SUPPLEMENTARY DATA

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

[4] Departments of Chemistry, Physics & Astronomy, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA

* 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

SUPPLEMENTARY ANALYSIS

Comparing the reproducibility of PBM and SELEX-seq data

Whereas PBM experiments have been instrumental in revealing the binding preferences for hundreds of TFs in a high-throughput manner (1,2), the typical experiment is designed to examine binding sites up to 8-bp in length. Although the Seed-and-Wobble approach was proposed to extend the binding site beyond 8-bp (1), motifs generated by the approach were found to greatly underperform relative to BEEML-PBM, putting its broad applicability into question (3). While fluorescence intensity and binding strength are correlated in uPBM data, unrelated factors may also affect fluorescence, such as the binding site location and orientation within the 60-bp probes. In a typical PBM experiment, each 8-mer occurs in at least 32 positions on the chip, so the effects from these secondary factors are thought to be minimal relative to the effects of binding strength (1). However, for moderate-to-low affinity sequences, these effects can lead to noisy or irreproducible measurements across different microarray designs (Supplementary Figure S3A). This also complicates the detection of small-scale preferences that may be found in the positions flanking the core. Our SELEX-seq experiment is able to address many of the discussed concerns simply due to the extreme depth of sequencing utilized. We ultimately collected around 35 million reads per sample for Fkh1 and Fkh2, and about 10 million reads per sample for Hcm1 and Fhl1.

Previously, Bulyk and co-workers derived binding profiles for 89 yeast TFs, including Fkh1, Fkh2, and Fhl1 using PBM experiments (4). For each TF, replicate experiments were performed using two separate microarrays designed with independent de Bruijn sequences. In comparing the Z-score normalized PBM enrichments between the two microarray designs probed with Fkh1, we detected weak correlations for the majority of 8-mers measured (Supplementary Figure S3A). This was especially apparent for 8-mers exhibiting low signal intensities. In comparing Fkh1 and Fkh2 intensities measured using the same microarray design, a stronger correlation was observed. This suggests that any differences in binding between Fkh1 and Fkh2 are more subtle than the variance that is observed across different microarray designs. For this reason, it is clear that a uPBM study of this design is not suitable for revealing the subtle differences in binding preferences between the two homologs. Alternatively, our SELEX-seq experiment produced reproducible enrichment measurements for Fkh1 across varying library designs while successfully revealing a greater degree of variance between Fkh1 and Fkh2 evaluated on the same library design (Supplementary Figure S3B). The specific differences are explored in detail using our core-based alignment framework as discussed in the main text.

Evaluating the sensitivity of our framework to the stopping rule

We determined our list of cores using a 95% stopping rule described in Materials and Methods. However, we wanted to evaluate how stable our observations would be given an extended set of core sequences. As an extreme case, we decided to realign our Fkh1 reads using our curated list of 49 cores plus all additional sequences from the list of the top 500 prioritized cores, ignoring our previously defined stopping rule. This allowed us to align 28.4% of the R1 data, and only 33.1% of the R2 data, likely due to overfiltering as a result of many reads containing multiple ‘cores’. We found $-\Delta\Delta G/RT$ measurements to be highly correlated with those collected when alignment was performed using only our curated list of 49 cores, with Pearson r^2 equal to 0.99 for core measurements, and 0.90 for flanking measurements (Supplementary Figure S6A). Looking at the flanking preferences of the additional cores alone, we find a drastic drop-off in flanking contributions when sorting by core enrichment, suggesting that many of the less-enriched cores are not faithfully identifying true binding sites (Supplementary Figure S6B). Therefore, while it may be possible to increase the number of cores analyzed using a less stringent stopping rule, it would also increase our false-positive rate.

Quantifying the interdependencies of flanking positions using MLR

One assumption in our calculation of flanking position contributions was that they contribute independently of each other to binding. To probe the validity of this assumption, we used Multiple Linear Regression (MLR) to model and predict *ΔΔG*/*RT* for long *k*-mers that cover multiple flanking positions. Since individual *k*-mer counts are larger for shorter sequences, their associated *ΔΔG*/*RT* measurements are less noisy. However, longer *k*-mers can be more informative regarding interdependencies between nucleotide positions. Using the set of core-aligned reads, the *ΔΔG/RT* can be calculated for extended core sequences by including additional positions flanking the core. We included at most six flanking positions covering four bp 5' and two bp 3' of the core, covering up to 13 bp in total. Instead of using a 100-count threshold for this length sequence, we filtered out sequences that do not appear in at least two independent alignment windows, allowing for the analysis or more sequences without compromising the correlation of *ΔΔG/RT* measured between subsets of the data. For each model, sequences were encoded to capture differing levels of information about interdependencies.

For the simplest encoding, flanking positions were encoded independently of each other and independently of the core using a one-of-four encoding, representing the four bases of DNA. For a flanking sequence of length *L*, the full encoding includes 4 x *L* features. These features are referred to as “Flank Mononucleotides”. Since the core was restricted to a relatively small list of predefined sequences, the core was encoded using a one-of-*C* encoding, with *C* representing the number of cores included during the alignment. This feature is later referred to as “Core Identity”. This allowed us to avoid the assumption that core positions contribute independently to binding.

Alternatively, flanking positions can be encoded in a core-dependent manner by matrix multiplication of the 4 x *L* flanking features by the one-of-*C* feature, resulting in a 4 x *L* x *C* set of features for each *k*-mer, referred to as “Core-Dependent Mononucleotides”. These features most closely represent the flanking contributions captured by our core-based alignment framework. The model can also be extended to include interdependencies between flanking positions by encoding all pairwise dinucleotides. This is done by matrix multiplication of the 4 x *L* flanking features by itself to produce a 16 * L^2 set of features for each *k*-mer. Since this matrix is symmetric, only the upper triangle was kept for modeling. These features are believed to capture stacking interactions between bp, as well as long-range interactions between distant positions. We refer to these features as “Flank Dinucleotides”. Models were evaluated using a nested 5-fold cross validation strategy with performance reported as the mean coefficient of determination, R^2. For each of the 5 training sets, hyperparameters were tuned using the LassoCV function from the sklearn python package with the maximum number of iterations set to one million, and all other settings unchanged from their default configuration (5).

Here, we used Fkh1 as an example since it exhibited the largest feature contributions of flanking positions. Additionally, we only utilized the round 2 data since a very small set of 13-mers are significantly enriched after 1 round of selection. As described previously, measurements of 13-mer enrichment are inherently noisy relative to shorter *k*-mers. If we split our dataset into two equal parts of randomly sampled sequences, ensuring that duplicate reads are in the same set, we can compare the –*ΔΔG*/*RT* measurements between the two to get a sense of how much of the variance is due to noise. We find the measurements moderately well correlated with a Pearson r^2 of 0.89. This puts an approximate upper bound on the performance of the model since it is not designed to explain noise.

This simplest model, including Core Identity and Flank Mononucleotide features, achieved an average cross-fold R^2 of 0.70 (Supplementary Figure S17A). Adding Flank Dinucleotides, we obtained an improved R^2 of 0.75. This suggests that flanking positions can interact, but these interactions only explain a small fraction of the overall variance. If we instead included Core-Dependent Mononucleotides, the model achieved an R^2 of 0.83. This improvement in performance confirms our previous observation that different cores slightly modulate flanking position contributions. Altogether, this indicates that flanking positions are dependent on the core as well as other flanking positions. With this in mind, we concluded that the best model would consider both types of interdependencies. Indeed, this model achieved the highest performance with an R^2 of 0.90, roughly matching the limit of the model approximated previously based on the noisiness of the data.

If we assume the 5' flanking positions contribute independently of the 3' flanking positions, we can repeat the above tests using less noisy 11-mer sequences including the four 5' flanking positions and the core. Between the same two equally sized halves mentioned previously, we observe an r^2 of 0.98. The simplest model achieved an R^2 of 0.89 (Supplementary Figure S17B). The model including dinucleotide interactions between flanking positions achieved an R^2 of 0.91. The model including core-dependent flanks achieved an R^2 of 0.93. Lastly, the model including dinucleotide interactions and core-dependent flanks achieved an R^2 of 0.96. As seen before, the best model included both types of interdependencies. Although both types of flanking interdependencies were beneficial to modeling, leaving them out did not severely diminish the power of the model. In the main text, we explore how local DNA shape features, which result from interdependencies between positions, may explain these preferences.

Evaluation of flanking preferences determined using a deep neural network

One of the main goals of our alignment-based framework is to increase the interpretability of our findings. If we were to unintentionally include a false core, which fails to align true binding sites, that would be apparent by the presence of atypical flanking preferences. Therefore, our framework can both identify and validate cores that are used for alignment. Furthermore, assigning reads to independent windows allows for greater confidence in observed flanking preferences since even small-scale differences can be considered significant if they are repeated across many windows.

Although deep learning models can accurately reproduce *k*-mer level enrichments, model parameters are highly complex and cannot be assigned meaningful confidence intervals. Additionally, deep learning models are not inherently informative of binding site positioning along a given *k*-mer, making them unsuitable for understanding position-specific interactions between the DNA and protein. One way to get around this limitation is to use predictions of the deep learning model to build a simpler model. For example, a PSAM can be generated by predicting the $\Delta\Delta G/RT$ of all sequences which are one mutation away from a given reference. Although this representation is easier to interpret, it once again suffers from the inability to account for interdependencies between positions.

Additionally, while we restrict our analysis to sequences containing one binding site, this is not the case for deep learning models. This makes it impossible to differentiate whether bases at flanking positions act to modulate the affinity of a given core, or act to create additional cores. To demonstrate this point, we modulated the BET-seq framework to investigate flanking preferences for our previously defined set of 7-bp cores (6,7). The goal of a BET-seq experiment is similar to a SELEX-seq experiment, except that a fixed 'core' is embedded within the randomized region of the library, allowing the user to specifically investigate flanking preferences for a given binding site. While this framework is designed to center binding on the given core, it does not prohibit the creation of additional cores outside the one that is given. Additionally, this framework requires the user to specify a previously known ‘core’ which limits the discovery of novel cores which we demonstrate with our analysis.

To summarize flanking preferences, the BET-seq framework employs a deep neural network to generate high-resolution estimates of binding energy for every unique sequence covering five bp upstream and downstream of the core. They then use these estimates to train a linear model based on 1-mer sequence features as input and use the feature weights to generate an interpretable energy logo. Adapting this framework, we first calculate the $-\Delta\Delta G/RT$ of all 13-mers which occur within our selected R2 reads, based on the original SELEX-seq protocol (8,9). Although these measurements are noisy on their own, we found that removing the 100-count threshold greatly improved coverage of moderate-to-low affinity cores, resulting in improved modeling outcomes. Next, we randomly split the data into a 30% testing set and a 70% training set which we used to train a model using the DeepBind framework (10,11) (Supplementary Figure S22). We then used this model to predict high-resolution estimates of $-\Delta\Delta G/RT$ for all 13-mers covering four positions 5’ and two positions 3’ of each of our 7-bp cores. This model achieved an R^2 of 0.85 on both the training and testing datasets.

For each core, Flank Mononucleotides are fed into an MLR model using Lasso regularization to predict the $-\Delta\Delta G/RT$ values provided by DeepBind. Model parameters were then used to represent flanking preferences in a grid-like format as was used for our alignment-based measurements. At each flanking position, the average feature weight is subtracted from the specific contribution of each bp in order to center the data. Looking down the list of flanking preferences predicted by MLR, sorted by the $-\Delta\Delta G/RT$ values of the core, we find that the magnitude of the flanking preferences decreases relative to the affinity of the core (Supplementary Figure S23A). Without our filtering framework, low affinity cores co-occur with stronger alternate cores. Therefore, they would not truly indicate the most likely binding site, and flanking positions would not be informative how binding to that core would be modulated. This results in the ‘dilution’ of signal, especially for low-affinity cores.

Additionally, we find that while MLR-based preferences exhibit similar patterns of specificity, there appears to be more noise across differing cores. Looking at energy logos plotted using the centered model parameters, we indeed find a preference for flanking positions which create high-affinity cores, rather than modify the affinity of the core of interest (Supplementary Figure S23B). For example, looking at the cores AAATACA and CAATACA, we see an elevated preference for ‘GT’ at the two positions 5’ of the core. Including these two bases creates the cores GTAAATA, our fourth strongest core, and GTCAATA, our thirteenth strongest core. The same is not observed for our alignment-based measurements because we eliminate reads which contain multiple cores. Altogether, this suggests that our method was both more sensitive and more accurate in revealing how flanking positions modulate the affinity for a wide number of cores.

SUPPLEMENTARY TABLES AND FIGURES

Oligo Name	Sequence
Library	GAGTTCTACAGTCCGACGATCCAG**NNNNNNNNNNNNNNNN**TCCGTATCGCTCCTCCAATG
+ control Fkh1/Fkh2	GAGTTCTACAGTCCGACGATCCAG**AAAAGGTAAACAAGAA**TCCGTATCGCTCCTCCAATG
+ control Hcm1	GAGTTCTACAGTCCGACGATCCAG**GCGAAATAAACAAAAC**TCCGTATCGCTCCTCCAATG
+ control Fhl1	GAGTTCTACAGTCCGACGATCCAG**AACCGACGCAAACAAA**TCCGTATCGCTCCTCCAATG
− control	GAGTTCTACAGTCCGACGATCCAG**AGAGTTAGCCGATGTT**TCCGTATCGCTCCTCCAATG
Forward Primer	GAGTTCTACAGTCCGACGATC
Reverse Primer	CATTGGAGGAGCGATACG
5’ adapter	AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA
3’ adapter – R0	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCACGTGATCATTGGAGGAGCGATAC
3’ adapter – R1 Fkh1	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCAACATCGCATTGGAGGAGCGATAC
3’ adapter – R1 Fkh2	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCAGCCTAACATTGGAGGAGCGATAC
3’ adapter – R2 Fkh1	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCACACTGTCATTGGAGGAGCGATAC
3’ adapter – R2 Fkh2	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCAATTGGCCATTGGAGGAGCGATAC
3’ adapter – R1 Hcm1	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCAATTGGCCATTGGAGGAGCGATAC
3’ adapter – R1 Fhl1	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCAACATCGCATTGGAGGAGCGATAC
3’ adapter – R2 Hcm1	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCAGATCTGCATTGGAGGAGCGATAC
3’ adapter – R2 Fhl1	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCAGCCTAACATTGGAGGAGCGATAC
GACAACA – CB	GCTCAATTGACAACATATCGG
GAAAACA – CB	GCTCAATTGAAAACATATCGG
GTAAACA – CB	GCTCAATTGTAAACATATCGG
StrongFlank – CB	GCTCAATTGTCAACATATCGG
WeakFlank – CB	GCTCGGCGGTCAACAGGTCGG
	1 10 20 30 40 50 60 70 80

Supplementary Table S1. Oligo sequences used for SELEX-seq and competitive binding (CB) experiments. The 16-bp variable region of the library is bolded. Positive and negative controls were used to determine the appropriate protein:DNA ratio to minimize non-specific binding, as described in Materials and Methods. Forward and reverse primers were used in the amplification of the library. Illumina adapters and barcodes were added to the final library products using four cycles of PCR with the 5’ and 3’ adapter oligos shown above. The flanking positions that were modulated for the competitive binding experiment are underlined.

Protein		Library	Volume
Fkh1	1488 fmol	5818 fmol	30 µL
Fkh2	1477 fmol	5818 fmol	30 µL
Hcm1	6000 fmol	6000 fmol	30 µL
Fhl1	3000 fmol	6000 fmol	30 µL

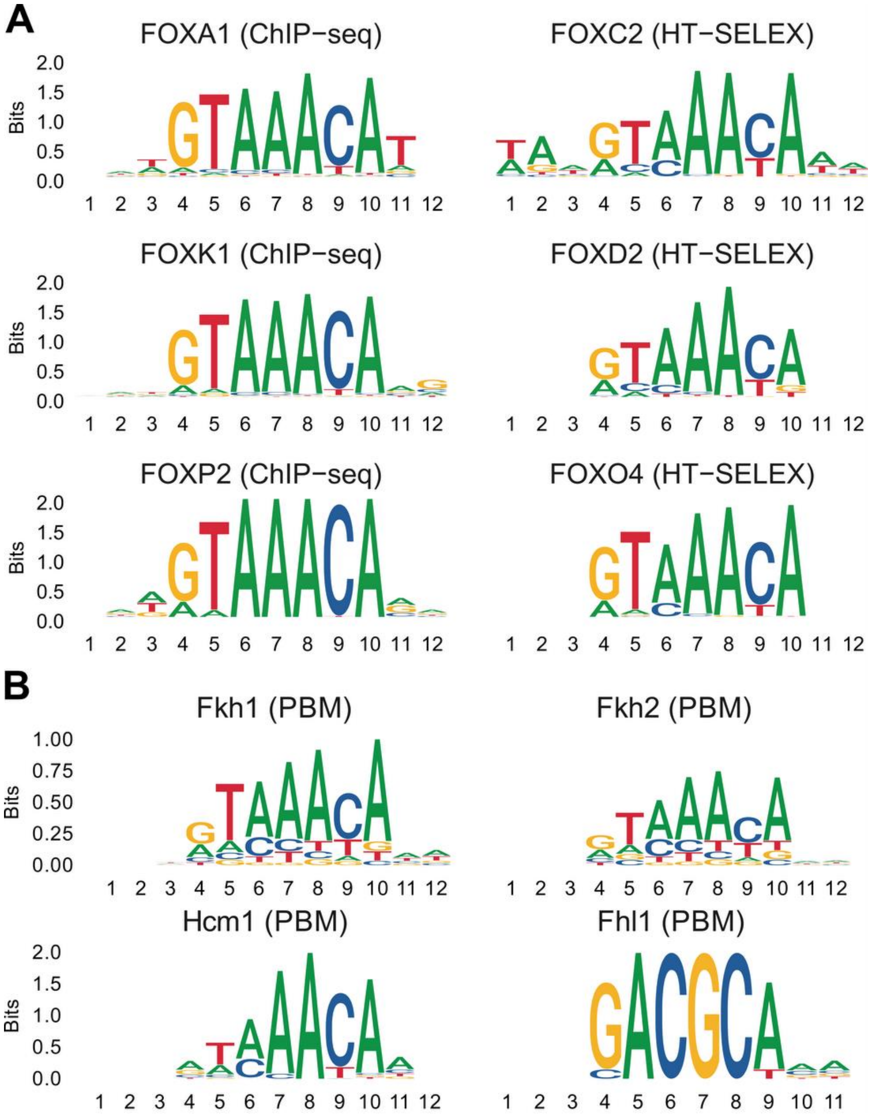
Supplementary Table S2. Table of molar amounts used for all SELEX-seq experiments pertaining to each homolog.

Protein		Probe		Competitor		Volume
Fkh1	2400 fmol	GACAACA - CB	600 fmol	GAAAACA - CB	0 - 4800 fmol	15 µL
Fkh1	2400 fmol	GAAAACA - CB	600 fmol	GACAACA - CB	0 - 2400 fmol	15 µL
Fkh1	1800 fmol	GTAAACA - CB	600 fmol	GAAAACA - CB	0 - 9600 fmol	15 µL
Fkh2	2400 fmol	GTAAACA - CB	600 fmol	GAAAACA - CB	0 - 9600 fmol	15 µL
Fkh1	1440 fmol	StrongFlank - CB	600 fmol	WeakFlank - CB	0 - 38400 fmol	15 µL

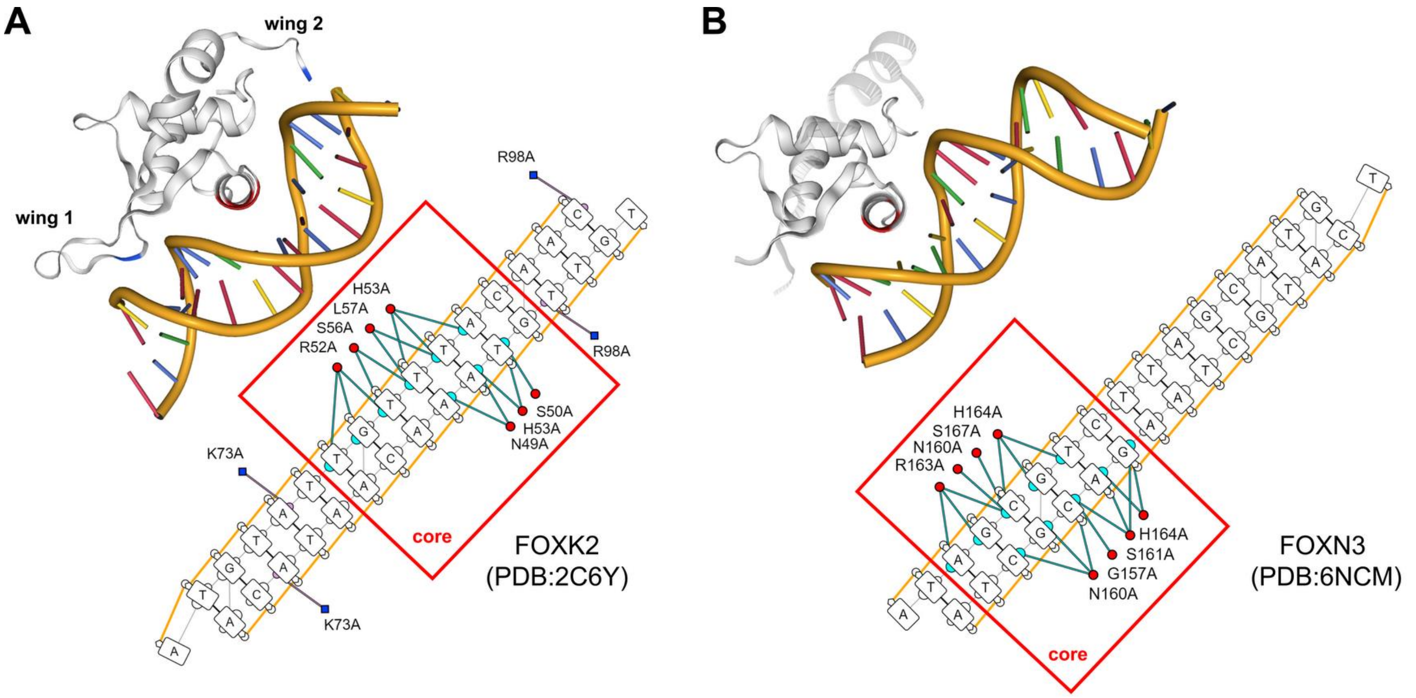
Supplementary Table S3. Table of molar amounts used for all competitive binding experiments. The amount of competitor varies, with molar ratios relative to the probe shown on the corresponding gel.

Core	Fkh1	Fkh2	Hcm1	Fhl1
GTAAACA	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	1.209 (0.105)
ATAAACA	0.536 (0.061)	0.524 (0.057)	0.191 (0.060)	1.266 (0.058)
GTCAACA	0.673 (0.060)	0.557 (0.070)	0.284 (0.064)	1.167 (0.076)
GTAAATA	0.835 (0.106)	0.550 (0.102)	1.205 (0.076)	1.456 (0.114)
ATCAACA	1.011 (0.065)	0.949 (0.072)	0.090 (0.058)	1.205 (0.057)
GTACACA	1.325 (0.072)	1.399 (0.103)	0.644 (0.076)	0.861 (0.061)
GACAACA	1.363 (0.080)	0.911 (0.083)	0.709 (0.062)	1.164 (0.078)
CACAACA	1.481 (0.059)	0.797 (0.066)	0.458 (0.084)	0.842 (0.080)
CAAAACA	1.483 (0.109)	0.790 (0.064)	0.783 (0.057)	1.166 (0.087)
CTAAACA	1.546 (0.123)	1.467 (0.091)	1.218 (0.085)	1.218 (0.078)
ATAAATA	1.548 (0.103)	1.298 (0.108)	1.610 (0.090)	1.441 (0.103)
AACAACA	1.610 (0.082)	0.934 (0.081)	0.235 (0.072)	1.129 (0.078)
GTCAATA	1.631 (0.074)	1.162 (0.114)	1.420 (0.064)	1.320 (0.096)
ACAAACA	1.692 (0.088)	1.299 (0.065)	0.761 (0.100)	0.976 (0.090)
ATACACA	1.862 (0.077)	1.793 (0.088)	0.707 (0.089)	0.904 (0.077)
ATATACA	1.881 (0.086)	1.764 (0.103)	1.314 (0.074)	1.176 (0.075)
GAAAACA	1.899 (0.096)	1.237 (0.082)	1.062 (0.071)	1.327 (0.134)
AATAACA	1.901 (0.068)	0.928 (0.070)	0.773 (0.077)	1.261 (0.068)
CCAAACA	1.926 (0.074)	1.349 (0.062)	0.958 (0.089)	0.862 (0.092)
CATAACA	1.938 (0.137)	1.072 (0.075)	0.871 (0.088)	1.037 (0.102)
AAAAACA	2.019 (0.104)	1.378 (0.067)	0.902 (0.078)	1.348 (0.121)
ATCAATA	2.028 (0.103)	1.699 (0.105)	1.335 (0.093)	1.338 (0.064)
GCAAACA	2.040 (0.066)	1.468 (0.083)	1.173 (0.074)	0.789 (0.075)
TACAACA	2.221 (0.105)	1.630 (0.103)	1.472 (0.070)	1.074 (0.077)
AGCAACA	2.285 (0.084)	1.966 (0.079)	1.114 (0.075)	0.842 (0.052)
CAAAATA	2.325 (0.126)	1.580 (0.104)	2.068 (0.094)	1.328 (0.073)
GGCAACA	2.426 (0.084)	1.975 (0.111)	1.635 (0.051)	1.075 (0.065)
AATAATA	2.440 (0.199)	1.511 (0.089)	2.071 (0.093)	1.436 (0.084)
CAATACA	2.475 (0.116)	2.014 (0.123)	1.453 (0.074)	1.065 (0.094)
GACAATA	2.475 (0.158)	1.952 (0.092)	2.190 (0.073)	1.327 (0.113)
GATAACA	2.482 (0.101)	1.484 (0.120)	1.448 (0.077)	1.196 (0.082)
ACCAACA	2.493 (0.093)	1.949 (0.087)	1.572 (0.069)	1.094 (0.104)
CACAATA	2.523 (0.180)	1.863 (0.098)	1.766 (0.055)	0.960 (0.079)
AAATACA	2.534 (0.242)	2.118 (0.183)	1.367 (0.101)	1.325 (0.118)
AGAAACA	2.577 (0.067)	2.432 (0.087)	1.516 (0.085)	1.291 (0.145)
GACGCA-	2.599 (0.129)	3.164 (0.132)	2.174 (0.114)	0.000 (0.000)
CATAATA	2.628 (0.124)	1.632 (0.107)	1.973 (0.081)	1.210 (0.095)
GAATACA	2.739 (0.103)	2.284 (0.091)	1.515 (0.098)	1.235 (0.070)
GTCGCA-	2.914 (0.192)	3.317 (0.172)	2.581 (0.104)	0.294 (0.051)
CATCACA	2.966 (0.199)	2.365 (0.102)	1.563 (0.084)	0.910 (0.075)
CACGCA-	3.048 (0.196)	3.168 (0.177)	2.272 (0.134)	0.070 (0.040)
AACGCA-	3.083 (0.234)	3.140 (0.238)	2.377 (0.108)	0.209 (0.044)
GACGCT-	3.209 (0.193)	3.383 (0.182)	2.904 (0.151)	0.385 (0.049)
TACGCA-	3.224 (0.307)	3.218 (0.191)	2.566 (0.130)	0.386 (0.062)
CGCGCA-	3.229 (0.161)	3.520 (0.148)	2.640 (0.153)	0.271 (0.050)
GACGCG-	3.233 (0.152)	3.444 (0.107)	2.738 (0.121)	0.194 (0.054)
GGCGCA-	3.406 (0.158)	3.565 (0.149)	2.694 (0.133)	0.360 (0.058)
GACTCA-	3.676 (0.310)	3.536 (0.178)	2.978 (0.164)	0.239 (0.046)

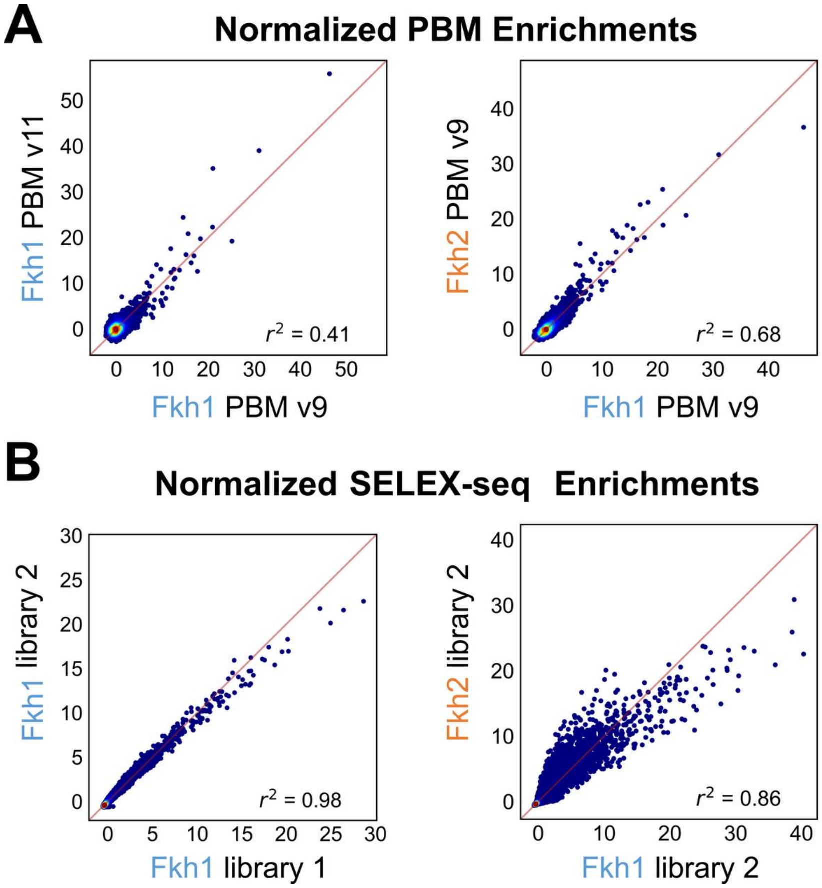
Supplementary Table S4. *ΔΔG/RT* measurements for our selected core sequences averaged over every window, with standard deviations shown in parentheses.



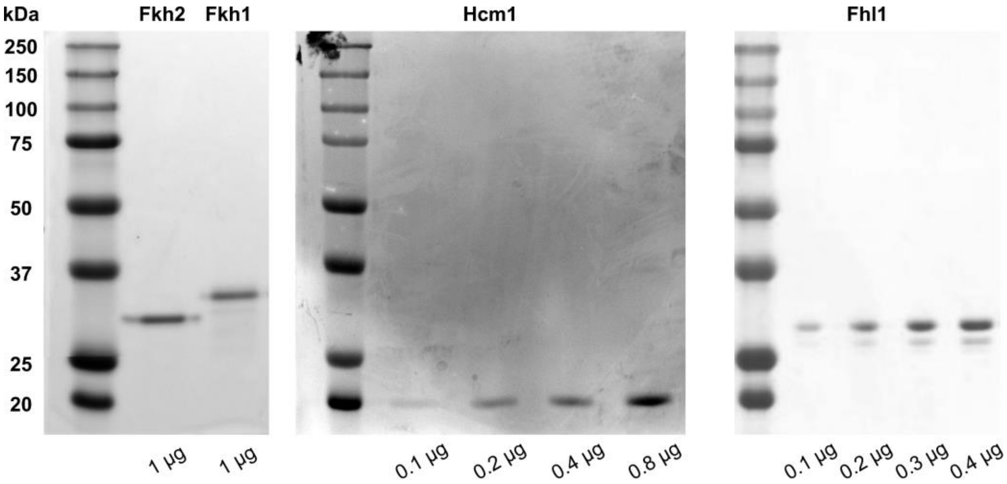
Supplementary Figure S1. Binding logos derived from ChIP-seq, HT-SELEX, and PBM experiments performed on FOX homologs found in (A) human or (B) *S. cerevisiae*. Data was downloaded from JASPAR (12) and UniPROBE (13) and then visualized using ggseqlogo.



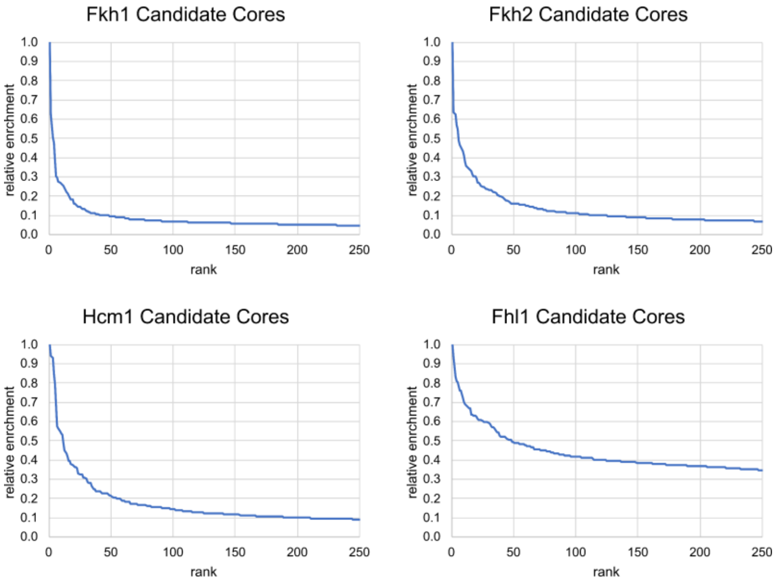
Supplementary Figure S2. Co-crystal structure of FOX protein bound to a fragment of DNA (14), with contact maps generated using DNAproDB (15,16). Teal half-circles represent major groove contacts, whereas purple half-circles represent minor groove contacts. Red circles indicate residues along the main recognition helix of FOXK2, while blue squares represent residues present within the winged regions. (A) Shows FOXK2 contacting a canonical GTAAACA binding site. In this structure, the wings were able to be resolved, revealing minor groove contacts outside of the core. (B) Shows FOXN3 contacting the sub-optimal Fhl motif, GACGCA. Many contacts in the core are maintained despite a drastically differing sequence.



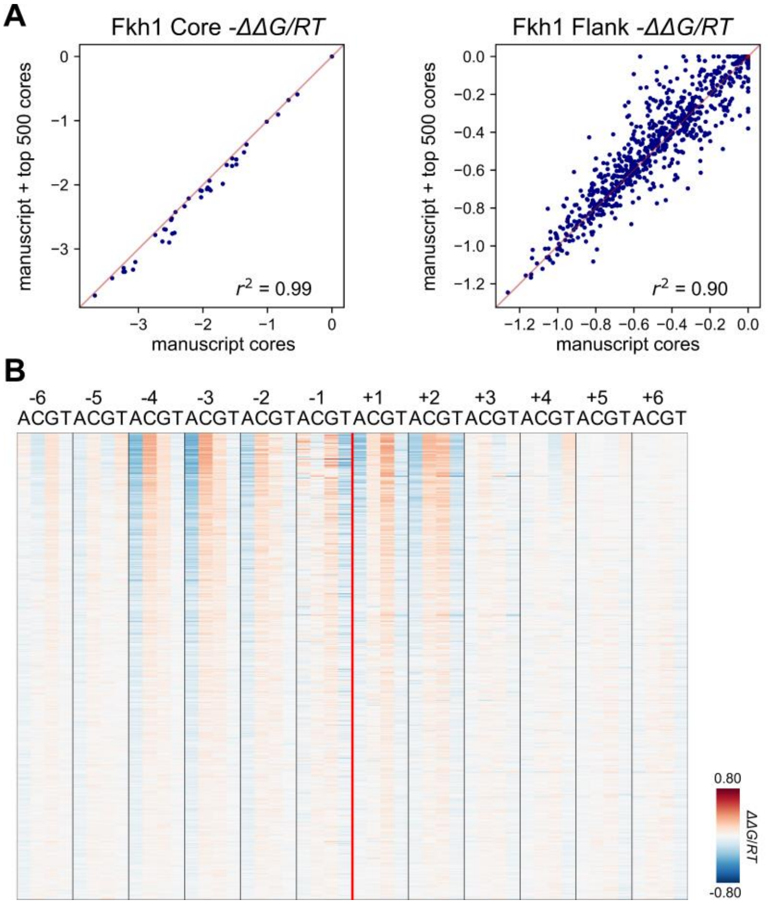
Supplementary Figure S3. (A) Comparison of Z-score normalized enrichment scores for 8-mer sequences according to two independent PBM microarray designs performed on Fkh1 and Fkh2. (13) (B) Comparison of Z-score normalized enrichment scores for 9-mer sequences derived from two separate SELEX-seq experiments with different library designs performed on Fkh1. Fkh2 was only evaluated using the second library design.



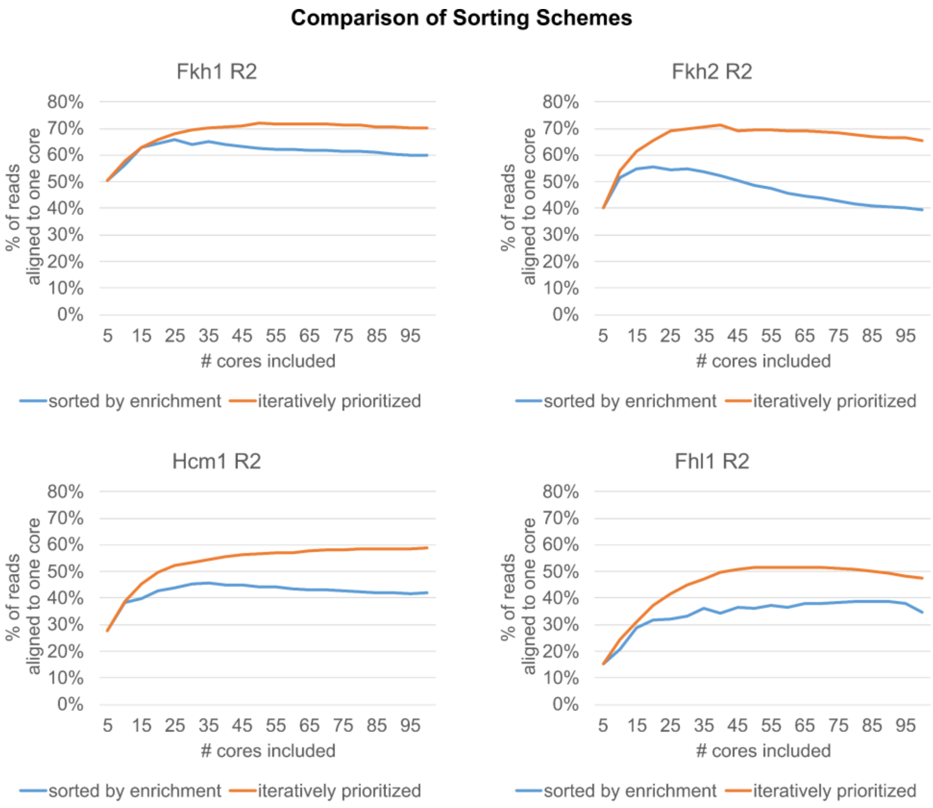
Supplementary Figure S4. SDS-PAGE gels of our purified constructs indicating the amount of construct loaded in each lane.



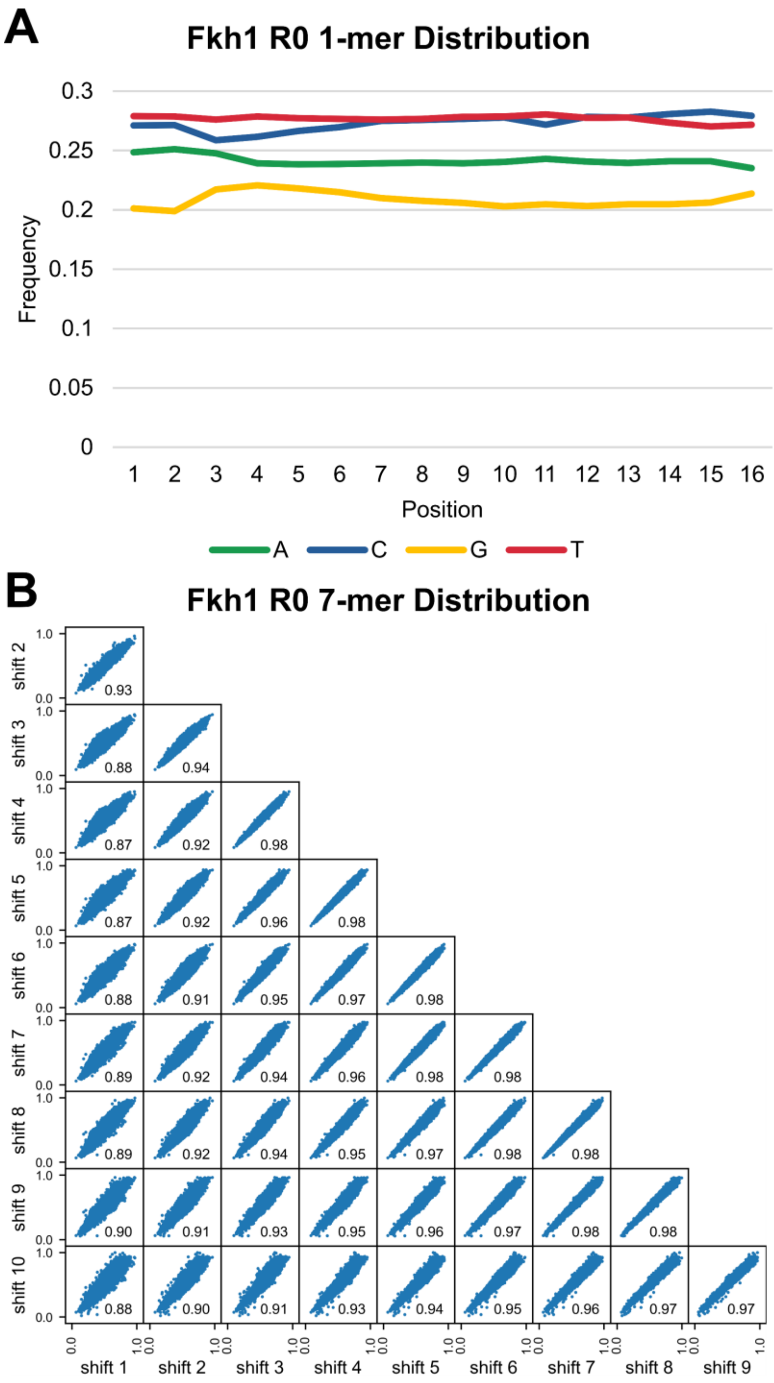
Supplementary Figure S5. Plots showing the relative enrichment of the top 250 candidate cores determined by Top-Down-Crawl (TDC) as described in Materials and Methods, showing that a small subset of cores are significantly more enriched than the majority.



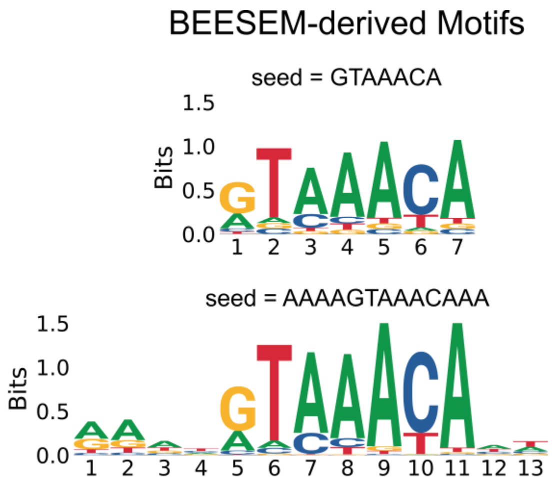
Supplementary Figure S6. (A) Comparison of core and flanking *−ΔΔG/RT* measurements when reads are aligned using our selected set of 49 cores, or aligned with the same set in addition to a list of the top 500 prioritized cores. (B) Average flanking contributions measured across the top 500 prioritized cores, showing a diminishing signal as false cores are more likely to be included.



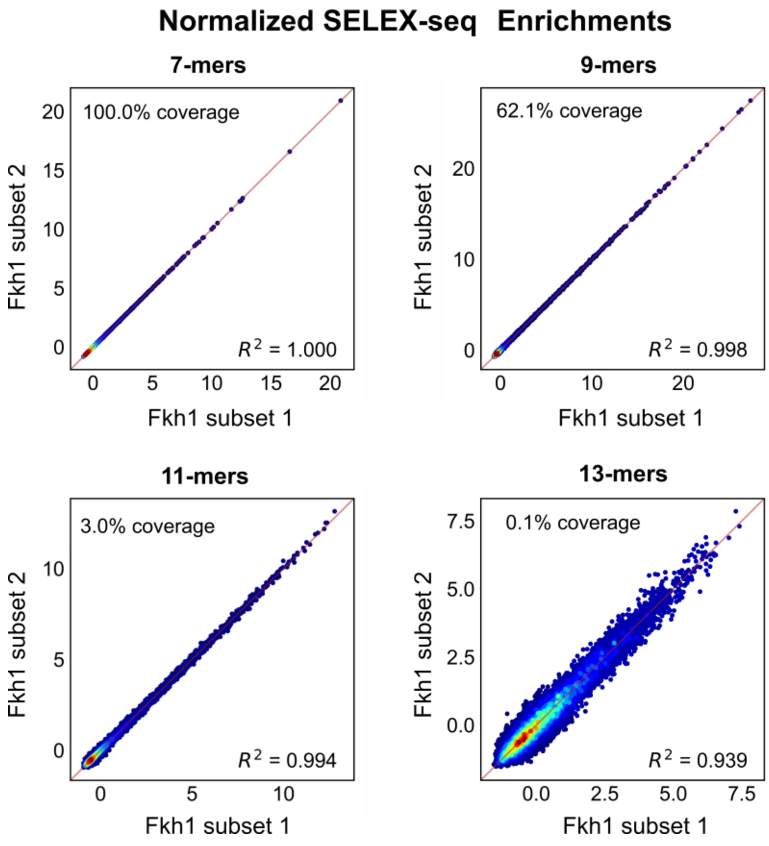
Supplementary Figure S7. Shows the relationship between the number of cores we included during alignment versus the fraction of reads that align to a single core. The list of candidate core sequences was sorted by raw enrichment or by iterative reprioritization as described in the manuscript. Using the top *k*-mers from the reprioritized lists allows for more sequences to be included in downstream analysis.



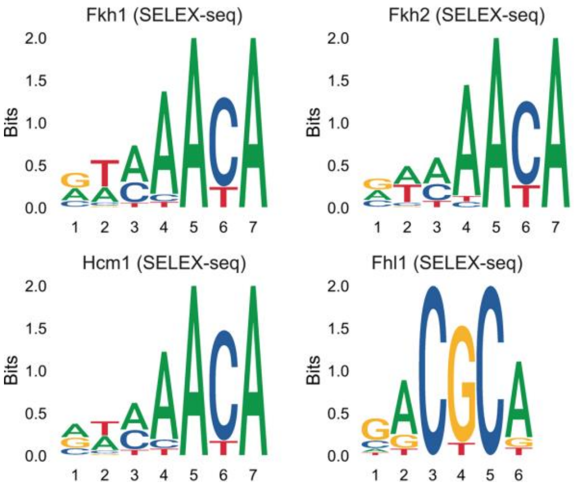
Supplementary Figure S8. (A) Distribution of the four bp at various positions along the 16-mer in the initial library. (B) Scatter plots comparing the relative frequencies of 7-mers occurring at different shifts. This figure exemplifies the need for a position-specific background model when determining enrichment.



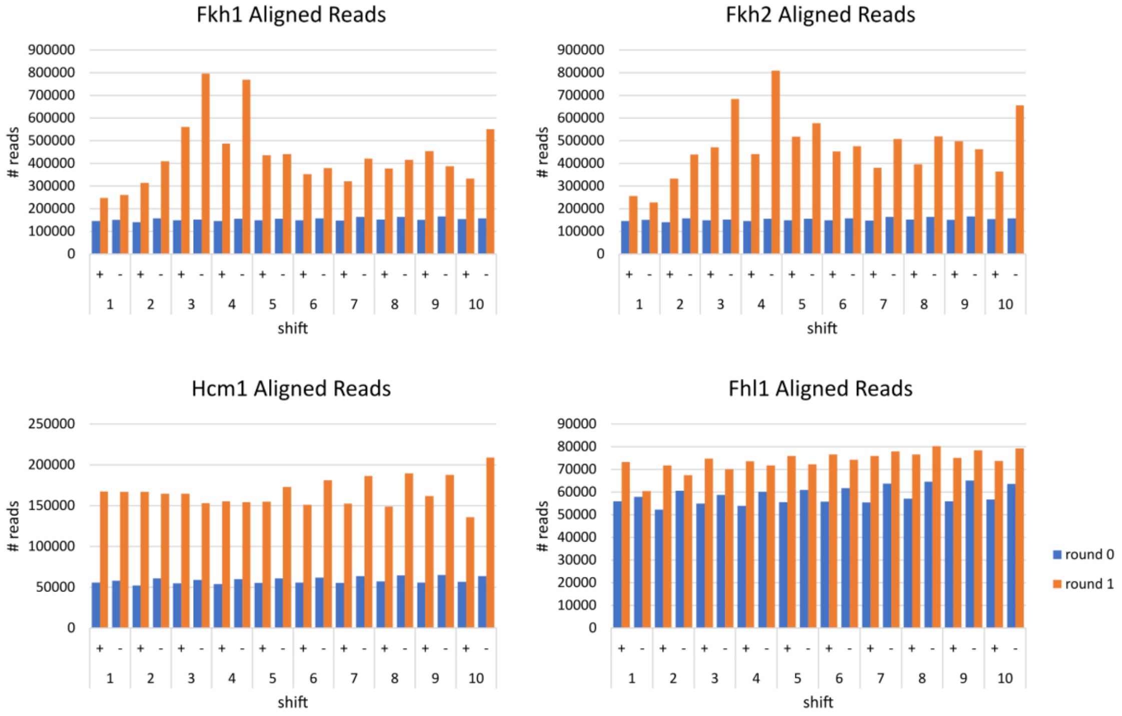
Supplementary Figure S9. BEESEM *(17)* motifs derived using either a seed of GTAAACA or AAAAGTAAACAAA, covering four positions 5' of the core and two positions 3' of the core. These motifs are used to predict the enrichment of cores, or flanking bp surrounding the core in ChIP-exo peaks.



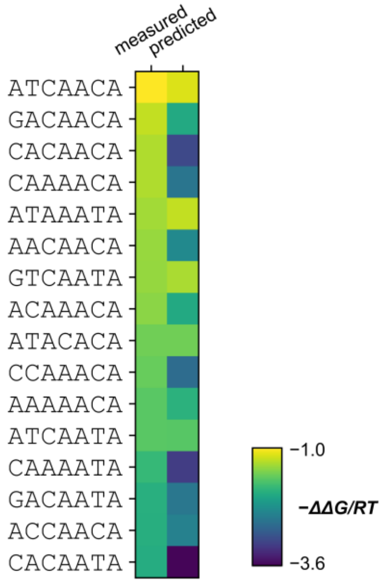
Supplementary Figure S10. Comparison of the Z-score normalized enrichment of various length *k*-mers derived from two equally sized sets of unique reads from Round 2 of the Fkh1 SELEX-seq experiment. The coverage represents the proportion of unique *k*-mers for which we could calculate the enrichment.



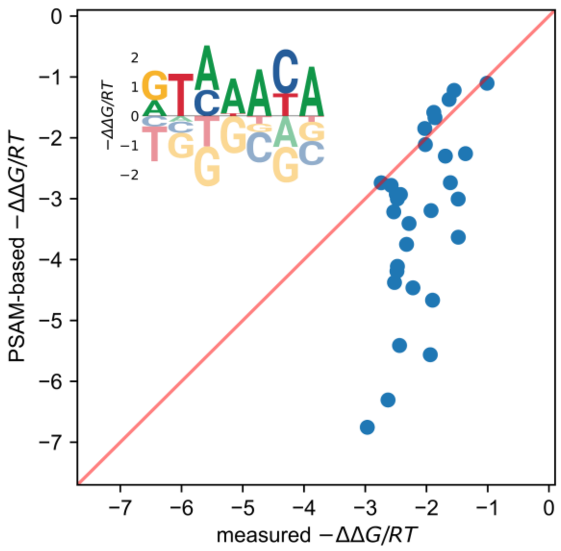
Supplementary Figure S11. PWMs generated from our selected 7-bp or 6-bp cores, weighting each sequence by its relative enrichment. We would like to reiterate that these are generated for comparison only, and we do not support the use of such PWMs in this case to broadly represent binding preferences.



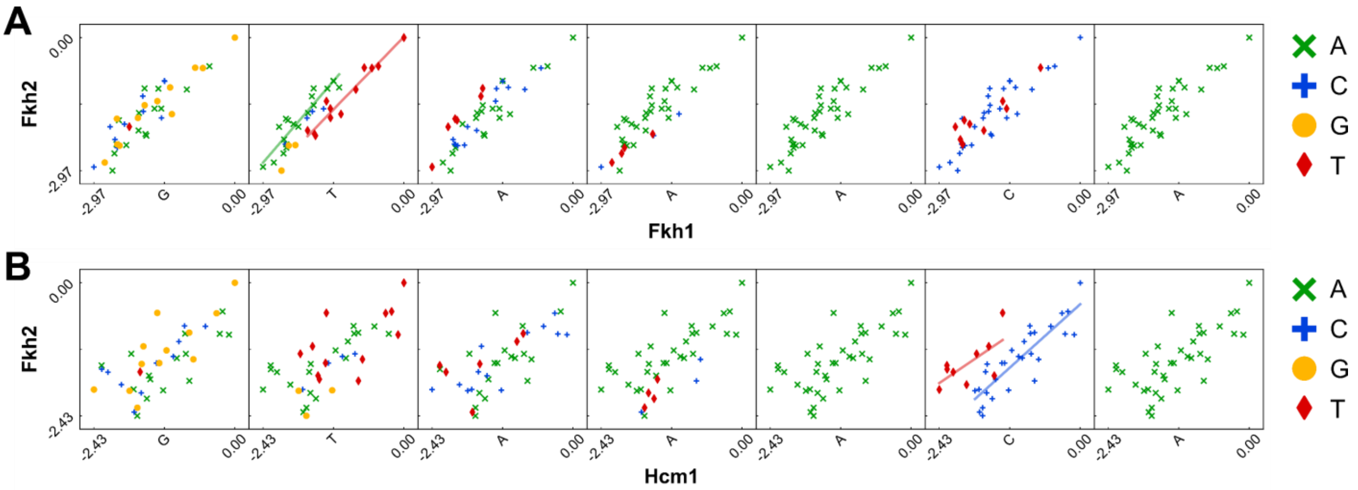
Supplementary Figure S12. Bar plots showing how reads are distributed across independent windows after alignment to our set of cores. The distribution changes more dramatically for homologs which are more sensitive to flanking positions.



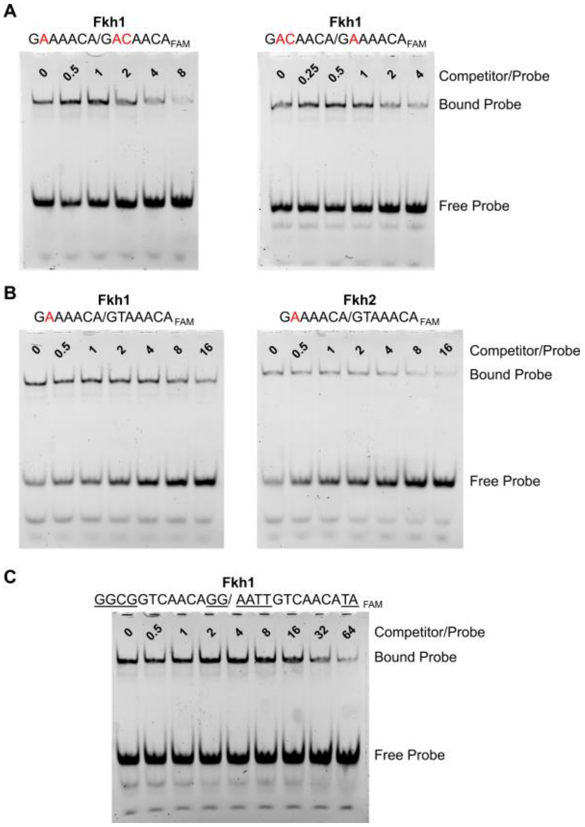
Supplementary Figure S13. Heatmap comparing the measured *–ΔΔG/RT* of 7-bp cores that are at least two mutations away from the reference, GTAAACA, to those values predicted by adding the *−ΔΔG/RT* measurements of cores that are one mutation away from the reference.



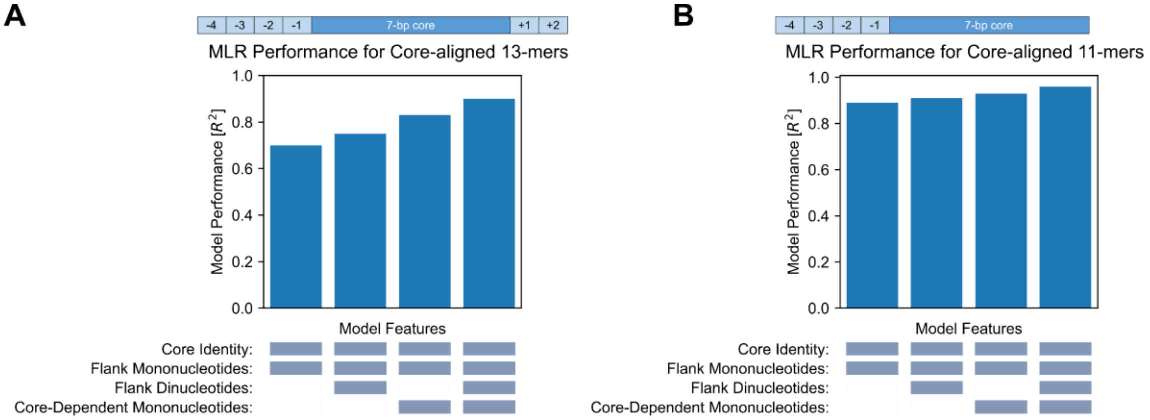
Supplementary Figure S14. Scatter plot comparing the measured *–ΔΔG/RT* of 7-bp cores that are at least two mutations away from the reference, GTAAACA, to those values predicted using a PSAM generated from the table of 7-mer enrichments from Fkh1 R2, as overlayed on the plot. PSAM-based predictions are consistently lower than our observed measurements.



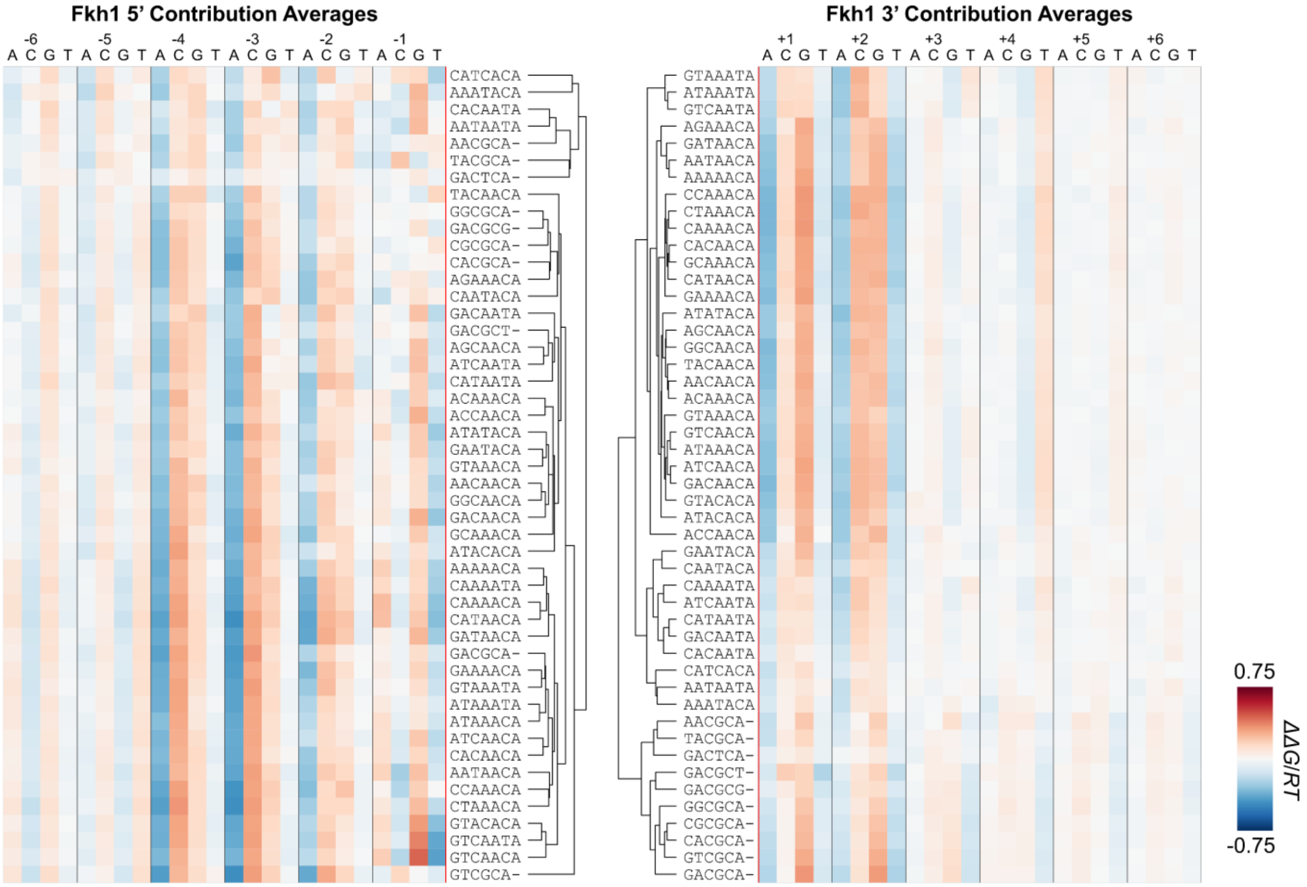
Supplementary Figure S15. Comparison of –ΔΔG/RT of 7-bp core sequences between (A) Fkh1 and Fkh2 or (B) Hcm1 and Fkh2 color-coded by bp identity at each position, with lines representing the base-specific least-squares fit at positions of interest, revealing differing base preferences between FOX homologs.



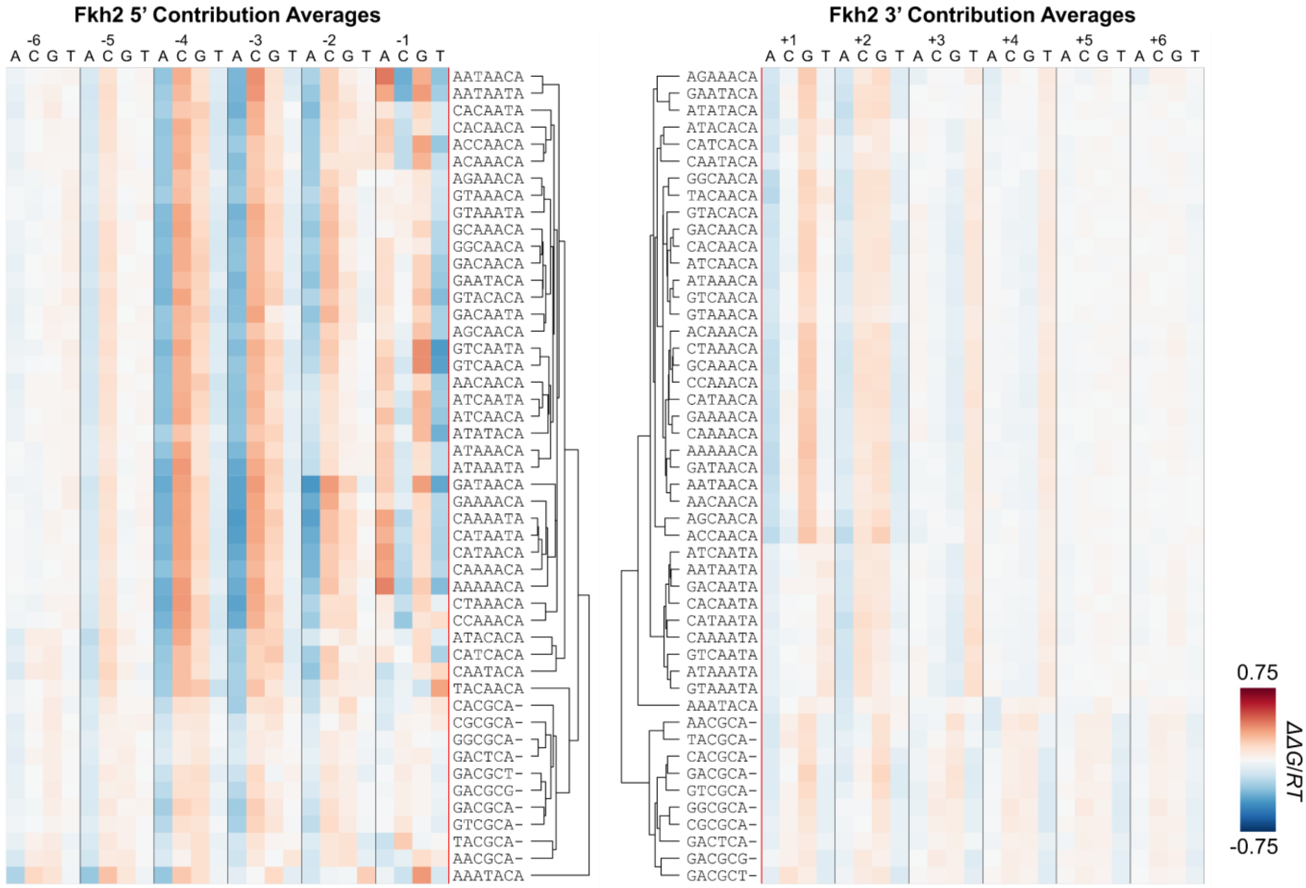
Supplementary Figure S16. Competitive binding assays, with conditions detailed in Supplementary Table S3. In each case, the probe is FAM-labeled, and the competitor is unlabeled. The molar ratio of competitor to probe is shown. (A) Experiments comparing the cores GAAAACA and GACAACA (B) Experiments comparing Fkh1 and Fkh2 binding to the cores GTAAACA and GAAAACA. (C) Experiment comparing a fixed core surrounded by optimal or suboptimal flanking sequences.



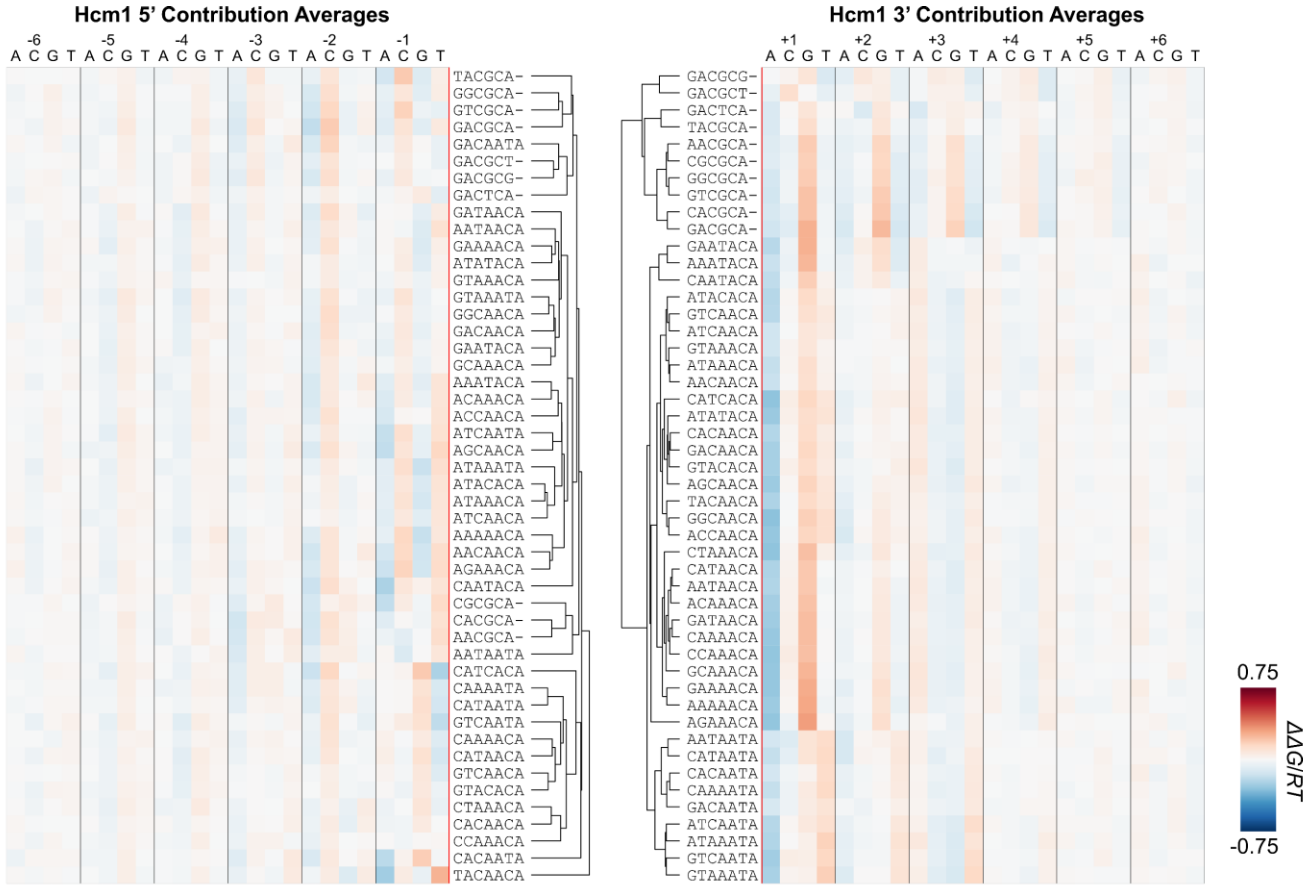
Supplementary Figure S17. Bar plots showing the performance of MLR models designed to interrogate the importance of interdependencies across flanking positions. Models were trained to predict (A) 13-mers covering four bp 5’ of the core and two bp 3’ of the core, or (B) 11-mers covering the four bp 5’ and zero bp 3’ of the core. Model features are described in Materials and Methods.



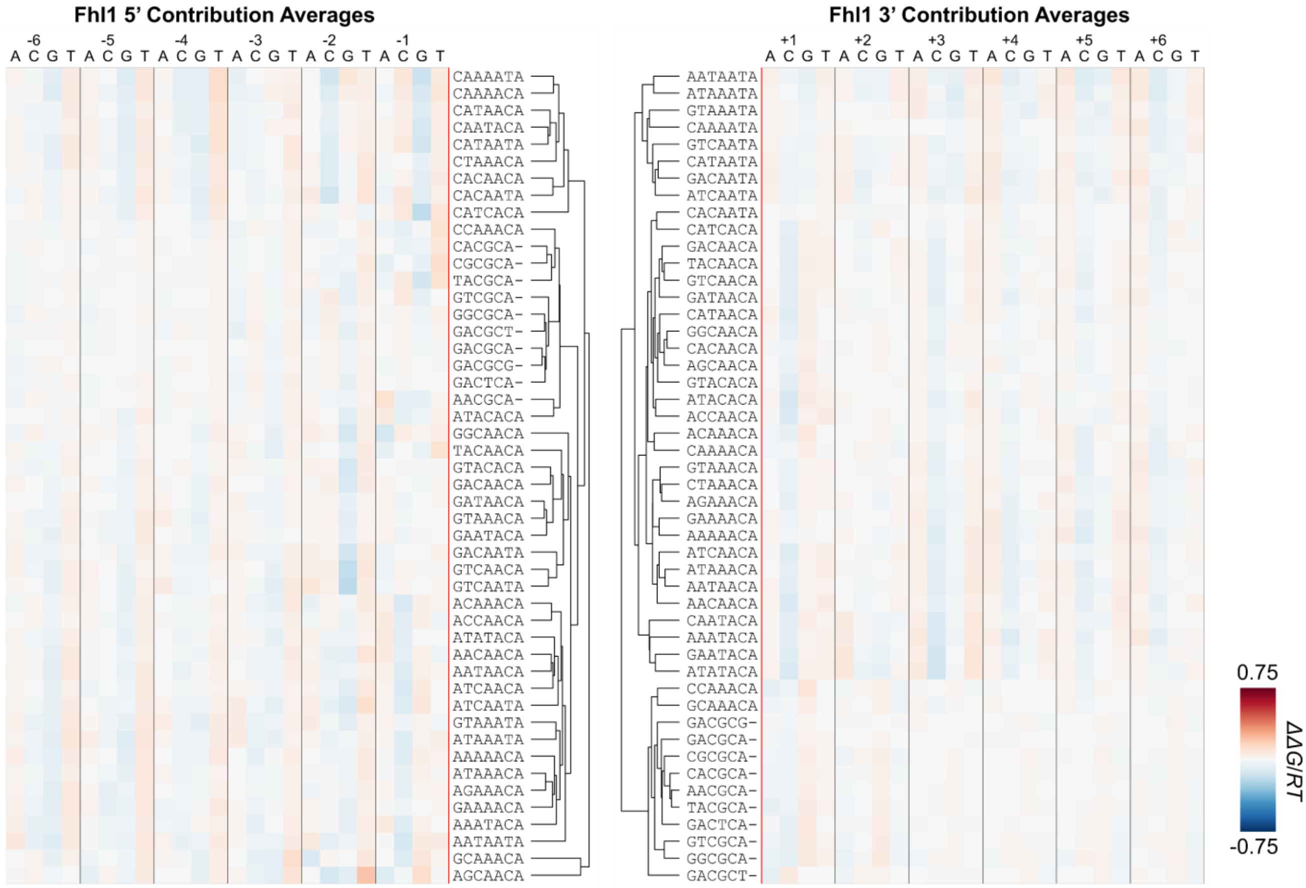
Supplementary Figure S18. *ΔΔG*/*RT* measurements for each aligned core averaged over the 40 independent sets of aligned reads from the Fkh1 SELEX-seq experiments. Rows are clustered with the UPGMA algorithm using Manhattan distance as the metric.



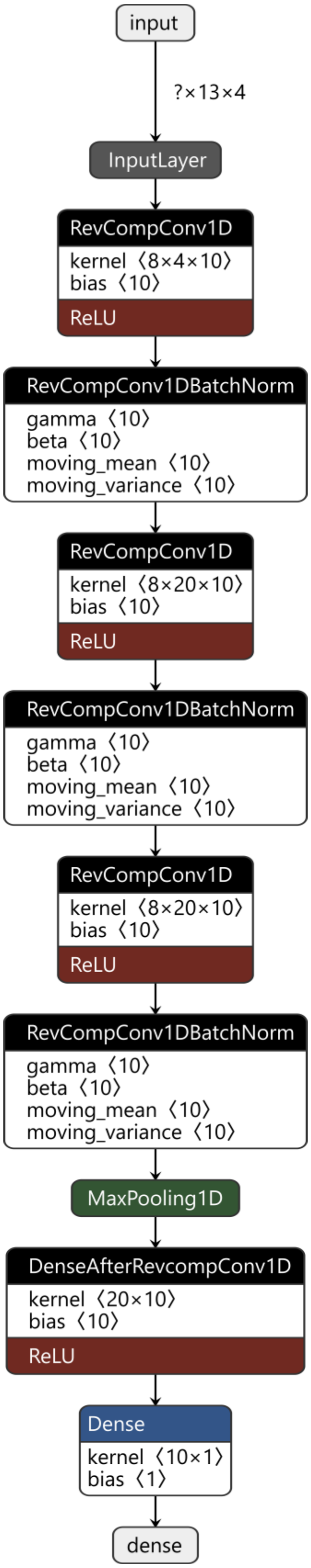
Supplementary Figure S19. *ΔΔG*/*RT* measurements for each aligned core averaged over the 40 independent sets of aligned reads from the Fkh2 SELEX-seq experiments. Rows are clustered with the UPGMA algorithm using Manhattan distance as the metric.



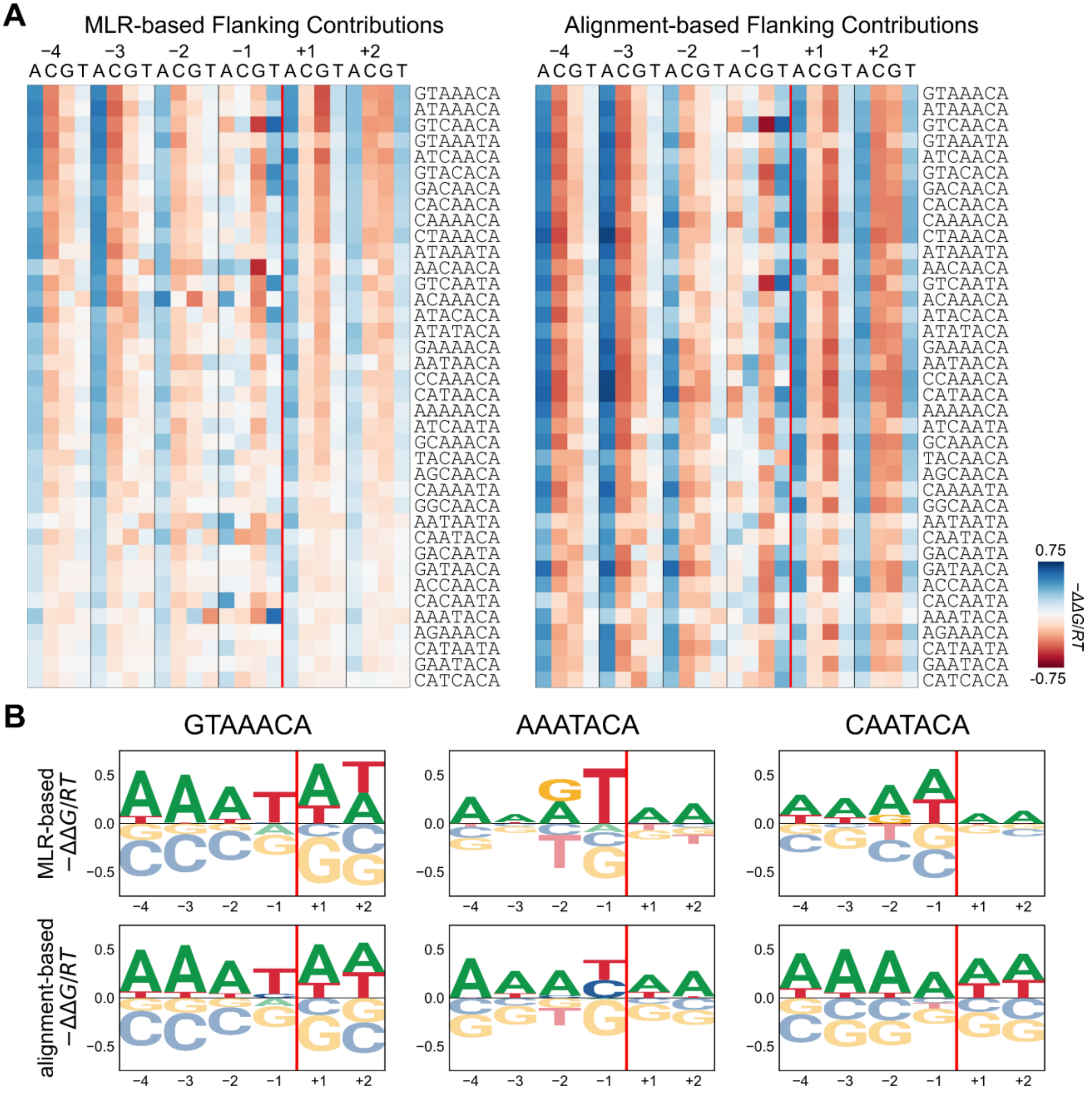
Supplementary Figure S20. *ΔΔG/RT* measurements for each aligned core averaged over the 40 independent sets of aligned reads from the Hcm1 SELEX-seq experiments. Rows are clustered with the UPGMA algorithm using Manhattan distance as the metric.



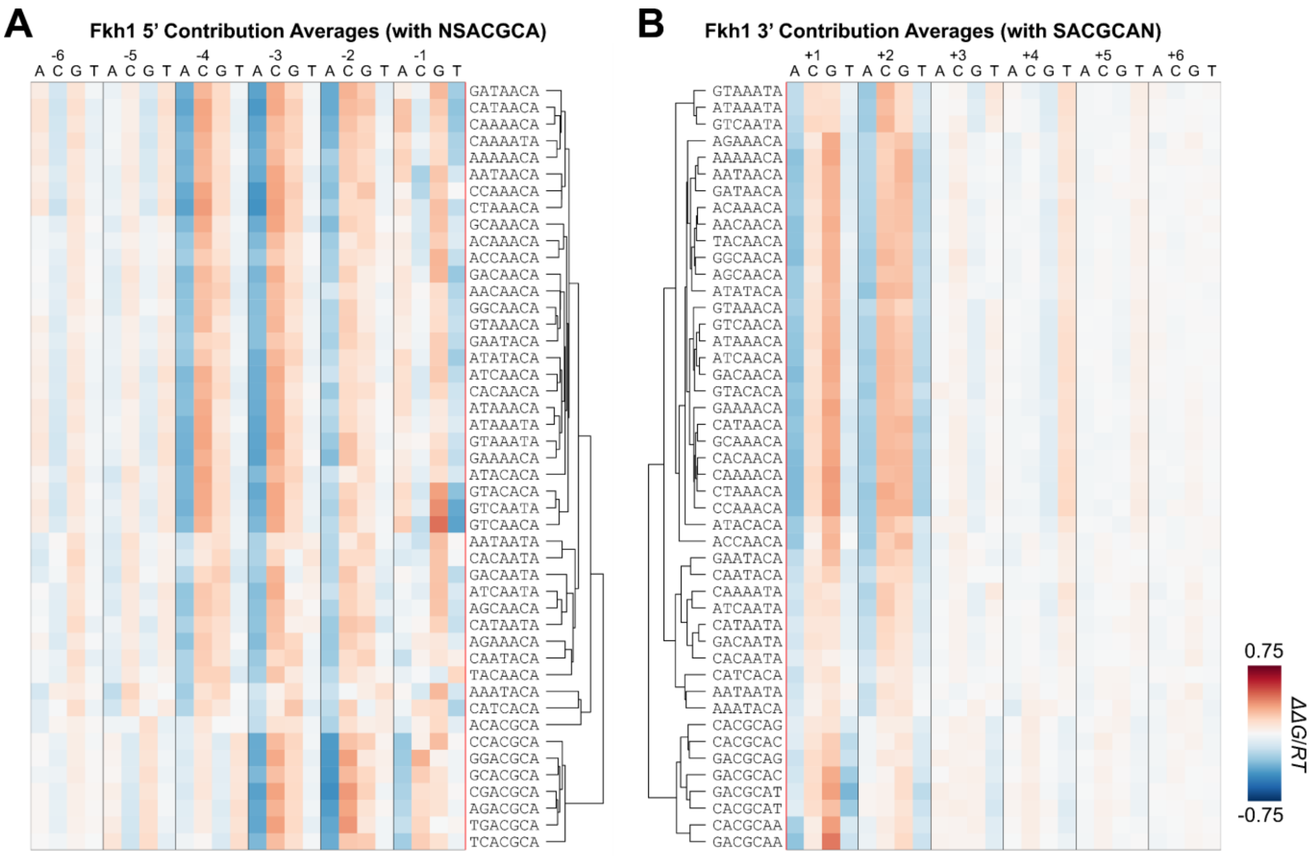
Supplementary Figure S21. *ΔΔG/RT* measurements for each aligned core averaged over the 40 independent sets of aligned reads from the Fhl1 SELEX-seq experiments. Rows are clustered with the UPGMA algorithm using Manhattan distance as the metric.



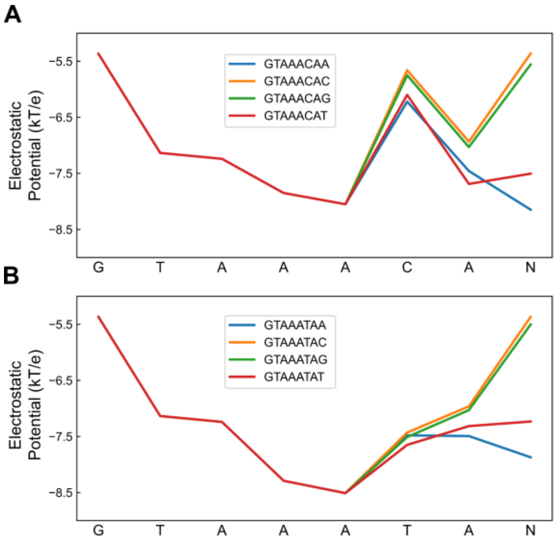
Supplementary Figure S22. Deep learning model architecture, employing reverse-complement parameter sharing layers from the keras-genomics package developed by the authors of DeepBind (10,11). The model was visualized using Netron (18).



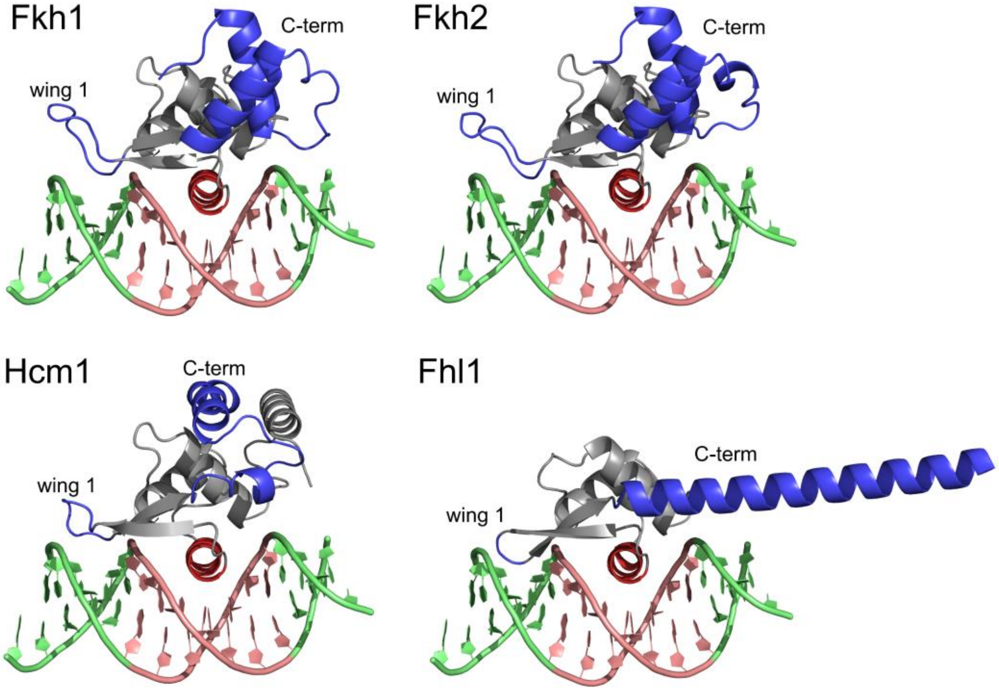
Supplementary Figure S23. Comparison of flanking contributions predicted using the DeepBind + MLR-based approach or measured using our alignment-based framework, (A) depicted as a matrix, sorted by the *−ΔΔG/RT* of the core or (B) depicted as an energy logo for specific cores references in the main text.



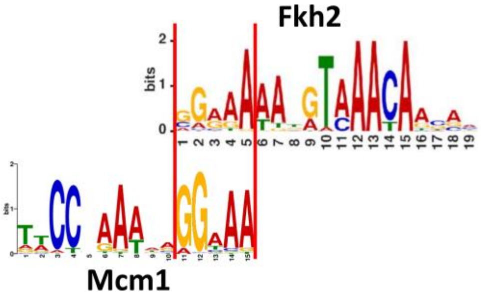
Supplementary Figure S24. *ΔΔG/RT* measurements for each aligned core averaged over the 40 independent sets of aligned reads from the Fkh1 SELEX-seq experiments. Rows are clustered with the UPGMA algorithm using Manhattan distance as the metric. In this case, Fhl1-based cores GACGCA and CACGCA are treated as 7-bp cores by padding each with one base either (A) 5' of the core or (B) 3' of the core.



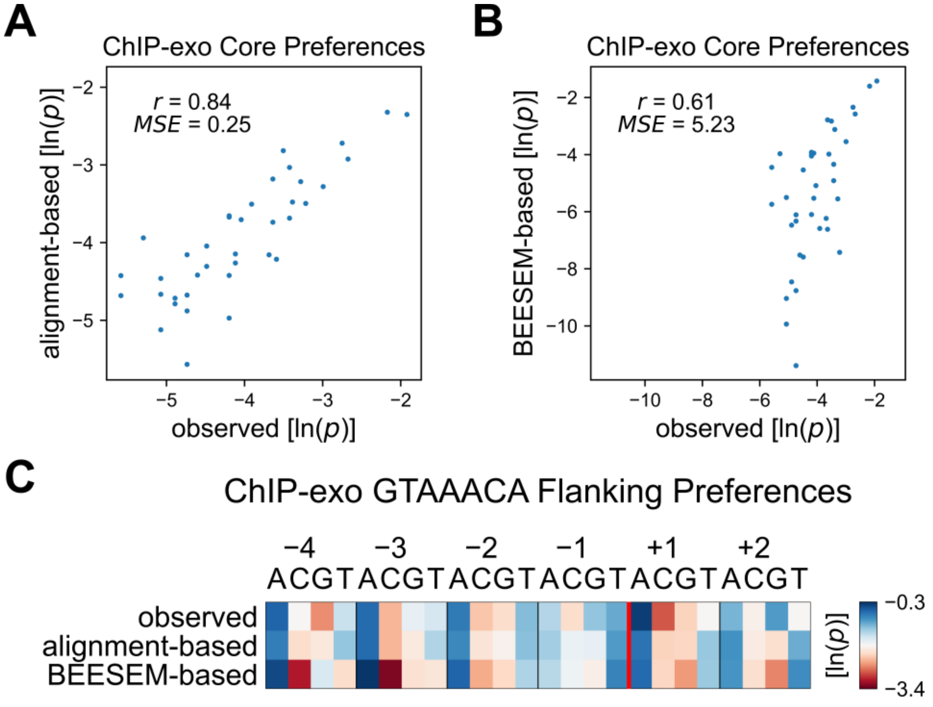
Supplementary Figure S25. (A/B) Plot of the electrostatic potential (EP) in the center of the minor groove (19) along the binding site given a C or T at position six and a variable bp at the first position 3' of the core.



Supplementary Figure S26. Structures of FOX homologs predicted using AlphaFold 2 (20,21) alongside DNA extracted from structural alignments with the FOXK2 co-crystal structure (PDB ID: 2C6Y). The recognition helix is indicated in red with labeled regions of interest indicated in blue. The core of the DNA binding site is indicated in pink and flanking bp are indicated in green.



Supplementary Figure S27. Depiction of the overlap between the Fkh2 motif adapted from the referenced ChIP-exo study (22) and an Mcm1 motif provided by the Yeast Epigenome Project (23).



Supplementary Figure S28. (A) Comparison of the observed relative frequency of cores and predicted relative frequencies based on *−ΔΔG/RT* measurements from our framework or (B) estimates based on a BEESEM-derived PWM. (C) Comparison of the observed relative frequency of bases flanking the core, GTAAACA, and predicted values based on *−ΔΔG/RT* measurements from our framework or a BEESEM-derived PWM.

SUPPLEMENTARY REFERENCES

1. Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd and Bulyk, M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429-1435.
2. Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A. and Chen, X. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720-1723.
3. Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126-134.
4. 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.
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825-2830.
6. 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.*, **115**, E3702-E3711.
7. 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.
8. 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.
9. 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.
10. 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.
11. Shrikumar, A., Greenside, P. and Kundaje, A. (2017) Reverse-complement parameter sharing improves deep learning models for genomics. *BioRxiv*, 103663.
12. Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D. *et al.* (2019) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **48**, D87-D92.
13. 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.
14. 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.
15. Sagendorf, J.M., Berman, H.M. and Rohs, R. (2017) DNAproDB: an interactive tool for structural analysis of DNA–protein complexes. *Nucleic Acids Res.*, **45**, W89-W97.
16. Sagendorf, J.M., Markarian, N., Berman, H.M. and Rohs, R. (2020) DNAproDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes. *Nucleic Acids Res.*, **48**, D277-D287.
17. Ruan, S., Swamidass, S.J. and Stormo, G.D. (2017) BEESEM: estimation of binding energy models using HT-SELEX data. *Bioinformatics*, **33**, 2288-2295.
18. Roeder, L. (2017) Netron, Visualizer for neural network, deep learning, and machine learning models.
19. Chiu, T.-P., Rao, S., Mann, R.S., Honig, B. and Rohs, R. (2017) Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein–DNA binding. *Nucleic Acids Res.*, **45**, 12565-12576.
20. 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.

21. 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.
22. 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.
23. 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.