Supplementary Information:  
Conserved nucleosome positioning defines replication origins

Matthew L. Eaton†, Kyriaki Galani‡, Sukhyun Kang,  
Stephen P. Bell, and David M. MacAlpine∗

†Department of Pharmacology and Cancer Biology  
Duke University Medical Center, Durham, NC 27710 USA

‡Howard Hughes Medical Institute, Department of Biology  
Massachusetts Institute of Technology, Cambridge, MA 02139

†These authors contributed equally.

∗To whom correspondence should be addressed; E-mail: david.macalpine@duke.edu

March 19, 2010
Supplementary Methods

Alignment of Illumina reads to reference genome

To map the genomic locations of Illumina reads, we used MAQ (Li et al. 2008) with a maximum mismatch of 3 bases and a minimum Phred quality score of 35.

Mapping ORC binding sites by ChIP-seq

We used an approach similar to that of MACS (Zhang et al. 2008) to identify ORC binding sites in the sequencing data. Briefly, for each replicate we scanned the genome with a window of size wsize (set to twice the estimated sonication fragment size) for areas with a high enrichment of reads (areas with greater than 3x the read density than expected by average based on chromosome size and the number of reads that mapped to it). Consecutive windows were combined. The negative reads within these windows were then shifted upstream until the maximum Pearson correlation between the positive and negative stranded reads within the window was found. The median optimal shift over all areas of enrichment was recorded as the expected ChIP-seq fragment size f. We then shifted the positive stranded reads downstream by \( \frac{f}{2} \) and the negative stranded reads upstream by \( \frac{f}{2} \) to create a read pile-up landscape. We then scanned the genome with a window size of \( 2f \), looking for areas that are significantly enriched above background. Background signal was estimated by sampling the distribution of reads within transcribed regions (David et al. 2006) and modeling that distribution as a negative binomial (the probability density function which we empirically determined to be the best fit for the tag distribution within ORFs). Areas with a p-value \( \leq 10^{-6} \) under the background distribution were identified as putative peaks of ORC enrichment.
Comparison of ORC ChIP-seq peaks with prior ChIP-chip peaks

To compare the ORC ChIP-seq peaks with prior ChIP-chip peaks (Xu et al. 2006), we first screened out any ORC peaks not identified in both replicates, leaving us with a total of 267 peaks. We then compared these enriched regions to the 396 regions identified in (Xu et al. 2006) as being ORC-only sites (47) or ORC-MCM sites (349). Our criterion for a match was simply that the areas of enrichment overlap. We found that 241 of the sites from (Xu et al. 2006) overlapped with 258 sites identified in this study. The lower resolution of the tiling arrays (especially in telomeric regions), contributed to the ChIP-chip study calling some very wide peaks, which accounts for the occasional single ChIP-chip peak overlapping multiple ChIP-seq peaks. Their analysis of ORC and MCM binding resulted in a set of predicted ARSs called the nimARS set. This set contains 529 areas of ORC and/or MCM enrichment. ACS calls (the nimACS set) were made at a subset of these nimARSs. Of the 396 ORC-enriched regions, 268 were included in their final set of 370 nimARSs for which an ACS call was made (the 102 other regions in the nimARS set for which an ACS was called were enriched only for MCM). Of this set of 268 ORC-enriched regions included in the nimARS set, 86 had more than one ACS call in their final nimACS set (the final set published in their supplemental materials had 504 unique ACS calls, corresponding to 370 nimARSs).

Identification of functional ACS matches

To locate the most likely ACS motif match bound by ORC for each ChIP-seq peak, we first searched for a common motif between the top 100 ChIP-seq peaks using cosmo (Bembom et al. 2007), an R (R Development Core Team 2008) implementation of MEME (Bailey and Elkan 1994). The motif we found is nearly identical to that shown in Figure 1C. We then scanned the genome, taking a log-odds ra-
ratio of the probability of matching the motif over the probability of matching yeast background sequence (represented by a 4th-order Markov chain trained on the entire yeast genome) at every position. We selected a minimum log-odds ratio cutoff of 4, which ruled out approximately 99.9% of genomic positions as potential ORC binding sites. We then estimated the parameters of a Gaussian distribution from the positions of the shifted reads for each ORC ChIP-seq peak. To select the most likely ACS per peak, we took the log-space product of the log of the density of this Gaussian to the log-odds ratio of the remaining ORC motif matches under each peak. We took the local maximum of this score under each peak to be the most likely ACS bound by ORC. Any ORC peak which didn’t have an ORC motif match above the log-odds ratio cutoff was discarded. We found three examples of ORC peaks in the first replicate which did not have an ORC motif match. These were weak peaks which were not reproduced in the second replicate.

ChIP-seq replicates were combined by accepting only those ACS calls which were called identically in both experiments. This set, the intersection of the ACS calls for our replicates, is the final ORC-ACS set.

**Mapping nucleosome positions**

To identify the location of stably positioned nucleosomes genome-wide, we first mapped the mononucleosomal micrococcal nuclease digested fragments back to the yeast genome using MAQ as described above. We then estimated an experimental mononucleosomal fragment size \( f \) by identifying the optimal shift between positive and negative reads which maximized the Pearson correlation between the two sets. To accurately locate each nucleosome, we generated two probability landscapes over the genome: one for positive reads and one for negative reads. To build these probability landscapes we used kernel density estimation (KDE) (Parzen 1962) with a Gaussian kernel \((\sigma = 20)\) (Boyle et al. 2008; Shivaswamy et al. 2008).
The area between a positive peak followed by a negative peak approximately $f$ bases downstream is interpreted as a potential nucleosome. To make sure that each positive peak is mapped to the most appropriate negative peak downstream, we use the Needleman-Wunsch alignment algorithm (banded to within 50bp of $f$) (Needleman and Wunsch 1970). Working under the assumption that two ends of the same nucleosome should show a similar distribution of tags, the product of the Euclidean distance and the Pearson correlation between the strands’ KDEs in a 40bp window around the peak was used as the score of an alignment, with a gap penalty of -2 (the median alignment score). This results in a mapping of 5’ nucleosome edges to 3’ nucleosome edges and allows for small false peaks that are a consequence of our high sequence depth to be discarded.

The scores associated with each nucleosome are determined by summing the kernelized distance of each read to its associated edge. The kernelized distance of a read to a nucleosome edge is defined as $k(d, \sigma) = e^{\frac{-1}{2} \times (\frac{d}{\sigma})^2}$ where $d$ is the chromosomal distance of a read to the nucleosome’s edge and $\sigma$ is the bandwidth used for KDE over the chromosome. These two sums (5’ and 3’) were added together to yield the full nucleosome score.

To normalize each experiment we generated as many in silico reads (drawn from a uniform distribution) as the experiment had Illumina reads (Shivaswamy et al. 2008). These reads were subjected to the same analysis as above, and the maximum nucleosome score was taken. This procedure was repeated twenty times and the average maximum was taken to be the scaling factor for the experiment. Each nucleosome score for the in vivo experiment was divided by this scaling factor. To further normalize between experiments, the distribution of nucleosome scores was quantile normalized. Since some experiments yielded a different average fragment size than others (estimated between 149 bp and 163 bp for the experiments described here) we represented each nucleosome as a dyad and then extended it
73 base pairs up- and downstream, yielding 147bp width nucleosomes.

**Visualization of nucleosomes**

When graphing the nucleosomes, in cases where two nucleosomes were found to overlap, we took the sum of the nucleosome scores in the area of overlap to be the score for that area, ignored any regions with a score < 0.2, and capped scores at 1 as in (Shivaswamy et al. 2008). We also ignored areas in rDNA regions (as defined in SGD) and telomeric regions (within 10 kb of the beginning or end of a chromosome), as the highly repetitive sequence content of these regions often made nucleosome mapping impossible with single-end reads.

In graphing the nucleosome dyad densities relative to a set of genomic features (Fig. 4C, SFig. 6, SFig. 9, SFig. 10, and SFig. 11), the nucleosome dyad positions (the center of each nucleosome) relative to the genomic features are collected. We then use Gaussian kernel density estimation with a bandwidth of 10 over the collection of relative dyad locations, and normalize the resulting density by the number of dyads and the number of genomic features to generate the data graphed. Peaks in this density may be interpreted as the most likely position of a nucleosome dyad relative to the genomic feature. To generate the nucleosome dyad density over the plasmid relative to the ARS1 ACS (Figs 4D & SFig 10A), we first called nucleosome locations on the plasmid with a much more fine-grained bandwidth than we did for the genomic nucleosomes. The combination of this fine-grained bandwidth and the extremely high sequencing coverage of the plasmid (on average 1m reads covering a 3750 bp plasmid) allowed us to call many overlapping potential nucleosome conformations. We weighted each of these potential conformations by the number of reads that contributed to it and then ran a similar kernel density estimation (with a bandwidth of 30) over these weighted potential nucleosome conformations. What results is a density map of potential nucleosome dyads relative to the ACS indicat-
ing both position of nucleosome dyads and the proportion of reads that support each position (by peak height).

Nucleosome positioning at genomic features

To display nucleosome positioning around genomic features (TSSs, ORC-ACSs, etc), we first needed to align the features. We took TSS coordinates from the transcript tiling array data of (David et al. 2006). The TSS was taken as the 5′ end of a segment which was categorized as overlapping a verified gene. The TSSs were then aligned such that transcription begins at base 0 and proceeds to the right. The first base of the transcript is annotated as position 0, the second base of the transcript is annotated as position 1, etc. The same was done for our ORC-ACS, although the decision of which way to align the set was made more arbitrary by the fact that an origin is a bi-directional feature. We thus aligned the ORC-ACS such that the first base of the T-rich match to the ACS is annotated as position 0, the second base is annotated as position 1, and the last base (the 33rd base), is annotated as position 32. Thus, the T-rich ACS match (Figure 1C) begins at position 0 and ends at position 32 regardless of whether it is present on the Watson or Crick strand. The same procedure was used for the non-replicative ACS set. The positions of the nucleosomes relative to the genomic features underwent the same transformation.

Construction of the nr-ACS set

The non-replicative ACS (nr-ACS) set was constructed by finding high-scoring matches to the ACS motif that did not show any signs of replication activity. Briefly, we scanned the genome for ACS motif matches in a manner identical to that described above. We then took all of the motif matches with a log-odds score higher than 9.15 (a number empirically determined to yield a set of motif matches of comparable size to the ORC-ACS set) and screened out any that were in genes or promoters (i.e.
within 160 bp of a gene defined by SGD), within 3 kb of an hydroxyurea-resistant origin (Bell lab, unpublished data), within 3 kb of a member of the nimACS set from (Xu et al. 2006) or contained within a ChIP-seq peak of ORC enrichment. We then screened out rDNA and telomeric (within 10 kb of the beginning or end of a chromosome) loci. The set of motif matches that survived this screening process composes the nr-ACS set.

Sequence analysis around genomic features

Given that 4–6 bp ‘islands’ of adenine bases (A’s) constitute a strong nucleosome excluding signal (reviewed in Segal and Widom 2009), we scanned the sequence local to a given genomic feature with a 6 bp window and noted the position of any windows that were homogeneously composed of 5 or 6 identical bases. We scanned only the genomic DNA strand on which the feature was located, and allowed the strand of the feature to define which direction was “upstream” (negative relative coordinates) and which was ‘downstream’ (positive relative coordinates). For example, if a promoter was located on the negative strand at position $x$, then we would scan the negative genomic strand around $x$ for nucleosome enrichment, and the genomic coordinates ($x - 400 .. x + 400$) would map to the local graph coordinates ($+400 .. -400$).

In the heatmap graphs (Figs. 3B, 3C & SFig. 7), these windows were given a color corresponding to their base composition identity (those with $\geq 5$ T’s are colored green, while those with $\geq 5$ A’s are colored red) and a transparency based on their homogeneity (those with 5 of the same base are half transparent, while those with 6 of the same base have no transparency).

In the nucleotide enrichment density plots, all points identified in the heatmap graphs were combined, and used as input to a kernel density estimation algorithm,
with a bandwidth of 20 bp. The resultant densities were then normalized both for the number of points and the number of genomic features considered, to make their y-axes quantitatively comparable.

**Purification of ORC and ABF1**

ORC was Flag-tagged at the N-terminus of ORC1 and purified from yeast G1 extracts as previously described (Tsakraklides and Bell 2010).

Abf1 was purified from Rosetta 2(DE3)pLysS (Novagen) harboring pET28a-N-His, C-2xFlag-Abf1. Abf1 expressed in this strain has an N-terminal His-Tag and a C-terminal 2xFlag-tag. Abf1 expression was induced by the addition of 1 mM IPTG at OD 0.4 for 4 hours. Cells were resuspended in Buffer A (50 mM HEPES-KOH [pH 7.4], 10% glycerol) containing 0.3M KCl (0.3K-BufferA) and frozen in liquid nitrogen. After thawing, cells were sonicated and then centrifuged for 30 min at 22,500 rpm. Abf1 was purified from the lysis supernatant by sequential use of TALON superflow metal affinity resin (Clontech), Anti-Flag M2 affinity gel (Sigma) and SP sepharose. First, lysis supernatant was mixed with TALON superflow metal affinity resin (Clontech) equilibrated with 0.3K-Buffer A. After a 1 hour incubation at 4°C, resin was washed with 20 column volumes of 0.3K-Buffer A and eluted with 250 mM imidazole in 0.3K-Buffer A. The eluted fraction was mixed with Anti-Flag M2 affinity gel (Sigma) equilibrated with 0.3K-Buffer A and incubated for 16 hours at 4°C. Resin was washed with 10 column volumes of 0.3K-Buffer A and 10 column volumes of 0.1K-Buffer A. Abf1 was eluted from the resin with 0.1 mg/ml 3xFlag peptide in 0.1K-Buffer A. The eluted fraction was loaded onto a SP sepharose column equilibrated with 0.1K-BufferA and subsequently eluted with 0.5K-buffer A.

ISWI complex and recombinant *S. cerevisiae* histone were purified as previously described (Vary et al. 2004).
References


Supplementary Figure Legends

**SFigure 1.**  Venn diagram of the overlap between ORC-ACS and nim-ACS (Xu et al. 2006) sites. Of the 258 ORC peaks identified in this study that overlapped with the ORC peaks from (Xu et al. 2006) (see Supplemental Methods), 13 were not considered in this graph for the following reasons: 3 did not have an associated ACS with a log odds score greater than 4 (see Supplemental Methods) and 10 were mapped to different ACSs between replicates.

**SFigure 2.**  Nucleosome organization at transcription start sites. Each row in the heatmap represents a transcription start site. Each is oriented such that the TSS is at position 0 and transcription proceeds to the right. A stably positioned nucleosome is evident at the TSS and a well-phased array of nucleosomes extends into the gene body. Upstream, those genes with open promoters have an NDR while others have a -1 nucleosome immediately adjacent to the +1 nucleosome. The average nucleosome signal is summarized below the heatmap as a line graph, with interpreted nucleosome positions represented by ovals.

**SFigure 3.**  ORC-ACS and nr-ACS sites have nearly identical motifs. A comparison of the sequence motif derived from the (A) ORC-ACS and (B) nr-ACS sets.

**SFigure 4.**  Comparison of nucleosome positioning at ORC-ACS and nimACS (Xu et al. 2006) sites. (A) Heatmap shows nucleosome positioning around the set of ACS calls from the two studies located in ORC peaks that both studies identified. Each putative ACS is categorized as in the sections of the Venn diagram in SFig. 1: those ACS calls that are common to both studies (top), specific to the nimACS set (middle), or specific to the ORC-ACS set (bottom). (B) A summary plot of the heatmap in A. The scores of the nucleosomes belonging to the agreed upon ACS calls (black), the ACS calls present only in nimACS set (blue) and those present only in the ORC-ACS set (red) are averaged and plotted as a function of distance.
from the beginning of the T-rich strand-oriented ACS.

**SFigure 5.** Nucleosome positioning in the absence of trans-acting factors. (A) Heatmap of *in vitro* nucleosome occupancy at ORC-ACS sites. (Kaplan et al. 2009) (B) *In vitro* nucleosome occupancy at nr-ACS sites. Vertical dashed lines represent center of the 33 bp ACS motif match. All ACS matches are oriented relative to the T-rich strand and aligned at position 0.

**SFigure 6.** A-rich islands correlate with the position of the downstream nucleosome. A- and T-island nucleotide density (green and red, respectively) are shown by distance to the downstream nucleosome dyad in the ORC-ACS set. Origins were divided into three bins of roughly equal size, based on the distance from the identified ACS to the dyad of the first downstream nucleosome. A kernel density estimation was calculated by noting the position of all A- and T-islands (5 or 6 As or Ts in a 6bp window) relative to the first base of the ORC-ACS. The vertical line represents the center of the 33bp ACS motif.

**SFigure 7.** Transcription start sites with open promoters exhibit T/A nucleotide polarity. Open yeast promoters show a T-island to A-island sequence polarity switch in the NFR much like that seen at functional ACSs (Fig. 3A). The heatmap shows nucleosome positions and is organized as in (SFig. 2). Superimposed over the heatmaps are A- and T-rich regions (green and red respectively). These A- and T-islands are summarized in the histogram above the heatmap, as in (Fig. 3).

**SFigure 8.** Orc1-161 is defective for origin binding at the non-permissive temperature. Locus specific ORC ChIP at *ARS305* and *URA3* was performed at the permissive (23°C) and restrictive (37°C) temperature in both a wild-type and mutant *orc1-161* strain. ORC dissociates from origin DNA at the restrictive temperature in the *orc1-161* strain.

**SFigure 9.** Comparison of nucleosome dyad locations across the different mnase-
seq experiments. WT=wildtype and MT=orc1-161 mutant. Replicates (dotted lines) were combined (solid lines) due to their high concordance. Additionally, the wildtype asynchronous experiment at 23C (purple) is indistinguishable from the wildtype G2 (nocodazole arrested) experiments at 37C (green). (A) Nucleosome dyad density around the ORC-ACS set. The density represents the likelihood of finding a nucleosome dyad in a given location relative to the ACS. Vertical line indicates the center of the 33bp ACS motif. (B) Nucleosome dyad density relative to the transcription start sites from (David et al. 2006).

**Figure 10.** Abf1 supersedes ORC in positioning the downstream nucleosome at ARS1 *in vitro*. A comparison of nucleosome dyad densities on the plasmid (A) and *in vivo* at ARS1. (A) Reconstituted chromatin assembly in the presence of ORC or ORC and Abf1p. The addition of Abf1p recapitulates the positioning of the downstream nucleosome observed *in vivo* at ARS1. (B) *In vivo* nucleosome positions found at ARS1. The downstream nucleosome is positioned by Abf1p *in vivo*.

**Figure 11.** The origin NFR is the same in G1 and G2. A comparison of nucleosome dyad locations around (A) ORC-ACS sites and (B) TSSs in the combined G2 (nocodazole arrested) data set (green) and the G1 (alpha-factor arrested) experiment (blue). The center of the 33bp ACS motif (A) and the TSS (B) are indicated by the dashed vertical lines.
SFigure 1. Venn diagram of the overlap between ORC-ACS and nim-ACS sites.
Figure 2. Nucleosome organization at transcription start sites.
Figure 3. ORC-ACS and nr-ACS sites have nearly identical motifs.
**Figure 4.** Comparison of nucleosome positioning at ORC-ACS and nimACS sites.
**SFigure 5.** Nucleosome positioning in the absence of *trans*-acting factors.
Figure 6. A-rich islands correlate with the position of the downstream nucleosome.
SFigure 7. Transcription start sites with open promoters exhibit T/A nucleotide polarity.
SFigure 8. *orc1-161* ts is defective for origin binding at the non-permissive temperature.
**Figure 9.** Comparison of nucleosome dyad locations across the different mnase-seq experiments.
**SFigure 10.** Abf1p positions the downstream nucleosome at ARS1.
SFigure 11. The origin NFR is the same in G1 and G2.