Supplemental Materials for
Genome-wide determinants of sequence-specific DNA binding of general regulatory factors

Matthew J. Rossi, William K. M. Lai, and B. Franklin Pugh

Table of Contents

Supplemental Methods

Antibodies

Santa Cruz antibody sc-7392 (3 µg per ChIP) was used against HA-Pho4 PB-exo samples. Genescript antibody A00683-100 (0.5 µg per ChIP) was used against TAP-tagged purified protein PB-exo and native PB-seq samples in which the TAP tag was cleaved by TEV protease. Indicated amounts of antibodies were conjugated to magnetic beads (10 µl slurry per ChIP of Protein G Mag Sepharose, GE Healthcare) by incubating in 100 µl of IP Dilution Buffer (20 mM Tris-HCl, pH 8.0, 2 mM EDTA, 150 mM NaCl, 1% Triton X-100, 0.1% SDS) for 2 hrs at 4°C prior to immunoprecipitation.

Rabbit IgG was conjugated to Dynabeads as described (Cristea and Chait 2011). Rabbit IgG (Sigma) conjugated to Dynabeads (40 µl of slurry per ChIP containing 68 µg of IgG and 200 µg of Dynabeads) was used against TAP-tagged strains and whole cell extracts, in which the intact TAP-tag containing Protein A was the target.

Genomic DNA purification

Saccharomyces cerevisiae genomic DNA was purified from a 4 g cell pellet of BY4741. After reaching an OD_{600}=2.0, the cells were pelleted and washed with water. Cells were pelleted again, and resuspended in 1.5 volumes of Lysis Buffer (100 mM TrisHCl, pH 7.5, 50 mM EDTA, 1% SDS). Cells were not frozen at any time during this procedure, as starting from frozen cells decreased the final yield of DNA. The resuspended cells were divided in 1.5 ml aliquots among 2-ml screwtop microcentrifuge tubes containing a 500 µl volume of 0.5mm zirconia/silica beads.

To lyse the cells, the samples were incubated for 5 min at 95°C and then disrupted for 5 min in a Biospec mini-beadbeater. The lysate was combined in a 50-ml Falcon tube. A stock of 7 M ammonium acetate, pH 7 was added to a final concentration of 2.5 M with mixing to precipitate proteins. The sample was incubated for 5 min at 65°C, followed by 5 min on ice, and then centrifuged at 8,000 g for 10 min. The supernatant was transferred to a new 50-ml falcon tube, and an equal volume of phenol:chloroform;isoamyl alchohol (25:24:1) (PCIA) was added. After vortexing, the sample was centrifuged at 8,000 g for 5 min. The aqueous layer was transferred to a new Falcon tube and ethanol precipitated. The resulting pellet was resuspended in 750 μl of water, transferred to a microcentrifuge tube, and extracted once more with PCIA to remove any remaining proteins. The aqueous layer was then treated with 5 μg of RNase at 37°C for 1 hr. After RNase treatment, the sample was precipitated with isopropanol, and resuspended in 300 μl of water. To remove the digested RNA, the intact genomic DNA was precipitated by adding 75 μl of PEG solution (25% polyethylene glycol (8,000), 2.5 M NaCl) and incubated for 30 min on ice. The DNA was pelleted in a tabletop microcentrifuge at maximum speed for 15 min at 4°C. The pellet was washed twice with 70% ethanol and centrifuged for 15 min at 4°C. The final pellet was resuspended in 60 μl of water and passed through a BioRad Micro Bio-spin 6 Chromatography column calibrated with 10 mM Tris-HCl, pH 7.5. This step was necessary to remove impurities that accumulated throughout the protocol. The typical yield for 4 g of cells was 100 ± 10 μg of intact genomic DNA with an A_{260}/A_{280} below 1.9.

Homo sapiens genomic DNA was obtained from 200 million MCF7 cells using the protocol for "Purification of Total DNA from Animal Blood or Cells (Spin Column)" with the optional RNAse treatment from a Qiagen Blood and Tissue Extraction Kit.

PB-exo

Genomic DNA was sonicated in Reaction Binding Buffer (20 mM HEPES-KOH, pH 7.5, 50 mM KCl, 5 mM $MgCl_2$, 100 μg/ml BSA, 1 mM DTT) using Diagenode Pico (10 cycles of 30 sec on/ 30 sec off pulses) to produce DNA fragments. When visualized on a 2% agarose gel, the vast majority of DNA molecules ranged from 200 to 500 bps. A typical binding reaction contained 8 μg of sonicated *S. cerevisiae* genomic DNA, and 50 nM to 1 μM purified transcription factor diluted to 400 μl with Reaction Binding Buffer. A protein titration (ranging from 10 nM to 4 μM) was carried out for each factor, and only the concentration that produced the best sequencing data, defined as the most bound sites with the best signal to noise ratio, was presented here. The optimized protein concentrations were: 50 nM for Pho4; 100 nM for Abf1, Reb1, and Rap1; and 1 μM for Cbf1, Mcm1, and Reb1 when the *H. sapiens* genome was used as a substrate.

The reaction was incubated for 30 min at 30°C. Cross-linking was achieved by adding 1% formaldehyde to a final concentration of 0.05% and incubated for 15 min at room temperature. Cross-linking was quenched by adding 2.5 M glycine to a final concentration of 125 mM and incubated for 5 min at room temperature. The sample was then passed through a BioRad Micro Bio-spin 6 Chromatography column calibrated with 10 mM Tris-HCl, pH 7.5 to remove formaldehyde byproducts and diluted to 1 ml with FA Lysis Buffer (50 mM HEPES-KOH, pH 7.5, 150 mM NaCl, 2 mM EDTA 0.1% sodium deoxycholate, 1% Triton X-100, 0.1% SDS).

This sample was then used as the input for the standard ChIP-exo protocol (Rhee and Pugh 2012) with slight modifications. To immunoprecipitate, a 100 µl suspension of IP Dilution buffer containing 10 µl-slurry of magnetic beads conjugated with the appropriate antibody was placed on a magnetic rack, the supernatant removed, and the beads were resuspended off the magnet with the 1 ml diluted PB-exo sample. The sample was incubated on a roto-torque overnight (~16 hrs) at 4°C. The immunoprecipitation was washed sequentially with 150 µl FAT Buffer (40 mM Tris-HCl, pH 8.0, 7 mM EDTA, 60 mM NaCl, 0.2% SDS, 0.375% Triton X-100, and complete protease inhibitors), 150 µl LiCl Buffer (100 mM Tris-HCl, pH 8.0, 500 mM lithium chloride, 1% NP-40, 1% sodium deoxycholate, and complete protease inhibitors), and 150 µl 10 mM Tris-HCl, pH 8.0.

The following enzymatic steps were then carried out with the immunoprecipitated sample on the resin. After each enzymatic step on resin, the sample was washed with LiCl Buffer and 10 mM Tris-HCl.

1. DNA polishing (60 µl reaction volume; 1X NEBuffer 2, 0.5X BSA (NEB; 50 µg/ml), 150 µM dNTPs, 1.5 U T4 DNA polymerase (NEB)) was incubated for 20 min at 12 °C.

2. Kinase (60 µl reaction volume; 1X T4 DNA ligase buffer (NEB), 10 U T4 PNK (NEB)) was incubated for 30 min at 37°C.

3. A-tailing (60 µl reaction volume; 1X NEBuffer 2, 100 µM dATP, 5 U Klenow fragment, -exo (NEB)) was incubated for 30 min at 37°C.

4. First Illumina adapter ligation (50 μl reaction volume; 1X T4 DNA ligase Buffer, 1X BSA, 15 pmol ExA2 adapter, 600 U T4 DNA ligase (Enzymatics)) was incubated for 2 hr at 25°C.

5. Phi29 DNA polymerase fill-in (1X phi29 reaction buffer (NEB), 2X BSA, 165 μM dNTPs, 10 U phi29 DNA polymerase (NEB) was incubated for 20 min at 30°C.

6. Second DNA polishing (60 μl reaction volume; 1X NEBuffer 2, 0.5X BSA, 150 μM dNTPs, 1.5 U T4 DNA polymerase (NEB)) was incubated for 20 min at 12 °C.

7. Second kinase (60 μl reaction volume; 1X T4 DNA ligase buffer, 10 U T4 PNK) was incubated for 30 min at 37°C.

8. Lambda exonuclease digestion (60 μl reaction volume; 1X lambda exonuclease buffer (NEB), 2X BSA, 5 U lambda exonuclease (NEB)) was incubated for 30 min at 37°C.

9. RecJ digestion (60 μl reaction volume; 1X NEBuffer 2, 2X BSA, 30 U RecJ$_f$ exonuclease (NEB)) was incubated for 30 min at 37°C.

Samples were then eluted from the resin by incubating with 40 μl of ChIP Elution Buffer (25 mM Tris-HCl, pH 7.5, 2 mM EDTA, 200 mM NaCl, 0.5% SDS) and the target protein digested with 20 μg Proteinase K incubated overnight (~16 hrs) at 65°C. AMPure (Agencourt) was used to purify the sample according to the manufacturer's instructions, and the sample was eluted into 10 μl of water. The following steps were then performed in solution.

10. Primer extension from ExA2 with phi29 DNA polymerase (20 μl total reaction volume; 1X phi29 reaction buffer, 2X BSA, 100 μM dNTPs, 0.5 μM FX-15 primer, 10 U phi29 DNA polymerase) was incubated without polymerase for 5 min

at 95°C, 10 min at 40°C to anneal the primer, then 5 min at 30°C. At this point, 1 µl of phi29 (10U) was added and the reaction incubated for 20 min at 30°C, then 10 min at 65°C to inactivate, and returned to 37°C.

11. A-tailing (30 µl total reaction volume; 1X NEBuffer 2, 100 µM dATP, 10 U Klenow fragment, -exo) was incubated for 30 min at 37°C, then 20 min at 75°C to inactivate, and returned to 25°C.

12.Second Illumina adapter ligation (ExA1), (40 µl total reaction volume; 1X T4 DNA ligase Buffer, 1X BSA, 15 pmol ExA1 adapter, 1200 U T4 DNA ligase (enzymatics)) was incubated for 2 hr at 25°C.

AMPure (Agencourt) was used a second time to purify the sample according to the manufacturer's instructions, and the sample was eluted into 13 µl of water. The library was then amplified for 24 cycles of polymerase chain reaction.

13.PCR (40 µl reaction volume; 1X Phusion HF Buffer (Thermo scientific), 200 µM dNTPs, 500 nM ExA1 primer, 500 nM ExA2 primer, 2 U Phusion Hot Start polymerase (Thermo scientific) was incubated for 2 min at 98°C to activate the enzyme, followed by 24 cyles of: 20 sec at 98°C to denature, 1 min at 52°C to anneal, and 1 min at 72°C to extend. A final 5 min at 72°C extension was performed before the sample was returned to 4°C.

Libraries were visualized on a 2% agarose gel, and the DNA fragments ranging from 200 to 500 bp were cut out of the gel and purified using a QiaQuick Gel Extraction Kit (Qiagen). Indexed libraries were quantified via qPCR, pooled, and sequenced (see below).

Native PB-seq

This protocol was adapted from (Guertin and Lis 2013); most importantly, formaldehyde is not used to capture protein-DNA interactions. The binding reaction for native PB-seq was identical to PB-exo. It was then added to magnetic resin already conjugated with antibody. Following an incubation for 30 min at 30°C, the resin was washed once with Reaction Binding Buffer to remove non-specifically bound DNA. After removing the wash, the resin was resuspended in 300 µl of ChIP Elution Buffer and incubated for 15 min at room temperature. The eluate was then extracted with equal volume of PCIA, ethanol precipitated, and resuspended in 10 µl of water.

The library construction was adapted from (Quail et al. 2008). All enzymatic steps were carried out in a single microcentrifuge tube. To the resuspended DNA was added components of a DNA polishing reaction with T4 DNA polymerase (NEB) for a reaction volume of 20 µl; this sample was incubated for 20 min at 12°C followed by 5 min at 65°C to inactivate, and returned to room temperature. Next, 5 µl of a kinase mixture with PNK (NEB) was added, raising the reaction volume to 25 µl. It was incubated for 30 min at 37°C, then 5 min at 65°C to inactivate, and returned to room temperature. Next, 5 µl of an A-tailing mixture with Klenow fragment, exo minus (NEB) was added, raising the reaction volume to 30 µl. It was incubated for 30 min at 37°C, then 5 min at 65°C to inactivate, and returned to room temperature. Next, 10 µl of an adapter ligation mixture containing both ExA1 and ExA2 adapters with T4 DNA ligase (NEB) was added, raising the reaction volume to 40 µl. It was incubated for 1 hr at 25°C, then 5 min at 65°C to inactivate, and returned to room temperature. The sample was purified with AMPure and then PCR amplified.

DNA sequencing

Sequencing was performed using the HiSeq 2000 and NextSeq500, generating either 40 bp single-end reads (tags) or 2x40 bp paired-end reads, respectively. Sequence reads were aligned to the yeast genome (sacCer3) using bwa-mem (version 0.7.9a) (Li 2013) using default parameters. Aligned reads were filtered to require only unique alignments to the genome.

Data normalization

All experiments were reproduced multiple times with the same conclusions evident in all replicates. Figures are presented with a merge of all replicates. All ChIP-exo samples were further filtered to remove duplicate PCR reads, defined as possessing identical read 1 and 2. In order to account for differences in depth of sequencing between experiments, aligned sequenced reads (tags) were set to be equal. This was calculated by dividing the total number of de-duplicated aligned sequence tags by genome size and then determining the scaling factor needed to set this ratio to 1 for each experiment. This scaling factor was then applied to tags pileups in all plots and Figures.

Datasets from other publications

H3 MNase-ChIP-seq data was from Batta et al. and aligned to the sacCer3 as previously described (Batta et al. 2011). Prior to analysis, aligned reads were shifted

75 bp from the 5’ to 3’ direction, binned into 3 bp bins, and then smoothed with a running average across 21 bins. Pho4 ChIP-seq in low phosphate growth conditions was downloaded from SRA (SRX065603) and aligned to sacCer3 (Zhou and O'Shea 2011).

Defining bound sites

Low stringency bound sites were called using the following logic. Firstly, the Genetrack (Albert et al. 2008) peak-calling algorithm was used on each strand (watson and crick) separately with the parameters size (s) 5 exclusion zone (e) 10 bp (s20 e40 for ChIP-seq data) and with only one tag required. Strand-separate peaks were then paired, requiring correct orientation with preference to higher occupied peaks. A motif was then required within ±30 bp from the midpoint of the peak-pair. High stringency bound sites were called using identical logic, except Genetrack peak calling stringency was increased from requiring one tag to requiring tags above a 1e^{-4} Poisson p-value threshold calculated for each sample depending on sequencing depth.

Defining ORF promoters

The list of all verified ORFs (5,114) was downloaded from SGD (Cherry et al. 2012). Promoters were leniently defined as the ORF ATG start codon to 500 bp upstream. Motif locations were defined using custom scripts.

Calculating cross-linking points

Predominant cross-linking points were calculated by defining peaks through visual identification of local maxima in the composite plot of strand-separated PB-exo tags (i.e. **Fig. 1B**). The most 5' motif strand (above x-axis) peak was then paired to first downstream opposite strand (below x-axis) peak. The step was repeated until all peaks were accounted for. This orientation of peaks is based on the 5' to 3' polarity of lambda exonuclease digestion. If the peak to peak distance was 15 bp or less, then it was concluded that one formaldehyde cross-linking point existed between the two peaks. The default position of the cross-linking point was set as the midpoint between the two peaks. If the peak to peak distance was greater than 15 bp, then it was concluded that two (or more) cross-links occurred between the identifiable peaks. In this orientation, the default cross-linking points were placed 6 bp downstream of a motif-strand peak, or 6 bp upstream of an opposite-strand peak; based on the empirically derived lambda exonuclease-generated peak pair distance of 12 bp.

The precise location of cross-linking points was further refined if an increase in G/C content was found to occur between the peaks and if that position did not appeared to be independent of the MEME-defined consensus motif. Based on our G/C content analysis, we hypothesize that the majority of formaldehyde cross-links can occur across a 2 – 3 bp window. Presumably, this window is limited by the range of the 5-carbon lysine side chain. Using this method, we believe the error is ±1 bp relative to the calculated cross-linking points.

PB-seq noise calculation

For native PB-seq signal to noise calculation, tags were summed (-200 to +200) relative to the motif midpoint divided by the summation of tags (-900 to -500) or (+500 to +900), whichever was greater. This was calculated only for the sites that had a significant peak pair.

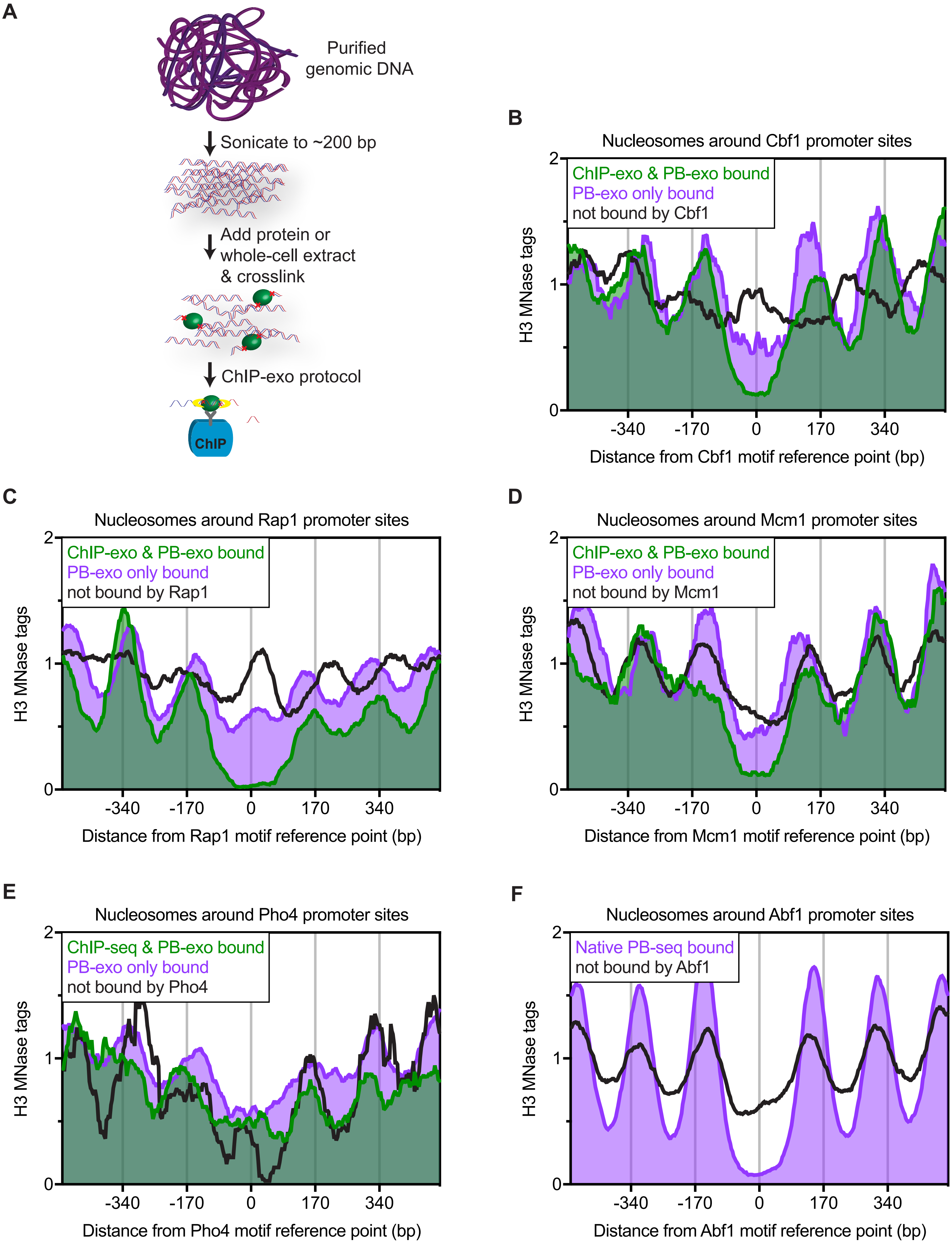
DNA shape analysis

DNA sequence was extracted relative to motif midpoints using the *bedtools getfasta* command (Quinlan and Hall 2010). Structural parameters were then predicted using the DNAShape web interface (Zhou et al. 2013). The Mann-Whitney *U* test was used to test for significance of structural profiles as described in (Gordan et al. 2013). The Z score reports the number of standard deviations of that calculation from the null hypothesis. The null hypothesis is that the two sets of sites possess the same distribution of DNA shape measurement at the indicated nucleotide.

Unsupervised hierarchical clustering

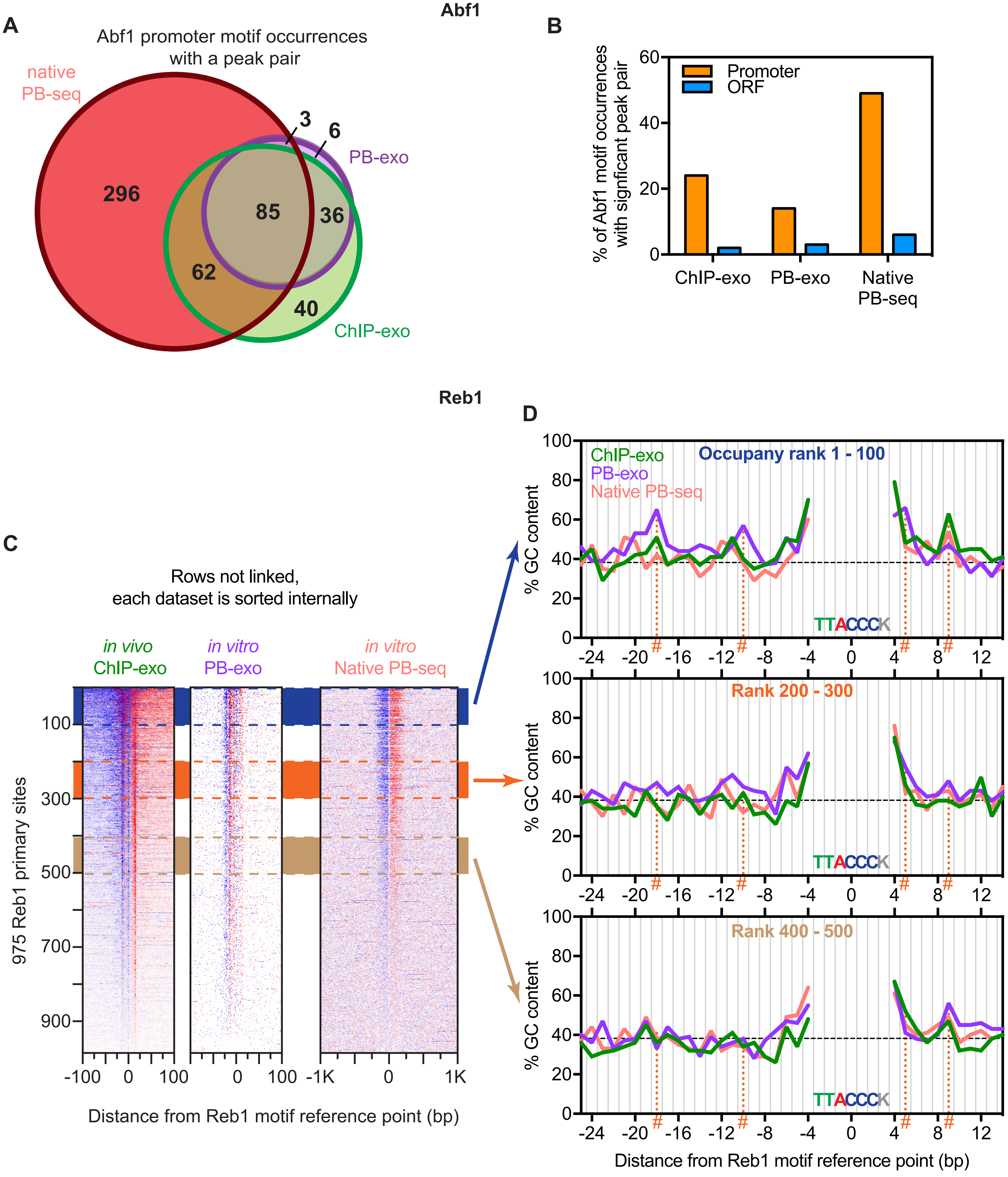
Unsupervised hierarchical clustering between Reb1 replicates was calculated with average Euclidean distance using the SciPy python package (Jones 2001).

Figure S1


A
Purified genomic DNA
Sonicate to ~200 bp
Add protein or whole-cell extract & crosslink
ChIP-exo protocol
ChIP
B
Nucleosomes around Cbf1 promoter sites
ChIP-exo & PB-exo bound
PB-exo only bound
not bound by Cbf1
H3 MNase tags
0
1
2
-340
-170
0
170
340
Distance from Cbf1 motif reference point (bp)
C
Nucleosomes around Rap1 promoter sites
ChIP-exo & PB-exo bound
PB-exo only bound
not bound by Rap1
H3 MNase tags
0
1
2
-340
-170
0
170
340
Distance from Rap1 motif reference point (bp)
D
Nucleosomes around Mcm1 promoter sites
ChIP-exo & PB-exo bound
PB-exo only bound
not bound by Mcm1
H3 MNase tags
0
1
2
-340
-170
0
170
340
Distance from Mcm1 motif reference point (bp)
E
Nucleosomes around Pho4 promoter sites
ChIP-seq & PB-exo bound
PB-exo only bound
not bound by Pho4
H3 MNase tags
0
1
2
-340
-170
0
170
340
Distance from Pho4 motif reference point (bp)
F
Nucleosomes around Abf1 promoter sites
Native PB-seq bound
not bound by Abf1
H3 MNase tags
0
1
2
-340
-170
0
170
340
Distance from Abf1 motif reference point (bp)

Supplemental Figure S1. *In vivo* GRF binding sites coincide with NFR midpoints.
(*A*) Scheme of PB-exo and WhIP-exo. Purified, sonicated genomic DNA is incubated with purified protein or whole cell extract and cross-linked with formaldehyde. That material is then immunoprecipitated and treated as starting material in the ChIP-exo assay.
(*B-F*) Composite plots of nucleosome midpoints generated by MNase H3 ChIP-seq located in promoters at different groups of motif occurrences for Cbf1 (*B*), Rap1 (*C*), Mcm1 (*D*), Pho4 (*E*), and Abf1 (*F*). Due to complications with Abf1 cross-linking (discussed in **Fig. 2**), Abf1 bound sites defined by ChIP-exo and PB-exo were ignored for this analysis. Instead, the native "PB-seq bound" motif occurrences represent our most comprehensive list of high confidence Abf1 binding sites.

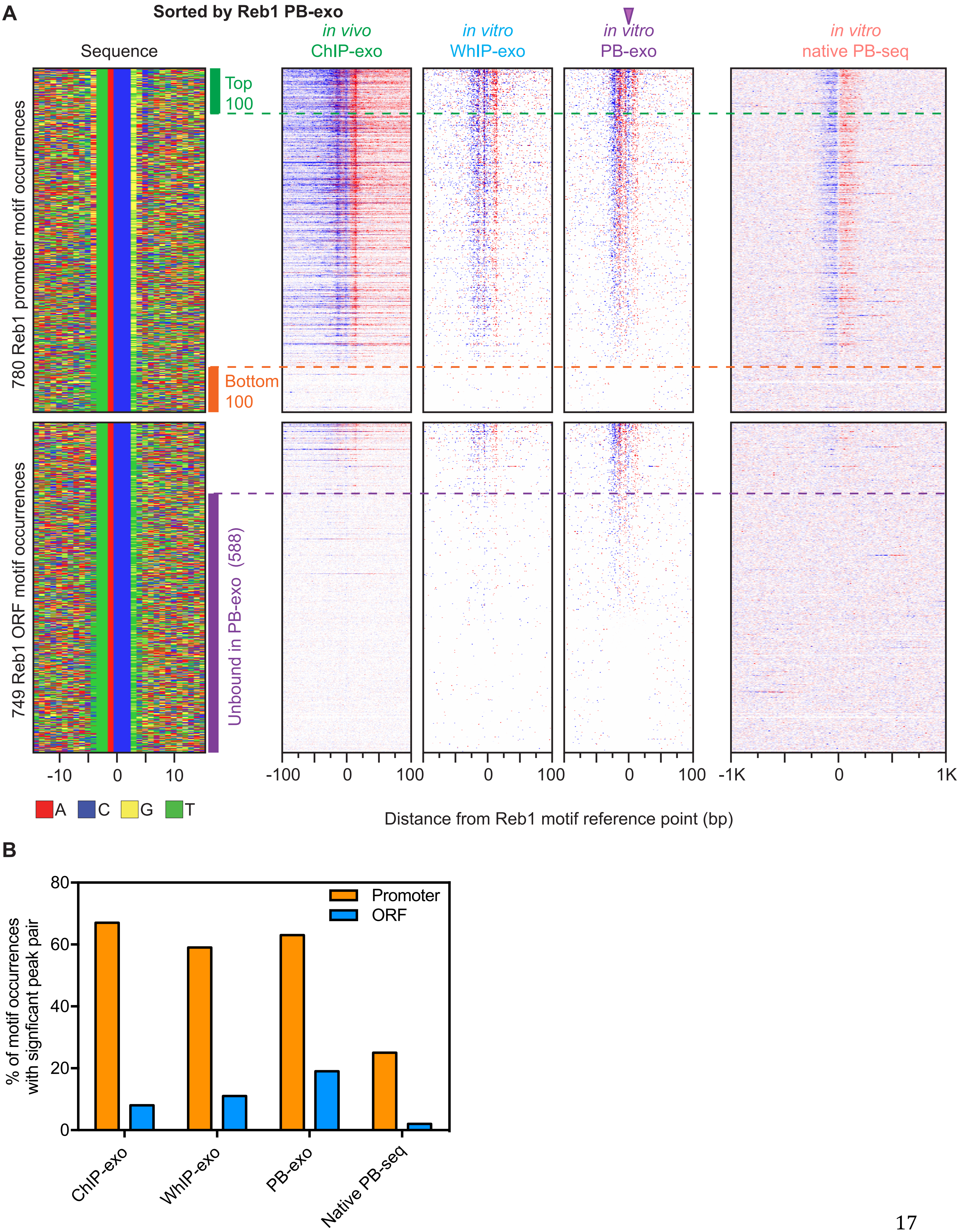
Figure S2



Abf1
A
Abf1 promoter motif occurrences with a peak pair
native PB-seq
PB-exo
ChIP-exo
296
3
6
85
36
62
40
B
% of Abf1 motif occurrences with signficant peak pair
Promoter
ORF
0
20
40
60
ChIP-exo
PB-exo
Native PB-seq
Reb1
C
Rows not linked, each dataset is sorted internally
in vivo ChIP-exo
in vitro PB-exo
in vitro Native PB-seq
975 Reb1 primary sites
100
300
500
700
900
-100
0
100
0
100
-1K
0
1K
Distance from Reb1 motif reference point (bp)
D
ChIP-exo
PB-exo
Native PB-seq
Occupany rank 1 - 100
Rank 200 - 300
Rank 400 - 500
% GC content
0
20
40
60
80
100
TTACCCK
#
-24
-20
-16
-12
-8
-4
0
4
8
12
Distance from Reb1 motif reference point (bp)

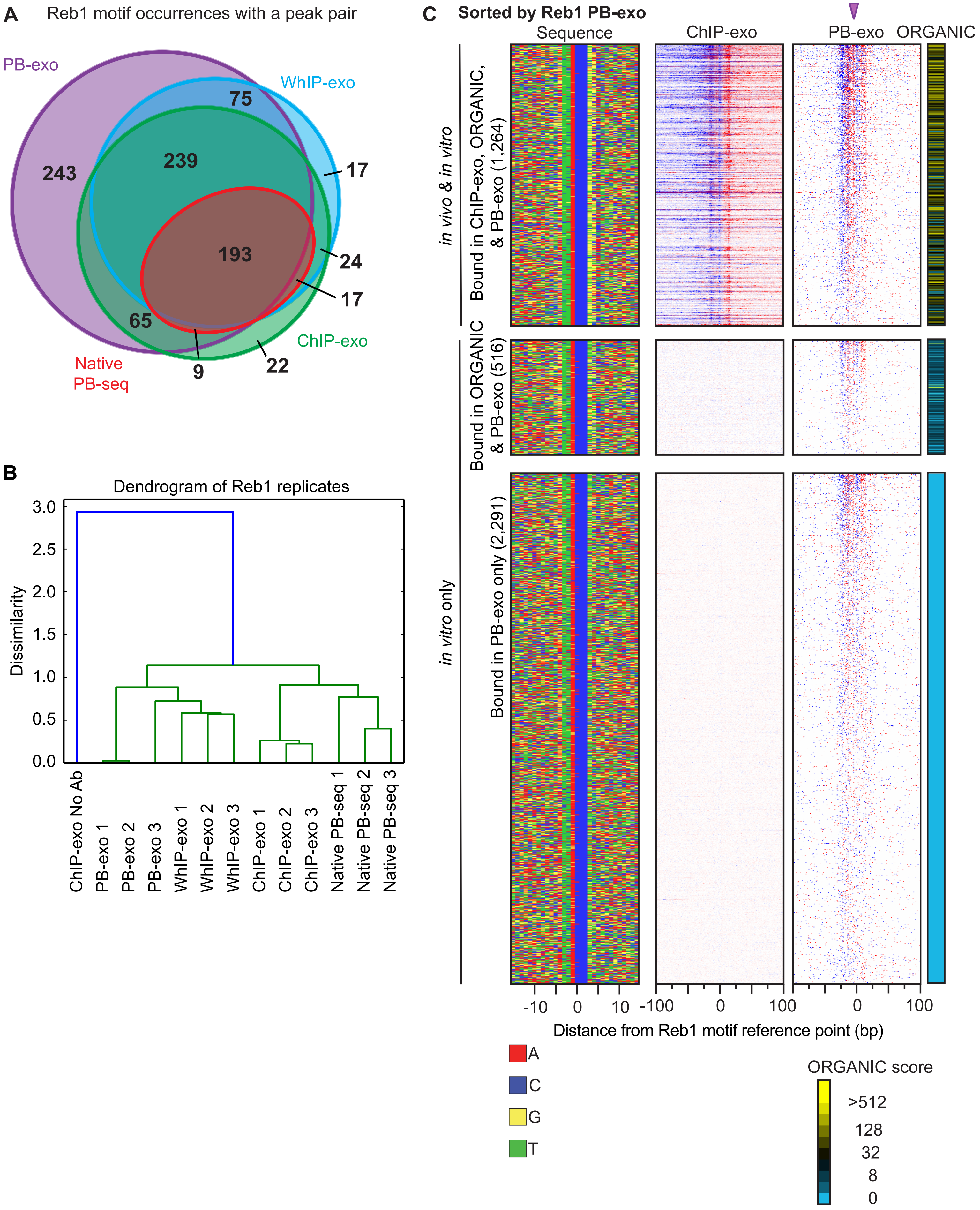
Supplemental Figure S2. Cross-linking specificity is mitigated in proteins with multiple cross-linking points. (*A*) Venn diagram representing the overlap of Abf1 sites that were bound in ChIP-exo, PB-exo, and native PB-seq. (*B*) Percentage of promoter (orange) or ORF (blue) Abf1 motif occurrences with significant peak pairs ±30 bp from the centered motif midpoint. (*C*) Heatmaps comparing ChIP-exo, PB-exo, and native PB-seq at 975 Reb1 primary sites (rows) (Rhee and Pugh 2011). Blue indicates tag 5' ends located on the motif strand, whereas red is on the opposite strand. Distances are from the motif midpoint. Datasets are sorted independently; rows are not linked. Colored bars and dashed lines highlight the rows rank ordered 1-100 (dark blue), 200 – 300 (orange), and 400 – 500 (tan) or each independently sorted dataset. (*D*) G/C frequency for sequences surrounding Reb1 motifs. Each panel is derived from datasets that were separately rank-ordered by Reb1 occupancy in (*C*), and having the indicated ranking range within their respective datasets. The dashed black line indicates the background G/C content. The orange hashtag and dashed lines represents the calculated cross-linking points. The top 100 occupied sites in ChIP-exo and PB-exo (top panel) had minor spikes in G/C content that corresponded with the calculated cross-linking sites (PB-exo at -18, -10, +5, and ChIP-exo at +9) seen in (**Fig. 1B**). G/C spikes were not evident within lower occupancy groups (bottom two line plots), indicating that the G/C-specificity of formaldehyde is realized to only a very minor extent with Reb1. This is likely due to there being at least four available cross-linking points in each binding event. The approximate probability of at least one of them being G/C is ~90% (assuming 40% G/C genome content and a 1 bp cross-link window). Thus, to a rough approximation the number of false negatives is expected to be low.

Figure S3


A
Sorted by Reb1 PB-exo
Sequence
in vivo
ChIP-exo
in vitro
WhIP-exo
in vitro
PB-exo
in vitro
native PB-seq
780 Reb1 promoter motif occurrences
Top 100
Bottom 100
749 Reb1 ORF motif occurrences
Unbound in PB-exo (588)
-10
0
10
-100
0
100
-1K
0
1K
A
C
G
T
Distance from Reb1 motif reference point (bp)
B
% of motif occurrences with signficant peak pair
0
20
40
60
80
Promoter
ORF
ChIP-exo
WhIP-exo
PB-exo
Native PB-seq

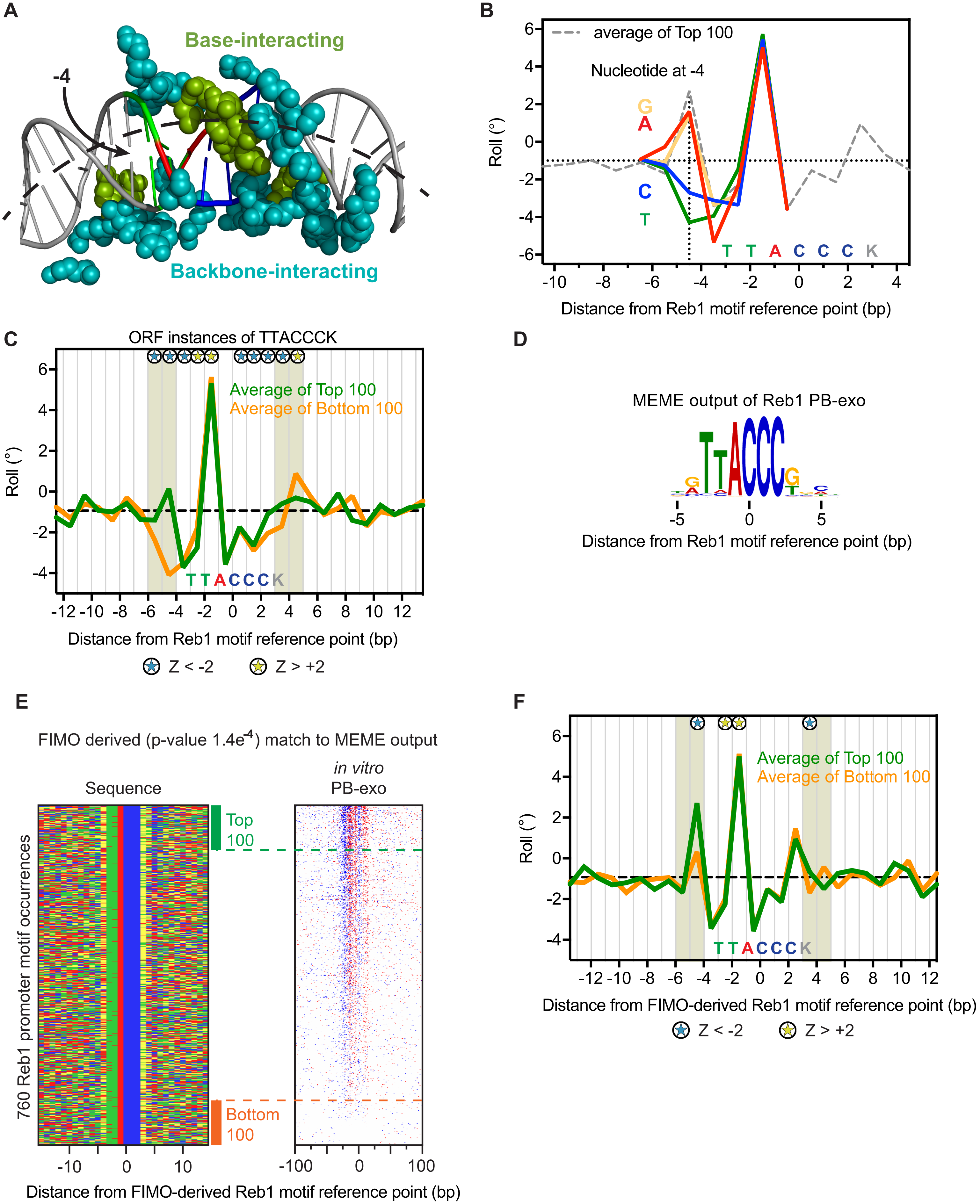
Supplemental Figure S3. Reb1 does not bind to every instance of the "perfect" core motif (TTACCCK)**.** (*A*) Four-color plot of sequences (left) centered on the Reb1 motif midpoint TTAC̲CCK for all instances within promoters (top) or ORFs (bottom). The remaining panels show tag 5' ends distributed around these motif occurrences. Panels were sorted by PB-exo tag counts (purple triangle), with dashed lines separating the top 100 (green) and bottom 100 (orange) occupied motifs. All motif occurrences below the purple dashed line were determined to be unbound in all assays. Note that sorting by the PB-exo data created the misleading appearance of additional enrichment in that dataset (below the dashed line) due to background values contributing to the sort. (*B*) Percentage of Reb1 motif occurrences in promoter (orange) or ORF (blue) regions across the *S. cerevisiae* genome with significant peak pairs (binding events) ±30 bp from the motif midpoint.

Figure S4


A
Reb1 motif occurrences with a peak pair
PB-exo
WhIP-exo
75
239
243
17
193
24
17
65
ChIP-exo
Native
PB-seq
9
22
B
Dendrogram of Reb1 replicates
Dissimilarity
3.0
2.5
2.0
1.5
1.0
0.5
0.0
ChIP-exo No Ab
PB-exo 1
PB-exo 2
PB-exo 3
WhIP-exo 1
WhIP-exo 2
WhIP-exo 3
ChIP-exo 1
ChIP-exo 2
ChIP-exo 3
Native PB-seq 1
Native PB-seq 2
Native PB-seq 3
C
Sorted by Reb1 PB-exo
Sequence
ChIP-exo
PB-exo
ORGANIC
in vivo & in vitro
Bound in ChIP-exo, ORGANIC, & PB-exo (1,264)
Bound in ORGANIC & PB-exo (516)
in vitro only
Bound in PB-exo only (2,291)
-10
0
10
-100
0
100
0
100
Distance from Reb1 motif reference point (bp)
A
C
G
T
ORGANIC score
>512
128
32
8
0

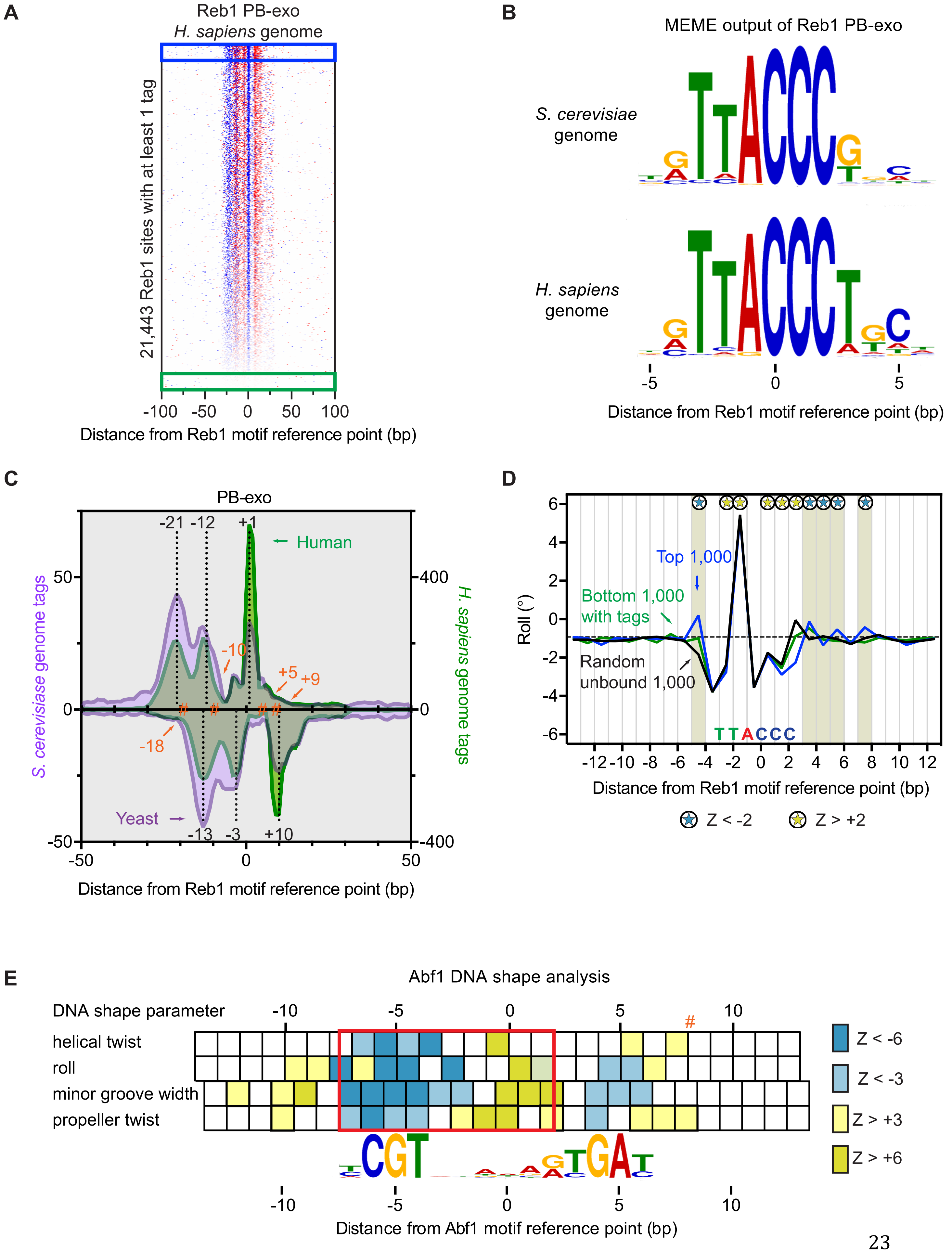
Supplemental Figure S4. Comparing sites identified as bound by Reb1 *in vitro* and *in vivo*. (*A*) Venn diagram representing the overlap of Reb1 sites that were bound in ChIP-exo, WhIP-exo, PB-exo, and native PB-seq. (*B*) Dendrogram of Reb1 replicates for each assay used in this study. Dissimilarity is based on tag counts at the Reb1 motif occurrences in Fig. S3A using unsupervised hierarchical clustering. (*C*) Comparison of all Reb1 sites (FIMO $p < 0.001$) bound in PB-exo to ChIP-exo and ORGANIC (Kasinathan et al. 2014). Four-color plot of sequences (left) centered on the Reb1 motif midpoint. The far-right column of panels displayed the “ORGANIC score” for the associated Reb1 motif occurrence. Motif occurrences were grouped based on binding classification in the various assays. Rows were sorted (purple triangle) by Reb1 PB-exo tag counts in each panel. Rows are linked.

Figure S5


A
Base-interacting
-4
Backbone-interacting
B
average of Top 100
Nucleotide at -4
G
A
C
T
T T A C C C K
Roll (°)
6
4
2
0
-2
-4
-6
-10
-8
-6
-4
-2
0
2
4
Distance from Reb1 motif reference point (bp)
C
ORF instances of TTACCCK
Average of Top 100
Average of Bottom 100
T T A C C C K
Roll (°)
-12 -10 -8 -6 -4 -2 0 2 4 6 8 10 12
Distance from Reb1 motif reference point (bp)
Z < -2
Z > +2
D
MEME output of Reb1 PB-exo
-5
0
5
Distance from Reb1 motif reference point (bp)
E
FIMO derived (p-value 1.4e^4) match to MEME output
Sequence
in vitro
PB-exo
760 Reb1 promoter motif occurrences
Top 100
Bottom 100
-10
0
10
-100
0
100
Distance from FIMO-derived Reb1 motif reference point (bp)
F
Average of Top 100
Average of Bottom 100
T T A C C C K
Roll (°)
-12 -10 -8 -6 -4 -2 0 2 4 6 8 10 12
Distance from FIMO-derived Reb1 motif reference point (bp)
Z < -2
Z > +2

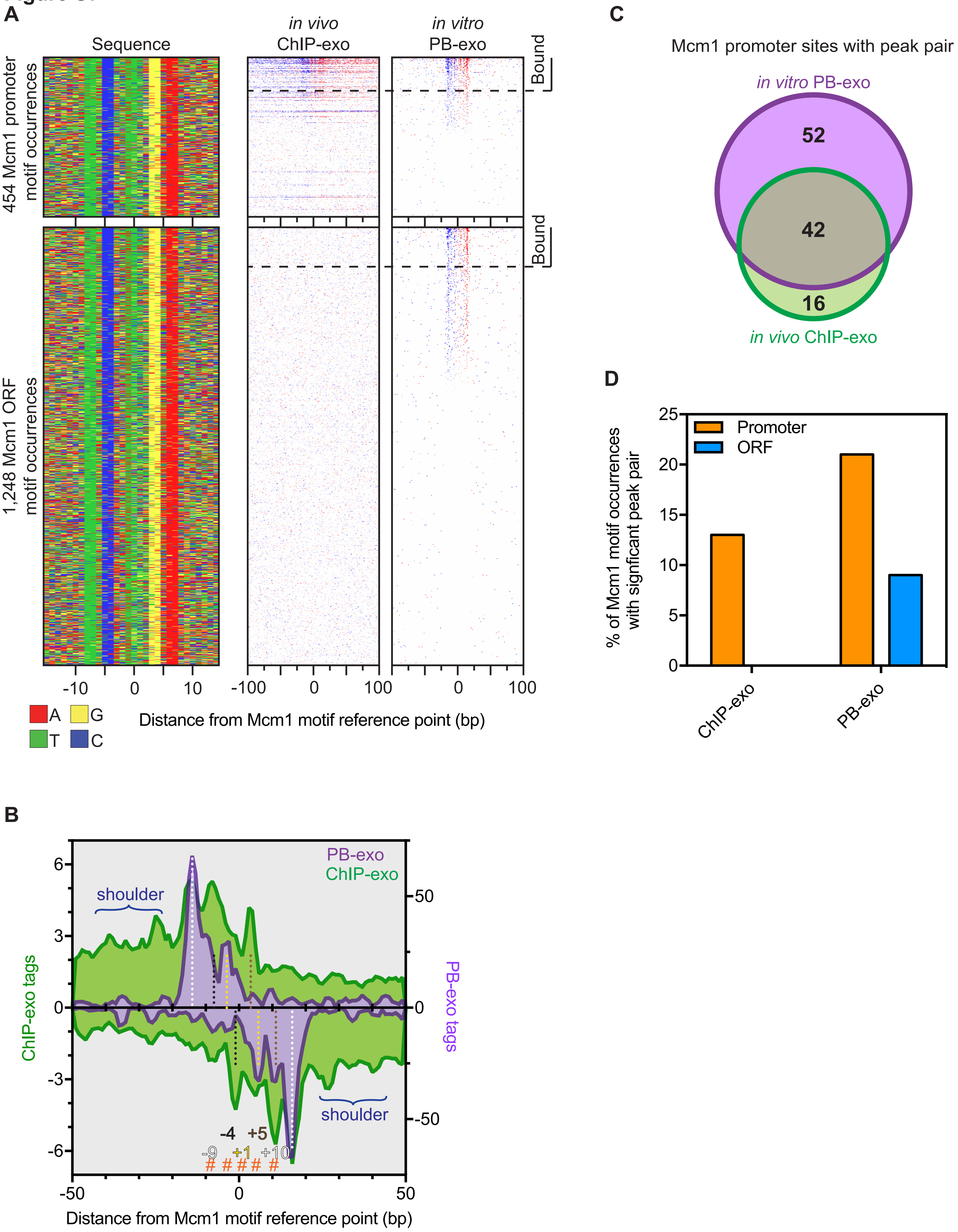
Supplemental Figure S5. Crystal structure of *S. pombe* Reb1 supports the role played by DNA shape in Reb1 binding. (*A*) Crystallographic-based model of amino acids involved in recognizing the Reb1 motif as reported in (Jaiswal et al. 2016) (PBD ID: 5EYB). Conserved base pairs in the Reb1 motif are colored, based on the composition of the top strand shown in the 5' to 3' direction: adenine (red), cytosine (blue), and thymine (green). Amino acids that make base-specific interactions (olive green spheres) are concentrated at the conserved nucleotides, whereas amino acids making DNA backbone interactions (cyan spheres) cluster at the motif flanks. The black dashed line indicates the approximate DNA path relative to the arc of a circle. (*B*) Line plots of variations in roll for Reb1 motifs when the nucleotide at the -4 position is varied. For each nucleotide variant, only the pentamer centered at the -4 position of the Reb1 motif is shown. These lines are shown in comparison to the average of the top 100 bound motifs (gray dashed line) in **Fig. 3B**. On average, variants of the Reb1 motif with A or G in the -4 position have intrinsic positive roll at the 5' end of the motif, whereas T in that position produces the most negative roll. Since these plots are of averages, they do not show that there are specific combinations of A or G in the -4 position that produce negative roll that would be unfavorable to Reb1 binding. However, all T variants produce negative roll at this position. (*C*) Line plots of variations in roll for top 100 (green) and bottom 100 (bottom) ORF motif occurrences defined by the sort in *Supplemental Fig. S3A*. The dashed black line indicates the median roll of all DNA sequences. The position of the consensus Reb1 motif is labeled along the x-axis. Blue and yellow stars represent positions with significant positive or negative roll (|Z|>2, Mann-Whitney *U* test), respectively, for the top 100 sites compared to the bottom 100 sites. The tan shaded areas highlight the nucleotides outside the core motif with significant differences in DNA shape in **Fig. 3C**. (*D*) MEME logo obtained from the top 500 peak-pairs from Reb1 PB-exo. (*E*) Four-color plot of sequences (left) centered on the Reb1 motif midpoint as defined by the MEME output for all instances within promoters (see panel *D*). The right panel shows tag 5' ends distributed around these motif occurrences. Panels were sorted by PB-exo tag counts, with dashed lines separating the top 100 (green) and bottom 100 (orange) occupied motifs. Importantly, this revised motif is closer to VTTA<u>C</u>CCKNH than TTA<u>C</u>CCK, as expected, and thus is more enriched with any shape component present within this motif. Consequently, the unbound or lowly occupied occurrences of this motif (thresholded at 1.4e^{-4}) will on average have an intrinsic shape that is closer to the bound version than that defined by the unbiased TTACCCK set. Nevertheless, we observed a statistical difference specifically at position -4 between bound and unbound motifs (see panel *F*). (*F*) Line plots of variations in roll for top 100 (green) and bottom 100 (bottom) promoter motif occurrences defined by the sort in panel **E**. Annotation descriptions are the same as in panel **C**.

Figure S6


A
Reb1 PB-exo
H. sapiens genome
21,443 Reb1 sites with at least 1 tag
-100 -50 0 50 100
Distance from Reb1 motif reference point (bp)
B
MEME output of Reb1 PB-exo
S. cerevisiae genome
H. sapiens genome
-5 0 5
Distance from Reb1 motif reference point (bp)
C
PB-exo
-21 -12 +1
Human
-10
+5 +9
-18
Yeast
-13 -3 +10
# # # #
S. cerevisiae genome tags
H. sapiens genome tags
50 0 -50
400 0 -400
-50 0 50
Distance from Reb1 motif reference point (bp)
D
Top 1,000
Bottom 1,000 with tags
Random unbound 1,000
TTACCC
Roll (°)
6 4 2 0 -2 -4 -6
-12 -10 -8 -6 -4 -2 0 2 4 6 8 10 12
Distance from Reb1 motif reference point (bp)
Z < -2
Z > +2
E
Abf1 DNA shape analysis
DNA shape parameter
-10 -5 0 5 10
#
helical twist
roll
minor groove width
propeller twist
Z < -6
Z < -3
Z > +3
Z > +6
-10 -5 0 5 10
Distance from Abf1 motif reference point (bp)

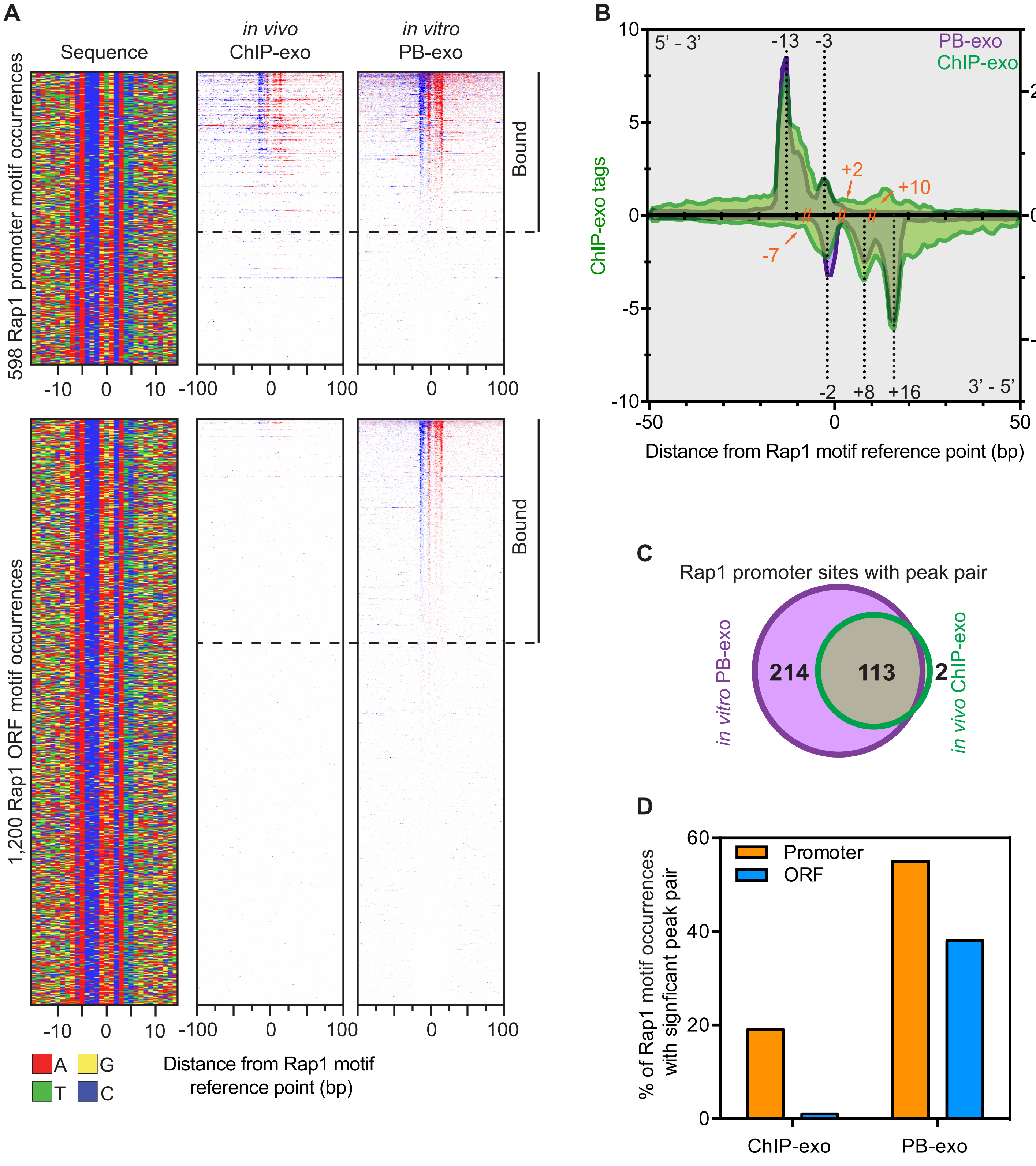
Supplemental Figure S6. *S. cerevisiae* Reb1 binding to *H. sapiens* DNA. (*A*) Tags distributed around Reb1 motif occurrences in the *H. sapiens* genome with at least one tag ±30 bp from the centered motif midpoint (underlined “C” in TTACCC). The blue and green boxes indicate the top 1,000 and bottom 1,000 rows, respectively, used in panel D. (*B*) MEME logos obtained from the top 1,000 peaks from Reb1 PB-exo using the S. *cerevisiae* or *H. sapiens* genome as substrate. Note that the change in preference for thymine over guanine at position +3 in the *H. sapiens* dataset is accounted for by the well know deficiency of CpG dinucleotides in human DNA (Li and Zhang 2014). (*C*) Composite tag 5’ ends for PB-exo using the *S. cerevisiae* (purple) or *H. sapiens* (green) genome as a substrate. Density above the x-axis represents tags on the same strand as the motif, and density below represents opposite strand tags. The orange hashtags represent prominent cross-linking sites calculated in **Fig. 1B**. Dashed lines represent the same PB-exo peaks as in **Fig. 1B**. (*D*) Line plots of variations in roll for Reb1 motif occurrences within the *H. sapiens* genome. The blue trace represents the average roll from the top 1,000 sites based on tag count as in panel A. The green trace represents the bottom 1,000 sites with non-zero tag counts, and the black trace represents 1,000 sites with no tags that still contain the core Reb1 motif (TTACCC). Blue and yellow stars represent positions with significant positive and negative roll (|Z|>2, Mann-Whitney U test), respectively, for the top 1,000 sites compared to a random 1,000 sites. The tan shading designates the positions where the sequence was allowed to vary and possessed significant differences in DNA shape. The position of the consensus Reb1 motif is labeled along the x-axis. (*E*) Heatmap representation of all four DNA shape values measured for Abf1. Colored boxes represent positions with significant Z-score deviations in DNA shape (Z-score, Mann-Whitney U test) for the combined sets of top 50 sites compared to the bottom 50 sites of the eight specific sequences listed in **Fig. 3E**. The position of the consensus Abf1 motif is labeled along the x-axis. The red box highlights the region of the motif with the greatest concentration of significant positions across all four DNA shape parameters. Helical twist and roll refer to values that occur between base pairs, and so are positioned as such.

Figure S7


A
Sequence
in vivo
ChIP-exo
in vitro
PB-exo
Bound
454 Mcm1 promoter
motif occurrences
1,248 Mcm1 ORF
motif occurrences
Bound
-10
0
10
-100
0
100
0
100
A
G
T
C
Distance from Mcm1 motif reference point (bp)
C
Mcm1 promoter sites with peak pair
in vitro PB-exo
52
42
16
in vivo ChIP-exo
D
% of Mcm1 motif occurrences
with signficant peak pair
25
20
15
10
5
0
Promoter
ORF
ChIP-exo
PB-exo
B
PB-exo
ChIP-exo
shoulder
shoulder
ChIP-exo tags
PB-exo tags
6
3
0
-3
-6
50
0
-50
-4
+5
-9
+1
+10
# # # # #
-50
0
50
Distance from Mcm1 motif reference point (bp)

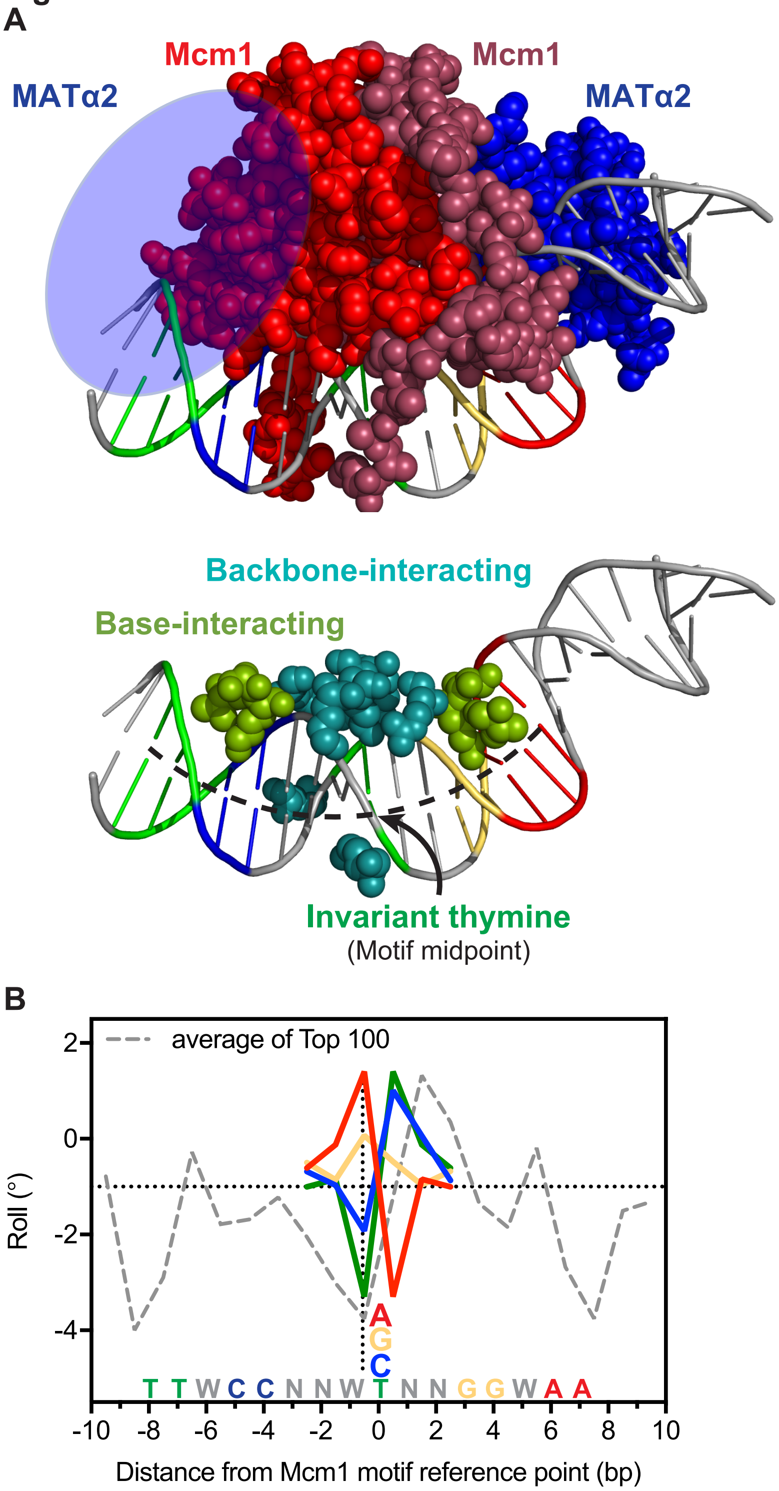
Supplemental Figure S7. Genome-wide *in vitro* Mcm1 binding locations. (*A*) The left panels show a four-color plot of 30-bp sequences centered on the Mcm1 motif midpoint. Each row represents a motif occurrence that passed our FIMO p-value threshold ($p<6.33e^{-5}$). The remaining panels show tags distributed around motif occurrences located in promoters (top) or ORFs (bottom) for each assay and sorted by PB-exo tag counts. Datasets are linked. (*B*) Composite tag count 5' ends for Mcm1 ChIP-exo (green) and PB-exo (purple) at Mcm1 promoter motif occurrences. Annotation descriptions are the same as in **Fig. 1B**, except applied to Mcm1. Dotted lines of the same color represent opposite-stranded paired peaks, that correspond to exonuclease stop sites around a common crosslink. The white dotted lines near the flanks represent peaks that do not have a visible pair in the composite, likely due to exonuclease encountering four more-5' blocks, which would prevent it from reaching the missing site. Additional flanking interactions ("shoulders") are evident *in vivo* that are absent *in vitro*; this density is consistent with Mcm1/MAT protein interactions. (*C*) Venn diagram representing the overlap of Mcm1 promoter sites identified as being bound in ChIP-exo and/or in PB-exo. (*D*) Percentage of promoter (orange) or ORF (blue) Mcm1 motif occurrences with significant peak pairs within ±30 bp from the centered motif midpoint.

Figure S8


A
Sequence
in vivo
ChIP-exo
in vitro
PB-exo
598 Rap1 promoter motif occurrences
Bound
-10
0
10
-100
0
100
0
100
1,200 Rap1 ORF motif occurrences
Bound
-10
0
10
-100
0
100
0
100
A
G
T
C
Distance from Rap1 motif reference point (bp)
B
5' - 3'
-13
-3
PB-exo
ChIP-exo
+2
+10
-7
-2
+8
+16
3' - 5'
ChIP-exo tags
PB-exo tags
10
5
0
-5
-10
20
0
-20
-50
0
50
Distance from Rap1 motif reference point (bp)
C
Rap1 promoter sites with peak pair
in vitro PB-exo
in vivo ChIP-exo
214
113
2
D
Promoter
ORF
% of Rap1 motif occurrences with signficant peak pair
60
40
20
0
ChIP-exo
PB-exo

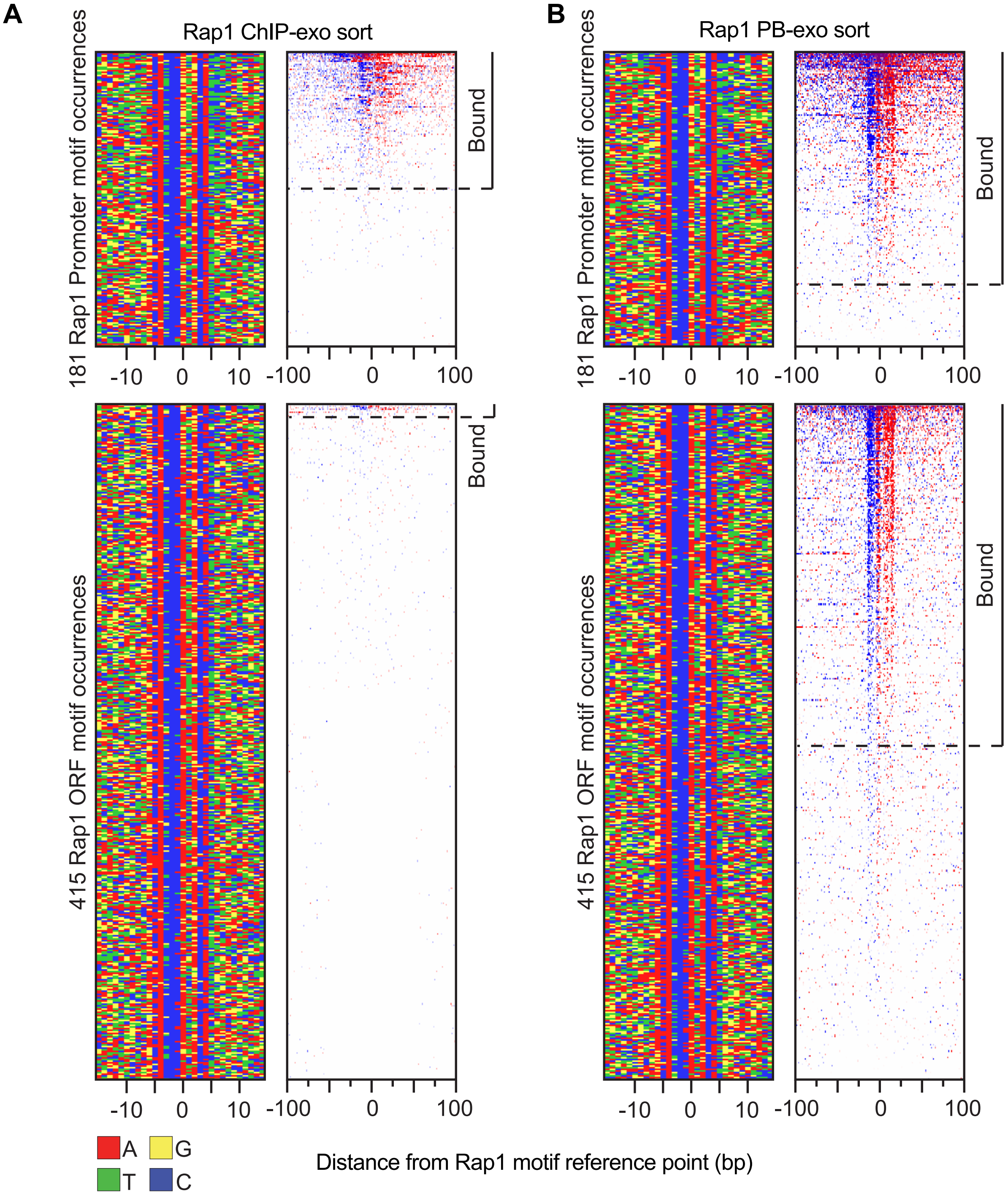
Supplemental Figure S8. Genome-wide *in vitro* Rap1 binding locations. (*A*) The left panels show a four-color plot of 30-bp sequences centered on the Rap1 motif midpoint. Each row represents a motif occurrence that passed our FIMO p-value threshold ($p<1.4e^{-4}$). The remaining panels show tags distributed around motif occurrences located in promoters (top) or ORFs (bottom) for each assay and sorted by PB-exo tag counts. Datasets are linked. (*B*) Composite tag count 5' ends for Rap1 ChIP-exo (green) and PB-exo (purple) at Rap1 promoter motif occurrences. Annotation descriptions are the same as in **Fig. 1B**, except applied to Rap1. (*C*) Venn diagram representing the overlap of Rap1 sites that were bound in ChIP-exo and PB-exo. (*D*) Percentage of promoter (orange) or ORF (blue) Rap1 motif occurrences with significant peak pairs ±30 bp from the centered motif midpoint.

Figure S9


A
Mcm1
Mcm1
MATα2
MATα2
Backbone-interacting
Base-interacting
Invariant thymine
(Motif midpoint)
B
average of Top 100
2
0
-2
-4
Roll (°)
A
G
C
T T W C C N N W T N N G G W A A
-10
-8
-6
-4
-2
0
2
4
6
8
10
Distance from Mcm1 motif reference point (bp)

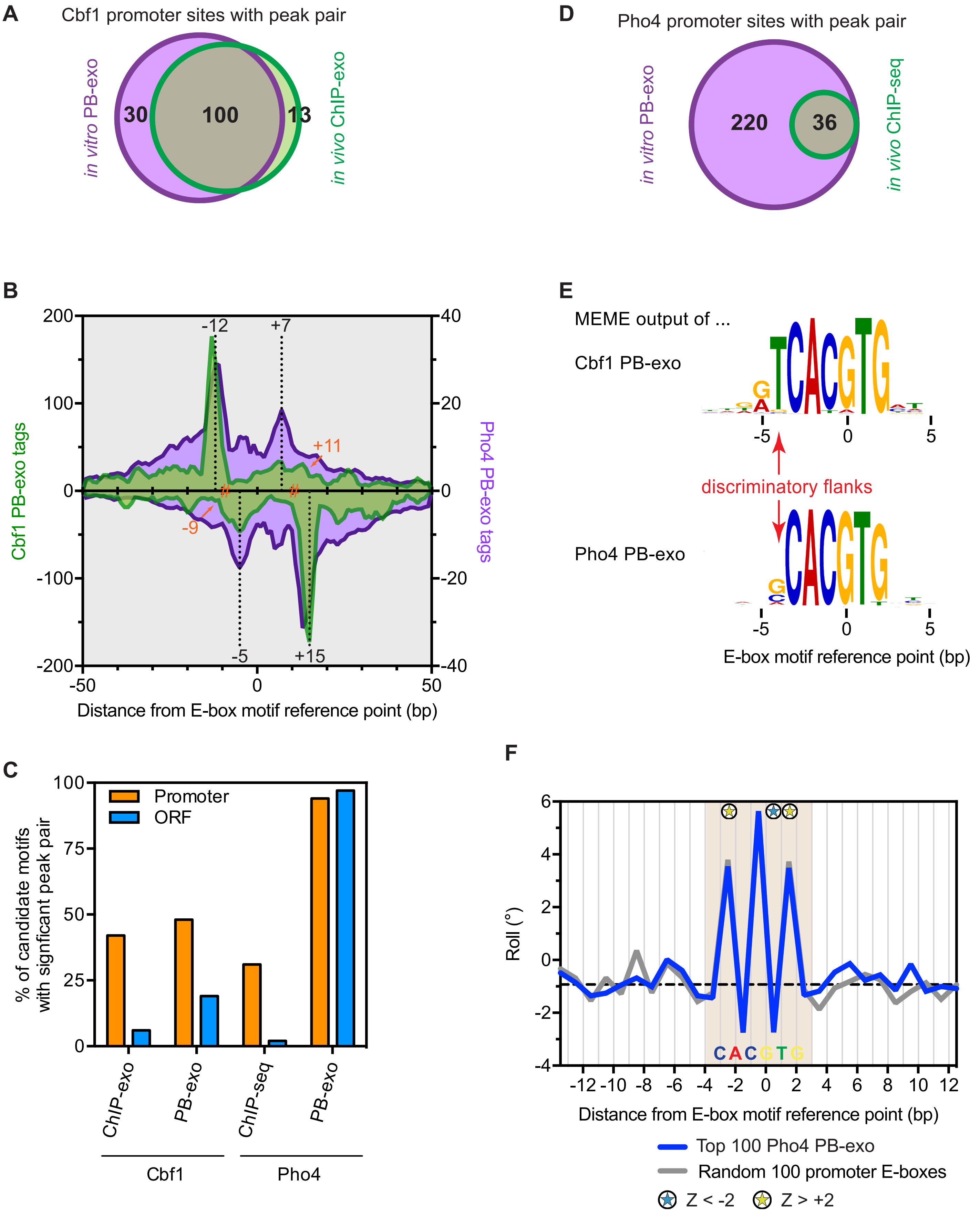
Supplemental Figure S9. Observations from the crystal structures of Mcm1 bound to DNA are consistent with the results of DNA shape analysis. (*A*) Upper panel: Crystallographic-based model of the Mcm1 (red and rose) and MATα2 (blue) heterodimer bound to its motif (Tan and Richmond 1998) (PBD ID: 1MNM). Blue oval represents the presumed location of the second MATα2 monomer if the construct had contained DNA that was long enough for the complex to bind. Conserved base pairs in the Mcm1 motif are colored based on the composition of the top strand shown in the 5' to 3' direction: adenine (red), cytosine (blue), guanine (yellow), and thymine (green). Lower panel: Only amino acids of Mcm1 that make direct contact with the DNA molecule are shown. Amino acids that make base-specific interactions (olive green) are concentrated at the two conserved regions of the motif. Amino acids that make DNA backbone interactions (cyan) are concentrated at the degenerate central core of the motif. The black arrow indicates the conserved thymine at the motif midpoint. The black dashed line emphasizes the bend in the DNA path. (*B*) Line plots of variations in roll for Mcm1 motifs when the nucleotide at the "0" position is varied. For each nucleotide variant, only the region of the Mcm1 motif that is influenced by the "0" position is shown. These lines are shown in comparison to the average of the top 100 motifs (gray line) bound in *Supplemental Fig. S7A*.

Figure S10


A
Rap1 ChIP-exo sort
181 Rap1 Promoter motif occurrences
Bound
-10
0
10
-100
0
100
415 Rap1 ORF motif occurrences
Bound
-10
0
10
-100
0
100
A
G
T
C
Distance from Rap1 motif reference point (bp)
B
Rap1 PB-exo sort
181 Rap1 Promoter motif occurrences
Bound
-10
0
10
-100
0
100
415 Rap1 ORF motif occurrences
Bound
-10
0
10
-100
0
100

Supplemental Figure S10. Rap1 analysis by motif p-value demonstrates that weak *in vitro* ORF Rap1 binding sites are devoid of Rap1 binding *in vivo*. Rap1 data is presented as in **Fig. 5A**, but was restricted to sites with a motif p-value between 2x10^{-5} and 6x10^{-5}. Sites were separated based on location in the promoter (top panels) or ORF (bottom panels), and then sorted by ChIP-exo (*A*) or PB-exo (*B*) tag counts. In this range, only 2% of ORF sites were bound *in vivo*, but 50% of promoter sites were bound (black dashed line). Approximately 50% of the ORF sites were also bound *in vitro*, which is consistent with the idea that Rap1 binding is blocked at these weaker sites *in vivo* by nucleosomes. (*C*) Crystallographic-based model of amino acids involved in recognizing the Rap1 motif as reported in (Le Bihan et al. 2013) (PBD ID: 4GFB). Conserved base pairs in the Rap1 motif are colored based on the composition of the top strand shown in the 5' to 3' direction: adenine (red), cytosine (blue). Amino acids that make base-specific interactions (olive green) are concentrated at the two conserved regions of the motif and the amino acids that make DNA backbone interactions (cyan) are concentrated at the degenerate core of the motif and well as in the flanking regions.

Figure S11


A
Cbf1 promoter sites with peak pair
in vitro PB-exo
in vivo ChIP-exo
30
100
13
B
Cbf1 PB-exo tags
Pho4 PB-exo tags
200
100
0
-100
-200
40
20
0
-20
-40
-12
+7
+11
-9
#
#
-5
+15
-50
0
50
Distance from E-box motif reference point (bp)
C
% of candidate motifs with signficant peak pair
100
75
50
25
0
Promoter
ORF
ChIP-exo
PB-exo
ChIP-seq
PB-exo
Cbf1
Pho4
D
Pho4 promoter sites with peak pair
in vitro PB-exo
in vivo ChIP-seq
220
36
E
MEME output of ...
Cbf1 PB-exo
Pho4 PB-exo
discriminatory flanks
-5
0
5
E-box motif reference point (bp)
F
Roll (°)
6
4
2
0
-2
-4
CACGTG
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
Distance from E-box motif reference point (bp)
Top 100 Pho4 PB-exo
Random 100 promoter E-boxes
Z < -2
Z > +2

Supplemental Figure S11. Despite strong overlap *in vitro*, Cbf1 and Pho4 bind different sets of sites *in vivo*. (*A*) Venn diagram representing the overlap of Cbf1 sites that were bound in ChIP-exo and PB-exo. (*B*) Composite tag 5' ends for Cbf1 PB-exo (green) and Pho4 PB-exo (purple) at E-box promoter motif occurrences. Density above the x-axis represents motif strand tags, and density below represents opposite strand tags. The orange hashtags represent prominent cross-linking points calculated by pairing adjacent peaks above and below the x-axis in the PB-exo plots. Dashed black lines represent peaks that are common in Cbf1 and Pho4 PB-exo. (*C*) Percentage of promoter (orange) or ORF (blue) E-box motif occurrences with significant Cbf1 or Pho4 peak pairs within ±30 bp from the centered motif midpoint. (*D*) Venn diagram representing the overlap of Pho4 sites, determined by ChIP-seq, that were bound under phosphate starvation conditions (Zhou and O'Shea 2011) compared to *in vitro* bound determined by PB-exo. (*E*) MEME logos obtained from the top 500 peak-pairs from Cbf1 and Pho4 PB-exo. (*F*) Line plots of variations in roll for the top 100 Pho4 PB-exo-bound E-box motif occurrences (blue) versus 100 random promoter E-boxes (gray). The dashed black line indicates the genome-wide median. Blue and yellow stars represent positions with significant large or small roll (|Z|>2, Mann-Whitney *U* test), respectively. The only positions calculated to be significantly different overlapped with the consensus sequence (tan shaded area), which suggests those differences are an artifact of the analysis and not biologically meaningful.

REFERENCES

Albert I, Wachi S, Jiang C, Pugh BF. 2008. GeneTrack--a genomic data processing and visualization framework. *Bioinformatics* **24**: 1305-1306.

Batta K, Zhang Z, Yen K, Goffman DB, Pugh BF. 2011. Genome-wide function of H2B ubiquitylation in promoter and genic regions. *Genes Dev* **25**: 2254-2265.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **40**: D700-705.

Cristea IM, Chait BT. 2011. Conjugation of magnetic beads for immunopurification of protein complexes. *Cold Spring Harb Protoc* **2011**: pdb prot5610.

Gordan R, Shen N, Dror I, Zhou T, Horton J, Rohs R, Bulyk ML. 2013. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep* **3**: 1093-1104.

Guertin MJ, Lis JT. 2013. Mechanisms by which transcription factors gain access to target sequence elements in chromatin. *Curr Opin Genet Dev* **23**: 116-123.

Jaiswal R, Choudhury M, Zaman S, Singh S, Santosh V, Bastia D, Escalante CR. 2016. Functional architecture of the Reb1-Ter complex of Schizosaccharomyces pombe. *Proc Natl Acad Sci U S A* **113**: E2267-2276.

Jones EO, T.; Peterson P. 2001. SciPy: Open source scientific tools for Python. In *http://wwwscipyorg/*.

Kasinathan S, Orsi GA, Zentner GE, Ahmad K, Henikoff S. 2014. High-resolution mapping of transcription factor binding sites on native chromatin. *Nat Methods* **11**: 203-209.

Le Bihan YV, Matot B, Pietrement O, Giraud-Panis MJ, Gasparini S, Le Cam E, Gilson E, Sclavi B, Miron S, Le Du MH. 2013. Effect of Rap1 binding on DNA distortion and potassium permanganate hypersensitivity. *Acta Crystallogr D Biol Crystallogr* **69**: 409-419.

Li E, Zhang Y. 2014. DNA methylation in mammals. *Cold Spring Harb Perspect Biol* **6**: a019133.

Li H. 2013. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**: 1005-1010.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**: 1408-1419.

Rhee HS, Pugh BF. 2012. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol* **Chapter 21**: Unit 21 24.

Tan S, Richmond TJ. 1998. Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. *Nature* **391**: 660-666.

Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, Di Felice R, Rohs R. 2013. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* **41**: W56-62.

Zhou X, O'Shea EK. 2011. Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Mol Cell* **42**: 826-836.