Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features

Christelle Cayrou[1], Philippe Coulombe[1], Alice Vigneron, Slavica Stanojcic[†], Olivier Ganier, Isabelle Peiffer, Eric Rivals[‡], Aurore Puy, Sabine Laurent-Chabalier, Romain Desprat[¶], and Marcel Méchali*

Institute of Human Genetics, CNRS, 141 Rue de la Cardonille, 34980 Montpellier, France; [‡]Laboratoire d'Informatique, de Robotique et de Microelectronique de Montpellier, UM2-CNRS, 161 Rue Ada, 34095 Montpellier, France;

[1]Co-first authors

Present addresses: [†]UMR1333 INRA, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
[¶] Albert Einstein College of Medicine, Bronx, NY 10461, USA

*Corresponding author
Tel.: +33 434 359 917
Fax :+33 434 359 920
e-mail: mechali@igh.cnrs.fr

Abstract

In metazoans, thousands of DNA replication origins (Oris) are activated to replicate DNA at each cell cycle. Although their timing of activation is better understood, their genomic organization and their genetic nature remain elusive. Here, we identified Oris by nascent strand (NS) purification and characterized their common features by performing a genome-wide analysis in both *Drosophila* and mouse cell lines. We show that in both species most CpG islands (CGI) contain Oris, although methylation is nearly absent in *Drosophila*, indicating that this epigenetic mark is not crucial for defining the initiation event. Initiation of DNA synthesis starts at the borders of CGIs, resulting in a striking bimodal distribution of NS, suggestive of a dual initiation event. We also found that Oris contain a unique nucleotide skew around NS peaks, characterized by G/T and C/A over-representation at the 5' and 3' of Ori sites, respectively. Repeated GC-rich elements were detected, which are good predictors of Oris, suggesting that common sequence features are part of metazoan Oris. In the heterochromatic chromosome 4 of *Drosophila*, Oris are strongly correlated with HP1 binding sites. At the chromosome level, regions rich in Oris are early replicating, whereas Ori-poor regions co-localize with late replicating domains during the cell cycle. The genome-wide analysis was coupled with a DNA combing analysis to unravel the organization of replication origins. The results indicate that Oris are present in a large excess, but their activation does not occur at random. They are organized in groups of site-specific but flexible origins that define replicons, where a single origin is activated in each replicon. This organization provides both site specificity and Ori firing flexibility in each replicon, allowing possible adaptation of DNA replication to environmental cues and cell fates.

INTRODUCTION

In metazoans, DNA replication is initiated at thousands of chromosomal sites during each S phase. These DNA replication origins (Oris) should be activated only once at each cell cycle to avoid any amplification and maintain genome integrity. This is an important task in human or mouse cells, where 30 000 replication Oris are activated at each cell division. In prokaryotes as well as in bacterial and animal viruses, Oris are sequence-specific. In *Saccharomyces cerevisiae*, Oris are identified by specific DNA elements, called Autonomous Replication Sequences (ARS), which have a common AT-rich 11 bp ARS Consensus Sequence (ACS). However, sequence-specificity identifies potential Oris, but does not determine their selection. Indeed, of the 12,000 ACS present in *S. cerevisiae* genome only 400 (3.3%) are functional (Nieduszynski et al. 2006). In *S. pombe*, ARS were also identified, but they do not share a specific consensus sequence and are characterized by AT-rich islands (Segurado et al. 2003; Dai et al. 2005; Heichinger et al. 2006) and polydA/dT tracks.

In multicellular organisms, how Oris are defined remains elusive despite considerable efforts to unravel a replication origin code. Until recently, only a few Oris were identified in metazoans. They appear to have variable features, since they can be extremely site-specific, as the human lamin B2 Ori (Abdurashidova et al. 2000), or have a broad site specification like the *DHFR* Ori (Dijkwel and Hamlin 1995). No consensus motif with predictive value has been found yet and therefore it has been proposed that some unknown epigenetic features could identify metazoan Oris. In agreement, transcription and chromatin status were found to influence Ori specification at specific gene domains (Aladjem 2007; Mechali 2010).

Unraveling common features of metazoan Oris requires a large-scale identification procedure, the development of which was hampered by the lack of a genetic test, like the yeast ARS test, and by the fact that methods to map Oris were not always adapted to robust genome-scale analysis. The first genome-scale studies to localize Oris in mouse and human cells (Cadoret et al. 2008; Sequeira-Mendes et al. 2009) have observed a correlation with unmethylated CpG islands (CGI) and some overlap with promoter regions (Delgado et al. 1998; Sequeira-Mendes et al. 2009). However, it was not clear whether CGI were a specific mark of Oris or of the associated promoters.

In order to identify new features of eukaryotic Oris, first we upgraded the method used to map nascent DNA strands (NS) at active Oris to a specificity and reproducibility compatible with genome-scale analysis. Then, we used this method in four cell systems: mouse embryonic stem cells (ES), mouse teratocarcinoma cells (P19), mouse embryonic fibroblasts (MEFs) and *Drosophila* cells (Kc). We characterized up to 2748 Oris on mouse chromosome 11 (P19 cells) and 6184 Oris in the *Drosophila* genome. The three mouse cell lines show common and also specific Oris, suggesting that Oris may contribute to cell identity specification. Ori-rich domains co-localize with the previously defined early replicating domains, whereas Ori-poor domains correspond to late replicating regions. Oris are also preferentially concentrated in transcription promoter regions in mouse cells. We found that Oris are strongly associated with CpG islands and exhibit a bimodal distribution that is suggestive of an asymmetric initiation event. Nucleotide sequence asymmetry is also present at the initiation site, and the analysis reveals specific GC-rich motifs in both mouse and *Drosophila* Oris at initiation sites. A strong correlation between HP1 binding sites and Oris was found at *Drosophila* chromosome 4, which is mainly constituted of heterochromatin. In combination with DNA combing data, our genome-scale results demonstrate that metazoan Oris are in large excess relative to their standard use, and computational simulations suggest that flexible Oris are clustered in groups that define the replicons, where activation of one Ori silences the others in the same group.

RESULTS

Genome-scale mapping of DNA replication origins: general features

Ori mapping by purification of RNA-primed NS is a well-established procedure. NS purification is achieved through (1) size selection of RNA-primed NS at Oris and (2) specific enrichment using lambda exonuclease (Figure 1A) as RNA-primed NS are resistant to this enzyme, while contaminating broken DNA is degraded efficiently. However, only a very small amount of NS (at most 20 ng per 10^8 cells (Cadoret et al. 2008) can be recovered and, consequently, the slightest contamination by broken DNA pieces would dramatically raise the background. In order to perform genome-scale mapping of Oris, we upgraded this procedure by improving NS purification (as detailed in Methods) and enrichment through two or three rounds of digestion with high specificity, custom-made exonuclease (Fermentas) to remove non RNA-primed NS. The efficiency of this step is crucial, as explained in Methods. In addition, we used tiling microarrays (Nimblegen) in which oligonucleotides were spaced on average every 40 bp and which, therefore, gave a 4- to 6-fold more accurate resolution than in the previous genome-scale studies. Altogether, this improved method allowed us to detect a much larger number of consecutive positive oligonucleotides that scanned each Ori, and to score Oris with increased confidence.

We used this procedure to obtain genome-scale maps of Oris in mouse ES, P19 and MEF cells as well as in *Drosophila* Kc cells in order to highlight conserved features between vertebrate and invertebrate Oris as well as to assess the impact of cell differentiation on the Oris repertoire. The full data set (obtained using up to 4 different biological replicates for each cell line) consisted of a continuous 60.4 Mbp fragment on mouse chromosome 11, which we considered representative of the mouse genome as it exhibits replication timing and transcription features that are comparable to those of the entire genome (Supplementary Figure 1 A), and of 118.3 Mbp of *Drosophila* genome. This allowed analyzing the overall distribution of Oris at a chromosome scale in comparison to the average 0.68Mb regions of the previous genome-scale studies (Cadoret et al. 2008). NS maps showed enrichment at specific genomic locations with a high degree of reproducibility (see examples in Figure 1B and Supplementary Figure 1 B-E). To control the rate of false positives, for each probe, the $\log_2$-ratio value was normalized and the p-value was computed by applying the false discovery rate (FDR) correction (Benjamin and Hochberg 1995). A probe was considered significant when the p-value was lower than 5% (level of significance).With the FDR correction, potential Oris could be identified with high confidence. Moreover, as the minimum size of purified NS was 0.5 kb, Oris should be theoretically at least 1 kb (2 X 0.5 kb for the general case of a bidirectional Ori). Oris were thus defined as positive regions ($\log_2$-ratio>0) of at least 1 kb containing significant probes (see details in Supplementary information).

We further validated the Ori maps by quantitative PCR (qPCR) analysis of known Oris in the mouse *Myc* gene and *Hoxb* domain (Supplementary Figure 1F, G) and *Drosophila Histone* gene locus (Supplementary Figure 2A), as well as of randomly chosen putative Oris validated in this study (Supplementary Figure 2 B, C). 17 out of 18 validated Oris showed significant NS enrichment. Only background signal was observed when total DNA or “NS” from mitotic cells was used for hybridization (Figure 1B and Supplementary Figure 1B, D and G), or when NS were RNase-treated before exonuclease digestion (data not shown), confirming the specificity of the purification procedure. Using ChIP-chip analysis, we also confirmed the presence of ORC2, a key component of Oris, at the mouse *Myc* Oris (Supplementary Figure 3A), with a profile similar to the one observed in human cells (Ghosh et al. 2006). We also found that the ORC2 signal profile strongly correlated with the profiles of the NS peaks (Figure 1d and Figure 3A, B).

By using exponentially growing cells we can potentially score all Oris activated during the entire S-phase and possible variations in Ori usage among cells would not prevent their detection. We identified 2412 Oris in ES, 2748 in P19, 2231 in MEF and 6184 in *Drosophila* Kc cells (Figure 1C). We noted that a large fraction of Oris (up to 44 %) was common to the three mouse cell lines (Supplementary Figure 4A). The Ori repertoires of ES and P19 cells, both pluripotent cells, were statistically more related to each other than to MEF Oris (see details in Supplementary Information and Supplementary Figure 4). We also observed that *Drosophila* cells had a denser Ori repertoire. Finally, Ori sequences were significantly more conserved compared to non-Ori sequences in both *Drosophila* and mouse cells ($p<2*10^{-16}$), suggesting that important evolutionary conserved elements might be present in Oris (data not shown).

Over-representation of Oris at transcriptionally active promoters

We next analyzed the association of Oris with genes. To determine if the observed correlation was significant, our datasets were compared with 1000 randomized Ori datasets, which contained the same number and length of Oris but randomly located, to evaluate association by chance. Oris were found in both intergenic and genic regions (Supplementary Figure 5A) with a significant preference for intra-genic localization (p<0.001). This association was not stochastic, as genes with Oris were significantly more actively transcribed than genes without Oris (Figure 2A). Compared to randomization, Oris showed a significant association with promoters (in mouse cells) and exons (in mouse and *Drosophila* cells) (Figure 2B and Supplementary Figure 5B). Ori distribution relative to gene length was also analyzed. Gene lengths were normalized on a scale of 0 to 100 (corresponding to the start and end of genes). Mouse Oris were found all along the genes, although they were over-represented at the start of genes (Figure 2C). Conversely, Ori concentration at promoter regions was not observed in *Drosophila*.

We then examined accurately the NS signal strength around transcription start sites (TSS) that had been aligned (see details in Supplementary information). In mouse cells, we observed a striking bimodal distribution with two major NS peaks located on each side of TSS (Figure 2D and Supplementary Figure 6A and B). This pattern was lost when the location of TSS along the chromosome was randomized (data not shown), indicating that the bimodal distribution of NS signal strength around TSS was significant. The two putative replication initiation sites were separated by a shallow valley centered on the TSS, with the NS peaks located around 600 bp upstream and downstream of the TSS. A careful analysis showed four categories of TSS-linked Oris (Supplementary Figure 7). Most of them (77%) exhibited the bimodal distribution. Other minor categories include unimodal TSS-linked Oris with the peak located upstream (11%), downstream (8%), or on the TSS (4%). We also noted that 67% of TSS with divergent transcription (Seila et al. 2008) contained a bimodal Ori, whereas the bimodal Ori-TSS association decreased to 35% when all TSS were considered. Conversely, in *Drosophila* cells, TSS were not enriched in Oris (Figure 2E) and NS did not show a bimodal distribution, but increased density within genes as opposed to promoter regions. This contrasting result led us to ask whether another element present at mammalian TSS, but not within TSS, was associated with Oris, both in *Drosophila* and mouse cells.

Replication origins are bimodal and enriched at CpG islands in mouse and *Drosophila*

Mammalian promoters, particularly those of highly expressed genes, are CpG-rich while genes that are highly regulated during development are often CpG-poor or -free (Cross and Bird 1995). CpG-rich sequences, known as CGI, are usually defined as regions of at least 200 bp in length with 60% of CG and a ratio of CpG observed/CpG expected >0.6. We asked whether such elements could explain the bimodal NS signal strength around TSS, and we analyzed TSS with CGI (n=820) and without CGI (n=434) separately. In the three mouse cell lines, the NS bimodal signal strongly associated with CGI-positive TSS, but not with TSS that did not contain CGI. This result suggests that the over-representation of Oris at promoter regions was in reality a consequence of their association with CGI (Figure 2F and Supplementary Figure 6C, D). In addition, although CGI represented only 1.3% of the mouse genome, most of them (up to 73%) were strongly (p<0.001) associated with Oris in the three mouse cell lines (Figure 3A and Supplementary Figure 6E, F), and the NS signal strength was bimodal around CGI as well (Figure 3B and Supplementary Figure 6E, F). Also, Oris that were common to the three mouse cell lines were significantly enriched in CGI (Figure 3E).

In higher metazoans, CpG dinucleotides are subject to cytosine methylation, which results in their depletion from the genome over time during evolution (Cross and Bird 1995; Illingworth and Bird 2009). Although cytosine methylation is almost inexistent in *D. melanogaster*, its genome contains regions with properties identical to those of mammalian CGI. We thus delimited about 20 000 ‘CGI-like’ regions that responded to the CGI definition and represented 5.9% of the *Drosophila* genome. 46% of these regions were significantly (p<0.001) associated with 59% of Oris (Figure 3C). Moreover, although the NS signal strength was not bimodal around *Drosophila* TSS (Figure 2E), it was bimodal around the *Drosophila* CGI-like regions (Figure 3D), like in mouse cells (Figure 3B). The bimodal curve was less accentuated than in the mouse, possibly linked to the fact that the CGI-like elements found in *Drosophila* were smaller (346 bp vs 606 bp in the mouse).

We conclude that CGI-related sequences are conserved determinants in a substantial part of mouse and *Drosophila* Oris, regardless of their genomic position. They do not need to be at promoter regions, or to rely on methylation, consistent with the presence of CGI-like sequences in exons in the *Drosophila* genome and the fact that CGI at mouse promoters are often demethylated. These results provide a novel possible function for CGI sequences in DNA replication that is conserved both in vertebrates and invertebrate species. Importantly, this role is, at least in *Drosophila,* independent of CpG methylation or of being localized close to a promoter region.

Oris are characterized by nucleotide asymmetry and CG-rich elements

We further investigated the GC/AT nucleotide composition of these Oris. Replication initiation sites (defined by the NS peaks) were located outside the central CGI, in a region with an AT content that was found similar to that of the whole genome (Figure 4A). We then asked whether the sequences flanking the NS peaks showed particular features. All *Drosophila* Ori sequences were aligned to their NS peaks (Figure 4B) and the frequency of each nucleotide was calculated in a 1000 bp region upstream and downstream of such peaks. This analysis revealed a clear nucleotide asymmetry with over-representation of T and G at the 5’ and of A and C at the 3’ of the NS peaks. This nucleotide bias could easily be visualized when individual Oris were aligned (Figure 4C). Similar results were obtained in mouse cells although the skew was more marked for C and G (Supplementary Figure 8). This asymmetry was not observed in randomized Oris (data not shown). We conclude that *Drosophila* and mouse initiation sites display a characteristic nucleotide asymmetry that is not observed at more upstream or downstream regions.

Although no consensus sequence has been associated with metazoan Oris ((Mechali 2010) for review), we investigated using the MEME suite (http://meme.sdsc.edu/meme4_4_0) whether some enriched motifs could be identified in Oris, Due to its smaller genome, *Drosophila* sequences were first analyzed. We submitted 2 kb stretches of *Drosophila* DNA sequences centered on NS peaks using specific parameters (see Methods). The more frequent motifs recovered in different MEME runs using different batches of *Drosophila* or mouse Oris were GC-rich motifs with a repetitive nature. In *Drosophila*, two motifs (Figure 4D) were associated with more than 60% Oris (data not shown). Interestingly, these motifs were often found in known Oris (data not shown). In mouse, G-rich motifs were also consistently recovered (Figure 4D).

We further characterized *Drosophila* Oris using the R'MES program (http://migale.jouy.inra.fr/outils/mig/rmes/), which investigates whether a motif is over-represented in a set of sequences (Hoebeke and Schbath 2006). As R'MES is limited to 13 nucleotide-motifs, we asked which among the 67,108,864 possible 13-mers occurred more frequently. $(TGC)_4T$, its cyclic permutations and complementary sequences were significantly over-represented in agreement with the MEME analysis (see motif 2 in Figure 4D). (TA)n sequences, which are very frequent in microsatellites, were not significantly associated with Oris (data not shown).

In summary, similar GC-rich motifs were found associated with metazoan Oris. Although it would be hazardous to conclude that there is a strong sequence specificity, these results altogether indicate that at least two general sequence features are associated with Oris: i) a bias towards GC-rich elements and ii) a clear nucleotide asymmetry upstream as well as downstream of these elements, at the position of NS synthesis.

In *Drosophila* heterochromatin, HP1 sites are associated with Oris

Drosophila chromosome 4 is unusual as it represents 1% of the genome and is organized mainly in heterochromatin that replicates early in S-phase in Kc cells (Schwaiger et al. 2009). The density of CGI-like regions was eight-fold lower than in other chromosomes (Figure 5A), but this feature cannot explain the early replication timing of chromosome 4. In fission yeast, the HP1 ortholog Swi6 is involved in early S-phase replication of heterochromatic pericentromeres and of the *MAT* locus (Hayashi et al. 2009). HP1 is believed not to be involved in Ori positioning, but rather in favoring Ori firing by recruiting DDK kinases (Hayashi et al. 2009). To evaluate a possible link between HP1 and replication in metazoans, Oris were correlated with reported HP1 binding sites (de Wit et al. 2007) and replication timing (Schwaiger et al. 2009) in *Drosophila* Kc cells. First, a strong positive correlation between early timing of replication and high HP1 binding was detected ($p<2x10^{-16}$, Figure 5B, left panel). This association was lost when the HP1 probes were randomized (Figure 5B, right panel). Moreover, 100% Oris were associated with HP1 sites (Figure 5C, D), indicating that, in CGI-poor regions and in heterochromatin, HP1 binding sites may contribute to Ori recognition.

DNA replication origins are organized in large, high-density domains

We then evaluated the higher order organization of Oris. In each cell line, Ori density along the chromosome was investigated (see details in Supplementary information). In mouse cells, Ori density along chromosome 11 was not uniform, with areas of low Ori density separated by large high density areas (Figure 6A). These regions were at similar positions in the three mouse cell lines (Figure 6A). We also found that Ori density correlated well with the replication timing domains (Figure 6A, B) which were previously identified by genome-scale analysis in mammals (White et al. 2004; Hiratani et al. 2008), suggesting that replication timing is controlled by Ori density. *Drosophila* Kc cells also showed Ori enrichment at early replication domains, albeit at a lower degree (data not shown).

Hierarchic organization of metazoan Oris

Genome-scale data score all the Oris that are activated in a given cell population and thus allow the identification of all potential sites which can serve as Oris, although they can vary from cell to cell within a given population. To study the actual Ori usage in *Drosophila* and mouse individual cells, DNA combing analysis was performed. Two consecutive pulses using different deoxynucleotide analogs allowed precise localization of Oris on single DNA molecules (Figure 7A, Methods, and Supplemental Methods). The size of the fibers analyzed ranged between 194 and 900 kb, and mouse cell lines presented similar inter-origin distances (136 kb in MEF and 139 kb in ES cells, Figure 7A). This is in agreement with the mean inter-origin distance (137 kb) recently found in the human MRC-5 fibroblast cell line at the *FRA3B* locus (Letessier et al. 2011) where longer domains were analyzed. Conversely, a near two-fold difference between the mouse (average 137.5 kb) and *Drosophila* (73 kb) inter-origin distances was observed (Figure 7A).

The pattern of Ori usage was further investigated by combining our genome-scale Ori data (from Kc, ES and MEF cells) and the inter-origin distances obtained by DNA combing. For simplicity, only the MEF replication dynamics will be explained in more detail. If all mapped Oris were activated in all cells (100% firing efficiency) the resulting, very short, inter-origin distance distribution would be significantly different from the distribution observed using DNA combing (Figure 7B). Indeed, the comparison of genome-scale and DNA combing data suggests that, in MEFs, 1 every 5 Oris on average was activated in a given DNA molecule (19.8 % firing efficiency, Figure 7B). Similar values were obtained for ES and Kc cells. Our results are consistent with the notion that metazoan Oris are redundant, and that only a small proportion of them is effectively used at each cell cycle.

We then assessed how different models of metazoan DNA replication could explain our data. We first considered the ‘Random Ori firing’ model, in which Oris are randomly activated at a density based on the DNA combing experiments. The mean inter-origin distance of fired Oris was then identical to the value obtained by DNA combing (Figure 7B); however the simulated inter-origin distance distribution (Figure 7C, in red, and Supplementary Figure 9A, B) was different from the experimental distribution obtained in combing experiments (Figure 7C in grey, and Supplementary Figure 9A, B). Specifically, the ‘Random Ori firing’ model led to populations of short and long inter-origin distances that were not observed in the DNA combing experiments. The presence of a group of large inter-origin distances is in agreement with the random gap problem (i.e., random firing leads to large gaps of unreplicated DNA that will persist at the end of S phase) (Laskey 1985; Hyrien et al. 2003).

Then, we evaluated the ‘Increasing Ori efficiency’ model (Rhind 2006). This model is based on the idea that Ori firing efficiency is not constant, but actually increases during S-phase (Figure 7D). This increase in efficiency ensures Ori firing in late S-phase in order to fill the remaining stretches of unreplicated DNA. To implement this model, fork speed was analyzed by DNA combing: replication forks in mouse cells were about twice faster than in *Drosophila* cells (1.77 vs 0.81 kb/min, Supplementary Figure 10A). Using these experimental values, Ori activation was simulated with increasing firing efficiency during S-phase progression

(Supplementary Figure 10B). The ‘Increasing Ori efficiency’ model solved the random-gap problem, as indicated by the disappearance of the population of large inter-origin distances observed in the ‘Random Ori’ model (Figure 7D). However, the simulated inter-origin distribution remained significantly different from the one observed by DNA combing (Figure 7 B, D, and Supplementary Figure 9). The simulated inter-origin distance distribution was wider and contained a much larger short-distance population than the one derived from the DNA combing data.

We then considered a ‘Flexible Replicon’ model in which adjacent Oris are functionally linked together over a defined distance that delineates a replicon, providing a multiple firing choice. In each replicon/group of Oris, one Ori is randomly activated and silences the others (Figure 7E). Oris were classified using hierarchical cluster analysis (see Supplementary information) to generate a dendrogram of Ori repartition along the chromosome. Clusters were obtained by cutting the dendrogram at the height that gave the strongest correlation with DNA combing data (Supplementary Figure 10C-E). The simulated inter-origin distances were very similar to the DNA combing values in these conditions (Figure 7B, E, and Supplementary Figure 9). This analysis suggests that replicons are on average 56 kb in length, contain 4.3 Oris and the inter-replicon distance is 117 kb in MEF cells (Supplementary Figure 10F). The ‘Flexible Replicon’ model was also applicable to ES and Kc cells (Supplementary Figure 9 and Supplementary Figure 10F). The model is thus robust and can accommodate changes in Ori density and firing efficiency.

Overall, these findings suggest that Oris are in large excess in metazoans and have a flexible use. However, Ori firing flexibility is not stochastic in the whole genome, but only inside each replicon. Metazoan replicons appear constituted of groups of potential and flexible adjacent Oris where activation of one Ori suppresses the surrounding Oris. The inter-origin distance is therefore the average distance between activated Oris in each group of flexible Oris.

DISCUSSION

This high resolution, genome-scale analysis of Oris allowed the identification of 13 575 Oris in four cell lines from two different metazoan species and the discovery of common organization and sequence features. The combined analysis of genome-scale and DNA combing data suggests that metazoan Oris are organized in replicons in which Ori flexibility is an essential feature.

Metazoan origins are bimodal and are enriched at actively transcribed genes and transcription start sites

In metazoans, replication timing and transcriptional activity are connected. Early activated Oris are in actively transcribed genes, whereas late replication is associated with poorly transcribed regions. We show here a strong correspondence between Ori density and timing of replication. Ori-rich regions are in early replicating domains; conversely Ori-poor regions correspond to late replicating domains. In mouse, but not in *Drosophila*, we found a significant enrichment of Oris at promoter regions, particularly at TSS, where most Oris have a bimodal structure, with two peaks of NS bordering the TSS regions. We then show that this bi-modal structure is mainly linked to CGI elements often found at promoter regions. We postulate that the bimodal nature of Oris is due to initiation with two start sites for the leading strand synthesis, separated by about 1 kb of sequence that might contain the Ori genetic determinants. The fusion of these two replication bubbles would rapidly lead to a single bubble at Oris (Figure 8A). Such mechanism is similar to the asymmetric bidirectional model of replication proposed for the human *DBF4* Ori (Romero and Lee 2008) and is reminiscent of initiation at the *E. coli* Ori (Fang et al. 1999), where the DNA helicase proceeds for at least 100 nucleotides before priming DNA synthesis.

We also observed that 67% of divergent TSS have an Ori (data not shown). Divergent transcription at TSS in mouse ES cells is associated with CpG-rich promoters, where antisense and sense short RNAs of 16-30 nucleotides are synthesized upstream and downstream of the TSS at two sites separated by 400-500 bp (Seila et al. 2008). The distance between the 3' ends of these transcripts is close to the mean distance we observed between the two NS peaks. It could be asked whether such short transcripts might be used for initiation of DNA replication, like in *E. coli* (Baker and Kornberg 1988; Skarstad et al. 1990), or Epstein-Barr Virus (Rennekamp and Lieberman 2011), where initiation of DNA replication is facilitated by transcription by RNA polymerases.

Metazoan origins exhibit common sequences features

We found a link between CGI and Oris in all cell types analyzed, as reported for a subset of human and mouse Oris (Delgado et al. 1998; Cadoret et al. 2008; Gomez and Antequera 2008; Sequeira-Mendes et al. 2009). CGI are essential elements for transcriptional control and imprinting in mammals and are regulated by DNA methylation. However, CGI-like regions are also present in *Drosophila* and they were significantly associated with Oris, which showed a bimodal NS distribution at these sites, like in mammals. Their association with Oris in *Drosophila* is intriguing as methylation is rather poor in this species. These data suggest that CGI, or some elements embedded in these regions, had a primary role in DNA replication and they further evolved to be used in transcriptional control. As the number of CGI elements in the genome is much lower than the number of potential Oris, other Ori classes must be present. However, CGI elements appear to be an important class of determinants of Ori localization, which explain why Oris are enriched at promoter regions without being obligatorily linked to transcription.

Two independent bioinformatics approaches showed that the Oris described in this study contain conserved features, which, although not as strict as the *S. cervisiae* ARS Consensus Sequence, reveal some bias toward GC-rich elements. NS enrichment peaks are not localized at the CpG-rich domain itself, but on its sides, suggesting that this domain might be a binding site for factors controlling NS synthesis upstream and downstream (Figure 8A). The NS synthesis sites are not GC-rich, but are characterized by more AT-rich sequences, in agreement with an easier opening of DNA (Figure 4A).

Another feature is the strong nucleotide skew we observed at NS positions, with a general bias for GT at the 5' side and for CA at the 3' side in both *Drosophila* and mouse Oris. Interestingly, bacterial Oris have a similar nucleotide skew (Lobry 1996), whereas *S. cerevisiae* ARS have only an (A/T) skew (Breier et al. 2004). We observed nucleotide asymmetry only around initiation sites and not in more upstream or downstream regions, another strong indication that these are true Oris. Nucleotide skew might be therefore a universal Ori property, possibly involved in the structure of DNA at Oris. It has been suggested that the nucleotide skew was a consequence of the mutational bias associated with DNA replication (Touchon et al. 2005).

Positive association of active Oris with HP1 in *Drosophila* chromosome 4

Heterochromatic DNA is generally believed to constitute late replicating domains. However, accumulating evidences indicate that a subset of heterochromatin DNA replicates early in S-phase (Hayashi et al. 2009) and this is the case for *Drosophila* chromosome 4 (Schwaiger et al. 2009). We detected a positive relationship between HP1 binding and early replication, as recently reported (Schwaiger et al. 2010). Moreover, we found a strong correlation between HP1 binding sites and Oris at this chromosome. This is in agreement with the interaction of HP1 with the Origin Recognition Complex (ORC) in higher eukaryotes (Pak et al. 1997). HP1 binding sites could help Ori recognition in compact heterochromatin regions. The fission yeast HP1 homolog Swi6 is required for early replication of heterochromatic regions (Hayashi et al. 2009). Swi6 stimulates Ori firing by recruiting DDK and facilitating pre-Initiation Complex formation. It would be interesting to investigate whether HP1 also stimulates Ori usage through DDK recruitment in higher eukaryotes. Although chromosome 4 represents only 1% of the *Drosophila* genome, this finding indicates that DNA replication could be facilitated by other means at specific chromatin domains, and strengthens the role of HP1 in Ori localization.

Replicons are groups of flexible origins

Our data show that, in metazoans, DNA replication firing appears at first as a relatively inefficient system as reported in yeast (Friedman et al. 1997; Dai et al. 2005; Heichinger et al. 2006). On average, there are four- to five-fold more potential Oris than used. Therefore, a replicon cannot be considered as the distance between two Oris. Rather, our data suggests that a replicon is a group of several adjacent and flexible potential Oris, in which only one Ori is activated per cell and per cell cycle, and the others are silenced (Figure 8B). The possibility to use several Oris in each replicon would increase their firing probability. In other words, flexibility is not stochastic in the whole genome but only inside each replicon. Such Ori flexibility and abundance in each replicon might be needed to respond to variations in growth conditions, problems encountered by the replication fork and to ensure complete duplication. For instance, when the concentration of nucleotides is decreased, new Oris are activated in the hamster *Gna13* domain (Anglana et al. 2003; Ge et al. 2007). This model is also in agreement with the conservation of replication foci (clusters of replicons) in subsequent cell cycles, since flexibility will be mainly inside replicons. It is also in agreement with the notion of initiation zone used for the *DHFR* domain, where multiple Oris can be found at close intervals (Dijkwel and Hamlin 1995). If some sites are deleted, others, close-by, become activated (Mesner et al. 2003). Several potential Oris per replicon might allow choosing the more suitable Ori to be activated in a given chromatin context, which could vary according to the transcriptional status or cell identity. The proposed "Jesuit Model": "Many are called, few are chosen" (DePamphilis 1993) appears therefore to apply to the Flexible Replicon model.

Pioneer former work ((Berezney et al. 2000) for a review) as well as our DNA combing experiments indicate that Oris are often synchronously activated in clusters which can form replication foci. We thus propose that a replication cluster includes consecutive groups of adjacent flexible Oris (each group constituting a replicon) that are activated synchronously (Figure 8C). The selection of a given Ori within each replicon might depend on the cell fate or the organization of the chromatin domain. The Ori interference mechanism has been described in yeast (Brewer and Fangman 1993; Lebofsky et al. 2006), where firing at one Ori inhibits close-by Oris and this phenomenon could lead to the 100-120 kbp average size of the replicon. Activation of one Ori might promote looping out of the replicon resulting in the silencing of the other potential Oris (Figure 8B).

DNA replication origins: a barcode defining cell fate and cell identity?

Our data show that although flexible, Oris are at specific positions that appear to be mostly conserved among different cells. Pluripotent cells (ES or P19) have slightly more Oris than differentiated cells (MEF), but the size of the replicon (Figure 7) and the length of S phase (data not shown) are similar. Pluripotent ES cells may have fewer constraints than differentiated cells thus allowing an extended Ori choice. This is in agreement with the changes in Ori choice observed during differentiation in *Xenopus* (Hyrien et al. 1995) *Physarum*, (Maric et al. 2003) *Sciara* fly development (Lunyak et al. 2002), human B cell development (Norio et al. 2005) and in the chicken Globi locus (Dazy et al. 2006). In contrast, *Drosophila* or *Xenopus* early embryos have no transcriptional constraints and can use all the potential Oris to accelerate S phase. Indeed, in early embryos, Oris are activated at very close intervals, every 10-20 kb in *Xenopus* (Hyrien and Méchali 1993; Walter and Newport 1997; Lemaitre et al. 2005) and every 8-12 kb in *Drosophila* (Blumenthal et al. 1974). These values are close to the use of every Ori. Since here we show that potential mouse and *Drosophila* Oris are at conserved specific sites in the same species, one could ask whether the Oris activated in *Xenopus* and *Drosophila* early embryos are as random as previously thought. A maximum usage of specific sites rather than random Ori usage might regulate embryonic chromosome replication. The Ori position in the genome might therefore define a barcode that organizes chromosomal replication patterns, in which the choice and usage of each bar (Ori position) is defined according to cell growth and fate.

METHODS

Cells and Cell Culture

MEFs derived from 13.5-day mouse embryos were cultured as previously described (Hiratani et al. 2008) and used at passage 4 or 5. P19 cells were cultured as previously described (Gregoire et al. 2006). The ES cell line CGR8 was cultured in standard ES cell medium. *Drosophila* Kc cells were cultured in Schneider's medium (Invitrogen) supplemented with 10 % insect cell culture-tested FBS (Sigma). When necessary, mouse cells were synchronized in prometaphase with 100 ng/ml nocodazole for 12 hours. Kc cells were synchronized in prometaphase by incubation with 4 mM thymidine overnight, release in fresh medium for 4 hours and incubation with 1 mg/ml nocodazole overnight.

Nascent strand (NS) preparation

DNA purification

Dividing cells (2.5-5 $\times 10^8$) were washed in PBS, harvested and lysed in 15 ml DNAzol (Invitrogen) at room temperature (RT) for 5 min. Samples were digested with 200 µg/ml proteinase K at 37°C for 2 hours. We found that combining the proteinase K treatment with DNAzol significantly improved the yield of NS. Insoluble material was discarded by centrifugation at 3000 g at 4°C for 15 min and genomic DNA was precipitated with 15 ml 100% ethanol at RT for 5 min. DNA was transferred to a new tube and washed with 5 ml 70% ethanol at RT for 5 min and air-dried. DNA was resuspended in 2 ml TEN20 (10 mM Tris-Cl pH 7.9, 2 mM EDTA, 20 mM NaCl, 0.1% SDS, 1000 U RNasin) at 70°C, boiled for 10-15 min and chilled on ice.

NS purification by sucrose gradient

1 ml of denatured genomic DNA was loaded onto a 30 ml neutral 5 to 30% sucrose gradient prepared in TEN300 (10 mM Tris pH 7.9, 2mM EDTA, 300 mM NaCl) and centrifuged in a Beckman SW28 rotor at 24000 RPM, 4°C, for 20 h. 1 ml fractions were withdrawn from the top of the gradient using a wide-bore pipette tip. 50 µl of each fraction was run with appropriate size markers on a 2% alkaline agarose gel at 40-50 volts, overnight at 4°C. The gel was neutralized with 1x TBE and stained with Gel Red (Interchim). Fractions corresponding to 0.5-2.5 kb were pooled and precipitated with 2.5 volumes ethanol at -80°C for 15 min. Pellets were washed with 1 ml 70% ethanol and suspended in 20 µl water with 100 U RNasin (NEB).

Lambda exonuclease treatment

After addition of 2 µl T4 polynucleotide kinase (PNK) 10X buffer (NEB), fractions were boiled for 5 min and chilled on ice. Phosphorylation with T4 PNK was performed in 1X PNK buffer containing 0.2 U/µl PNK in a volume of 100 µl at 37°C for 1 hour. After heat inactivation at 75°C for 15 min, DNA was precipitated with 2.5 volumes ethanol-0.3 M sodium acetate (Na-acetate) at -80°C for 15 min. Pellets were washed with 1 ml 70% ethanol and suspended in 50 µl water with 100 U RNasin. Digestion with 100 U lambda exonuclease was in exo buffer (67 mM Glycine-KOH pH 9.4, 2.5 mM MgCl2, 50 µg/ml BSA) in 100 µl total volume at 37°C overnight. We found that the quality of the lambda exonuclease is crucial, and deserves to be always tested before use. For the experiments described here, we used a custom-made preparation by Fermentas (20U/µl). NS were extracted once with phenol/chloroform/isoamylalcohol and once with chloroform/isoamylalcohol, then precipitated with 2.5 volumes ethanol-0.3M Na-acetate at -80°C for 15 min. Pellets were washed with 1 ml 70% ethanol and suspended in 20 µl water. NS were subjected to one or two further cycles of T4 PNK phosphorylation and exonuclease digestion. We observed that the second round of exonuclease treatment significantly improves the NS preparation. Aliquots of digested and undigested DNA were run on a 2% agarose gel to confirm the efficiency of the exonuclease treatment. Finally, NS were purified using the CyScribe GFX Purification Kit (GE Healthcare) and eluted in 60 µl water.

NS amplification and Chip data analysis

10 µl of purified NS were amplified using the WGAII kit (Sigma), omitting the first step of fragmentation. Amplification products were purified with NucleoSpin columns (Machrey Nagel). Proper unbiased amplification was monitored by qRT-PCR. Hybridization, washing and scanning of microarrays were done by Nimblegen Service Laboratory. Details about the Nimblegen microarrays used and the data analysis are available as Supplemental data. Tiling array data are available on Gene Expression Omnibus (GEO) series GSE29183.

Quantitative Real-Time PCR analysis

qRT-PCR analysis of NS samples was performed using the SYBR Green PCR master mix (Roche) in a Lightcycler 480 real-time PCR thermocycler (Roche). For relative quantification, dilutions of total genomic DNA were used to construct the standard curves. One µl NS or genomic DNA was used per reaction and all experiments were done in triplicate.

DNA Combing

The complete procedure is detailed in Supplemental Data. Briefly, asynchronous cell populations were labeled with 40 mM IdU for 20 min followed by a second 20-min pulse with 40 mM CldU. After staining of proteinase K-treated DNA plugs with YOYO-1 (Molecular Probes) and digestion with agarase (New England Biolabs), DNA fibers were combed on silanized cover slips (Michalet et al. 1997). Immunodetection was done with mouse anti-BrdU (Becton Dickinson) and rat anti-BrdU (Sera Lab) antibodies and DNA was stained with the anti-ssDNA antibody (Chemicon). Image acquisition was performed with a fully motorized Leica DM6000B microscope equipped with a CoolSNAP HQ CDD camera and controlled by MetaMorph (Roper Scientific). Inter-origin distances and fork speed were measured manually using the MetaMorph software.

Motif search

Enriched motifs in Oris were identified using the MEME bioinformatics suite (http://meme.sdsc.edu/meme4_4_0/cgi-bin/meme.cgi). The settings were: zero or one occurrence, motif length between 6 (minimum) and 50 pb (maximum). A 5th order Markov model was generated as a background distribution model to take into account repetitive sequences. From 5% of all Oris, 2 kb of DNA sequences centered on the NS peak were randomly selected. Independent analyses were performed which showed that the results were not dependent on the Oris sample. As an additional negative control, randomly selected genomic sequences were also analyzed. For each motif, an E-value was computed. E-values are commonly used for assigning significance to the optimal reported motifs. When the E-value is high, the confidence in the motif prediction is low, whereas low E-values are significant. Genomic frequencies of motifs were generated with the help of the FIMO server (http://meme.sdsc.edu/meme4_4_0/cgi-bin/fimo.cgi). Occurrences having p-value $p<1X10^{-5}$ were used in this study. Over-represented motifs were searched with R'MES (Hoebeke and Schbath, 2006). The motif length was set to the maximum (13 nucleotides); as background distribution, we used for *Drosophila* Oris a Markov model of the

6th order, in which the expected number of occurrences of each motif was estimated using the compound Poisson distribution.

Acknowledgments

We would like to thank E. Andermarcher for critical reading of this manuscript, Benjamin Leblanc and Aubin Thomas for their proficient help with bioinformatic analyses, Giacomo Cavalli for providing *Drosophila* Kc cells as well as our laboratory members for helpful suggestions. We acknowledge the Montpellier DNA Combing Facility for providing silanized surfaces. CC was supported by the 'Fondation pour la Recherche Médicale' (FRM) and PC was supported by the FRM and the 'Association pour la Recherche contre le Cancer' (ARC). The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 233339". This work was also supported by the 'Agence Nationale de la Recherche' (ANR) (ANR-08-BLAN-0092-0), the ARC and the 'Ligue Nationale Contre le Cancer' (LNCC).

REFERENCES

LEGENDS TO FIGURES

Figure 1: Genome-scale mapping of replication origins by Nascent Strand (NS) Chip

(A) NS isolation schematic. 0.5-2.5 kb NS were isolated from total genomic DNA by denaturation and sucrose gradient centrifugation. NS enriched by lambda-exonuclease treatment were hybridized against total genomic DNA on high-density tiling arrays (see Supplementary information). **(B)** Example of the distribution of replication origins in mouse (upper panel) and *Drosophila* cells (lower panel) along a 200 kb region. The $\log_2$-ratio between NS and total genomic DNA is shown. For genes, the position of the start site (high bar bordering the gene), exons (large blue boxes) and introns (thin blue boxes) is indicated. **(C)** Origin number and density per genome. (**D**) Immunoprecipitation of chromatin associated with ORC2 was carried out in P19 cells as described in Methods. Compilation of ORC2 signal strength data and correlation with the NS peaks is shown.

Figure 2: Replication origins in metazoans are linked to expressed genes

(A) Replication origins are significantly associated with transcribed genes (*: p<0.001) in both mouse MEF and *Drosophila* Kc cells. **(B)** Association of replication origins with gene partitions in MEFs (left panel) and *Drosophila* Kc cells (right panel). Replication origins are found more frequently at gene promoters (mouse cells) and exonic sequences (mouse and *Drosophila* cells, *: p<0.001). **(C)** Distribution of mouse replication origins along a gene. The position of each origin is allocated depending on the length of the gene adjusted to 100%. **(D)** Nascent strand signal strength at TSS in ES cells and (**E**) *Drosophila* Kc cells. The enrichment value is the $\log_{10}$ of the combined p-value associated with NS signal (see Supplementary information. **(F)** NS signals in mouse ES cells are associated with CGI-positive TSS but not with CGI-negative TSS.

Figure 3: Association of replication origins with CGI in metazoans

Example of replication origins associated with CGI in **(A)** mouse ES cells and **(C)** *Drosophila* cells. The percentage of CGI/replication origin association is also shown. **(B)** NS signal strength around all CGIs in mouse ES cells and **(D)** CGI-like regions in *Drosophila* Kc cells. The average size of CGI is shown in scale. **(E)** Common origins in mouse cells are strongly associated with CGI regions. The proportion of CGI-positive origins in the indicated groups of origins is shown.

Figure 4: Nucleotide skew and GC-rich elements at replication origins

(A) Origins were centered on *Drosophila* CGI-like regions. The mean AT and GC percentages of centered Oris are shown. Genome-scale NS signal strengths are represented by a black line. Note that the NS peaks (putative replication initiation sites) are not enclosed in the central CG-rich region. **(B)** Genome-scale nucleotide distribution of all *Drosophila* origins centered on the NS peak. Note the skew in nucleotide distribution with GT and AC enrichment at the 5' and 3' end of the origin peak, respectively. **(C)** Nucleotide distribution at and around the origin peak for origins in *Drosophila* Kc cells. 200 bp sequences of 300 replication origins were stacked and aligned around the NS peak. Four colors were used: green for A, red for T, yellow for G and blue for C. The exact sequence can be read by enlarging the Figure in Supplemental Data. A clear bias is observed for C or G, and A or T around the NS peak. **(D)** Motifs frequently found in *Drosophila* (top panel) and mouse (bottom panel) replication origins. The E-value is indicated (see Methods).

Figure 5: Positive link between HP1 and origin firing/early replication in heterochromatic regions
(A) Density of CGI-like regions in the whole genome and on chromosome 4 in *Drosophila*. **(B)** Positive correlation between HP1 binding and early S-phase replication timing in *Drosophila* chromosome 4. Scatter plots between experimental and randomized HP1 data sets and replication timing are shown. The p-value is indicated at the bottom of the panels. **(C)** Significant association between HP1 binding and origins on the entire *Drosophila* chromosome 4. **(D)** A 300 kb region of chromosome 4 showing the relationship between origin firing, HP1 binding and early replication timing.

Figure 6: Early replication domains are characterized by high origin density
(A) Shown is the origin density in the three mouse cell lines calculated using a 100 kb sliding window along the chromosomal region. The computed gene and CGI densities are also illustrated. Origin density is positively correlated with early replication domains (*: p<0.001). **(B)** Origin number is also positively correlated (*: p<0.0001) with the early replication timing observed in mouse ES cells (Hiratani et al. 2008).

Figure 7: Replication origins are organized in a functional hierarchical manner along the chromosome
(**A)** DNA combing analysis performed in *Drosophila* Kc (top panel) and mouse (bottom panel) cells after two consecutive labeling pulses of IdU and CldU. **(B)** Summary of the experimental and simulated inter-origin distance distributions for MEF cells. For the 'Increasing Ori efficiency' model, the values for the firing efficiency represent the initial and final origin firing efficiency during simulations. **(C)** 'Random Ori firing' model. In this model, origins are completely independent and are activated randomly (red circles). Very short and long inter-origin distances are observed. **(D)** In the 'Increasing Ori efficiency' model, origins are completely independent and activated randomly, but with increasing firing efficiency throughout S phase progression. **(E)** 'Flexible Replicon' model. In this model, origins are linked within functional units where activation of one origin silences the others in the same group.
The bottom panels present the computer-simulated results for each model. The grey profile is the distribution of inter-origin distances obtained by DNA combing of MEF cells. The red line represents the simulated distribution of inter-origin distances according to each model. The "Flexible Replicon" model is the only to yield a simulated distribution of inter-origin distances that is statistically indistinguishable from the DNA combing data.

Figure 8: Origins, replicons and replicon clusters
(A) The presence of a CpG island or CGI-like region allows the positioning of two potential initiation sites upstream and downstream of the region. **(B)** Replicons are organized as functional units containing several potential DNA replication origins. Activation of one origin within a replicon silences the others. The origin choice within each replicon can occur either stochastically or be dictated by specific cell fates. Replicon clusters include several consecutive replicons which are activated simultaneously (Berezney et al. 2000). **(C)** Representation of replicons as chromatin loops where activation of one origin silences the other origins contained in the same replicon.

SUPPLEMENTARY FIGURES

Supplementary Figure 1: Biological replicates of nascent strands isolated from mouse and *Drosophila* cells and confirmation of the microarray data by qPCR

(A) Box plots showing that mouse chromosome 11 is comparable to the entire mouse genome concerning replication timing (left panel) and transcription activity (right panel). **(B)** Alignment of four entirely independent biological replicates of microarray data for P19 cells, and **(C)** representative Scatter Plots with computation of the Pearson correlation (R^2) of two biological replicates. **(D)** Alignment of two entirely independent biological replicates of microarray data for *Drosophila* Kc cells and **(E)** representative Scatter Plots of two biological replicates. **(F)** Nascent strand (NS) preparations from mouse cells were validated using a known origin by qPCR with different sets of primers localized along the *Myc* gene. **(G)** Microarray data for mouse cells were confirmed at the *Hoxb* locus by qPCR with different sets of primers localized along the locus.

Supplementary Figure 2: Confirmation of microarray data by qPCR

(A) qPCR confirmation of the *Histone* gene repeat origins in *Drosophila* Kc cells. Different sets of primers localized along various loci of the *Drosophila* (**B**) or mouse (**C**) genome were used for qPCR measurements of nascent strand enrichment.

Supplementary Figure 3: Association of replication origins with ORC

Immunoprecipitation of chromatin associated with ORC2 was carried out in P19 cells as described in Methods. **(A)** DNA fragments were analyzed by microarrays and validated by qPCR at the *Myc* gene. **(B)** Alignment of origins and ORC2 sites on a representative region in P19 cells.

Supplementary Figure 4: Common origins in the three mouse cell lines

(A) The percentage of replication origins' overlap in the different mouse cells is shown. **(B)** Common origins in the three mouse cell lines. Origins conserved between two mouse cell lines where compared to each other. The proportion of conserved origins between ES and P19 cells was significantly higher than between ES and MEF or MEF and P19 cells.

Supplementary Figure 5: Distribution of origins along genes

(A) Intragenic or intergenic distribution of origins. **(B)** Origins are enriched at gene promoters and exon sequences (*=p<0.001) compared to randomized data sets (dashed white boxes) in P19 (left panel) and MEF cells (right panel).

Supplementary Figure 6: Origins in MEF and P19 cells are frequently associated with TSS and CGI

Patterns of NS strength at TSS in P19 **(A)** and MEF cells **(B). (C)** Association of origins with TSS which contain or not CGI in P19 **(C)** and MEF **(D)** cells. **(E)** Origins found by microarrays are highly associated with CGI in MEF **(E)** and P19 cells **(F).** The percentage of the CGI-origin association is also shown.

Supplementary Figure 7: Analysis of bimodal origins located at TSS in mouse cells.

The NS profile of individual TSS associated with an origin was examined. Examples of NS profiles for each class of TSS are shown. TSS were scored as bimodal if the $\log_2$-ratio increased both upstream and downstream of the TSS.

Supplementary Figure 8: Nucleotide asymmetry of origins in mouse cells
Nucleotide composition along a 3 kb region (A-green, T-red, G-black and C-blue) centered on the origin peaks in mouse ES **(A)**, P19 **(B)** and MEF **(C)** cells. In these cells lines, an asymmetric distribution of G/T versus A/C is observed, like in *Drosophila* Kc cells.

Supplementary Figure 9: Hierarchical organization of origins in *Drosophila* Kc cells and mouse ES cells
The grey profile is the distribution of inter-origin distances obtained by DNA combing of ES **(A)** and Kc **(B)** cells. The red line represents the simulated distribution of inter-origin distances according to each model. The 'Flexible Replicon' model is the only to yield a simulated distribution of inter-origin distances that is statistically indistinguishable from that obtained from DNA combing data for these cell lines.

Supplementary Figure 10: Characterization of the models of origin organization in metazoans
(A) Distribution of fork speed (measured by DNA combing) in mouse and *Drosophila* cell lines (see Methods). **(B)** Simulated origin firing efficiency in the ‘Increasing origin efficiency’ model for ES cells. Note the increase in firing efficiency as replication takes place. Similar profiles were obtained for MEF and Kc cells. **(C)** Dendrogram illustrating how origin clusters (replicons) were defined in the 'Flexible Replicon' model. Origins were grouped based on their closeness along the chromosome. Clusters were defined by cutting the tree at a specific height. Shown are three cluster generations based on different height (h) cuts. Clustered origins are highlighted. **(D-E)** Clusters were exhaustively generated by cutting the tree every 1000 steps. For every cluster generation, origin firing was performed (100 simulations). The simulated inter-origin distribution was compared with the DNA combing data and a p-value was calculated with the Kolmogorov-Smirnov test. The p-values were plotted in function of the cutting height. A cubic smoothing spline function was applied to the data (grey curve). The significance value ($p=0.05$) is indicated. The minimal (h_{min}), optimal (h_{opt}) (where the simulated inter-origin distribution is not statistically different from the DNA combing data) and maximal (h_{max}) cutting heights are highlighted. **(F)** Statistics on the clusters generated in the Flexible Replicon model. The average number of origins/cluster, length of clusters and the inter-cluster distance are indicated for ES, MEF and Kc cells for the optimal cutting height (h_{opt}). The values in brackets are for the h_{min} and h_{max} clusters.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Description of the genomic features

Gene databases were Flybase (for *Drosophila*) and RefSeq (for *Mus Musculus*). CGI were defined as a DNA region of at least 200 pb with a GC content greater than 60% and the (observed CpG/expected CpG) ratio equal to or greater than 0.6 (classical definition of a CpG island, (Gardiner-Garden and Frommer 1987). *Drosophila* HP1 binding sites were determined from DamID data (de Wit et al. 2007). Replication timing data for Kc and ES cells were from (Schwaiger et al. 2009) and (Hiratani et al. 2008) respectively. Divergent transcription start sites (TSS) used for ES cells were described in (Sailo et al. 2008).

Nascent Strand-ChIP Data Analyses

Microarray Design

Drosophila melanogaster samples were hybridized using 2.1M Nimblegen microarrays (Design ID 6262). These tiling arrays contain in total 2,164,511 oligonucleotide probes representing the non-repetitive regions of the *Drosophila melanogaster* genome (chromosome 2L, 2R, 3L, 3R, 4 and X; Flybase release 4.3).

To analyze the data, 1,807,015 oligonucleotide probes were selected (909,279 for the top strand and 897,736 for the bottom strand) with an average length of 50 bp for oligonucleotides and for inter-oligo spacing. All the processed data were generated using the BDGP/Flybase release 4 of the *Drosophila melanogaster* genome assembly (UCSC dm2, April 2004).

Mouse samples were hybridized using the Nimblegen 389K tiling arrays (Design ID 4095) which cover 60.4 Mb of non-repetitive DNA sequences in chromosome 11 (56.6-117 Mb). In total, 385,496 probes were analyzed with an average coverage of one 50 bp-probe each 100 bp. All the processed data were generated using the UCSC mm8 (NCBI Build 36, February 2006) of the *Mus musculus* genome assembly.

Correlation between biological replicates

The degree of correlation between biological replicates was evaluated using a scatter plot and computing the Pearson's Correlation Coefficient (R^2).

Data normalization and determination of significant probes

Experimental (Cy5) and control (Cy3) signal intensities quantified and provided by Nimblegen were converted into $\log_2$-ratios ($\log_2$ (Cy5/Cy3)). The Lowess normalization method was applied to eliminate intensity-dependent variations in dye bias (Yang et al. 2002). A sliding median window with a length of 5 oligonucleotide probes was used to smooth the signal. Mode (m) and s (median absolute deviation) of normalized $\log_2$-ratios were computed. Assuming that the normal distribution (specified by m and s) covered the entire background noise (non-significant signals), for each probe, one p-value was computed by applying the false discovery rate (FDR) correction (Benjamin and Hochberg 1995). Two biological independent samples for *Drosophila*, four independent samples for P19 cells, three independent samples from ES and three for MEF cells were used. Normalized $\log_2$-ratios of replicate samples were combined by averaging the values at the corresponding genomic positions and the corrected p-values were combined using a Chi-Square distribution (Fisher 1932). Thus, one probe was denoted as significant if the combined p-value was lower than 5% (level of significance).

Origin definition

The minimum size of purified NS is 0.5 kb. Thus, potential Oris should be at least 1 kb (2 X 0.5 kb for a bidirectional origin). We defined Oris as regions that have at least one significant probe ($p<0.05$ with FDR correction) in an area containing a minimum of 10 consecutive positive probes (showing NS enrichment with a $\log_2$-ratio>0). For *Drosophila* cells, two significant probes (because they are twice denser in *Drosophila* than in mouse Chips) and at least ten consecutive positive probes should be detected. If two enriched regions were separated by <1 kb, they were merged into one. These conditions were used to minimize false positive events by excluding over-hybridization signals of single probes or small regions, and to score as Oris only sites with multiple consecutive positive values.

Comparative analysis of Oris and genome features

For each profile (*Drosophila* and mouse cells), 1000 bootstrap samples of random Oris were generated. Random Oris contained the same number of Oris with the same length, but each origin segment was randomly picked in the chromosome region with the condition that the segments did not overlap. For each profile and each studied genome feature (CGI or CGI-like, TSS, etc …), one permutation test with theoretical expectation under a null hypothesis was performed from the 1000 random Oris to compute the statistical significance of the Ori positions relative to the studied genome feature.

NS signal strength around specific features

For each profile (*Drosophila* and mouse cells), specific feature positions (TSS, middle of CGI or CGI-like) were taken as 'Local center' (Lcent). For each nucleotide position around every Lcent (Lcent - 5 kb to Lcent + 5 kb), p-values (previously calculated) were retrieved. P-values were merged in a matrix (rows representing the nucleotide coordinate/position and columns representing Lcent). The strand was also considered. Thus, for TSS from the minus strand, nucleotide positions and associated p-values were reversed. To obtain only one overall p-value distribution around the set of Lcent, p-values were combined using a Chi-Square distribution (Fisher 1932). To visualize the combined p-value distribution around the specific features, results were plotted using the transformation '-log(p-value)' and labeled as 'NS signal strength'.

Analysis of Bimodal TSS in mouse cells

For both upstream and downstream regions (TSS -2 kb to TSS + 2 kb) of each TSS overlapping one Ori, the highest '-log(p-values)' (noted '$-\log(\text{p-value})_{\text{upstream}}$ and '$-\log(\text{p-value})_{\text{downstream}}$') were retrieved. In the same way, the lowest '-log(p-value)' (noted '$-\log(\text{p-value})_{\text{middle}}$') around each TSS (TSS -0.1 kb to TSS + 0.1 kb), was collected. Note that, high '-log(p-values)' corresponds to high $\log_2$-ratio of NS/total genomic DNA. In this analysis, the orientation of TSS was considered.

Four classes were created:

- If the '-log(p-value)' increased both upstream and downstream of the feature, the TSS was scored as bimodal.
  More precisely, this category corresponds to TSS in which:
  $-\log(\text{p-value})_{\text{upstream}} > -\log(\text{p-value})_{\text{middle}}$ and $-\log(\text{p-value})_{\text{downstream}} > -\log(\text{p-value})_{\text{middle}}$

- If the '-log(p-value)' increased only upstream of the feature, the TSS was scored as unimodal with NS enrichment at the 5' of the feature. More precisely, this category corresponds to TSS in which:
  $-\log(\text{p-value})_{\text{upstream}} > -\log(\text{p-value})_{\text{middle}}$ and $-\log(\text{p-value})_{\text{downstream}} \leq -\log(\text{p-value})_{\text{middle}}$

- If the '-log(p-value)' increased only downstream of the feature, the TSS was scored as unimodal TSS with NS enrichment at the 3' of the feature. More precisely, this category corresponds to TSS in which:
  $-\log(\text{p-value})_{\text{upstream}} \leq -\log(\text{p-value})_{\text{middle}}$ and $-\log(\text{p-value})_{\text{downstream}} > -\log(\text{p-value})_{\text{middle}}$

- Otherwise, TSS was associated with one Ori exhibiting a more symmetrical NS profile around the TSS.

Sequence distribution around specific regions

For each profile (*Drosophila* and mouse cells), the sequence distribution was centered on the middle of the CGI or CGI-like regions associated with Oris and taken as the 'Local center' (Lcent). The 3-kb sequence around each Lcent (Lcent – 1.5 kb to Lcent + 1.5 kb) was retrieved. The resulting sequences were merged in a matrix (rows representing nucleotide coordinate/position and columns representing Lcent). The number and the percentage of A/T/C/G nucleotides were computed. Results were plotted using a sliding mean window to fit the signal.

The same analysis was performed to represent the sequence distribution centered on the probe with maximum intensity for Oris.

Organization of Oris

Computer simulations were performed to model Ori organization. For each model (Random, Increasing Firing Efficiency and Flexible Replicon), inter-origin distances from 100 simulations were calculated. Importantly, the firing density (e. g. the number of activated Oris/Mb) was identical to the density observed in DNA combing experiments. The simulated inter-origin distribution was compared with the inter-origin distribution of DNA combing data by calculating the p-value with the Kolmogorov-Smirnov test (Massey 1951). A high p-value (p>0.05) indicates that the two distributions cannot be considered as statistically different. The different models were evaluated as follow.

1) *Random Ori firing model*

In the random model, Oris are fired in a purely stochastic manner. In this model, firing efficiency is supposed to be constant. The inter-origin distances for each of the 100 simulations were calculated.

2) Increasing Ori efficiency model

This model is based on the hypothesis that Ori firing efficiency increases during S-phase progression (Rhind 2006). Also, the advancing replication fork passively suppresses replicated Ori regions. During each cycle (simulation of time), one Ori is randomly selected. The resulting bidirectional replication fork was simulated using the mean fork speed obtained from DNA combing experiments. The duration of each cycle was optimized to achieve a firing efficiency identical to the one of the single DNA molecule experiments. The model stops when the entire DNA is replicated. Inter-origin distances between fired origins from 100 simulations were collected. The simulated Ori firing efficiency was also calculated.

3) Flexible Replicon Model

This hierarchical clustering model is based on the hypothesis that Oris are functionally grouped and that activation of one Ori suppresses the firing of other Oris within the same group. In this model, Ori firing is randomly selected and the firing efficiency is supposed constant. The steps to obtain groups of Oris, called clusters, are described below.

First, Oris were classified using hierarchical cluster analysis with Euclidean distance as the distance metric to determine how the similarity of two elements was calculated, and average linkage clustering to determine the distance between sets of observations (Brian et al. 2001). Then, by cutting the dendrogram at different heights different clusters were defined. For each height cut, 100 simulations were collected and the distribution of inter-origin distances was compared with DNA combing data. The range of selected height cuts corresponded to heights where the p-values were the highest (p>0.05).

Precisely, the retained clusters of Oris were obtained by cutting the dendrogram at the optimum height of 26,560 bp for *Drosophila,* 66,374 bp for mouse ES cells and 71,032 bp for mouse MEF cells. For each profile, cluster characteristics (inter-cluster distance, cluster length, etc.) were calculated at the optimum height.

Density of origins and other genome features

Density analysis was used to compare specific data distribution along the genome at large scale. The coordinates of the specific regions and the genome positions were retrieved. Each nucleotide inside specific regions was flagged as 1 (if belonging to one Ori) and 0 (if not belonging to one Ori). A sliding window was used to compute the frequency of data per window.

For each profile (*Drosophila* and mouse cells), the window size was based on the optimal height cut from the hierarchical cluster of Oris.

Conserved regions

To test whether Oris were in conserved regions, conservation scores were downloaded for alignments of 14 insect genomes with the *Drosophila melanogaster* genome and 16 vertebrate genomes with the *Mus Musculus* genome from the UCSC Website. Conservation data were divided in two groups called "inside origins" and "outside origins". The Wilcoxon-Mann-Whitney (Mann and Whitney 1947) test was used to determine whether the conservation scores between the two groups were significantly different.

Comparison of the Ori repertoires in mouse cells

Considering as reference the Oris from P19 cells, two proportions of “common Oris” were calculated: “common Oris” between P19 and ES cells and “common Oris” between P19 and MEF cells. Then, the difference between these two proportions (Newcombe 1998) was tested by computation of the p-value (p). A p-value <0.05 indicates that the two proportions are significantly different. The same analysis was carried out considering as reference the Oris from ES and then from MEF cells as well.

Comparison of Ori coverage in early and late replication timing regions

For each profile of mouse cells, Ori coverage in early and late replication timing domains was calculated. To compare the coverage values, a test of difference between the two groups was performed (Newcombe 1998).

Software

All Nascent Strand-ChIP data analyses were carried out using the software R, version 2.11.1 (www.R-project.org) (R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0).

Cell Cycle Analysis

For cell cycle analysis a Beckman Coulter flow cytometer was used. Cells were fixed with 70% ethanol in PBS at -20C° for at least 20 min. After one wash in PBS, cells were incubated in propidium iodide (PI, Sigma) at room temperature for 30 min before treatment with DNase-free RNase A (Sigma).

DNA Combing

Cells were sequentially labeled with iodo-deoxyuridine (IdU) and chloro-deoxyuridine (CldU). Asynchronous cell populations were first labeled with 40 mM IdU for 20 minutes and then with 40 mM CldU for another 20 minutes, without intermediate wash. Cells were then washed with 1x phosphate-buffered saline (PBS), trypsinized, pooled, counted and 100 000 cells were resuspended in 100 ml of 1x PBS with 1% low-melting agarose in order to make agarose plugs with imbedded cells. Plugs were incubated in 0.5 ml 0.5 M EDTA with 1% N-lauryl-sarcosyl and 1 mg/ml proteinase K and incubated at 50°C for 2 days (fresh solution added after the first day). Complete removal of digested proteins and other degradation products was performed by washing the plugs in 0.5M EDTA and TE buffer several times. Protein-free DNA plugs were then stored in 0.5 M EDTA at 4°C or used immediately for combing. Agarose plugs were stained with YOYO-1 fluorescent dye (Molecular Probes) in TE buffer for 2 h, washed with TE buffer, resuspended in 100 μl of TE buffer and melted at 65°C for 15 minutes. The solution was maintained at 42°C for 15 minutes and treated overnight with agarase (New England Biolabs). After digestion, 4 ml of 50 mM MES (2-(*N*-morpholino)ethanesulfonic acid, pH 5.7) were added very gently to the DNA solution and then combing of DNA fibers on silanized cover slips was performed as described (Michalet et al. 1997). Combed DNA was denatured in 1N NaOH for 20 minutes and washed several times in PBS. After denaturing, silanized cover slips with DNA were blocked with 1% BSA in PBS, 0.1% Triton X100. Immunodetection was done with antibodies diluted in PBS, 0.1% TritonX100, 1% BSA and incubated at 37°C in a humid chamber for 30 min. Each step of incubation with antibodies was followed by extensive washes with PBS. Immunodetection was with a mouse anti-BrdU antibody (1/50 dilution, Becton Dickinson) and a rat anti-BrdU antibody (1/25 dilution, Sera Lab) that recognize the IdU and CldU tracks, respectively, goat anti-rat antibody coupled to Alexa 488 (1/50 dilution, Molecular Probes), goat anti-mouse IgG1 coupled to Alexa 546 (1/50 dilution, Molecular Probes), anti-ssDNA antibody (1/100 dilution, Chemicon) and goat anti-mouse IgG2a coupled to Alexa 647 (1/50 dilution, Molecular Probes). Cover slips were mounted with 20 μl of Prolong Gold Antifade (Molecular Probes), dried at room temperature for 2 hr and processed for image acquisition using a fully motorized Leica DM6000B microscope equipped with a CoolSNAP HQ CDD camera and controlled by MetaMorph (Roper Scientific). Images were acquired with a 40x objective: 1 pixel was equal to 340 bp. Inter-origin distances were measured manually using MetaMorph. Statistical analysis of inter-origin distances was performed with Prism 5.0 (GraphPad).

ORC2 ChIP on Chip and qPCR analysis.

Briefly, approximately 1.5 × 10^8 P19 cells were treated with 100 ng/ul nocodazole for three hours and seeded after shaking off, followed by three washes with PBS. After 30 min (cells in G2\M phase by flow cytometry analysis), cells were cross-linked by adding fresh 0.5% paraformaldehyde solution to the medium at 37°C for 15 minutes. Paraformaldehyde was neutralized by adding 250 mM glycine at room temperature for 10 min. Cells were washed twice with 1× phosphate-buffered saline (PBS), scraped off the plates, and nuclei were isolated with NE buffer (50 mM HEPES at pH 7.6, 350 mM sucrose, 0.1% Tween20, 5 mM MgCl2, 1 mM EDTA and protease inhibitors). After centrifugation, nuclei were lysed in 1 ml RIPA buffer (50 mM HEPES at pH 7.6, 150 mM NaCl, 0.3% SDS, 0.5% NaDoc, 1% TritonX100 and protease inhibitors) and sonicated into fragments ranging from 300 to 1000 bp using the Bioruptor (Diagenode). The chromatin solution was clarified by centrifugation at 15 000g at 4°C for 5 min. The supernatant was pre-cleared with 50 µl of Dynabeads protein A for 2 h at 4°C. Pre-cleared chromatin was separated in two fractions and incubated at 4°C overnight with 50 µl of Dynabeads protein A, blocked with 0.05% bovine serum albumin/PBS and pre-incubated with 30 µg of ORC2 antibody (home-made with recombinant mouse ORC2) or with 30 µg of pre-immune antibody (from the same rabbit used for generating the ORC2 antibody, but before injection) for 2 h. After extensive washing with RIPA buffer, cross-linking of each immune complex was reversed by incubation of the eluate at 65 °C in 50 mM Tris pH 8, 1% SDS, 10 mM EDTA overnight. After digestion with RNaseA at 37°C for 1 h and proteinase K at 50°C for 2 h, DNA was purified by phenol–chloroform extraction and precipitated with ethanol. The amount of DNA in the immunoprecipitates and in the input was quantified by real-time PCR with primers localized along the *Myc* gene and promoter. ChIP data are reported as the percentage of the total input that was immunoprecipitated. Quantitative PCR was performed on a Roche LightCycler 480 machine using LightCycler® 480 SYBR Green I Master (Roche). DNA from immunoprecipitates was amplified using the WGAII kit (Sigma). Amplification products were purified with NucleoSpin columns (Machrey Nagel). Hybridization, washing and scanning of microarrays were done by the Nimblegen Service Laboratory. For this experiment, the Nimblegen 389K tiling arrays (Design ID 4095) were used. ChIP on chip signals were analyzed in the same manner as the data from hybridization with nascent strands (see above).

Supplementary References

Benjamin Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSSB* 57: 289-300.

Brian S, Everitt, Landau S, Leese M. 2001. Cluster Analysis 4th Edition. *Oxford University Press*.

de Wit E, Greil F, van Steensel B. 2007. High-resolution mapping reveals links of HP1 with active and inactive chromatin components. *PLoS Genet* 3(3): e38.

Fisher RA. 1932. Statistical Methods for Research Workers. *Oliver and Boyd* London.

Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. *J Mol Biol* 196(2): 261-282.

Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, Chang CW, Lyou Y, Townes TM, Schubeler D, Gilbert DM. 2008. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol* 6(10): e245.

Mann HB, Whitney DR. 1947. On a test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics* 18(1): 50-60.

Massey FJ. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* 46(N° 253): 68-78.

Michalet X, Ekong R, Fougerousse F, Rousseaux S, Schurra C, Hornigold N, van Slegtenhorst M, Wolfe J, Povey S, Beckmann JS et al. 1997. Dynamic molecular combing: stretching the whole human genome for high-resolution studies. *Science* 277(5331): 1518-1523.

Newcombe RG. 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat Med* 17(8): 873-890.

Rhind N. 2006. DNA replication timing: random thoughts about origin firing. *Nat Cell Biol* 8(12): 1313-1316.

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(5): 589-601.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 30(4): e15.