Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting

Jay R Hesselberth[1,6], Xiaoyu Chen[2,6], Zhihong Zhang[1,3,5], Peter J Sabo[1], Richard Sandstrom[1], Alex P Reynolds[1], Robert E Thurman[1], Shane Neph[1], Michael S Kuehn[1], William S Noble[1,2], Stanley Fields[1,3] & John A Stamatoyannopoulos[1,4]

The orchestrated binding of transcriptional activators and repressors to specific DNA sequences in the context of chromatin defines the regulatory program of eukaryotic genomes. We developed a digital approach to assay regulatory protein occupancy on genomic DNA *in vivo* by dense mapping of individual DNase I cleavages from intact nuclei using massively parallel DNA sequencing. Analysis of >23 million cleavages across the *Saccharomyces cerevisiae* genome revealed thousands of protected regulatory protein footprints, enabling *de novo* derivation of factor binding motifs and the identification of hundreds of new binding sites for major regulators. We observed striking correspondence between single-nucleotide resolution DNase I cleavage patterns and protein-DNA interactions determined by crystallography. The data also yielded a detailed view of larger chromatin features including positioned nucleosomes flanking factor binding regions. Digital genomic footprinting should be a powerful approach to delineate the *cis*-regulatory framework of any organism with an available genome sequence.

The binding of transcriptional regulators to specific sites on DNA is the fundamental mechanism of actuating gene expression, DNA replication, environmental response and other basic cellular processes. Delineation of the complete set of genomic sites bound *in vivo* by these proteins is therefore essential to understand genome function. The discovery more than 35 years ago that regulatory proteins protect their underlying DNA sequences from nuclease attack[1,2] has been widely exploited to define *cis*-regulatory elements in diverse organisms. Although conceptually simple, classical DNase I 'footprinting'[3], which reveals a DNA sequence protected from nuclease cleavage relative to flanking exposed nucleotides, is laborious and particularly challenging to apply systematically to the study of *in vivo* protein binding in the context of native chromatin. Current genomic approaches for localizing sites of regulatory factor–DNA interaction *in vivo* such as chromatin immunoprecipitation (ChIP) coupled to DNA microarrays[4] or to high-throughput DNA sequencing[5,6], although more readily executed on a large scale, require both prior knowledge of binding factors and factor-specific reagents, yet do not provide nucleotide-level resolution.

Regulatory factor binding to DNA in place of canonical nucleosomes results in markedly increased accessibility of the DNA template both immediately surrounding the factor binding regions and over neighboring chromatin. This accessibility is manifest as DNase I–hypersensitive sites in chromatin, which comprise a structural signature of the regulatory regions of eukaryotic genes from yeast to humans[7]. Within hypersensitive sites, cleavages occur at nucleotides that are not protected by protein binding. We therefore reasoned that these binding sites could be detected systematically, provided sufficiently dense local sampling of DNase I cleavage sites.

Here we coupled DNase I digestion of intact nuclei with massively parallel sequencing and computational analysis of cleavage patterns at single-nucleotide resolution to disclose the *in vivo* occupancy sites of DNA-binding proteins genome-wide. The resulting maps provided gene-by-gene views of transcription factor binding and related *cis*-regulatory phenomena at the resolution of individual factor binding sites. This degree of detail was sufficient to define regulatory factor binding motifs *de novo*, and to correlate factor occupancy patterns with higher-level features such as chromatin remodeling, gene expression and chromatin modifications.

RESULTS

The digital genomic footprinting strategy

To visualize regulatory protein occupancy across the *S. cerevisiae* genome, we coupled DNase I digestion of yeast nuclei with massively parallel DNA sequencing to create a dense whole-genome map of DNA template accessibility at nucleotide level. We analyzed a single well-studied environmental condition, yeast **a** cells treated with the pheromone α-factor, which synchronizes cells in the G1 phase of the cell cycle. We isolated yeast nuclei and treated them with an amount of DNase I sufficient to release short (<300 base pair (bp)) DNA fragments while maintaining the bulk of the sample as high-molecular-weight species (**Supplementary Fig. 1**

[1]Department of Genome Sciences, [2]Department of Computer Science, [3]Howard Hughes Medical Institute, [4]Department of Medicine, University of Washington, Seattle, Washington, USA. [5]Present address: Illumina, Inc., San Diego, California, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to J.A.S. (jstam@u.washington.edu).

RECEIVED 12 NOVEMBER 2008; ACCEPTED 19 FEBRUARY 2009; PUBLISHED ONLINE 22 MARCH 2009; DOI:10.1038/NMETH.1313

online). These small fragments derived from two DNase I 'hits' in close proximity, and therefore their isolation minimized contamination by single fragment ends derived from random shearing[8]. Because each end of the released DNase I 'double-hit' fragments represented an *in vivo* DNase I cleavage site, the sequence and hence genomic location of these sites could be readily determined by sequencing (**Supplementary Methods** online).

Using an Illumina Genome Analyzer I, we obtained 23.8 million high-quality 27 bp end-sequence reads that could be localized uniquely within the *S. cerevisiae* genome after filtering for duplicated sequences such as telomeric regions, transposable elements, tRNA genes, rDNA genes and other paralogous elements (**Supplementary Methods**). The DNase I cleavages mapped by these 23.8 million sequences were confined to 6.4 million unique positions within the yeast genome. We computed both the density of DNase I cleavage sites across the genome using a 50-bp sliding window and the number of times that individual nucleotide positions had been cleaved by DNase I (per-nucleotide cleavage). To control for possible DNase I cleavage bias, we digested naked DNA from the same cells in parallel to yield an equivalent fragment-size distribution. We obtained 3.27 million DNase I cleavages mapping to distinct genomic positions, from which we computed background cleavage rates for all possible dinucleotide pairs flanking the DNase I cleavage sites (**Supplementary Table 1** online). We then used these background propensities to normalize the per-nucleotide cleavage counts obtained from *in vivo* DNase I treatment (**Supplementary Fig. 2** and **Supplementary Methods** online).

Systematic identification of DNase I footprints

Data from an exemplary 100-kilobase (kb) region (**Fig. 1a**) showed that regional peaks in DNase I cleavage density concentrated in yeast intergenic regions (**Fig. 1b**), where they coincided with

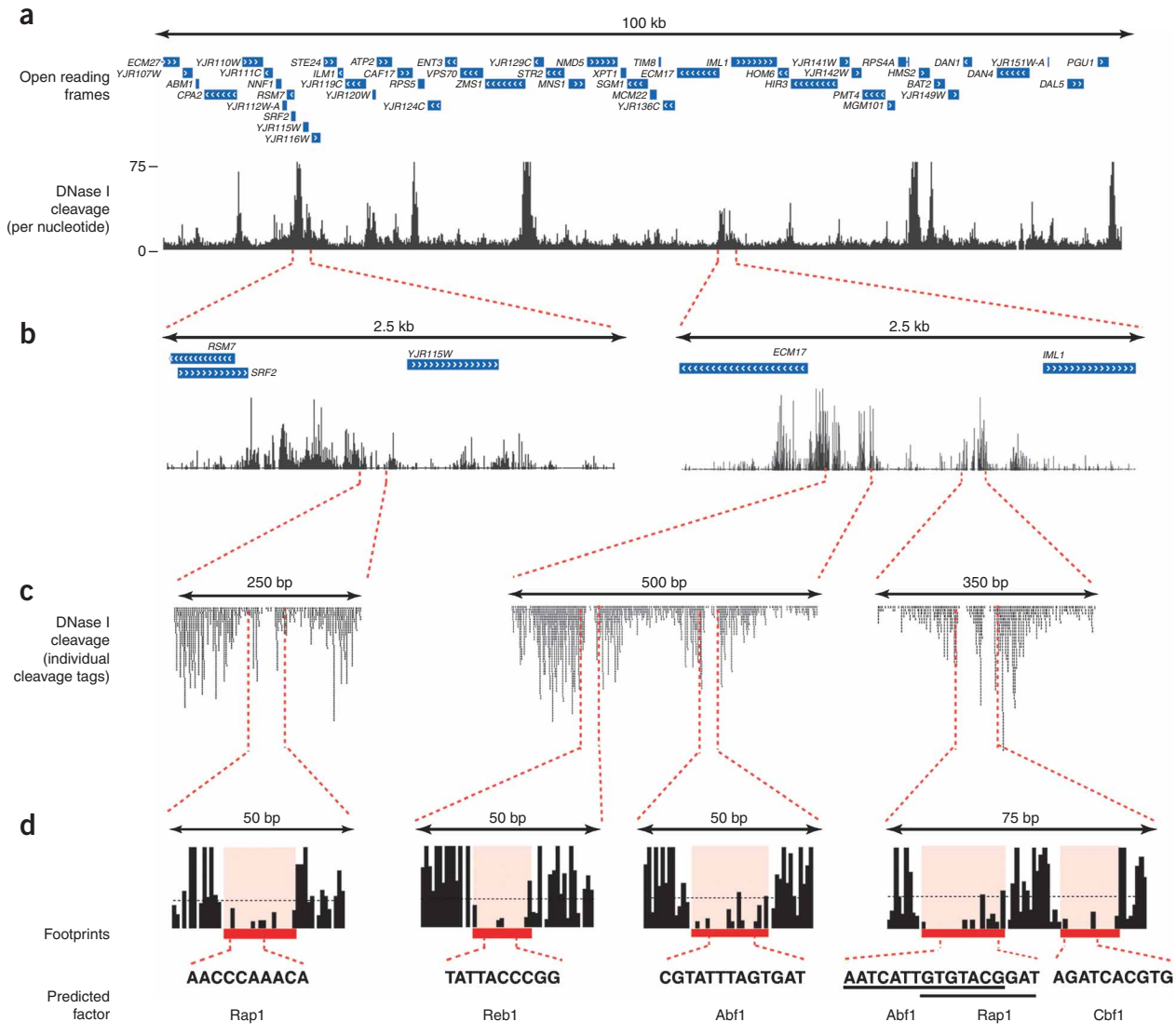

Figure 1 | Digital DNase I analysis of yeast chromatin structure from chromosomal to nucleotide resolution. (**a**) Per-nucleotide DNase I cleavage density across an exemplary 100-kilobase region of chromosome 10 (postions 625000–725000) containing ~50 open reading frames (blue boxes with arrows). (**b**) Magnification of exemplary ~2.5-kb regions containing *RSM17-YJR115W* and *ECM17-IML1* intergenic intervals. (**c**) Additional magnification showing positions of individual DNase I cleavage events (stacked vertical black tick marks), revealing DNase I footprints. (**d**) Resolution of individual DNase I footprints (red) with known motifs for yeast regulatory factors Rap1, Reb1, Abf1 and Cbf1. The dashed black line indicates the average amount of DNase I cleavage throughout the genome (~2 cleavages per base pair).

contiguous stretches of individual nucleotides that had been digested repeatedly by DNase I (**Fig. 1c**). Within the upstream regions of yeast genes, individual nucleotide positions were cleaved tens to hundreds of times.

DNase I cleavage patterns upstream of transcriptional start sites (TSSs) were punctuated by short stretches of protected nucleotides, consistent with the footprints of DNA-binding proteins, and in many cases individual footprints could be matched to known DNA-binding motifs (**Fig. 1d**). We also examined the extent to which computationally predicted factor binding sites within yeast intergenic regions exhibited DNase I protection. For any given factor, we expected computational predictions to contain a mixture of true and false positive sites. We ranked DNase I cleavage patterns surrounding 907 computationally predicted[9] Reb1 binding sites (±25 bp) within yeast intergenic regions by the ratio of DNase I cleavage flanking the motif to that within the motif (**Fig. 2a**). This analysis revealed that a substantial proportion of predicted Reb1 sites were protected from DNase I cleavage, consistent with protein binding *in vivo* and, moreover, that the DNase I protection patterns were specifically localized to the motif region. We observed analogous patterns for other motifs, with considerable variation in the fraction of computationally predicted motif instances that evidenced DNase I protection (data not shown), commensurate with the expectation that many, if not most, binding sites predicted from motif scans alone are not actuated *in vivo*.

To detect footprints systematically across the yeast genome, we developed a computational algorithm to identify short regions (8–30 bp) over which the DNase I cleavage density was substantially reduced compared with the immediately flanking regions (**Supplementary Methods**). To assess significance and compute a false discovery rate (FDR) for footprint predictions, we compared predictions with a randomly shuffled local background distribution (**Supplementary Methods**). Using this approach, we identified 4,384 footprints within the intergenic regions of the yeast genome at a false discovery rate (FDR) threshold of 5% or 0.05 (**Supplementary Table 2** online). We identified at least one footprint in the proximal promoter region of 1,778 genes, and 630 genes contained two or more footprints. At an FDR threshold of 10%, we identified 6,056 footprints distributed across 2,929 gene promoters, with 1,048 of them evincing more than two footprints.

Identification of sequence motifs in DNase I footprints

We categorized the 4,384 footprints identified at an FDR threshold of 5% by deriving sequence motifs *de novo* using the MEME

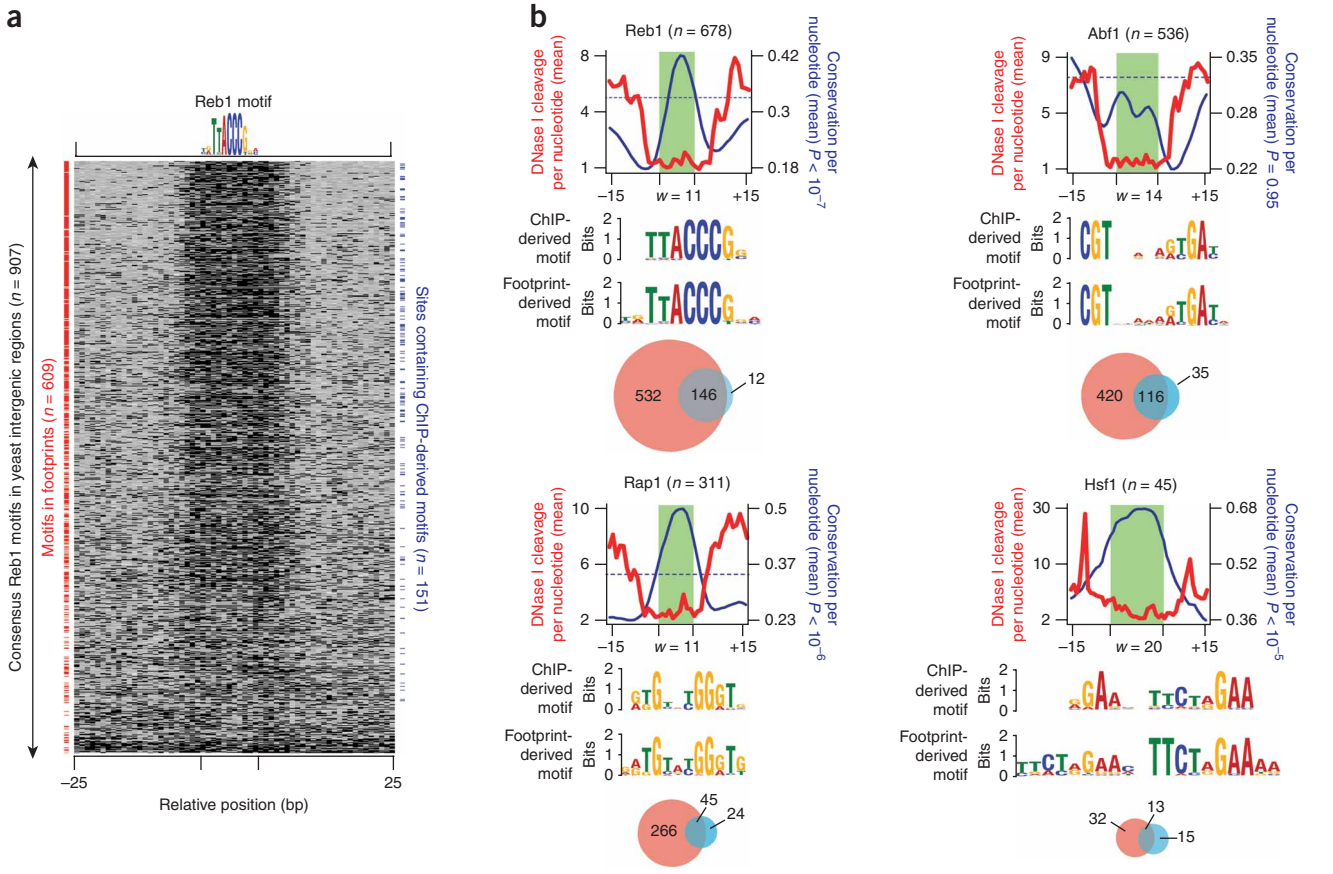

Figure 2 | Detection of footprints and corresponding sequence motifs. (**a**) Visualization of DNase I protection (footprinting) around 907 computationally predicted Reb1 sites in a heat map. Rows show extent of DNase I cleavage 25 bp upstream and downstream of each motif instance and are sorted by the ratio of mean cleavage over flanking regions to that within the motif itself. Red ticks indicate motif instances ($n = 609$) that coincide with footprints (FDR = 0.05) containing *de novo*–derived Reb1 motifs. Blue ticks indicate motif instances coinciding with those identified by ChIP[10]. (**b**) Mean per-nucleotide DNase I cleavage and evolutionary conservation (significance of observed conservation pattern is also indicated) calculated for footprints that match the Reb1, Abf1, Rap1 and Hsf1 motifs. Green (w, motif width) delineates the consensus motifs derived from the footprinted region; motifs derived from ChIP and footprinting are shown below the graph. Venn diagrams depict the overlap of motifs derived from mapping footprints (FDR = 0.05; red) versus ChIP (blue).

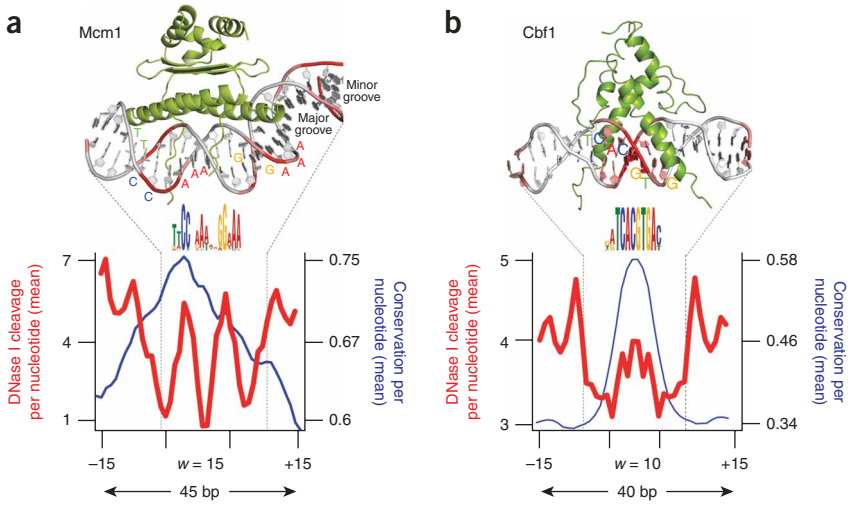

Figure 3 | Mean nucleotide-level accessibility parallels protein-DNA interactions. (**a**,**b**) Structures (green) of Mcm1 (ref. 16; adjacent Matα2 was removed for clarity; **a**) and of the human homolog of CBF1 (ref. 17; **b**) bound to a single DNA recognition site. Labeled DNA bases correspond to positions within the footprint-derived motifs (middle; w, motif width), and the extent of observed DNase I cleavage in 88 Mcm1 sites and in 243 Cbf1 sites is reflected by red shading of the DNA backbone and is plotted in the graph below. Mean nucleotide-level conservation is also shown in the graph (Mcm1, $P < 10^{-5}$; Cbf1, $P < 10^{-3}$).

software package[9] and comparing the results with previously described factor-binding motifs. The predicted numbers of *in vivo* binding sites across the yeast genome for different regulators vary by nearly two orders of magnitude[10]. However, MEME readily recovered high-quality motifs corresponding to many important yeast regulators including Reb1, Abf1, Hsf1, Rap1, Mcm1 and Cbf1 (**Supplementary Table 3** and **Supplementary Methods** online).

Beyond the factor binding sequences recovered *de novo* using relatively stringent thresholds, footprints were markedly enriched (versus yeast intergenic regions) for a broad range of regulators (**Supplementary Table 4** online), indicating that the footprinted space was densely populated with previously recognized protein binding sites. Collectively, 35.2% of the footprints with an FDR of 0.05 overlapped a conserved factor binding site inferred from ChIP data[10]. To assess the effect of stringently thresholded footprint detection, we computed factor motif-specific receiver-operator characteristic curves for various regulators (**Supplementary Fig. 3** online). All curves were well above the diagonal, indicating strong enrichment of previously recognized factor binding sites near the $P < 0.05$ threshold. This observation implies that many additional real sites exist in the data and simply do not meet the selected detection threshold.

As footprints identified at an FDR threshold of 5% were well-distinguished from local background, we speculated that these might be enriched for factors with strong binding specificities, whereas many more weakly binding factors might not have yet achieved requisite coverage depth for detection using our algorithm. We therefore predicted that protection of the underlying DNA sequence from nuclease attack should be roughly inversely proportional to the binding affinity of the overlying regulatory factors. To test this, we compared the information content (a measure of the size and complexity of the predicted binding site[10]) of 117 known factor motifs with the amount of DNase I protection within all predicted matches of each motif genome-wide, and found them to be significantly anticorrelated ($P < 10^{-6}$; **Supplementary Fig. 4** online). This result suggested that high information content of a binding site was a good predictor of the affinity of a factor for its cognate DNA sequences and consequently its propensity to generate footprints detectable at the FDR = 0.05 cutoff given the current depth of sequence sampling. The result also indicated that weaker motifs should be progressively recovered with greater DNase I cleavage sampling whereupon their cognate footprints may become reliably distinguished from the background.

To visualize consensus nucleotide-level DNase I protection patterns for motifs corresponding to the most abundant footprints, we computed aggregate per-nucleotide DNase I cleavage and evolutionary conservation (PhastCons[11]) across all instances of each motif (**Fig. 2b**). This showed that several footprint-derived consensus sequences were more information-rich than prior predictions based on inference from ChIP and conservation data alone[10,12] (**Fig. 2b**). For example, the previously characterized motif weight matrix for Reb1 spans 8 nucleotides[10], whereas the footprint-derived consensus fine-tuned the motif core and extended it an additional 3 nucleotides (**Fig. 2b**). In some cases, such as for Hsf1, the *de novo* footprint-derived motif was more complex than previous predictions (**Fig. 2b**).

Nucleotide-level DNase I protection patterns closely paralleled evolutionary conservation for virtually all factors, attesting to the biological relevance of the footprints and their derived cognate motifs (**Fig. 2b**). The conserved region was typically larger than the previously derived consensus sequence but closely matched the footprint-derived consensus. To assess the significance of the aggregate conservation patterns for each motif, we used a permutation approach to compare the observed patterns to random samples from yeast intergenic regions (**Fig. 2b** and **Supplementary Methods**). These calculations confirmed the significance of the patterns seen for factors such as Reb1, Rap1, Mcm1 and others (**Figs. 2b**, **3** and **Supplementary Fig. 5** online), paralleling previous results from analyses of factor binding sites across yeast species[13,14]. Although the majority of individual footprints genome-wide were well-conserved, many lacked obvious conservation, consistent with the known potential for some sites to undergo rapid evolutionary turnover[15].

In comparison with binding site catalogs based on ChIP and conservation data[10], digital footprinting revealed 678 Reb1 sites versus the 158 Reb1 sites previously predicted, 536 versus 151 sites for Abf1, and 311 versus 42 sites for Rap1 (ref. 10; **Fig. 2b**). These discrepancies are partly a reflection of the FDR and significance thresholds applied both to earlier and the present data, though they suggest an important contribution of condition-specific binding.

DNA 'structural motifs' parallel protein-DNA interactions

A striking feature of the DNase I cleavage and protection profiles for many factors is the presence of complex patterns within and surrounding the derived consensus motif sequence. For example,

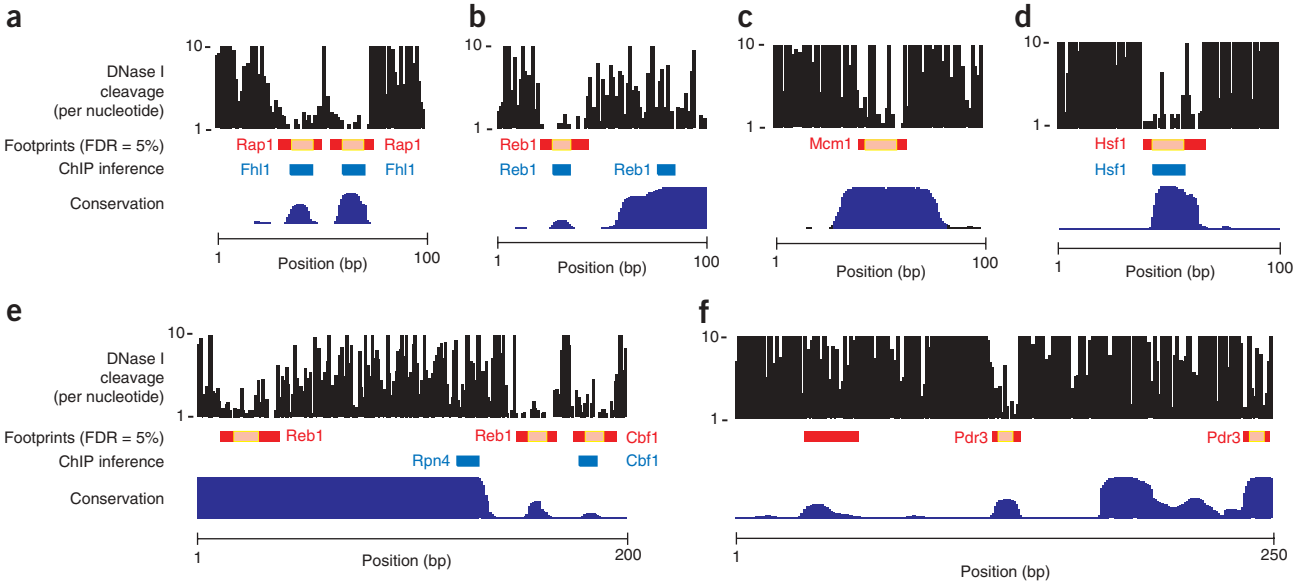

Figure 4 | Individual yeast regulatory regions and factor binding sites. (**a**–**f**) Each panel shows per nucleotide DNase I cleavage, detected footprints (red boxes), assigned motifs (pink boxes), binding sites inferred from ChIP experiments (blue boxes), and evolutionary conservation (dark blue, bottom).

Mcm1 sites had a characteristic multiphasic cleavage pattern, with three short protected regions alternating with two accessible regions (**Fig. 3a**). Analogously, Cbf1 sites comprise a large protected region with a central zone of accessibility (**Fig. 3b**). We surmised that these and other stereotypical 'structural motifs' reflected patterns of interaction of each factor with the DNA helix. To examine this in detail, we aligned the nucleotide-level DNase I accessibility motifs, the corresponding sequence motifs and crystal structures of Mcm1 (ref. 16) and a Cbf1 homolog[17] (**Fig. 3a**,**b**), This revealed striking correspondence between mean nucleotide-level DNase I accessibility and the pattern of protein-DNA contacts. Mcm1 is a MADS box factor that binds DNA through long α helices that make many contacts along the major groove[16]. Mcm1 binding introduces bends into the DNA helix, which distort the opposing minor grooves, rendering them more susceptible to nuclease attack[18]. These effects were evident in the nucleotide-level DNase I cleavage patterns, which revealed many nuclease attack sites opposite the Mcm1 α helices (**Fig. 3a**). Similarly, in the case of the helix-loop-helix protein Cbf1, alignment of the DNase I cleavage profile to the crystal structure of the human homolog (which shares the same DNA-binding residues) revealed protection of nucleotides by the opposite α helices, separated by a central region of increased accessibility (**Fig. 3b**). Taken together, these data suggest that the mean nucleotide-level DNA accessibility patterns derived from digital genomic footprinting of specific factors represent structural motifs that parallel protein-DNA interactions *in vivo*.

Footprints in individual regulatory regions

Digital genomic footprinting data were sufficiently dense to allow analysis of regulatory factor occupancy patterns for individual regulatory regions. The examples in **Figure 4** and **Supplementary Figure 6** online provide snapshots of a diverse population of regulators and binding-site contexts. In many cases, high-confidence footprints agreed with previous predictions for specific regulators (**Fig. 4a**,**b**,**d**,**e** and **Supplementary Fig. 6a**). However, we also observed discordance with previous predictions (**Fig. 4b**,**e**), possibly reflecting condition-specific binding. For example, at the *REB1* promoter (**Fig. 4e**), we detected footprints at two previously identified evolutionarily conserved Reb1 binding sites[19], neither of which had been identified under conditions used in prior ChIP experiments. Conversely, ChIP data annotated a nearby Rpn4 site that does not fall within a footprint with an FDR of 0.05.

The data also illustrate variability in the extent to which a given regulator protects different cognate recognition sites (**Fig. 4a**,**e**,**f**). In some cases, the identification of footprints matching characterized regulators could be used to revise gene annotations. For example, we identified a Rap1 site upstream of *RPS30B* that is situated within the hypothetical open reading frame for *FYV12*. However, the marked DNase I sensitivity and general lack of evolutionary conservation within this region suggest that *FYV12* is not a gene but rather the promoter of the neighboring *RPS30B* (**Supplementary Fig. 7** online).

High-resolution mapping of chromatin architecture

We next visualized patterns of DNase I cleavage and protection in extended promoter domains. We extracted DNase I cleavage data from −1 kb to +1 kb intervals around the TSSs of ~5,000 yeast genes and performed hierarchical clustering (**Fig. 5a**). This revealed that 93% of yeast genes could be organized into four distinct clusters, ranging from low to high mean chromatin accessibility (**Fig. 5a**). For genes in the lowest-accessibility cluster, chromatin accessibility was maximal over the −100 bp region. Even at this resolution, an ~10 bp footprint centrally positioned within the −100 bp region could be discerned at a surprising proportion of genes (**Fig. 5a**). A prominent feature of the larger-scale DNase I cleavage patterns is the presence of regular undulations in accessibility, with a period of ~175 bp symmetrically flanking the central high-accessibility zone (**Fig. 5a**,**b**). This pattern is consistent with the presence of phased nucleosomes. We also observed that the periodic pattern emanated from the boundaries of the central high mean chromatin accessibility region, even though this region varied in size between the four clusters. This observation suggested

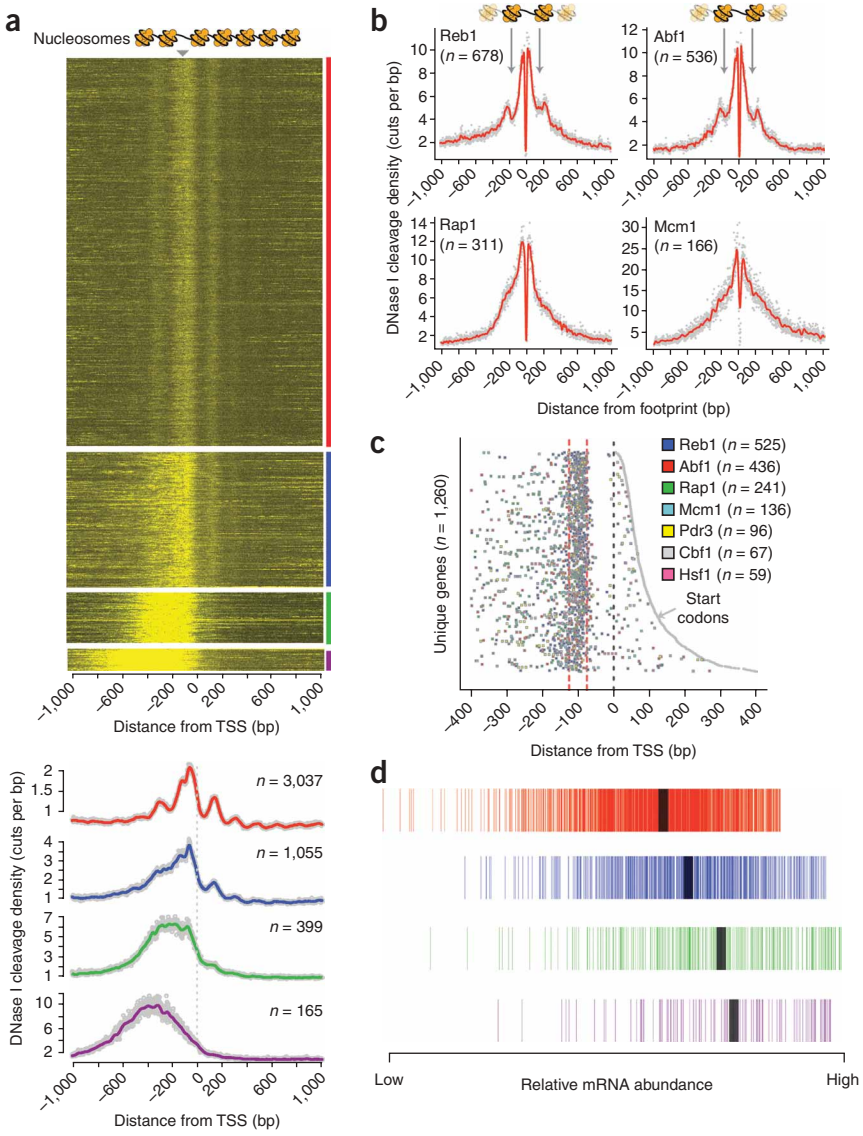

Figure 5 | Higher-order patterns of DNA accessibility. (**a**) Mapped DNase I cleavages relative to 5,006 TSSs[29]. Four major clusters were exposed by *k*-means analysis. In the cluster marked by a red bar, maximal DNase I cleavage occured in a stereotypic ~50 bp band ~100 bp upstream of the TSS (arrowhead, top). In clusters marked by blue, green and purple bars, the extent and intensity of DNase I cleavage upstream of the TSS extended to the −1, −2 and −3 nucleosomes, respectively. (**b**) DNase I cleavage profiles aligned relative to Reb1, Abf1, Rap1 and Mcm1 footprints. (**c**) Distribution of footprints matching Reb1, Abf1, Rap1, Mcm1, Pdr3, Cbf1 and Hsf1 relative to the TSSs (dashed black line) and start codons of 1,260 genes sorted by the length of the 5′ UTR. Enrichment within an ~50 bp region centered ~100 bp upstream of the TSS (dashed red lines). (**d**) mRNA abundance for genes found in each of the four clusters correlate with the accessibility of the promoters of those genes (colors as in **a**; median expression denoted by black bars).

that phased nucleosomes were in fact distributed relative to central sites occupied by factors.

To explore the relationship between nucleosomal features and factor occupancy, we examined the long-range distribution of DNase I cleavages surrounding footprints of individual regulators across the genome. The distribution of DNase I cleavages relative to footprints for Reb1 and Abf1 revealed periodic undulations (**Fig. 5b**), consistent with phased nucleosome arrays symmetrically distributed relative to the factor-binding sites. However, Rap1 and Mcm1 exhibited less prominent patterns (**Fig. 5b**), suggesting that some factors (for example, Reb1 and Abf1) have a more determinative role in establishing chromatin architecture at promoters[20]. Collectively, these data are consistent with statistical positioning of nucleosomes relative to factor binding–induced 'barrier' events[21,22].

We also observed that the binding of many factors appears to be positionally constrained relative to transcriptional start sites. For six factors (Reb1, Abf1, Rap1, Mcm1, Cbf1 and Pdr3), footprints exhibited tight clustering into an ~50 bp zone centered ~100 bp upstream of the TSS (**Fig. 5c**). Furthermore, the region immediately 3′ to the –100 bp region is generally depleted of footprints (**Fig. 5c**), consistent with the presence of a positioned nucleosome. These results are compatible with the existence of a common focal point for the organization of promoter architecture of a substantial fraction of yeast genes[22,23].

High-resolution chromatin architecture and gene expression

We next asked whether the four chromatin structural clusters (**Fig. 5a**) were correlated with expression of their constituent genes. We observed that the average gene expression from each cluster increased monotonically with the extent of chromatin disruption upstream of the TSS (**Fig. 5d**). This organization is most readily explained by the size of the domain over which factor binding takes place. For the genes in the lowest mean accessibility cluster, factor binding was largely restricted to the –100 bp region, with a prominent −1 nucleosome situated around −200 bp. By contrast, for genes in the second cluster the accessible factor-binding region extended from the TSS to approximately −360 bp, with a 5′ shift in the position of the –1 nucleosome. For genes in the two clusters with the highest chromatin accessibility (**Fig. 5a**), the factor-binding region extended to −450 bp and −750 bp, respectively. Taken together, these observations suggest that, rather than simple gain or loss of an upstream nucleosome[23–26], high expression of yeast genes may involve increases in both the number and longitudinal extent of regulatory factors bound in the upstream region. Conversely, many genes expressed at a low level nonetheless had high chromatin accessibility across their promoter regions, with attendant footprints indicative of factor binding. The existence of such promoters parallels reports of binding by well-described regulators such as Hsf1, Gal4, Abf1 and Pdr1/Pdr3 under conditions in which they do not activate transcription[27]. These results emphasize the heterogeneous nature of factor binding and consequent control of gene expression, requiring gene-level analyses of factor occupancy.

DISCUSSION

Application of DNase I footprinting to the study of *in vivo* interactions has proven difficult, and only a handful of studies have been reported for highly targeted loci such as individual *cis*-regulatory elements[28]. The digital genomic footprinting approach we describe now enables genome-scale detection of the *in vivo* occupancy of genomic sites by DNA-binding proteins. Although detection of individual binding events is dependent on the depth of sequence coverage at a given position, the approach takes advantage of the concentration of cleavages within DNase I–hypersensitive regions. In the case of mammalian genomes, DNase I cleavage is highly targeted to DNase I–hypersensitive sites, which comprise only 1–2% of the genome in each cell type. As such, although the human genome is ~250-fold larger than the yeast genome, it is potentially addressable using this method with only modest scale-up.

To date, genome-scale localization of regulatory factor binding sites has largely relied on a top-down approach centered on ChIP. Whereas ChIP requires that each DNA-binding protein first be interrogated by genome-wide location analysis and can be carried out for only one protein at a time, DNase I footprinting addresses all factors simultaneously in their native state and detects regions of direct binding with nucleotide precision. However, many regulatory factors share common binding sequences, and ChIP offers definitive identification of the protein of interest. The joint application of digital genomic footprinting with ChIP should therefore provide particularly rich information concerning the fine-scale architecture of *cis*-regulatory circuitry.

Digital genomic footprinting is also a powerful tool for annotation of the genomes of diverse organisms, about which little is known beyond the genome sequence itself. In these contexts, top-down approaches to regulatory factor binding site localization are limited. By contrast, digital genomic footprinting can be applied to develop rapidly both a gene-by-gene map and a lexicon of major regulatory motifs.

Cis-regulatory alterations accompanying different growth conditions or cell differentiation and cycling impact multiple regulators simultaneously and are difficult to study. Our method is readily extensible to the analysis of such changes across the genome by sampling sequential time points to visualize *cis*-regulatory dynamics. Digital genomic footprinting therefore has the potential to expose and probe the *cis*-regulatory regulatory framework of virtually any sequenced organism in a single experiment, regardless of prior functional characterization.

METHODS

Detection of footprints within digital DNase I data. Footprints were identified using a computational algorithm. Descriptions of digital DNase I library production, sequencing, computational identification of footprints and other analyses are available in **Supplementary Methods**. Software used for this analysis is available as **Supplementary Software** online.

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

ACKNOWLEDGMENTS

We thank the staff of the University of Washington Genome Sciences High-Throughput Genomics Unit for technical assistance with Illumina-Solexa sequencing, and members of the Stamatoyannopoulos and Fields laboratories for many helpful discussions. This work was supported by US National Institutes of Health grants R01GM071923 and U54HG004592 to J.A.S., and P41RR11823 to S.F. and W.S.N.; X.C. was supported by a fellowship from the Natural Sciences and Engineering Research Council of Canada (NSERC PGS D3).

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

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

1. Maniatis, T. & Ptashne, M. Structure of the lambda operators. *Nature* **246**, 133–136 (1973).
2. Gilbert, W. in *Polymerization in Biological Systems* 245–259 (Elsevier, North-Holland, Amsterdam, 1972).
3. Galas, D.J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).
4. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
5. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
6. Wei, C.L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**, 207–219 (2006).
7. Gross, D.S. & Garrard, W.T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).
8. Sabo, P.J. *et al.* Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. *Nat. Methods* **3**, 511–518 (2006).
9. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
10. MacIsaac, K.D. *et al.* An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7**, 113 (2006).
11. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
12. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
13. Cliften, P. *et al.* Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301**, 71–76 (2003).
14. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
15. Borneman, A.R. *et al.* Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815–819 (2007).
16. Tan, S. & Richmond, T.J. Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. *Nature* **391**, 660–666 (1998).
17. Ferre-D'Amare, A.R., Pognonec, P., Roeder, R.G. & Burley, S.K. Structure and function of the b/HLH/Z domain of USF. *EMBO J.* **13**, 180–189 (1994).
18. Acton, T.B., Zhong, H. & Vershon, A.K. DNA-binding specificity of Mcm1: operator mutations that alter DNA-bending and transcriptional activities by a MADS box protein. *Mol. Cell. Biol.* **17**, 1881–1889 (1997).
19. Wang, K.L. & Warner, J.R. Positive and negative autoregulation of REB1 transcription in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **18**, 4368–4376 (1998).
20. Planta, R.J., Goncalves, P.M. & Mager, W.H. Global regulators of ribosome biosynthesis in yeast. *Biochem. Cell Biol.* **73**, 825–834 (1995).
21. Boeger, H., Griesenbeck, J. & Kornberg, R.D. Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. *Cell* **133**, 716–726 (2008).
22. Mavrich, T.N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18**, 1073–1083 (2008).
23. Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* **39**, 1235–1244 (2007).
24. Shivaswamy, S. *et al.* Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* **6**, e65 (2008).
25. Yuan, G.C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309**, 626–630 (2005).
26. 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).
27. Jakobsen, B.K. & Pelham, H.R. Constitutive binding of yeast heat shock factor to DNA *in vivo*. *Mol. Cell. Biol.* **8**, 5040–5042 (1988).
28. Strauss, E.C. & Orkin, S.H. *In vivo* protein-DNA interactions at hypersensitive site 3 of the human beta-globin locus control region. *Proc. Natl. Acad. Sci. USA* **89**, 5809–5813 (1992).
29. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. USA* **103**, 5320–5325 (2006).