Published online 16 April 2015 *Nucleic Acids Research, 2015, Vol. 43, No. 9* ***4429–4446***
doi: 10.1093/nar/gkv334

Survey of protein–DNA interactions in *Aspergillus oryzae* on a genomic scale

Chao Wang[†], Yangyong Lv[†], Bin Wang[†], Chao Yin, Ying Lin and Li Pan[*]

School of Bioscience and Bioengineering, South China University of Technology, Guangzhou, Guangdong, 510006, China

Received May 21, 2014; Revised March 20, 2015; Accepted March 31, 2015

ABSTRACT

The genome-scale delineation of *in vivo* protein–DNA interactions is key to understanding genome function. Only ∼5% of transcription factors (TFs) in the *Aspergillus* genus have been identified using traditional methods. Although the *Aspergillus oryzae* genome contains >600 TFs, knowledge of the *in vivo* genome-wide TF-binding sites (TFBSs) in aspergilli remains limited because of the lack of high-quality antibodies. We investigated the landscape of *in vivo* protein–DNA interactions across the *A. oryzae* genome through coupling the DNase I digestion of intact nuclei with massively parallel sequencing and the analysis of cleavage patterns in protein–DNA interactions at single-nucleotide resolution. The resulting map identified overrepresented *de novo* TF-binding motifs from genomic footprints, and provided the detailed chromatin remodeling patterns and the distribution of digital footprints near transcription start sites. The TFBSs of 19 known *Aspergillus* TFs were also identified based on DNase I digestion data surrounding potential binding sites in conjunction with TF binding specificity information. We observed that the cleavage patterns of TFBSs were dependent on the orientation of TF motifs and independent of strand orientation, consistent with the DNA shape features of binding motifs with flanking sequences.

INTRODUCTION

The genus *Aspergillus* includes human and plant pathogens and beneficial species that produce foodstuffs and industrial enzymes. Within this genus, *Aspergillus oryzae* is used to manufacture Asian fermented foods and is regarded as a suitable host for homologous and heterologous protein production. The genome sequencing of *A. oryzae* has led to genomic-scale studies (1,2). In *A. oryzae*, genomic-scale transcription analyses have been performed using expressed sequence tags (3,4), microarray platforms (5,6), RNA sequencing (7) and proteomic analyses (8) based on two-dimensional electrophoresis and high-throughput chromatographic separations from protein mass spectrometry. These tools have been used to globally monitor thousands of transcripts or proteins to systematically determine their physiological states. However, the activity of each gene is regulated through the binding of regulatory proteins to DNA sequences. The delineation of the complete set of genomic sites through *in vivo* protein interactions with *cis*-regulatory DNA elements remains a major challenge in understanding biological macromolecule functions. The availability of the *Aspergilli* genome sequence facilitates the identification of a large number of putative genes encoding DNA-binding proteins (1,9). Approximately 5% of the transcription factors (TFs) in the *Aspergillus* genus have been identified (1,9). Most of the current knowledge concerning *Aspergillus* TFs and their binding sites is derived from traditional *in vitro* and *in vivo* approaches, such as electrophoretic mobility shift assays, DNA footprinting using DNase I or dimethylsulphate, and promoter deletion analyses coupled with reporter gene assays. Although these classical approaches are precise and complementary, these techniques are laborious, low-throughput and challenging for the study of *in vivo* protein binding across the entire genome. Identifying genome-wide binding sites for TFs *in vivo* is a critical step toward deciphering genome function. The identification of TF-binding sites (TFBSs) is challenging because the *cis*-regulatory DNA elements recognized by TFs are often short and dispersed throughout the genome and the *in vivo* target loci of a TF vary depending on physiological conditions. The current knowledge of genome-wide TF binding events in *Aspergilli* remains limited.

Both computational and experimental techniques have been developed to identify the location of TFBSs on a genomic scale. Computational predictions (10) based on scanning the genome sequence for DNA motifs represented through a position-specific scoring matrix (11) have been used to analyze TFBSs. Additional information, such as the conservation of TFBSs and co-expressed genes, improves

[*] To whom correspondence should be addressed. Tel: +86 20 39380616; Fax: +86 20 39380698; Email: btlipan@scut.edu.cn

[†] The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint first authors.

prediction accuracy. Because most DNA motifs are four to eight bases in length, annotations are highly prone to false-positive predictions (12). Furthermore, none of these computational methods can be used to study condition-dependent dynamic TF-binding activities (12). Chromatin immunoprecipitation (ChIP) coupled with DNA microarray (ChIP-chip) (13) and massively parallel sequencing (ChIP-seq) (14) could be used to localize genome-wide TF–DNA interaction sites *in vivo* and have become the gold standard for the genome-wide identification of TFBSs in higher eukaryotes. However, ChIP assays are limited because these methods only survey the binding location of a single TF per experiment (15) and do not resolve protein–DNA interactions at a base-pair resolution. ChIP assays also require a high-quality factor-specific reagent. The *A. oryzae* genome has >600 TFs (1) for which high-quality antibodies are lacking. Gene transformation and knockout technologies are also inefficient in *A. oryzae.* Thus far, no genome-wide TF-binding maps have been generated for *A. oryzae* using ChIP-seq and ChIP-chip approaches.

A common characteristic of genomic regulatory regions is the binding of TFs at locations of canonical nucleosomes, resulting in hypersensitivity to DNase I cleavage (16). Steric hindrance of DNase I access to DNA has been associated with TF occupancy (16). DNase I digestion coupled with high-throughput sequencing (DNase-seq) (17) and tiling DNA microarrays (18) are powerful tools for mapping genome-wide DNase I hypersensitive sites (DHSs) at a single-base resolution. DNase-seq has been applied to identify a variety of *cis*-regulatory DNA elements *in vivo* and simultaneously monitor the genome-wide binding sites of many TFs in *Saccharomyces cerevisiae* (19,20), humans (21–24), *Arabidopsis* (25) and the prokaryote *Bacillus subtilis* (26). A high-depth sequencing technique can be used to identify depleted narrow regions in the DHS regions of a genome corresponding to a single TF footprint, referred to as genomic-scale digital genomic footprinting (DGF) (20,21). With sufficient sequencing data, DGF can be used to identify single protein-binding events and narrow DNA footprints with significant enrichment for known motifs and *de novo* motif discovery (20,21).

Here, we describe the landscape of *in vivo* protein–DNA interactions in the *A. oryzae* genome using DNase I cleavage profiles by coupling the DNase I digestion of intact nuclei with massively parallel sequencing. The resulting map identified overrepresented *de novo* TF-binding motifs from genomic footprints and correlated chromatin remodeling patterns in the neighboring regions of transcription start sites (TSSs), the 5′ untranslated regions (5′-UTRs) of target genes and their expression, and the distribution of transcriptional regulators. The active TFBSs of 19 known *Aspergillus* TFs were further identified based on DNase I digestion data surrounding candidate binding sites in conjunction with TF binding specificity information. Furthermore, the DNase I cleavage patterns of TFs in *A. oryzae* were consistent with the DNA shape features of binding motifs with flanking sequences.

MATERIALS AND METHODS

Strains and culture conditions

A. oryzae strain RIB 40 was obtained from the NITE Biological Resource Center (NBRC) in Japan. For nutrient-rich culture conditions (DPY conditions), 1 × 10^8 spores were inoculated into 100 ml DPY liquid medium (2% dextrin, 1% peptone, 0.5% yeast extract, 0.5% KH_2PO_4, 0.05% $MgSO_4{\cdot}7H_2O$ and 0.02% KCl) and cultivated at 30°C with shaking at 200 rpm for 24 h. For endoplasmic reticulum (ER) stress culture conditions (unfolded protein response (UPR) conditions), 1 × 10^8 spores were inoculated into 100 ml CD medium (2% glucose, 0.1% KH_2PO_4, 0.05% $MgSO_4{\cdot}7H_2O$, 0.02% KCl, 0.3% $NaNO_3$, and 0.001% $FeSO_4{\cdot}H_2O$) and cultivated at 30°C with shaking at 200 rpm for 40 h. DTT was added to the culture at a final concentration of 20 mM for ER-stress induction for 2 h. Mycelia were harvested for nuclei extraction.

DNase I digestion of nuclei extracts

Nuclei extracts were prepared according to a previously described method with some modifications (27). *A. oryzae* mycelia were harvested through filtration, immediately washed twice with cold sterile water and ground to a powder in liquid nitrogen. Four grams of ground mycelium powder was dissolved in 80 ml of nuclease isolation buffer (250 mM sucrose, 60 mM KCl, 15 mM NaCl, 0.5 mM DTT, and 15 mM Tris-HCl, pH 7.5) and homogenized through ∼20 strokes in a Dounce homogenizer to release the nuclei. The homogenized suspension was centrifuged at 4000 × g for 20 min at 4°C. The pellets were washed twice with 20 ml of nuclease isolation buffer and centrifuged at 12 000 × g for 5 min at 4°C. The pellets were resuspended in 4.8 ml of nuclease digestion buffer (250 mM sucrose, 60 mM KCl, 15 mM NaCl, 0.05 mM $CaCl_2$, 3 mM $MgCl_2$, 0.5 mM DTT, and 15 mM Tris-HCl, pH 7.5). DNase I (Roche) digestion was initiated immediately after resuspension through the addition of different amounts of the enzymes to aliquots of the suspension. The mixture was incubated for 5 min at 25°C with the optimal enzyme concentration of 100 U/ml. The reactions were terminated through the addition of an equal volume of stopping buffer (40 mM EDTA and 2% SDS). Digested fragments were extracted with phenol-chloroform and examined through gel electrophoresis. DNase I digestion produced a smear of low-molecular-weight DNA fragments with the bulk of high-molecular-weight fragments. Fragments of ≤500 bp were recovered from gels and purified using MinElute spin columns (QIAGEN, Dusseldorf, Germany) to construct a sequencing library.

Validation of DNase I digestion products

Real-time quantitative PCR (RT-qPCR) was performed to determine the degree of DNase I digestion. Three regions were selected for validation: a protein-coding region of S1 endonuclease, a random intergenic region and the promoter region of bipA. *A. oryzae* RIB40 samples for RT-qPCR were obtained as for DNase-seq, and the sample concentrations were determined through spectrophotometry. The primers for RT-qPCR are listed in Supplementary Table S9.

RT-qPCR reactions, including 10 μl of reaction mixture with 50 ng DNase I digestion product, 2 × 0.4 μl primers (forward and reverse, 10 μM), 0.2 μl ROX reference dye II (50×), 5 μl SYBR Premix Ex Taq II (2×) and dH_2O, were amplified using an Applied Biosystems 7500 Real-time PCR System for 1 min at 95°C, followed by 40 cycles of 95°C for 5 s, 55°C for 20 s and 72°C for 34 s. The degree of DNase I digestion was determined based on changes in Ct values.

Construction and sequencing of the DNase I library

Briefly, purified DNase I digestion fragments (<500 bp) were subjected to end repair using T4 DNA polymerase, Klenow fragment and T4 polynucleotide kinase. Following end repair, an 'A' base was added to the 3′ end of the blunt phosphorylated DNA fragments using Klenow fragment. After purification, Illumina adapters were ligated to the ends of the DNA fragments. Adapter-ligated DNA fragments were purified and further enriched through PCR amplification using only 16 cycles. The amplified libraries were purified and assessed using a library quality test. The qualified DNase I library was sequenced on an Illumina HiSeq 2000 System with a read length of 90 bp. The quality standard of the Illumina sequencing reads was controlled by BGI Shenzhen (China): low-quality reads containing adapters, uncertain nucleotides (Ns) > 9, and bases with sequencing quality scores ≤ 5 for more than half of the read length (90 bp) were discarded. The resulting DNase-seq data of *A. oryzae* cultured under DPY and UPR conditions were deposited in Sequence Read Archive (SRA) at NCBI under accession numbers SRX607943 and SRX610905.

Computing DNase I dinucleotide cleavage

The resulting reads were aligned to the *A. oryzae* RIB40 genome retrieved from AspGD (version s01-m06-r03, www.aspergillusgenome.org) using Bowtie (28) version 0.12.5 with the following parameters: *bowtie -a -m 1 –best –strata -n 2 –trim3 54* for Illumina HiSeq sequencer runs. The parameter *–trim3 54* indicates trimming 54 bp from the 3′ (right) end of the reads, and the leaving 36-bp reads were aligned to the reference genome, which is consistent with *A. oryzae* genome mappability data of the 36-bp reads. Only uniquely mapped reads of high quality were used, and reads with multiple mapping positions or >3 mismatches were discarded. A density of DNase I cleavage per nucleotide was calculated based on the number of uniquely mapped sequence tags with 5′ ends mapping to a certain genome position.

Uniquely mappable nucleotide positions in the *A. oryzae* genome

The mappability map was constructed for each base pair in the genome, counting the number of locations with which a subsequence starts at that position. The genome mappability, which defines uniquely mappable nucleotide positions in the reference genome, masked multiple mapped and repeated regions. The 36-bp reads in the *A. oryzae* genome were used to generate mappability data by Mappability_Map (29). The resulting data of genome unmappability and mappability were used for further analysis of genomic footprints and hotspot algorithms, respectively.

Identification of DNase I footprints

To identify DNase I footprints across the genome, we used a computation algorithm applied previously to the yeast genome (20) to identify short regions (8–30 bp) with low DNase I cleavage rates relative to adjacent flanking regions (150 bp). The scripts of DGF were downloaded from the Noble Research Lab at the following website: http://noble.gs.washington.edu/proj/footprinting/. Only the maxColNum parameter, delimiting the maximum size of an intergenic region, was changed to 22 000 in the footprinting_run_all.sh Linux shell file, as the maximum size of the intergenic region in *A. oryzae* genome is 22 kb.

The result was used for the subsequent analysis of footprints overlapping and motif discovery. The overlap between the UPR and DPY footprint, with its length >10% of the footprint length, was considered as an overlap. Furthermore, the overlaps, with their lengths >50% of the corresponding footprints length, accounted for 95% and 92% in all overlapping footprints with the false-discovery rate (FDR) thresholds of 0.05 and 0.1, respectively (Supplementary Table S2).

De novo motif discovery from DNase I footprints

Genomic footprints were used for *de novo* motif discovery. Footprints of start and stop coordinates were extended by 10 bp upstream or downstream to include potential footprint boundaries. We used the MEME software package (30) to identify overrepresented motifs ($E > 0$) in the set of footprints. Overrepresented motifs were used to search original footprint regions (FDR < 0.05) without 10 bp upstream or downstream regions using FIMO (31) ($P < 10^{-4}$) to identify all motif instances in footprints. *De novo* motifs were assigned to known motifs available in databases, including fungi and yeast databases, using TOMTOM (32). To evaluate the potential impact of *de novo* motifs on gene regulation, we defined a gene as the target of a motif if it contained a motif instance within a 1-kb flanking region of the nearest TSS. Gene Ontology (GO) analyses for target genes of the same motif were performed using Cytoscape with the Cluego plug-in (33).

Identification of DHSs

Uniquely mapped reads of 36 bp derived from *A. oryzae* cultured under DPY and UPR conditions were considered to identify DHSs using the HotSpot algorithm version 3 (34). DNase I cleavage tags represented the 5′ DNase I cleavage site from each uniquely mapped read. Hotspot was used to identify regions (hotspot regions) of local enrichment of DNase I cleavage tags mapped to the genome using a binomial distribution model (34). DNase I tag densities were summed every 20 bp in a 150-bp sliding window, and peak-finding of the density in each hotspot region was performed (34). DHSs were identified as significant tag density peaks within DNase I sensitive regions.

Analysis of known motifs enriched in DHSs

Known motifs of TFs were used to scan intergenic regions of the *A. oryzae* genome for candidate binding sites using FIMO software (31). We buffered (±100 nucleotides) a motif instance of the candidate binding site, and at each base position, we counted the number of uniquely mapped DNase I tags with 5′ ends mapping to the position as specific experimental data. We combined genomic information with specific experimental data using the MILLIPEDE model (35) to estimate the probability that the candidate binding site was bound. We focused on DHSs (defined as 150 nucleotides in length) identified based on cleavage density peaks within high-cleavage-density regions (DNase I hypersensitive regions, DHS) using the Hotspot algorithm (34). We filtered out the motif instances of the candidate binding sites with MILLIPEDE probability values greater than 0.5, which occurred in DHSs with FDR of 1%. MILLIPEDE was run using a partially unsupervised method (35).

High-throughput DNA shape prediction

For each position of a TF motif instance with a 10-bp upstream and downstream region, we predicted four structural features of the DNA shape: minor groove width (MGW) and propeller twist (ProT) for each nucleotide position, and roll and helix twist (HelT) as base pair-step parameters. The DNA shape analysis was performed using a high-throughput prediction approach (36) to infer structural features from a library of 512 unique pentanucleotides derived from Monte Carlo simulations of 2121 DNA fragments. The Pearson correlation coefficient (PCC) was used to measure the linear dependency between DNA shape parameters and DNase I cleavage patterns.

Strand-specific RNA-seq for TSS reannotation of the *A. oryzae* genome

The total RNA from *A. oryzae* cultured under DPY medium was extracted using RNAiso^{TM} Plus (TaKaRa, Japan), treated with RNase-free DNase I (TaKaRa, Japan) and purified. The RNA integrity was analyzed using an Agilent Technologies 2100 Bioanalyzer.

A strand-specific RNA-seq library was prepared according to the method of Parkhomchuk *et al.* with some modifications (37). Briefly, Sera-mag magnetic Oligo(dT) beads were used to isolate poly(A)-mRNA from purified total RNA. Random hexamer primers were used to synthesize first-strand cDNA using SuperScript II. After purification, dUTP-containing second-strand cDNA was synthesized after the addition of buffer, dNTPs with dTTP replaced by dUTP, RNaseH and DNA polymerase I. Double-stranded cDNA fragments were subjected to end repair, phosphorylation and 3′-adenylation, and Illumina sequencing adapters were ligated to the 3′-adenylated cDNA fragments. Uracil-DNA glycosylase was added to digest dUTP-containing second-strand cDNA in alkalescent medium at a high temperature. The remaining first-strand cDNA was enriched by 15 rounds of PCR amplification. After purification and quality control, the strand-specific RNA-seq cDNA library was sequenced on the Illumina HiSeq2000 platform using a 90-bp paired-end sequencing strategy. The quality standard of the Illumina sequencing reads were controlled by BGI Shenzhen (China): low-quality reads containing adapters, uncertain nucleotides (Ns) > 9 and bases with sequencing quality scores ≤ 5 for more than half of the read length (90 bp) were discarded. The resulting Illumina sequencing data of *A. oryzae* cultured under DPY medium were deposited in SRA at NCBI under accession number SRX610909.

The resulting reads were mapped to the *A. oryzae* RIB40 genome and genes annotated by AspGD (version s01-m06-r03, www.aspergillusgenome.org) using SOAP2 (38) version 2.21.

To identify 5′-UTRs, the upstream sequences of *A. oryzae* genes were searched to identify TSSs with read counts of zero. The region between the TSS and translation start codon was designated as the 5′-UTR. The RPKM values of gene expression were calculated based on the number of reads per kilobase pairs per million mapped reads, according to the number of reads that were uniquely mapped to genes (39). The FPKM values of gene expression were measured using Cufflinks based on the number of fragments per kilobase of exons per million mapped reads, where the fragment means the two reads that comprise a paired-end read were counted as one (40). TOPHAT (41) was applied to map resulting reads to the *A. oryzae* RIB40 genome with the optional parameters of -r 20 –mate-std-dev 40 –solexa1.3-quals. The mapping results from TOPHAT were fed to Cufflinks (40) with default parameters for calculating FPKM values.

Analysis of the DNase I cleavage pattern relative to TSSs

DNase I cleavage counts among +/− 1 kb regions relative to TSSs were extracted for all genes with annotated 5′-UTRs based on strand-specific RNA-seq data. The average count of DNase I cleavages for each nucleotide position was calculated. DNase I cleavage patterns at +/− 1 kb TSS flanking regions were sorted according to 5′-UTR length and expression level using heat maps generated through Java Treeview (42). To illustrate the spatial distribution of footprints surrounding the TSS, the digital footprints ($FDR < 0.05$) among −500/+1000-bp regions relative to TSSs (black vertical line) sorted according to the 5′-UTR length of its targeted genes were plotted to accompany translation start sites (gray curved line) of the corresponding genes.

RESULTS

Genome-scale identification of DNase I footprints from the DNase-seq library

DNase I digestion of *A. oryzae* nuclei coupled with massively parallel sequencing was used to create a whole-genome DNase I cleavage library under conditions including nutrient-rich culture (DPY condition) and ER-stress induction (UPR condition). Based on the ssRNA-seq data obtained from *A. oryzae* cultures under DPY conditions, 39.3% of 346 TFs in six families (including bHLH, bZIP, C2H2 zinc finger, GATA type zinc finger, Zn(II)2Cys6 zinc finger and CCAAT-binding complex (CBC)) showed high-level expression from 10 to 1400 RPKM (Supplementary Table S6). The TFs that regulate responses under ER stress were analyzed under UPR culture conditions. The expressions of 24 TFs in six TF families are up-regulated by 1.5-fold (P-value < 0.01) over control conditions (7). *A. oryzae* nuclei were isolated and treated with DNase I to release DNA fragments of <500 bp derived from double-hit DNase I cutting events occurring close to each other. RT-qPCR was used to verify the sensitivity of the DNase I library to ensure that the PCR amplification of active regions containing regulatory elements from DNase I-treated DNA templates generated the same quantity of PCR products with additional PCR cycles as the control of naked genomic DNA without DNase I digestion, measured based on the ΔCt value (Supplementary Figure S1) (43). We employed ultra-deep sequencing of the *A. oryzae* DNase I library to obtain 163.96 million and 106.26 million reads of an average length of 90 bp from cells grown under DPY medium condition and UPR induction condition, respectively (Supplementary Table S1). Overall, 97.77% (160.30 M) and 95.36% (100.20 M) of the reads derived from DPY and UPR conditions, respectively, could be uniquely mapped to the *A. oryzae* genome, where the length of 90-bp reads were trimmed to 36 bp (Supplementary Table S1). DNase I cleavage sites mapped by these uniquely mapped reads were confined to 30.71 and 27.10 million unique positions within the *A. oryzae* genome. The genome mappability data of the 36-bp reads generated using Mappability_Map (29) masked multiple mapped and repeated regions in the *A. oryzae* genome.

To identify DNase I footprints *in vivo* across the *A. oryzae* genome, we used a computational algorithm previously applied to the yeast genome (20) to identify 4544 DNase I footprints from cells grown under DPY condition and 4148 DNase I footprints from cells grown under UPR induction condition in the intergenic regions of the *A. oryzae* genome with an FDR threshold of 0.05 (Supplementary Table S2). We identified 2288 genes under DPY condition and 2346 genes under UPR induction condition with at least one footprint in the 1-kb upstream region. There are 310 genes under DPY condition and 288 genes under UPR induction condition containing two or more footprints. At an FDR threshold of 0.1, we identified 8125 DNase I footprints under DPY condition and 8894 footprints under UPR induction condition (Supplementary Table S2) distributed in 1-kb upstream regions of 3835 genes under DPY condition and 4341 genes under UPR induction condition. More than two footprints were identified for 800 genes under DPY condition and 970 genes under UPR induction condition. Furthermore, 698 and 1583 DNase I footprints with FDR thresholds of 0.05 and 0.1, respectively, were overlapped between DPY condition and UPR induction condition, which were located in upstream regions of 672 and 1449 genes (Supplementary Table S2). The GO functional enrichment analysis revealed that genes targeted by overlapped footprints with an FDR threshold of 0.05 were primarily involved in 'cellular components of cytosolic ribosome, cytoplasmic membrane-bounded vesicle, proteasome complex, nuclear envelope and site of polarized growth' (Supplementary Figure S2 and Supplementary Table S2). The biological processes of regulation of protein transport, ribose phosphate biosynthetic process, nucleosome organization, negative regulation of reproductive process and response to organic substance were also enriched in these genes (Supplementary Figure S2 and Supplementary Table S2).

De novo identification of motif sequences through genomic DNase I footprinting

To identify TF motifs from the genomic footprints, we extracted 8125 footprints under DPY condition and 8894 footprints under UPR induction condition with an FDR threshold of 0.1 to assay *de novo* sequence motifs using MEME (30). The results were compared with previously described TF motifs. MEME recovered 12 overrepresented motifs from the footprints set under DPY condition and 11 overrepresented motifs from the footprints set under UPR induction condition, corresponding to known *Aspergillus* TFs, including SltA, CpcA and the E-box of bHLH factors (Supplementary Table S3). All instances of each motif were identified using FIMO (31) to search for DNA sequences in 4544 DPY footprints and 4148 UPR footprints with an FDR threshold of 0.05 (Supplementary Table S3). The mean per-nucleotide DNase I cleavage rates across all instances of each motifs were computed in the footprint regions (Figure 1A and Supplementary Figure S3).

DPY_motif 3 (5′-WGCCTSAGGCAKM-3′), UPR_motif 2 (5′-KGCCTSAGGCMDMMA-3′), DPY_motif 7 (5′-DHTGCCTGAAHHHKMAGGCMK-3′) and UPR_motif 5 (5′-GCCTKRWRDHTCAGGCADCMA-3′) displayed similar core elements, including two reverse complementary sequences AGGCA/TGCCT (Supplementary Table S3) matching the binding site AGGCA of the Cys2/His2-type zinc finger protein AO090005001502 (SltA) (44–47). The DNase I cleavage density per nucleotide for all motif instances showed that the DNase I patterns of DPY_motif 3 and DPY_motif 7 were symmetrical in inverted palindromes AGGCA/TGCCT and similar to UPR_motif 2 for DPY_motif 3 and UPR_motif 5 for DPY_motif 7 (Figure 1A). The inverted palindromes AGGCA/TGCCT were spaced by one base in DPY_motif 3 and UPR_motif 2 and by eight bases in DPY_motif 7 and UPR_motif 5 (Supplementary Table S3). These results suggested that the SltA TF-binding motifs contained two different binding patterns, including at least two SltA binding sites (5′-AGGCA-3′ or 5′-TGCCT-3′). The prediction of conserved domains from SMART and InterPro databases showed that AO090005001502, a homologue of SltA, contains three tandem repeats of adjacent Cys2/His2-like zinc finger domains bound to the major groove of DNA. The amino acid encoded by the 421–423-bp region constitutes a classical Cys2-His2 zinc finger with two additional unusual zinc fingers from the amino acids encoded by the 447–475-bp and 482–507-bp regions near the C-terminus of the protein. Cys2/His2 zinc finger-type proteins interact with both strands of DNA with most contacts through a subsite of the G-rich strand with the 5′-to-3′ direction of the subsite antiparallel to the N-to-C direction of the zinc finger protein (48). The binding patterns of the four SltA homologue motifs derived from DPY and UPR conditions contained two G-rich binding sites (5′-AGGCA-3) distributed throughout the two strands. We hypothesized that the two SltA homologue TFs might co-localize in

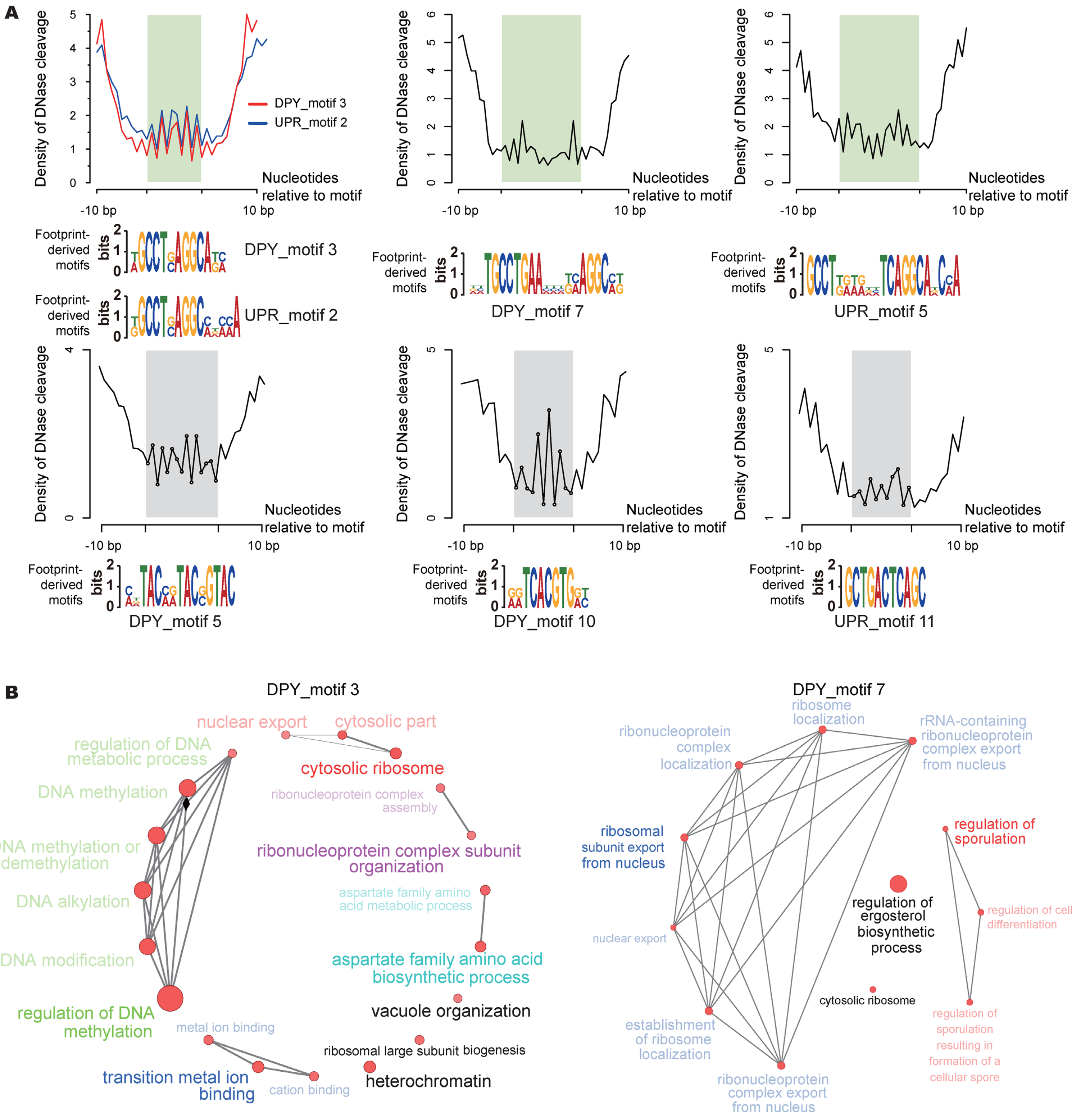


Figure 1. Diversity of DNase I cleavage patterns and function annotation of target genes for the overrepresented motifs in genomic footprints. (**A**) DNase I cleavage density per nucleotide calculated for footprint instances from two culture conditions. Shaded regions delineate the overrepresented motifs derived from the footprint region. The MEME logo of overrepresented motifs derived from footprints is shown below the graph. (**B**) GO function enrichment for the target genes under the DPY_motif 3 and DPY_motif 7. The genes containing at least one motif instance inside the 1-kb region of the annotated TSSs were selected. The genes under the same motif were analyzed using ClueGo. Functional group networks are represented by nodes linked with each other based on their kappa score level (>0.3). The node size represents the percentage of associated genes with the enrichment significance of the term (Term *P*-value < 0.05, red color). The most significant term of each group is shown by the size and color of the caption.

these footprints. The DPY_motif 3 and UPR_motif 2 binding patterns, in which only one base occurred between two complementary binding sites, indicated the potential homodimerization of the two SltA homologue TFs to yield a close binding pattern (Figure 1A) via protein–protein interaction mediated by Cys2/His2 zinc finger domains (49). In the DPY_motif 7 and UPR_motif 5 binding patterns, eight bases occurred between two complementary binding sites to separate the two co-localized TFs of the SltA homologue (Figure 1A). We assumed that SltA homologue TFs regulate different gene functions in *A. oryzae* using two co-localization binding patterns. Two types of binding motifs were scanned in the region 1-kb upstream of TSSs to identify putative target genes, and their associated functions were analyzed (Figure 1B, Supplementary Table S4 and Supplementary Figure S4). The genes under the control of two types of TFs with SltA-binding patterns were enriched for functions involving ribosomes such as 'rRNA and ribosomal subunit export from the nucleus and ribosomal subunit assembly' (Figure 1B and Supplementary Figure S4). The targeted genes with separated binding patterns, such as DPY_motif 7 and UPR_motif 5 binding patterns, were enriched in the function of *asexual sporulation*, resulting in the formation of cellular spores (Figure 1B and Supplementary Figure S4) (47). These results confirmed that the two co-localization binding patterns of SltA homologues in *A. oryzae* regulated genes with different functions. Furthermore, the genes targeted by DPY_motif 3, UPR_motif 2 and UPR_motif 5 were also implicated in heterochromatin and sister chromatid segregation (Figure 1B and Supplementary Figure S4).

DPY_motif 10 (5′-T**CACGTG**-3′) contained the signature sequence CACGTG (Supplementary Table S3), the canonical E-box motif in the binding sites of bHLH E-box family TFs, such as SclR, StuA, SrbA, PacA, DevR and AnBH1, in *A. oryzae*. DPY motif_10 had a 5′ T residue flanking the CACGTG sequence. The DNase I cleavage pattern of DPY_motif 10 (Figure 1A) was similar to patterns of yeast bHLH TF Cbf1 in yeast genomic footprints (20). The genes targeted by DPY_motif 10 were primarily enriched in *protein import*, *nuclear transport* and *coupled ATPase activity* (Supplementary Figure S4 and Supplementary Table S4). The transcriptional activation of sulfur amino acid metabolism in yeast depends on a complex functional factor derived from bHLH TF Cbf1 with two other bZIP TFs Met4 and Met28, which forms over the **TCACGTG** sequence (50). A 5′ T residue flanking the CACGTG sequence in regulatory regions determines major specificity by preventing the binding of other bHLH TFs, such as yeast Pho4 (51,52).

DPY_motif 5 contained the repeated submotif of the palindromic sequence GTAC (Supplementary Table S3), and the DNase I cleavage profile displayed tandem cleavage patterns per the submotif sequence GTAC (Figure 1A). The genes targeted by DPY_motif 5 were enriched in the molecular function of *response to osmotic stress* and *dephosphorylation* (Supplementary Figure S4 and Supplementary Table S4). The top overrepresented motifs in placozoan ribosomal protein gene promoters contain GTAC submotifs once or in repeats, and >97% of placozoan ribosomal protein gene promoters have at least one GTAC motif (53). However, the known motifs in the TRANSFAC database do not possess similar TFBSs (53). The palindromic GTAC core motif has also been identified as the core binding motif of SQUAMOSA promoter binding proteins (SBP) in plants (54–57), although SBP-domain proteins have not been identified in *Aspergillus*. Therefore, another TF is likely to recognize this motif in fungi.

UPR_motif 11 (5′-GC**TGACTCA**GC-3′) comprised the palindromic sequence **TGACTCA** with GC elements on both sides (Supplementary Table S3). This core sequence was identified in the binding site of the transcriptional activator CpcA (AO090009000459) in *A. oryzae* and AP-1 binding site of GCN4 in yeast. Genes targeted by UPR_motif 11 were enriched in the molecular function of *ribosomal subunit export from the nucleus* for releasing ER stress (Supplementary Table S4 and Supplementary Figure S4). CpcA transcription was the first highest of all *A. oryzae* TFs under DPY conditions (Supplementary Table S6). Studies have also shown that CpcA expression is higher under UPR induction conditions compared with control conditions (7).

DNase I cleavage patterns and the distribution of digital footprints near TSSs

To precisely define TSS annotation in *A. oryzae*, we used strand-specific RNA-seq data generated under DPY conditions to determine genome-wide transcription levels. A total of 26 287,168 pair-end reads of 90 bp were mapped to the *A. oryzae* RIB40 genome (2). Approximately 89.79% of all reads were uniquely mapped to the reference genome with a tolerance of 5 bp mismatches, with a perfect match for 80.31% of mapped reads (Supplementary Table S5). Of all reads, 61.17% were uniquely mapped to annotated genes (Supplementary Table S5). The gene expression in the *A. oryzae* RIB40 genome was calculated according to RPKM and FPKM values (Supplementary Table S5). The correlation coefficient (R) between the logarithmic RPKM and FPKM values of *A. oryzae* gene expression derived from strand-specific RNA-seq data generated under DPY conditions was 0.8738 (Supplementary Figure S5). Based on the strand-specific RNA-seq data, specific TSSs were assigned for 5050 genes, accounting for 43.16% of the total 11 702 genes in the *A. oryzae* genome (Supplementary Table S6). The median length of the 5′ UTR in 5050 genes was 152 bp (Supplementary Table S6 and Supplementary Figure S6). Normality test revealed that the logarithmic RPKM values of TSS-annotated genes exhibited the Gaussian distribution (Supplementary Figure S7). However, the correlation between the expression of *A. oryzae* genes and the length of the corresponding UTRs was only 0.21 (Supplementary Figure S8). In contrast to conventional RNA-seq technology, the ssRNA-seq data could precisely decode the accurate TSS information based on RNA polarity information, which effectively reduced the artifact in the previous analysis of TSSs (7).

To investigate the protein–DNA interaction at transcript initiation sites, we extracted DNase I cleavage information from the DNase-seq data for −1 kb and +1 kb TSS flanking regions of 5050 *A. oryzae* genes (Supplementary Table S6). The pattern of the average DNase I cleavage in 5050

A. oryzae genes was organized around the TSSs in the sequential arrangement of the −1 nucleosome (the first one upstream of the TSS), the 5′-nucleosome free region (5′-NFR between the −1 nucleosome and the TSS), the TSS and the +1 nucleosome (the first one downstream of the TSS) (Figure 2A). The range of the 5′-NFR, which was depleted of nucleosomes, was 140 bp, shorter than the normal oscillatory distance (167 bp, Figure 2A). The nucleosome signal was strongly increased at the +1 nucleosome and gradually phased after the +1 or before the −1 nucleosome (Figure 2A). Furthermore, K-means clustering was performed for DNase I cleavage patterns of 5050 genes at +/− 1 kb TSS flanking regions. The results revealed that the 5050 *A. oryzae* genes could be divided into four distinct clusters (Figure 2B). For the 1567 genes in the third cluster, the nucleosomes were highly phased around the TSSs with −1 nucleosome situated around −200 bp and a −140 bp of the short 5′-NFR (Figure 2B, green color). The presence of regular undulations in the DNase I cleavage pattern was a prominent feature in these cluster genes (Figure 2B, green color). The nucleosomes phasing was fading around TSSs with −1 nucleosome situated around −280 bp and a −210 bp of the long 5′-NFR in the 1509 genes of the second cluster (Figure 2B, blue color). The DNase I cleavage density of the second cluster genes at +/− 1 kb TSS flanking regions was highest among all clusters of genes, representing high mean chromatin accessibility (Figure 2B, blue color). For the 804 genes in the first cluster, the DNase I cleavage patterns had irregular undulations in low chromatin accessibility over the upstream of TSSs and only phased in the downstream region +1 kb of the TSSs (Figure 2B, red color). The phased nucleosomes were not detected from −1 kb to +1 kb intervals around the TSSs in the 1161 genes of the fourth cluster (Figure 2B, purple color). We compared the gene expression levels among genes of the four clusters evaluated with an ANOVA, and the results showed significantly different levels of gene expression (*P*-value = 6.77E-04) (Figure 2D). This result showed that the mean (186 RPKM) and median (41 RPKM) gene expression in the second cluster were the highest among all clusters (Figure 2D), and the mean expression of the genes from the third cluster was 94 RPKM, which was lower than the expression from the other three clusters. We also observed that the 5′ UTR lengths of genes among the four clusters were significantly different, as determined by ANOVA (*P*-value = 2.16E-30), and the mean and median lengths of gene 5′ UTRs in the third cluster were shorter than those in the other three clusters (Figure 2E). GO analysis indicated that only genes in the third cluster could be overrepresented in cellular components of 'protein complex, intracellular non-membrane-bounded organelle, ribonucleoprotein complex, nucleoplasm part, mitochondrial inner membrane and endomembrane system.'

To evaluate the relationship between protein–DNA interactions at TSSs and the transcriptional activity of targeted genes, the chromatin accessibility patterns of 5050 genes at +/− 1 kb regions flanking TSSs were sorted based on the expression level in a heat map (Supplementary Figure S9). The logarithmic RPKM values of TSS-annotated genes were classified into four groups based on the mean (μ) and the standard deviation (σ) of the distribution (Supplementary Figure S7). Among all TSS-annotated genes, 86.55% were assigned to three expression groups: log RPKM > $\mu+\sigma$, μ < log RPKM < $\mu+\sigma$ and $\mu-\sigma$ < log RPKM < μ, in which 5′-NFRs were located upstream of TSSs. Particularly for the group with high expression (log RPKM > $\mu+\sigma$), the 5′-NFRs had extensions greater (Supplementary Figure S9). However, in the group with the lowest expression (logRPKM < $\mu-\sigma$), marked with a purple bar, typical periodic and oscillatory patterns of nucleosomes gradually faded. No enrichment for GO molecular functions could be observed.

To observe the spatial distribution of the digital footprints upstream of genes, we computed the number of the overrepresented footprints in the regions from the translation start codon to 500 bp upstream of TSSs. The 1186 footprint instances of the overrepresented footprints were near the TSSs of 1035 genes with annotated TSSs (Figure 2C). We observed that 626 (52.78%) footprint instances from the overrepresented footprints concentrated within the +/−100 bp region surrounding TSSs, particularly from −50 to −100 bp. The genes with short UTRs, particularly <200 bp, had 554 footprints located between the start codon and −500 bp upstream of their TSSs (Figure 2C). Among these footprints, 54.69% were located in the region −200 bp upstream of the TSS. The spatial distribution of overrepresented footprints showed a strong positional preference relative to the TSS in *A. oryzae*. DNase I cleavage patterns and the distribution of digital footprints in regions near TSSs also defined the stereotyped chromatin structural signature of transcription initiation and regulation in *A. oryzae* via *in vivo* protein–DNA interactions.

Genome-scale identification of active binding sites for the known TFs

To identify active TFBSs based on *in vivo* genome-scale protein–DNA interactions, we built the DNase I cleavage matrix of all discovered motif instances (±100 nucleotides) of a candidate binding site, which was determined by position weight matrices scanning in the intergenic regions of the *A. oryzae* genome. The MILLIPEDE model (35) was used to estimate the probability that candidate binding sites of the known TFs were bound. DHSs (defined as 150 nucleotides in length) with an FDR of 1% were identified using the hotspot algorithm (34) (Supplementary Table S7). The active binding sites of the known TFs were further identified by filtering TFBS instances using the MILLIPEDE probability > 0.5 in DHSs with an FDR of 1% (Supplementary Table S8).

We gathered the available binding motif sequences of 19 known *Aspergillus* TFs, for which the DNase I cleavage patterns of the orientation-specific motifs were derived from mapping tags to the plus and minus strands based on DNase-seq data (Figure 3A and Supplementary Figure S10). We also plotted heat maps of the typical TFs of five families across all predicted instances of each motif (Figure 3A). The DNase I cleavage patterns of the 19 known TFs showed an imbalance between sense and antisense strands within and outside of the binding-motif sequences derived from the DNA strand-specific alignment information of DNase-seq data (Figure 3A and Supplementary Fig-

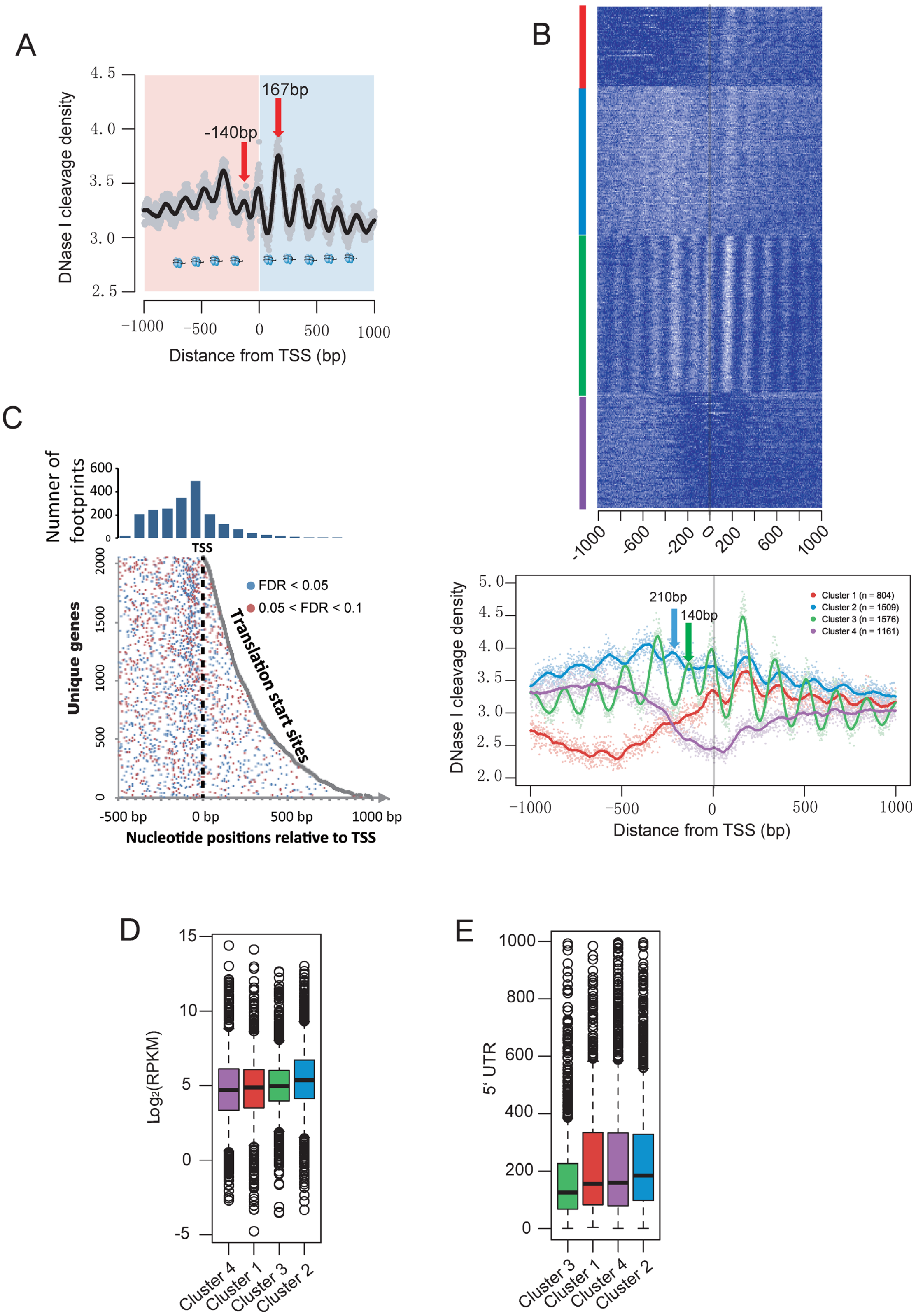
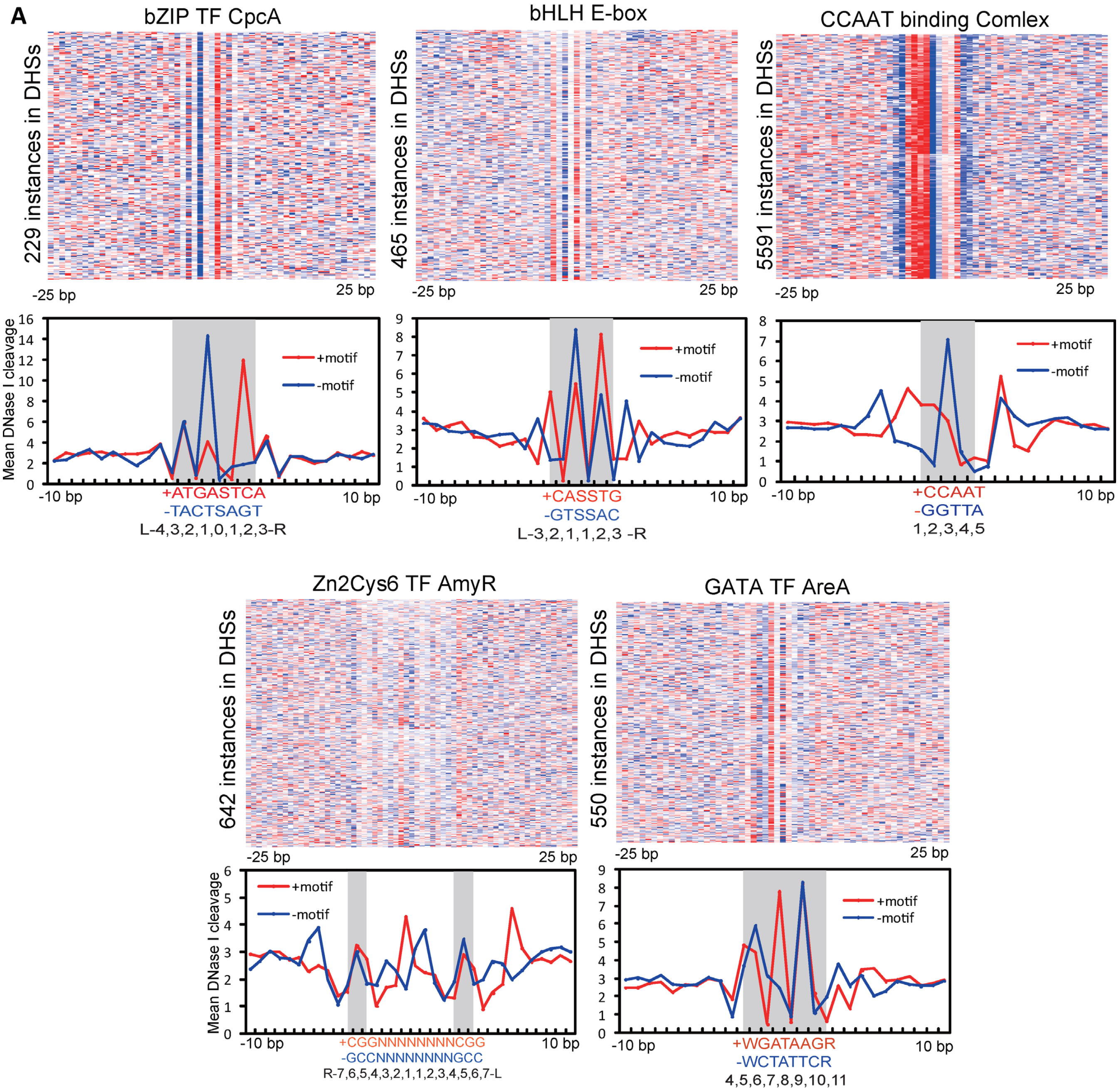


Figure 2. DNase I cleavage patterns and footprint distribution for overrepresented footprints surrounding TSSs. (**A**) Mean per-nucleotide DNase I cleavage profile from aligning the annotated TSSs of 5050 genes (+/− 1 kb regions). (**B**) Top heat map plotted for DNase I cleavage patterns of 5050 genes at +/− 1 kb TSS flanking regions by K-means clustering, which were subsequently divided into four distinct clusters, marked with red, blue, green and purple bars. The bottom mean DNase I cleavage patterns derived from four distinct clusters, where the line colors correspond to the marked colors of the heatmap. (**C**) Distribution of digital footprints (FDR < 0.05 marked with blue, and $0.05 <$ FDR < 0.1 marked with red) relative to TSSs (black vertical line) and translation start sites (gray curved line) of genes sorted according to 5′-UTR length. (**D**) Expression levels ($\log_2$RPKM) for the genes observed in each of the four clusters correlated with the targeted genes. (**E**) The length of the 5′ UTR for the genes identified in each of the four clusters correlated with the targeted genes.


A
bZIP TF CpcA
229 instances in DHSs
-25 bp
25 bp
Mean DNase I cleavage
16
14
12
10
8
6
4
2
0
+motif
-motif
-10 bp
+ATGASTCA
-TACTSAGT
L-4,3,2,1,0,1,2,3-R
10 bp
bHLH E-box
465 instances in DHSs
-25 bp
25 bp
9
8
7
6
5
4
3
2
1
0
+motif
-motif
-10 bp
+CASSTG
-GTSSAC
L-3,2,1,1,2,3 -R
10 bp
CCAAT binding Comlex
5591 instances in DHSs
-25 bp
25 bp
8
7
6
5
4
3
2
1
0
+motif
-motif
-10 bp
+CCAAT
-GGTTA
1,2,3,4,5
10 bp
Zn2Cys6 TF AmyR
642 instances in DHSs
-25 bp
25 bp
Mean DNase I cleavage
6
5
4
3
2
1
0
+motif
-motif
-10 bp
+CGGNNNNNNNNNCGG
-GCCNNNNNNNNNGCC
R-7,6,5,4,3,2,1,1,2,3,4,5,6,7-L
10 bp
GATA TF AreA
550 instances in DHSs
-25 bp
25 bp
9
8
7
6
5
4
3
2
1
0
+motif
-motif
-10 bp
+WGATAAGR
-WCTATTCR
4,5,6,7,8,9,10,11
10 bp

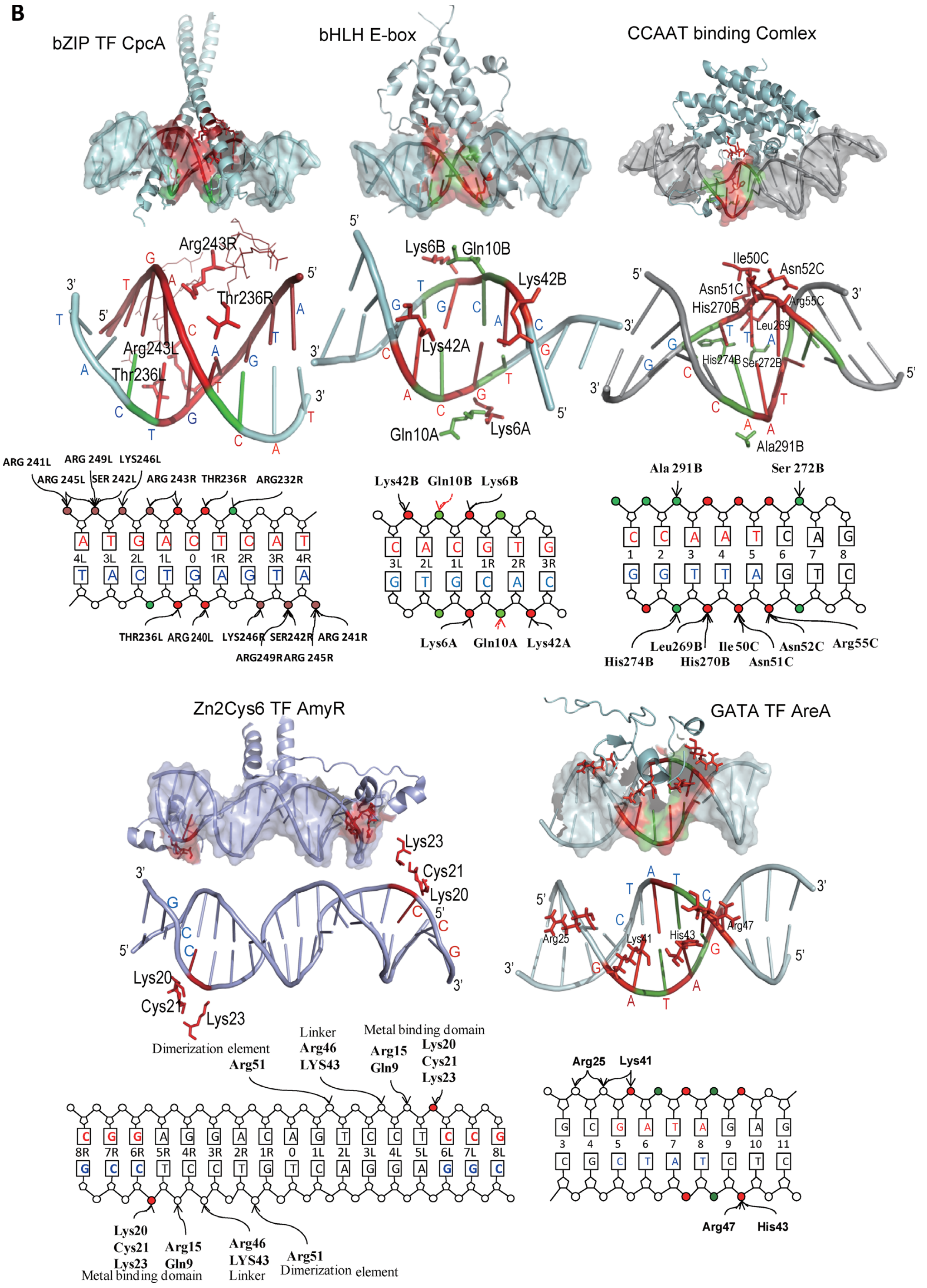


Figure 3. The DNase I cleavage patterns of five family types of TFs parallel the co-crystal structures of protein and DNA interaction. (**A**) Strand-specific DNase-seq signal for DNase I cleavage imbalance between the plus and minus motif sequences of five family types of the TFs independent of strand orientation. The upper panels show the heat maps of per-nucleotide DNase I cleavage derived from all instances of plus (red) and minus (blue) TFBS motifs within DHSs under DPY conditions ranked according to the probability of MILLIPEDE (FIMO $P < 10^{-4}$, MILLIPEDE probability > 0.5). The lower panels show the average per-nucleotide DNase I cleavage patterns of plus (red line) and minus (blue line) motif sequences of the TFs and its flanking sequences. (**B**) The co-crystal structures of the known TFs or yeast homologues bound to the DNA recognition sites are aligned with DNase I cleavage patterns relative to the motif orientation. Upper panels: the shadows of DNA backbones and surfaces of amino acids (red) of TFs that contact with the DNA backbones, the marked depression in DNase I cleavage, are indicated in red on the crystal structure. The green color represents high-level DNase I accessibility in the crystal structure. The plus and minus motif sequences are indicated as red and blue characters, respectively. Bottom panels: the labeled amino acids in the bottom graph contact with the DNA backbones. The deoxyribose sugar rings are indicated as pentagons, and the phosphates are indicated as circles. The colors represent the same indication in upper lanes. L and R represent the binding motif sequences contacted by the left and right monomers of the TF dimer, respectively.

ure S10). The DNase-seq profiles outside of the binding-motif sequences did not always exhibit a peak/trough/peak footprint shape in the aggregate plots. The cleavage patterns in the 19 known TFs' binding motifs depended on TF motif orientation-specific information and were independent of the specificity of strand orientation (Figure 3A and Supplementary Figure S10). Each TF contained a distinct DNase cleavage profile visible in aggregate plots derived from different culture conditions. The binding sites of the dimerization of two TF monomers, such as bZIP CpcA and bHLH E-box, had reversely symmetrical patterns between the plus and minus motif sequences (Figure 3A). A marked depression of DNase I cleavage was observed in the opposite 5′-phosphate groups of DNA backbones located in two monomer-overlapping sites of TFs, and high-level DNase I accessible regions occurred in the plus and minus motif sequences near two monomer-overlapping sites (Figure 3A). Similarly, other DNase I cleavage patterns of the dimerization of Zn(II)2Cys6 amyR also showed approximate motif-orientation-specific symmetry, with DNase I inaccessibility between the CGG region and the central zone (Figure 3A). However, the binding sites of monomer GATA TFs and CBCs were obviously asymmetric and imbalanced in plus and minus motif sequences (Figure 3A).

Structure features of *trans*-elements of the known TFBSs

To survey the TF structures associated with protein–DNA interactions on the genomic scale, we gathered available DNA–protein co-crystal structures of the five family TF homologues in *Aspergillus* and *S. cerevisiae*. We aligned the DNase I cleavage patterns at individual nucleotide positions onto the DNA backbones of the co-crystal model (Figure 3B). Because DNase I hydrolyses the 3′O-P bond between the phosphorus and the 3′ oxygen of the deoxyribose sugar within one strand of the recognition sequence, we analyzed the correspondence between motif-orientation patterns of DNase I cleavage and contact patterns of amino acid residues in TF monomers to 5′-phosphate groups of plus and minus motifs (Figure 3B). In the available DNA–protein co-crystal structures of the yeast TF GCN4 (PDB ID: 1YSA) (58), the hydrogen bonds binding to the 5′-phosphate groups of the plus and minus motifs are donated by amino acid residues (Thr236, Arg240, Arg-241, Arg245 and Lys246) from the left and right monomers. A marked depression of DNase I cleavage was located in the intersection region (+AST/−TSA) of two bZIP CpcA monomers, which are contacted simultaneously by amino acid residues (Thr236, Arg240) from the left and right monomers (Figure 3B). Similarly, in the case of the bHLH E-box, the yeast Pho4 co-crystal model (59) showed that Lys6A and Lys6B of the bHLH TF dimer interacted with phosphate groups of the double-strand DNA backbone of the opposite G/C of motif center bases, which also exhibited low-level DNase I accessibility (Figure 3B). In another type of dimerization of Zn(II)2Cys6 TF monomers, the co-crystal model of the *S. cerevisiae* Zn(II)2Cys6 factor Gal4 (PDB ID: 1D66) (60) shows that Lys20, Cys21 and Lys23 of each monomer are in contact with DNA backbone 5′-phosphate groups in the end bases C of the plus and minus motifs (61). The depression of DNase I cleavage was also observed between the opposite G/C ends of two CGG binding sites and central DNase I accessibility N8 elements (Figure 3B). The motif central bases of the dimerization TFs trend to be the opposite G/C, which always resisted DNase I digestion. Furthermore, in the case of CBCs, the co-crystal model of the CBC from *Aspergillus nidulans* (PDB ID: 4G92) (62) showed that most amino acid residues (Leu269, His270, Ile50, Ala51, Arg55 and Asn52) donated from the HapB and HapC subunits were hydrogen bonded to the 5′-phosphates of the DNA backbone of the minus motif (Figure 3B). The available solution structure of the fungal AreA protein–DNA complex (PDB ID: 4GAT) determined via multidimensional nuclear magnetic resonance spectroscopy (60) also showed that electrostatic interactions and contact patterns with phosphate groups lining the edge of the major groove are formed by a triad of positively charged side-chains of Arg25, Lys41, Arg47 and His43 (60) (Figure 3B). The contact patterns between the charged amino acids of the monomer TF and the complex TF were imbalanced between the plus and minus motif sequences (Figure 3B). However, certain 5′-phosphate groups in the TF binding motifs of five family TFs in contact with amino acid residues were cleaved by DNase I at high rates (Figure 3B). The other 5′-phosphate groups without the protection of amino acid residues were DNase I inaccessible (Figure 3B). These results revealed that the phosphodiester backbones in contact with the charged amino acids of TFs were not protected from DNase I cleavage.

DNA shape features of the known TFBSs

To assay the DNase I cleavage patterns of protein–DNA interactions with DNA shape features in TFBSs, we used a high-throughput DNA shape prediction approach (36) to analyse groove geometry parameters (MGW and ProT) and helical parameters (Roll and HelT) that reflect the DNA structure shapes of the TFBSs of the five TF family types. In the alignment of the cleavage patterns to the mean DNA shape features in the TFBSs of the five family types, the narrow minor grooves, defined by a groove width of 5.0 Å (compared with 5.8 Å in ideal B-DNA), were consistent with the marked DNase I inaccessible regions in TFBSs (Figure 4A). MGWs at the central S_0 base of the bZIP CpcA motif, $C_{1L}G_{1R}$ bases of bHLH E-box motif and the T5 base of CBC motif were enriched in narrow minor grooves, resulting in DNase I inaccessibility (Figures 3A and 4A). By contrast, MGWs at G_{2L}/C_{2R} bases of the bZIP CpcA motif, $C_{3L}A_{2L}/T_{2R}A_{3R}$ bases of the bHLH E-box motif, the T_7 base of the GATA AreA motif, the T_5 base of CBC motif and G_{6L}/C_{6R} bases of the Zn(II)2Cys6 amyR motif widened the minor groove and were cleaved by DNase I at higher rates (Figures 3A and 4A). PCCs were used to measure the correlation between DNase I cleavage and DNA shape features per base pair in the TFBSs of the five TF family types (Table 1). The PCC of the CBC between the MGW parameter and DNase I cleavage was 0.44, the highest in all five TF families, which only binds to the minor groove. The PCC between the roll parameter and the DNase I cleavage per base pair was the highest in four DNA shape feature parameters in all five TF families (Table 1). These data suggested that the base pair-step preferentially bends

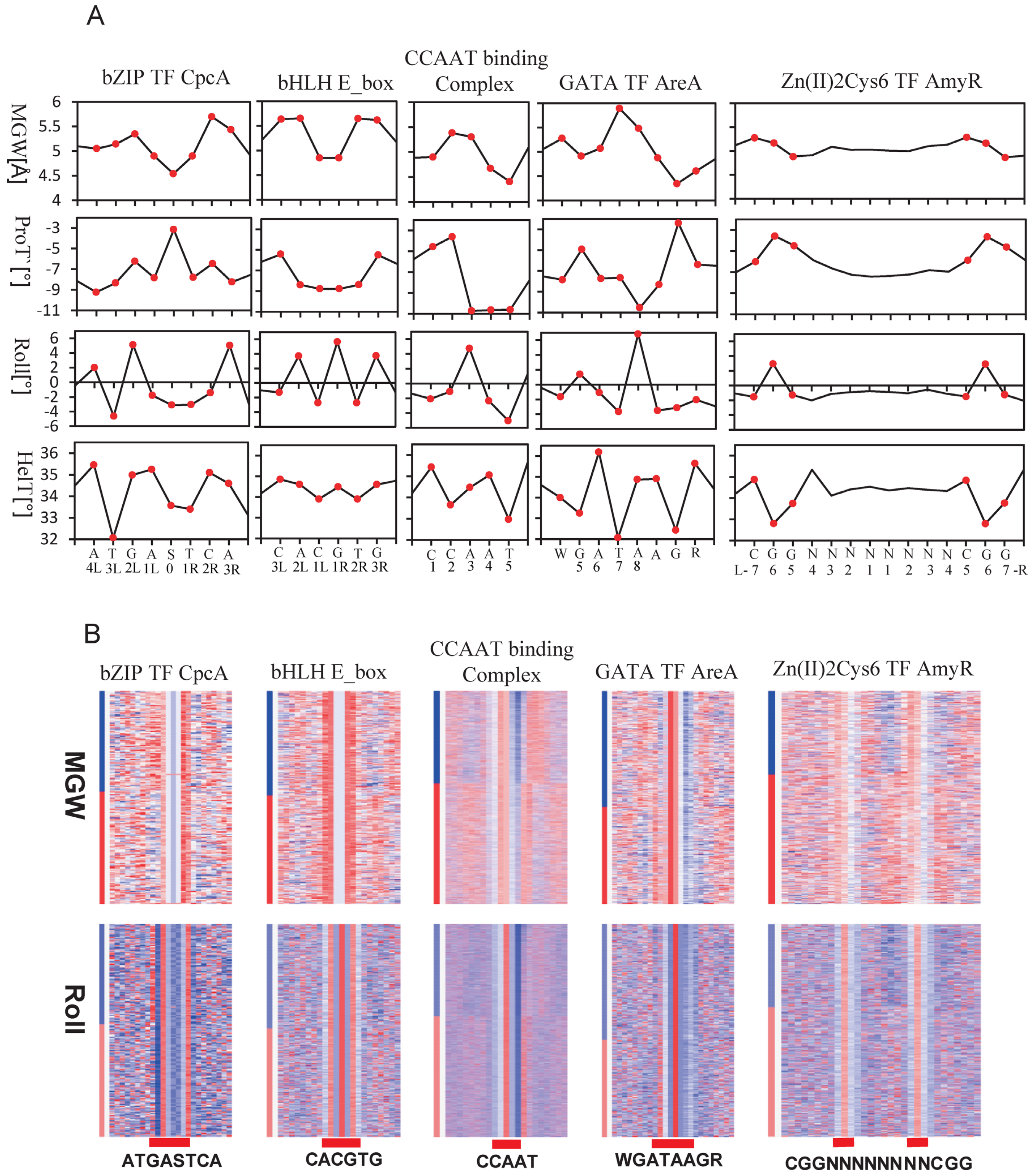


Figure 4. The DNA shape features are characteristic for the known TF motifs and its flanking sequences. (**A**) The plots display the average parameters of DNA shape features (MGW, propeller twist, Propeller twist, roll and helix twist) per base which are calculated in the active binding motif instances for the known TFs of the five family types. L and R represent the binding motif sequences contacted by the left and right monomers of the TF dimer, respectively. (**B**) Heat maps show the average MGW (upper) and Roll (down) for sequences derived from each motif instance for the known TFs of the five family types within a 10-bp flanking region. The sequences of each motif according to the average MGW were clustered and sorted into two major groups indicated with blue and red bars.

via roll was susceptible to the DNase I cleavage in TF motif sequences. MGW and roll parameters of TF motif instances in 10-bp upstream and downstream flanking regions were further clustered to analyse the variations of DNA shape features for the TFBSs of the five TF family types. Clustering two major groups according to MGW parameters showed variation at the flanking sequences of the five TF family motifs (Figure 4B). The MGWs of the flanking sequence for bZIP CpcA, bHLH E-box, CBC, GATA AreA factor and amyR tended to be wide grooves (Figure 4B). The DNase I cleavage patterns of TFs were consistent with the variation in DNA shape features, providing more information about *in vivo* protein–DNA interactions.

DISCUSSION

We surveyed the *in vivo* TF occupancy of sequence elements across the *A. oryzae* genome using DGF (20) based on DNase-seq data. The ability to resolve individual binding events depends on the specificity of DNase I digestion and the sequencing depth of the mapped DNase I cleavage (20,21). A considerable technical challenge for DGF analysis is that the DNase-seq sample must be sequenced at great depth to allow the reliable detection of locally protected regions within the DHS site (20,21,63). Ultradeep sequencing data of our *A. oryzae* DNase-seq include >100 M sequence reads, which is deep enough to identify DNase I footprints based on genome size compared with the previous sequencing depth of the yeast (20), human (21) and *Adipogenesis* (64) DNase-seq libraries for identifying digital footprints. To maximize the discovery of DNase I footprints in *A. oryzae* on a genome scale, we carry out our experiment design of sequencing a single sample to a greater depth. But in experiments design for the comparison of multiple samples, sequencing biological replicates to a lower sequencing depth can make an assessment of data variability between samples. And the lower sequencing data of replicates are supposed to be combined together to meet the requirement for identifying DNase I footprints. The 'double-hit' protocol of DNase-seq showed that DNA fragments are enriched at the regulatory element with TFBSs. The size fractionation of double-hit DNA fragments in *A. oryzae* DNase-seq was performed as previously described for yeast (20), human (21) and *Adipogenesis* (64). However, recently, it was reported that the effect of the fragment size selection of DNase-seq for the identification of TFBSs is dependent on the TF positions relative to the nucleosome (65). For TFBSs flanked by arrays of nucleosomes, it is more efficient to use shorter fragments (<100 bp) (65). For TFBSs flanked by loosely packed and unorganized nucleosomes, both long and short fragments are equally effective (65). Longer fragments tend to span the entire nucleosome (65), and might be more efficient for identifying TFBSs in nucleosomes.

DGF, as a motif-free approach (20), applies a computational algorithm to identify short regions (8–30 bp) across the intergenic regions of the genome. Hesselberth *et al.* presented a greedy algorithm to detect footprints across the yeast genome (20), and this approach was also applied to the *A. oryzae* genome. To eliminate false-positive footprint identification from aggregations of multiple mapped regions, we generate the genome mappability data of the 36-bp reads. Longer reads (>50 bp), which could be mapped into repetitive regions, result in lessening masked multiple mapped and repeated regions in genome (66). Furthermore, the yeast and *A. oryzae* genome exhibited a more compact pattern compared with the human genome, and the maximum lengths of the intergenic regions in yeast and *A. oryzae* genomes were 15 kb and 22 kb, respectively. However, this footprinting methodology does not scale well for large genomes (67) and is adapted to the human genome in the ENCODE project using metric values to calculate the ratio of DNase I cleavages within a binding site to those outside (the Footprint Occupancy Score) (21). We used a DGF approach (20) to identify two sets of DNase I footprints across the *A. oryzae* genome under DPY and UPR culture conditions. The overlapped footprints between DPY and UPR conditions regulate the targeted genes enriched in functions with protein translation and secretory pathway, consistent with high-level production of proteins under nutrient-rich culture (DPY condition) and induction mechanism of ER stress under UPR condition.

The advantage of the motif-free approach is that it does not require prior knowledge of the TF binding motifs. *De novo* motifs were further identified and overrepresented from our two sets of *A. oryzae* genomic footprints using MEME under a relatively stringent threshold, corresponding to the known *Aspergillus* TFs, including SltA, CpcA and the E-box of bHLH factors. However, *de novo* footprint-derived motifs are more complex than previously predicted (20) and rely on the accuracy of the footprint algorithm under high-depth DNase I sequencing (68). In yeast, genomic footprints are densely populated with previously recognized TFBSs by enriching ChIP-defined motifs (20). The application of genomic footprinting in combination with ChIP should provide rich information to recover a broad range of TF binding motifs. However, the information concerning the genome-scale localization of TFBSs using ChIP technology in *A. oryzae* remains scarce.

Based on *de novo* motifs recovered from genomic footprints, we observed that the SltA TF-binding motifs contained two co-localization binding patterns including at least two SltA binding sites (5′-AGGCA-3′ or 5′-TGCCT-3′) to regulate genes with different functions. The functional analysis of genes under the control of two types of SltA-binding patterns suggested that TF SltA could regulate the function of ribosomes to release and balance protein translation under ER stress. The SltA deletion mutant of *A. nidulans* has been reported to reduce colony growth and conidial production (47), consistent with the functional analysis of genes under the control of SltA binding patterns of DPY_motif 7 or UPR_motif 5.

A. oryzae nucleosomes are organized around the TSSs in the sequential arrangement of the −1 nucleosome, 5′-NFR, TSS and +1 nucleosome. The 167-bp periodicity in the DNase I cleavage pattern is consistent with the sum of the length of a 147-bp nucleosomal unit with a 20-bp linker region (69), representing phased nucleosome occupation on both sides of the TSS. The strong nucleosome signal at the +1 nucleosome indicates a stable nucleosome at the transcription initiation site as a barrier for a subsequent nucleosome (70). The ∼140-bp 5′-NFR is depleted of nucleosomes because it is shorter than the single-nucleosome-occupancy

Table 1. Correlation between DNase I cleavage and DNA shape features

DNase I cleavage	PCCs			
	MGW	ProT	Roll	HelT
bZIP TF cpcA	0.12**	– 0.10**	– 0.29**	0.03
bHLH E-box	– 0.08**	– 0.19**	– 0.71**	– 0.44**
CCAAT-Box Binding Factor	0.42**	0.03**	0.54**	0.07**
GATA TF areA	0.19**	– 0.03*	– 0.26**	– 0.21**
Zn(II)2Cys6 TF amyR	0.05**	0.17**	0.17**	– 0.32**

Pearson correlations were determined from the R matrix of correlation coefficients calculated for the DNase I cleavage matrix and DNA shape feature matrix of TFBSs. The Mantel test was used to measure the correlation between two matrices typically containing measures of distance. The matrix of *P*-values were tested for hypotheses of no correlation. *P*-values for the correlations were zero with significant power, except for correlations of DNase I cleavage with HelT of cpcA, with values of greater than 0.05.
**$P < 0.01$; *$P < 0.05$.

size. A ∼140-bp 5′-NFR (71) was also located at ∼95% of all genes in the *S. cerevisiae* Genome Database (as of May 2007) (72). GO analysis also confirmed that the DNase I cleavage patterns of the genes in the cluster of highly phased nucleosomes around the TSSs and 140-bp 5′-NFRs might be associated with the constitutive transcription of certain housekeeping genes (73). Furthermore, a 210-bp region of long 5′-NFRs observed in the genes of the highest-chromatin accessibility cluster might lead to high transcription activity through the extension of chromatin disruption upstream of the TSS over which TF binding occurs. The result of the chromatin accessibility patterns around the TSSs of 5050 genes sorted based on expression level also confirms that the 5′-NFRs of the group under high expression levels have extension greater than the 5′-NFRs of the group under low expression levels. The transcriptional variability is correlated with various nucleosome occupancy patterns surrounding the TSSs (74). Furthermore, the spatial distributions of the overrepresented footprints display a strong positional preference relative to the TSS. The data showing DNase I cleavage patterns and the distribution of digital footprints near TSSs can be used to define the stereotyped chromatin-structure signatures for transcription initiation and regulation in *A. oryzae*.

The DNase I cleavage profiles of TF motifs reveal imbalances of cleavages between sense and antisense strands and motif-orientation-specific patterns between plus and minus TF binding motifs. The DNase-seq profiles of TFs were not always present as sequence motif-centered footprints, likely reflecting weaker or indirect TF/DNA binding and DNase-seq profiles as a combination of background and footprint profiles. The DNA–protein co-crystal models can provide information about the contact patterns of amino acid residues in TFBS motifs. The binding motif ATGASTCAT of yeast GCN4 in the bZIP TF family is completely identical to that of *Aspergillus* TF CpcA (ATGASTCA). The core binding sequences of the bHLH E-box Pho4 and the Zn(II)2Cys6 factor Gal4 in yeast are CACGTG and CGGs binding motifs, respectively, consistent with those of TF homologues in *Aspergillus*. The available DNA–protein co-crystal structures of TF homologues in *S. cerevisiae* can be used to analyse DNA–protein interactions in *Aspergillus*. DNase I hydrolyses the 3′ O-P bond of phosphates of deoxyribose sugars within one of the strands. However, the binding of charged amino acids of TFs to phosphate groups on the backbone is not sequence specific. Some of the phosphodiester backbones in the TF binding-motif were cleaved by DNase at a high rate, where charged amino acids bind to the phosphate group in the backbone of the TF co-crystal model. The DNA cleavage specificity in the TF binding motif could not provide direct information concerning the 'structural motifs' of the DNA–protein interface, which does not support the previous report of the 'structural motifs' in yeast (20) and human (21).

The DNA structure shapes of the TFBSs instances of the five TF family types in *A. oryzae* are consistent with those of TFBS homologues in *S. cerevisiae* (75). In the alignment of cleavage patterns to DNA shape features, narrow and wide minor grooves (MGW < 5.0 Å) in the TFBSs are consistent with marked DNase I inaccessibility and accessibility regions, respectively. DNase I is used as a structural probe for DNA shape features and steric hindrance via interactions with the minor groove of DNA (76). The rate of DNase I cleavage is significantly associated with the width of the minor groove (77). The phosphodiester backbones of dinucleotides with the widened minor groove are cleaved by DNase I at a high rate (78), whereas a narrow minor groove resists DNase I digestion (77). According to the three-dimensional structures of protein–DNA complexes, the binding of arginine residues to a narrow minor groove is a widely used mode for protein–DNA recognition (79). A narrow MGW enhances the negative electrostatic potential in the minor groove, which can positively attract charged amino acids (77). Sequence dependencies on the DNase I cleavage rate might reflect differences in the DNA shape, which plays an important role in protein–DNA recognition. DNA shape features in the flanking sequences of all TF motif instances vary, likely reflecting different binding affinities in all TF motif instances. For the bHLH TFs Cbf1 and Tye7 in yeast, the flanking sequences of E-box binding sites contribute to the specificity by influencing the three-dimensional structure of DNA-binding sites (80). The MGWs of the flanking sequences tend to be widened compared with the sites of low-binding affinity (80). Furthermore, Roll describes the opening of a dinucleotide to either the minor or major groove. TF motifs are preferentially bent via roll and are susceptible to DNase I cleavage. CpG methylation leads to an increased roll angle at the CpG step, which regulates protein–DNA interaction strength via changes in the DNA shape (77). Moreover, the dimerization and monomer TFs, such as bZIP, CpcA, bHLH E-box, Zn(II)2Cys6 AmyR and GATA AreA, binding to the major groove, and TF complexes, such as the binding of the CAAT-binding complex, binding to the minor groove, can distort the conformation of the backbone, resulting in the variation in the DNase I cleavage rate in the TF binding motif.

We coupled the DNase I digestion of intact nuclei with massively parallel sequencing to survey the landscape of *in vivo* protein–DNA interaction across the *A. oryzae* genome. Motif-free and motif-based approaches were used to identify motifs from *A. oryzae* DNase-seq data via ultra-deep sequencing. DNase I cleavage patterns and the distribution of digital footprints in regions near TSSs define the stereotyped chromatin structural signature of transcription initiation and regulation of the *in vivo* protein–DNA interaction. These data can improve our knowledge of genome-wide TF binding events in *aspergilli,* particularly when no available genome-wide TF-binding data using ChIP-seq and ChIP-chip can be obtained in *A. oryzae.*

ACCESSION NUMBERS

NCBI-SRA SRX607943, SRX610905 and SRX610909.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

863 (Hi-tech Research and Development Program of China) program [2012AA022108]; National Science and Technology Foundation during the 'Twelfth Five-Year Plan' period for the rural development in China [2011AA100905]; Guangdong Provincial Department of Science and Technology Research Project [2012B010900028, 2012A080800013]; National Natural Science Foundation of China [31100026]. Funding for open access charge: The 863 (Hi-tech Research and Development Program of China) program under contract no. 2012AA022108.

Conflict of interest statement. None declared.

REFERENCES

1. Kobayashi,T., Abe,K., Asai,K., Gomi,K., Juvvadi,P.R., Kato,M., Kitamoto,K., Takeuchi,M. and Machida,M. (2007) Genomics of *Aspergillus oryzae*. *Biosci. Biotechnol. Biochem.*, **71**, 646–670.
2. Machida,M., Asai,K., Sano,M., Tanaka,T., Kumagai,T., Terai,G., Kusumoto,K., Arima,T., Akita,O., Kashiwagi,Y. *et al.* (2005) Genome sequencing and analysis of *Aspergillus oryzae*. *Nature*, **438**, 1157–1161.
3. Vongsangnak,W., Olsen,P., Hansen,K., Krogsgaard,S. and Nielsen,J. (2008) Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*. *BMC Genomics*, **9**, 245.
4. Akao,T., Sano,M., Yamada,O., Akeno,T., Fujii,K., Goto,K., Ohashi-Kunihiro,S., Takase,K., Yasukawa-Watanabe,M., Yamaguchi,K. *et al.* (2007) Analysis of expressed sequence tags from the fungus *Aspergillus oryzae* cultured under different conditions. *DNA Res.*, **14**, 47–57.
5. Tamano,K., Sano,M., Yamane,N., Terabayashi,Y., Toda,T., Sunagawa,M., Koike,H., Hatamoto,O., Umitsuki,G., Takahashi,T. *et al.* (2008) Transcriptional regulation of genes on the non-syntenic blocks of *Aspergillus oryzae* and its functional relationship to solid-state cultivation. *Fungal Genet. Biol.*, **45**, 139–151.
6. Andersen,M.R., Vongsangnak,W., Panagiotou,G., Salazar,M.P., Lehmann,L. and Nielsen,J. (2008) A trispecies *Aspergillus* microarray: comparative transcriptomics of three *Aspergillus* species. *Proc. Natl Acad. Sci. U.S.A.*, **105**, 4387–4392.
7. Wang,B., Guo,G., Wang,C., Lin,Y., Wang,X., Zhao,M., Guo,Y., He,M., Zhang,Y. and Pan,L. (2010) Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing. *Nucleic Acids Res.*, **38**, 5075–5087.
8. Oda,K., Kakizono,D., Yamada,O., Iefuji,H., Akita,O. and Iwashita,K. (2006) Proteomic analysis of extracellular proteins from *Aspergillus oryzae* grown under submerged and solid-state culture conditions. *Appl. Environ. Microbiol.*, **72**, 3448–3457.
9. Galagan,J.E., Calvo,S.E., Cuomo,C., Ma,L.J., Wortman,J.R., Batzoglou,S., Lee,S.I., Basturkmen,M., Spevak,C.C., Clutterbuck,J. *et al.* (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature*, **438**, 1105–1115.
10. Das,M.K. and Dai,H.K. (2007) A survey of DNA motif finding algorithms. *BMC Bioinformatics*, **8**(Suppl. 7), S21.
11. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
12. Won,K.J., Ren,B. and Wang,W. (2010) Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol.*, **11**, R7.
13. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
14. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, **316**, 1497–1502.
15. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
16. Gross,D.S. and Garrard,W.T. (1988) Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.*, **57**, 159–197.
17. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
18. Sabo,P.J., Kuehn,M.S., Thurman,R., Johnson,B.E., Johnson,E.M., Cao,H., Yu,M., Rosenzweig,E., Goldy,J., Haydock,A. *et al.* (2006) Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. *Nat. Methods*, **3**, 511–518.
19. Chen,X., Hoffman,M.M., Bilmes,J.A., Hesselberth,J.R. and Noble,W.S. (2010) A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics*, **26**, i334–i342.
20. Hesselberth,J.R., Chen,X., Zhang,Z., Sabo,P.J., Sandstrom,R., Reynolds,A.P., Thurman,R.E., Neph,S., Kuehn,M.S., Noble,W.S. *et al.* (2009) Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.
21. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
22. Song,L., Zhang,Z., Grasfeder,L.L., Boyle,A.P., Giresi,P.G., Lee,B.K., Sheffield,N.C., Graf,S., Huss,M., Keefe,D. *et al.* (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.*, **21**, 1757–1767.
23. Pique-Regi,R., Degner,J.F., Pai,A.A., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.
24. Boyle,A.P., Song,L., Lee,B.K., London,D., Keefe,D., Birney,E., Iyer,V.R., Crawford,G.E. and Furey,T.S. (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.*, **21**, 456–464.
25. Zhang,W., Zhang,T., Wu,Y. and Jiang,J. (2012) Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in *Arabidopsis*. *Plant Cell*, **24**, 2719–2731.
26. Chumsakul,O., Nakamura,K., Kurata,T., Sakamoto,T., Hobman,J.L., Ogasawara,N., Oshima,T. and Ishikawa,S. (2013) High-resolution mapping of *in vivo* genomic transcription factor

binding sites using *in situ* DNase I footprinting and ChIP-seq. *DNA Res.*, **20**, 325–338.

27. Gonzalez,R. and Scazzocchio,C. (1997) A rapid method for chromatin structure analysis in the filamentous fungus *Aspergillus nidulans*. *Nucleic Acids Res.*, **25**, 3955–3956.
28. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
29. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.
30. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
31. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
32. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
33. Bindea,G., Mlecnik,B., Hackl,H., Charoentong,P., Tosolini,M., Kirilovsky,A., Fridman,W.H., Pages,F., Trajanoski,Z. and Galon,J. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**, 1091–1093.
34. John,S., Sabo,P.J., Thurman,R.E., Sung,M.H., Biddie,S.C., Johnson,T.A., Hager,G.L. and Stamatoyannopoulos,J.A. (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genet.*, **43**, 264–268.
35. Luo,K. and Hartemink,A.J. (2013) Using DNase digestion data to accurately identify transcription factor binding sites. *Pac. Symp. Biocomput.*, 80–91.
36. Zhou,T., Yang,L., Lu,Y., Dror,I., Dantas Machado,A.C., Ghane,T., Di Felice,R. and Rohs,R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.*, **41**, W56–W62.
37. Parkhomchuk,D., Borodina,T., Amstislavskiy,V., Banaru,M., Hallen,L., Krobitsch,S., Lehrach,H. and Soldatov,A. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.*, **37** ,e123.
38. Li,R., Yu,C., Li,Y., Lam,T.W., Yiu,S.M., Kristiansen,K. and Wang,J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.
39. Mortazavi,A., Williams,B.A., Mccue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
40. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
41. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
42. Saldanha,A.J. (2004) Java Treeview–extensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248.
43. McArthur,M., Gerum,S. and Stamatoyannopoulos,G. (2001) Quantification of DNaseI-sensitivity by real-time PCR: quantitative analysis of DNaseI-hypersensitivity of the mouse beta-globin LCR. *J. Mol. Biol.*, **313**, 27–34.
44. Hagiwara,D., Kondo,A., Fujioka,T. and Abe,K. (2008) Functional analysis of C2H2 zinc finger transcription factor CrzA involved in calcium signaling in *Aspergillus nidulans*. *Curr. Genet.*, **54**, 325–338.
45. Spielvogel,A., Findon,H., Arst,H.N., Araujo-Bazan,L., Hernandez-Ortiz,P., Stahl,U., Meyer,V. and Espeso,E.A. (2008) Two zinc finger transcription factors, CrzA and SltA, are involved in cation homoeostasis and detoxification in *Aspergillus nidulans*. *Biochem. J.*, **414**, 419–429.
46. Calcagno-Pizarelli,A.M., Hervas-Aguilar,A., Galindo,A., Abenza,J.F., Penalva,M.A. and Arst,H.N. Jr (2011) Rescue of *Aspergillus nidulans* severely debilitating null mutations in ESCRT-0, I, II and III genes by inactivation of a salt-tolerance pathway allows examination of ESCRT gene roles in pH signalling. *J. Cell Sci.*, **124**, 4064–4076.
47. Shantappa,S., Dhingra,S., Hernandez-Ortiz,P., Espeso,E.A. and Calvo,A.M. (2013) Role of the zinc finger transcription factor SltA in morphogenesis and sterigmatocystin biosynthesis in the fungus *Aspergillus nidulans*. *PLoS One*, **8**, e68492.
48. Pavletich,N.P. and Pabo,C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, **252**, 809–817.
49. Brayer,K.J. and Segal,D.J. (2008) Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell Biochem. Biophys.*, **50**, 111–131.
50. Kuras,L., Barbey,R. and Thomas,D. (1997) Assembly of a bZIP-bHLH transcription activation complex: formation of the yeast Cbf1-Met4-Met28 complex is regulated through Met28 stimulation of Cbf1 DNA binding. *EMBO J.*, **16**, 2441–2451.
51. Zhou,X. and O'Shea,E.K. (2011) Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Mol. Cell*, **42**, 826–836.
52. Fisher,F. and Goding,C.R. (1992) Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. *EMBO J.*, **11**, 4103–4109.
53. Perina,D., Korolija,M., Roller,M., Harcet,M., Jelicic,B., Mikoc,A. and Cetkovic,H. (2011) Over-represented localized sequence motifs in ribosomal protein gene promoters of basal metazoans. *Genomics*, **98**, 56–63.
54. Klein,J., Saedler,H. and Huijser,P. (1996) A new family of DNA binding proteins includes putative transcriptional regulators of the *Antirrhinum majus* floral meristem identity gene SQUAMOSA. *Mol. Gen. Genet.*, **250**, 7–16.
55. Cardon,G., Hohmann,S., Klein,J., Nettesheim,K., Saedler,H. and Huijser,P. (1999) Molecular characterisation of the *Arabidopsis* SBP-box genes. *Gene*, **237**, 91–104.
56. Kropat,J., Tottey,S., Birkenbihl,R.P., Depege,N., Huijser,P. and Merchant,S. (2005) A regulator of nutritional copper signaling in *Chlamydomonas* is an SBP domain protein that recognizes the GTAC core of copper response element. *Proc. Natl Acad. Sci. U.S.A.*, **102**, 18730–18735.
57. Yamasaki,H., Hayashi,M., Fukazawa,M., Kobayashi,Y. and Shikanai,T. (2009) SQUAMOSA promoter binding protein-like7 is a central regulator for copper homeostasis in *Arabidopsis*. *Plant Cell*, **21**, 347–361.
58. Ellenberger,T.E., Brandl,C.J., Struhl,K. and Harrison,S.C. (1992) The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. *Cell*, **71**, 1223–1237.
59. Shimizu,T., Toumoto,A., Ihara,K., Shimizu,M., Kyogoku,Y., Ogawa,N., Oshima,Y. and Hakoshima,T. (1997) Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *EMBO J.*, **16**, 4689–4697.
60. Starich,M.R., Wikstrom,M., Arst,H.N. Jr, Clore,G.M. and Gronenborn,A.M. (1998) The solution structure of a fungal AREA protein-DNA complex: an alternative binding mode for the basic carboxyl tail of GATA factors. *J. Mol. Biol.*, **277**, 605–620.
61. Todd,R.B. and Andrianopoulos,A. (1997) Evolution of a fungal regulatory gene family: the Zn(II)2Cys6 binuclear cluster DNA binding motif. *Fungal Genet. Biol.*, **21**, 388–405.
62. Huber,E.M., Scharf,D.H., Hortschansky,P., Groll,M. and Brakhage,A.A. (2012) DNA minor groove sensing and widening by the CCAAT-binding complex. *Structure*, **20**, 1757–1768.
63. Yardimci,G.G., Frank,C.L., Crawford,G.E. and Ohler,U. (2014) Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.*, **42**, 11865–11878.
64. Siersbaek,R., Baek,S., Rabiee,A., Nielsen,R., Traynor,S., Clark,N., Sandelin,A., Jensen,O.N., Sung,M.H., Hager,G.L. *et al.* (2014) Molecular architecture of transcription factor hotspots in early adipogenesis. *Cell Rep.*, **7**, 1434–1442.
65. He,H.H., Meyer,C.A., Hu,S.S., Chen,M.W., Zang,C., Liu,Y., Rao,P.K., Fei,T., Xu,H., Long,H. *et al.* (2014) Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods*, **11**, 73–78.
66. Derrien,T., Estelle,J., Marco Sola,S., Knowles,D.G., Raineri,E., Guigo,R. and Ribeca,P. (2012) Fast computation and applications of genome mappability. *PLoS One*, **7**, e30377.

67. Baek,S., Sung,M.H. and Hager,G.L. (2012) Quantitative analysis of genome-wide chromatin remodeling. *Methods Mol. Biol.*, **833**, 433–441.
68. Barozzi,I., Bora,P. and Morelli,M.J. (2014) Comparative evaluation of DNase-seq footprint identification strategies. *Front. Genet.*, **5**, 278.
69. Lee,W., Tillo,D., Bray,N., Morse,R.H., Davis,R.W., Hughes,T.R. and Nislow,C. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.
70. Albert,I., Mavrich,T.N., Tomsho,L.P., Qi,J., Zanton,S.J., Schuster,S.C. and Pugh,B.F. (2007) Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature*, **446**, 572–576.
71. Yuan,G.C., Liu,Y.J., Dion,M.F., Slack,M.D., Wu,L.F., Altschuler,S.J. and Rando,O.J. (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, **309**, 626–630.
72. Mavrich,T.N., Ioshikhes,I.P., Venters,B.J., Jiang,C., Tomsho,L.P., Qi,J., Schuster,S.C., Albert,I. and Pugh,B.F. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, **18**, 1073–1083.
73. Jiang,C. and Pugh,B.F. (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.*, **10**, 161–172.
74. Tirosh,I. and Barkai,N. (2008) Two strategies for gene regulation by promoter nucleosomes. *Genome Res.*, **18**, 1084–1091.
75. Yang,L., Zhou,T., Dror,I., Mathelier,A., Wasserman,W.W., Gordan,R. and Rohs,R. (2014) TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.*, **42**, D148–D155.
76. Weston,S.A., Lahm,A. and Suck,D. (1992) X-ray structure of the DNase I-d(GGTATACC)2 complex at 2.3 A resolution. *J. Mol. Biol.*, **226**, 1237–1256.
77. Lazarovici,A., Zhou,T., Shafer,A., Dantas Machado,A.C., Riley,T.R., Sandstrom,R., Sabo,P.J., Lu,Y., Rohs,R., Stamatoyannopoulos,J.A. *et al.* (2013) Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl Acad. Sci. U.S.A.*, **110**, 6376–6381.
78. Brukner,I., Jurukovski,V. and Savic,A. (1990) Sequence-dependent structural variations of DNA revealed by DNase I. *Nucleic Acids Res.*, **18**, 891–894.
79. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–1253.
80. Gordan,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.*, **3**, 1093–1104.