6064–6075 *Nucleic Acids Research, 2009, Vol. 37, No. 18* *Published online 10 August 2009*
doi:10.1093/nar/gkp631

Open chromatin encoded in DNA sequence is the signature of 'master' replication origins in human cells

Benjamin Audit[1,2], Lamia Zaghloul[1,2], Cédric Vaillant[1,2], Guillaume Chevereau[1,2], Yves d'Aubenton-Carafa[3], Claude Thermes[3] and Alain Arneodo[1,2,*]

[1]Université de Lyon, F-69000 Lyon, [2]Laboratoire Joliot-Curie and Laboratoire de Physique, CNRS, Ecole Normale Supérieure de Lyon, F-69007 Lyon and [3]Centre de Génétique Moléculaire, CNRS, F-91198 Gif-sur-Yvette, France

Received June 16, 2009; Revised July 10, 2009; Accepted July 14, 2009

ABSTRACT

For years, progress in elucidating the mechanisms underlying replication initiation and its coupling to transcriptional activities and to local chromatin structure has been hampered by the small number (approximately 30) of well-established origins in the human genome and more generally in mammalian genomes. Recent *in silico* studies of compositional strand asymmetries revealed a high level of organization of human genes around 1000 putative replication origins. Here, by comparing with recently experimentally identified replication origins, we provide further support that these putative origins are active *in vivo*. We show that regions ~300-kb wide surrounding most of these putative replication origins that replicate early in the S phase are hypersensitive to DNase I cleavage, hypomethylated and present a significant enrichment in genomic energy barriers that impair nucleosome formation (nucleosome-free regions). This suggests that these putative replication origins are specified by an open chromatin structure favored by the DNA sequence. We discuss how this distinctive attribute makes these origins, further qualified as 'master' replication origins, priviledged loci for future research to decipher the human spatio-temporal replication program. Finally, we argue that these 'master' origins are likely to play a key role in genome dynamics during evolution and in pathological situations.

INTRODUCTION

In eukaryotes, transmission of genetic information requires the precise and complete duplication of genomic DNA. A number of experimental studies have shown that in metazoan, replication initiates from a large number of origins according to a temporal program modulated by the type of tissue and/or by the developmental stage (1–4). However, our knowledge of the mechanisms that control this program remains sparse and understanding how these origins are distributed along the genome and how their activation is controlled and coordinated constitutes one of the main challenges of molecular biology (5). For years, the small number (approximately 30) of well-established origins in the human genome and more generally in mammalian genomes has been an obstacle to fully appreciate the genome-wide organization of replication in relation to gene expression and local chromatin structure (6–8). Despite considerable experimental efforts to determine origin positions at the genome scale (9–11), much remains to be understood about the impact of the DNA sequence on origin activity in human cells in parallel to epigenetic controls (8,12–15). In that context, an *in silico* analysis of the strand composition asymmetry (skew) profile of the human genome allowed us to identify 1060 putative replication origins, likely conserved in mammalian genomes, that border 678 large (mean length of 1.2 Mb) genomic domains labeled N-domains as their skew profile is shaped like an N (Figure 1A) (16–19). These data correspond to the largest set of origin predictions for human genome available to date. While the origin of the skew itself remains debated (20), these domains were shown to be linked to

*To whom correspondence should be addressed. Tel: +33 4 7272 8757; Fax: +33 4 7272 8080; Email: alain.arneodo@ens-lyon.fr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

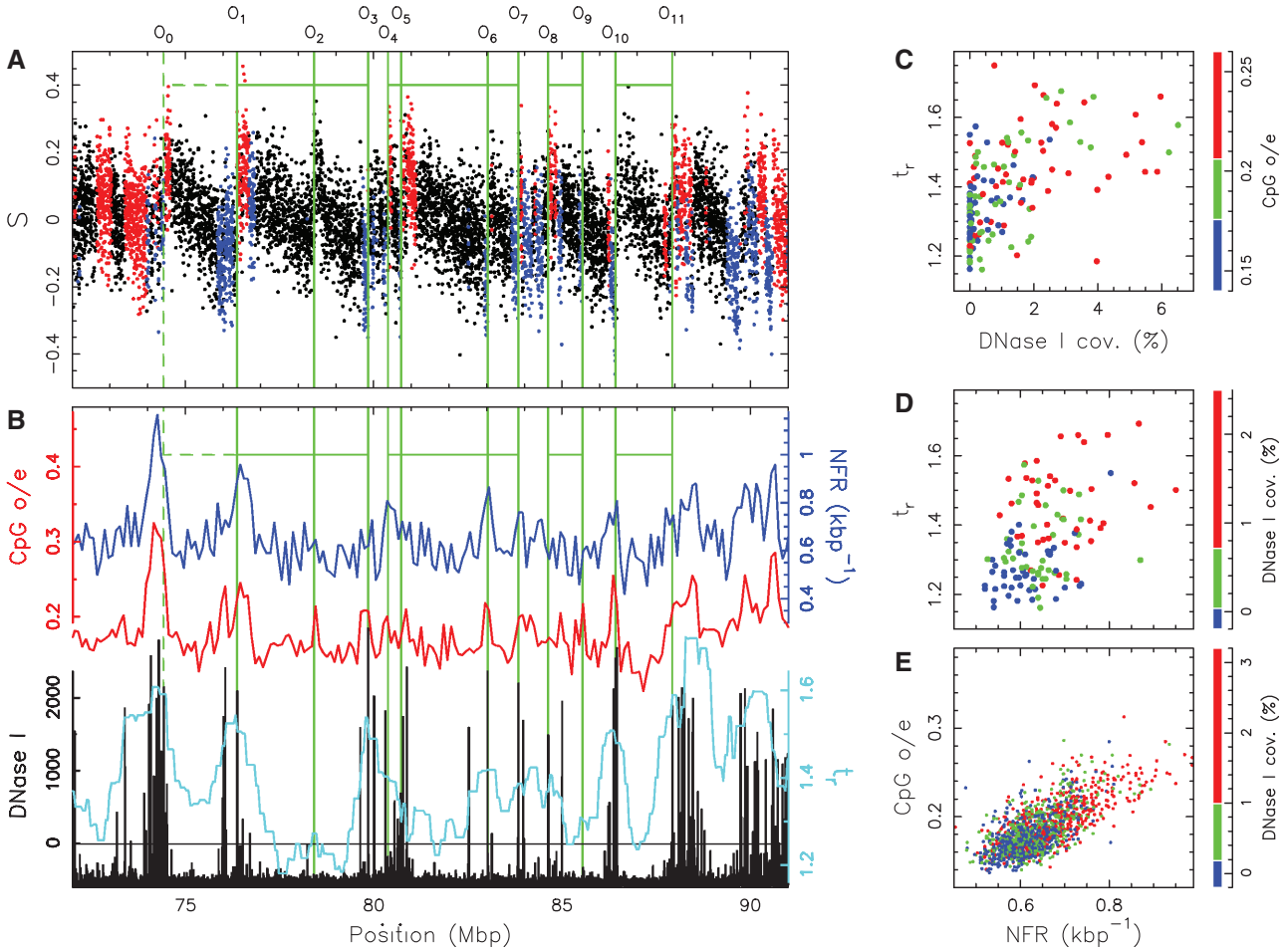

Figure 1. Open chromatin markers along N-domains. (**A**) Nucleotide compositional asymmetry profile S along a 19 Mb long fragment of human chromosome 6 that contains seven replication N-domains (horizontal green lines) bordered by 11 distinct putative replication origins O_1 to O_{11} (vertical green lines). Each dot corresponds to the compositional asymmetry [Equation (1)] calculated for a window of 1 kb of repeat-masked sequence (18). The black color corresponds to intergenic regions; red, sense (+) genes; blue, antisense (−) genes. (**B**) The different colored profiles correspond to the DNase I HS score (black, resolution 1 kb), the NFR density (blue, resolution 100 kb), the CpG o/e (red, resolution 100 kb) and the replication timing ratio t_r (light blue, inhomogeneous spatial resolution ~300 kb). The vertical dashed green line marks the location O_0 of the closest upward jump to the region where we observed the concomitant occurrence of a prominent burst in DNase I HS data, NFR numerical density and CpG o/e, in a region where the replication timing ratio is high ($t_r \sim 1.6$). (**C**) Correlation between replication timing ratio t_r and DNase I HS sites coverage (window size = 300 kb) along the 54 N-domains identified in chromosome 6 ($r = 0.56$, $P = 1.3 \times 10^{-14}$); dots are color coded according to CpG o/e value. (**D**) Correlation between replication timing ratio t_r and NFR density (window size = 300 kb, GC content ≤41%) along the 54 N-domains in chromosome 6 ($r = 0.41$, $P = 1.7 \times 10^{-6}$); dots are color coded according to DNase I HS sites coverage. (**E**) Correlation between CpG o/e and NFR density (window size = 300 kb, GC content ≤41%) along the 678 N-domains identified along the 22 human autosomes ($r = 0.71$, $P < 10^{-15}$); dots are color coded according to DNase I HS sites coverage. In (C, D and E) the three color code is provided as a lateral color bar; each color corresponds to one-third of the data points.

the organization of replication and transcription (16,17). Indeed, the comparison of recent high-resolution timing data for chromosome 6 (21) with N-domains provided a first experimental evidence for their relationship to the replication program (16,17). A majority of N-domains borders replicate earlier in the S phase than their surrounding regions and are thus likely associated with early replicating origins, while N-domain central regions are late replicating. Hence, most N-domains correspond to units of replication where timing decreases when going from borders to center (17). In higher eukaryotes, extensive connections have been established between replication timing, genome organization and gene transcriptional state; early replication tends to colocalize with active transcription and, in mammals, with gene-dense GC-rich isochores (21–27) and with transcription initiation early in development (28). Interestingly, the putative origins at N-domain borders were also shown to be at the heart of a remarkable gene organization (16); in a close neighborhood, genes are abundant and broadly expressed and their transcription is mainly directed away from the borders. This preferential orientation was interpreted in relation to replication fork directionality (16). All these features weaken progressively with the distance to domain borders. Altogether, these results suggest that N-domain borders are landmarks of the human genome organization and possible triggers of the replication program; they correspond to early replicating origins

separated by large distance (~1 Mb) around which replication and transcription are highly coordinated. Genome-wide investigation of chromatin architecture has revealed that, at large scales (from 100 kb to 1 Mb), regions enriched in open chromatin fibers correlate with regions of high gene density (29); whereas, at small scales (≲1 kb), DNA accessibility, nucleosome distribution and modifications are important determinant for transcriptional activity (30–34). Moreover, there is a growing body of evidence that transcription factors are regulators of origin activation [reviewed in (35)]. In this context, we ask to which extent the remarkable genome organization observed around N-domain borders is mediated by particular chromatin structure favorable to specification of early replication origins (16). In particular, the recent genome-wide mapping of DNase I hypersensitive (HS) sites (34) provides the unprecedented opportunity to study open chromatin in relation to the observed nucleosome-depleted regions (30–34) that look very similar to the nucleosome-free regions (NFRs) previously observed at yeast promoters (36,37). Here, we map experimental and numerical chromatin mark data in the 678 replication N-domains and we show that a significant subset of N-domain borders corresponds to particular open chromatin regions, permissive to transcription, which may have been imprinted in the DNA sequence during evolution. Our results suggest that the putative replication origins located at N-domain borders likely deserve a distinctive status as 'master' origins of the replication program.

MATERIALS AND METHODS

Sequence and annotation data

Sequence and annotation data were retrieved from the Genome Browsers of the University of California Santa Cruz (UCSC) (38). Analyses were performed using the human genome assembly of May 2004 (NCBI35 or hg17) except when specified otherwise. As human gene coordinates, we used the UCSC Known Genes table. When several genes presenting the same orientation overlapped, they were merged into one gene whose coordinates corresponded to the union of all the overlapping gene coordinates. This resulted in 19 543 distinct genes over the 22 human autosomes (where replication N-domain data were available; see below). We used CpG islands (CGIs) annotation provided in UCSC table 'cpgIslandExt'.

GC content

GC content was computed over the native sequence. We checked whether the results remained qualitatively similar when considering the GC content computed over the repeat-masked sequence or when masking CGIs.

CpG observed/expected ratio

CpG observed/expected ratio (CpG o/e) was computed as $\frac{n_{CpG}}{L-l} \times \frac{L^2}{n_C n_G}$, where n_C, n_G and n_{CpG} are the number of C, G and dinucleotides CG, respectively, counted along the sequence, L is the number of nonmasked nucleotides of the sequence and l is the number of masked nucleotide gaps plus one, i.e. $L-l$ is the number of dinucleotide sites. The CpG o/e was computed over the sequence after masking annotated CGIs. We checked that the results remained qualitatively similar while also masking the repeat sequences.

Replication domains and putative replication origins

The detection of human replication N-domains is based on the mammalian replication domain model that imposes an N-shaped profile for the nucleotide compositional strand asymmetry

$$S = \frac{G-C}{G+C} + \frac{T-A}{T+A} \quad \mathbf{1}$$

between two successive fixed replication origins (18,19). Using the wavelet transform as a multi-scale (the distance between origins is highly variable) shape detector, the human genome was segmented into candidate replication domains where the skew S (when calculated in nonoverlapping 1-kb windows) displays the characteristic N-shaped pattern (16). Note that this segmentation strategy is less efficient in GC-rich regions of the genome. Indeed, the smaller N-domain size and the high gene density in these regions make it difficult to distinguish replication-related from transcription-related strand asymmetry (16).

The coordinates of the 678 human replication N-domains were obtained directly from the authors (16). There are 1060 N-domain borders since in 296 cases, a border is shared by two consecutive domains. N-domain detection was performed for the 22 human autosomes, where they cover 30% of the sequenced genome length and 18% of the genes (3431 gene starts are in an N-domain).

Genome coordinates

We used the LiftOver coordinate conversion tool from the UCSC website to map N-domains determined on Human May 2004 (hg17) assembly to Human March 2006 (hg18) coordinates and we kept only the N-domains that had exactly the same size before and after conversion. This resulted in 663 unambiguous N-domain assignments on hg18. The analyses of genomic data available for the hg18 assembly were then performed using this hg18 N-domain database.

Correlation analysis

For the correlation analyses, we reported the Pearson's product moment correlation coefficient r and the associated P-value for no association ($r = 0$). In every case, we checked that Kendall tau rank correlation coefficient provided the same statistical diagnosis. All statistical computations were performed using the R software (http://www.r-project.org/).

DNase I HS site data

The 95 723 experimental DNase I HS site data (UCSC 'dukeDnaseCd4Sites' track) correspond to genome-wide DNase I HS sites as determined for human $CD4^+$ T cells using DNase sequencing and DNase chip (34). A total of 21 066 (22%) DNase HS sites overlap with a replication N-domain.

Replication timing data

We used the high-resolution timing ratio data obtained with a lymphoblastoid cell line using an array of overlapping tile path clones for human chromosome 6 (21). Data for clones completely included in another clone were removed after checking for timing ratio value consistency, leaving 1648 data points. The timing ratio value at each point was chosen as the median over the four closest data points to remove noisy fluctuations, so that the spatial resolution is ~300 kb.

Genome-wide nucleosome positioning data

We used the genome-wide map of nucleosome positioning in resting human $CD4^+$ T cells obtained from direct sequencing of nucleosome ends using the Solexa high-throughput sequencing technique (32). Nucleosome score profiles for human genome assembly hg18 were downloaded from http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellnucleosomes.html.

Chromatin fiber density data

We used open over input chromatin ratio data obtained by co-hybridization to a genomic microarray of open chromatin purified using sucrose gradient fractionation and of input chromatin from human lymphobastoid cells (29). Data were obtained directly from the authors.

Genome-wide maps of Pol II binding and tri-methylation of histone 3 lysine 4

We used Pol II binding and H3K4me3 data for human $CD4^+$ T cells obtained using direct sequencing analysis of ChIP DNA samples using Solexa 1G genome analyzer (ChIP-Seq) (33). Summary Bed files for hg18 assembly with tag counts in 400-bp windows were downloaded from http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.html.

RESULTS AND DISCUSSION

N-domain borders correspond to experimental replication origins mapped on ENCODE regions

Previous analyses of nucleotide strand compositional asymmetries have shown that, out of the nine experimentally identified replication origins, seven (78%) presented an upward jump in the asymmetry profile analog to those bordering N-domains (18,19). Recently, the localization of replication origins has been experimentally investigated along 1% of the human genome (ENCODE regions) by hybridization to Affymetrix ENCODE tiling arrays of purified small nascent DNA strands and of restriction fragments containing small replication bubble (9,10). Out of the seven N-domain borders that reside within an ENCODE region, four match with an experimental replication origin ($P = 4 \times 10^{-3}$): three to the bubble trapping dataset (Bubble-HL, $P = 0.017$) and one to a nascent strand purification dataset (NS-HL-2, $P = 0.09$) (Table 1). Actually, as previously noted (9), a second N-domain border is located within 1 kb of a NS-HL-2 origin (Table 1). Hence, there is direct experimental evidence that 5/7 (71%) N-domain borders correspond to active replication origins at a few kb resolution. These results are all the more significant considering rather low overlap between the experimental datasets. For example, only 69 (25%) of the NS-HL-2 origins overlap with a Bubble-HL origin, only 4 (1.4%) with the second nascent strand dataset in HeLa cells (NS-HL-1) and only 12 (4.3%) even when extending NS-HL-1 and NS-HL-2 origins by 1-kb on both sides.

Table 1. Correspondence between N-domain borders and experimental replication origins datasets along ENCODE regions a binomial test (*P*-values <0.02 are in bold)

Method	Number	Coverage (%)	Match with N-domain borders (*P*-value)	Reference
Bubble-HL	234	8.6	3 (**0.017**)	(10)
NS-GM	758	1.0	0	(10)
NS-HL-1	434	0.6	0	(10)
NS-HL-2	282	1.4	1 (0.093)	(9)
All		11	4 (**0.004**)	
NS + 1 kb-HL-2		3.2	2 (**0.019**)	
+ Bublle-HL		11.2	5 (**0.0003**)	

Main characteristics of replication origin prediction along ENCODE regions based on purified restriction fragments containing replication bubble (Bubble) (11) or purified small nascent strands (NS) (58). First column indicates the experimental method (Bubble or NS) and the cell type (HL: HeLa cells and GM: GM06990 cell lines). They are two independent NS-HL datasets labeled 1 and 2. 'All' corresponds to the four initial datasets considered together. NS + 1 kb-HL-2 corresponds to the NS-HL-2 dataset when extending replication origins by 1 kb on both sides. + Bublle-HL corresponds to merging the NS + 1 kb-HL-2 and Bubble-HL datasets. We provide the number of replication origins, their total coverage of ENCODE regions, the number of N-domain borders out of seven within ENCODE regions that match with one of the experimental replication origins and the corresponding *P*-value using a binomial test.

N-domain borders are HS to DNase I digestion

By combining DNase sequencing and DNase tiled microarray strategies, high-resolution DNase I HS sites were identified in human primary $CD4^+$ T cells as markers of open chromatin across the genome (34). The resulting library contains 94 925 DNase I HS sites covering 60 Mb (2.1%) of the human genome. We first observed that only 22.7% of the autosomal DNase I HS sites fall within an N-domain, whereas replication N-domains cover 30% of the human autosomes. This is significantly lower than expected if the DNase I HS sites were uniformly distributed along the human genome ($P < 10^{-15}$ using a binomial test). When mapping DNase I HS sites inside the 678 replication N-domains previously

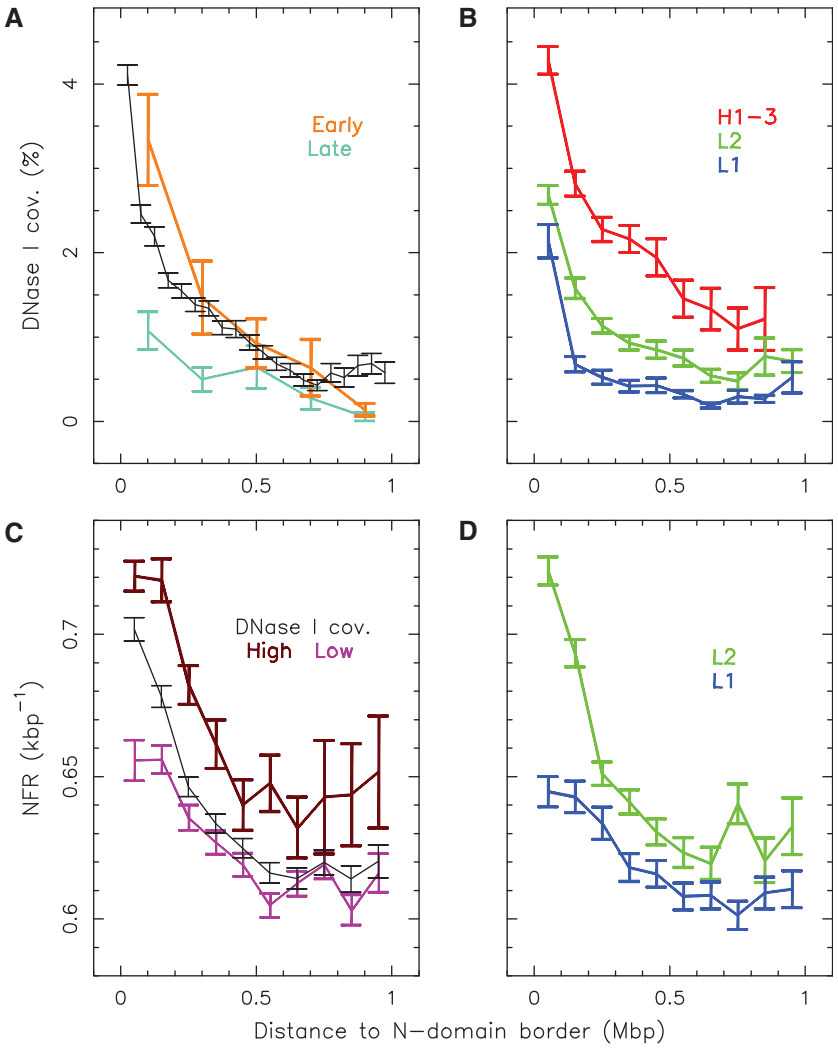

Figure 2. Over representation of DNase I HS sites and *in silico* NFRs at N-domain borders. Mean profiles of DNase I HS sites coverage (**A** and **B**) and NFR density (GC content <41%) (**C** and **D**) over the 678 replication N-domains identified in the human genome (16,17) as a function of the distance to the closest N-domain border. Black lines in (A and C) correspond to the overall average; orange (resp. light blue) line in (A) corresponds to the average over the half N-domains bordered by the 25 earliest (resp. the 25 latest) chromosome 6 putative replication origins (out of a total of 83); brown (resp. purple) line in (C) corresponds to the average over loci presenting a high DNase I HS sites coverage >1% (resp. low <0.2%). In (B and D), color lines correspond to the average over loci belonging to different isochores—blue lines: GC < 37% (L1), green lines: 37 < GC < 41% (L2) and red lines GC > 41% (H1–3).

identified in the human autosomes (16), we observed that the mean site coverage is maximum at the N-domain extremities and decreases significantly from the extremities to the center that is rather insensitive to DNase I cleavage (Figure 2A). This decrease extends over ~150 kb suggesting that N-domain extremities are at the center of an open chromatin region of ~300 kb. This is illustrated by a 19 Mb long fragment of human chromosome 6 containing seven N-domains (Figure 1B) and showing peaks of DNase I hypersensitivity that colocalize within 3 kb for seven ($O_1, O_3, O_6, O_7, O_8, O_{10}$ and O_{11}) out of the 11 distinct N-domain borders (from O_1 to O_{11}). Similar observations were made for each of the 22 human autosomes (Supplementary Figure S1).

When examining high-resolution replication timing data previously measured in chromosome 6 (21), we observed a significant correlation between DNase I HS sites coverage and the replication timing ratio showing that DNase I HS sites are preferentially located in early replicating regions ($r = 0.56$, $P = 1.3 \times 10^{-14}$, Figure 1C). We then compared the mean DNase I HS sites coverage profiles computed around (i) the 25 earliest replicating ($t_r > 1.51$) and the 25 latest replicating ($t_r < 1.4$) N-domain borders among the 83 putative replication origins that border the 54 N-domains predicted along chromosome 6 (17) (Figure 2A). In contrast to the earliest borders located within ±150 kb regions characterized by a high sensitivity to DNase I cleavage, the regions around the 25 latest borders do not display such an enrichment in DNase I HS sites. This suggests that these putative origins lie in a less accessible chromatin environment similar to the one observed at the N-domains centers which replicate late and where genes are rare and expressed in a few tissues (16,17). This was further

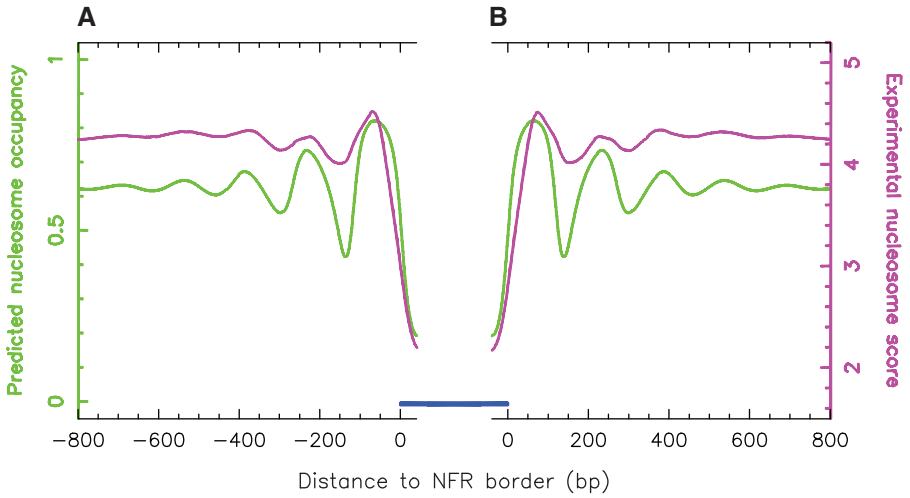

Figure 3. Nucleosome occupancy profiles around *in silico* NFRs. (**A**) and (**B**) Average theoretical nucleosome occupancy probability (green) and experimental nucleosome score (32) (purple) around the 1 017 747 predicted NFRs in low GC content regions (≤41%) when aligned on their 5′ (resp. 3′) borders. The blue bars represent the theoretical NFR predictions (Supplementary Data).

illustrated by the observation that among the 11 N-domain borders predicted in the human chromosome 6 fragment previously examined (Figure 1A), only two (O_2 and O_9) that were found without any DNase I HS site in a close neighborhood presented low timing ratios ($t_r = 1.2$ and 1.25, respectively), as the signature of late replication. For comparison, the seven putative origins that turned out to be HS to DNase I cleavage are all early replicating with $t_r \gtrsim 1.5$.

Recently, it was observed that the density of DNase I HS sites is positively correlated with the GC content, indicating a significant compositional preference in the accessibility of chromatin to DNase I (39). We examined the DNase I HS site coverage around the set of putative replication origins when conditioning the analysis by the GC content, 100 kb windows being grouped into three classes according to their GC level (Figure 2B). Consistently with previous results (39), we observed an overall increase of DNase I HS sites coverage with the GC content. Yet, whatever the GC class, a significant decrease of the DNase I HS sites coverage with the distance to the N-domain borders is robustly observed over a similar ±150 kb distance around the putative replication origins bordering the replication N-domains. This observation demonstrates that DNase I HS at putative replication origins is not a simple consequence of sequence composition.

DNA sequence codes for the accumulation of NFRs around N-domain borders

Previous analysis revealed that promoter regions for protein-coding genes are extremely HS to DNase I digestion (34). These regions were shown to be nucleosome depleted (30–34), very much like the NFRs observed at yeast promoters (36,37). Recent numerical studies revealed that, to a large extent, these NFRs are coded in the DNA sequence via high-energy barriers that impair nucleosome formation (40–42). Furthermore, these excluding genomic energy barriers were shown to play a fundamental role in the collective nucleosomal organization observed over rather large distances along the chromatin fiber (40). Here we used the same physical modeling of nucleosome formation energy based on sequence-dependent bending properties as previously introduced for modeling nucleosome occupancy profiles in the yeast genome (Supplementary Data, Physical Modeling) (40,42). Since the GC content of *Saccharomyces cerevisiae* is rather homogeneous around 39% as compared with the heterogeneous isochore structure of the human genome (43), we restricted our modeling of nucleosome positioning to the light isochores L1 and L2 (GC < 41%). Combining the nucleosome occupancy probability profile and the original energy profile, we identified nucleosome NFRs as the genomic energy barriers that are high enough to induce a nucleosome depleted region in the nucleosome occupancy profile (Supplementary Data, Physical Modeling). When averaging the nucleosome occupancy profiles around the predicted NFR positions along human light isochores, we observed a striking correlation between the profiles obtained using our physical modeling and an experimental genome-wide nucleosome mapping (32), which confirms the relevance of our physical nucleosome modeling to the human genome (Figure 3). Moreover, the concomitant observation of ∨-shape experimental occupancy profiles coinciding with high energy barriers in the sequence-derived nucleosomal formation energy landscape indicates that the regions depleted in nucleosome *in vivo* are likely to be encoded, at least to some extent, in the DNA sequence.

The distribution of NFRs along the 678 N-domains shows a mean density profile that is maximum at N-domain extremities (~0.7 NFR/kb) and that decreases from extremities to center where some NFR depletion is observed (~0.62 NFR/kb) (Figure 2C). This decay over a characteristic length scale ~150 kb is strikingly similar to that displayed by DNase I HS sites coverage (Figure 2A). The excess of NFRs at putative replication origins is robustly observed when conditioning the

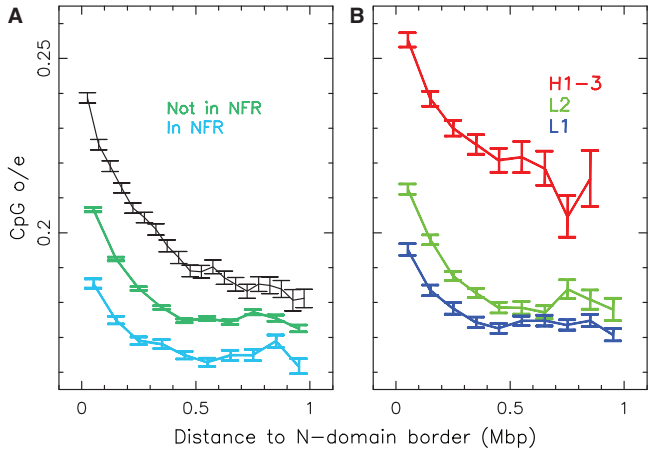

Figure 4. N-domain borders are hypomethylated. Mean profiles of the CpG o/e over the 678 replication N-domains identified in the human genome (16,17) as a function of the distance to the closest N-domain border. In (**A**), black line corresponds to the overall average; blue (resp. green) line corresponds to the average over NFR (resp. non-NFR) loci. In (**B**), colors have the same meaning as in Figure 2B and D.

analysis by the GC content but to a lesser extent in the neighborhood of the GC-poorest origins that likely replicate late in S phase (Figure 2D). Indeed, very much like the DNase I HS sites coverage (Figure 1C), the *in silico* NFR density displays strong correlation with the replication timing ratio data in human chromosome 6 ($r = 0.41$, $P = 1.7 \times 10^{-6}$, Figure 1D). This is in agreement with the fact that no excess of NFRs was observed at the putative replication origins O_2 and O_9 that fire late in the S phase ($t_r = 1.2$ and 1.25, respectively) and where no DNase I HS sites was found in a close neighborhood (Figure 1B).

Altogether, these results show that the NFR density profile displays the same characteristic increase around N-domain borders, as the experimental DNase I HS sites coverage profile. In fact, when comparing the NFR density profiles obtained over loci presenting high (resp. low) DNase I HS sites coverage, we confirmed that NFR enrichment is correlated to higher sensitivity to DNase I (Figure 2C). If this correlation was expected, the fact that we recovered it using a sequence-based modeling of nucleosome occupancy suggests that putative replication origins that border the N-domains are located within regions of accessible open chromatin state that are likely to be encoded in the DNA sequence via excluding energy barriers that inhibit nucleosome formation and participate to the collective ordering of the nucleosome array (40).

DNA hypomethylation is associated with N-domain borders

Cytosine DNA methylation is a mediator of gene silencing in repressed heterochromatic regions, while in potentially active open chromatin regions, DNA is essentially unmethylated (44). DNA methylation is continuously distributed in mammalian genomes with the notable exceptions of CGIs, short unmethylated regions rich in CpGs and of certain promoters and transcription start sites (TSSs) (45). Since there was no genome-wide map of DNA methylation available, we investigated the distribution of DNA methylation using instead indirect estimators calculated directly from the genomic sequence. Methyl-cytosines being hypermutable, prone to deamination to thymines, we considered the CpG o/e ratio as an estimator of DNA methylation (46). Using data from the Human Epigenome Project (47), we confirmed that hypomethylation in sperm corresponded to high values of the CpG o/e outside CGIs (Supplementary Data, Cytosine Methylation). We also observed that the hypomethylation level of CGI's extends to about 1 kb in flanking regions (Supplementary Figure S5), so that the sequence coverage by CGIs enlarged 1 kb at both extremities provided a complementary marker for hypomethylated regions.

When computing CpG o/e after removing CGIs from the analysis along the 19 Mb long reference fragment of human chromosome 6, we found that 9 out of the 11 putative replication origins that border the seven N-domains correspond to a well-defined local maximum of the CpG o/e profile (Figure 1B). Since CpG o/e values are known to be positively correlated with the GC content (48), we determined the CpG o/e profiles for fixed GC content. When averaging over the 678 N-domains, the overall mean CpG o/e profile (Figure 4A), as well as the mean CpG o/e profiles obtained for each class of GC content (Figure 4B), present a maximum at origin positions, as the signature of hypomethylation, and decrease over a characteristic distance ~150 kb, similar to the one found for DNase I HS sites coverage and NFR density profiles (Figure 2), from the extremities to the center of N-domains where a minimal level of CpG o/e is attained. These data show that the peak of CpG o/e observed around the putative replication origins, even when CGIs were removed from the genome, cannot be attributed to some peculiar GC-content environment but more likely to a hypomethylated open chromatin state where CpG o/e is correlated with DNase I HS sites coverage ($r = 0.35$, $P < 10^{-15}$) and NFR density ($r = 0.71$, $P < 10^{-15}$) (Figure 1E). The decrease of the CpG o/e with the distance to N-domain border was robustly observed when considering separately NFR and non-NFR regions (Figure 4A) showing that the gradient of hypomethylation over ~150 kb around N-domain borders is not specific to either of these regions. The correlation measured between CpG o/e and replication timing ($r = 0.30$, $P = 1.1 \times 10^{-4}$, Figure 1C) further indicates that this property is significantly associated with the putative replication origins located in early-replicating regions. The complementary analysis using the 1-kb-enlarged CGI coverage as hypomethylation marker provided exactly the same diagnosis (Supplementary Figure S4). We observed that each of the 11 N-domain borders presented in Figure 1B corresponds to a peak in the 1-kb-enlarged CGI coverage profile (data not shown) and that the average over the 678 N-domains decreases from borders to center over a similar characteristic distance ~150 kb. These observations are consistent with the hypothesis (49) that CGIs are protected from methylation due to the colocalization with replication origins.

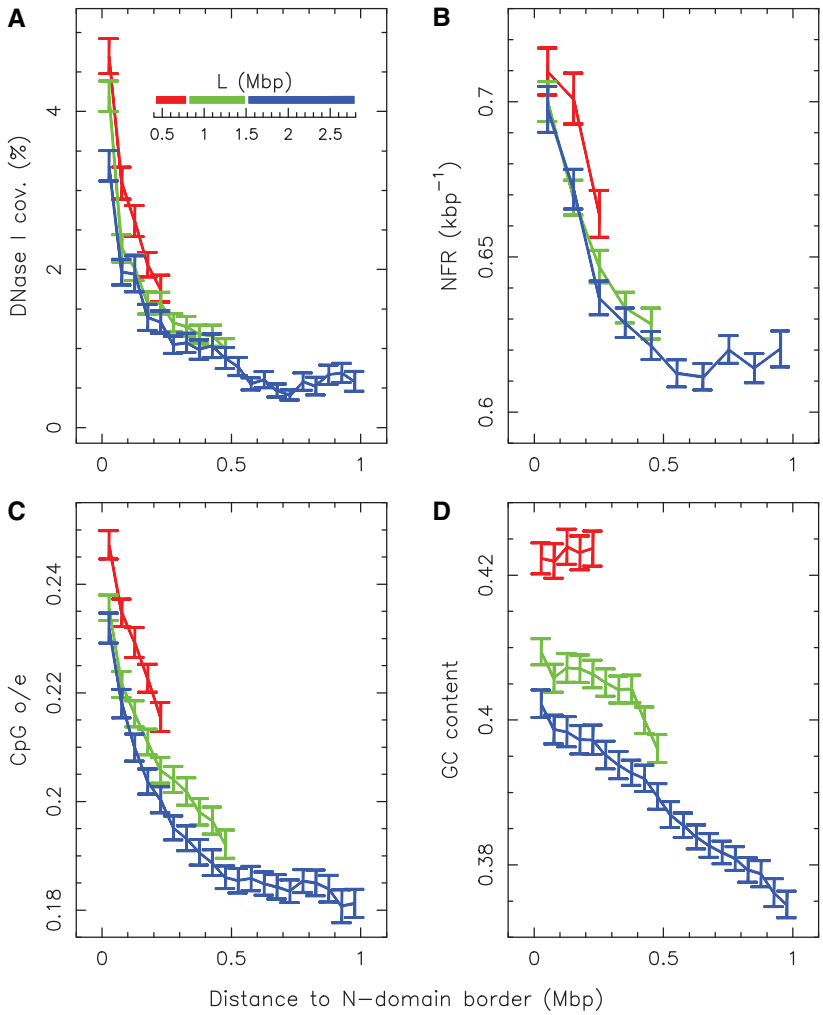

Figure 5. Open chromatin regions around N-domain borders have a characteristic size. Mean profiles of DNase I HS sites coverage (**A**), NFR density (GC content <41%) (**B**), CpG o/e (**C**) and GC content (**D**) as a function of the distance to the closest N-domain border over the 678 replication N-domains identified in the human genome (16,17) for three N-domain size categories: $L < 0.8$ Mb (red), $0.8 < L < 1.5$ Mb (green) and $L > 1.5$ Mb (blue).

Open chromatin regions around N-domain borders have a characteristic size

The decreasing behavior over a characteristic distance of ±150 kb from N-domain borders common to the mean DNase I HS site coverage profile, the mean NFR density profile, the mean CpG o/e profile and the average 1-kb-enlarged CGI coverage was observed whatever the size of the replication N-domains (Figure 5A,B and C; Supplementary Figure S4A). This contrasts with the GC-content profile that behaves quite differently (Figure 5D). For small N-domains of size ($L < 0.8$ Mb), the GC profile is rather flat all along the domains, whereas for the larger sizes ($L > 0.8$ Mb), it decreases very slowly toward the N-domain center. These results confirm that the excess of CpG o/e observed around the putative replication origins does not simply reflect some localized high-GC environment but more likely some open chromatin state with a ~300 kb mean characteristic size. Chromatin structure has also been analyzed at the fiber level using separation by sucrose gradient sedimentation (29). We observed that the proportion of microarray clones presenting an open/input ratio >1.5 decreased 5-fold from N-domain borders to centers (Figure 6B). This result provides additional support for the peculiar property of chromatin in the neighborhood of N-domain borders.

Both intergenes and TSSs present active open chromatin marks close to N-domain borders

It was previously reported that N-domain borders correspond to a high gene concentration (16), TSS density profiles presenting, as expected, a strong similarity with 1-kb-enlarged CGI coverage (Supplementary Figure S4A and B). We observed a concomitant increase in the proportion of small genes and small intergenes in proximity of N-domain borders (Supplementary Figure S4C and D). Since the previous open chromatin markers have been associated, at least to some extent, with genes [e.g. 16% of all DNase I HS sites are in the first exon or at the TSS of a gene, and 42% are found inside a gene (34)],

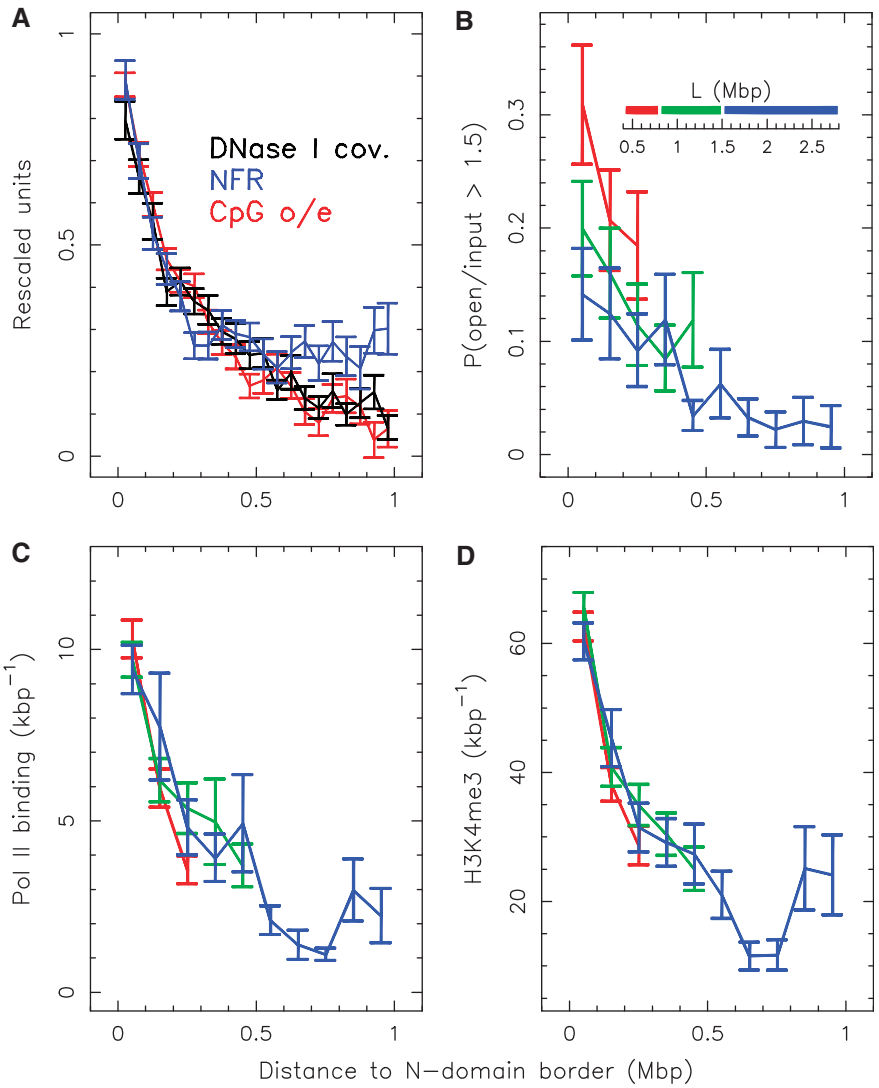

Figure 6. Both intergenes and TSSs present open chromatin marks close to N-domain borders. (**A**) Mean profiles of DNase I HS sites coverage (black), NFR density (GC content <41%, blue) and CpG o/e (red) as a function of the distance to the closest N-domain border after masking CGIs and genes extended by 2 kb at both extremities. (**B**) Proportion of clones presenting a ratio of 'open' over input chromatin >1.5 versus the distance to the closest N-domain border. (**C** and **D**) Mean profile of Pol II and H3K4me3 Chip-Seq tag density ±2 kb around TSS versus the distance to the closest N-domain border. In (B, C and D) colors correspond to three N-domain size categories: $L < 0.8$ Mb (red), $0.8 < L < 1.5$ Mb (green) and $L > 1.5$ Mb (blue).

we reproduced the analysis of their distribution along the N-domains after masking the genes extended by 2 kb at both extremities and the CGIs (Figure 6A). The fact that the mean DNase I HS sites coverage, NFR density and CpG o/e profiles still present the decaying behavior over ~150 kb demonstrates that the excess observed around the putative replication origins does not simply reflect the rather packed gene organization at the N-domain borders. Additionally, we assessed the transcriptional potential of genes as a function of their distance to N-domain borders. We observed that Pol II binding and H3K4me3 Chip-Seq tag density ±2 kb around TSS also presents a strongly decaying mean profile over a length of ~150 kb from N-domain borders to center (5- and 3-fold, respectively, Figure 6C and D). The presence of these two marks at the TSS have been shown to correlate with gene activity (33). These results thus indicate that the open chromatin regions around putative replication origins are prone to transcription, whereas N-domain central regions appear transcriptionally silent.

Open chromatin around N-domain borders are potentially fragile regions involved in chromosome instability

Since chromatin accessibility and openness are possible factors responsible for fragility and instability, N-domain borders could also play a key role in genome dynamics during evolution and genome instability in pathologic situations like cancer. A study of evolutionary breakpoint regions (breakage of synteny) along human chromosomes (50) shows that they appear more

frequently near N-domain borders than in their central regions (51), suggesting that the distribution of large-scale rearrangements in mammals reflects a mutational bias toward regions of high transcriptional activity and replication initiation (Supplementary Figure S6). Furthermore, the fact that chromosome anomalies involved in the tumoral process like at the RUNX1T1 oncogene locus (Supplementary Figure S6) coincide with replication N-domain extremities raises the possibility that the replication origins detected *in silico* are potential candidate loci susceptible to breakage in some cancer cell types.

N-domain borders: a subset of 'master' replication origins

Comparison to experimentally identified replication origins provided further support that most N-domain borders are replication origins active early in S-phase. Our findings show that these putative replication origins are located within a ~300 kb region extremely sensitive to DNase I cleavage, presenting hypomethylation marks and enriched in open chromatin fibers, suggesting that these regions present an open chromatin structure. This accessible chromatin organization is to some extent encoded in the DNA sequence via an enrichment in nucleosome excluding energy barriers (NFRs). The additional observation that the densities of Pol II binding and H3K4me3 around TSSs (±2 kb) close to N-domain borders significantly exceeds the density found around TSSs in N-domain central regions (Figure 6C and D), suggesting that this local chromatin structure is associated with transcriptional activity. However, the fact that some N-domain borders (like O_2) neither present an open chromatin signature nor an early replication timing in the cell line used experimentally, but still exhibit a sharp upward jump in the skew profile, raises the question of whether they have a different status or are associated with open chromatin and early replication only in the germline.

Are these traits shared by many other replication origins? In metazoans, recognition of replication origins by the origin recognition complex (ORC) does not involve simple consensus DNA sequence. Initiation sites do not share common genetic entities but seem to be favored by various factors that can differ from one origin to another and be required or dispensable under different conditions (4). Specification of initiation sites can be favored by negatively supercoiled DNA (52) (possibly resulting from the removal or displacement of nucleosomes), interacting proteins that chaperone ORC to specific chromatin sites (53), by the transcriptional activity (54) or open chromatin to which ORC might bind in a nonspecific way (55). A recent study performed on 283 replication origins identified in the ENCODE regions showed that, besides a strong association with CGIs, only 29% overlap a DNase I HS site and that half of these origins do not present open chromatin epigenetic marks and are not associated with active transcription (9). The particular open chromatin state associated with N-domain borders suggest that these putative early replication origins present properties that are only shared by a subset of origins. These properties likely contribute to the specification of this peculiar subset of origins that will be further qualified as 'master' replication origins.

In conclusion, analyses of experimental and numerical open chromatin markers suggest the existence of 'master' replication origins likely to be active in germline as well as somatic human cells. These privileged loci were identified as upward jumps in the strand asymmetry profiles accumulated during evolution, which attest that they are well positioned in the germline. We show that they are located within a ~300 kb wide region of open chromatin, encoded in the DNA sequence via an enrichment of nucleosome excluding energy barriers. Interestingly, location O_0 that was not identified as an N-domain border (16) but displays all these open chromatin characteristics (Figure 1B) actually corresponds to a sharp upward jump in the skew profile (Figure 1A), as the hallmark of the presence of a 'master' replication origin. Such a strong gradient of accessible and open chromatin environment is not observed around a large fraction of the replication origins experimentally identified along ENCODE regions (9). The typical inter-origin distance in the human somatic cells has been estimated to be of the order of 50–100 kb (9,56), a value significantly smaller that the typical size (1 Mb) of N-domains. We propose that replication would initiate in early S phase at these privileged open chromatin locations and that the replication timing gradients observed from 'master' origins (17) would correspond to the diverging replication forks progression triggering secondary origins that supress in a 'domino' cascade manner. As structural defects (bursts of 'openness') in the chromatin fiber, these 'master' replication origins might also be central to the tertiary structure of eukaryotic chromatin into rosette-like structures (57). The present data suggest that they are likely to be associated with structuring chromatin elements playing an essential role in the spatio-temporal replication program.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank C. Chen, M. Huvet, O. Hyrien, S. Nicolay and M. Touchon for helpful discussions.

FUNDING

Conseil Régional Rhône-Alpes (Emergence 2005); Agence Nationale de la Recherche (ANR) [projects HUGOREP (NT05_3_41825) and DNAnucl (ANR_06_PCVI_0026)]. Funding for open access charge: ANR (NT05_3_41825).

Conflict of interest statement. None declared.

REFERENCES

1. Hyrien,O. and Méchali,M. (1993) Chromosomal replication initiates and terminates at random sequences but at regular intervals in the

ribosomal DNA of Xenopus early embryos. *EMBO J.*, **12**, 4511–4520.
2. Berezney,R., Dubey,D.D. and Huberman,J.A. (2000) Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. *Chromosoma*, **108**, 471–484.
3. Bell,S.P. and Dutta,A. (2002) DNA replication in eukaryotic cells. *Annu. Rev. Biochem.*, **71**, 333–374.
4. Gilbert,D.M. (2004) In search of the holy replicator. *Nat. Rev. Mol. Cell Biol.*, **5**, 848–855.
5. Schwaiger,M. and Schubeler,D. (2006) A question of timing: emerging links between transcription and replication. *Curr. Opin. Genet. Dev.*, **16**, 177–183.
6. Gerbi,S.A. and Bielinsky,A.K. (2002) DNA replication and chromatin. *Curr. Opin. Genet. Dev.*, **12**, 243–248.
7. Lemaitre,J.-M., Danis,E., Pasero,P., Vassetzky,Y. and Mechali,M. (2005) Mitotic remodeling of the replicon and chromosome structure. *Cell*, **123**, 787–801.
8. Courbet,S., Gay,S., Arnoult,N., Wronka,G., Anglana,M., Brison,O. and Debatisse,M. (2008) Replication fork movement sets chromatin loop size and origin choice in mammalian cells. *Nature*, **455**, 557–560.
9. Cadoret,J.-C., Meisch,F., Hassan-Zadeh,V., Luyten,I., Guillet,C., Duret,L., Quesneville,H. and Prioleau,M.-N. (2008) Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc. Natl Acad. Sci. USA*, **105**, 15837–15842.
10. The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
11. Mesner,L.D., Crawford,E.L. and Hamlin,J.L. (2006) Isolating apparently pure libraries of replication origins from complex genomes. *Mol. Cell*, **21**, 719–726.
12. Hamlin,J.L., Mesner,L.D., Lar,O., Torres,R., Chodaparambil,S.V. and Wang,L. (2008) A revisionist replicon model for higher eukaryotic genomes. *J. Cell Biochem.*, **105**, 321–329.
13. McNairn,A.J. and Gilbert,D.M. (2003) Epigenomic replication: linking epigenetics to DNA replication. *Bioessays*, **25**, 647–656.
14. Méchali,M. (2001) DNA replication origins: from sequence specificity to epigenetics. *Nat. Rev. Genet.*, **2**, 640–645.
15. Bogan,J.A., Natale,D.A. and Depamphilis,M.L. (2000) Initiation of eukaryotic DNA replication: conservative or liberal? *J. Cell Physiol.*, **184**, 139–150.
16. Huvet,M., Nicolay,S., Touchon,M., Audit,B., d'Aubenton-Carafa,Y., Arneodo,A. and Thermes,C. (2007) Human gene organization driven by the coordination of replication and transcription. *Genome Res.*, **17**, 1278–1285.
17. Audit,B., Nicolay,S., Huvet,M., Touchon,M., d'Aubenton-Carafa,Y., Thermes,C. and Arneodo,A. (2007) DNA replication timing data corroborate in silico human replication origin predictions. *Phys. Rev. Lett.*, **99**, 248102.
18. Touchon,M., Nicolay,S., Audit,B., Brodie of Brodie,E.-B., d'Aubenton-Carafa,Y., Arneodo,A. and Thermes,C. (2005) Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc. Natl Acad. Sci. USA*, **102**, 9836–9841.
19. Brodie of Brodie,E.-B., Nicolay,S., Touchon,M., Audit,B., d'Aubenton-Carafa,Y., Thermes,C. and Arneodo,A. (2005) From DNA sequence analysis to modeling replication in the human genome. *Phys. Rev. Lett.*, **94**, 248103.
20. Necsulea,A., Guillet,C., Cadoret,J.-C., Prioleau,M.-N. and Duret,L. (2009) The relationship between DNA replication and human genome organization. *Mol. Biol. Evol.*, **26**, 729–741.
21. Woodfine,K., Beare,D.M., Ichimura,K., Debernardi,S., Mungall,A.J., Fiegler,H., Collins,V.P., Carter,N.P. and Dunham,I. (2005) Replication timing of human chromosome 6. *Cell Cycle*, **4**, 172–176.
22. Karnani,N., Taylor,C., Malhotra,A. and Dutta,A. (2007) Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. *Genome Res.*, **17**, 865–876.
23. Hiratani,I., Ryba,T., Itoh,M., Yokochi,T., Schwaiger,M., Chang,C.-W., Lyou,Y., Townes,T.M., Schubeler,D. and Gilbert,D.M. (2008) Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.*, **6**, e245.
24. Farkash-Amar,S., Lipson,D., Polten,A., Goren,A., Helmstetter,C., Yakhini,Z. and Simon,I. (2008) Global organization of replication time zones of the mouse genome. *Genome Res.*, **18**, 1562–1570.
25. MacAlpine,D.M., Rodriguez,H.K. and Bell,S.P. (2004) Coordination of replication and transcription along a Drosophila chromosome. *Genes Dev.*, **18**, 3094–3105.
26. Schübeler,D., Scalzo,D., Kooperberg,C., vanSteensel,B., Delrow,J. and Groudine,M. (2002) Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. *Nat. Genet.*, **32**, 438–442.
27. Goldman,M.A., Holmquist,G.P., Gray,M.C., Caston,L.A. and Nag,A. (1984) Replication timing of genes and middle repetitive sequences. *Science*, **224**, 686–692.
28. Sequeira-Mendes,J., Diaz-Uriarte,R., Apedaile,A., Huntley,D., Brockdorff,N. and Gomez,M. (2009) Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet.*, **5**, e1000446.
29. Gilbert,N., Boyle,S., Fiegler,H., Woodfine,K., Carter,N.P. and Bickmore,W.A. (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell*, **118**, 555–566.
30. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Calcar,S.V., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
31. Ozsolak,F., Song,J.S., Liu,X.S. and Fisher,D.E. (2007) High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.*, **25**, 244–248.
32. Schones,D.E., Cui,K., Cuddapah,S., Roh,T.-Y., Barski,A., Wang,Z., Wei,G. and Zhao,K. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.
33. Barski,A., Cuddapah,S., Cui,K., Roh,T.-Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
34. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
35. Kohzaki,H. and Murakami,Y. (2005) Transcription factors and DNA replication origin selection. *Bioessays*, **27**, 1107–1116.
36. Lee,W., Tillo,D., Bray,N., Morse,R.H., Davis,R.W., Hughes,T.R. and Nislow,C. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.
37. Yuan,G.-C., Liu,Y.-J., Dion,M.F., Slack,M.D., Wu,L.F., Altschuler,S.J. and Rando,O.J. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, **309**, 626–630.
38. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC genome browser database. *Nucleic Acids Res.*, **31**, 51–54.
39. Di Filippo,M. and Bernardi,G. (2008) Mapping DNase-I hypersensitive sites on human isochores. *Gene*, **419**, 62–65.
40. Vaillant,C., Audit,B. and Arneodo,A. (2007) Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys. Rev. Lett.*, **99**, 218103.
41. Mavrich,T.N., Ioshikhes,I.P., Venters,B.J., Jiang,C., Tomsho,L.P., Qi,J., Schuster,S.C., Albert,I. and Pugh,B.F. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, **18**, 1073–1083.
42. Miele,V., Vaillant,C., d'Aubenton-Carafa,Y., Thermes,C. and Grange,T. (2008) DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.*, **36**, 3746–3756.
43. Bernardi,G. (2001) Misunderstandings about isochores. Part 1. *Gene*, **276**, 3–13.
44. Bird,A.P. and Wolffe,A.P. (1999) Methylation-induced repression–belts, braces, and chromatin. *Cell*, **99**, 451–454.
45. Suzuki,M.M. and Bird,A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.
46. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.

47. Eckhardt,F., Lewin,J., Cortese,R., Rakyan,V.K., Attwood,J., Burger,M., Burton,J., Cox,T.V., Davies,R., Down,T.A. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**, 1378–1385.
48. Duret,L. and Galtier,N. (2000) The covariation between TpA deficiency, CpG deficiency, and G + C content of human isochores is due to a mathematical artifact. *Mol. Biol. Evol.*, **17**, 1620–1625.
49. Antequera,F. and Bird,A. (1999) CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr. Biol.*, **9**, R661–R667.
50. Lemaitre,C., Tannier,E., Gautier,C. and Sagot,M.-F. (2008) Precise detection of rearrangement breakpoints in mammalian chromosomes. *BMC Bioinformatics*, **9**, 286.
51. Lemaitre,C., Zaghloul,L., Sagot,M.-F., Gautier,C., Arneodo,A., Tannier,E. and Audit,B. (2009) Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relations to genome organisation. *BMC Genomics*, **10**, 335.
52. Remus,D., Beall,E.L. and Botchan,M.R. (2004) DNA topology, not DNA sequence, is a critical determinant for Drosophila ORC-DNA binding. *EMBO J.*, **23**, 897–907.
53. Schepers,A., Ritzi,M., Bousset,K., Kremmer,E., Yates,J.L., Harwood,J., Diffley,J.F. and Hammerschmidt,W. (2001) Human origin recognition complex binds to the region of the latent origin of DNA replication of epstein-barr virus. *EMBO J.*, **20**, 4588–4602.
54. Danis,E., Brodolin,K., Menut,S., Maiorano,D., Girard-Reydet,C. and Méchali,M. (2004) Specification of a DNA replication origin by a transcription complex. *Nat. Cell Biol.*, **6**, 721–730.
55. Vashee,S., Cvetic,C., Lu,W., Simancek,P., Kelly,T.J. and Walter,J.C. (2003) Sequence-independent DNA binding and replication initiation by the human origin recognition complex. *Genes Dev.*, **17**, 1894–1908.
56. Conti,C., Sacca,B., Herrick,J., Lalou,C., Pommier,Y. and Bensimon,A. (2007) Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Mol. Biol. Cell*, **18**, 3059–3067.
57. St-Jean,P., Vaillant,C., Audit,B. and Arneodo,A. (2008) Spontaneous emergence of sequence-dependent rosettelike folding of chromatin fiber. *Phys. Rev. E*, **77**, 061923.
58. Gerbi,S.A. and Bielinsky,A.K. (1997) Replication initiation point mapping. *Methods*, **13**, 271–280.