# **METHODS**

### **Preparation of histone octamer**

A vector encoding *Xenopus laevis* H2A with a N-terminal streptavidin binding peptide (SBP) tag was cloned using 'Round-the-horn site-directed mutagenesis. *X. laevis* histones were expressed and purified as described previously60. Inclusion bodies were resuspended by using a Dounce tissue grinder (Sigma-Aldrich). Purified histones were aliquoted, flash-frozen, lyophilized, and stored at -80 °C prior to use. The lyophilized histones were resuspended in unfolding buffer (7 M guanidine hydrochloride and 10 mM DTT in 20 mM Tris-Cl, pH 7.5) to a concentration of 1.5 mg/ml. N-terminally tagged H2A, H2B, H3 and H4 were then combined at a molar ratio of 1.2:1.2:1:1. The sample was incubated on ice for 30 minutes before it was dialyzed against three times 600 ml refolding buffer (2 M NaCl, 1 mM EDTA, and 5 mM β-mercaptoethanol in 10 mM Tris-Cl, pH 7.5). The sample was recovered after dialysis and applied to a GE S200 16/600 pg size exclusion column (GE Healthcare, Little Chalfont, United Kingdom). Peak fractions were analyzed by SDS-PAGE. Fractions containing the octamer were pooled and concentrated. Both the histone expression and octamer formation have been quality-controlled (**Extended Data Fig. 1a**, **b**).

### **Clones, protein expression and purification for TFs**

We essentially followed Yin *et al.*18 to clone, express and purify the proteins from *E. coli* cells. Gateway recipient vectors having a pETG20A backbone were employed in the bacterial protein expression. The vectors incorporate a Thioredoxin-6×His tag in the N-terminal of the insertion cassette, and a SBP or a 3×FLAG tag in the C-terminal. According to our previous work18, the identity of affinity tag only have minor effect on SELEX results. Insertions for these expression vectors were derived either from PCR clones or from gene synthesis. The sequences and domains for all TFs are listed in **Supplementary Table 1**. The non-full-length constructs contain extended DNA binding domains (eDBDs), with a design rationale reported previously19.

### **NCAP-SELEX ligand design**

Approximately 70% of eukaryotic DNA is packaged into nucleosomes, separated from each other by free DNA linker sequences of 10–80 bp61-63. It is well established that TFs compete with nucleosomes for available genomic DNA sequences, and that this competition has a major influence on gene expression. Although the DNA binding specificities of many TFs and the nucleosome itself are relatively well-characterized18,19,40,64-71, there is little information on how the nucleosome affects TF binding. In this work, we developed a new method, nucleosome CAP-SELEX (NCAP-SELEX), to study TF-nucleosome interactions.

The NCAP-SELEX protocol has two steps of selection, respectively for ligands bound by the nucleosome and by individual TFs. To facilitate PCR amplification and sequencing, we designed DNA ligands with flanking adaptors based on Illumina’s Truseq library (**Supplementary Table 2**). The adapter lengths were 24 bp at the left side and 22 bp at the right side. The total lengths of the ligands are 200 bp (lig200), and 147 bp (lig147), and they contain 101 bp and 154 bp random sequences, respectively; both are flanked with the Illumina adapters. The lig200 can accommodate only one nucleosome but is longer than required (147 bp), and therefore it harbors both nucleosomal DNA and free DNA, which allows a direct comparison between the two. Although the nucleosome is delocalized on lig200 (**Extended Data Fig. 3a**), lig200 allows a direct comparison between nucleosomal DNA and free DNA. Thus, lig200 is ideal for studying TF’s nucleosome preference, and is also preferable for investigating the low-resolution or positionally independent TF binding modes. In contrast, lig147 harbors only nucleosomal DNA, but is advantageous for studying high-resolution TF binding modes, because nucleosome is uniquely positioned on it. Since lig147 is fully occupied by the nucleosome, it allows a TF to compare only between nucleosomal DNA sites. Such comparison is impossible in the presence of free DNA, because most TFs bind stronger to free DNA. The single-stranded oligos of lig147 and lig200 were purchased from IDT (Ultramer DNA oligos). A PCR reaction with primers binding to the adapters (**Supplementary Table 2**, PCR\_primers) was used to obtain double-stranded DNA from the synthetic oligos, and was also used to amplify the libraries between SELEX cycles. For sequencing, the ligands were amplified with the multiplexing primers (**Supplementary Table 2**, PE\_PCR\_primers). In this study, positions on the ligands were indexed with increasing numbers from left to right according to the sequences in **Supplementary Table 2**.

The amount of input DNA used in SELEX contains almost all possible 20-bp consecutive or gapped subsequences. Such complexity well suffices the specificity studies of human TFs, whose binding is associated with ~15 bits of information on average65. For nucleosome, the complexity allows the study of optimal sequences around each histone-DNA contact, but might not capture all the specificities as the nucleosome-favored or disfavored sequences may include cooperation spanning a large length of DNA, e.g., the phased successive bending or the rigidity of a long segment.

### **NCAP-SELEX selection**

In NCAP-SELEX, first, recombinant histone octamers containing streptavidin-binding protein tagged H2A proteins were loaded onto the DNA ligands (**Extended Data Fig. 1c, d**) in 384-well microplates, by decreasing the salt concentration in a stepwise fashion. Specifically, 100–200 ng double-stranded DNA ligand and 25 ng histone octamer were mixed in 2 M KCl solution (5 µl) and incubated for 30 min. The mixture was then diluted stepwise60, with a dilution buffer (TE buffer supplemented with 1 mM tris(2-carboxyethyl)phosphine (TCEP) and a cocktail of protease inhibitors (05892970001, Roche)). For the stepwise dilution, the volumes of the subsequent buffer additions were 5 µl, 2.5 µl, 2.5 µl, 2.5 µl, 2.5 µl, 30 µl, 25 µl, each followed by a 1 h incubation. Microplates were used also for nucleosome reconstitution because in cycle 2-5 each well already enriched signals for different TFs. After reconstitution, the reconstituted nucleosome was incubated for 30 min with 1.2 µl streptavidin-coated magnetic sepharose beads (28-9857-99, GE Healthcare; pre-blocked with the blocking buffer containing 25mM Tris, 0.5% BSA, 0.1% tween 20, 0.02% NaN3), and shaken at 1900 rpm with a microplate shaker (13500-890, VWR). The beads were then washed 15 times with a microplate washer (HydrospeedTM, Tecan). The nucleosome was eluted with 10 mM biotin (Sigma) in the incubation buffer (50–140 mM KCl, 5 mM NaCl, 2 mM MgSO4, 3 μM ZnSO4, 100 μM EGTA, 1 mM K2HPO4, in 20 mM HEPES, pH 7.0), and incubated with 10–200 ng purified hexahistidine-tagged TFs for 20 min. The TF-bound species were pulled down with 1.8 µl of nickel magnetic sepharose beads (28-9799-17, GE Healthcare; pre-blocked with the blocking buffer) and washed 15 times. The bead suspension were used for PCR as previously described by Jolma *et al*.19 This process was repeated for a total of five times. Ligands were amplified and sequenced after each cycle as well as the input. When incubating the nucleosome with TF, we initially used 140 mM of monovalent cations. The physiological salt concentration resulted in relatively high nonspecific adsorption of the nucleosome to the sepharose beads. To improve the assay, lower salt concentrations (50 mM to 75 mM) were used in subsequent experiments. Most effects were robust to the changes in the salt concentration; discussion in the main text is limited to observations that were detected under multiple salt concentrations. Moreover, in SELEX, each cycle is essentially an independent replicate of the experiment. The reported effects all show enrichment across multiple SELEX cycles.

 As a control, HT-SELEX (i.e., SELEX using nucleosome-free DNA) with lig147 or with lig200 was performed according to the previous protocol18,65 with the same purified TF proteins as those used in NCAP-SELEX. In addition, to verify that TFs have little preference to the edge of a nucleosome, we performed NCAP-SELEX also on a 293-bp ligand that positions nucleosome at the center and contains 70-bp random sequences at both ends (Lig70Nlinker, **Extended Data Fig. 3c**, sequence in **Supplementary Table 2**). We also confirmed that TF concentration has a minor effect on the results (**Extended Data Fig. 3d**). Moreover, we also examined nucleosome’s sequence preference as a control, where the reconstituted nucleosome was purified and the associated ligands were amplified prior to the next cycle. To enrich the nucleosome-disfavored sequences, an EMSA followed by gel-extraction of the free-DNA band was performed after the nucleosome reconstitution. Comparing the nucleosome-favored and disfavored sequences suggest nucleosome prefers subsequences with moderate G/C contents (**Extended Data Fig. 1e**).

 To interrogate whether the binding of TFs facilitates the dissociation of nucleosome, we carried out the fifth cycle and separated the TF-bound species into libraries unbound and bound by nucleosome as follows. The TF-bound species associated with nickel beads were eluted with 300 mM imidazole (Sigma), and consecutively pulled down for four times by incubating 15 min with magnetic streptavidin beads (1.2 µl, 28-9857-99, GE Healthcare) to deplete the nucleosome-bound species. The streptavidin beads incubated in the first pulldown was washed and eluted to collect the nucleosome-bound ligands, which is sequenced as the bound library. After all four pulldowns, the DNA ligands remaining in the final supernatant were sequenced as the unbound libraries. Both the nucleosome-bound and unbound libraries from the final cycle were sequenced. As a control, the cycle five nucleosome was also allowed to dissociate in the absence of TFs; the bound library and the unbound library were collected as described above. The result shows little difference of E-MI between the bound and unbound libraries (**Extended Data Fig. 9a**, right). We also verified that TF concentration only has minor effect on the results (**Extended Data Fig. 9c**). In cycle 1-4, the NCAP-SELEX process captures TFs bound to DNA in the presence of nucleosomes, either by displacing the nucleosome, or by binding to DNA together with it. Both types of sequences are enriched, and there is thus no selection for nucleosome dissociation *per se* during the four initial rounds. The final round is then performed to determine whether the TF induces nucleosome dissociation or not. Although more nucleosome dissociation is expected for earlier cycles, the dissociation assay is carried out at the end to enrich and study the sequences that have considerable affinity to nucleosome, but get dissociated by the TFs. In the first cycle, the dissociation of nucleosome is likely affected more by ligands’ affinity to the nucleosome, and less by the binding events of the TFs. In the dissociation experiment, the salt concentration is around 100 mM thus nucleosome cannot reform once dissociated. Multiple binding events were reported to bind and destabilize nucleosome cooperatively72,73. This is expected but difficult to detect using NCAP-SELEX, because far more single motif matches are found in random sequences compared to multiple matches.

 The NCAP-SELEX and HT-SELEX library for each TF contains hundreds of thousands of unique reads. Under this sample size, if a TF is binding nucleosomal DNA without restrictions, any non-random pattern of TF binding that has a biologically meaningful effect size (as observed in our study) can only occur with an extremely small p-value.

### **Sequencing and pre-processing**

The SELEX ligands amplified with multiplexing primers were purified with AMPure beads (Beckman Coulter), and sequenced using Illumina Hiseq 2000 or Hiseq 4000, with >80 bp paired-end settings. Raw sequences were demultiplexed with bcl2fastq (v2.16.0.10). In general hundreds of thousands of reads were obtained for each TF.

 The R1 and R2 reads of paired-end sequencing were merged with PEAR74 requiring 5 bp overlap at minimum. The merged sequences were discarded if their variable region length is not the same as the ligand design. The obtained sequences were then trimmed for adaptor and for quality by Trim Galore (version 0.4.3). All trimmed sequences were subsequently removed. The sequences were further cleaned for adaptor sequences by removing all sequences that contained a 14-bp overlap with Illumina sequences. For further analysis, we removed the PCR duplicates and used only the unique reads.

### **TF signal analysis with E-MI**

TF binding on the SELEX ligands were analyzed for the NCAP-SELEX libraries, and as a control, also for the HT-SELEX libraries. We initially analyzed the data using the MI-based approach, which is not dependent on prior knowledge of the TF motifs. Specifically, the binding signals were evaluated by the MI between 3-mer distributions at two non-overlapping positions of the ligand **(Fig. 1b)**. The underlying rationale is that if a binding event contacts two continuous or spaced 3-bp wide positions of the SELEX ligand at the same time, the 3-mer distributions at these two positions will be correlated in the enriched library. This biased joint distribution is then detected as an increase in MI between the positions.



where *P(3+3-mer)* is the observed probability of a 3-mer pair (i.e. gapped or ungapped 6 mer) from position 1 and position 2. *Ppos1(3-mer)* and *Ppos2(3-mer)*, respectively, are the marginal probabilities of the constitutive 3-mers at position 1 and position 2. Their product represents the expected probability of the 3-mer pair. Sums are over all 3-mer pairs.

In the heatmap showing MI from all possible 3-mer pairs, TF signal is visible in the vicinity of the diagonal (e.g., HSF1; **Fig. 1b**, left), as TF binding causes correlation between the distributions of closely spaced 3-mers. However, the nucleosome binding preference also appears as stripes with ~ 10 bp spacing, as histones contact DNA at ~ 10 bp intervals29,40,69,75. Because nucleosomes can bind to most sequences, whereas TFs bind to only a few specific sequences, it is possible to almost completely separate the TF signals from the nucleosome signals (**Fig. 1b**, right; **Extended Data Fig. 1g**), by limiting the MI measure to the top 10 most highly enriched 3-mer pairs (enriched 3-mer pair based mutual information; E-MI).



In the MI-based analysis, we have empirically chosen to use the kmer of length 3 for a balance between the signal/noise (S/N) ratio and the resolution. This is because a short kmer gives better resolution and less noisy result (more counts for each kmer), whereas a long kmer gives stronger E-MI signal as long as the kmer counts are sufficient. For E-MI, we focused on only 10 of the most enriched 3-mer pairs because such criterion efficiently filters out the nucleosome signal but retains most of the TF signals.

We employed the E-MI approach in most of the analyses because it is not affected by potential changes in TF motifs and presents all binding signals of a TF at the same time for a quick overview. In addition, E-MI also helps to locate TF-DNA contacts; as E-MI detects correlations between 3-mer pairs, its signal is usually strongest at positions corresponding to direct TF-DNA contacts. Although a modified motif matching (e.g. scoring each position of the hit based on the motif information content) is also possible to locate such contacts, the E-MI based measure is more straightforward. The E-MI measure also offers a convenient way to visualize all possible dimer modes of a TF and to compare selectivity between TFs, because it captures all binding signals without predefined assumptions (i.e. not restricted by motifs).

In the 2D E-MI plots, the randomized region is 154 bp for lig200, and thus contains 149 windows for MI calculation between neighboring 3 mers. For lig147, the randomized region is 101 bp, and contains 96 windows for MI calculation. Because TFs rarely bind or cooperate across a large span of DNA, most E-MI signals are visible only near the diagonal of the 2D E-MI plot (**Fig. 1b**, right). In most analysis we focus on the E-MI diagonals (*pos2* = 3 + *pos1*;**Supplementary Table 3**) which reflect the footprints of TFs on the SELEX ligands. For clustering analysis, each E-MI diagonal was first linearly normalized by subtracting the minimum and then dividing with the maximum. Next, clustering of the E-MI diagonal was performed using the cosine distance metric and ward.D2 linkage of the *hclust* function in R. The clustering is performed without any sequence information and only considers the diagonal of the 2D E-MI data. The circular representation of the classification result was generated using the *circlize* R package76. To calculate the penetration of E-MI for each TF, the diagonal of E-MI was first LOESS smoothed with a span of 0.45; next, for each half of the diagonal, the maximum E-MI value among the half was identified; after that, the positions where the E-MI decreases to half of the E-MI maximum were taken as the penetration depth. The final penetration depth is the average value for both halves of the E-MI diagonal.

To check whether the gyre-spanning mode of TF T is preferring nucleosomal DNA, for both its bound and unbound libraries of cycle 5, the E-MI strength of Type 2 binding was evaluated by summing E-MI from 3-mer pairs spaced 77–83 bp, the E-MI strength of the background was evaluated by summing E-MI from 3-mer pairs spaced 50–70 bp. For both the binding signal and the background, Log2 ratios of E-MI strength between the bound and unbound libraries were calculated for four independent replicates of NCAP-SELEX using TF T. The obtained ratio indicates whether the signal (or the background) has a different strength between the two libraries.

When comparing E-MI between the bound and the unbound libraries from cycle five, only TFs with the 3×FLAG tag were considered.

### **Motif discovery and motif matching**

Primary motifs were manually curated for each TF’s NCAP-SELEX and HT-SELEX libraries (**Supplementary Data 1**; **Supplementary Table 7**). Seeds for the PFMs (positional frequency matrix) were selected according to our previous criteria65. PFMs were then generated using Autoseed19,77 with a multinomial of 1. Comparison between motifs on the nucleosomal and free DNA was performed similarly. The motif analyses revealed that most TFs bind with similar specificity in the presence and absence of nucleosome (**Supplementary Data 1**). However, consistent with earlier observations21, we also found a few cases where the binding specificities of the TFs were detectably different; the difference was confirmed by analysis of the enriched subsequences (**Extended Data Fig. 5b**).

Motif matching with the curated motifs18,65 were conducted using MOODS78,79 with the p-value set to 0.0001. Motif hits from both strands were combined unless indicated.

To supplement the systematic E-MI analysis with motif matching results (**Supplementary Data 2**), we used an automatic pipeline that discovers a 20-bp-long primary motif for each TF’s NCAP-SELEX library, and then used the discovered motif to perform motif matching. This automatic pipeline is only used for analyses in **Supplementary Data 2**. The manually curated PFMs were not used because they are of different lengths, and their curation criteria also vary slightly between individual TFs. To discover the primary motif, first we counted all 10-mers (with a 0–10 bp gap in the middle) for each cycle 4 NCAP-SELEX library. Compared to cycle 0, the most enriched heteropolymeric 10-mer (at least hamming distance 2 from any homopolymer run) is selected as the seed. The seed is then used to derive a PFM with multinomial 1, and subtracted for background carryover according to our previous approach47. The length of the derived PFM is trimmed to 20 bp centered at the seed. Subsequently, the motif matching is performed using MOODS with p=0.0001.

### **Quality control of the SELEX experiments**

The successful TFs were called by manually checking the E-MI and motif discovery results for each TF. The successful TFs have detectably stronger E-MI between neighboring 3-mer pairs than that between 3-mer pairs far away from each other, and show enriched motifs that are not contaminations from unrelated TFs. We also found that most of the failed TFs are due to the lack of active protein expression.

We also noticed that in addition to the TF-nucleosome interaction, additional sources such as the adaptor sequence, adaptor length, and the oligonucleotide bias during ligand synthesis may as well affect TFs’ positioning on the SELEX ligand, and lead to asymmetric E-MI patterns in both the NCAP-SELEX (**Extended Data Fig. 2a**; **Fig. 3a**) and the HT-SELEX (**Extended Data Fig. 3b**, **6a**). However, the conclusions are reliable based on the different positional preferences observed between the NCAP-SELEX and HT-SELEX (**Extended Data Fig. 2a** vs. **Extended Data Fig. 3b**; **Fig. 3a** vs. **Extended Data Fig. 6a**; **Supplementary Data 2a**). The reliability is further supported by the agreement of TFs’ patterns on lig200 to those on lig147 (**Extended Data Fig. 6f**,weaker positional preferences on lig200 due to nucleosome delocalization). Moreover, TFs’ compatibility with nucleosomal DNA inferred from lig147 and lig200 data also correlates with each other (**Extended Data Fig. 6e)**. However, a direct comparison between lig147 and lig200 is impossible because lig200 has delocalized nucleosome, and that on lig200 TF signals from the free DNA region will overwhelm the signals inside the nucleosome-occupied region.

### **Evaluation of nucleosome-induced orientational preference of TFs**

Because DNA is double-stranded, TFs can bind to it in two different orientations. For TFs that bind non-palindromic sites, their binding orientation can be determined from the bound sequences. On free DNA, a binding sequence has the same affinity for TF-binding irrespective of its orientation. This is not true when DNA is wrapped onto a nucleosome as the nucleosome breaks DNA’s 2-fold rotational symmetry. Depending on the sequence’s relative orientation to nucleosome, the affinity can also differ (**Fig. 2e** and **Extended Data Fig. 4**). We systematically examined this asymmetric effect between binding orientations by comparing the strand-wise distributions of top 8-mers for lig200.

For each TF, we first calculated the binding energy difference (∆∆*G*) between the two relative orientations for each of the most enriched non-palindromic 8-mers (top 40 used). The ligands in this TF’s SELEX library were divided into two halves according to the dyad position. The two halves were calculated separately and then averaged. Similarly to previous studies80,81, we assumed a low TF concentration and that the dissociation during wash is insignificant for high-affinity 8-mers. Consequently, for each 8-mer and for each half of the ligands, the ∆∆*G* of the 8-mer between the two relative orientations is



Here *C5’* and *C3’* are counts of this 8-mer, respectively for the DNA-strands with their free ends located at the 5’ and the 3’ (the other end is at the dyad where we divide). The count ratio *C5’*/*C3’* in cycle *r* was normalized with the count ratio in cycle 0, taken the *r*th root to account for the exponential enrichment in SELEX, and subsequently converted into energy difference. The directional energy difference for each 8-mer was then averaged for the two halves of the ligands, and the absolute value is used to represent the orientational asymmetry of this 8-mer



Orientational asymmetry of the TF is then represented by the mean of the 40 most enriched 8-mers’ orientation asymmetry.

To rule out any potential orientational bias induced by the adaptors of the SELEX ligands, we also calculated theorientation asymmetry values for 8-mers in the HT-SELEX library. For each TF, the 8-mers used for its HT-SELEX library are the same 8-mers as used for its NCAP-SELEX library. After obtaining the 8-mers’ orientation asymmetry values for both the NCAP-SELEX library and the HT-SELEX library of the TF, we used a one-tailed t-test to examine if the orientation asymmetry values in the NCAP-SELEX library are larger than those in the same TF’s HT-SELEX library, and obtained the p-value.

Signal enrichment in each TF’s library was represented using the median fold change of the 8-mers that are most enriched (top 10 8-mers). The fold change for each 8-mer was calculated using log2(cycle 4 count / cycle 0 count). This study observed the orientational asymmetry for ELF factors. Such asymmetry occurs on nucleosomal DNA, but also extends to the adjacent linker DNA (**Extended Data Fig. 4f**). It is worth noting that although nucleosome breaks the local rotational symmetry of DNA, the nucleosome as a whole remains pseudo-symmetric against a pseudo 2-fold axis (dyad axis). Therefore, the distributions of motif matches in the two orientations are symmetric with regard to the dyad.

### **Fast Fourier Transformation (FFT) analysis and structure alignment**

The diagonal of E-MI for each TF’s library was subtracted with the mean, windowed by Welch’s function, and then subjected to FFT. The obtained power spectrum was further divided with the mean of the E-MI diagonal and the length of the diagonal. We next calculated FFT-AUC (area under the curve) from the power spectrum and used it as an indicator for the ~10 bp periodicity induced by nucleosome. The FFT-AUC was calculated for frequencies ranging from 0.08–0.12 bp-1 and subtracted with the baseline level (estimated from 0.14–0.3 bp-1). The phase of FFT was examined at 0.102 bp-1. The same process was applied to the TA dinucleotide counts across all positions of the ligand for the NCAP-SELEX library of all individual TFs. The median value of TA phase in all NCAP-SELEX libraries is used as a reference (**Fig. 4a**). TFs’ groove preference can be inferred by comparing the phase of E-MI and that of the TA counts, because high E-MI signal usually occurs at positions that correspond to direct TF amino-acid to DNA contacts. On the other side, TA-enriched positions on nucleosomal DNA correspond to positions where histones contact DNA4,40, which are also positions where the DNA major groove is facing towards the solvent. Thus, TFs that bind to the major groove tend to show E-MI maximums in phase with TA, and TFs that bind to the minor groove commonly display E-MI maximums out of phase with TA. The FFT analysis also shows that a few TFs (e.g. ETS factors) may have periodic preference in the absence of nucleosome (**Fig. 4a**; **Supplementary Data 1**). This likely emerges from the allosteric effect of DNA82.

To mimic the in-phase and out-of-phase bindings of TFs with periodic preferences relative to the preferred TA positions on nucleosome, the available structure of TF-DNA complex was aligned (with UCSF Chimera49) to the nucleosome by matching the center of the TF’s core binding sequence either to the TA step (in phase), or to a step 5-bp downstream of the TA step (out of phase). The 6-bp core binding sequence in the structure of TF-DNA complex is defined according to the most enriched 6-mers in this TF’s NCAP-SELEX library. To make the alignment, C1–C4 on all deoxyribose rings were matched between the 6-bp core binding sequence and the 6-bp nucleosomal DNA centered in-phase or out-of-phase to the TA step. The protein structure models are visualized with UCSF Chimera49 and DNAproDB83.

### **TF Classification according to the identified binding modes**

The successful TFs in NCAP-SELEX were systematically classified based on strength of each TF-nucleosome interaction mode. The five modes include TF’s positional preferences (end, periodic, and dyad) on nucleosomal DNA, gyre-spanning binding mode, and orientational asymmetry. Custom classifiers are defined based on expert analyses and calculated for each TF. Then, discrete classes are called by applying thresholds to each classifier. The classifiers for positional preferences are E-MI penetration (end), FFT amplitude at 0.102 bp-1 (periodic), and a simple "fraction of E-MI at dyad" classifier (dyad). The E-MI from 3-mer pairs spaced 75–85 bp is used as classifier for the gyre-spanning mode. Whereas the other classifiers are based on TF’s behavior on lig147, the orientational asymmetry classifier uses lig200 results (the distribution asymmetry of 8-mers, as defined in the “Evaluation of nucleosome-induced orientational preference of TFs” section). In addition, we also classified TFs according to their effect on nucleosome stability, by the log ratio of the E-MI diagonal (bound vs. unbound). The thresholds for the classifiers, the classifier values and class-calls of each TF are curated in **Supplementary Table 5**.

We also tried pure data-driven approach for TF classification. Dimension-reduction approaches like PCA (principal component analysis) and NMF (non-negative matrix factorization) also give dimensions corresponding to the identified positional preferences of TFs (**Extended Data Fig. 6b**; **Supplementary Table 6**). The correlation between such dimensions and the E-MI diagonal also serve as a classifier to evaluate TF’s positional preferences (left of **Extended Data Fig. 6c**) but is less optimal than the custom classifiers. This is because the dimensions are still conflated with multiple features, and cannot account for the continuous phase of the periodic preference (right of **Extended Data Fig. 6c**). Therefore we prefer the custom classifiers, which are based both on the data and the physical understanding of the nucleosome structure. The PCA and NMF analyses were respectively performed using the FactoMineR (v1.39) and NMF (v0.20.6) R packages.

### **MNase-seq**

In MNase-seq, the LoVo cell line from ATCC was used (CCL-229, tested to be free of mycoplasma infection by Hoechst staining). MNase-seq was performed as described previously84. Specifically, 107 cells were harvested and washed twice with 10 ml cold DPBS (Dulbecco's phosphate-buffered saline), spinned down with 350 g for 5 min at 4 °C. The cells were next crosslinked with 10 ml of 1.1% formaldehyde for 10 min in DPBS, tumbling end over end. The crosslinking reaction was quenched with 50 µl 2.5 M glycine and further tumbled for 2 min, and washed twice with cold DPBS. Lysis of the cells was performed with 20 ml 0.5× PBS containing 0.5% Triton X-100 for 3 min on ice; the nuclei were then collected by centrifugation (350 g, 5 min). Before MNase digestion, the nuclei were washed three times with 1× MNase digestion buffer, resuspended with 1 ml of the same buffer containing 100 µg/ml RNase A. An aliquot of 100 µl was used for MNase digestion. MNase digestion was carried out with 100 units of MNase (M0247S, NEB) at 37 °C for 8 min, quenched with 100 µl stop buffer (40 mM EDTA, 40 mM EGTA, 1% SDS, 1.5 mg/ml proteinase K) at 65 °C o/n. The MNase fragments with length of 100–1000 bp were selected using Ampure beads (Beckman Coulter), and subjected to the library preparation workflow of Illumina (E7370L, NEB). The paired-end sequencing (2 × 86 bp) was performed using Illumina HiSeq 4000. MNase-seq for HEK293 cell line is performed similarly as controls for the MNase-ChIP.

MNase-sequencing data from K562 cell line were downloaded from GEO accession GSE78984. Three titration series (20.6U, 79.2U and 304U) of MNase were selected.

### **Combined analysis of MNase-seq and ChIP-seq**

For MNase-seq data, the raw sequencing reads were quality and adapter trimmed with cutadapt version 1.12 in Trim Galore (version 0.4.3). Low-quality ends trimming was done using Phred score cutoff 30. Adapter trimming was performed using the first 13 bp of the standard Illumina paired-end adapters with default parameters. Raw sequencing reads were mapped to the human reference genome (hg19) using bwa85 with default parameters. Duplicates were removed with samtools (v 1.3.1) rmdup function. Insert size distribution was calculated based on 10000 reads that were aligned to autosomes. After duplicate removal, data from K562 titration series were merged.

Coverage of MNase fragments with length >140 bp was calculated at ChIP-seq peaks of 20 TFs in K562 cell line. We selected 500 highest signal ChIP-seq peaks that had respective TF’s motif match site and did not overlap with hg19 blacklist genomic regions. BEDtools (v2.26.0) genomecov and intersect functions were utilized in calculations. ENCODE narrowPeak calls including two replicates were used from March 2012 freeze (UCSC wgEncodeAwgTfbsUniform track) release for ATF3, CEBPB, CTCF, ELF1, GATA2, JUND, SRF, USF2 and YY1, and from later releases (the ENCODE Portal http://www.encodeproject.org, accessed 07/12/2017) for ATF2, CREB3L1, CREM, ELF4, HMBOX1, MYBL2, NFATC3, PKNOX1, RFX1, SREBF1 and YBX128. All TFs with ENCODE data and lig200 NCAP-SELEX data were included in the analysis without selection. Genomic sites recognized by each motif retrieved from previous HT-SELEX runs were searched from the human genome using program MOODS78 with a p-value cut-off of 10-4 and a score cut-off of 5. Final MNase fragment coverage values were calculated by taking the average MNase-seq coverage across multiple motifs for each TF, and correlated with E-MI penetration values with Pearson’s method.

To visualize nucleosome distribution near TF sites *in vivo*, MNase-fragments aligned to autosomes were used. MNase-fragments overlapping with hg19 blacklist genomic regions were excluded. The LoVo ChIP-seq data (for ELF1 and 2) from Yan *et al.*17 were downloaded from GEO accession GSM1239499 and GSM1208610. Genomic sites recognized by each motif were searched from the human genome using MOODS78 with a p-value cut-off of 10-4 and a score cut-off of 5. Coordinate of the center-point of MNase-fragments was calculated relative to the motif sites within ChIP-peaks using BEDtools (v2.26.0) closest function. The motif sites were oriented by the strand information from MOODS. To verify the significance of the asymmetry around TF sites, we also examined the nucleosome distribution around genomic sites of 30 random 9-mers. Nucleosome occupancy at the two sides of the target motif or 9-mer sites was estimated using MNase-fragments that are 140–170 bp in length, and centering at 83–183 bp from the target sites.

### **MNase-ChIP and data analysis**

The HEK293 cell line from ATCC was used (CRL-1573, tested to be free of mycoplasma infection by Hoechst staining). Cells were first cultured in T-160 flasks until 60% confluency is reached. Subsequently, the cells were transfected with mammalian expression vectors of RFX5 or HOXB13 containing a C-terminal V5 tag. For transfection, 12 µg plasmid was mixed with 750 µl Opti-MEM (31985062, ThermoFisher) and 60 µl FuGENE (E2311, Promega), incubated for 10 min under room temperature, and added into the cell culture. Transfected cells were further grown for 36 h and used for MNase-ChIP.

MNase-ChIP was performed according to the previous protocol86. Specifically, two million cells were harvested and washed twice with 1 ml cold DPBS, spinned down with 350 g for 5 min at 4 °C. The cells were next crosslinked with 1 ml of 0.5% formaldehyde for 10 min in DPBS, tumbling end over end. The crosslinking reaction was quenched with 1 ml 0.25 M glycine and further tumbled for 2 min, and washed twice with cold DPBS. Lysis of the cells was performed with 1 ml 0.5× PBS containing 0.5% Triton X-100 for 3 min on ice; the nuclei were then collected by centrifugation (350 g, 5 min). The collected nuclei were resuspended in 100 µl 1× MNase digestion buffer containing 100 µg/ml RNase A. MNase digestion was performed with 40 units of MNase (M0247S, NEB) at 37 °C for 8 min, and quenched with 100 µl stop buffer (40 mM EDTA, 40 mM EGTA, 5 mg/ml BSA, 150mM LiCl, 2 mM TCEP). The quenched reaction is then filtered through a membrane (0.65 µm pores) to remove cell debris. After taking a 20 µl aliquot as input, the remaining solution is incubated with 20 µl Protein G Dynabeads (10003D, ThermoFisher) and 1 µg V5 antibody (R96025, ThermoFisher) for 1 h at room temperature. The beads were then collected by magnet and washed three times with the wash buffer (0.1% SDS, 1% Triton, 2 mM EDTA, 0.5% sodium deoxycholate, 75 mM LiCl, 75mM NaCl in 20 mM Tris, pH8). The retrieved DNA was subjected to the library preparation workflow of Illumina (E7370L, NEB). The paired-end sequencing (2 × 91 bp) was performed using Illumina HiSeq 4000. As a control, the mock ChIP of HEK293 without TF transfection, the input of each TF’s ChIP (i.e. MNase-seq 36 h after TF transfection), as well as the MNase-digested fragments of HEK293 before transfection were also subjected to library preparation and sequenced. The MNase-seq controls are necessary because although not always the case, many TFs have been reported to interact with chromatin remodelers87-89. For such TFs, their TF-nucleosome complexes are expected to be unstable in cells. Two replicates are available for each MNase-ChIP.

The sequencing reads for MNase-ChIP and for the controls (MNase-seq) were processed with the same pipeline as described above for MNase-seq. Peaks of MNase-ChIP were called using MACS90 (fold enrichment 2-50, q-value < 0.05; **Supplementary Table 8**). The top 1000 peaks according to the q-value were selected for motif discovery. Peaks with abnormally high occupancies were excluded (greater than 40 fold of the median coverage of the top 20 peaks). Distribution of the MNase-seq and MNase-ChIP fragments near TF motifs (in peaks of the MNase-ChIP) were visualized as described in the prior section (“combined analysis of MNase-seq and ChIP-seq”).

Fragmentation of ChIP can be achieved either by MNase digestion or by sonication. We prefer MNase-digestion here due to the high resolution reported for MNase-ChIP51, and more importantly, due to the fact that in MNase-ChIP typical TF-bound fragments (of sub-nucleosomal size) are distinguishable from nucleosome-bound fragments according to fragment length (**Extended Data Fig. 8c** and previous reports50,51). Thus it provides a powerful tool to identify TFs that bind on a nucleosome. Although combination of high-resolution maps respectively for nucleosome and for TF can also identify positional relationships between the two69,70, however, such approach is agnostic about whether the sites are at the same time bound by both a nucleosome and a TF.

### **Electrophoretic mobility shift assay (EMSA)**

Nucleosomes were formed essentially as described previously60 from the histone octamers and the modified Widom 60142 DNA sequence CTGGAGAATCCCGGTCTGCAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAAACGCACGTACGGTATTGTTTATTTTGTTCCTCCGCCAAGGGGATTACTCCCTAGTCTCCAGGCACGTGTCAGATATATACATCCTGT. A SOX11-binding segment (GGTATTGTTTATTTTGTTCCT) is positioned at the center of the modified Widom 601 DNA. The sequence of the inserted segment is extracted from a ligand in the cycle 4 NCAP-SELEX library of SOX11. The inserted segment position on Widom 601 is the same as the segment’s original position on the SELEX ligand. Nucleosomes were reconstituted using this modified Widom 601 ligand and subsequently heat-shifted at 55°C for 30 min. Next the nucleosomes (containing 1 µg DNA) were incubated on ice with purified SOX11 eDBD in a 40 µl volume. As a control, the SOX11 eDBD were also directly incubated with 1 µg modified Widom 601 ligand in 40 µl volume. The samples were then subjected to EMSA. A 0.8% agarose gel was cast and run in the 0.2x Tris–Boric acid–EDTA (TBE) buffer. EMSA was performed in native conditions at 4°C for 1 h at 120 V, and later the gel was post-stained in DNA Stain Clear G (Serva). The DNA ladder 100 bp (NEB) was used as the marker.

### **Binding affinity comparison by competition analysis**

To compare RFX5’s affinity to free-DNA ligands with its affinity to nucleosomal-DNA ligands, we performed a competition assay. First, four cycles of NCAP-SELEX were carried out for RFX5 (full-length) using barcoded 147-bp ligands (Lig147\_ATTA and Lig147\_TAAT, **Supplementary Table 4**). In the last cycle, nucleosomes were reconstituted with Lig147\_ATTA, purified, and then mixed with Lig147\_TAAT. The mixture (input) was incubated with purified RFX5 for 15 min in the incubation buffer for NCAP-SELEX. The TF-bound species were then pulled down with nickel magnetic sepharose beads and washed 15 times. The ligands associated with TF-bound species, as well as the input mixture, were amplified and sequenced. The sequencing reads were processed as described in “Sequencing and pre-processing”.

RFX5 sites in both the TF-bound library and the input library were identified using MOODS78 with a p-value cut-off of 10-4. Motif hits at each position of the ligand were counted for both libraries, separately for Lig147\_ATTA (Nucleosomal DNA) and Lig147\_TAAT (free DNA). The affinity difference between RFX5’s nucleosomal-DNA sites and free-DNA sites was estimated for each position of the ligand by comparing the counts (*C*) of motif hit on different barcodes:

$$Affinity difference=ln\frac{C\_{bound}^{ATTA}/C\_{input}^{ATTA}}{C\_{bound}^{TAAT}/C\_{input}^{TAAT}}$$

### **Binding of nuclear proteins to reconstituted nucleosome**

The workflow is adapted from the active TF identification method91. The nucleosomes were first reconstituted using lig147 and immobilized onto streptavidin beads followed by washing. Nuclear proteins extracted from mouse liver (using the kit from Life Technologies, 87790) were then added and incubated for 90 min at room temperature. The reaction volume is 20 µl, containing 0.5 pmol nucleosome, 20 µg nuclear extract, and 5% w/v PEG4000 in the binding buffer (140 mM KCl, 5 mM NaCl, 2 mM MgSO4, 3 μM ZnSO4, 100 μM EGTA, 1 mM K2HPO4, in 20 mM HEPES, pH 7.0). Then, the nucleosomes were eluted using 10 mM biotin, and subjected to EMSA to separate the ligands that were also bound by proteins from the nuclear extract. The collected ligands were PCR-amplified and used to reconstitute nucleosome again. The reconstituted nucleosomes were immobilized and incubated with nuclear extract as mentioned above. After incubation, the dissociated ligands in the solution were collected (the unbound library), and the ligands bound by both nucleosome and proteins from the nuclear extract were collected by EMSA as well (the bound library). Both libraries were PCR-amplified and sequenced. Homeodomain sites in both the bound and the unbound libraries were identified and counted using MOODS78 with a p-value cut-off of 10-4. To compare the binding frequency between the two libraries, the raw counts in each library were further normalized by dividing each library’s total reads number.

### **Determination of nucleosome positions on DNA ligand**

Nucleosome positions on DNA ligands were determined for lig200 NCAP-SELEX libraries by MNase digestion. First, nucleosomes were reconstituted using amplified cycle 4 libraries of each TF, and purified as described for the NCAP-SELEX workflow. The nucleosomes were then digested with 0.15 unit MNase (M0247S, NEB) at 37°C for 8 min in 25 µl digestion buffer (10 mM biotin, 5 mM CaCl2 and 1 mM TCEP in 10 mM Tris-HCl, pH 8). The MNase reaction is stopped by adding 15 µl stop buffer (40 mM EDTA, 40 mM EGTA, 1% SDS, 1.5 mg/ml proteinase K) and incubated overnight. Next, the digested DNA fragments were recovered with 3× AMPure beads (Beckman Coulter). The retrieved DNA was subjected to the library preparation workflow of Illumina (E7370L, NEB). The paired-end sequencing (2 × 91 bp) was performed using Illumina HiSeq 4000.

The sequencing reads were demultiplexed with bcl2fastq (v2.16.0.10). The R1 and R2 reads of paired-end sequencing were merged with PEAR74 requiring 5 bp overlap at minimum. Then for each TF’s library, 100,000 merged reads were sampled and aligned back to the design of lig200 (“N” used for bases in the variable region). In the alignment, only correct matches and matches to “N” are allowed (no gaps and mismatches allowed), and only reads with a length <150 bp were considered, because the remnant undigested ligands (with an apparent insert length of 154 bp in sequencing) would undermine the analysis if included. Reads with at least 6 bp aligned to the adaptors are defined as adaptor-overlapping reads, and used to determine the preferred nucleosome positions on the ligand. Reads with no more than 2 bp alignment to the adaptors are defined as reads from the variable region.

### **Statistical analysis**

No statistical methods were used to predetermine sample size. All used statistical tests and the results are indicated in the individual figure legends. Samples were analyzed directly and individually for each TF, and not randomized to experimental groups. Investigators were not blinded.

### **Reporting summary**

Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

### **Code availability**

All of the computer programs and scripts used are either published or available upon request.

### **Data availability**

All next-generation sequencing data have been deposited in the European Nucleotide Archive (ENA) under accession PRJEB22684. The relevant processed data are included as supplementary data and tables.

60 Dyer, P. N. *et al.* Reconstitution of nucleosome core particles from recombinant histones and DNA. *Methods Enzymol.* **375**, 23-44 (2004).

61 Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516-520 (2011).

62 Wang, J. P. *et al.* Preferentially quantized linker DNA lengths in Saccharomyces cerevisiae. *PLoS Comput. Biol.* **4**, e1000175 (2008).

63 Felsenfeld, G. & Groudine, M. Controlling the double helix. *Nature* **421**, 448-453 (2003).

64 Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-1723 (2009).

65 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327-339 (2013).

66 Hughes, A. L. & Rando, O. J. Mechanisms underlying nucleosome positioning in vivo. *Annu. Rev. Biophys.* **43**, 41-63 (2014).

67 Radman-Livaja, M. & Rando, O. J. Nucleosome positioning: how is it established, and why does it matter? *Dev. Biol.* **339**, 258-266 (2010).

68 Fernandez, A. G. & Anderson, J. N. Nucleosome positioning determinants. *J. Mol. Biol.* **371**, 649-668 (2007).

69 Brogaard, K., Xi, L. Q., Wang, J. P. & Widom, J. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**, 496-501 (2012).

70 Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362-366 (2009).

71 Olson, W. K. & Zhurkin, V. B. Working the kinks out of nucleosomal DNA. *Curr. Opin. Struct. Biol.* **21**, 348-357 (2011).

72 Adams, C. C. & Workman, J. L. Binding of Disparate Transcriptional Activators to Nucleosomal DNA Is Inherently Cooperative. *Mol. Cell. Biol.* **15**, 1405-1421 (1995).

73 Polach, K. J. & Widom, J. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J. Mol. Biol.* **258**, 800-812 (1996).

74 Zhang, J. J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614-620 (2014).

75 Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659-675 (1986).

76 Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. *circlize* Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812 (2014).

77 Nitta, K. R. *et al.* Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* **4**, e04837 (2015).

78 Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181-3182 (2009).

79 Pizzi, C., Rastas, P. & Ukkonen, E. Finding significant matches of position weight matrices in linear time. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**, 69-79 (2011).

80 Slattery, M. *et al.* Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147**, 1270-1282 (2011).

81 Riley, T. R. *et al.* SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Methods Mol. Biol.* **1196**, 255-278 (2014).

82 Kim, S. *et al.* Probing allostery through DNA. *Science* **339**, 816-819 (2013).

83 Sagendorf, J. M., Berman, H. M. & Rohs, R. DNAproDB: an interactive tool for structural analysis of DNA-protein complexes. *Nucleic Acids Res.*, doi: 10.1093/nar/gkx1272 (2017).

84 Mieczkowski, J. *et al.* MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat. Commun.* **7**, 11485 (2016).

85 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

86 Skene, P. J. & Henikoff, S. A simple method for generating high-resolution maps of genome-wide protein binding. *Elife* **4** (2015).

87 Nguyen, T. T. *et al.* Cdx2 regulates gene expression through recruitment of Brg1-associated Switch-Sucrose Non-fermentable (SWI-SNF) chromatin remodeling activity. *J. Biol. Chem.* **292**, 3389-3399 (2017).

88 Fryer, C. J. & Archer, T. K. Chromatin remodelling by the glucocorticoid receptor requires the BRG1 complex. *Nature* **393**, 88-91 (1998).

89 King, H. W. & Klose, R. J. The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatroy element funciton in mouse embryonic stem cells. *Elife* **6** (2017).

90 Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9** (2008).

91 Wei, B. *et al.* A protein activity assay to measure global transcription factor activity reveals determinants of chromatin accessibility. *Nat. Biotechnol.* **36**, 521-529 (2018).