Published online 13 May 2023

Nucleic Acids Research, 2023, Vol. 51, No. 11 ***5621–5633***
https://doi.org/10.1093/nar/gkad372

DNA binding specificity of all four *Saccharomyces cerevisiae* forkhead transcription factors

Brendon H. Cooper[1], Ana Carolina Dantas Machado[1], Yan Gan[1,2], Oscar M. Aparicio[2,3] and Remo Rohs[1,3,4,*]

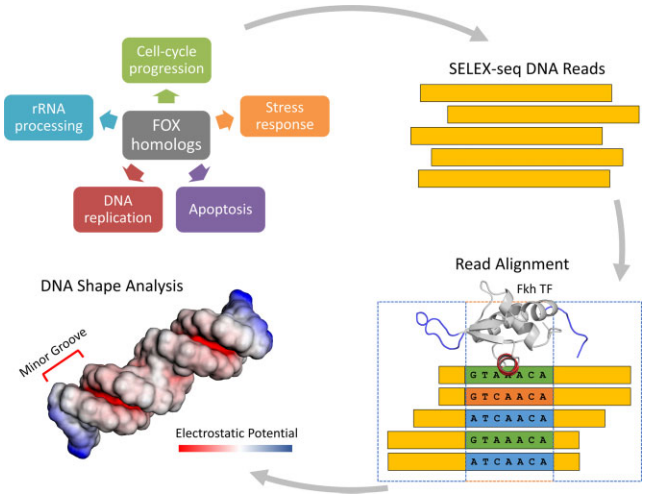
[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA, [2]Molecular and Computational Biology Section, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA, [3]Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA 90033, USA and [4]Departments of Chemistry, Physics & Astronomy, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA

Received May 30, 2022; Revised April 19, 2023; Editorial Decision April 20, 2023; Accepted April 27, 2023

ABSTRACT

Quantifying the nucleotide preferences of DNA binding proteins is essential to understanding how transcription factors (TFs) interact with their targets in the genome. High-throughput *in vitro* binding assays have been used to identify the inherent DNA binding preferences of TFs in a controlled environment isolated from confounding factors such as genome accessibility, DNA methylation, and TF binding cooperativity. Unfortunately, many of the most common approaches for measuring binding preferences are not sensitive enough for the study of moderate-to-low affinity binding sites, and are unable to detect small-scale differences between closely related homologs. The Forkhead box (FOX) family of TFs is known to play a crucial role in regulating a variety of key processes from proliferation and development to tumor suppression and aging. By using the high-sequencing depth SELEX-seq approach to study all four FOX homologs in *Saccharomyces cerevisiae*, we have been able to precisely quantify the contribution and importance of nucleotide positions all along an extended binding site. Essential to this process was the alignment of our SELEX-seq reads to a set of candidate core sequences determined using a recently developed tool for the alignment of enriched *k*-mers and a newly developed approach for the reprioritization of candidate cores.

GRAPHICAL ABSTRACT



INTRODUCTION

The Forkhead box (FOX) transcription factor (TF) family consists of over 43 homologs in human characterized by a highly conserved winged-helix DNA binding domain (DBD). Members within the family play crucial roles in regulating a variety of key processes from proliferation and development, to tumor suppression and aging (1). Because of this, it is not surprising that members within the FOX family recognize distinct DNA binding sites *in vivo* (2). Despite this, previous attempts to characterize DNA binding preferences across the FOX family using *in vitro* methods such as protein binding microarray (PBM) or high-throughput (HT)-SELEX approaches have revealed little variability in the position weight matrices (PWMs) between family members (Supplementary Figure S1A). Although several *in vivo* conditions can modulate the DNA binding specificity and

*To whom correspondence should be addressed. Tel: +1 213 740 0552; Fax: +1 213 821 4257; Email: rohs@usc.edu
Present address: Ana Carolina Dantas Machado, Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA.

activity of a TF, it is important to understand inherent binding preferences and the limitations of previous attempts to characterize them.

PBM and HT-SELEX binding assays have unveiled the motifs of hundreds of TFs. Despite these important advances, these methods are often inadequate to capture precise information about moderate-to-low affinity sequences. For PBM, also known as universal PBM (uPBM), the fluorescence of low-affinity probes is often indistinguishable from the background or not included in the array design, and HT-SELEX typically utilizes a dramatically shallower sequencing depth per TF compared to more targeted methods such as SELEX-seq (3,4). However, low-affinity binding sites frequently make up actively bound regions that modulate transcription of their target genes *in vivo* (5). In closely related homologs, such as those within the Hox family of TFs, differential binding to suboptimal sites can help distinguish family members (6).

To explore this phenomenon in the context of the FOX family, we chose to further investigate the binding of all four paralogs in yeast: Fkh1, Fkh2, Hcm1 and Fhl1. Based on DNA binding profiles published in UniPROBE (7), calculated using BEEML-PBM (8), Fkh1 and Fkh2 exhibit virtually indistinguishable binding preferences (Supplementary Figure S1B). While these TFs are capable of binding many of the same loci *in vivo*, hundreds of non-shared sites have also been identified using ChIP-chip (9). Unfortunately, ChIP-chip and ChIP-seq data is notoriously noisy, with broad peaks that are often unable to pinpoint the nucleotides bound by the TF.

To further explore the binding preferences of these proteins, we have performed SELEX-seq with deep sequencing to collect over ten million reads per sample after up to two rounds of selection. With this framework, we were able to collect more information for moderate-to-low affinity binding sites that would have been exponentially diluted had we utilized many rounds of selection. Because of the extreme sequencing depth utilized, enrichment measurements were remarkably consistent across independent windows and between subsampled sets of reads. Furthermore, these precise measurements were able to reveal small-scale differences between Fkh1 and Fkh2 homologs that were undetectable using the previously published PBM data (Supplementary Analysis, Supplementary Figure S2).

One of the biggest challenges in analyzing SELEX-seq data is identifying the location of the binding site or binding sites along the length of a given read. The original SELEX-seq protocol provides a method to calculate the relative binding affinity between *k*-mers (10,11), but the location of the binding site within these *k*-mers is unknown. This is especially problematic since the sequence context of any given *k*-mer is lost during this process, so interdependencies with positions outside of the *k*-mer would be lost. Because of this, enriched *k*-mers are likely to include incomplete binding sites, or even multiple overlapping binding sites. This makes it difficult to understand the biophysical processes of binding site recognition without further data processing.

Although many methods have been published for the analysis of HT-SELEX and SELEX-seq data (12,13), binding preferences are often presented in the form of a position weight matrix (PWM)—even for models which derive predictions from more complicated sets of features, such as convolutional neural networks (14–16). Unfortunately, PWMs inherently assume that base pairs (bp) contribute independently to binding (17). When this assumption is violated, the alignment of false binding sites can dilute the signal of nucleotide positions with a relatively small impact on binding. Dinucleotide PWMs can capture some of these interdependencies (18–20), but may still be inadequate in summarizing binding for the vast number of moderate-to-low affinity sequences.

With respect to the FOX family, nucleotide positions flanking the core of the binding site may be particularly insightful in differentiating the binding of closely related homologs (21). The winged-helix domain of FOX proteins consists of a highly conserved recognition helix that binds in the major groove, making several base-specific contacts to its target (Figure 1, Supplementary Figure S3). These positions contact a 6–7-bp region of the binding site that we refer to as the 'core', which exhibits high sequence specificity across FOX homologs (Supplementary Figure S1). For some co-crystal and NMR structures, these regions form contacts with the minor groove regions flanking the core of the binding site (21–24) (Figure 1, Supplementary Figure S3).

To further explore the role of these flanking positions, as well as interdependencies between positions within the core, we employed a recently developed tool, Top-Down Crawl (TDC) (25), as part of a multi-step alignment framework that allows us to align full-length reads and precisely detect the contributions of flanking positions to binding specificity in a core-specific manner. For Fkh1 and Fkh2, this approach has allowed us to expand the canonical 7-bp binding site to 13 bp by revealing the contribution of six positions flanking the core of the binding site. Although previously ignored, the cumulative impact of these positions is able to reduce the binding affinity of the best core to be lower than the worst. Comparatively, Hcm1 and Fhl1 exhibited minimal dependence on positions flanking the core.

MATERIALS AND METHODS

Protein expression and purification

The DBD of Fkh1 and its surrounding residues (amino acids 243–484) were cloned into the pET-28b(+) DNA vector such that a His-Tag is added to the N-terminal end of the polypeptide (Millipore Sigma: 69865). The product was then transformed into Rosetta 2(DE3) Competent Cells (Millipore Sigma: 71405) and expression was induced overnight at 16°C using 0.2 mM IPTG. Cells were centrifuged at 4°C and resuspended in an MCAC-0 buffer (20 mM Tris–Cl pH 8.0, 0.5 M NaCl, 10% glycerol, 1 μg/ml pepstatin A, 50 μg/ml TPCK, 1 mM benzamidine, 1 mM PMSF) on ice. The cells were then broken by sonication and the lysate was clarified by centrifugation at 4°C. The lysate was then incubated with a Ni-NTA resin (Qiagen: 30210) equilibrated in MCAC-0. Next, the resin was sequentially washed with MCAC buffer containing 10, 20, 30, 40 mM imidazole, followed by elution with 250 mM imidazole. Salts were then removed by buffer exchange using Amicon Ultra Centrifugal Filters (Millipore Sigma: UFC901024,

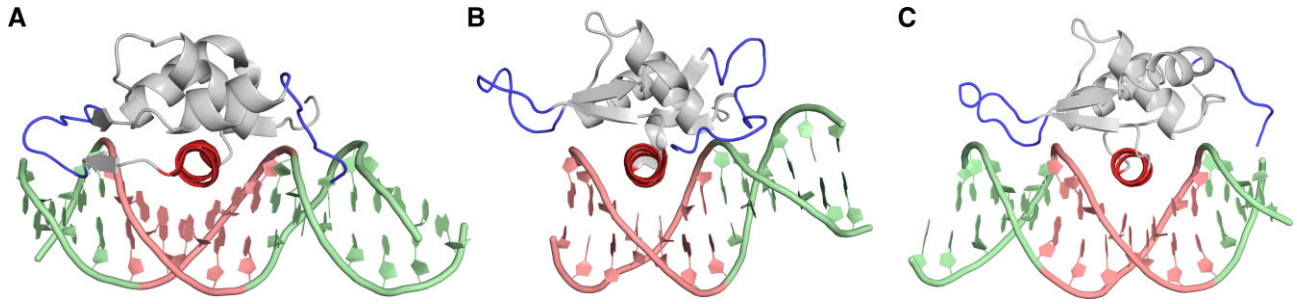


Figure 1. Representative structures of FOX DBDs bound to DNA. Winged regions are indicated in blue, and the main recognition helix is indicated in red with the core of the DNA binding site indicated in pink and flanking bp indicated in green. (**A**) *Rattus norvegicus* FOXD3 (PDB ID: 2HDC) (22), (**B**) human FOXA3 (PDB ID: 1VTN) (23), (**C**) human FOXK2 (PDB ID: 2C6Y) (24).

UFC800324). The final protein products were verified using SDS-PAGE (Supplementary Figure S4). The same procedure was followed for the expression and purification of the DBDs of Fkh2 (amino acids 280–520), Hcm1 (amino acids 41–213), and Fhl1 (amino acids 401–638), using pET-28a(+) as the DNA vector (Millipore Sigma: 69864).

Oligonucleotide synthesis

The library and all other DNA oligonucleotides were synthesized by Integrated DNA Technologies and purified by standard desalting (Supplementary Table S1).

SELEX-seq binding assay

The SELEX-seq procedure was carried out following the original protocol by Slattery *et al.* (10,11). The library was designed with a 16-bp randomized region surrounded by fixed adapters. Purification was carried out using Qiagen's MinElute PCR Purification Kit (Cat: 28004). Binding reactions were performed in a binding buffer consisting of 10% glycerol, 50 mM KCl, 20 mM Tris HCl (pH 7.9), 5 mM $MgCl_2$, 0.1 mg/ml BSA and 3 mM DTT. Since we were interested in the affinity of moderate-to-low affinity binding sites, we avoided the use of a non-specific polymer such as poly(dI·dC), and instead opted for a higher concentration of $MgCl_2$. Additionally, for each construct, we determined an optimum molar ratio of protein:DNA such that we observed preferential binding to a positive control over the negative control, both with the same fixed adapters as our library (Supplementary Table S1). These ratios were determined to be around 1:4 for Fkh1 and Fkh2, 1:1 for Hcm1, and 1:2 for Fhl1. Total amounts of protein and DNA used for each SELEX-seq experiment are shown on Supplementary Table S2. We used a 6-FAM labeled library for tracking, and all gels were visualized on Invitrogen's iBright™ CL1000 Imaging System. Bound fragments were excised and purified by phenol chloroform extraction followed by ethanol precipitation. We sequenced samples from round zero (R0), round one (R1) and round two (R2) of selection.

Competitive binding assay

Competitive binding assays were performed to compare the binding strengths of FOX homologs to varying sequences as described in the text (Supplementary Table S1). In this binding assay, a fixed amount of 6-FAM labeled probe was incubated with a limited amount of binding protein such that binding was visible but non-specific binding was prevented (Supplementary Table S3). An increasing amount of unlabeled competitor was included in the reaction mixture until binding by the labeled probe was visibly reduced by at least 50%, representing the half maximal inhibitory concentration (IC_{50}). This value was estimated using the local background corrected intensities as provided by Thermo Fisher Scientific's iBright™ Analysis Software. Large values of IC_{50} represent a greater difference in binding affinity between the probe and competitor. However, since free protein is abundant in many of the experimental conditions used, we expect the IC_{50} to be an overestimate of the relative affinity of the probe. Instead, we provide the IC_{50} as a comparative measurement between similar experiments as discussed in Results and Discussion.

Multi-step alignment

Rather than using a PWM-based method for alignment, we aligned the 16-bp variable region of full-length reads to a predetermined set of 6-bp cores for Fhl1, and 7-bp cores for Fkh1, Fkh2 and Hcm1. The short *k*-mers allow for sequence diversity in the flanking regions while remaining long enough to accurately pinpoint true binding sites. This also allows us to specifically focus on identifying the contributions of positions flanking the core. The alignment is performed over a multi-step process summarized in Figure 2.

Identification of candidate cores. First, the relative enrichment of every *k*-mer was calculated as described in the original SELEX-seq protocol (10,11). We then performed a preliminary alignment of the longest set of *k*-mers which exhibited a high degree of coverage given a 100-count cutoff. For every dataset we tested, >95% of all unique 9-mers met this threshold, compared to an approximate 30% coverage of unique 10-mers. The preliminary alignment is performed using our recently published approach named Top-Down Crawl (TDC) (25). This framework was created specifically for the alignment of quantitative binding data, using highly enriched sequences as a template to explain the binding of those that are less enriched. To determine a set of candidate cores, this preliminary alignment is then trimmed

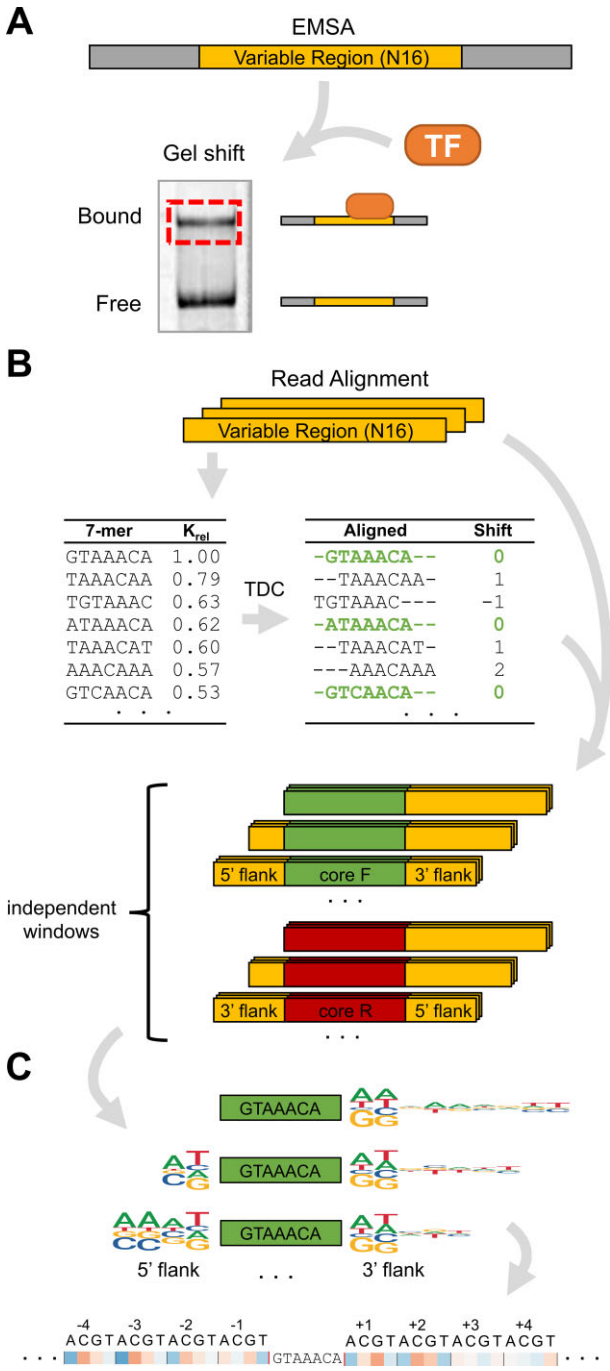


Figure 2. Experiment and analysis overview. (**A**) A library containing a 16-bp fully randomized region is incubated with the TF of interest and run on a non-denaturing polyacrylamide gel. The bound DNA is then extracted, purified, and amplified. Aliquots are sent for sequencing or used for additional rounds of selection. (**B**) Enriched 7-mers are identified and aligned as described in Materials and Methods to identify core sequences to use for alignment of the full 16-mers. Each 16-mer containing a hit to exactly one core sequence is assigned to a group based on where the core sequence begins. (**C**) For each position outside of each core, the relative enrichments between gapped 8-mers are used to calculate $\Delta\Delta G/RT$. The averages across all windows are plotted in a condensed view to allow for easier comparison across cores.

to the 6–7-bp region covering the canonical binding sites GTAAACA for Fkh1, Fkh2 and Hcm1, and GACGCA for Fhl1. Unique sequences falling within this region are treated as candidate cores to be used in the next step, which we refer to as iterative reprioritization. This allows us to study flanking positions using a minimal set of core sequences.

Reprioritization of candidate cores. To avoid complications resulting from combinatorial effects between multiple binding sites, we restrict our analysis to reads that only align to one core. This is also key to ensuring that observed flanking preferences are acting to modulate the given core rather than creating additional cores. However, this creates a trade-off between the number of cores we choose to analyze and the number of reads we can align. Therefore, rather than including the entire list of candidate cores aligned to the most enriched *k*-mer, we must prioritize a subset of these sequences. The most obvious way to do this is to simply rank them by their enrichment; however, it is important to consider that the observed enrichment of any given *k*-mer is a result of its activity when bound as a core in addition to its activity when bound as an optimal flank to an adjacent or overlapping core, since the context is lost during the counting process. Since we aim to prioritize sequences based solely on their contribution as a core, we developed an iterative approach to reprioritize candidate cores as described below.

First, the highest priority core was assigned as the most enriched *k*-mer as calculated previously. Next, all reads containing a match to that *k*-mer on the forward or reverse strand were removed from the dataset and *k*-mer enrichments were recalculated from the remaining reads. The next core to be prioritized was then the most enriched *k*-mer from this reduced set of reads, such that the *k*-mer belongs to our list of candidate cores determined previously. As before, all reads containing a match to that *k*-mer are subsequently removed, followed by recalculation of *k*-mer enrichments. This process is iteratively repeated until a sufficient number of cores have been identified. This framework allows us to identify sequences which are independently enriched, rather than being 'carried' by a more enriched overlapping *k*-mer.

To avoid overfiltering of our input data, our goal was to only include cores which identify binding sites with a high degree of confidence. We expect such cores to be enriched well above the background. Based on our observations, and previous work with SELEX-seq data, *k*-mer enrichment versus rank plots typically follow a negative exponential distribution (Supplementary Figure S5A). Based on this, we designed a stopping rule intended to identify the point at which enrichment levels off and is no longer significantly above the background. More specifically, we repeated the iterative process until the enrichment of the most recently prioritized *k*-mer is measured to be at least 95% of the average enrichment of the five previously removed *k*-mers. We used the average in order to smooth out noise over the measurements. We evaluated our framework's sensitivity to this stopping rule in Supplementary Analysis (Supplementary Figure S6).

After this process, every read from the full dataset was re-aligned to the set of prioritized cores, discarding reads with multiple hits. For comparison, we separately aligned the full dataset to the original list of candidate cores prioritized by raw enrichment. Shown in Supplementary Figure S7, the reprioritized list consistently allowed for the alignment of more reads compared to the original list. To simplify later comparisons between Fkh1, Fkh2, Hcm1 and Fhl1, we used the same set of core sequences for the alignment of every dataset. To generate this list, we took the union of the core sequences prioritized by each. The final list contained 49 sequences including ten 6-bp Fhl1-based cores and 39 7-bp cores.

Relative enrichment and free energy determination

Given a core sequence of length k, there will be $16–k+1$ windows across the 16-bp variable region where that k-mer can occur. Including occurrences on the reverse-complement strand, this number doubles. Because our alignment framework only allows for one core per read, each read can only be assigned to one window. Additionally, this makes the enrichment of any given sequence within one window independent of its enrichment in any other window (Figure 2B). To determine the enrichment of a core at a given window for a given round, its proportion is divided by its proportion at that same window in round zero (R0). The relative enrichment is then the measured enrichment for each sequence divided by the maximum across all sequences. As described in the original SELEX-seq protocol (10,11), the r^{th} root of the relative enrichment, with r representing the number of rounds of selection that have taken place, represents a close approximation of the true relative affinity (10).

Differences in binding affinity can also be represented as $\Delta\Delta G/RT$ by taking the negative natural log of the relative affinity (26) (Figure 3). This represents the difference in the free energy released by binding scaled by $1/RT$. In this case, a more positive value represents a less favorable binding interaction relative to the most enriched core. Analysis of $\Delta\Delta G/RT$ rather than relative affinity also helps accentuate changes between moderate-to-low affinity binding sites. The measurements were then averaged over all windows, removing outliers that are greater than the third quartile plus 1.5 times the interquartile range (IQR), or less than the first quartile minus 1.5 times the IQR.

To measure how flanking positions modulate binding site affinity, we then calculated $\Delta\Delta G/RT$ at every variable nucleotide position outside of the core. Given a core sequence of length k, each window will contain $16–k$ flanking positions across the 16-bp variable region. These can occur 5' and/or 3' of the core sequence depending on the window being analyzed. Assuming positions flanking the core contribute to binding independently of each other, their effects can be measured by looking at the relative enrichment between gapped $(k+1)$-mers, including one flanking position and a fixed 6-bp or 7-bp core. By looking at gapped $(k+1)$-mers, we are able to calculate the core-specific effects of flanking positions up to at least nine bp away from the core. This core-specific approach also allows us to avoid the dilution of flanking contributions if false binding sites are included during the alignment process, and to determine to what extent these contributions are dependent on the core.

Like the core, we measure flanking contributions in terms of $\Delta\Delta G/RT$. For a given window, a position frequency matrix (PFM) is generated by counting the occurrence of each bp at every variable nucleotide position outside of each core. These are the gapped $(k+1)$-mer counts mentioned previously. The enrichment of each bp is then determined in a position specific manner by dividing the frequency of each bp by its frequency observed in R0 at that same position (Supplementary Figure S8A). The enrichment of each bp is divided by the mean enrichment at every position to get relative enrichments. As described in the original SELEX-seq paper (10), the relative enrichment of samples collected after r rounds of selection is only equivalent to the relative affinity if the rth root is taken. However, this scaling factor may vary slightly depending on the efficiency of separating bound and unbound fractions. Instead, we estimate the scaling factor by dividing the mean log relative enrichment of cores from R2 by that from R1. This scaling factor was equal to 1.89, 1.97, 1.76 and 1.85 for Fkh1, Fkh2, Hcm1 and Fhl1, respectively. The scaled relative enrichments are then converted to $\Delta\Delta G/RT$ by applying the negative natural log. Positive values indicate bp that are more disruptive to binding than on average, and negative values indicate bp which facilitate binding. We represent the values using a heat map instead of a traditional motif logo in order to facilitate comparisons between independent windows and cores (Figure 2C). This entire process is repeated for every window along the 16-bp variable region for every core (Figure 4A). The $\Delta\Delta G/RT$ measurements are then averaged across all windows (Figure 4B).

Analysis of ChIP-exo data

In previously published work, researchers performed ChIP-exo experiments targeting Fkh1 in *S. cerevisiae* (27). A merged set of identified peaks are provided in Supplementary Data. We restricted our analysis to regions spanning 50 bp upstream and downstream of each peak's center. Across all these regions, we counted the total number of unique occurrences of each of our 7-bp cores and divided them by the total to obtain the relative frequency of each core. The same was performed across the entire SacCer3 reference genome to get a background frequency of each core. The genomic values were then multiplied by the relative enrichments, as predicted using the exponential function of the $-\Delta\Delta G/RT$ values, and divided by the sum across all cores to get the predicted relative frequencies of each core. These values were then compared with the observed values calculated previously to obtain the Pearson correlation and Mean Squared Error (MSE). A similar process was performed to compare the observed and predicted relative frequencies of bp at the four positions 5' and two positions 3' of the GTAAACA core.

Alternatively, we generated predictions using PFMs derived from BEESEM, a method for the generation of binding motifs from SELEX-seq data (12). A PFM was used to calculate the expected distribution of our cores by multiplying bp probabilities and dividing by the sum of proba-

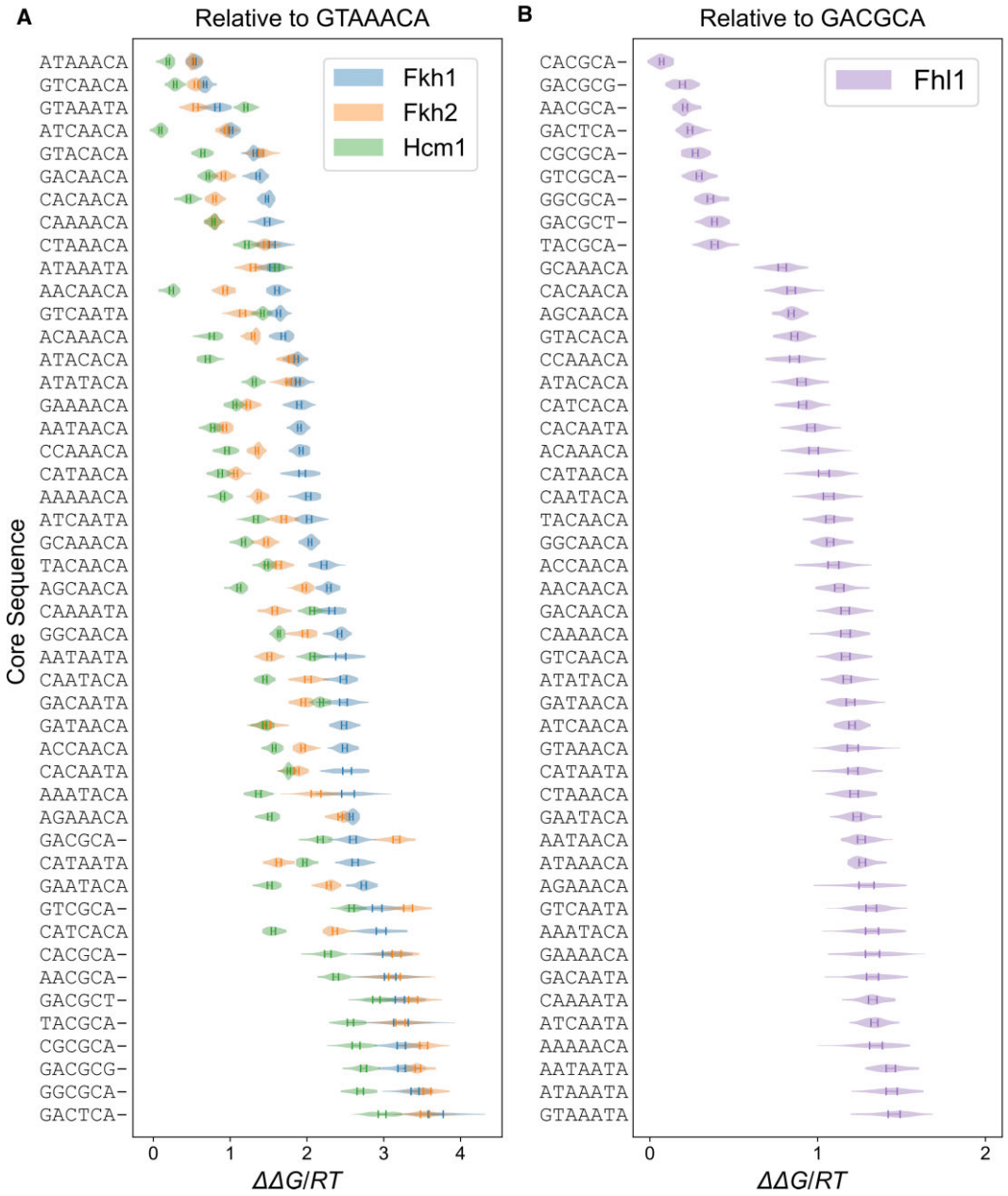


Figure 3. Violin plot of $\Delta\Delta G/RT$ estimates from every window resulting from two rounds of SELEX-seq for 48 selected core sequences relative to the most enriched core, (**A**) GTAAACA for Fkh1, Fkh2 and Hcm1, and (**B**) GACGCA for Fhl1. Larger values indicate a greater disruption to binding relative to the reference. Error bars show 95% confidence intervals.

bilities across all cores. Due to the high computational cost, BEESEM was trained using only 10% of the R1 reads with 10% of the R0 reads as background. For predicting the relative core frequencies, BEESEM was trained with the seed, GTAAACA, and was provided the sequences of the fixed adapters used in our library design (Supplementary Figure S9). Even with the reduced input size, this process required nearly 11 hours of compute time and 52 GB of RAM on a 16-processor compute node. For flanking positions, we generated a 13-bp motif covering all six flanking positions of interest, which required 122 GB of RAM, and a similar compute time. This was done using the seed, AAAAGTAAACAAA (Supplementary Figure S9). The motif was then used to predict the observed flanking relative frequencies as described previously.

RESULTS AND DISCUSSION

Core-based alignment of full-length reads

To explore distal positions flanking the core binding site, we used a library with a 16-bp variable region. With a *k*-mer length of 16 bp, approximately 4.3 billion unique sequence permutations are possible. With the size of a typical sequencing run, only a small subset of these permutations can be captured and the count for each is often too low to

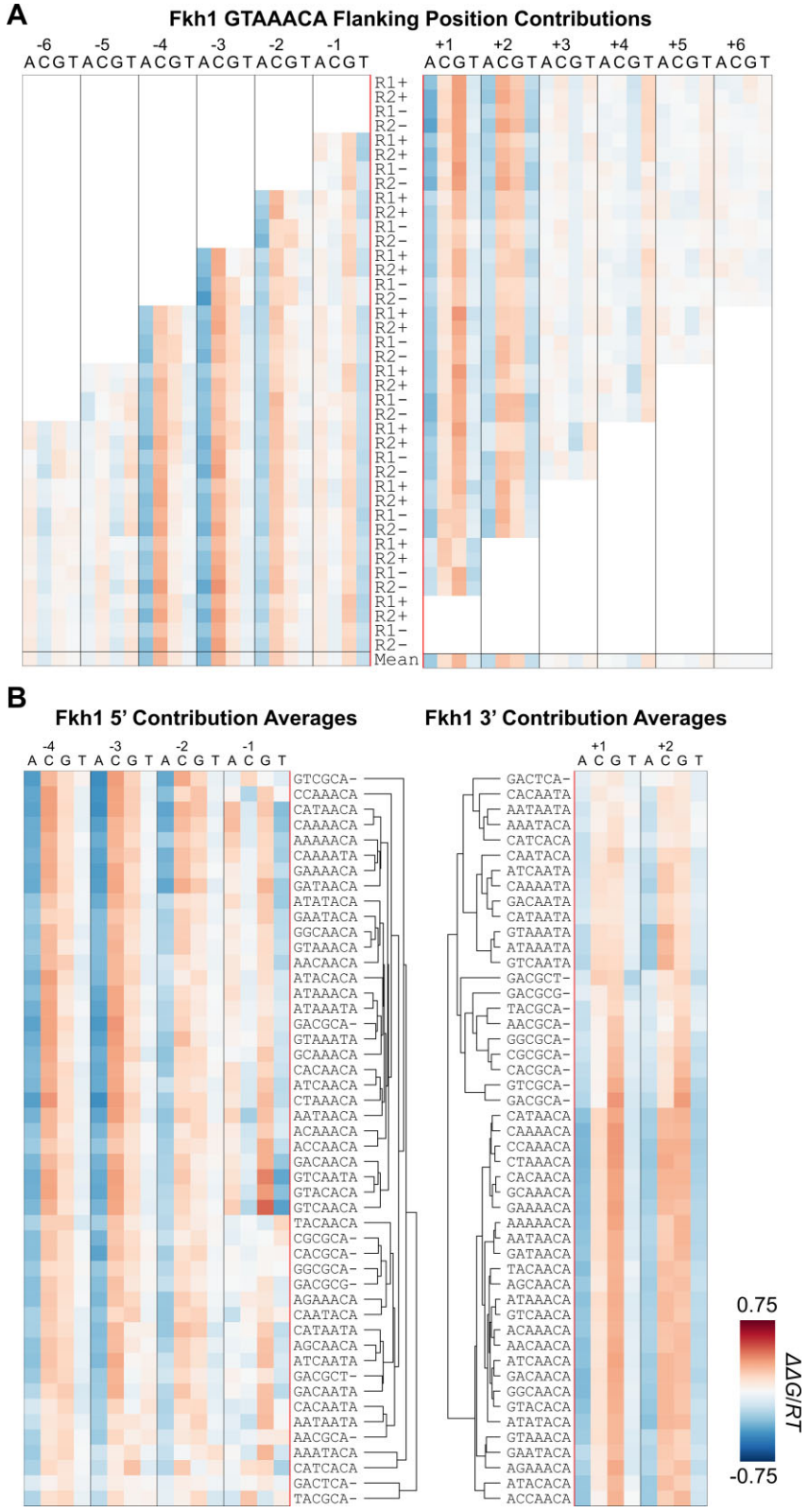


Figure 4. (**A**) Graphic representation of $\Delta\Delta G/RT$ for every possible bp at positions outside of the core binding site, given a core sequence of GTAAACA. Rows represent the 40 independent sets of aligned 16-mers, each with independent measurements. Larger values indicate greater destabilization of binding relative to other possible bp at that position. Contributions appear to be highly consistent across samples. (**B**) $\Delta\Delta G/RT$ measurements for each aligned core averaged over the 40 independent sets of aligned 16-mers. Rows are clustered with the UPGMA algorithm using Manhattan distance as the metric. The measurements suggest that flanking contributions are largely independent of the core.

provide a meaningful measurement of binding affinity. By looking at the enrichment of shorter *k*-mers within the randomized regions, we can collapse the number of possible permutations several fold and greatly increase the number of occurrences for each. This creates a tradeoff between *k*-mer length, noise, and sequence coverage given a 100-count threshold (Supplementary Figure S10).

In previous studies analyzing SELEX-seq data (10,28), an enrichment table is determined by counting every *k*-mer across a sliding window and dividing it by the expected count according to a Markov Model generated from R0 *k*-mer counts. Although a sliding window increases the number of observed counts for each *k*-mer, it comes at the cost of losing positional information amongst the original reads. Additionally, we found that R0 biases are dependent on the position, so representing the bias universally with a position-agnostic Markov model may not be appropriate (Supplementary Figure S8). This is particularly true for larger *k*-mers, from which positional biases may be compounded. Furthermore, by counting *k*-mers using a sliding window, smaller subsequences are counted multiple times, so *k*-mer counts cannot be considered independent of each other. For example, if the most enriched core is GTAAACA, then we know that *k*-mers of the form TAAACAN will also be highly enriched, even if those sequences are not intrinsically able to promote binding in any other context. Consequently, any given *k*-mer may be enriched due to its activity as a strong core binding site, optimal flanking sequence, or a combination of both. By limiting the number of cores per sequence to one, as described in Materials and Methods, we can be confident that the observed core is acting as the most likely binding site amongst each read, and that flanking positions will be aligned accordingly. Additionally, if multiple binding sites were to be permitted, then flanking positions may be biased to prefer the creation of additional cores, rather than by modulating the affinity of the aligned core.

From data shown in Supplementary Figure S1, we know that the most conserved region of the binding site, referred to as the 'core,' is 7-bp long for most FOX proteins. Considering a 7-bp core, and a 16-bp randomized region, there are a total of 20 different positions in which the core may reside, including ten on the forward strand and ten on the reverse strand (Figure 2B). Assuming the data reflects sequence-specific binding, we expect every 16-mer to have at least one predominant binding site. Since the core sequence is the most influential region in determining binding, we sought to create a table of putative core sequences that could be used to identify and align the binding sites for the largest number of 16-mers. Details of how this table is generated are described in Materials and Methods.

We ultimately decided to use a list of 49 core sequences including ten 6-bp Fhl1 cores (Supplementary Table S4). For the R1 data, we were able to align 25.9%, 28.6%, 28.0%, and 13.2% of the reads bound by Fkh1, Fkh2, Hcm1, and Fhl1, respectively. We found that 63.4%, 60.0%, 59.3% and 79.5% of the reads bound by Fkh1, Fkh2, Hcm1 and Fhl1, respectively, were removed because no core was detected, meaning that these reads would be uninformative of core-specific binding parameters. For R2, we aligned a more substantial 63.3%, 63.7%, 46.1%, and 19.9% of the reads bound by Fkh1, Fkh2, Hcm1 and Fhl1, respectively.

Unlike PWM-based methods, core-based alignment does not assume any interdependencies within the core and reveals many high-affinity sequences with surprisingly high dissimilarity from the most enriched core. Perhaps most importantly, this framework aligns full-length reads which enables the analysis of nucleotide positions at least nine positions away from the core on either side. For comparison to traditional methods, we generated a PWM weighting of each core by its relative enrichment. Using only our reduced set of 7-bp cores for Fkh1, Fhh2 and Hcm1, and the 6-bp cores for Fhl1, the generated PWMs (Supplementary Figure S11) are highly similar to the uPBM-derived motifs published previously (Supplementary Figure S1B) (29).

Core binding sites exhibit interdependencies and shape preferences

For every alignment window, $\Delta\Delta G/RT$ was calculated for each core sequence relative to the most preferred core sequence, as described in Materials and Methods. Since each window consists of an independent set of sequences that were selected by the protein independently, they can be treated as independent samples. For a core length of seven bp, each sequence can be measured across 40 samples, including ten samples per strand per round. In the initial libraries, aligned reads are distributed similarly across windows, but this changes after selection, particularly for Fkh1 and Fkh2 which may be more sensitive to fixed adapter positions (Supplementary Figure S12). $\Delta\Delta G/RT$ measurements for our set of cores exhibit little variability across independent samples, following normal distributions, resulting in narrow confidence intervals for the averaged values (Figure 3, Supplementary Table S4). Measurements spanned a $\Delta\Delta G/RT$ of 3.7, 3.5 and 3.0 for Fkh1, Fkh2, and Hcm1, respectively. For Fhl1, binding was generally less specific, spanning a $\Delta\Delta G/RT$ of 1.5 across all cores, and 0.4 across 6-bp cores. One core, GTATACA, was excluded from analysis since we were concerned it may be filtered at a much higher rate. Adding just one thymine to the first position 5' of this core results in a palindromic sequence containing two cores. Although our list of cores only included one such core, our alignment analysis script provided on GitHub will automatically detect palindromic cores and cores which are one flanking mutation away from the creation of an additional core and remove them from downstream analysis. Nevertheless, it is important to include these cores in the previous alignment step since they can still serve as valid cores to indicate overlapping binding sites.

For Fkh1, Fkh2 and Hcm1, the most enriched sequence was GTAAACA. It becomes clear upon further inspection that nucleotide contributions in the core do not appear to contribute independently. Using Fkh1 as an example, we considered the four core sequences shown in Figure 5. If only the third position is mutated from an A to a C, we see a $\Delta\Delta G/RT$ value of 0.67. If we mutate the second position from a T to an A, we see a large $\Delta\Delta G/RT$ value of 1.90 If the effects of these mutations acted independently, then mutating both positions would result in a $\Delta\Delta G/RT$ value of 2.57. Instead, we observe a relatively modest effect of 1.36,

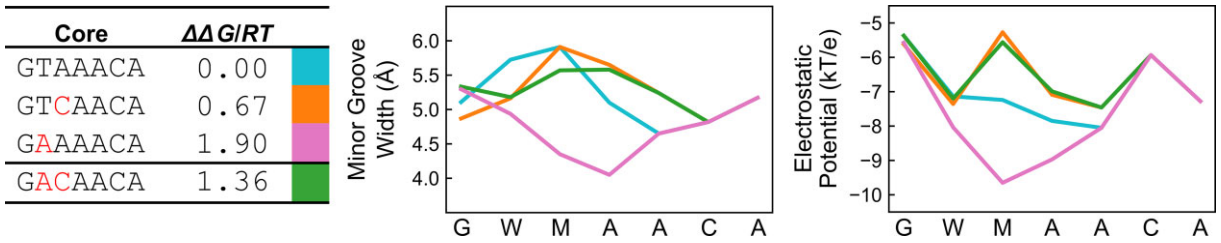


Figure 5. Comparison of the Fkh1 $\Delta\Delta G/RT$ measurements for different mutations of the core DNA target sequence with respect to the reference sequence GTAAACA, exemplifying a non-additive relationship. Out of this set of sequences, the three most enriched cores exhibit similar minor groove width (MGW) and electrostatic potential (EP) profiles as predicted using DNAshapeR (IUPAC: W = A/T; M = A/C).

nearly half of its expected value, and even lower than the single mutation from T to A. This phenomenon is repeated throughout the table of cores analyzed, exemplifying complex interdependencies between positions within the core (Supplementary Figure S13). This further suggests that a PWM-based representation would not accurately describe binding preferences in this system.

To explain these interdependencies, we examined DNA shape features for the four core sequences discussed previously. The DNAshapeR method (30) predicts several DNA shape features using a 5-bp sliding window with values based on previously run Monte Carlo simulations of free DNA (31). Although DNAshapeR is able to predict 13 different DNA shape features, we are most interested in minor groove width and electrostatic potential due to their biophysical origin arising from the identity of multiple bp beyond dinucleotides and their potential to influence interactions with charged residues within the winged regions of the DBD (32). Although, the GAAAACA core is only one mutation away from the reference core, it is disadvantaged in Fkh1 binding compared to GACAACA, which contains an additional mutation at the third position. Although this second mutation may disrupt some preferred bp-specific contacts, it appears to increase the minor groove width and electrostatic potential of the DNA so that it is more similar to the preferred reference (Figure 5). It is worth noting that this secondary mutation disrupts what would be a 4-bp A-tract, a feature known to cause intrinsic DNA bending (33).

Alternatively, we evaluated the use of a position specific affinity matrix (PSAM) for the prediction of $-\Delta\Delta G/RT$ for every 7-bp core containing more than one mutation from the reference, including 30 enriched sequences. The PSAM is generated using the $-\Delta\Delta G/RT$ for every sequence that is one point away from the reference, GTAAACA. The PSAM-based predictions were only weakly correlated ($r^2 = 0.34$; Supplementary Figure S14) with the measured values, with observed values often much larger than predicted. This further emphasizes the importance of using full-length 7-mers to define a list of cores to use for alignment, rather than a PWM-based simplification.

Core binding sites exhibit differing selectivity

The $\Delta\Delta G/RT$ measurements from every window for each core are displayed in a violin plot in Figure 3, and all averages are provided in Supplementary Table S4. To identify any sequence-specific difference in DNA binding specificity between Fkh1 and Fkh2, we plotted $-\Delta\Delta G/RT$ of every 7-bp core and color-coded each point by the bp identity at each nucleotide position along the 7-mer (Supplementary Figure S15A). Although only a few positions exhibited variability in sequence, we noticed a sequence-dependent shift of approximately 0.5 units at the second position of the core. At this position, Fkh2 exhibited a greater tolerance for adenine relative to Fkh1. Comparing Hcm1 with Fkh2, we find that Hcm1 is less tolerant of a thymine at position 6 of the core, with the line of best fit shifted by about 0.8 units compared to cytosine (Supplementary Figure S15B). Across these three homologs, the base-contacting residues, based on co-crystal structures for other FOX proteins, appear to be highly conserved. This suggests that the observed differences in preferences may be a result of a higher-order feature such as DNA shape.

Experimental validation of core preferences

Based on our SELEX-seq experiment, we found that the double mutant GACAACA to be bound preferentially over the single mutant GAAAACA. This is a particularly interesting example because it confirms the importance of interdependencies within the core. We confirmed this finding using two competitive binding assays, in which we alternate the labeled probe and the unlabeled competitor. In every case, cores are surrounded by an optimal flanking sequence context (Supplementary Table S1). In order to see binding to the probes, a 4-fold excess of protein was included relative to a fixed amount of probe.

In the first experiment, we used GACAACA as the labeled probe and GAAAACA as the unlabeled competitor (Supplementary Figure S16A). We found the IC_{50} to be around 4, meaning that a four-fold excess of competitor was needed to displace nearly 50% of the labeled probe. However, because we are using more protein than probe in the starting reaction, we expect the IC_{50} to be an overestimate of the relative affinity of the probe due to the presence of free protein, which can be bound by the competitor prior to direct competition. Based on our measurements, the $-\Delta\Delta G/RT$ values differ by 0.536, corresponding to a nearly 1.7-fold change in binding affinity. Because the affinities are similar, we can perform the reciprocal experiment without drastically changing input concentrations. In this case, we used GAAAACA as the labeled probe, with GACAACA as the competitor, and found the IC_{50} to be around 2. These experiments confirm our original expectations, since GACAACA was harder to displace than GAAAACA.

In looking at monomeric differences in binding specificity within the core, we found Fkh2 to tolerate adenine at the second position to a greater extent than Fkh1 (Supplementary Figure S15A). To confirm this, we performed a competitive binding assay using GTAAACA as the labeled probe, and GAAAACA as the unlabeled competitor, using either Fkh1 or Fkh2 as the DNA-binding protein. We found the IC_{50} to be about 16 for Fkh1, and between 4 and 8 for Fkh2 (Supplementary Figure S16B). As expected, this experiment confirmed that Fkh2 exhibits an increased tolerance for the thymine to adenine mutation at the second position of the core compared to Fkh1.

Flanking sequence contributions across differing cores

To investigate the effects of flanking sequences on binding, we plotted $\Delta\Delta G/RT$ for every possible bp outside of the core for every window. This type of analysis assumes that edge positions contribute to binding independently of each other. The validity of this assumption is evaluated using Multiple Linear Regression (Supplementary Analysis, Supplementary Figure S17). For Fkh1, Fkh2 and Hcm1, measurements were not only consistent across different windows (Figure 4A), but also highly consistent across different cores (Figure 4B, Supplementary Figures S18–S21). This suggests that our multi-step approach to identifying core sequences and aligning reads is valid, since flanking preferences are also aligned. Alternatively, we modified the BET-seq (34,35) framework to predict flanking contributions using a deep learning framework based on DeepBind (14,36) (Supplementary Analysis, Supplementary Figure S22- S23). In this case, flanking contributions were less consistent across cores and were apparently attenuated for moderate-to-low affinity cores, suggesting that our framework is better able to detect these flanking contributions, likely due to the single-core requirement during alignment.

We found it interesting that flanking contributions were consistent even for cores of differing length. The decision to include the Fhl1-based cores in the alignment of all homologs was supported by previous crystallographic evidence of FOXN3 that showed conserved amino acid contacts between a 6-bp Fhl motif, GACGCA (37), and the standard 7-bp Fkh motif (38). As described by the authors, the three-dimensional DNA shape of the shorter core was altered in order to align two 'registration positions' at the edges of each motif (38). For Fkh1, in particular, the 6-bp and 7-bp cores exhibited similar flanking preferences both upstream and downstream of the core, further supporting the discovery of registration positions that could align contacts flanking the core (Figure 4B).

In most cases, we found flanking preferences for 6-bp cores to be distinctly clustered from the 7-bp cores (Supplementary Figures S18–S21). We therefore wanted to determine whether an alternate method of aligning reads with Fhl1-based cores could better align flanking preferences. This is explored by comparing the gapped 6-bp alignment, with a 7-bp alignment including one bp (A/C/G/T) 5' or 3' of the 6-bp cores. For demonstration, we only consider the two strongest Fhl1 cores, GACGCA and CACGCA, denoted SACGCA (Supplementary Figure S24). This was applied to the Fkh1 dataset which shows the greatest level of sensitivity to positions 5' and 3' of the core. When we included an extra bp 5' of the cores, we see a corresponding shift in the 5' flanking preferences, with the 3' preferences relatively unchanged (Supplementary Figure S24A). When the extra bp is 3' of the cores, we see similar preferences at the first flanking position 3' of the core, but altered preferences at the second position, relative to the other 7-bp cores (Supplementary Figure S24B). These shifts confirm that the flanking preferences of Fhl1-based cores are best aligned when treated as gapped 6-mers rather than 7-mers (Supplementary Figure S18).

We describe the sensitivity to flanking positions as the difference in $\Delta\Delta G/RT$ between the most and least favored bp at each position, using the averages over all cores (Figure 6). For Fkh1 and Fkh2, we found the largest flanking contributions at the four positions 5' of the core and two positions 3' of the core. For Hcm1, only one position 3' of the core was found to have a similarly large impact on binding. It is interesting to see such a stark difference in sensitivity to flanking positions even though all three homologs share similar preferences for the core. For Fhl1, flanking positions did not appear to contribute significantly to binding. Although the 6-bp cores are far more enriched than those that are 7-bp, the overall difference in $\Delta\Delta G/RT$ between the best and worst core is on a much smaller scale than we see across Fkh1, Fkh2, and Hcm1 cores. This suggests that Fhl1 exhibits less specific binding to its preferred cores, which could impact alignment by allowing 'false' binding sites to be aligned.

Looking at the two nucleotide positions 3' of the core, referred to as the +1 and +2 positions, we see a slightly diminished range in binding affinities by Fkh1 when the sixth position of the core is a thymine, rather than a cytosine (Figure 4B). Likewise, we find that the +1 position appears to have a smaller impact on the predicted electrostatic potential at positions 5 and 6 of the core when there is a thymine at the sixth position, using GTAAA(C/T)A as an example (Supplementary Figure S25). This may explain why several core sequences that contain a thymine at position 6 appear to be less sensitive to variations at the +1 position. More broadly, this observed preference for a more negative electrostatic potential in the 3' flanking region is consistent with the hypothesis that positively charged residues in the winged regions of Fkh proteins act to stabilize binding to the DNA. Compared to Fkh1, Fkh2 exhibits a substantial reduction in specificity at the +1 and +2 positions across all cores (Figure 6). This observation suggests reduced sensitivity to electrostatic potential by Fkh2 in this region.

Structural analysis of flanking sequence contributions

To further investigate these observations, we analyzed structures of the DBDs and their surrounding residues predicted using AlphaFold2 (39,40). To understand where contacts may lie relative to the DNA, predicted structures were aligned to a previously published Human FOXK2 co-crystal structure (PDB ID: 2C6Y) and the FOXK2 protein was removed (Supplementary Figure S26). In all structures, wing 1 is in the proximity of the 3' flank relative to the core GTAAACA. Fkh1 and Fkh2 both contain long, structurally similar wings enriched with three

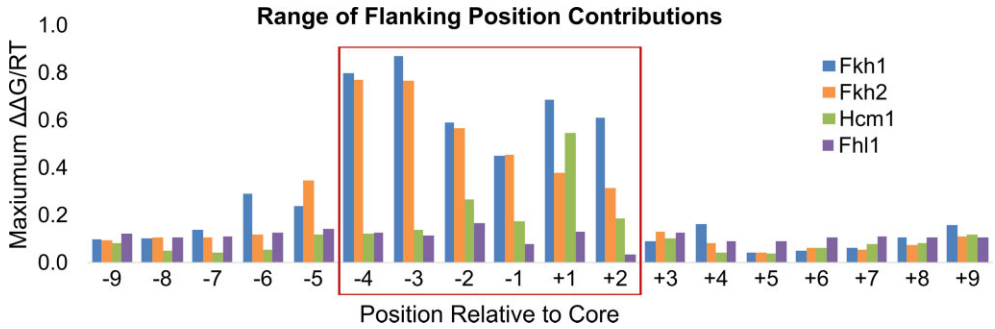


Figure 6. Maximum $\Delta\Delta G/RT$ values averaged across all cores for each flanking position. The red box highlights the positions that had the largest impact on Fkh1 binding.

positively charged residues each, including arginines and lysines. These residues are particularly interesting because they are known to interact with electrostatically negative minor groove surfaces (41). In addition, wing 1 of Fkh2 contains one negatively charged residue, glutamate, which may have an opposing effect. This small change could explain why Fkh2 was found to be less sensitive to 3' modulations compared to Fkh1. Wing 1 of Hcm1 is shorter and contains two lysine residues and a single negatively charged residue, aspartate. Lastly, Fhl1 has the shortest wing, and exhibits undetectable flanking preferences.

Looking at potential contacts 5' of the core, we highlight the C-terminal region beyond the final β-sheet of the DBD. Interestingly Fkh1 and Fkh2 both exhibit a helix-loop-helix structure which may act to stabilize a secondary wing which could contact the minor groove. Between the two helices, both proteins contain four positively charged residues and no negatively charged residue. This is an interesting finding given the extensive preferences identified 5' of the core. Alternatively, Hcm1 and Fhl1 exhibit minimal sensitivity to mutations 5' of the core. In the structure for Hcm1, we observe a C-terminal helix which interacts with an N-terminal helix to pull the disordered region farther from the minor groove. Although this region contains four positively charged residues, it also contains two negatively charged residues. For Fhl1, a rigid helix is presumably formed, which may restrict movement of positively charged residues into the minor groove. Although we are skeptical of the complete validity of this structural feature, we found it interesting that the prediction was made with a high degree of confidence by the AlphaFold algorithm (39). While these insights provide a potential explanation of our observed flanking preferences, validation is necessary to confirm the accuracy of the predicted structures.

Experimental validation of flanking contributions

Based on the assumption of independence between nucleotide positions, mutating the flanking positions to their most unfavorable bp results in an average increase in $\Delta\Delta G/RT$ of 4.0 across all cores for Fkh1 and 3.24 for Fkh2. This corresponds to a roughly 55-fold and 26-fold reduction in binding affinity, respectively. Alternatively, mutating the core to the lowest affinity sequence included in the alignment, an Fhl1-based core, resulted in a $\Delta\Delta G/RT$ of 3.68 for Fkh1 and 3.57 for Fkh2.

We experimentally validated the importance of the flanking positions using a competitive binding assay to compare the binding of Fkh1 to the core, GTCAACA, surrounded by either optimal or suboptimal flanking nucleotides. Starting with a 2.4-fold excess of protein to probe, and using the sequence with the optimal probe, we measured an IC_{50} around 32 (Supplementary Figure S16C). Out of all our competitive binding assays, this is by far the largest value measured. This aligns with our expectations well and further emphasizes the importance of including flanking positions in discriminating the affinity of identified binding sites. In this case, a reciprocal experiment was not feasible since it would require an extreme excess of protein in order to visualize binding to the suboptimal probe.

Applications to *in vivo* binding site prediction

To evaluate whether our SELEX-seq experimental data and findings could be applied to *in vivo* binding site prediction, we analyzed the peaks reported from a previously published ChIP-exo dataset targeting Fkh1 and Fkh2 (27). Because Fkh2 interacts with the cofactor Mcm1 *in vivo* (42), we expected altered preferences compared to our controlled SELEX-seq experiment. This is further supported by looking at the overlap between the provided Fkh2 ChIP-exo motif and a previously published motif for Mcm1 provided by the Yeast Epigenome Project (43) (Supplementary Figure S27).

Using the set of all peaks identified by the original study, we extracted all non-overlapping sequences ±50 bp from the center of each peak and counted the total number of occurrences of every core from our previously defined set. We also counted the relative frequency (p) of each core across the genome to use as a background. Using our alignment-based $\Delta\Delta G/RT$ measurements from the SELEX-seq experiment, we then calculated the expected relative frequency of every core given the genomic background, given selection by Fkh1. Alternatively, we used a BEESEM-derived motif to predict the expected relative frequencies of the same set of cores, given the same genomic background. Since the BEESEM-derived motif would not be able to provide dependable predictions for the 6-bp Fhl1 motifs, they were removed from subsequent comparisons. Furthermore, natural logarithms of the relative frequencies were taken for comparison with the ChIP-exo observations in order to better compare differences across a wide range of affinities. We

found the observed values to be remarkably well-correlated with and on a similar scale as our $\Delta\Delta G/RT$ based predictions (Pearson $r = 0.84$, MSE $= 0.25$; Supplementary Figure S28A). Comparatively, the BEESEM-based predictions were only modestly correlated with the observed values ($r = 0.61$, MSE $= 5.23$; Supplementary Figure S28B). We found that BEESEM underestimates the enrichment of many core sequences, as was the case for the PSAM-based analysis described previously. This further supports the importance of considering interdependencies within the core both *in vitro* and *in vivo*. It should also be noted that, in total, we identified a total of 798 binding sites across the ChIP-exo regions, representing a much smaller sample size than what can be collected *in vitro*.

To investigate flanking preferences, we focused on positions surrounding the GTAAACA core, of which 117 sites were identified. For four positions 5' and two positions 3' of the core, we calculated the natural log of the relative frequency of each bp and compared it to predictions using the alignment-based measurements of $\Delta\Delta G/RT$ or using a BEESEM-derived motif. As before, we found the observed values at the six flanking positions to be well-correlated with and on a similar scale as our $\Delta\Delta G/RT$ based predictions ($r = 0.79$, MSE $= 0.16$; Supplementary Figure S28C). The BEESEM-based predictions were found to have weaker correlations ($r = 0.66$, MSE $= 0.40$; Supplementary Figure S28C). Together, these findings show that our quantitative measurements collected *in vitro* can be used to predict the enrichment of binding sites found *in vivo*.

CONCLUSIONS

With our multi-step alignment approach, we have been able to thoroughly explore how flanking nucleotide positions contribute to binding site affinity in a way that previous approaches cannot (reviewed in (44)). By focusing on the alignment of full-length reads, we have revealed patterns of flanking nucleotide preferences that are highly consistent across independent nucleotide windows and across drastically different cores. Although the impact of each nucleotide position may be small, their combined effect can greatly impair binding to a putative DNA target. Additionally, these contributions are often lost in traditional PWM-based analytical frameworks, for which the alignment of false binding sites can dilute their effect. By using a restricted set of cores, we can pinpoint the effects of mutating flanking positions without assuming independence between positions of the core.

In this study, we explored the binding preferences of all four *Saccharomyces cerevisiae* forkhead TFs, Fkh1, Fkh2, Hcm1 and Fhl1, revealing small-scale, but consistent differences that have not been characterized previously. Including flanking contributions, we were able to expand the binding sites of Fkh1 and Fkh2 to cover a 13-bp window including four bp 5' of the core and two bp 3' of the core. Alternatively, Hcm1 and Fhl1 only exhibited minor flanking preferences, despite similarities in the DBD. The framework can be adapted to fully capture the impact of flanking positions for other TFs whose binding has been measured using next generation sequencing. In this work, we selected a list of candidate cores using Top-Down Crawl, which has recently been published as a method applied to other datasets (25), and iterative reprioritization. The computational analysis framework is flexible and can be applied to any list of cores desired by the researcher, even when those cores are not of the same length.

DATA AVAILABILITY

SELEX-seq data was collected for Fkh1, Fkh2, Hcm1, and Fkl1, as described in Materials and Methods, and submitted to the Gene Expression Omnibus (GEO) with accession number GSE178811. A small-scale test run targeting Fkh1 was included as well as the large-scale runs discussed throughout this work.

Workflow and scripts to perform the multi-step alignment approach as well as supplementary analyses can be found at https://github.com/bhcooper/multi-step-align and https://doi.org/10.5281/zenodo.7865759. All steps of the approach can be performed by following the publicly available workflow provided. The Top-Down Crawl alignment method is publicly available at https://topdowncrawl.usc.edu (25).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Peter Z. Qin, Tsu-Pei Chiu, and members of the Rohs lab for valuable suggestions.

FUNDING

National Institutes of Health [R35GM130376 to R.R., R01GM065494 to O.M.A.]; Human Frontier Science Program [RGP0021/2018 to R.R.]; USC Provost Fellowship [to B.H.C.]; USC Dornsife Bridge Institute Catalyst grant [to R.R.]; DNA sequencing was performed in the USC Norris Cancer Center supported by the National Cancer Institute [P30CA014089]. Funding for open access charge: National Institutes of Health [R35GM130376 to R.R.].
Conflict of interest statement. None declared.

REFERENCES

1. Golson,M.L. and Kaestner,K.H. (2016) Fox transcription factors: from development to disease. *Development*, **143**, 4558–4570.
2. Lalmansingh,A.S., Karmakar,S., Jin,Y. and Nagaich,A.K. (2012) Multiple modes of chromatin remodeling by Forkhead box proteins. *Biochim. Biophys. Acta.*, **1819**, 707–715.
3. Zhou,T., Shen,N., Yang,L., Abe,N., Horton,J., Mann,R.S., Bussemaker,H.J., Gordân,R. and Rohs,R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 4654–4659.
4. Yang,L., Orenstein,Y., Jolma,A., Yin,Y., Taipale,J., Shamir,R. and Rohs,R. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.*, **13**, 910.
5. Tanay,A. (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*, **16**, 962–972.
6. Crocker,J., Abe,N., Rinaldi,L., McGregor,A.P., Frankel,N., Wang,S., Alsawadi,A., Valenti,P., Plaza,S. and Payre,F. (2015) Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*, **160**, 191–203.

7. Hume,M.A., Barrera,L.A., Gisselbrecht,S.S. and Bulyk,M.L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Res.*, **43**, D117–D122.
8. Zhao,Y. and Stormo,G.D. (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.*, **29**, 480–483.
9. Ostrow,A.Z., Nellimoottil,T., Knott,S.R., Fox,C.A., Tavare,S. and Aparicio,O.M. (2014) Fkh1 and Fkh2 bind multiple chromosomal elements in the S. cerevisiae genome with distinct specificities and cell cycle dynamics. *PLoS One*, **9**, e87647.
10. Slattery,M., Riley,T., Liu,P., Abe,N., Gomez-Alcala,P., Dror,I., Zhou,T., Rohs,R., Honig,B. and Bussemaker,H.J. (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270–1282.
11. Riley,T.R., Slattery,M., Abe,N., Rastogi,C., Liu,D., Mann,R.S. and Bussemaker,H.J. (2014) SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Methods Mol. Biol.*, **1196**, 255–278.
12. Ruan,S., Swamidass,S.J. and Stormo,G.D. (2017) BEESEM: estimation of binding energy models using HT-SELEX data. *Bioinformatics*, **33**, 2288–2295.
13. Rube,H.T., Rastogi,C., Feng,S., Kribelbauer,J.F., Li,A., Becerra,B., Melo,L.A.N., Do,B.V., Li,X., Adam,H.H. *et al.* (2022) Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. *Nat. Biotechnol.*, **40**, 1520–1527.
14. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
15. Yuan,H., Kshirsagar,M., Zamparo,L., Lu,Y. and Leslie,C.S. (2019) BindSpace decodes transcription factor binding signals by large-scale sequence embedding. *Nat. Methods*, **16**, 858–861.
16. Asif,M. and Orenstein,Y. (2020) DeepSELEX: inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs. *Bioinformatics*, **36**, i634–i642.
17. Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A. (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.*, **10**, 2997–3011.
18. Ruan,S. and Stormo,G.D. (2018) Comparison of discriminative motif optimization using matrix and DNA shape-based models. *BMC Bioinf*, **19**, 86.
19. Rube,H.T., Rastogi,C., Kribelbauer,J.F. and Bussemaker,H.J. (2018) A unified approach for quantifying and interpreting DNA shape readout by transcription factors. *Mol. Syst. Biol.*, **14**, e7902.
20. Sharon,E., Lubliner,S. and Segal,E. (2008) A feature-based approach to modeling protein–DNA interactions. *PLoS Comput. Biol.*, **4**, e1000154.
21. Li,J., Dantas Machado,A.C., Guo,M., Sagendorf,J.M., Zhou,Z., Jiang,L., Chen,X., Wu,D., Qu,L., Chen,Z. *et al.* (2017) Structure of the forkhead domain of FOXA2 bound to a complete DNA consensus site. *Biochemistry*, **56**, 3745–3753.
22. Jin,C., Marsden,I., Chen,X. and Liao,X. (1999) Dynamic DNA contacts observed in the NMR structure of winged helix protein-DNA complex. *J. Mol. Biol.*, **289**, 683–690.
23. Clark,K.L., Halay,E.D., Lai,E. and Burley,S.K. (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature*, **364**, 412–420.
24. Tsai,K.-L., Huang,C.-Y., Chang,C.-H., Sun,Y.-J., Chuang,W.-J. and Hsiao,C.-D. (2006) Crystal structure of the human FOXK1a-DNA complex and its implications on the diverse binding specificity of winged helix/forkhead proteins. *J. Biol. Chem.*, **281**, 17400–17409.
25. Cooper,B.H., Chiu,T.P. and Rohs,R. (2022) Top-Down Crawl: a method for the ultra-rapid and motif-free alignment of sequences with associated binding metrics. *Bioinformatics*, **38**, 5121–5123.
26. Riley,T.R., Lazarovici,A., Mann,R.S. and Bussemaker,H.J. (2015) Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. *Elife*, **4**, e06397.
27. Mondeel,T.D.G.A., Holland,P., Nielsen,J. and Barberis,M. (2019) ChIP-exo analysis highlights Fkh1 and Fkh2 transcription factors as hubs that integrate multi-scale networks in budding yeast. *Nucleic Acids Res*, **47**, 7825–7841.
28. Abe,N., Dror,I., Yang,L., Slattery,M., Zhou,T., Bussemaker,H.J., Rohs,R. and Mann,R.S. (2015) Deconvolving the recognition of DNA shape from sequence. *Cell*, **161**, 307–318.
29. Zhu,C., Byers,K.J., McCord,R.P., Shi,Z., Berger,M.F., Newburger,D.E., Saulrieta,K., Smith,Z., Shah,M.V. and Radhakrishnan,M. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res*, **19**, 556–566.
30. Chiu,T.P., Comoglio,F., Zhou,T., Yang,L., Paro,R. and Rohs,R. (2016) DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, **32**, 1211–1213.
31. 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.
32. Azad,R.N., Zafiropoulos,D., Ober,D., Jiang,Y., Chiu,T.P., Sagendorf,J.M., Rohs,R. and Tullius,T.D. (2018) Experimental maps of DNA structure at nucleotide resolution distinguish intrinsic from protein-induced DNA deformations. *Nucleic Acids Res.*, **46**, 2636–2647.
33. Haran,T.E. and Mohanty,U. (2009) The unique structure of A-tracts and intrinsic DNA bending. *Q. Rev. Biophys.*, **42**, 41–81.
34. Le,D.D., Shimko,T.C., Aditham,A.K., Keys,A.M., Longwell,S.A., Orenstein,Y. and Fordyce,P.M. (2018) Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E3702–E3711.
35. Aditham,A.K., Shimko,T.C. and Fordyce,P.M. (2018) In: Fletcher,D.A., Doh,J. and Piel,M. (eds). *Methods Cell Biol*. Academic Press, Vol. **148**, pp. 229–250.
36. Shrikumar,A., Greenside,P. and Kundaje,A. (2017) Reverse-complement parameter sharing improves deep learning models for genomics. bioRxiv doi: https://doi.org/10.1101/103663, 27 January 2017, preprint: not peer reviewed.
37. Badis,G., Chan,E.T., van Bakel,H., Pena-Castillo,L., Tillo,D., Tsui,K., Carlson,C.D., Gossett,A.J., Hasinoff,M.J., Warren,C.L. *et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell*, **32**, 878–887.
38. Rogers,J.M., Waters,C.T., Seegar,T.C.M., Jarrett,S.M., Hallworth,A.N., Blacklow,S.C. and Bulyk,M.L. (2019) Bispecific forkhead transcription factor FoxN3 recognizes two distinct motifs with different DNA shapes. *Mol. Cell*, **74**, 245–253.
39. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
40. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A. *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
41. 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.
42. Kumar,R., Reynolds,D.M., Shevchenko,A., Shevchenko,A., Goldstone,S.D. and Dalton,S. (2000) Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase. *Curr. Biol.*, **10**, 896–906.
43. Rossi,M.J., Kuntala,P.K., Lai,W.K.M., Yamada,N., Badjatia,N., Mittal,C., Kuzu,G., Bocklund,K., Farrell,N.P., Blanda,T.R. *et al.* (2021) A high-resolution protein architecture of the budding yeast genome. *Nature*, **592**, 309–314.
44. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordân,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.