HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2009 October 01.

Published in final edited form as:

Nat Methods. 2009 April ; 6(4): 283–289. doi:10.1038/nmeth.1313.

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], and **John A. Stamatoyannopoulos**[1,4,#]

[1] Dept. of Genome Sciences, University of Washington, Seattle, WA 98195

[2] Dept. of Computer Science, University of Washington, Seattle, WA 98195

[3] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195

[4] Dept. of Medicine, University of Washington, Seattle, WA 98195

Abstract

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 as well as the identification of hundreds of novel binding sites for major regulators. We observed striking correspondence between nucleotide-level 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 provides a powerful approach to delineate the *cis*-regulatory framework of any organism with an available genome sequence.

Background

The binding of transcriptional regulators to specific sites on DNA provides the fundamental mechanism for actuating genomic programs of 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 for an understanding

[#] Correspondence: jstam@u.washington.edu.
[5]Present address: Illumina, Inc., San Diego, CA
[6]equal contribution

Additional methods: Additional methodological details are provided in Supplementary Methods.

The authors declare no competing financial interests.

EdSumm: Dense mapping of DNase I cleavage sites across the whole yeast genome by next generation sequencing reveals a global view of the binding of regulatory proteins to genomic DNA. The high resolution allows the identification of new binding sites for known factors as well as the *de novo* derivation of factor binding motifs.

of genome function. The discovery more than 35 years ago that regulatory proteins protect their underlying DNA sequences from nuclease attack1,2 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 in practice 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 coupled to DNA microarrays4 or to high-throughput DNA sequencing5,6, while 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 humans7. Within hypersensitive sites, cleavages accumulate 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. We present an analysis of binding sites on yeast DNA based on over 23 million DNA sequence reads mapped back to the genome.

Here we couple DNase I digestion of intact nuclei with massively parallel sequencing and computational analysis of nucleotide-level patterns to disclose the *in vivo* occupancy sites of DNA-binding proteins genome-wide. The resulting maps provide gene-by-gene views of transcription factor binding and related *cis*-regulatory phenomena at the resolution of individual factor binding sites. This level of detail is 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 genome of *Saccharomyces cerevisiae*, 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 a DNase I concentration sufficient to release short (<300 bp) DNA fragments while maintaining the bulk of the sample in high molecular weight species (Supplementary Fig.1). These small fragments derive from two DNase I "hits" in close proximity, and therefore their isolation minimizes contamination by single fragment ends derived from random shearing8. Because each end of the released DNase I 'double-hit' fragments represents an *in vivo* DNase I cleavage site, the sequence and hence genomic location of these sites can be readily determined by sequencing (Supplementary methods).

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 following 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, as well as 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 at particular nucleotide combinations, we carried out a parallel experiment with naked DNA from the same cells digested 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 (*i.e.*, tetranucleotides straddling) the DNase I cleavage sites (Supplementary Table 1). We then used these background propensities to normalize the per nucleotide cleavage counts obtained from *in vivo* DNase I treatment (Supplementary Fig. 2, Supplementary methods).

Systematic identification of DNase I footprints

Data from an exemplary 100 kb region (Fig.1a) showed that regional peaks in DNase I cleavage density concentrate in yeast intergenic regions (Fig.1b), where they coincide with contiguous stretches of individual nucleotides that had been struck repeatedly by DNase I (Fig.1c). Within the upstream regions of yeast genes, individual nucleotide positions were cleaved tens to hundreds of times.

On close inspection, we observed that 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 that in many cases individual footprints could be matched to known DNA-binding motifs (Fig.1d). We also examined the degree to which computationally predicted factor binding sites within yeast intergenic regions exhibited DNase I protection. For any given factor, computational predictions are expected to contain a mixture of true- and false-positive sites. Fig.2a shows the DNase I cleavage patterns surrounding 907 computationally-predicted9 Reb1 binding sites (+/-25 bp) within yeast intergenic regions, ranked by the ratio of DNase I cleavage flanking the motif to that within the motif. This analysis showed that a significant proportion of predicted Reb1 sites exhibited DNase I protection 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 (between 8 and 30 bp) over which the DNase I cleavage density was significantly reduced compared with the immediately flanking regions (Supplementary methods). To assess statistical 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 of 5% (FDR=0.05; Supplementary Table 2). At least one FDR=0.05 footprint was identified in the proximal promoter region of 1,778 genes, with 630 genes harboring two or more footprints. At a false discovery rate of 10%, we identified 6,056 footprints distributed across 2,929 gene promoters, with 1,048 of them evincing >2 footprints.

Identification of sequence motifs in DNase I footprints

We categorized the 4,384 FDR=0.05 footprints by deriving sequence motifs *de novo* using MEME9, 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 magnitude10. 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).

Beyond the factor binding sequences recovered *de novo* using relatively stringent thresholds, footprints were significantly enriched (vs. yeast intergenic regions generally) for a broad range of regulators (Supplementary Table 4), indicating that the footprinted space was densely populated with previously recognized protein binding sites. Collectively, 35.2% of the FDR=0.05 footprints overlapped an occurrence of a conserved factor binding site inferred from ChIP data10. To assess the effect of stringently thresholded footprint detection, we computed factor motif-specific receiver-operator characteristic (ROC) curves for a variety of regulators (Supplementary Fig.3). 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.

Since footprints identified at the FDR=0.05 level are well-distinguished from their local background, we speculated that these might be generally enriched in factors with strong binding specificities, while many more weakly binding factors might not have yet achieved requisite coverage depth for detection using our algorithm. In both cases, we 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 site10) of 117 known factor motifs with the level of DNase I protection within all predicted matches of each motif genome-wide, and found them to be significantly anticorrelated ($P<10^{-16}$; Supplementary Fig.4). 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 level given the current depth of sequence sampling. The result also indicates that weaker motifs should be progressively recovered at deeper levels of 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 (PhastCons11) 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 (Fig.2b)10,12. For example, the previously-characterized motif weight matrix for Reb1 spans 8 nucleotides10, whereas the footprint-derived consensus fine-tunes the motif core and extends it an additional 3 nucleotides (Fig.2b). In some cases, such as Hsf1, the *de novo* footprint-derived motif is substantially more complex than previous predictions (Fig. 2b).

We observed that nucleotide-level DNase I protection closely parallels evolutionary conservation for virtually all factors, further attesting to the significance of the footprints and their derived cognate motifs (Fig.2b). The width of the conserved region is typically broader than the span of previously-derived consensus sequence, but matches closely 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 (Supplementary Fig.5), paralleling previous results from analyses of factor binding sites across yeast species13,14. Although the majority of individual footprints genome-wide were well-conserved, many lacked significant conservation, consistent with the known potential for some sites to undergo rapid evolutionary turnover15.

In comparison with binding site catalogues based on ChIP and conservation data10, digital footprinting revealed 678 Reb1 sites vs. the 158 previously predicted; 536 vs. 151 Abf1 sites; and 311 Rap1 footprints vs. 42 previously predicted10 (Fig.2b). These discrepancies are partly a reflection of the statistical 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, Mcm1 sites display a characteristic multi-phasic cleavage pattern, with three short protected regions alternating with two accessible regions (Fig.3a). Analogously, Cbf1 sites evince a broad 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 co-crystal structures of Mcm116 and a Cbf1 homolog17 (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 numerous contacts along the major groove16. Mcm1 binding introduces significant bends into the DNA helix, which distort the opposing minor grooves, rendering them more susceptible to nuclease attack18. These effects are evident in the nucleotide-level

cleavage patterns which show a concentration of nuclease attack opposite the Mcm1 alpha 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 homologue (which shares the same DNA-binding residues) reveals protection of nucleotides by the opposite alpha 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 are sufficiently dense to enable analysis of regulatory factor occupancy patterns at the level of individual regulatory regions. The examples in Fig. 4 and Supplementary Fig.6 provide snapshots of a diverse population of regulators and binding site contexts. In many cases, high-confidence footprints agree with previous predictions for specific regulators (Fig.4a,b,d,e and Supplementary Fig.6a). However, we also observed numerous examples of discordance (Fig.4c,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 sites19, neither of which were identified under conditions used in prior ChIP experiments. Conversely, ChIP data annotated a nearby Rpn4 site that does not fall within an FDR=0.05 footprint.

The data also illustrate considerable variability in the degree to which a given regulator protects different cognate recognition sites (compare pairs of Rap1, Reb1, and Pdr3 sites in Fig.4a,e, and f, respectively). 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).

High-resolution mapping of chromatin architecture

We next sought to visualize patterns of DNase I cleavage and protection at the level of 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 (red cluster) to high (purple) mean chromatin accessibility (Fig.5a). For genes in the red cluster, chromatin accessibility was maximal over the -100 region, visualized in Fig.5a as a prominent central vertical yellow stripe. Even at this resolution, a ∼10 bp footprint centrally positioned within the -100 region could be discerned at a surprising proportion of genes (Fig.5a). A prominent feature of the 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). This pattern is consistent with the presence of phased nucleosomes. We further observed that the periodic pattern emanated from the boundaries of the central high-accessibility region, even though this region varied in size between the four clusters. This observation suggested that phased

nucleosomes were in fact distributed relative to central sites occupied by factors. To explore further the relationship between nucleosome-level 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, consistent with phased nucleosome arrays symmetrically distributed relative to the factor-binding sites. However, Rap1 and Mcm1 exhibited less prominent patterns (Fig.5c), suggesting that some factors (*e.g.* Reb1 and Abf1) have a more determinative role in establishing chromatin architecture at promoters20. Collectively, these data are consistent with statistical positioning of nucleosomes relative to factor binding-induced 'barrier' events21,22.

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 exhibit tight clustering into a ∼50 bp zone centered ∼100 bp upstream of the TSS (Fig.5b). Furthermore, the region immediately 3′ to the -100 region is generally depleted of footprints (Fig.5b), 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 genes22,23.

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 found that the average expression level of genes from each cluster increases 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 red cluster, factor binding is largely restricted to the -100 region, with a prominent -1 nucleosome around -200. By contrast, for genes in the blue cluster, the accessible factor-binding region extends from the TSS to approximately -360, with a 5′ shift in the -1 nucleosome. For genes in the green and purple clusters, the factor-binding region extends to -450 bp and -750, respectively. Taken together, these observations suggest that, rather than simple gain or loss of an upstream nucleosome23-26, 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 exhibited 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 Heat Shock Factor (Hsf1), Gal4, Abf1 and Pdr1/Pdr3 under conditions in which they do not activate transcription27. These results emphasize the heterogeneous nature of factor binding and consequent control of gene expression, requiring gene-level analyses of factor occupancy.

Discussion

DNase I footprinting has long been used in an *in vitro* context to interrogate protein-DNA interactions. However, application of this approach 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 elements28. By coupling DNase I digestion of intact nuclei with massively parallel sequencing and computational analysis of nucleotide-level patterns, 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, the collective span of human DNase I hypersensitive sites is only 1-2% of the genome, and therefore potentially addressable 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 chromatin immunoprecipitation. Several limitations of this approach are addressed by digital genomic footprinting. Whereas ChIP requires prior knowledge of each DNA-binding protein to be interrogated by genome-wide location analysis, and can be carried out on only one protein at a time, DNase I footprinting addresses all factors simultaneously in their native state, and detects regions of direct binding at nucleotide precision vs. inference based on motif enrichment analysis. 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 also provides 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. The approach described herein 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 its prior level of functional characterization.

Methods

Detection of footprints within digital DNase I data

Footprints were identified using a computational algorithm that evaluates short regions (between 8 and 30 bp) over which the DNase I cleavage density was significantly reduced compared with the immediately flanking regions (Supplementary Methods). FDR thresholds were assigned by comparing *p*-values obtained from real and shuffled cleavage data.

Software and data used for this analysis are available at http://noble.gs.washington.edu/proj/footprinting/.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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 labs for many helpful discussions. This work was supported by NIH grants R01GM071923 and U54HG004592 to J.S. and P41RR11823 to S.F. S.F. is an Investigator of the Howard Hughes Medical Institute. X.C. was supported by a fellowship from the Natural Sciences and Engineering Research Council of Canada (NSERC PGS D3).

References

1. Maniatis T, Ptashne M. Structure of the lambda operators. Nature. 1973; 246:133–6. [PubMed: 4586104]
2. Gilbert, W. Polymerization in Biological Systems. Elsevier; North-Holland, Amsterdam: 1972. p. 245-259.
3. Galas DJ, Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 1978; 5:3157–70. [PubMed: 212715]
4. Ren B, et al. Genome-wide location and function of DNA binding proteins. Science. 2000; 290:2306–9. [PubMed: 11125145]
5. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007; 316:1497–502. [PubMed: 17540862]
6. Wei CL, et al. A global map of p53 transcription-factor binding sites in the human genome. Cell. 2006; 124:207–19. [PubMed: 16413492]
7. Gross DS, Garrard WT. Nuclease hypersensitive sites in chromatin. Annu Rev Biochem. 1988; 57:159–97. [PubMed: 3052270]
8. Sabo PJ, et al. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods. 2006; 3:511–8. [PubMed: 16791208]
9. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994; 2:28–36. [PubMed: 7584402]
10. MacIsaac KD, et al. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics. 2006; 7:113. [PubMed: 16522208]
11. Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005; 15:1034–50. [PubMed: 16024819]
12. Harbison CT, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004; 431:99–104. [PubMed: 15343339]
13. Cliften P, et al. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science. 2003; 301:71–6. [PubMed: 12775844]
14. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 2003; 423:241–54. [PubMed: 12748633]
15. Borneman AR, et al. Divergence of transcription factor binding sites across related yeast species. Science. 2007; 317:815–9. [PubMed: 17690298]
16. Tan S, Richmond TJ. Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. Nature. 1998; 391:660–6. [PubMed: 9490409]
17. Ferre-D'Amare AR, Pognonec P, Roeder RG, Burley SK. Structure and function of the b/HLH/Z domain of USF. Embo J. 1994; 13:180–9. [PubMed: 8306960]

18. Acton TB, Zhong H, Vershon AK. DNA-binding specificity of Mcm1: operator mutations that alter DNA-bending and transcriptional activities by a MADS box protein. Mol Cell Biol. 1997; 17:1881–9. [PubMed: 9121436]
19. Wang KL, Warner JR. Positive and negative autoregulation of REB1 transcription in Saccharomyces cerevisiae. Mol Cell Biol. 1998; 18:4368–76. [PubMed: 9632820]
20. Planta RJ, Goncalves PM, Mager WH. Global regulators of ribosome biosynthesis in yeast. Biochem Cell Biol. 1995; 73:825–34. [PubMed: 8721998]
21. Boeger H, Griesenbeck J, Kornberg RD. Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. Cell. 2008; 133:716–26. [PubMed: 18485878]
22. Mavrich TN, et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res. 2008; 18:1073–83. [PubMed: 18550805]
23. Lee W, et al. A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet. 2007; 39:1235–44. [PubMed: 17873876]
24. Shivaswamy S, et al. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS Biol. 2008; 6:e65. [PubMed: 18351804]
25. Yuan GC, et al. Genome-scale identification of nucleosome positions in S. cerevisiae. Science. 2005; 309:626–30. [PubMed: 15961632]
26. Raisner RM, et al. Histone variant H2A.Z marks the 5′ ends of both active and inactive genes in euchromatin. Cell. 2005; 123:233–48. [PubMed: 16239142]
27. Jakobsen BK, Pelham HR. Constitutive binding of yeast heat shock factor to DNA in vivo. Mol Cell Biol. 1988; 8:5040–2. [PubMed: 3062378]
28. Strauss EC, Orkin SH. In vivo protein-DNA interactions at hypersensitive site 3 of the human beta-globin locus control region. Proc Natl Acad Sci U S A. 1992; 89:5809–13. [PubMed: 1631062]
29. David L, et al. A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci U S A. 2006; 103:5320–5. [PubMed: 16569694]

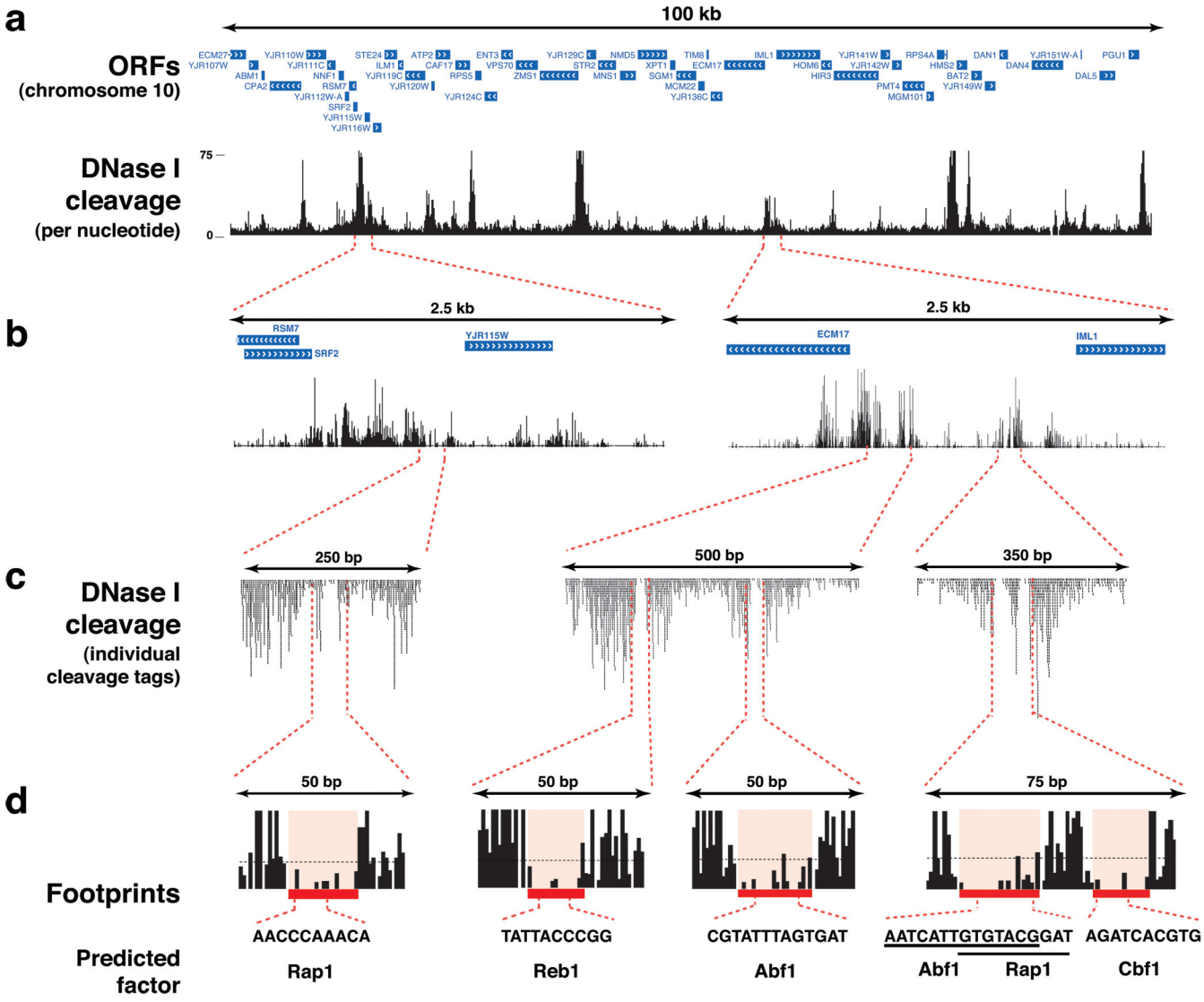

Fig. 1. Digital DNase I analysis of yeast chromatin structure from chromosomal to nucleotide resolution

(**a**) Per-nucleotide DNase I cleavage density across an exemplary 100-kilobase region of chromosome 10 (chr10:625,000-725,000) containing ~50 open reading frames (ORFs). (**b**) Magnification of exemplary ~2.5 kb regions containing *RSM17*/*YJR115W* and *ECM17*/*IML1* intergenic intervals. (**c**) Further magnification showing positions of individual DNase I cleavage events (stacked vertical black tick marks), revealing short regions protected from DNase I cleavage (DNase I "footprints"). (**d**) Resolution of individual DNase I footprints (red shading) with known motifs for yeast regulatory factors Rap1, Reb1, Abf1 and Cbf1. The dashed black line indicates the average level of DNase I cleavage throughout the genome (avg. ~2 cleavages per bp).

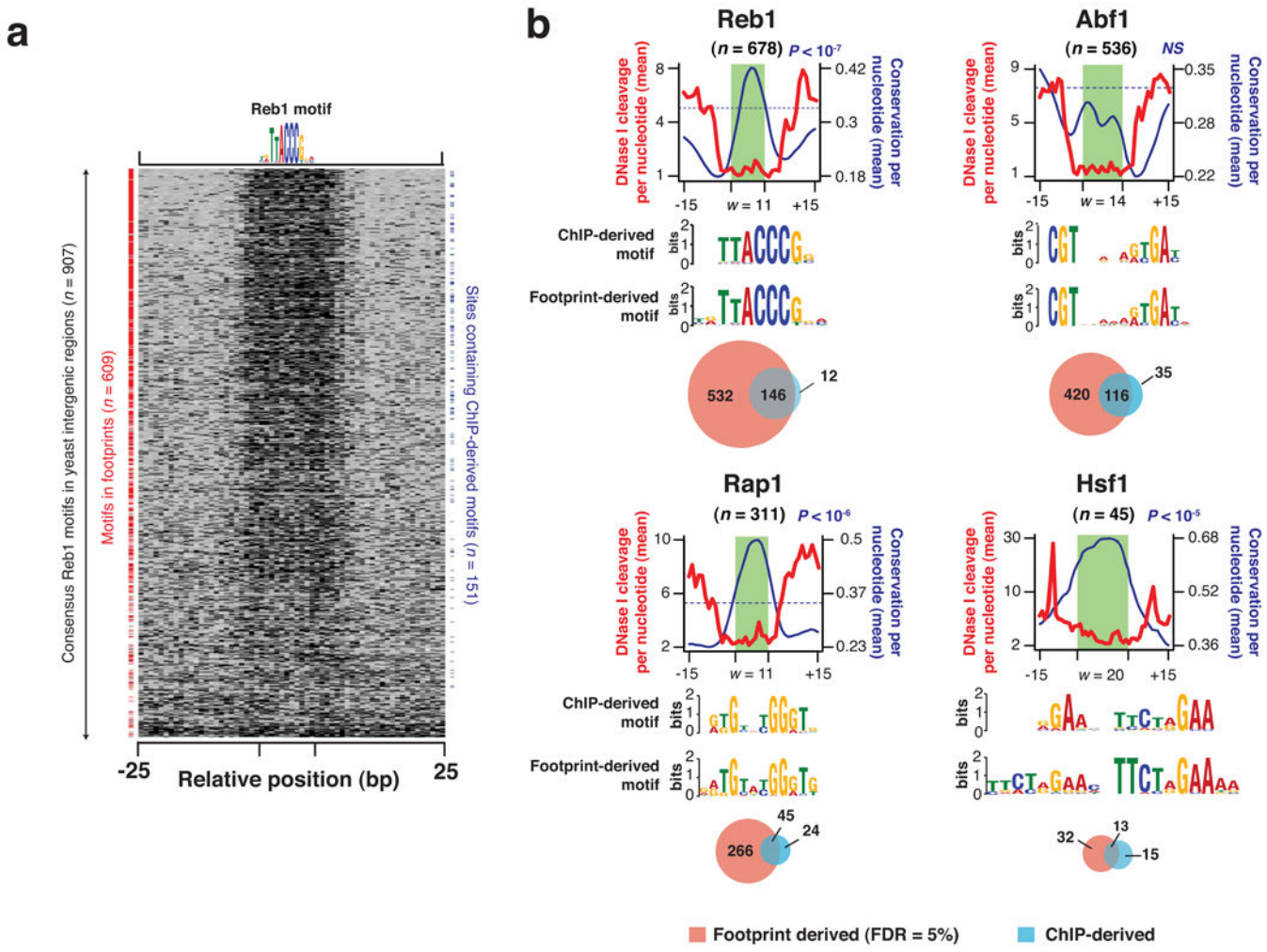

Fig. 2. Detection of footprints and corresponding sequence motifs

(**a**) Visualization of DNase I protection (footprinting) around 907 computationally-predicted Reb1 sites in a heat map. Rows show levels of DNase I cleavage 25 bp up- 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 (at left) indicate motif instances (n = 580) that coincide with footprints (FDR = 0.05) containing *de novo*-derived Reb1 motifs. Blue ticks (right) indicate motif instances (n = 151) coinciding with those identified by ChIP10. All motif instances are uniquely mappable within the yeast genome. (**b**) Mean per nucleotide DNase I cleavage (red) and evolutionary conservation (Phastcons?; blue) calculated for footprints that match the Reb1, Abf1, Rap1 and Hsf1 motifs (subpanel vertical axes). Significance of observed conservation patterns (blue text) (Supplementary Methods), extent of consensus motifs derived from the footprinted region (green shading), motifs derived from ChIP and footprinting below. Venn diagrams depict the overlap of motifs derived from and mapping to footprints (red) vs. ChIP (blue).

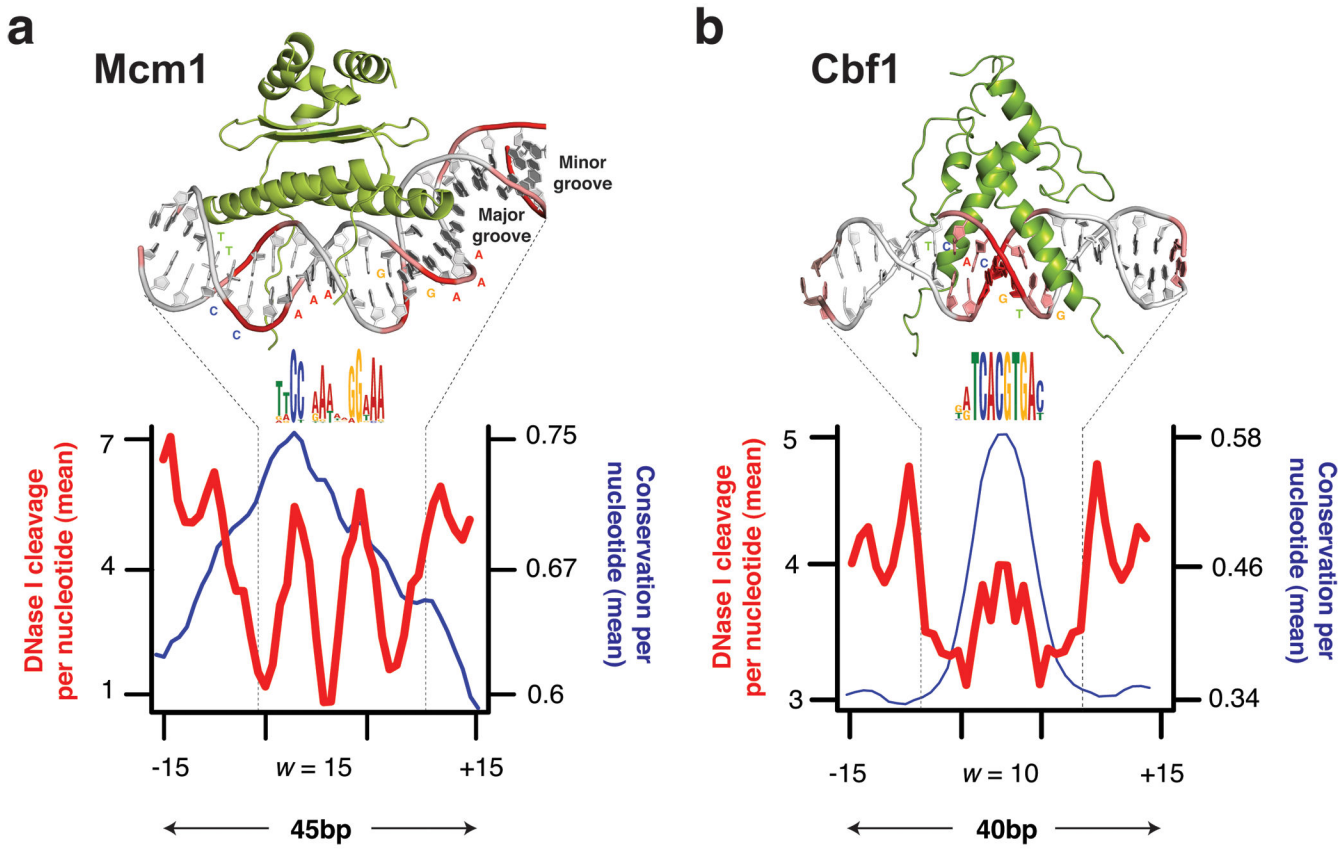

Fig. 3. Mean nucleotide-level accessibility parallels protein:DNA interactions

(**a**) Structure of Mcm1 (green) bound to a single DNA recognition site16 (adjacent Matα2 was removed for clarity). Colored DNA bases correspond to positions within the footprint-derived Mcm1 motif (below), and red DNA backbone coloration reflects the extent of observed DNase I cleavage across 88 Mcm1 sites (red trace in subpanel). Mean nucleotide-level conservation by PhastCons11 is shown in parallel (blue trace in subpanel; $P < 10^{-5}$). (**b**) Structure of the human homolog of CBF117 is shown relative to the mean nucleotide level cleavage and conservation ($P < 10^{-3}$) across 243 Cbf1 sites.

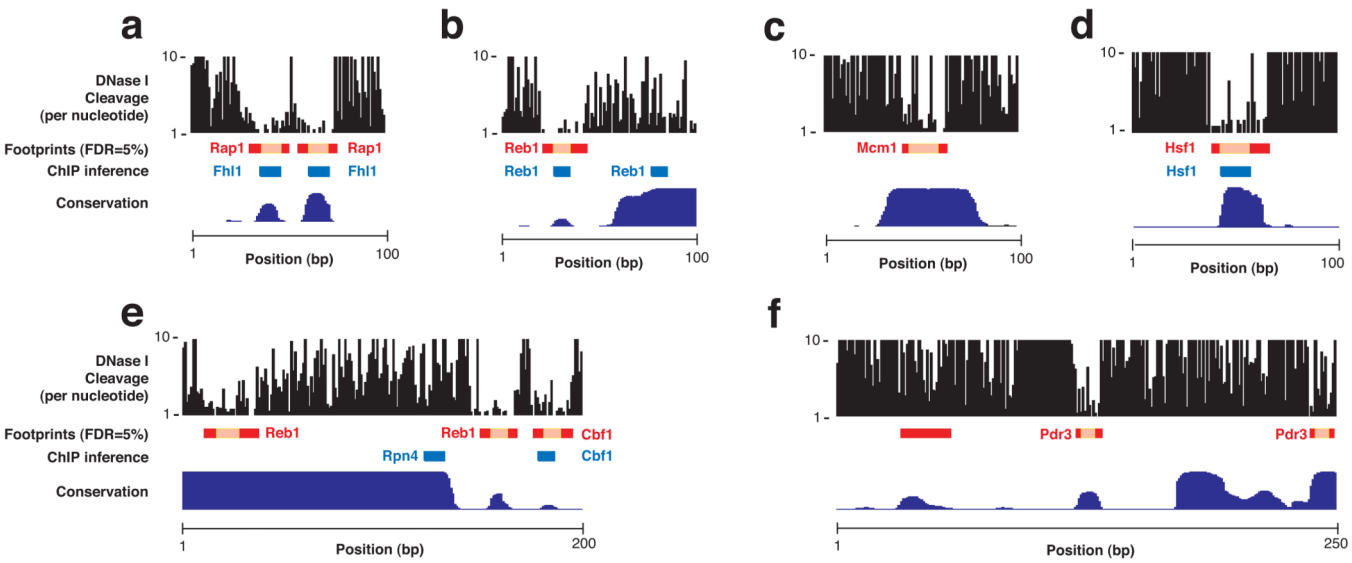

Fig. 4. Individual yeast regulatory regions and factor binding sites

(**a**) Rap1 binds to two adjacent sites also predicted from ChIP experiments upstream of *RPS6A* (chr16:378,775-378,874). (**b**) Reb1 binds to a canonical site upstream of *TUF1* (chr15:683,707-683,806) but a non-canonical site upstream is only inferred from ChIP data (**c**) Mcm1 site upstream of *MFA1* (chr4:1,384,893-1,384,993) exhibits hypersensitive nucleotides illustrated in Fig. 3a. (**d**) Hsf1 site identified by ChIP in *BTN2* promoter (chr7:772,068-772,167) is identified as a footprint. (**e**) Two Reb1 binding sites in the *REB1* promoter (chr2:336,885-337,084) are identified as footprints; a Cbf1 site predicted by ChIP shows a footprint, but a Rpn4 site defined by ChIP does not. (**f**) Two Pdr3 sites in the *PDR5* promoter (chr15:619,227-619,476) are identified as footprints, in addition to an evolutionarily conserved region further upstream. 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, Phastcons, bottom).

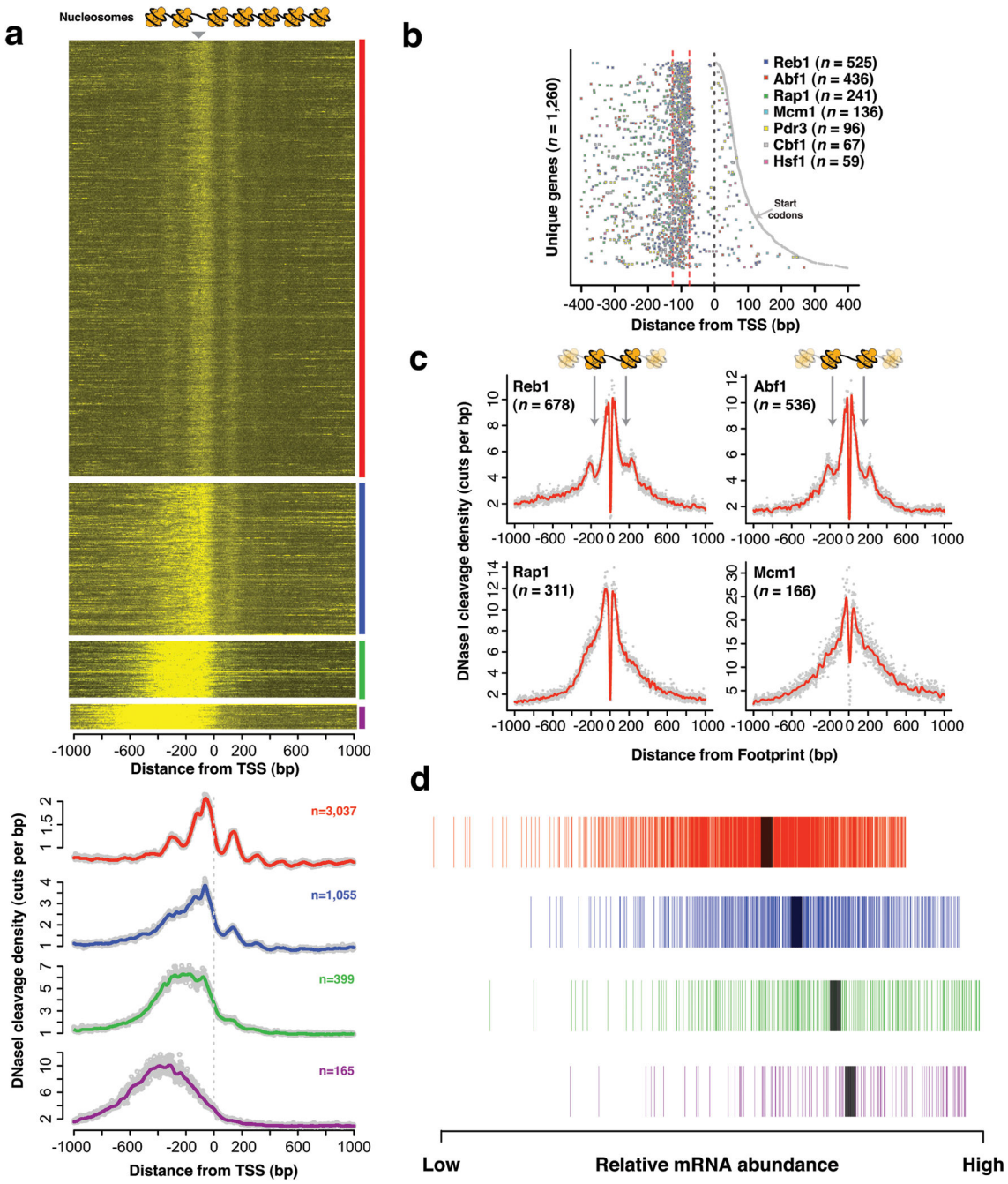

Fig. 5. Higher-order patterns of DNA accessibility

(**a**) Mapped DNase I cleavages relative to 5,006 TSSs29. Four major clusters are exposed by *k*-means analysis (red, blue, green and purple bars, respectively). In the red cluster, maximal DNase I cleavage occurs in a stereotypic ∼50 bp band ∼100 bp upstream of the TSS (grey arrowhead, top). In the blue, green and purple clusters, the extent and intensity of DNase I cleavage upstream of the TSS widens to the -1, -2, and -3 nucleosomes (respectively). (**b**) Spatial restriction of footprints near TSSs. Distribution of footprints matching Reb1, Abf1, Rap1, Mcm1, Pdr3, Cbf1 and Hsf1 relative to the TSSs (dashed black lines) and start codons

of 1,260 genes sorted by the length of the 5′UTR. Enrichment within a ∼50 bp region centered ∼100 bp upstream of the TSS (dashed red lines). (**c**) DNase I cleavage profiles aligned relative to Reb1, Abf1, Rap1 and Mcm1 footprints. (**d**) mRNA abundance for genes found in each of the four clusters correlates with the accessibility of the promoters of those genes (colors as in **a**; median expression denoted by a black bar).