure of the yeast *CHA1* promoter: analysis of activation-dependent chromatin remodeling of an RNA-polymerase-II-transcribed gene in TBP and RNA pol II mutants defective *in vivo* in response to acidic activators. *EMBO J.* **17,** 6028–6038 (1998).
25. Shimizu, M., Roth, S. Y., Szent-Gyorgyi, C. & Simpson, R. T. Nucleosomes are positioned with base pair precision adjacent to the α2 operator in *Saccharomyces cerevisiae*. *EMBO J.* **10,** 3033–3041 (1991).
26. Kent, N. A., Tsang, J. S., Crowther, D. J. & Mellor, J. Chromatin structure modulation in *Saccharomyces cerevisiae* by centromere and promoter factor 1. *Mol. Cell Biol.* **14,** 5229–5241 (1994).
27. Verdone, L., Camilloni, G., Di Mauro, E. & Caserta, M. Chromatin remodeling during *Saccharomyces cerevisiae ADH2* gene activation. *Mol. Cell Biol.* **16,** 1978–1988 (1996).
28. Almer, A., Rudolph, H., Hinnen, A. & Horz, W. Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements. *EMBO J.* **5,** 2689–2696 (1986).
29. Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genet.* **36,** 900–905 (2004).
30. Bernstein, B. E., Liu, C. L., Humphrey, E. L., Perlstein, E. O. & Schreiber, S. L. Global nucleosome occupancy in yeast. *Genome Biol.* **5,** R62 (2004).
31. Yuan, G. C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309,** 626–630 (2005).
32. Guillemette, B. *et al.* Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS Biol.* **3,** e384 (2005).
33. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–4257 (2000).
34. Kim, J. & Iyer, V. R. Global role of TATA box-binding protein recruitment to promoters in mediating gene expression profiles. *Mol. Cell Biol.* **24,** 8104–8112 (2004).
35. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25,** 25–29 (2000).
36. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126 (2000).
37. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292,** 929–934 (2001).
38. Sudarsanam, P., Iyer, V. R., Brown, P. O. & Winston, F. Whole-genome expression analysis of *snf/swi* mutants of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **97,** 3364–3369 (2000).
39. Casolari, J. M. *et al.* Genome-wide localization of the nuclear transport machinery couples transcriptional status and nuclear organization. *Cell* **117,** 427–439 (2004).
40. Robert, F. *et al.* Global position and recruitment of HATs and HDACs in the yeast genome. *Mol. Cell* **16,** 199–209 (2004).
41. Li, H., Rhodius, V., Gross, C. & Siggia, E. D. Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl Acad. Sci. USA* **99,** 11772–11777 (2002).
42. Basehoar, A. D., Zanton, S. J. & Pugh, B. F. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116,** 699–709 (2004).
43. Cui, Y. & Bustamante, C. Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure. *Proc. Natl Acad. Sci. USA* **97,** 127–132 (2000).
44. Thåström, 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).

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

Acknowledgements We thank A. Travers for providing the chicken nucleosome core DNA sequences; M. Kubista for providing selected mouse DNA sequences; O. Rando for providing access to their nucleosome data before publication; J. Lieb, E. Nili and P. Jones for sharing their respective unpublished data; Y. Lubling for creating the supplementary website; and H. Chang, N. Friedman, U. Gaul, A. Matouschek, B. Meyer, M. Ptashne, E. Siggia and A. Tanay for useful comments on the manuscript. E.S. was supported by a fellowship from the Center for Studies in Physics and Biology at Rockefeller University and by an NIH grant. J.W. thanks the Center for their hospitality during a sabbatical. J.-P.Z.W. acknowledges support from an NIH grant and J.W. acknowledges support from two NIH grants. E.S. is the incumbent of the Soretta and Henry Shapiro career development chair.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.S. (eran.segal@weizmann.ac.il) or J.W. (j-widom@northwestern.edu).