Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading

Heather K. MacAlpine, Raluca Gordân, Sara K. Powell, et al.

Genome Res. 2010 20: 201-211 originally published online December 7, 2009
Access the most recent version at doi:10.1101/gr.097873.109

Letter

Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading

Heather K. MacAlpine,[1,3] Raluca Gordân,[2,3] Sara K. Powell,[1] Alexander J. Hartemink,[2] and David M. MacAlpine[1,4]

[1]*Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina 27710, USA;*
[2]*Department of Computer Science, Duke University, Durham, North Carolina 27708, USA*

The origin recognition complex (ORC) is an essential DNA replication initiation factor conserved in all eukaryotes. In *Saccharomyces cerevisiae*, ORC binds to specific DNA elements; however, in higher eukaryotes, ORC exhibits little sequence specificity in vitro or in vivo. We investigated the genome-wide distribution of ORC in *Drosophila* and found that ORC localizes to specific chromosomal locations in the absence of any discernible simple motif. Although no clear sequence motif emerged, we were able to use machine learning approaches to accurately discriminate between ORC-associated sequences and ORC-free sequences based solely on primary sequence. The complex sequence features that define ORC binding sites are highly correlated with nucleosome positioning signals and likely represent a preferred nucleosomal landscape for ORC association. Open chromatin appears to be the underlying feature that is deterministic for ORC binding. ORC-associated sequences are enriched for the histone variant, H3.3, often at transcription start sites, and depleted for bulk nucleosomes. The density of ORC binding along the chromosome is reflected in the time at which a sequence replicates, with early replicating sequences having a high density of ORC binding. Finally, we found a high concordance between sites of ORC binding and cohesin loading, suggesting that, in addition to DNA replication, ORC may be required for the loading of cohesin on DNA in *Drosophila*.

[Supplemental material is available online at http://www.genome.org. The sequence data from this study have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession nos. GSE17282, GSE17279, GSE17285, and GSE18942.]

To duplicate a eukaryotic genome completely during S phase, DNA replication must initiate at multiple start sites along each chromosome. These replication start sites, or origins of replication, are marked by the origin recognition complex (ORC) (for review, see Bell and Dutta 2002). During G_1 the replicative helicase, MCM2-7 complex, is assembled onto the DNA in an ORC-dependent manner to form the pre-replicative complex (pre-RC). The ordered assembly of the pre-RC in G_1 and its subsequent activation in S phase are tightly regulated to ensure that the genome is copied precisely once and only once. ORC is not only critical for genetic inheritance, but is also involved in epigenetic inheritance (Bell et al. 1993; Foss et al. 1993) and chromosome segregation (Gillespie and Hirano 2004; Takahashi et al. 2004, 2008; Shimada and Gasser 2007). Despite the conservation of ORC in all eukaryotes, very little is known regarding the mechanisms by which ORC binding sites are selected to ensure the inheritance of both genetic and epigenetic information (for review, see Cvetic and Walter 2005).

ORC localization to the chromosome is essential for the establishment of DNA replication origins. In yeast, ORC specifically binds to the ARS consensus sequence (ACS), a degenerate A/T-rich motif that is necessary, but not sufficient, for ORC binding and origin function (for review, see Bell and Dutta 2002). Despite the clear evolutionary conservation of all six ORC subunits in eukaryotes, the mechanism by which ORC is recruited to DNA in other organisms remains elusive. Typically, ORC interacts with many different sequences with little change in sequence-dependent specificity. Interestingly, the topological properties of the DNA influence ORC's affinity for DNA (Remus et al. 2004). Specifically, the induction of negative supercoils can increase the affinity of *Drosophila* ORC for DNA 30-fold without changing the sequence specificity.

Although no clear consensus sequence for ORC binding has emerged in higher eukaryotes, known initiator sequences have been identified in multiple species (for review, see Aladjem and Fanning 2004). These initiator sequences have the ability to recruit ORC and function as origins of replication even when placed at ectopic locations. One of the best-studied initiator sequences is the *Drosophila* origin, Ori-beta, found at the chorion (egg shell protein) locus on chromosome 3L (for review, see Tower 2004). During oogenesis Ori-beta undergoes multiple rounds of ORC-dependent replication initiation in a single cell cycle, resulting in the 60-fold amplification of the chorion locus in the developing follicle cell. The Ori-beta origin is only used at a very specific stage in development and does not appear to function as a canonical origin in normal mitotic cell cycles (MacAlpine et al. 2004). An attractive hypothesis is that the lack of sequence specificity for metazoan ORC is a critical feature that allows developmental plasticity of ORC binding and origin function. Unlike *Saccharomyces cerevisiae*, metazoan organisms have to coordinate the replication program with developmental programs, the most extreme example being the high density of replication origins required for the rapid cell divisions during early embryogenesis (Kriegstein and Hogness 1974; Hyrien et al. 1995).

ORC's role in the nucleus is not limited to DNA replication. ORC is critical for the establishment of silencing at the cryptic mating type loci HML and HMR in *S. cerevisiae* (Bell et al. 1993; Foss et al. 1993). In *Drosophila* a population of ORC is associated with heterochromatin and SU(VAR)205 (also known as HP1),

[3]**These authors contributed equally to this work.**
[4]**Corresponding author.**
E-mail david.macalpine@duke.edu; fax (919) 681-9567.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.097873.109.

suggesting a role for ORC in maintaining heterochromatic silencing (Pak et al. 1997; Huang et al. 1998). Recently, ORC has been implicated both directly and indirectly in sister chromatid cohesion. Cohesion between the sister chromatids is mediated by the cohesin complex, which, like the pre-RC, is loaded onto the DNA in G_1 before the initiation of S phase (Nasmyth 2002). Recent biochemical studies using *Xenopus* egg extracts have established that loading of the cohesin complex is dependent on the formation of the pre-RC (Gillespie and Hirano 2004; Takahashi et al. 2004). Specifically, the loading of cohesin on chromatin templates requires dbf4-dependent kinase (DDK) activity and ORC-dependent loading of the MCM2-7 complex (Takahashi et al. 2008). In contrast, the loading of cohesin in *S. cerevisiae* is independent of pre-RC formation, as cohesion is still established in a *cdc6* mutant that fails to assemble the pre-RC (Uhlmann and Nasmyth 1998). In addition, genome-wide studies in yeast have revealed that cohesin complex members are not loaded at sites of ORC binding (Glynn et al. 2004; Lengronne et al. 2004). However, multiple *S. cerevisiae* ORC mutants exhibit a G_2/M delay (Dillin and Rine 1998), and Orc2p appears to be important for maintaining sister-chromatid cohesion perhaps through a second mechanism that is independent of cohesin loading (Shimada and Gasser 2007).

We describe the genome-wide localization of approximately 5000 ORC binding sites in the *Drosophila* Kc167 cell line. ORC localizes to specific chromosomal locations, many of which function as early activating origins of replication, and preferentially localizes to dynamic chromatin near the transcription start sites of actively transcribed genes. The density of ORC binding across the genome appears to be a key determinant of the replication timing program. The large array of ORC binding sites allowed us to use machine learning algorithms to identify sequence features that are predictive of ORC binding sites and, more generally, open chromatin. Finally, integration of our ORC data with cohesin complex binding data suggests that ORC may be critical for the nucleation of cohesin onto the DNA in *Drosophila*.

Results

Genome-wide localization of ORC

Formation of the pre-RC at origins of replication is essential for progression into S phase. Despite conservation of pre-RC components in all eukaryotes, little is known about how origins are selected in metazoan organisms. Recent advances in genomic technology have allowed for the mapping of origins in mammalian cells (Lucas et al. 2007; Cadoret et al. 2008; Sequeira-Mendes et al. 2009), but these studies lack the precise mapping possible with chromatin immunoprecipitation (ChIP). We have used ChIP in combination with Agilent genome-wide tiling arrays (one 60-mer probe every 100 bp) to identify approximately 5000 ORC binding sites in the *Drosophila* genome from asynchronous Kc167 cells. ORC-associated chromatin was immunoprecipitated with a polyclonal antibody that specifically recognizes the *Drosophila* ORC subunit, ORC2 (Austin et al. 1999; Royzman et al. 1999; MacAlpine et al. 2004). These experiments significantly expand on our prior efforts using a PCR product tiling array to map ORC binding sites across chromosome 2L of *Drosophila* (MacAlpine et al. 2004).

Experiments in *Drosophila* and *Xenopus* suggest that ORC does not exhibit any sequence specificity in vitro (Vashee et al. 2003; Remus et al. 2004). Despite the lack of apparent sequence specificity, we find ORC binding sites distributed across the *Drosophila* genome at specific chromosomal locations (Fig. 1A, black). ORC binding sites were mapped genome-wide from two independent biological replicates. Enrichment along the chromosome was calculated from the raw microarray data using the MA2C software program, which normalizes the raw intensity data and corrects for probe-specific hybridization effects (Song et al. 2007). The normalized data are plotted as a function of chromosomal position. Peaks of significant ORC binding were subsequently identified using a sliding window approach, and 5135 ORC binding sites were identified in the *Drosophila* genome ($P < 1 \times 10^{-5}$). We estimated a false discovery rate of <0.24% based on the number of significant peaks in which the control channel signal was greater than the ORC-enriched signal.

It is important to validate the results of genome-wide localization experiments using independent approaches. To this end, we have confirmed our ORC localization data by using a TAP-tagged allele of *Orc2*. Specifically, we performed ChIP using IgG sepharose beads from ORC2-TAP-expressing and control cells. We found a similar distribution of ORC enrichment by the IgG pull-down of the ORC2-TAP-expressing cells as we did with the polyclonal antibody ($R = 0.43$) (Supplemental Fig. 1; Supplemental Table 1). In contrast, no significant enrichment was detected by the IgG pull-down in the absence of the TAP tag ($R = 0.042$).

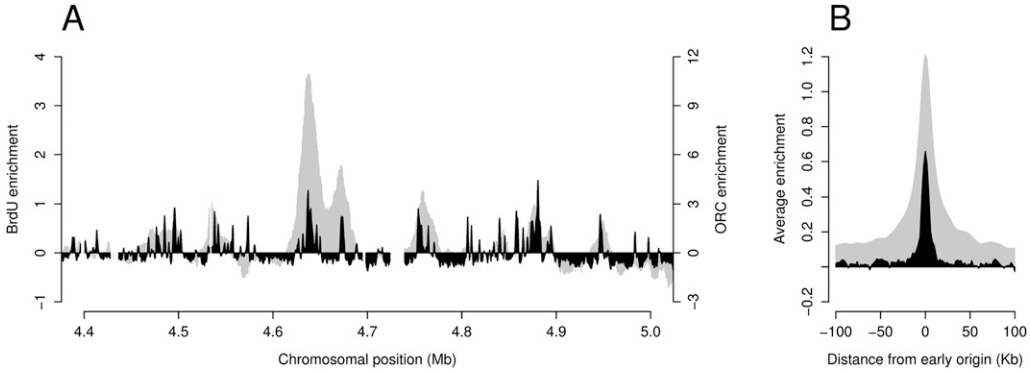

Figure 1. ORC localizes to specific chromosomal locations. (*A*) A subset of ORC binding sites map to early origins of replication. (Black) MA2C-normalized $\log_2$ ratios of ORC2 enrichment from chromatin immunoprecipitations are plotted as a function of chromosomal position for a 600-kb window of chromosome 3R. (Gray) Early origins were identified by immunoprecipitation of BrdU-containing replication intermediates during an HU arrest. (*B*) ORC is enriched at early origins of replication. The summits of 630 HU-resistant early origins were determined, and the average BrdU (gray) or ORC (black) enrichment is plotted relative to the distance from the early origin summit.

A second biological validation of our ORC localization was obtained by mapping early-activating origins of replication. Because ORC is an essential initiator complex, we expect to find ORC localized to origins. To identify start sites of DNA replication, we treated Kc167 cells with the drug hydroxyurea (HU) in the presence of the thymidine analog 5-bromo-2-deoxyuridine (BrdU). Treatment with HU inhibits ribonucleotide reductase, resulting in a depletion of nucleotide pools and activation of the intra-S-phase checkpoint (Santocanale and Diffley 1998; Shirahige et al. 1998). Thus, in the presence of HU, BrdU incorporation will be limited to sequences immediately adjacent to early-activating origins.

To identify sites of BrdU incorporation throughout the genome in the presence of HU, we used a mouse monoclonal antibody specific for BrdU to immunoprecipitate and enrich for replication intermediates as described previously (MacAlpine et al. 2004). The BrdU-enriched replication intermediates and control sequences were differentially labeled and hybridized to a lower-resolution (one 60-mer probe every 400 bp) Agilent genomic tiling microarray. In Figure 1A, we also show BrdU enrichment along the chromosome (gray). We find ORC localized to 89% of the peaks of BrdU incorporation, with the apex of BrdU incorporation most often centered directly over the ORC binding site. Not unexpectedly, we also find ORC at sequences that are not located at early activating origins. Only 30% (1570/5135) of the ORC binding sites are associated with early origins of replication. The remaining ORC binding sites may represent potential sites of late replication initiation or may be activated in a stochastic manner in a limited number of cells.

To demonstrate that ORC was specifically enriched at early-activating origins of replication, we identified the peaks of 630 early-activating origins throughout the *Drosophila* genome, at a false discovery rate of <0.54% ($P < 1 \times 10^{-3}$). We then plotted the average BrdU and ORC enrichment as a function of distance from the apex of BrdU enrichment (Fig. 1B). ORC specifically localizes to early activating origins, providing an important biological validation of our ORC mapping using genomic arrays.

In order to duplicate the genome completely during S phase, multiple start sites of DNA replication must be distributed across each chromosome. We sought to identify the distances between ORC binding sites, as these would represent the lengths of the minimum potential replicons. We found the median distance between ORC binding sites to be 11 kb (Fig. 2A). Strikingly, we found that many distinct ORC peaks were clustered over just a few thousand base pairs. These clusters of ORC binding sites may represent potential initiation zones and might increase the probability of a local initiation event.

The distribution of ORC binding sites throughout the genome is critical for establishing the replication timing program. As the distance between two ORC binding sites increases, the intervening sequence between them must be replicated by forks that initiated at distant origins. Therefore, we would expect to observe a later time of replication for sequences located between two distant ORC binding sites. To examine this, we first determined the relative time of replication for all unique sequences in the *Drosophila* genome using previously developed methods (MacAlpine et al. 2004). Briefly, early- and late-replicating intermediates were isolated and hybridized to genomic tiling arrays. The replication timing value for the mid-point of each inter-ORC segment was determined and plotted versus the inter-ORC distance (Fig. 2B). We found that as the inter-ORC distance increased, sequences between ORC binding sites were likely to replicate significantly later in S phase.

Is the density of ORC binding along the chromosome a determinant of replication timing? The previous results suggest that sequences distal from ORC binding sites replicate later in S phase. We wanted to determine if the density of ORC binding sites was correlated with the time at which sequences replicated in S phase. We binned the replication timing data into the earliest and latest replicating quartiles and found that the density of ORC binding sites was almost three times greater in the earliest quartile than in the latest quartile (Fig. 2C). This suggests that the replication timing program is, at least in part, established at the level of ORC binding along the chromosome.

ORC localizes to active promoters

A common theme to emerge from genome-wide studies of DNA replication in higher eukaryotes is that a correlation exists between

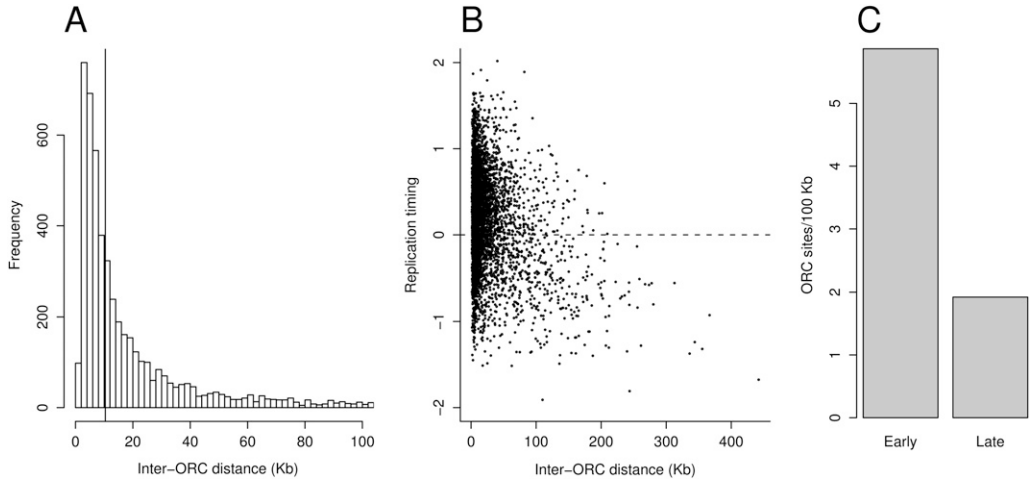

Figure 2. Chromosomal density of ORC is a determinant of replication timing. (*A*) A histogram of inter-ORC distance throughout the genome. The median distance between ORC binding sites is 11 kb. (*B*) Time of replication is significantly delayed for sequences between distant ORC binding sites. The relative time of replication for the sequences equidistant from two ORC binding sites was plotted against the distance between the two ORC binding sites. (*C*) Early-replicating regions of the chromosome have an increased density of ORC binding. ORC binding sites per 100 kb were determined for the earliest and latest quartile of replication timing domains.

the time at which a sequence replicates and its transcriptional activity (Schübeler et al. 2002; MacAlpine et al. 2004; Karnani et al. 2007; Hiratani et al. 2008). Sequences that are transcriptionally active tend to replicate early in S phase, whereas transcriptionally inactive sequences replicate later in S phase. This correlation is not at the level of individual genes but rather applies to the transcriptional activity of multiple genes integrated over broad chromosomal domains (MacAlpine et al. 2004). The establishment of origins of replication in the vicinity of actively transcribed genes is a potential mechanism to ensure that these regions are duplicated early in S phase. We investigated the genome-wide distribution of ORC relative to gene promoters in the *Drosophila* genome and found that almost two-thirds of the ORC binding sites overlapped with transcription start sites (Fig. 3A).

In *S. cerevisiae*, ORC is involved in silencing of the cryptic mating type loci (for review, see Rusche et al. 2003). Based in part on ORC's role in silencing, we wanted to examine the potential role of ORC in regulating the transcription program. We isolated mRNA from asynchronous Kc167 cells and examined the transcription program by hybridization to Affymetrix expression arrays. Transcripts were binned into five equal groups based on their relative expression levels (Supplemental Fig. 2). For each quintile of expression, we plotted the average enrichment for ORC as a function of distance to the transcription start site. We found a clear peak of ORC enrichment at active promoters in the top three quintiles of expression, while no significant enrichment was observed for ORC at the promoters of inactively transcribed genes (Fig. 3B). We conclude that ORC localizes to the promoters of actively transcribed genes.

Numerous biochemical interactions have been detected between ORC and various transcription factors (MYB, RBF, DM, E2F) (Bosco et al. 2001; Beall et al. 2002; Dominguez-Sola et al. 2007), suggesting that transcription factors may act as specificity factors in recruiting ORC to the DNA. If specific transcription factors are participating in ORC recruitment, then we might expect to see enrichment of specific functional classes of genes, presumably regulated by a common set of transcription factors. We used Gene Ontology to categorize the ORC-associated promoters (Supplemental Table 2) and found that ORC interacted with many diverse classes of genes with no significant patterns. These data suggest that the local chromatin environment surrounding actively transcribed genes, and not any specific transcription factor, is responsible for ORC recruitment.

ORC associates with open chromatin

A hallmark of actively transcribed genes is the presence of open chromatin in the vicinity of their promoters (for review, see Rando and Chang 2009). We explored the hypothesis that ORC may preferentially localize to regions of open chromatin by using *Drosophila* nucleosome data generated from *Drosophila* S2 cells (Mito et al. 2007; Henikoff et al. 2009). In *Drosophila*, active promoters and regulatory elements are marked by the replacement histone variant, H3.3 (Mito et al. 2007), which is assembled into nucleosomes and deposited on the DNA in a replication-independent manner (Ahmad and Henikoff 2002). To justify the comparison of nucleosomes in S2 cells and our ORC binding sites in Kc167 cells, we note a high concordance between expression profiles for these two cell lines (Celniker et al. 2009; S Celniker, pers. comm.), and we find similar replication timing patterns and distribution of early activating origins of replication (Celniker et al. 2009), suggesting that these cell lines will, at the gross level, have similar chromatin landscapes.

To investigate the chromatin state near ORC binding sites, we plotted the aggregate enrichment of H3.3 as a function of distance from the ORC binding site for all previously identified ORC binding sites (Fig. 4A). Not unexpectedly, we found a clear peak of H3.3 enrichment at ORC binding sites that overlapped with the promoters of active genes (solid line). However, we also found a similar distribution of H3.3 enrichment at ORC sites that were not associated with an active promoter (dashed line). Therefore, open and dynamic chromatin as marked by H3.3 appears to be a determinant of ORC localization, and not simply a consequence of ORC colocalizing to promoter elements.

In *Drosophila*, the single histone variant, H2Av, functions as both histone H2A.Z and H2A.X (van Daal and Elgin 1992). Like H2A.Z in *S. cerevisiae* (Dion et al. 2007), H2Av is enriched at the promoters of active transcripts, specifically at the +1 nucleosome (Henikoff et al. 2009). We find a broad peak of enrichment for H2Av at those ORC sites that overlap with promoter elements (Supplemental Fig. 3A). The broad H2Av peak is a result of it occupying the +1 nucleosome and the fact that we did not differentiate the direction of ORC-associated promoters. However, unlike H3.3, we observe that surrounding H2Av levels are lower around non-promoter-associated ORC binding sites, which is likely due to decreased gene density. The lack of strong local H2Av enrichment at ORC sites distal from promoters refutes a counter

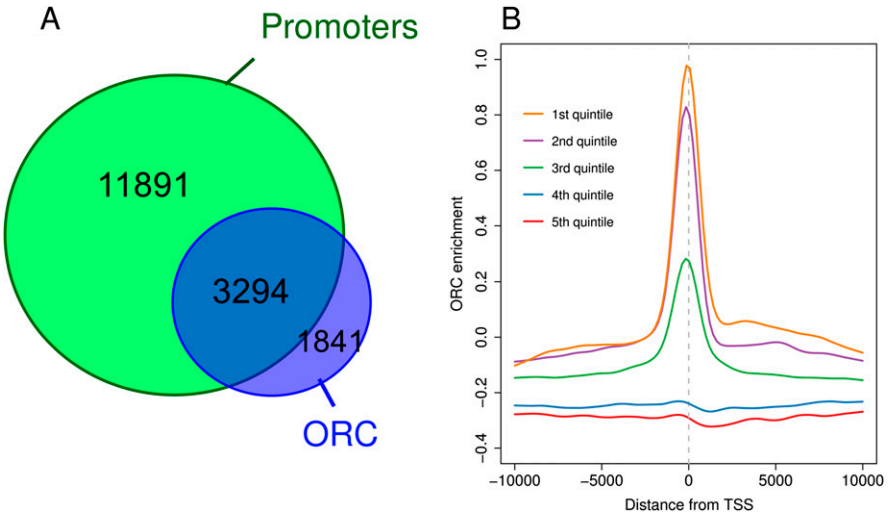

Figure 3. ORC associates with promoters of active genes. (*A*) Venn diagram depicting the overlap between ORC-associated sequences and annotated transcription start sites. (*B*) ORC localizes to the promoter regions of actively transcribed genes. The mRNA expression levels for approximately 15,000 *Drosophila* transcripts were determined by Affymetrix expression analysis. The mRNA expression levels for each gene were binned into five equal quintiles (Supplemental Fig. 2). The average ORC enrichment for each quintile of expression is plotted relative to the transcription start site.

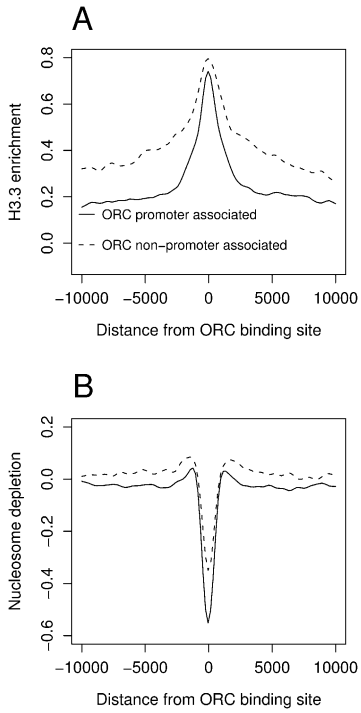

Figure 4. ORC localizes to open dynamic chromatin. (*A*) ORC binding sites are enriched for the histone variant H3.3. The average enrichment of H3.3 was plotted relative to the position of all ORC binding sites. The ORC binding sites were classified as either proximal or distal to annotated transcription start sites. Proximal ORC binding sites directly overlap with annotated promoters. (*B*) ORC binding sites are depleted for nucleosomes. The average depletion of nucleosomes was plotted relative to the position of promoter-proximal or -distal ORC binding sites.

hypothesis that a subset of our promoter distal ORC binding sites might exist at poorly annotated genes.

Prior chromatin mapping studies have demonstrated that sequences marked by H3.3 are also depleted for bulk nucleosomes (Mito et al. 2007). Not surprisingly, we find that ORC-associated sequences at sites both proximal (overlapping) and distal to promoters are indeed depleted for bulk nucleosomes as mapped by Mito et al. (2007) (Fig. 4B). To further explore the chromatin landscape in the vicinity of ORC binding sites, we turned to recent studies from the modENCODE consortium that profiled the dynamic state of chromatin using salt extraction of S2 cell nuclei (Supplemental Fig. 3B; Henikoff et al. 2009). We find that ORC-bound sequences are enriched in soluble nucleosome fractions following 80 mM and 150 mM salt washes. These fractions typically represent open and active chromatin and are enriched for regulatory elements and promoters. Conversely, no enrichment of ORC-associated sequences was observed after extraction with 600 mM salt, which solubilizes the bulk of chromatin (Henikoff et al. 2009). The remaining insoluble chromatin pellet was also enriched for ORC associated sequences and has been shown to be over-enriched for oligomeric nucleosomes and large protein complexes.

Cohesin complex members colocalize with ORC

Cohesion between the sister chromatids is mediated by the cohesin complex, which, like the pre-RC, is loaded onto the DNA in G_1 before the initiation of DNA replication (for review, see Nasmyth 2002). Recent biochemical studies using *Xenopus* egg extracts have established that, at least in *Xenopus*, the loading of the cohesin complex is dependent on the formation of the pre-RC (Gillespie and Hirano 2004; Takahashi et al. 2004, 2008). If, in *Drosophila*, the establishment of cohesion is also dependent on pre-RC assembly, then one might expect to find cohesin subunits at ORC binding sites throughout the genome.

To address this question, we compared our genome-wide mapping of ORC binding sites to a previous genomic study where multiple cohesin complex members were mapped in different *Drosophila* cell lines (Misulovin et al. 2008). We found a very high concordance between ORC binding sites and sites of cohesin binding in Kc167 cells. In Figure 5A, we depict our ORC2 localization data with SMC1, a cohesin subunit, in a browser window from the modENCODE project (Celniker et al. 2009). Both the ORC2 and SMC1 tracks are depicted as density plots, with intense blue indicating a binding site. Furthermore, it appeared that ORC was centered in the SMC1 peak and that the cohesin complex was spreading out from the ORC binding sites (Fig. 5B). Together these data suggest that ORC (and the pre-RC) may function to nucleate cohesin binding on the DNA.

The colocalization of ORC and cohesin is consistent with the *Xenopus* data and suggests that in *Drosophila*, the establishment of cohesin may be dependent on pre-RC assembly. We used a chromatin association assay to directly assess, in vivo, the dependence of cohesin loading on pre-RC assembly. Briefly, cellular extracts from Kc167 cells were fractionated into cytoplasmic, soluble nuclear, and insoluble (pellet) chromatin fractions, which were subsequently immunoblotted for subunits of both pre-RC and cohesin complexes. Because the assembly of the pre-RC is an ordered process that is tightly coupled with the cell cycle, we developed methods to arrest Kc167 cells at various points in the cell cycle to assess pre-RC formation and cohesin assembly.

We evaluated cohesin and pre-RC assembly in asynchronous cells, cells arrested in G_2 by treatment with 3% DMSO, cells arrested at the G_1/S transition with HU, and finally, cells arrested in G_1 by RNAi depletion of DUP (also known as CDT1), a critical pre-RC intermediate required for MCM2-7 loading (Supplemental Fig. 4; Maiorano et al. 2000; Wohlschlegel et al. 2000). However, prolonged exposures to DUP dsRNA (>24 h) resulted in a population of cells with less than 2N DNA, presumably mediated by a reductional anaphase.

We observed chromatin bound ORC in all of the samples, consistent with ORC remaining associated with the chromatin throughout the cell cycle (Fig. 5C). The MCM2-7 complex was observed on the chromatin only in asynchronous cells and cells arrested in early S phase. In the absence of DUP, the association of MCM2-7 with chromatin was markedly reduced, consistent with the critical role of DUP in assembling the pre-RC. The MCM2-7 levels on chromatin were also substantially reduced in G_2 cells because S phase had been completed. The total protein levels of pre-RC components and cohesin complex members were constant throughout the cell cycle in a whole cell extract (Supplemental Fig. 4F).

We found that the bulk loading of cohesin complex members onto DNA was not dependent on formation of the pre-RC. We observed constant levels of multiple cohesin complex members (SMC1, SA, Nipped-B) associated with chromatin at each stage of the cell cycle (Fig. 5C). Critically, no reduction in cohesin bulk loading was observed in DUP-depleted cells despite their clear inability to load the MCM2-7 complex. Taken together, these data suggest that pre-RC formation is not essential for cohesin loading in *Drosophila* cells. It is important to note, however, that these

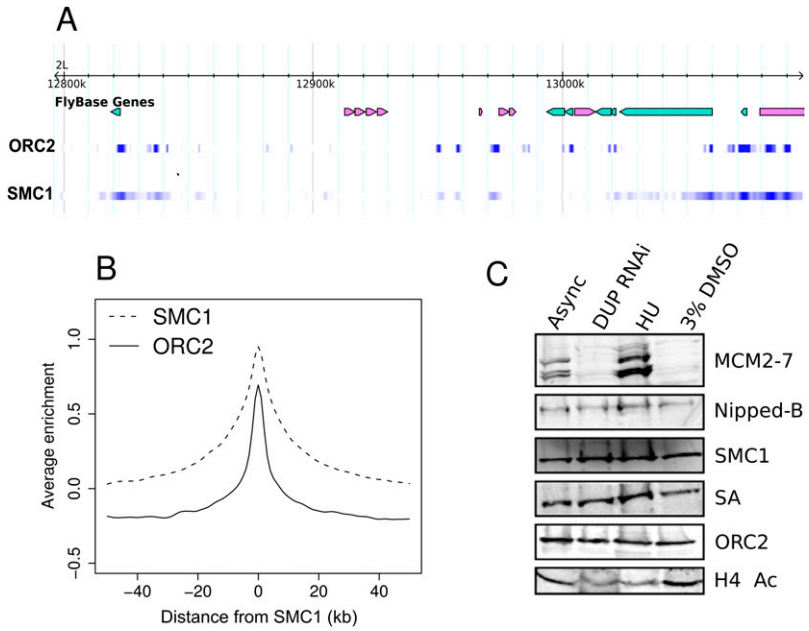

Figure 5. Cohesin loading occurs at ORC binding sites independent of pre-RC assembly. (*A*) ORC2 and SMC1 binding sites colocalize. ORC2 and SMC1 binding are indicated as blue density tracks for a 300-kb region of chromosome 2L. (*B*) ORC appears to nucleate cohesin binding. The average enrichments for ORC2 and SMC1 are plotted relative to the peak of all SMC1 binding sites in the *Drosophila* genome. (*C*) Cohesin loading is independent of pre-RC formation. Chromatin-associated MCM2-7, ORC, acetylated H4, and cohesin subunits at various points in the cell cycle (Supplemental Fig. 4).

chromatin association experiments are only examining the bulk level of DNA binding proteins on chromatin. Although bulk cohesin levels are not affected by DUP depletion, we have not examined where in the genome these cohesin complexes are being loaded. It remains possible that in the absence of pre-RC assembly, cohesin complex members may be loaded by alternative mechanisms and perhaps at different locations.

Sequence determinants of ORC localization

The large array of ORC binding sites gave us an opportunity to search for sequence elements that might participate either directly or indirectly in recruiting ORC to the chromosome. Sequences that directly participate in ORC binding would be reminiscent of the ACS in *S. cerevisiae*—that is, sequences that are necessary but not sufficient for ORC binding. Alternatively, certain transcription factor motifs may indirectly recruit ORC via intermediate protein–protein interactions. Using the motif identification tool PRIORITY (Narlikar et al. 2007), we found several motifs that are over-represented in the DNA sequences bound by ORC. When using different parameter settings, however, different motifs are reported, many of them repetitive. Although all these motifs may be important for ORC binding and/or origin function, possibly through indirect effects such as preventing nucleosome occupancy or facilitating DNA unwinding, we could not conclusively identify any single motif that had clear predictive value in identifying ORC binding sites.

Given the inconclusive results obtained using the motif finding approach, we turned to a more basic question: Are there any signals in the DNA sequence itself that guide ORC toward its binding locations? To answer this question, we used a machine learning approach. We selected 5050 euchromatic sequences that were associated with ORC. All these sequences, henceforth referred to as "positives," are 1 kb long and centered on peaks of ORC enrichment. We selected an equal number of 1-kb-long ORC-free sequences, henceforth referred to as "negatives," which do not overlap any of the ORC peaks. About two-thirds of the ORC-positive sequences overlapped with transcription start sites (TSSs); thus, it was important to select a negative set that had the same proportion of sequences overlapping transcription start sites as the positive set. Balancing the ratio of TSS-proximal versus TSS-distal sequences in both the positive and negative sets ensured that we were not inadvertently training on promoter signatures.

We used a support vector machine (SVM) to classify sequences as ORC-associated or ORC-free. Briefly, an SVM is a classification algorithm. It receives as input a set of labeled training examples. Each example corresponds to a DNA sequence, and it is represented as a vector of features (i.e., a vector of *k*-mer frequencies). Each example is labeled as "positive" (bound by ORC) or "negative" (not bound by ORC). Given the input examples, the SVM algorithm generates a model that can be applied to classify any new example (in our case as "positive" or "negative"). The SVM was trained using 80% of the data; the remaining 20% was used to validate and assess the generalization capabilities of different SVMs. All of the SVM classifiers we employed used the frequencies of sequence elements (*k*-mers; for $k = 1$ to K) as features, ignoring orientation (e.g., ACC and GGT were considered the same *k*-mer). The performance of the SVM classifiers, measured using the area under the receiver-operating characteristic curve (auROC), is shown in Figure 6, A and B. The ROC curve is a plot of the sensitivity or true-positive rate as a function of 1-specificity or false-positive rate. Surprisingly, when using only AT-content information (i.e., when $K = 1$), the classifier cannot distinguish well between the positive (ORC-associated) and negative (ORC-free) sequences: the auROC is just 0.68, only marginally better than the auROC of a random classifier (0.5). As we increased K and thus added more sequence elements, we were able to better discriminate between ORC-associated and ORC-free sequences, with the best performance achieved for $K = 6$: a 10-fold cross-validation accuracy of 81% and an auROC of 0.88.

Not all 2772 sequence features used by this SVM classifier were important for discriminating between the ORC-associated and ORC-free sequences: if we simply choose the features that are most correlated with the class, we can reduce the size of the feature set to 200 or 500 without a loss in accuracy. We sorted the features in decreasing order of the absolute value of their Pearson correlation with the class (i.e., +1 for positives and −1 for negatives). We tested the performance of the classifiers when using the top 20, 50, 100, 200, or 500 most discriminating sequence features. As shown in Figure 6B, the best SVM classifier was obtained when using just the top 500 features, with an auROC of 0.92. Using this classifier, at a false-positive rate of only 15%, we can accurately classify 83% of the ORC-associated sequences in our independent test set.

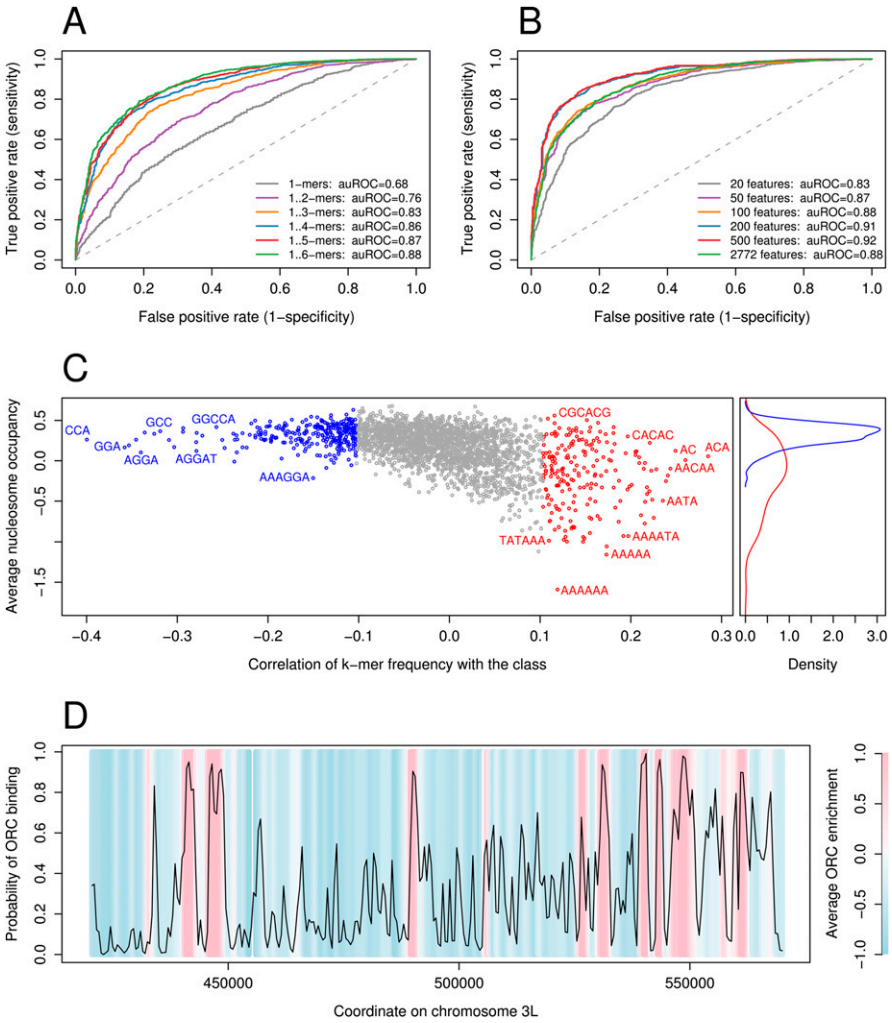

Figure 6. Sequence elements can discriminate between ORC-associated and ORC-free sequences. (*A*) ROC curves for SVM classifiers trained using as features the frequencies of *k*-mers for $k = 1$ to *K*, for increasing values of *K*. (*B*) ROC curves for SVM classifiers using subsets of features of increasing size. For each size of the feature set, we selected features with the highest absolute Pearson correlation with the class value (+1 or −1). All ROC curves were computed on the test set. (*C*) Correlation between *k*-mer features and class value, versus the average nucleosome occupancy over *k*-mers. The blue and red dots in the *left* plot correspond to the 500 features used in our analysis. The remaining features are shown in gray. The *right* plot shows the distributions of average nucleosome occupancy for the features with very high (red curve) and very low (blue curve) correlation coefficients. The average nucleosome occupancy values were computed from high-resolution in vitro nucleosome positioning data (Kaplan et al. 2009). (*D*) Predicted ORC binding for an arbitrary region on chromosome 3L. The background shows the average ORC2 ChIP enrichment; (blue) low enrichment; (pink) high enrichment. (Black curve) The posterior probability of ORC binding is in good agreement with the average ORC2 enrichment.

The high classification accuracy and predictive power of the SVM classifier indicate that there are signals in the DNA sequence itself that guide ORC toward its binding locations. However, it is not yet clear whether the sequences have a direct impact on ORC binding (i.e., there is a specific DNA binding motif for ORC), or an indirect effect (e.g., by preventing nucleosome formation, or by facilitating DNA unwinding). To begin to answer this question, we analyzed the 500 features used by the best SVM classifier. Interestingly, as shown in Figure 6C, the features that have the highest positive correlation with the class (i.e., are specific to ORC-associated regions) tend to have a low nucleosome occupancy, while features that are negatively correlated with the class (i.e., are specific to ORC-free regions) are in general occupied by nucleosomes. This suggests that many of the features that best distinguish between ORC-associated and ORC-free regions may have an indirect effect on ORC binding by preventing nucleosome formation. Features with a high positive correlation coefficient include ACA, AC, AACAA, AATA, AAAATA, CACAC, AAAAA, CGCACG, AAAAAA, TATAAA, etc. Some of these sequence elements (e.g., AAAAA, TATAAA, and other dA–dT sequences) have been shown to prevent nucleosome formation (Iyer and Struhl 1995; Suter et al. 2000; Ozsolak et al. 2007; Kaplan et al. 2009).

To further test the predictive power of the SVM classifier, we held out all the sequences from chromosome 3L, trained the classifier on the other chromosomes using the set of 500 features, and then used it to predict the probability of ORC binding to each 1-kb region on chromosome 3L. Our predictions are in good agreement with the experimental ORC data, as illustrated in Figure 6D, for a representative part of chromosome 3L (nucleotides 420,000 to 570,000). In summary, we can use primary sequence information to systematically identify sequences with a potential to bind ORC; however, not all of these potential sequences will interact with ORC. For example, active transcription through a sequence may physically prevent ORC association.

Discussion

In *Drosophila*, ORC localizes to specific regions of the chromosome, many of which function as origins of replication. The underlying chromosomal feature that appears to define ORC's association with the chromosome is open and active chromatin. As a class, ORC binding sites are enriched for the replication-independent histone variant H3.3, depleted for bulk nucleosomes, and often found immediately upstream of the transcription start sites of active genes. The median distance between ORC binding sites is 11 kb. Changes in the local density of ORC binding dictate whether a sequence will be replicated early or late in S phase, with regions of sparse ORC density replicating later in S phase. The integration of our ORC data with the genome-wide mapping of cohesin complex subunits (Misulovin et al. 2008) revealed extensive colocalization between ORC and multiple cohesin subunits, suggesting that, in *Drosophila*, ORC may participate in the loading of cohesin on the DNA. Finally, we used machine learning approaches to identify sequence elements that can accurately discriminate between ORC-associated and ORC-free sequences.

ORC binds to specific chromosomal locations in all eukaryotes; however, sequence specificity for ORC has only been

observed in the budding yeast, *S. cerevisiae*. We have identified approximately 5000 ORC-associated sequences distributed throughout the *Drosophila* genome. Despite being able to accurately resolve the peaks of ORC enrichment to within 1 kb, we were not able to identify a simple sequence motif like that found for many transcription factors. Instead, we found that a complex code of short sequence *k*-mers (≤6 bases) was sufficient to discriminate between ORC-associated and ORC-free sequences (Fig. 6A,B). These short oligomeric sequences likely contribute to a local environment that is favorable for ORC binding. We found that many of the short sequences that contributed to ORC binding were exclusionary to nucleosomes (Fig. 6C), suggesting that local nucleosome density may contribute to ORC association. Although open chromatin appears to be important for ORC localization, it is clearly not sufficient for ORC localization as not all predicted sites of low nucleosome occupancy are associated with ORC (Supplemental Fig. 5). Undoubtedly, ORC itself likely contributes some specificity as it has recently been shown that *Drosophila* ORC6 has a preference for poly(dA) sequences (Balasov et al. 2007), which are indeed enriched in our top 500 SVM sequence features.

Biochemical studies from multiple metazoan systems have clearly demonstrated that ORC exhibits little sequence specificity in vitro despite a high affinity for DNA (Vashee et al. 2003; Remus et al. 2004). This suggests that a primary determinant of ORC binding will be accessibility to the DNA. A critical difference between the interaction of ORC with DNA in vitro and in the nucleus is that the majority of DNA in the nucleus is packaged into nucleosomes and higher-order chromatin structure. The accessibility of chromatin is an inherited feature established, in part, by primary DNA sequence, transcription factors, insulator elements, transcription machinery, and histone variants. Indeed, we found that ORC preferentially localizes to open chromatin marked by transcription start sites, the histone variant H3.3, and depletion of bulk nucleosomes (Fig. 4).

The lack of inherent sequence specificity for ORC may be a unique feature of higher eukaryotes, a feature required to ensure genetic inheritance throughout multiple developmental stages and tissue types. For example, during embryogenesis an increased density of origins is required to ensure the complete duplication of the genome within the confines of a very short cell cycle (Kriegstein and Hogness 1974; Hyrien et al. 1995). By using open chromatin to define potential origins, the cell can harness the plasticity of the transcription program and rapidly reprogram origin selection. Thus, changes in the transcription program will likely affect ORC localization and origin usage. Indeed, numerous examples exist in the literature of origin activity being modulated by transcription (Danis et al. 2004; Saha et al. 2004; Norio et al. 2005).

Open chromatin appears to be necessary for ORC binding, but it is clearly not sufficient as many sites of open chromatin are not associated with ORC. Additional proteins may participate in targeting ORC to specific chromosomal locations. Several studies have shown direct interactions between ORC and various transcription factors, including MYB, RBF, and E2F (Bosco et al. 2001; Beall et al. 2002; Dominguez-Sola et al. 2007). Although we cannot rule out the possibility that these and other *trans*-acting factors are cooperating with open chromatin to localize ORC to the DNA at specific locations, we do not think this is a general mechanism. For example, we did not identify a common motif near ORC binding sites. If ORC were being recruited to the DNA by a limited number of transcription factors, we would have found evidence for conserved transcription factor motifs at the ORC-associated promoters. We also explored the gene ontologies associated with the ORC-associated promoters. No significant functional categories emerged, suggesting that there was not a common regulatory element among the ORC-associated promoters (Supplemental Table 2). A consequence of locating the majority of origins at promoter regions is that the replication and transcription forks will be traveling in the same direction. This will eliminate potential head-to-head collisions between DNA and RNA polymerases. This coordination of transcription and replication to avoid direct polymerase collisions is readily apparent in the organization of prokaryotic genomes (Kunst et al. 1997).

The local topology of DNA is a likely factor contributing to the localization of ORC at regions of open chromatin. Negative supercoiling increases ORC's affinity for DNA in vitro by at least an order of magnitude (Remus et al. 2004). Negative supercoils are likely to be found behind the promoters of actively transcribed genes where ORC is often localized. Recent genome-wide RNA profiling experiments using sequencing-based platforms have identified the presence of short abortive divergent transcripts at many promoters in mammalian genomes (Core et al. 2008; Seila et al. 2008). This divergent transcription would create a very high local concentration of negative supercoils in the vicinity of the promoter. It will be interesting to see if the subset of promoters that are divergently transcribed in mammalian genomes are enriched for ORC and origin activity. Finally, it is tempting to note the parallels between bidirectional DNA replication and bi-directional transcription.

The DNA replication program is defined, in part, by where potential origins are established and the time at which they are activated in S phase. Studies in *S. cerevisiae*, *Drosophila*, and human cells have clearly shown that the local chromatin environment influences origin activation (Vogelauer et al. 2002; Goren et al. 2008; Knott et al. 2009; Schwaiger et al. 2009). For example, tethering the histone deacetylase, RPD3, to Ori-beta at the chorion locus inhibited origin function (Aggarwal and Calvi 2004). Here, we show that the density of ORC on the chromosome also contributes to the overall temporal pattern of DNA replication (Fig. 2). Not unexpectedly, sequences distant from ORC binding sites replicate later in S phase and are limited to gene-poor regions of the genome. The localization of ORC to promoters of actively transcribed genes results in a higher density of ORC in gene-rich regions, which guarantees their duplication early in S phase. Mammalian replication timing studies have shown that the rate of mutagenesis is significantly higher in late S phase (Stamatoyannopoulos et al. 2009). By duplicating gene-rich regions in early S phase, the cell can specifically maintain the integrity of these regions of the genome.

Our genome-wide studies of ORC binding in *Drosophila* revealed extensive colocalization between ORC and multiple subunits of the cohesin complex. ORC was most often found at the center of cohesin peaks, suggesting that ORC may participate directly or indirectly in sister-chromatid cohesion. Prior genome-wide mapping experiments in mammalian cells implicated the insulator element, CTCF, in sister-chromatid cohesion (Parelho et al. 2008; Wendt et al. 2008). In contrast, in *Drosophila*, <15% of the CTCF sites are coincident with the cohesin complex member, SA; however, there is significant overlap between the CP190 insulator element and cohesin binding sites (Bartkuhn et al. 2009). In egg extracts from *Xenopus*, ORC and pre-RC assembly are required for cohesin loading on sperm chromatin. Despite inhibiting pre-RC assembly by depleting DUP by RNAi, we were unable to perturb the levels of chromatin-associated Nipped-B, SMC1, and SA (Fig. 5C). These results suggest that in *Drosophila* the loading of

cohesin is not dependent on pre-RC formation. We were unable to directly test the role of *Drosophila* ORC in sister-chromatid cohesion because of compounding cell cycle phenotypes induced by RNAi depletion of ORC subunits. Alternatively, in the absence of pre-RC formation, cohesion may still be established, possibly at different locations, via an alternative pathway. Future experiments will address the localization of cohesin subunits in the absence of pre-RC assembly.

Methods

Cell growth

Kc167 cells were cultured in 150-mm plates at a density of 1×10^6 cells/mL in Schneider's Insect Cell Medium (Invitrogen) supplemented with 10% FBS and 1% penicillin/streptomycin/glutamine (Invitrogen). All cell growth in this study was conducted at 25°C. For G_1 arrests, DUP was depleted using RNAi. For G_1/S arrests, HU (Sigma) was added to 1 mM final concentration. For G_2 arrests, DMSO (Sigma) was added to 3% final concentration. Cell cycle position was determined by flow cytometry.

Chromatin immunoprecipitation

For each experiment, ~2–3 $\times$ 10^8 asynchronous Kc167 cells were collected. The cells were cross-linked at 25°C with 1% formaldehyde (Ted Pella) for 10 min and quenched with 125 mM glycine for 5 min. The cross-linked cells were centrifuged at 1000 rpm and washed twice in cold 1× PBS, and the pellet was frozen in liquid nitrogen and stored at −80°C. The pellet was thawed on ice, resuspended in lysis buffer 1 (50 mM HEPES-KOH at pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100, Complete MINI protease inhibitor cocktail tablet, Roche), and incubated with rotation for 10 min at 4°C. The extract was centrifuged at 1350*g* for 5 min at 4°C. The pellet was resuspended in lysis buffer 2 (10 mM Tris-HCl at pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, Complete MINI protease inhibitor cocktail tablet), incubated with rotation for 10 min at 25°C, and centrifuged as noted above. The pellet was resuspended in lysis buffer 3 (10 mM Tris-HCl at pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, 0.5% N-lauroylsarcosine, Complete MINI protease inhibitor cocktail tablet). The extract was sheared by sonication to an average size of 1 kb. Triton X-100 was added to 1%, vortexed gently, and centrifuged at 20,000*g* for 10 min at 4°C. The chromatin was immunoprecipitated with anti-dORC2 antibody, as described in Austin et al. (1999).

Labeling of DNA with fluorescent nucleotides

The immunoprecipitated DNA was labeled with either fluorescent Cy5- or Cy3-conjugated dUTP (Perkin Elmer), using Sequenase (US Biochemicals) and a random nonamer oligo (IDT). The immunoprecipitated DNA was dried in a speed-vac and resuspended in 10 μL of Sequenase primer mix (1× Sequenase buffer, 5 μg of random nonamer, 5 mM dUTP-Cy*). The samples were heat-denatured and cooled to 4°C in a thermocycler before adding 5 μL of Sequenase reaction mix (1× Sequenase buffer, 1.5 mM dATP, 1.5 mM dCTP, 1.5 mM dGTP, 0.75 mM dTTP, 500 ng of BSA, 3.5 mM DTT, and 13 U of Sequenase). The reaction temperature was slowly ramped to 37°C and incubated for 30 min. Following incubation, the sample was heat-denatured and cooled to 4°C, and fresh Sequenase (4 U) was added for the second and final round of labeling. Following the labeling, unincorporated nucleotides, oligo, and dye were removed using Microcon filters (Millipore).

Array hybridization and washing

All experiments were performed using biological replicates. For the ORC ChIP experiments, the labeled DNA was hybridized across four genomic tiling arrays, which contained probes spanning the entire *Drosophila* genome. The slides were hybridized and washed as per Agilent recommendations.

Data analysis

The resulting raw data for each slide were loess-normalized using the LIMMA package in R. Quantile normalization was then used to normalize probe values across the four-array set. Finally, the normalized values across the four arrays were output into a pseudo-NimbleGen format for downstream analysis and peak calling by the python MA2C package (Song et al. 2007). Peaks were identified at a cutoff of $P > 1 \times 10^{-5}$ using a windowing function of 300 bp. Analysis of each biological replicate independently identified 3666 and 4533 peaks in replicate 1 and replicate 2, respectively. Eighty-four percent of the peaks in replicate 1 were also found in replicate 2. The Pearson correlation between replicate 1 and replicate 2 was $R = 0.71$. Additional analysis details are provided in the metadata associated with the data submission files (see below).

Data submission

All genomic data (with accompanying metadata) are publicly available at the NCBI GEO data repository. The accession numbers are: GSE17282 for Kc167 ORC2 localization, GSE17279 for Kc167 replication timing, GSE17285 for Kc167 early origins of replication, GSE18942 for Kc167 ORC2-TAP, IgG control data, and expression data.

TAP tag chromatin immunoprecipitation

Primers 5′-GGTGGTGGATCCATGAGTGCCAGCAACAAAGG and 3′-GGTGGTACTAGTTTTCTTCTCCTGCTCCTCGAG were used to amplify the *Orc2* gene from the FastBac-DmOrc2 plasmid. The PCR product was sequentially digested with BamHI and SpeI (New England Biolabs) and ligated into a BamHI/SpeI-digested pMK33CTAP plasmid (gift of the Artavanis-Tsakonas laboratory, Harvard Medical School). The recombinant ORC2-TAP plasmid was sequenced and found to be in-frame with the TAP tag and have no errors. The ORC2-TAP plasmid was transfected into asychronous Kc167 cells in 100-mm tissue culture dishes using the Effectene Transfection Reagent (QIAGEN). 7.0 $\times$ 10^6 cells were transfected with 2.0 μg of ORC2-TAP plasmid and 2.5× vol Effectene reagent. Day 2 post-transfection, 125 μM hygromycin was added to cells and allowed to grow for ~2 wk. Transfected cells were induced with 500 μM $CuSO_4$ and allowed to grow for 3 d before harvesting. Chromatin isolation and cross-linking procedures were conducted as outlined above. Chromatin was immunoprecipitated using IgG Sepharose beads (GE Healthcare Bio-Sci Corp), overnight at 4°C with rotation. Samples were washed, and DNA was isolated as described in Austin et al. (1999). Labeling, microarray hybridization, and washing were performed as described above.

Expression arrays

Total RNA was prepared from ~5 $\times$ 10^6 cells using the RNeasy Kit (QIAGEN). mRNA was labeled and hybridized to Affymetrix *Drosophila* Genome 2.0 Arrays by the Duke core facility.

RNAi

Primers 5′-TTAATACGACTCACTATAGGGAGACTATCAGTATCAAGAACAGGCG and 3′-TTAATACGACTCACTATAGGGAGATGCTTTCCACCAGACTG were used to amplify an ~700-bp fragment

from the *dup* ORF. dsRNA was prepared from the PCR product using the T7 Ribomax Express Large Scale kit (Promega). 2×10^7 Kc167 cells were collected, centrifuged at 1000 rpm for 5 min, washed with serum-free medium (SFM), and resuspended in 5 mL of SFM. Fifteen micrograms of DUP dsRNA was added to 1×10^6 Kc167 cells per mL and allowed to incubate for 1 h. Five milliliters of 2× medium (Schneider's insect cell medium, 20% FBS, 2% PSG) and DMSO was added and allowed to incubate for 24 h. Cells were released from DMSO and allowed to incubate for an additional 24 h.

Chromatin fractionation

Chromatin fractionation was performed as described (Wysocka et al. 2001). Soluble and chromatin fractions were incubated for 5 min at 95°C and loaded on a 6% polyacrylamide gel for subsequent Western blotting to nitrocellulose. Blots were blocked in 5% milk/TBS-T for 20 min at 25°C. anti-dORC2 antibody was diluted 1:3000 in TBS-T, anti-MCM (AS1.1) 1:100 in TBS-T, and cohesin subunits 1:3000 in 5% milk/TBS-T and incubated overnight at 4°C. All secondary antibodies were diluted 1:10,000 in 5% milk/TBS-T with 0.005% SDS. ORC and SMC1 were detected with Alexa Fluor 680 goat anti-rabbit IgG (Invitrogen), MCM2-7 with IRDye 800 conjugated anti-mouse IgG (Rockland Immunochemicals), and Nipped-B and SA with IRDye 800 conjugated anti-guinea pig IgG (Rockland Immunochemicals). Blots were scanned on a LICOR Imaging system.

SVM classification

We used the LIBSVM tool (C-C Chang and C-J Lin, http://www.csie.ntu.edu.tw/~cjlin/libsvm) to train different support vector machine (SVM) classifiers on the ORC binding data. We generated feature vectors from 5050 positive (ORC associated) and 5050 negative (not associated with ORC) 1-kb sequences. The set of ORC negative sequences was randomly chosen, but biased to have two-thirds of the sequence set overlap transcription start sites. Of the 5050 positive sequences and 5050 negative sequences, we used 80% as the training set and 20% as the test set. All the ROC curves shown in Figure 6 were computed on the test set. For each SVM classifier, we optimized the parameters (the cost parameter and the width of the kernel's radial basis function) by performing 10-fold cross-validation on the training set and choosing the parameters that gave the best cross-validation accuracy. In each of the 10 rounds of cross-validation, we used 90% of the training set for training and the remaining 10% of the training set for testing.

Correlation of *k*-mer features with the class

For each of the 2772 sequence features used by the SVM classifiers, we computed the Pearson correlation coefficient between the feature and the class value. The class value was set to +1 for all positive (ORC-associated) sequences and to −1 for all negative (ORC-free) sequences. The correlation coefficient ranges between −1 and +1. Features with a positive correlation coefficient are specific to ORC-associated sequences, while features with a negative correlation coefficient are specific to ORC-free sequences.

Nucleosome occupancy for SVM analysis

To compute the average nucleosome occupancy over *k*-mers we used the in vitro nucleosome occupancy data of Kaplan et al. (2009). For each *k*-mer, we identified all of its occurrences in the yeast genome and then computed the average nucleosome occupancy over all occurrences of the *k*-mer.

Acknowledgments

We thank members of the MacAlpine laboratory for critical reading of the manuscript and Dale Dorsett for cohesin antibodies and advice. This work was supported by a CAREER award from the National Science Foundation 0347801 (A.J.H.), an Alfred P. Sloan Research Fellowship (A.J.H.), a Whitehead Foundation Scholar Award (D.M.M.), and the National Institutes of Health grants P50-GM081883-01 (A.J.H.) and HG004279 (D.M.M.).

References

Aggarwal BD, Calvi BR. 2004. Chromatin regulates origin activity in *Drosophila* follicle cells. *Nature* **430:** 372–376.

Ahmad K, Henikoff S. 2002. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol Cell* **9:** 1191–1200.

Aladjem MI, Fanning E. 2004. The replicon revisited: An old model learns new tricks in metazoan chromosomes. *EMBO Rep* **5:** 686–691.

Austin R, Orr-Weaver T, Bell S. 1999. *Drosophila* ORC specifically binds to ACE3, an origin of DNA replication control element. *Genes & Dev* **13:** 2639–2649.

Balasov M, Huijbregts RP, Chesnokov I. 2007. Role of the Orc6 protein in origin recognition complex-dependent DNA binding and replication in *Drosophila melanogaster*. *Mol Cell Biol* **27:** 3143–3153.

Bartkuhn M, Straub T, Herold M, Herrmann M, Rathke C, Saumweber H, Gilfillan GD, Becker PB, Renkawitz R. 2009. Active promoters and insulators are marked by the centrosomal protein 190. *EMBO J* **28:** 877–888.

Beall E, Manak J, Zhou S, Bell M, Lipsick J, Botchan M. 2002. Role for a *Drosophila* Myb-containing protein complex in site-specific DNA replication. *Nature* **420:** 833–837.

Bell S, Dutta A. 2002. DNA replication in eukaryotic cells. *Annu Rev Biochem* **71:** 333–374.

Bell S, Kobayashi R, Stillman B. 1993. Yeast origin recognition complex functions in transcription silencing and DNA replication. *Science* **262:** 1844–1849.

Bosco G, Du W, Orr-Weaver T. 2001. DNA replication control through interaction of E2F–RB and the origin recognition complex. *Nat Cell Biol* **3:** 289–295.

Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H, Prioleau MN. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc Natl Acad Sci* **105:** 15837–15842.

Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* **459:** 927–930.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848.

Cvetic C, Walter JC. 2005. Eukaryotic origins of DNA replication: Could you please be more specific? *Semin Cell Dev Biol* **16:** 343–353.

Danis E, Brodolin K, Menut S, Maiorano D, Girard-Reydet C, Mechali M. 2004. Specification of a DNA replication origin by a transcription complex. *Nat Cell Biol* **6:** 721–730.

Dillin A, Rine J. 1998. Roles for ORC in M phase and S phase. *Science* **279:** 1733–1737.

Dion MF, Kaplan T, Kim M, Buratowski S, Friedman N, Rando OJ. 2007. Dynamics of replication-independent histone turnover in budding yeast. *Science* **315:** 1405–1408.

Dominguez-Sola D, Ying CY, Grandori C, Ruggiero L, Chen B, Li M, Galloway DA, Gu W, Gautier J, Dalla-Favera R, et al. 2007. Nontranscriptional control of DNA replication by c-Myc. *Nature* **448:** 445–451.

Foss M, McNally F, Laurenson P, Rine J. 1993. Origin recognition complex (ORC) in transcriptional silencing and DNA replication in *S. cerevisiae*. *Science* **262:** 1838–1844.

Gillespie PJ, Hirano T. 2004. Scc2 couples replication licensing to sister chromatid cohesion in *Xenopus* egg extracts. *Curr Biol* **14:** 1598–1603.

Glynn EF, Megee PC, Yu HG, Mistrot C, Unal E, Koshland DE, DeRisi JL, Gerton JL. 2004. Genome-wide mapping of the cohesin complex in the yeast *Saccharomyces cerevisiae*. *PLoS Biol* **2:** e259. doi: 10.1371/journal.pbio.0020259.

Goren A, Tabib A, Hecht M, Cedar H. 2008. DNA replication timing of the human beta-globin domain is controlled by histone modification at the origin. *Genes & Dev* **22:** 1319–1324.

Henikoff S, Henikoff JG, Sakai A, Loeb GB, Ahmad K. 2009. Genomewide profiling of salt fractions maps physical properties of chromatin. *Genome Res* **19:** 460–469.

Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, Chang CW, Lyou Y, Townes TM, Schubeler D, Gilbert DM, et al. 2008. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol* **6:** e245. doi: 10.1371/journal.pbio.0060245.
Huang DW, Fanti L, Pak DT, Botchan MR, Pimpinelli S, Kellum R. 1998. Distinct cytoplasmic and nuclear fractions of *Drosophila* heterochromatin protein 1: Their phosphorylation levels and associations with origin recognition complex proteins. *J Cell Biol* **142:** 307–318.
Hyrien O, Maric C, Mechali M. 1995. Transition in specification of embryonic metazoan DNA replication origins. *Science* **270:** 994–997.
Iyer V, Struhl K. 1995. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J* **14:** 2570–2579.
Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458:** 362–366.
Karnani N, Taylor C, Malhotra A, Dutta A. 2007. Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. *Genome Res* **17:** 865–876.
Knott SR, Viggiani CJ, Tavare S, Aparicio OM. 2009. Genome-wide replication profiles indicate an expansive role for Rpd3L in regulating replication initiation timing or efficiency, and reveal genomic loci of Rpd3 function in *Saccharomyces cerevisiae*. *Genes & Dev* **23:** 1077–1090.
Kriegstein H, Hogness D. 1974. Mechanism of DNA replication in *Drosophila* chromosomes: Structure of replication forks and evidence for bidirectionality. *Proc Natl Acad Sci* **71:** 135–139.
Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390:** 249–256.
Lengronne A, Katou Y, Mori S, Yokobayashi S, Kelly GP, Itoh T, Watanabe Y, Shirahige K, Uhlmann F. 2004. Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature* **430:** 573–578.
Lucas I, Palakodeti A, Jiang Y, Young DJ, Jiang N, Fernald AA, Le Beau MM. 2007. High-throughput mapping of origins of replication in human cells. *EMBO Rep* **8:** 770–777.
MacAlpine D, Rodriguez H, Bell S. 2004. Coordination of replication and transcription along a *Drosophila* chromosome. *Genes & Dev* **18:** 3094–3105.
Maiorano D, Moreau J, Mechali M. 2000. XCDT1 is required for the assembly of pre-replicative complexes in *Xenopus laevis*. *Nature* **404:** 622–625.
Misulovin Z, Schwartz YB, Li XY, Kahn TG, Gause M, MacArthur S, Fay JC, Eisen MB, Pirrotta V, Biggin MD, et al. 2008. Association of cohesin and Nipped-B with transcriptionally active regions of the *Drosophila melanogaster* genome. *Chromosoma* **117:** 89–102.
Mito Y, Henikoff JG, Henikoff S. 2007. Histone replacement marks the boundaries of *cis*-regulatory domains. *Science* **315:** 1408–1411.
Narlikar L, Gordan R, Hartemink A. 2007. A nucleosome-guided map of a transcription factor binding sites in yeast. *PLoS Comput Biol* **3:** e215. doi: 10.1371/journal.pcbi.0030215.
Nasmyth K. 2002. Segregating sister genomes: The molecular biology of chromosome separation. *Science* **297:** 559–565.
Norio P, Kosiyatrakul S, Yang Q, Guan Z, Brown NM, Thomas S, Riblet R, Schildkraut CL. 2005. Progressive activation of DNA replication initiation in large domains of the immunoglobulin heavy chain locus during B cell development. *Mol Cell* **20:** 575–587.
Ozsolak F, Song J, Liu X, Fisher D. 2007. High-throughput mapping of the chromatin structure of human promoters. *Nat Biotechnol* **25:** 244–248.
Pak DT, Pflumm M, Chesnokov I, Huang DW, Kellum R, Marr J, Romanowski P, Botchan MR. 1997. Association of the origin recognition complex with heterochromatin and HP1 in higher eukaryotes. *Cell* **91:** 311–323.
Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, Canzonetta C, Webster Z, Nesterova T, et al. 2008. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132:** 422–433.
Rando OJ, Chang HY. 2009. Genome-wide views of chromatin structure. *Annu Rev Biochem* **78:** 245–271.
Remus D, Beall E, Botchan M. 2004. DNA topology, not DNA sequence, is a critical determinant for *Drosophila* ORC-DNA binding. *EMBO J* **23:** 897–907.
Royzman I, Austin R, Bosco G, Bell S, Orr-Weaver T. 1999. ORC localization in *Drosophila* follicle cells and the effects of mutations in dE2F and dDP. *Genes & Dev* **13:** 827–840.
Rusche LN, Kirchmaier AL, Rine J. 2003. The establishment, inheritance, and function of silenced chromatin in *Saccharomyces cerevisiae*. *Annu Rev Biochem* **72:** 481–516.
Saha S, Shan Y, Mesner L, Hamlin J. 2004. The promoter of the Chinese hamster ovary dihydrofolate reductase gene regulates the activity of the local origin and helps define its boundaries. *Genes & Dev* **18:** 397–410.
Santocanale C, Diffley JF. 1998. A Mec1- and Rad53-dependent checkpoint controls late-firing origins of DNA replication. *Nature* **395:** 615–618.
Schübeler D, Scalzo D, Kooperberg C, van Steensel B, Delrow J, Groudine M. 2002. Genome-wide DNA replication profile for *Drosophila melanogaster*: A link between transcription and replication timing. *Nat Genet* **32:** 438–442.
Schwaiger M, Stadler MB, Bell O, Kohler H, Oakeley EJ, Schubeler D. 2009. Chromatin state marks cell-type- and gender-specific replication of the *Drosophila* genome. *Genes & Dev* **23:** 589–601.
Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. *Science* **322:** 1849–1851.
Sequeira-Mendes J, Diaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, Gomez M. 2009. Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet* **5:** e1000446. doi: 10.1371/journal.pgen.1000446.
Shimada K, Gasser SM. 2007. The origin recognition complex functions in sister-chromatid cohesion in *Saccharomyces cerevisiae*. *Cell* **128:** 85–99.
Shirahige K, Hori Y, Shiraishi K, Yamashita M, Takahashi K, Obuse C, Tsurimoto T, Yoshikawa H. 1998. Regulation of DNA-replication origins during cell-cycle progression. *Nature* **395:** 618–621.
Song JS, Johnson WE, Zhu X, Zhang X, Li W, Manrai AK, Liu JS, Chen R, Liu XS. 2007. Model-based analysis of two-color arrays (MA2C). *Genome Biol* **8:** R178. doi: 10.1186/gb-2007-8-8-r178.
Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41:** 393–395.
Suter B, Schnappauf G, Thoma F. 2000. Poly(dA·dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. *Nucleic Acids Res* **28:** 4083–4089.
Takahashi TS, Yiu P, Chou MF, Gygi S, Walter JC. 2004. Recruitment of *Xenopus* Scc2 and cohesin to chromatin requires the pre-replication complex. *Nat Cell Biol* **6:** 991–996.
Takahashi TS, Basu A, Bermudez V, Hurwitz J, Walter JC. 2008. Cdc7/Drf1 kinase links chromosome cohesion to the initiation of DNA replication in *Xenopus* egg extracts. *Genes & Dev* **22:** 1894–1905.
Tower J. 2004. Developmental gene amplification and origin regulation. *Annu Rev Genet* **38:** 273–304.
Uhlmann F, Nasmyth K. 1998. Cohesion between sister chromatids must be established during DNA replication. *Curr Biol* **8:** 1095–1101.
van Daal A, Elgin SC. 1992. A histone variant, H2AvD, is essential in *Drosophila melanogaster*. *Mol Biol Cell* **3:** 593–602.
Vashee S, Cvetic C, Lu W, Simancek P, Kelly T, Walter J. 2003. Sequence independent DNA binding and replication initiation by the human origin recognition complex. *Genes & Dev* **17:** 1894–1908.
Vogelauer M, Rubbi L, Lucas I, Brewer B, Grunstein M. 2002. Histone acetylation regulates the time of replication origin firing. *Mol Cell* **10:** 1223–1233.
Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, et al. 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451:** 796–801.
Wohlschlegel JA, Dwyer BT, Dhar SK, Cvetic C, Walter JC, Dutta A. 2000. Inhibition of eukaryotic DNA replication by geminin binding to Cdt1. *Science* **290:** 2309–2312.
Wysocka J, Reilly PT, Herr W. 2001. Loss of HCF-1-chromatin association precedes temperature-induced growth arrest of tsBN67 cells. *Mol Cell Biol* **21:** 3820–3829.

Received July 11, 2009; accepted in revised form November 23, 2009.