nature genetics

A high-resolution atlas of nucleosome occupancy in yeast

William Lee[1,2], Desiree Tillo[3], Nicolas Bray[3], Randall H Morse[4], Ronald W Davis[1,2], Timothy R Hughes[3,5,6] & Corey Nislow[3,5,6]

We present the first complete high-resolution map of nucleosome occupancy across the whole *Saccharomyces cerevisiae* genome, identifying over 70,000 positioned nucleosomes occupying 81% of the genome. On a genome-wide scale, the persistent nucleosome-depleted region identified previously in a subset of genes demarcates the transcription start site. Both nucleosome occupancy signatures and overall occupancy correlate with transcript abundance and transcription rate. In addition, functionally related genes can be clustered on the basis of the nucleosome occupancy patterns observed at their promoters. A quantitative model of nucleosome occupancy indicates that DNA structural features may account for much of the global nucleosome occupancy.

The genomes of all eukaryotic organisms are packaged into nucleosomes, comprising 147-bp segments of DNA wrapped around a histone octamer. Nucleosomes are separated by linker DNA and form the basis of higher order packaging of genomes into chromatin[1,2]. Much information has been collected regarding the relationship between chromatin structure, histone modifications and the control of gene expression[3]. For example, in-depth analysis of the *PHO5* and *GAL1–GAL10* promoters in yeast shows that nucleosome positioning can be a chief determinant in regulating gene expression[4–6]. Large-scale studies of nucleosome positions have shown that promoters are often depleted of nucleosomes[7–9], and computational models suggest there may be intrinsic cues for nucleosome occupancy encoded across the genome[10,11].

Here we present the first complete, experimentally obtained, high-resolution nucleosome map of a eukaryotic genome. Our analysis confirms that intergenic regions are depleted of nucleosomes relative to genes[7–9]. We identify a general pattern of nucleosome occupancy that borders the transcriptional unit and is anchored at the transcription start site (TSS)[8]. Our atlas of nucleosome occupancy permits analysis of nucleosome positions across the genome. For example, clustering reveals that functionally related genes share nucleosome occupancy patterns in their promoters. By combining these data with global studies of transcription[12], protein-binding sites[13,14] and computational models[10,11], we considerably extend the characterization of genome structure and gene architecture in yeast. Indeed, taking elements from each of these data sources enables us to generate a model that predicts a large fraction of the nucleosome occupancy of the genome.

RESULTS

Data collection

To prepare enriched nucleosomal DNA, we treated genomic chromatin with micrococcal nuclease[8], which preferentially digests linker segments between nucleosomes. Primarily mononucleosomal samples were prepared from haploid yeast collected in the logarithmic phase of growth in YPD medium. Three independent samples of nucleosomal and total genomic DNA were hybridized to an Affymetrix tiling microarray with 4-bp resolution[12].

As an initial validation of the data, we confirmed that auxotrophic gene-deletion markers of the yeast strain BY4741 were detectable on the array as contiguous regions of no signal (**Supplementary Fig. 1** online). We also confirmed that nucleosomes at several well-characterized loci, including the benchmark *HIS3* and *CHA1* promoters[15,16], were consistent with previously established positions (**Fig. 1a**,**b**). Specifically, the local peaks of signal intensity in our data corresponded well with the centers of previously mapped nucleosomes. Manual comparison also revealed congruence with a similar study that assayed nucleosome positioning with 20-bp resolution on only a fraction of the genome (chromosome 3 and select promoters) (**Fig. 1**, blue graph; ref. 8). Direct comparison of our 4-bp resolution data with the previous data confirmed that linker regions between nucleosomes are often short (<50 bp) and, as a consequence, resolution benefits from additional probe coverage. In some cases, for example, 20-bp probe coverage provides only one or two probes that span an identified segment of linker DNA, suggesting that our greater probe coverage substantially improves data quality (**Fig. 1**).

To allow an overall comparison to the Yuan *et al.* data set[8] (which, in addition to being lower resolution, seems to have a nonlinear scaling relationship to our data; **Fig. 1**), we obtained the Hidden Markov Model (HMM) code used in their study and adapted the parameters to our higher-resolution data (see Methods). On the basis of the HMM nucleosome calls (see **Fig. 1** for examples), 81% of the chromosome 3 sequence analyzed by Yuan *et al.*[8] is occupied by nucleosomes in our study. Among this sequence, 84% overlaps with

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305-5120, USA. [2]Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304, USA. [3]Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. [4]Wadsworth Center, New York State Department of Health, Albany, New York 12201-2002, USA. [5]Terrence Donnelly Centre for Cellular & Biomolecular Research, 160 College Street, Toronto, Ontario M5S 3E1, Canada. [6]Banting and Best Department of Medical Research, 112 College St., Toronto, Ontario M5S 1L6, Canada. Correspondence should be addressed to C.N. (corey.nislow@utoronto.ca).

Received 23 March; accepted 16 July; published online 16 September 2007; doi:10.1038/ng2117

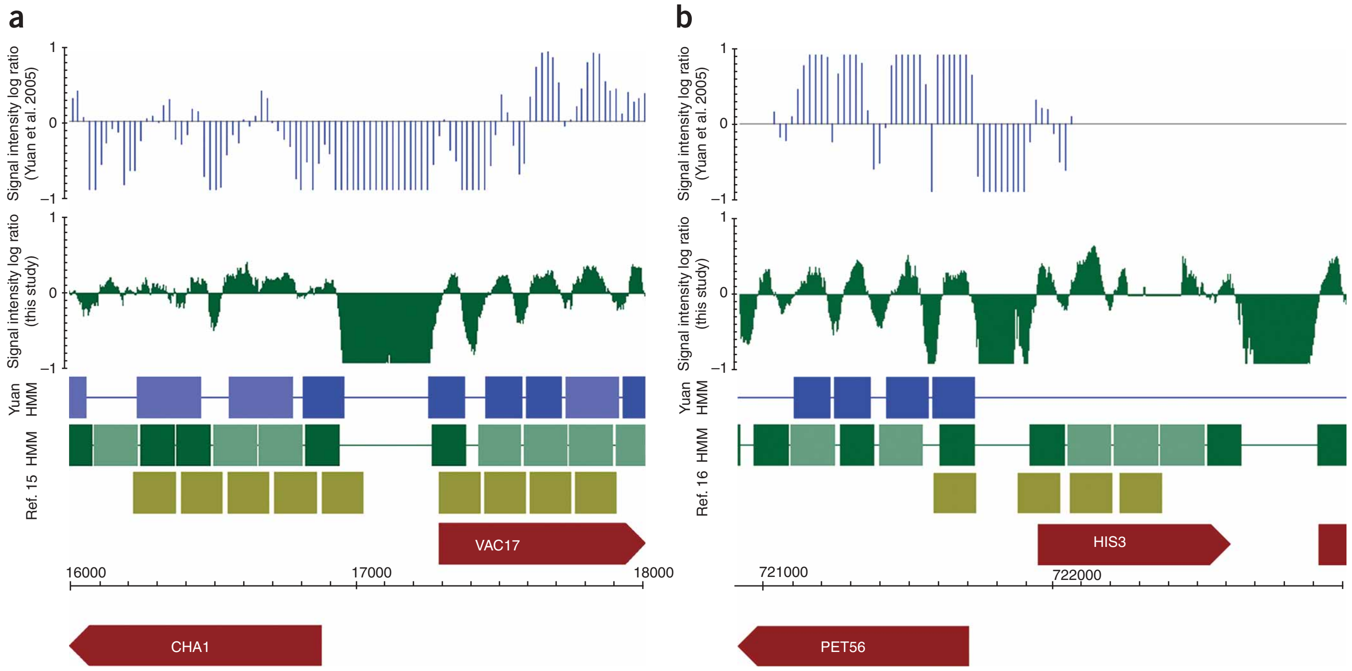


Figure 1 Distribution of nucleosomes around the *CHA1* and *HIS3* promoters. (**a**) A 2-kb region on chromosome 3 surrounding the *CHA1* promoter. Blue graph is derived from Yuan *et al.*[8]; each vertical line represents the average probe intensity log ratio between nucleosomal and whole-genome DNA at that position. Ratios are represented on a $\log_2$ scale (the graph has been truncated at -1 to allow closer inspection of positioned nucleosomes); a positive signal represents nucleosomal occupancy. Green graph represents data from this study; individual probes are represented by vertical lines but, owing to the data density, the probes are too close to distinguish. Blue and green boxes represent HMM-predicted nucleosome positions derived from Yuan *et al.*[8] and this study, respectively; edges have been trimmed slightly to make them more distinct, and the more lightly shaded boxes represent delocalized nucleosome calls. Gold boxes represent the position of nucleosome locations determined by Moreira and Holmberg[15]. Red boxes indicate annotated ORFs; the arrows represent the direction of transcription. (**b**) Same as in **a**, but showing the *HIS3* region on chromosome 15 (nucleosome locations determined by Sekinger *et al.*[16]).

regions occupied in the Yuan *et al.* study[8] (versus the 60% expected; hypergeometric $P < 2.22 \times 10^{-308}$), indicating a significant correspondence between the two data sets. However, only 32% of the centers of our well-positioned nucleosomes are within 10 bp (51% within 20 bp) of the centers of well-positioned nucleosomes in the Yuan *et al.* study[8], further indicating that the data sets are not identical (**Fig. 1**).

The HMM also generated an index of nucleosome positions across the whole genome (**Table 1**), detecting 40,095 well-positioned and 30,776 'fuzzy' nucleosomes. These 70,871 nucleosomes encompass 9.8 Mb or 81% of the non-repetitive genome. Nucleosomes encompass 87% of transcribed sequence but only 53% of intergenic sequence. Similar estimates are obtained by counting the proportion of probes detecting above the average linker signal intensity learned by the HMM ($\log(\text{ratio}) = -0.66$): 91.6% of transcribed sequence was above this linker threshold as compared with 57.3% of intergenic sequence.

Because the HMM nucleosome calls do not explicitly capture many of the subtleties in the relative occupancy data, including differences in internucleosomal occupancy (for example, **Fig. 1**), our subsequent analyses focused on occupancy ratios from the smoothed tiling path data.

General features

Several independent studies of nucleosome occupancy at varying degrees of resolution concur that intergenic regions are depleted of nucleosomes as compared with coding regions[7,9], a trend that is also obvious in our data (**Figs. 1** and **2a**). For example, Yuan *et al.*[8] identified nucleosome-depleted regions (NDRs) of ~150 bp positioned, on average, ~200-bp upstream of annotated genes. If promoter depletion of nucleosomes is associated with promoter activity, then we expect NDRs to align with TSSs as opposed to start codons. To examine this, we calculated the average nucleosome occupancy of probes within a region 50-bp upstream and 50-bp downstream of the TSS of 5,015 high-confidence transcripts (see Methods) derived from a high-resolution map of the yeast transcriptome[12]. We found that the 50-bp region that lies upstream of the TSS of verified transcripts is far more depleted of nucleosomes than are the corresponding downstream regions (**Fig. 2a**). This observation is consistent with the view that promoters of active genes are depleted of nucleosomes, presumably to permit protein-binding events required for gene expression activity[7–9,17].

Gene expression has been reported to correlate inversely with nucleosome occupancy in promoters: highly expressed genes contain

Table 1 Nucleosome content of the genome

	Coverage (bp)	Fraction of total genome (total intergenic / total transcribed)	Number of nucleosomes
Array probe coverage	12,068,004	1 (1 / 1)	N/A
Well-positioned nucleosomes	4,970,908	0.41 (0.36 / 0.42)	40,095
Delocalized (fuzzy) nucleosomes	4,801,292	0.4 (0.17 / 0.45)	30,776
Total nucleosomal DNA	9,772,200	0.81 (0.53 / 0.87)	70,871
Non-nucleosomal ('linker') DNA	2,295,804	0.19 (0.47 / 0.13)	32.4 bp average length

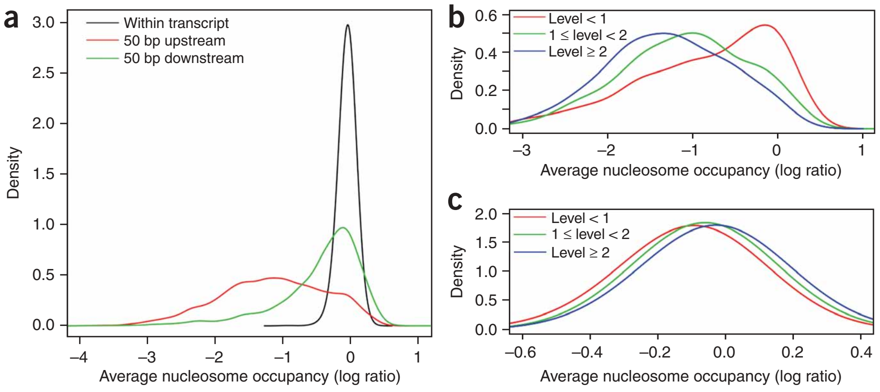


Figure 2 TSSs are demarcated by NDRs. (**a**) Kernel density plot showing the distribution of nucleosome occupancy for regions surrounding verified transcription segments that overlap ≥50% of a verified gene on the 5′ end, as defined in ref. 12 ($n = 5{,}015$). Red shows the distribution of average nucleosome occupancy within a region 50-bp upstream of a verified transcript, green shows occupancy within a region 50-bp downstream, and black shows occupancy within the transcript. (**b**) Nucleosome occupancy within a region 50-bp upstream of verified transcription segments as separated by transcription level. Red shows the distribution of average nucleosome occupancy for promoters of segments with expression level < 1 ($n = 759$), green shows that for segments with expression level between 1 and 2 ($n = 1{,}859$), and blue shows the most highly expressed genes with level ≥ 2 ($n = 2{,}397$). (**c**) Same as in **b**, but showing average nucleosome occupancy within verified transcripts.

prominent NDRs[7,8,17], and genes expressed at low levels tend to have promoters that are more occupied by nucleosomes. We find a similar trend in our data (**Fig. 2b**). When we divided the 5,015 verified transcripts into three classifications on the basis of transcript abundance[12] and tested whether expression correlated with a particular nucleosome distribution, we found that the distribution of nucleosome occupancy in the promoters of these three sets of transcripts is distinct.

Our data also show that nucleosome occupancy within coding regions correlates with transcription level, but in the opposite manner. Specifically, highly expressed genes are significantly more occupied by nucleosomes than are genes expressed in small amounts or not at all (**Fig. 2c**). This observation also holds true when genes are separated by transcriptional frequency rather than by steady-state mRNA levels (ref. 18 and data not shown). When the nucleosome occupancies of genes expressed in different amounts are compared, the distinctions are statistically significant (t-test, $P < 1 \times 10^{-15}$ for all three expression levels). A possible explanation for the observed patterns of occupancy is that the act of transcription promotes or requires formation of the ordered nucleosome structures that we observe within genes, perhaps by increasing residence time of the Rpd3S complex[19,20].

When we examined regions of the genome on the basis of their annotations, we found that nucleosome occupancy preferences generally depend on the genomic feature. For example, coding regions and centromeres are the most highly occupied, whereas unique and divergent intergenic regions (that is, unidirectional and bidirectional promoters) are the most depleted of nucleosomes (data not shown). Here, we define intergenic regions as sequences that do not encode protein (including 5′ and 3′ UTRs). Although all intergenic regions are nucleosome depleted relative to protein-coding regions, convergent intergenic segments that are unlikely to be promoters are significantly more occupied than unique and divergent regions (Mann-Whitney test, $P < 2.2 \times 10^{-16}$), consistent with the idea that the most depleted intergenic regions are promoters.

TSSs define a nucleosome occupancy signature

The hypothesis that promoters define the boundary of nucleosome-free regions is supported by gene-by-gene observations. For example, the *SAC7* gene shows an extended 5′ UTR of greater than 500 bp when its ORF is aligned with its corresponding transcription segment[12]. In this case, the NDR is directly upstream of the TSS and a nucleosome is positioned precisely at the start of the transcript (**Fig. 3a**). On a global scale, when the nucleosome occupancy patterns of all promoters are averaged, the NDR is evident and most genes contain a consistent ladder of well-positioned nucleosomes at their 5′ ends immediately downstream from the NDR (**Fig. 3**). By precisely aligning nucleosome occupancy signal by TSSs and averaging all genes, the average nucleosome signature is clearly oriented at TSSs (**Fig. 3b**,

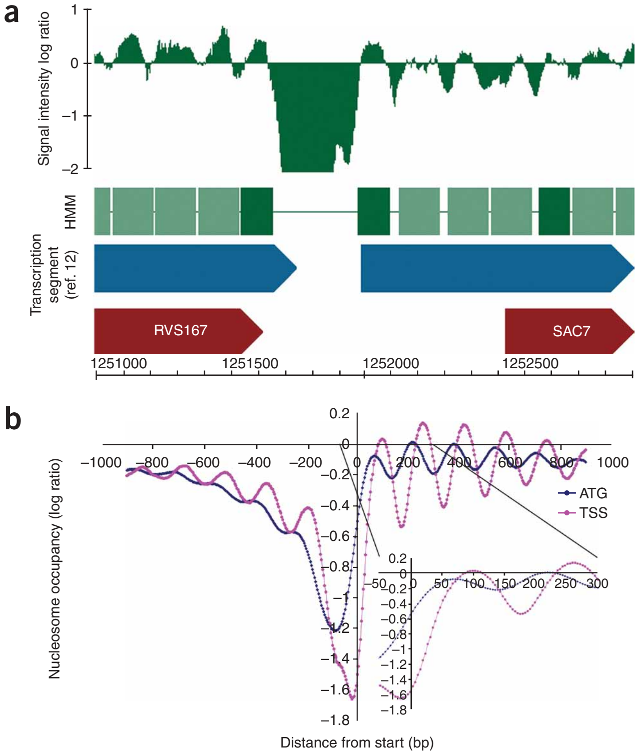
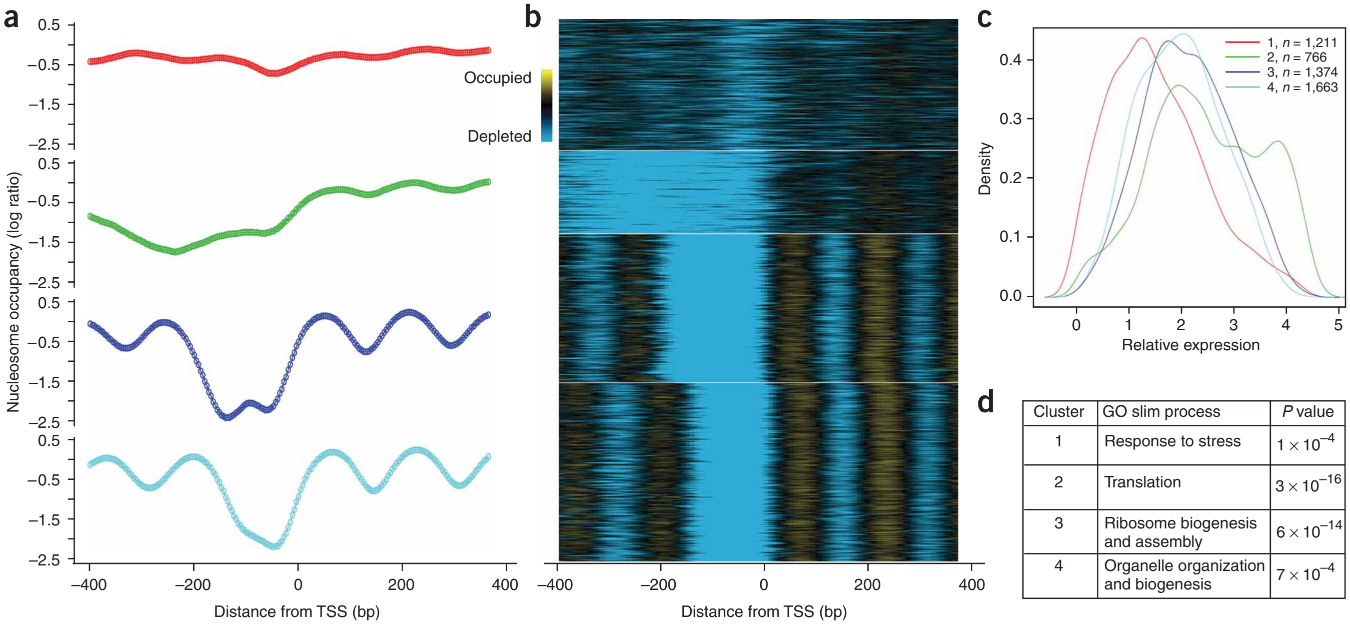


Figure 3 NDRs align with TSSs and not with translation start sites. (**a**) Nucleosome occupancy within a 2-kb region on chromosome 4. Green graph represents average probe intensity ratio ($\log_2$ scale), dark green boxes show the location of HMM-called well-positioned nucleosomes, lighter green boxes are delocalized nucleosomes, blue boxes are transcription segments from David *et al.*[12], red boxes are annotated ORFs (derived from the Saccharomyces Genome Database), and arrows denote the direction of transcription. *SAC7* shows a 5′ UTR of 536 bp (ref. 12). (**b**) Average nucleosome occupancy surrounding TSSs for the ensemble of verified transcripts. This graph shows the log ratio of nucleosome occupancy plotted against genomic coordinate relative to the ATG start codon of an ORF (blue) or the TSS[12] (magenta). Inset highlights the data between −50 and +300 of the TSS.



Cluster	GO slim process	*P* value
1	Response to stress	1×10^{-4}
2	Translation	3×10^{-16}
3	Ribosome biogenesis and assembly	6×10^{-14}
4	Organelle organization and biogenesis	7×10^{-4}

Figure 4 Clustering promoter nucleosome signatures. (**a**) Average nucleosome occupancy for each cluster, covering ~800 bp surrounding the TSS. The clustergram was generated with *k*-means clustering using the Euclidean distance metric. The four clusters contain 1,211, 766, 1,374 and 1,663 transcripts. (**b**) *k*-Means clustergram for the set of ~5,000 verified transcripts with known annotations. Blue represents areas depleted for nucleosomes; yellow areas are more occupied. (**c**) Kernel density plot showing distribution of expression levels for transcripts in each cluster. (**d**) The GO Slim biological process term most overrepresented by genes in each cluster and the hypergeometric *P* value for each term.

magenta graph). This transcript-aligned, promoter-specific signature comprises an NDR centered immediately (<50 bp) upstream of the TSS flanked by nucleosomes centered at about −200 bp and +100 bp. Aligning nucleosome occupancy signals by ORF start codons reveals a similar, less-pronounced signature and an NDR further upstream from the ATG start (**Fig. 3b**, blue graph).

The nucleosomes downstream from the NDR are consistently positioned; in particular, the first nucleosome immediately downstream from the NDR in the coding region is very well positioned. Other studies have shown that there is a genome-wide bias for the positioning of this initial nucleosome[10,11], and experimental data indicate the histone variant H2A.Z is preferentially deposited in nucleosomes flanking nucleosome-free regions[21]. The resolution and sensitivity of the array also enable us to detect a nucleosomal ladder downstream from the transcript start in the global nucleosome signature. The frequency of nucleosome occupancy upstream of the TSS is unchanged, but the amplitude of the signature is lower.

Nucleosome occupancy signatures correlate with transcript level

To categorize promoters across the genome, we aligned and clustered genes on the basis of nucleosome occupancy surrounding their TSSs, and tested whether genes with a similar nucleosome signature shared other attributes. We used *k*-means clustering to group nucleosome occupancy profiles in a window of ~800 bp surrounding the TSS of 5,015 verified transcripts (**Fig. 4**). We found only four significantly distinct clusters of nucleosome occupancy; increasing the number of clusters resulted in a larger number of smaller clusters, but had no qualitative effect on the observed Gene Ontology (GO) enrichments (**Supplementary Fig. 2** online). *k*-Means clustering with values of $k > 3$ consistently produced a discrete cluster whose genes seem to lack a significant NDR and are instead highly occupied with nucleosomes (cluster 1, **Fig. 4** and **Supplementary Fig. 2b**,**c**). These genes are interesting because their promoters (**Fig. 4a**) are clearly different from the typical genome-wide signature (**Fig. 3b**) and the cluster is statistically enriched for genes of unknown function (cumulative hypergeometric $P = 1.7 \times 10^{-13}$). Removing these uncharacterized genes does not, however, qualitatively change the results (**Supplementary Fig. 3a–c** online). Hypergeometric distribution analysis of GO Slim annotations reveals that cluster 1 is also enriched for genes involved in response to stress ($P = 1\times\ 10^{-4}$; **Fig. 4d**); these genes are unlikely to be expressed during logarithmic growth in rich media. The other three clusters showed enrichment of at least one GO Slim process annotation[22]. Cluster 2 shows a bimodal distribution of expression (**Fig. 4c**), which is explained by the strong enrichment of genes involved in protein biosynthesis in this cluster, particularly the highly expressed ribosomal protein genes[18]. This enrichment was confirmed by repeating the analysis after removing ribosomal protein genes (**Supplementary Fig. 3d**). Cluster 3 is enriched for genes involved in ribosome biogenesis and assembly. Cluster 4 is the largest cluster and is most enriched for genes involved in protein modification with a significant enrichment of DNA and RNA metabolism genes ($P < 1 \times 10^{-3}$).

Our clustering results differ from those obtained from a similar analysis based on computationally calculated nucleosome occupancy[10]. That study found three clusters that segregated primarily on the basis of the distance between a predicted TATA sequence and start codon. Our clustering assesses nucleosome occupancy from experimental data, suggesting that nucleosome occupancy often reflects function (on the basis of GO) and gene expression.

Nucleosome occupancy correlates with base composition

Having found a set of characteristic nucleosome occupancy patterns, we sought to identify mechanisms that could best explain the observed

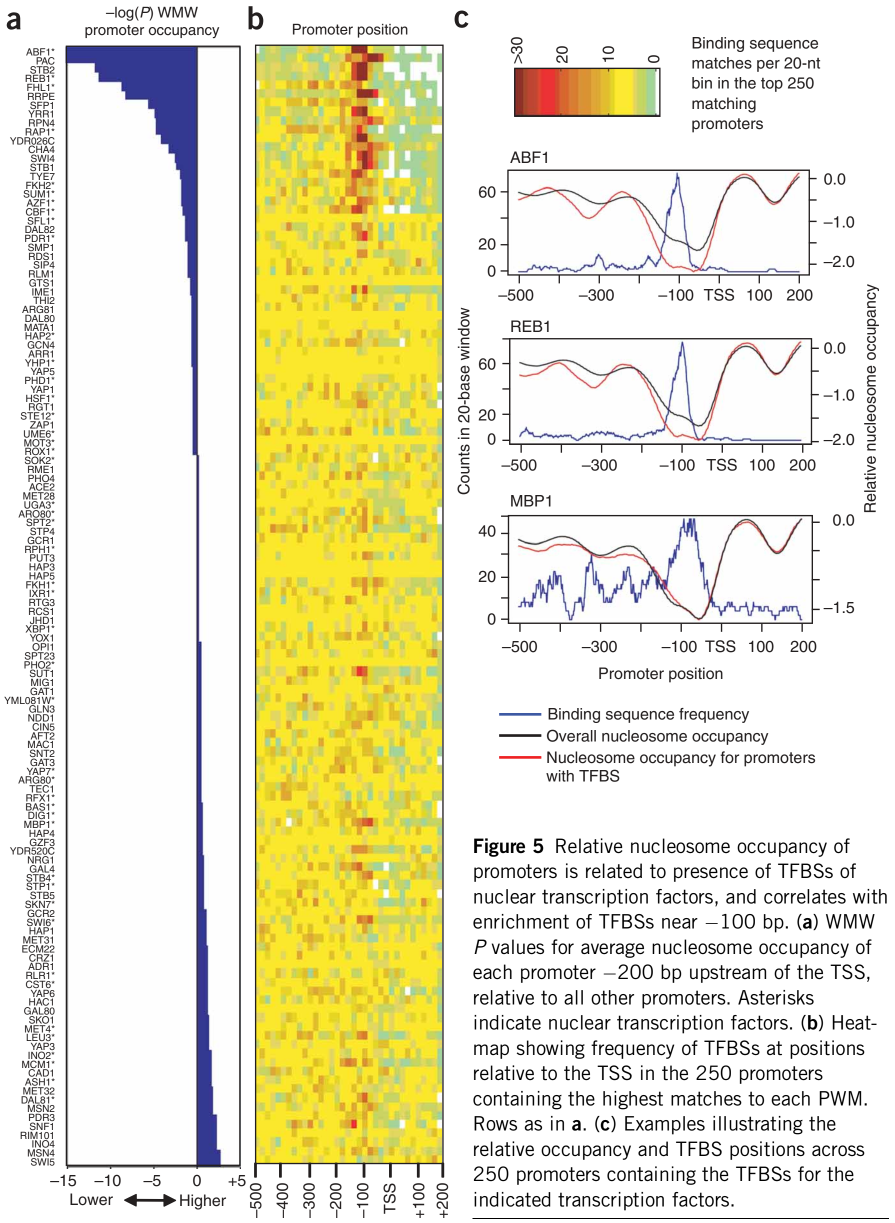


Figure 5 Relative nucleosome occupancy of promoters is related to presence of TFBSs of nuclear transcription factors, and correlates with enrichment of TFBSs near –100 bp. (**a**) WMW *P* values for average nucleosome occupancy of each promoter –200 bp upstream of the TSS, relative to all other promoters. Asterisks indicate nuclear transcription factors. (**b**) Heatmap showing frequency of TFBSs at positions relative to the TSS in the 250 promoters containing the highest matches to each PWM. Rows as in **a**. (**c**) Examples illustrating the relative occupancy and TFBS positions across 250 promoters containing the TFBSs for the indicated transcription factors.

nucleosome positioning and occupancy. Previous work in predicting nucleosome positions focused on periodic dinucleotide patterns, which were identified on the basis of DNA sequences with high affinity for DNA *in vitro*[10,11,23]. These patterns seem to explain some nucleosome positioning *in vivo*; however, the correlation is not sufficiently strong to infer exclusive causation[11]. Other structural and simple-sequence features of DNA have been shown or proposed to influence nucleosome affinity for specific genomic sites. For example, CTG repeats are the strongest reported naturally occurring nucleosome positioning sequences[24], whereas poly(dA-dT) tracts act as nucleosome-excluding sequences[25]. Z-DNA, which forms as a result of DNA unwinding from the nucleosome during transcription, also inhibits nucleosome formation[26]. In addition, sequence-specific DNA-binding transcription factors must compete and/or interact with nucleosomes in the formation of chromatin[27]. Indeed, a key trend in our data and previous analyses[7–9] is a prominent depletion of nucleosome occupancy just upstream of TSSs (**Fig. 3**). Consistent with the idea that transcription factors have a role in nucleosome positioning, we identified binding sequences for Abf1 and Reb1 by Gibbs sampling among the least-occupied sequences[28] (data not shown). Both of these transcription factors have been previously implicated in dictating chromatin architecture[29–31].

We therefore tested whether the presence of transcription factor binding sequences (TFBSs) alone (in the absence of data on actual binding, we considered whether the transcription factors are active in YPD medium) correlates with nucleosome occupancy, and if the locations of TFBSs reflect the stereotypic nucleosome architecture of promoters. We first scored all yeast promoters for a match to the position weight matrix (PWM) of 126 known and predicted TFBSs (see Methods). We used transcription factor localization data[32] as a proxy for transcription factor activity, assuming that the 46 transcription factors localized exclusively in the nucleus are active and all of the others are inactive. Although at best an approximation, some transcription factors are known to be regulated by localization or synthesis (for example, Crz1, Gcn4, Yap1, Swi5 and Pho4; Saccharomyces Genome Database). In addition, many of the nuclear transcription factors regulate constitutive activities such as the cell cycle and protein synthesis, and five of seven of the essential transcription factors with a PWM (Abf1, Hsf1, Mcm1, Rap1 and Reb1) are categorized as nuclear.

We calculated Wilcoxon-Mann-Whitney (WMW) *P* values for the average promoter occupancy across the promoters with the top 250 matches to the PWM for each transcription factor (assuming that the highest-scoring binding site in a promoter was the only binding site) as compared with the occupancy of all other promoters (**Fig. 5a**). Two main trends are evident. First, there is a strong statistical correspondence between nucleosome occupancy and presence of a binding sequence for a nuclear transcription factor ($P < 10^{-15}$). Second, the TFBSs for transcription factors whose binding sequences are least occupied by nucleosomes tend to cluster at a position 80–100 bp upstream of the TSS, and this clustering is also observed to a lesser degree for TFBSs in general (**Fig. 5b**). In promoters containing these TFBSs, the location of the TFBS corresponds to part, albeit not all, of the trough in nucleosome occupancy (examples are shown in **Figure 5c**). Although the trend is statistically significant, the separation between promoters with binding sites of nuclear transcription factors and all other promoters is not nearly as great as the separation between promoters and transcribed sequences in general (**Figs. 3** and **5c**; and data not shown). Thus, although TFBSs do, in some cases, have a strong positional trend relative to the TSS, and this position is within the nucleosome-free promoter region, it seems unlikely that displacement of nucleosomes by transcription

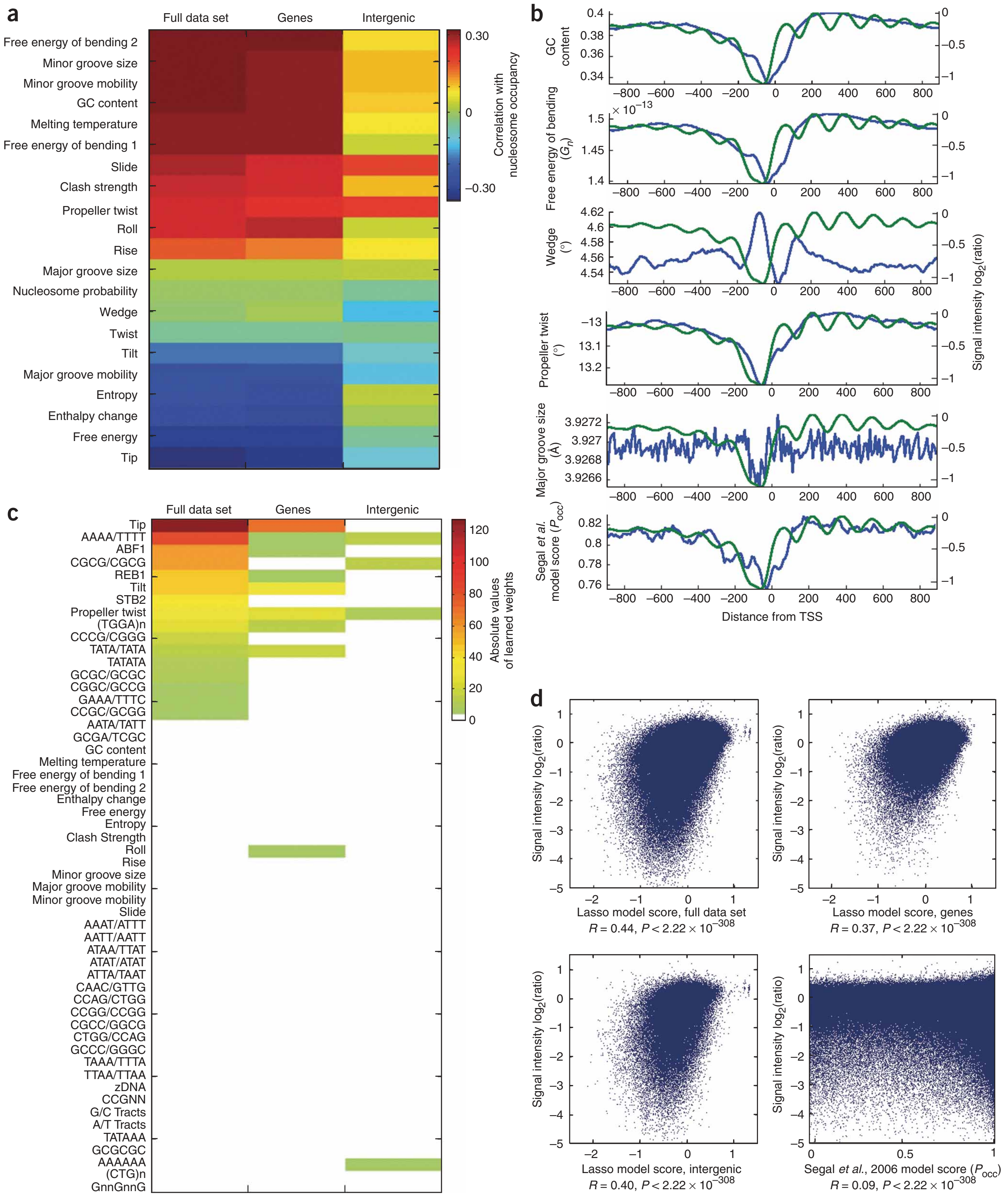


Figure 6 Correlations between nucleosome occupancy measurements and local DNA properties. (**a**) Pairwise correlations between smoothed nucleosome occupancy and selected DNA properties, as indicated, over all measurements. (**b**) Average shape of smoothed nucleosome occupancy data (green line) and selected DNA properties (blue line) over 5,037 promoters. (**c**) Weights learned by Lasso for indicated DNA properties over all measurements for chromosomes 1–8. (**d**) Correlations between the Lasso model and Segal *et al.*'s model[11] and actual smoothed measurements over chromosomes 9–16. *P* values are from rank correlations.

factors can explain nucleosome occupancy patterns across the genome. This observation is consistent with previous data indicating that nucleosome occupancy does have a strong influence on transcription factor occupancy[17,33].

We therefore tested whether other sequence features could explain the observed data. To develop a predictive model that combines several features, we made the assumption that such features should follow the same additive formula for any stretch of nucleosome-sized DNA. Our model ignores the effects of neighboring nucleosomes[34]; however, it has the advantage that the inputs can be weighted by using well-established regression methods. In addition, the existing Lasso algorithm used here[35] can choose between related features that have overlapping impact, giving greater weight to the feature that provides the greatest explanation. If the alternative feature provides additional information, it will receive an appropriate weight.

We used Lasso to build a linear model (in other words, a formula) that includes both TFBSs and other features known or thought to influence nucleosome positioning. We first selected those features with high independent correlation or discrimination power, as judged by Pearson correlations and/or WMW *P* values. For example, G+C content and many features that are based on dinucleotide composition correlate with nucleosome occupancy (**Fig. 6a**); in fact, most of these features have an average profile across all promoters that is correlated with nucleosome occupancy (**Fig. 6b**). We then trained Lasso, using these features and all of our data from chromosomes 1–8, to produce a linear model in which the input features predict the microarray measurements.

The weights that Lasso assigned to each feature are shown in **Figure 6c**. Different results were obtained depending on whether Lasso was run on all sequence, on transcripts alone or on intergenic regions alone, suggesting that relative nucleosome occupancy may be dictated by different mechanisms in intergenic and genic regions. Nonetheless, in all cases, the highest weights were assigned to DNA structural features (tip, tilt and/or propeller twist), suggesting that energetic costs of DNA deformation account for much of the overall nucleosome occupancy across the yeast genome. Tip and tilt are among the most periodic parameters in the conformation of nucleosomal DNA[2]. The propeller twist capacity, which describes the angle of the two aromatic bases in a base pair, was selected in all three cases; this parameter is the most strongly correlated with nucleosome occupancy in intergenic regions (**Fig. 5**). In general, dinucleotides with highly negative propeller twist angles are more rigid than dinucleotides with smaller propeller twist angles[36] and, consistent with rigidity having a role in nucleosome exclusion from promoters[8], 'AAAA' was also selected by Lasso in all three cases; AAAA is the most rigid of all possible tetranucleotides. Among the TFBSs considered, only three (ABF1, REB1 and STB2) were selected as independently informative. These factors have each been previously implicated in chromatin function[31,37].

After training on chromosomes 1–8, we tested the model on chromosomes 9–16. On a probe-by-probe basis, our model has a strong correlation to the data (correlation coefficient, $R = 0.44$; **Fig. 6d**). We note that, although our approach eliminated Segal *et al.*'s[11] periodic dinucleotide-based occupancy model in the feature selection stage (in favor of other DNA structural parameters with greater correlation), the dinucleotide-based model does bear some relationship to our data (**Fig. 6d**). Nonetheless, other DNA structural features correlate more highly with both the 'trough' of nucleosome occupancy at promoters and much of the nucleosome occupancy within genes (**Fig. 6**), indicating that, genome-wide, these parameters may contribute more generally to both nucleosome occupancy and positioning *in vivo*.

DISCUSSION

We have presented the first complete, high-resolution, nucleosome map of any organism. These data provide a benchmark both for gene-by-gene and global analysis of chromatin architecture and for understanding the genomic changes induced by environmental or genetic perturbation. We uncovered several genome-wide features that were not observed in previous studies, including a stronger nucleosome positioning signature relative to the TSS than to the ATG start codon. In addition, clustering the nucleosome occupancy signatures of individual genes reveals that genes with similar patterns of promoter occupancy tend to share similar functions.

More detailed analysis of our data indicates that nucleosome occupancy in promoters as compared with genes may be dictated primarily by DNA structural features. In particular, the propeller twist capacity seems to be relevant across genomic features, although we note that many transformations of G+C content also correlate highly with nucleosomal occupancy. Occupancy correlates with TFBSs in promoters, suggesting that a general selection for nucleosome-refractory DNA structural regions at a fixed position from the TSS and/or around TFBSs may explain the NDRs seen at promoters[38].

The well-established periodic (AA/TT/TA)-dominated dinucleotide nucleosome positioning pattern seems to have much less correlation with global nucleosome occupancy than other features. Because this pattern is clearly relevant *in vitro* and the signal is present across the genome[11], it is curious that the 199 sequences used to train Segal *et al.*'s model[11] have a nearly random distribution of occupancy ratios in our data and do not correspond to well-positioned nucleosomes (data not shown). These apparent discrepancies can be reconciled if nucleosome occupancy across the genome is directed more often by exclusion signals (which would include almost all of the parameters in **Fig. 6c**), whereas local 'translational and rotational' settings, in addition to strongly positioned nucleosomes, are specified by periodic signals[11,39]. In support of this possibility, the periodic AA/TT/TA signal is apparent in purified nucleosomal DNA fragments from *Caenorhabditis elegans*, although overall these dinucleotides are depleted in nucleosomal DNA relative to adjacent DNA[40].

The clustering of TFBSs at a position 100-bp upstream from the TSS, which has been clarified by mapping of TSS positions[12], is marked. Because the TFBS locations are a fixed feature, we conclude that the trends that we observe in **Figure 5** cannot be an artifact of our growth conditions. Applying the same methodology used for TFBSs, which resulted in an overall peak at ∼100 bp, we obtained a peak occurrence of TATA at about −80 bp (data not shown).

What explains the characteristic average distribution of nucleosomes within transcribed regions (for example, to the right of the TSS in **Fig. 3b**)? This phenomenon could be mechanistically explained by the known activity of RNA polymerase II in displacing and then replacing nucleosomes[41], and the gradual decay could correspond to a lack of strong precision in the nucleosome replacement pattern. However, a decaying sine-wave function of the same periodicity, which mimics almost perfectly the aggregate data from −200 to + 800 bp, in fact has little correlation with individual probe measurements. Thus, although we cannot rule out the possibility that RNA polymerase II repositions nucleosomes—an idea bolstered by the strength of the positioning pattern in the highest expressed genes—there seems to be substantial variation in the pattern from individual genes.

Work on genome-wide chromatin structure in metazoans is at its beginning[40]. A pioneering study[42] has used tiling microarrays to examine nucleosome positioning at 3,692 promoters in seven human cell lines. Despite the much greater compactness of the yeast genome as compared with the human genome, both genomes show a prominent NDR at the TSS. It seems likely that in both genomes these sites correlate with partial or complete formation of pre-initiation complexes, but whether complex assembly follows or causes NDRs remains to be determined[42,43]. Metazoan genomes differ from yeast in having gene regulatory elements that can be thousands of nucleotides distant from proximal promoter regions. It will be interesting to determine whether TFBSs play a larger or smaller part in determining nucleosome occupancy in metazoans, and how well features of chromatin structure can be used to help to predict TFBSs and other functional attributes of chromatin. We anticipate that genomic approaches similar to that used here and in other studies[7–11,39,40,42] will provide answers to these questions in the near future.

METHODS

Microarray design. The microarray was designed in collaboration with Affymetrix (PN 520055). It contains 25-nucleotide probes spaced every 8 bp covering one strand of the whole *Saccharomyces cerevisiae* genome sequence and a second set of probes offset 4 bp from the first set to cover the other strand. When combined, these probes therefore provide whole-genome coverage at 4-bp resolution for double-stranded hybridization samples. The array contains 6.5 million 5-μm oligonucleotide features and is compatible with commercially available Affymetrix scanners.

Nucleosomal DNA isolation. We adapted a previously described nucleosomal DNA preparation procedure[8]. In brief, 5-ml cultures of *Saccharomyces cerevisiae* strain BY4741 (parent strain of the Mat-a yeast knockout collection) were grown overnight and then diluted to an absorbance at 600 nm (A_{600}) of 0.2/ml into 50 ml of YPD media in a 250-ml flask. These 50-ml cultures were then grown at 30 °C to an A_{600} of 0.8–1.0/ml.

Cells were cross-linked by addition of methanol-free formaldehyde to a final concentration of 2% for 30 min while shaking at 30 °C. The reaction was quenched by addition of glycine to a final concentration of 125 mM for 5 min. Cells were collected, washed once with 20 ml of PBS, and resuspended in 6 ml of 1 M sorbitol in 50 mM Tris (pH 7.4) with freshly added 10 mM β-mercaptoethanol in a 15-ml conical tube. Zymolyase (20T) was added to a final concentration of 0.25 mg/ml and cells were spheroplasted at 30 °C while gently rolling for 30 min. After zymolyase treatment, cells were collected and resuspended in 2 ml of 1 M sorbitol, 50 mM NaCl, 10 mM Tris (pH 7.4), 5 mM $MgCl_2$, 1 mM $CaCl_2$ and 0.075% Nonidet P40, with freshly added 1 mM β-mercaptoethanol and 500 μM spermidine. Spheroplasts were divided into 6 aliquots of 300 μl and transferred into 1.5-ml Eppendorf tubes. Micrococcal nuclease (Sigma) dissolved in water at 0.1 U/μl was added to the tubes at concentrations of 0, 0.05, 0.1, 0.2, 0.3 and 0.5 U per sample. The digestion reactions were incubated at 37 °C for 45 min, and reactions were stopped with 75 μl of 5% SDS and 50 mM EDTA. We added 3 μl of proteinase K solution (20 mg/ml; Qiagen) to each tube and incubated them at 65 °C overnight.

DNA from each sample was prepared by phenol-chloroform extraction and concentrated by ethanol precipitation. Samples were then treated with RNase and analyzed in a 2% agarose gel to quantify nucleosomal content.

Microarray labeling and hybridization. Samples were prepared in pairs comprising a 0-U MNase genomic DNA control sample matched to a mononucleosomal sample. Each DNA sample was further fragmented by nuclease digestion in a solution containing 1× One-Phor-All buffer (GE Healthcare) and 1 μl of 1:16 DNase I mix (Invitrogen, Amp Grade) at 37 °C for 2 min. Fragmented samples were separated on a 2% agarose gel and stained with SYBR green (Molecular Probes) to confirm that digestion produced a distribution of fragments of less than 100 bp and an average of 50 bp. These fragments were labeled with terminal deoxynucleotidyl transferase (Amersham/GE Healthcare) and biotinylated ddATP (Perkin Elmer, NEL508) at 37 °C for 2 h. Labeled samples were hybridized at 45 °C for 72 h, washed and stained according to Affymetrix protocol EukGE-WS2v4_450 in an Affymetrix Fluidics Station 450, and then scanned in an Affymetrix 7G scanner.

Data analysis. Raw data from Affymetrix GCOS software (.CEL format) were analyzed with Affymetrix Tiling Analysis Software (TAS) v1.1 and visualized with Affymetrix Integrated Genome Browser. A tiling analysis group (.TAG file) for a two-sample analysis containing the three nucleosomal experiments as the 'treatment' and the three whole-genome samples as the 'control' was created in the Tiling Analysis Software. The data were normalized together with the built-in quantile normalization and probe-level analysis with both perfect match and mismatch probes and run with a bandwidth of 20. All log ratios used for subsequent analysis were output directly from the Tiling Analysis Software[44] probe signal intensity analysis without further manipulation. In brief, each probe was given a value calculated from the average of all probe intensities within the specified window (defined by the bandwidth parameter). These values were first averaged across replicates and then a ratio was calculated for the two-sample analysis. The resulting log ratio was then output for each probe position, where the probe position is defined as the center (13th) base coordinate for each 25-nucleotide probe.

Genome annotations were downloaded from Saccharomyces Genome Database on October 2006. Kernel density plots were generated in R.

Defining high-confidence transcription segments. We filtered the transcription segments from the complete published set[12] of $>$9,000 poly(A)-terminated transcripts and considered only segments that overlap $>$50% of a non-dubious annotated coding region on the 5′ end (to account for genes with multiple exons). The resulting 5,015 verified transcription segments comprise a comprehensive set of high-confidence transcripts with associated TSSs.

Clustering. *k*-Means clustering was computed with Cluster 3.0 using the Euclidean distance metric and 20 repetitions. Clusters were visualized with Java Treeview.

Features in genomic DNA. PWMs were downloaded from the URLs 'An improved map of conserved regulatory sites for *S. cerevisiae*' and 'Selected motifs in *S. cerevisiae*'[45]. We scored TFBSs as log odds ratios. We scored DNA structural features using data obtained from the DNA 'PROPERTY' database[46]. Free energy of bending was calculated with the free-energy equation and persistence length values in ref. 47. We calculated the propensity to form Z-DNA (*Z*-score) with the ZHUNT program[48]. We determined a threshold for the raw PWM scores for the TFBSs by taking the top 250 PWM hits in promoter regions for a given transcription factor. We assigned PWM scores equal to or above this threshold a value of 1, and all other PWM scores a value of 0. We calculated the average of each structural and base composition feature in 75-base windows with a step size of 10 across the whole yeast genome because this window size gave the highest correlation with the $\log_2$ ratios in the data (data not shown), with the exception of sequence motifs (such as tetranucleotide copy number, A/T tracts, G/C tracts and propensity to form Z-DNA), which we scored in 150-base windows.

We ran the program 'Nucleosomes Position Prediction by Genomic Sequence' on the whole yeast genome using the published yeast model[11]. This program makes a prediction for each base across the genome. The P_{occ} (average occupancy for a base pair, or probability of a base pair being occupied) was used in the comparisons to our data set, where the corresponding genomic coordinate of each P_{occ} score was taken to be the midpoint coordinate of a tiling array probe.

Statistical analysis. We calculated WMW *P* values in **Figure 5a** by considering the top 250 promoters matching the PWM for each transcription factor, and by assuming that the highest-scoring binding site in a promoter was the only binding site. We also calculated histograms in **Figure 5b,c** by using 250 hits per transcription factor. We downloaded a MATLAB implementation of Lasso[35,49]. Input features were selected as follows: tetranucleotides (copy number) and nucleosome excluding/including motifs (length, Z-DNA *Z*-score): AUC $\leq$ 0.45 and AUC $>$ 0.54; base composition and structural features: $>$0.10 correlation with $\log_2$ signal ratio; TFBSs: WMW score less than -10. We ran Lasso on the selected sequence features from chromosomes 1–8, and selected the optimal weights by means of *k*-fold cross validation ($k = 10$). Essentially similar results were obtained using the alternative regression methods Elastic Net and SVR.

Defining nucleosome occupancies. An HMM approach for automated detection of nucleosome positions from hybridization data has been described[8]. To accommodate our higher resolution data, we modified the topology of this HMM such that the model contained 78 distinct hidden states: 38 well-positioned nucleosome states (which span 31–38 probes), 39 delocalized (fuzzy) nucleosome states (spanning 39 probes), and one linker state. We trained the model on several characterized key loci in sliding windows of 100 consecutive probes using the forward-backward algorithm. These key loci were Chr 16, 833335–834655 (**Fig. 4a**); Chr 16, 29155–30195 (**Fig. 4b**); Chr 3, 21845–22925 (**Fig. 4c**); Chr 3, 38745–39785 (ref. 8; **Fig. 4d**); Chr 3, 15798–18000 (CHA1-VAC17); Chr 15, 720947–722609 (HIS3-PET56); Chr 3, 10873–14558 (HMLa); Chr 3, 292427–295317 (HMRa); Chr 3, 27898–31264 (KAR4); Chr 4, 1251000–1253000 (RVS167- SAC7); Chr 13, 873760–875248 (ADH2); Chr 2, 430782–431668 (PHO5); Chr 2, 277712–279412 (GAL10). We averaged the learned HMM parameters from all windows and applied the Viterbi algorithm to compute the most-likely states of all contiguous (separated by 4 bp) tiling array probes.

URLs. Saccharomyces Genome Database, http:// www.yeastgenome.org; ftp://ftp.yeastgenome.org/yeast; Mat-a yeast knockout collection, http://www.sequence.stanford.edu/group/yeast_deletion_project/deletions3.html; Affymetrix Tiling Analysis Software, http://www.affymetrix.com/support/developer/downloads/TilingArrayTools/index.affx; Affymetrix Integrated Genome Browser, http://www.affymetrix.com/support/developer/tools/download_igb.affx; R, http://www.r-project.org/; Cluster, http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm; Java Treeview, http://jtreeview.sourceforge.net/; 'Improved map of conserved regulatory sites for *S. cerevisiae*', http://fraenkel.mit.edu/improved_map; 'Selected motifs in *S. cerevisiae*', http://atlas.med.harvard.edu/cgi-bin/compareace_motifs.pl; DNA 'PROPERTY' database, http://srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-page+FieldInfo+-id+6F1Iv1SJDVl+-lib+PROPERTY+-bf+PropertyName; ZHUNT, http://gac-web.cgrb.oregonstate.edu/zDNA/; Nucleosomes Position Prediction by Genomic Sequence, http://genie.weizmann.ac.il/pubs/nucleosomes06/segal06_exe.html. All supplemental figures and data sets are available for downloading at our supporting information website (http://chemogenomics.stanford.edu/supplements/03nuc/). All DNA features (DNA properties and TFBSs as binary data) can be found on our website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/tillo/nucleosomes/).

Accession code. Array data has been deposited in the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/) under accession code E-MEXP-1172.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank G. Giaever for scientific advice; members of the HIP-HOP labs at Stanford and Toronto for advice and encouragement; C. Palm for help with data analysis; S.J. Altschuler (Univ. Texas Southwester Medical Center) and G.-C. Yuan (Harvard Univ.) for software for the HMM; and Z. Zhang, Q. Morris and E. Chan for discussions. W.L. is supported by a training grant from the US National Institutes of Health, and the work was supported by grants from the National Human Genome Research Institute (R.W.D. and C.N.), the National Science Foundation (R.H.M.), Genome Canada and the Ontario Genomics Institute (N.B., D.T. and T.R.H.).

AUTHOR CONTRIBUTIONS

W.L. preformed experiments, collected and analyzed the data, prepared figures and co-wrote the paper. R.W.D. provided essential input into the study design and interpretation. D.T. designed the Lasso model, analyzed all of the data, ran the HMM, and prepared figures. N.B. did data analysis on TFBSs. R.H.M. contributed to the initial study design, provided essential experimental advice, and contributed text. T.R.H. directed the detailed data analysis, and contributed figures and text. C.N. designed the study, directed the work and wrote the paper.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Noll, M. & Kornberg, R.D. Action of micrococcal nuclease on chromatin and the location of histone H1. *J. Mol. Biol.* **109**, 393–404 (1977).
2. Richmond, T.J. & Davey, C.A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
3. Jenuwein, T. & Allis, C.D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
4. Lohr, D. & Lopez, J. GAL4/GAL80-dependent nucleosome disruption/deposition on the upstream regions of the yeast GAL1–10 and GAL80 genes. *J. Biol. Chem.* **270**, 27671–27678 (1995).
5. Martinez-Campa, C. *et al.* Precise nucleosome positioning and the TATA box dictate requirements for the histone H4 tail and the bromodomain factor Bdf1. *Mol. Cell* **15**, 69–81 (2004).
6. Straka, C. & Horz, W. A functional role for nucleosomes in the repression of a yeast promoter. *EMBO J.* **10**, 361–368 (1991).
7. 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).
8. Yuan, G.C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae. Science* **309**, 626–630 (2005).
9. Lee, C.K., Shibata, Y., Rao, B., Strahl, B.D. & Lieb, J.D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat. Genet.* **36**, 900–905 (2004).
10. Ioshikhes, I.P., Albert, I., Zanton, S.J. & Pugh, B.F. Nucleosome positions predicted through comparative genomics. *Nat. Genet.* **38**, 1210–1215 (2006).
11. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
12. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. USA* **103**, 5320–5325 (2006).
13. MacIsaac, K.D. *et al.* An improved map of conserved regulatory sites for *Saccharomyces cerevisiae. BMC Bioinformatics* **7**, 113 (2006).
14. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
15. 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).
16. 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).
17. 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).
18. Holstege, F.C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
19. Carrozza, M.J. *et al.* Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123**, 581–592 (2005).
20. Keogh, M.C. *et al.* Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell* **123**, 593–605 (2005).
21. 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).
22. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
23. Satchwell, S.C., Drew, H.R. & Travers, A.A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675 (1986).
24. Wang, Y.H., Amirhaeri, S., Kang, S., Wells, R.D. & Griffith, J.D. Preferential nucleosome assembly at DNA triplet repeats from the myotonic dystrophy gene. *Science* **265**, 669–671 (1994).
25. Suter, B., Schnappauf, G. & Thoma, F. Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. *Nucleic Acids Res.* **28**, 4083–4089 (2000).
26. Wong, B., Chen, S., Kwon, J.A. & Rich, A. Characterization of Z-DNA as a nucleosome-boundary element in yeast *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA* **104**, 2229–2234 (2007).
27. Morse, R.H. Getting into chromatin: how do transcription factors get past the histones? *Biochem. Cell Biol.* **81**, 101–112 (2003).
28. Roth, F.P., Hughes, J.D., Estep, P.W. & Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**, 939–945 (1998).
29. 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).
30. Fedor, M.J., Lue, N.F. & Kornberg, R.D. Statistical positioning of nucleosomes by specific protein-binding to an upstream activating sequence in yeast. *J. Mol. Biol.* **204**, 109–127 (1988).
31. Yarragudi, A., Miyake, T., Li, R. & Morse, R.H. Comparison of ABF1 and RAP1 in chromatin opening and transactivator potentiation in the budding yeast *Saccharomyces cerevisiae. Mol. Cell. Biol.* **24**, 9152–9164 (2004).
32. Huh, W.K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
33. Liu, C.L. *et al.* Single-nucleosome mapping of histone modifications in *S. cerevisiae. PLoS Biol.* **3**, e328 (2005).
34. Kornberg, R.D. & Stryer, L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.* **16**, 6677–6690 (1988).
35. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).

36. el Hassan, M.A. & Calladine, C.R. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.* **259**, 95–103 (1996).
37. Kasten, M.M. & Stillman, D.J. Identification of the *Saccharomyces cerevisiae* genes STB1–STB5 encoding Sin3p binding proteins. *Mol. Gen. Genet.* **256**, 376–386 (1997).
38. Hogan, G.J., Lee, C.K. & Lieb, J.D. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS Genet.* **2**, e158 (2006).
39. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446**, 572–576 (2007).
40. Johnson, S.M., Tan, F.J., McCullough, H.L., Riordan, D.P. & Fire, A.Z. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res.* **16**, 1505–1516 (2006).
41. Studitsky, V.M., Kassavetis, G.A., Geiduschek, E.P. & Felsenfeld, G. Mechanism of transcription through the nucleosome by eukaryotic RNA polymerase. *Science* **278**, 1960–1963 (1997).
42. Ozsolak, F., Song, J.S., Liu, X.S. & Fisher, D.E. High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.* **25**, 244–248 (2007).
43. Zanton, S.J. & Pugh, B.F. Full and partial genome-wide assembly and disassembly of the yeast transcription machinery in response to heat shock. *Genes Dev.* **20**, 2250–2265 (2006).
44. Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342 (2004).
45. Macisaac, K.D. *et al.* A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. *Bioinformatics* **22**, 423–429 (2006).
46. Ponomarenko, J.V. *et al.* Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics* **15**, 654–668 (1999).
47. Sivolob, A., De Lucia, F., Alilat, M. & Prunell, A. Nucleosome dynamics. VI. Histone tail regulation of tetrasome chiral transition. A relaxation study of tetrasomes on DNA minicircles. *J. Mol. Biol.* **295**, 55–69 (2000).
48. Champ, P.C., Maurice, S., Vargason, J.M., Camp, T. & Ho, P.S. Distributions of Z-DNA and nuclear factor I in human chromosome 22: a model for coupled transcriptional regulation. *Nucleic Acids Res.* **32**, 6501–6510 (2004).
49. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Ann. Stat.* **32**, 407–451 (2004).