The Relationship between DNA Replication and Human Genome Organization

Anamaria Necşulea, Claire Guillet,* Jean-Charles Cadoret,† Marie-Noëlle Prioleau,† and Laurent Duret**

*Université de Lyon, Lyon; Université Lyon 1; CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Villeurbanne, France; HELIX, Unité de recherche INRIA; and †Institut Jacques Monod, CNRS, Université Paris7, Université Pierre et Marie Curie, Paris, France

Assessment of the impact of DNA replication on genome architecture in Eukaryotes has long been hampered by the scarcity of experimental data. Recent work, relying on computational predictions of origins of replication, suggested that replication might be a major determinant of gene organization in human (Huvet et al. 2007. Human gene organization driven by the coordination of replication and transcription. Genome Res. 17:1278–1285). Here, we address this question by analyzing the first large-scale data set of experimentally determined origins of replication in human: 283 origins identified in HeLa cells, in 1% of the genome covered by ENCODE regions (Cadoret et al. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc Natl Acad Sci USA. 105:15837–15842). We show that origins of replication are not randomly distributed as they display significant overlap with promoter regions and CpG islands. The hypothesis of a selective pressure to avoid frontal collisions between replication and transcription polymerases is not supported by experimental data as we find no evidence for gene orientation bias in the proximity of origins of replication. The lack of a significant orientation bias remains manifest even when considering only genes expressed at a high rate, or in a wide number of tissues, and is not affected by the regional replication timing. Gene expression breadth does not appear to be correlated with the distance from the origins of replication. We conclude that the impact of DNA replication on human genome organization is considerably weaker than previously proposed.

Introduction

During each cell cycle, the genome must be accurately replicated and transcribed. These essential cellular mechanisms are likely to be the source of strong functional constraints on the DNA sequence. The extent to which the superposition of these constraints affects genome organization is, in many species, yet to be evaluated.

In Bacteria, for which the term "replication-related organization" has been proposed (Rocha 2004), DNA replication has undoubtedly a profound impact on genome architecture. Bacterial chromosomes generally have a single origin and terminus of replication that define the boundaries of two replichores equal in size but unequal in many other aspects. The most conspicuous asymmetric feature between the two replichores is nucleotide composition; the position of the origin and terminus of replication can be reliably predicted from the variation in composition bias along the chromosome (Lobry 1996a, 1996b).

DNA replication also appears to influence the chromosomal arrangement of genes. For most bacterial species, genes are preferentially coded on the leading strand for replication (McLean et al. 1998; Mrazek and Karlin 1998), especially those that are essential for the organism (Rocha and Danchin 2003). As a consequence of this gene orientation bias, the frequency of deleterious frontal collisions between replication and transcription polymerases is greatly reduced (Nomura and Morgan 1977; Brewer 1988). Another gene distribution bias is present in fast-growing species: highly expressed genes, such as those involved in transcription and translation, tend to cluster near the origin of replication (Ardell and Kirsebom 2005; Couturier and Rocha 2006). As genes near the origin are present in multiple copies during replication, this preferential positioning is believed to increase the level of expression through gene dosage effects (Sousa et al. 1997).

In Eukaryotes, the relationship between DNA replication and genome organization has been more elusive, not least because the precise identification of origins and termini is considerably more difficult than in prokaryotes. The positions of replication origins are best known for yeast (Fangman and Brewer 1991; Raghuraman et al. 2001; Wyrick et al. 2001). However, studies of the yeast genome have failed to reveal the strong association between DNA replication and genome structure encountered in prokaryotes: asymmetric nucleotide composition is present only in subtelomeric regions (Gierlik et al. 2000), and no significant association between replication and transcription has been found (Nieduszynski et al. 2006).

Recently, unexpected evidence in favor of a replication-related organization in Eukaryotes has come from analyses of the human genome (Huvet et al. 2007). Under the assumption that the replicon size is of the order of 100 kb (Huberman and Riggs 1968), the human genome should possess several thousands of origins of replication, but until recently, only a minuscule fraction had been experimentally determined. Although scarce, the available data have provided the basis for the genome-wide prediction of replication origins: analysis of nucleotide composition asymmetry, or skew, around experimentally determined origins revealed that in most cases the skew displays an abrupt sign switch at the origin, as it does in prokaryotes (Touchon et al. 2005). The nucleotide skew decreases linearly between two consecutive origins of replication, from positive values in 3′ of the first origin to negative values in 5′ of the second one (Touchon et al. 2005). These observations have led to the development of a computational method for the prediction of putative origins of replication, based on the identification of chromosomal regions with linearly decreasing skews: the so-called N-domains, bordered by putative origins of replication (Touchon et al. 2005; Huvet et al. 2007). This method has provided the first large-scale set of replication unit predictions: around 1,000 putative origins, distributed over more

Key words: origin of replication, genome organization, transcription, nucleotide composition, polymerase collision.

E-mail: duret@biomserv.univ-lyon1.fr.

Mol. Biol. Evol. 26(4):729–741. 2009
doi:10.1093/molbev/msn303
Advance Access publication January 6, 2009

than one-quarter of the human genome, have been detected (Touchon et al. 2005; Huvet et al. 2007).

Surprisingly, it appeared that gene distribution along the predicted replication domains is highly ordered (Huvet et al. 2007). Gene density decreases from the borders to the center of the N-domains and so does gene expression breadth. The transcription orientation is also nonrandom: genes seem to be preferentially coded on the leading strand for replication, especially near the putative origins of replication. These findings provided support for a new model of human genome organization, defined by the interplay of replication and transcription (Huvet et al. 2007).

In silico prediction of replication origins clearly represents a major step toward understanding the impact of DNA replication on human genome organization. Nevertheless, independent validation of the results obtained in silico remains necessary and is now possible, thanks to the recent publication of the first large-scale experimental data set, consisting of 283 human replication origins (Cadoret et al. 2008). Initial analyses of this data set confirmed that the distribution of origins of replication is correlated with other aspects of genome structure, such as GC content variation and distribution of transcriptional regulatory elements (Cadoret et al. 2008), but no direct comparison was made with the structural features of computationally predicted replication domains. Here, we analyze the characteristics of both experimentally and in silico determined origin data sets. Our results confirm that the genomic distribution of origins of replication is nonrandom. However, we find no evidence in favor of a relationship between replication, transcription orientation, and expression breadth. The influence of DNA replication on human genome organization may therefore be less important than previously suggested.

Materials and Methods

Sequence Data Set

All analyses were done on the March 2006 assembly of the human genome (versions NCBI 36, hg18). When necessary, conversions between different assembly versions were done using the liftOver utility of the University of California–Santa Cruz Genome Browser (Karolchik et al. 2003). Human genome annotations were extracted from the Ensembl database, release 50 (Hubbard et al. 2007). For the analyses presented here, we only analyzed genes that have at least one corresponding transcript in the RefSeq database (Pruitt et al. 2007). We used Galaxy (Giardine et al. 2005) to extract the coordinates of intergenic and intronic sequences from Ensembl annotations.

Origins of Replication

The positions of 283 experimentally determined origins of replication were taken from Cadoret et al. (2008). We will refer to this data set as *OriExp*. These origins of replication were determined experimentally, with a method based on hybridization of short nascent strands on DNA microarrays designed for ENCODE regions (The ENCODE Project Consortium 2007). The origin length varies between ≈600 and ≈4,500 bp, with an average of ≈1,500 bp.

The positions of 678 replication domains (also termed N-domains) were taken from Huvet et al. (2007). These domains were predicted in silico, with a computational method based on the association of nucleotide composition asymmetry with DNA replication. The borders of the N-domains correspond to 1,060 putative origins of replication. We will refer to this data set as *OriIS*. Unlike the experimental approach, the computational method used by Huvet et al. (2007) does not provide an estimate of the origin length, instead putative origins are represented by a single nucleotide position. To avoid potential biases arising from this discrepancy between the two methods, experimental origins are also represented here as a single nucleotide position, corresponding to the central point of the origin segment.

CpG Islands Data Set

The CpG islands ("CGI") were determined using the procedure proposed by Ponger et al. (2001). The CGI are defined as sequences longer than 500 bp, with a ratio of observed over expected number of CpG dinucleotides > 0.6 and average G + C content > 0.5. The analysis was done on the unmasked genome sequence; therefore, CGI can overlap with repetitive sequences.

Gene Expression Data Set

We used expressed sequence tag (EST) and SAGE data to determine expression patterns for human genes, following the procedure described by Semon and Duret (2006). We extracted 8,137,901 human ESTs from GenBank release 166 (July 2008, Benson et al. [2008]). We extracted tissue/organ annotations for each EST and we pooled together EST libraries from the same tissues/organs. We excluded pooled libraries with less than 10,000 ESTs, as well as ESTs from tumors, cell cultures, embryonic tissues, and pooled or unidentified tissues. When a whole organ and parts of it were present (e.g., whole brain, hippocampus, thalamus, substantia nigra, etc.), we removed the library corresponding to the whole organ to avoid redundancy. The final data set consisted of 3,199,559 ESTs from 47 normal tissues. We used nucleotide Blast (Altschul et al. 1990) to determine the correspondence between ESTs and annotated genes, with a similarity threshold of at least 95% identity and 100 nt length. With this procedure, we identified 21,887 genes with at least one EST in our data set. We consider that a gene is "broadly expressed" if transcribed in at least 23 of the 47 tissues and that it is "narrowly expressed" if transcribed in at most five tissues.

For the SAGE data set, we used the short tag libraries provided by SAGE Genie (downloaded from http://cgap.nci.nih.gov/SAGE in July 2008, Boon et al. [2002]). We retrieved 92 libraries with at least 20,000 tags, corresponding to a total of 26 adult nontumoral tissues. For each annotated gene, we extracted the two most likely short tags (10 nt downstream of the two most 3′ *Nla*III restriction sites). We removed from the analysis all tags that could not be

unambiguously assigned to a single gene. The final data set consisted of 13,623 genes with unambiguous tags. We pooled libraries corresponding to the same tissue. We consider that a gene is broadly expressed if transcribed in at least 13 of the 26 tissues and that it is narrowly expressed if transcribed in at most three tissues.

We also extracted from the Gene Expression Omnibus (Edgar et al. 2002), an Affymetrix microarray-based data set that investigates the gene expression profile during the cell cycle for T89G cells (Litovchick et al. 2007). We retrieved the information present in the GSM211871 sample that concerns specifically genes expressed in the S-phase. Each gene is represented by several microarray probesets. An estimation of the probability of presence/absence of the transcript, computed with the Affymetrix MAS5 algorithm, is given for each probeset. We considered that a gene is transcribed during the S-phase if at least one of its probesets was classified as "present." Conversely, we considered that a gene is inactive during the S-phase if all of its probesets were classified as "absent."

To analyze the gene orientation bias as a function of the level of expression in HeLa cells, we extracted from the Gene Expression Omnibus the Affymetrix data set provided by Scotto et al. (2008), investigating gene expression profile in various types of cervical tumors. The information specific to HeLa cells is given in the GSM246123 sample. We classified probesets according to the value of the hybridization signal. We considered that genes are expressed at a high level if at least one of their corresponding probesets was in the top 33% of the hybridization values. Conversely, a gene is considered to be expressed at a low level if all its probesets are in the bottom 33% of the hybridization values.

Replication Timing Data Set

We used the replication timing data set provided by Karnani et al. (2007) and extracted through the UCSC Genome Browser. This data set is specific to ENCODE regions. In this data set, genomic regions are characterized as early, mid, late, or PanS replicating (the last class corresponds to regions that are found to replicate at more than one point during the S-phase, probably due to allelic differences in replication timing). Out of 283 *OriExp* origins, 65 were in early-replicating regions, 75 in mid-replicating, 36 in late-replicating, and 87 in PanS-replicating regions.

Leading and Lagging Strand Definition

We defined leading and lagging strand segments with respect to the published sequence strand. Leading strand segments are delimited in 5′ by an origin for replication and in 3′ by a terminus. Conversely, lagging strand segments are delimited in 3′ by an origin and in 5′ by a terminus. In eukaryotes, termination is believed to occur at the meeting point between two convergent replication forks (Zhu et al. 1992). Under the simplifying assumption that all origins are equally active, replication termini can thus be assigned to the midpoint of the segment between two consecutive origins. We applied this reasoning to determine leading and lagging strand segments for *OriExp* and *OriIS*. Because the interorigin distance is much higher for the latter to limit the discrepancy between the two data sets, we set the maximum size of the segments at 50 kb.

For the analysis of the coorientation between replication and transcription, we computed for each gene the number of nucleotides for which the sense strand for transcription corresponds to either leading or lagging strand for replication; transcribed nucleotides that fall in ambiguous (with respect to replication) regions were ignored. Nucleotides are counted only once for each gene, even when they belong to several alternative transcripts.

Analysis of Sequence Conservation

We used the UCSC Genome Browser to extract the coordinates of sequences that are highly conserved among placental mammals (Siepel et al. 2005). For this analysis, we used a homogenized origin definition for the two data sets, by reducing the experimental origins to a single base pair, chosen as the center of the initially given segments. We then analyzed sequence conservation in 2 kb segments centered around the origin. Using Ensembl annotations, we masked exonic segments and we computed for each 2 kb segment the nonexonic fraction that overlaps with highly conserved sequences. We divided the origins into three classes: CGI TSS (at less than 2 kb from both CGI and annotated transcription start sites ["TSS"]), CGI NonTSS (at less than 2 kb from CGI and more than 2 kb from annotated TSS), and NonCGI NonTSS (at more than 2 kb from both CGI and annotated TSS). As there are only few origins close to TSS but not to CGI, we excluded these origins from the analysis. For each class of origins, we computed the average conserved fraction, weighted by the length of nonexonic origin segments.

To obtain the expected distributions, we computed the average conserved fraction in simulated data sets consisting of 100 2 kb segments, drawn at random in ENCODE regions (for *OriExp*) and in the whole genome (for *OriIS*). The segments were chosen so that they do not intersect with *OriExp* or *OriIS* positions. We did 1,000 simulations for each origin class.

Analysis of Nucleotide Composition Asymmetry

We computed the GC skew ($S_{GC} = 100 \times \frac{(G-C)}{(G+C)}$) and the TA skew ($S_{TA} = 100 \times \frac{(T-A)}{(T+A)}$), and the global skew ($S = S_{GC} + S_{TA}$) on the leading and lagging strand segments defined above, separately for nontranscribed regions, forward-transcribed intronic regions, and reverse-transcribed intronic regions. Repetitive sequences were masked with RepeatMasker (Smit et al. 1996–2004) before computing the nucleotide composition. Only segments with at least 100 nt of nonrepetitive sequences were considered when computing the skew.

Randomization Procedures for Statistical Significance

We developed a randomization procedure to test whether the regions surrounding origins of replication

display unusual features as compared with the rest of the genome. For the experimental data set, we randomly sample 283 positions in ENCODE regions by setting the number of sampled origins per ENCODE region equal to the number observed in the real data set. For the in silico data set, we divided the chromosomes in 1 Mb regions and we randomly sampled 1,060 positions by setting the number of sampled origins per 1 Mb region equal to the number observed in the real data set. With this randomization procedure, we account for different sequence characteristics (gene density, GC content, etc.) that may be specific to the regions where *OriExp* or *OriIS* origins were sampled.

For the analysis of replication–transcription coorientation, we developed an additional randomization procedure by constraining not only the number of origins per region but also the number of origins that are close to annotated TSS per region. The positions of the TSS were taken from Ensembl 50 and were restricted to transcripts that are present in RefSeq.

Results

We analyzed two data sets of human origins of replication, one experimentally determined (Cadoret et al. 2008) and one predicted in silico (Huvet et al. 2007). Throughout the text, we will refer to the first data set as *OriExp* and to the second one as *OriIS*.

The *OriExp* data set consists of 283 origins of replication identified in HeLa cell cultures, using short nascent strands hybridization on DNA microarrays (Cadoret et al. 2008). The origins were specifically searched in ENCODE regions (The ENCODE Project Consortium 2007). Although the fraction of the genome covered by this data set is small (only 1%), conclusions drawn from this analysis are expected to be general as ENCODE regions were selected in order to provide a representative sample of the whole genome.

There are 44 ENCODE regions of sizes comprised between 500 and 2,000 kb. The GC content of the regions varies between 35% and 56%, with the median at 43%. The density of origins of replication displays a strong positive correlation with the GC content (Cadoret et al. 2008). The distance between two consecutive origins varies widely between 1 and >500 kb: whereas in 75% of the cases, the interorigin distance is lower than 81 kb, six 500 kb regions are devoid of origins (Cadoret et al. 2008).

The *OriIS* data set consists of 678 N-domains, bordered by 1,060 predicted origins of replication (Huvet et al. 2007). The N-domain size varies between 195 and 2,788 kb, with an average of 1,197 kb; the N-domains represent a total of ≈812 Mb or ≈27% of the genome. The average GC content of the N-domains varies between 35% and 58%, with the median at 40%.

The overlap between the genomic regions covered by the two data sets is disappointingly small: among the 1,060 *OriIS* origins, only seven fall within ENCODE regions. Out of these seven putative origins, two were at less than 1 kb from experimentally determined initiation sites, one was at ≈17 kb, and the remaining four were more than 50 kb away from *OriExp* origins (Cadoret et al. 2008). The colocalization between *OriIS* and *OriExp* origins is highly significant: in simulated data sets where *OriIS* positions are randomized, the probability to obtain an overlap for two out of seven predictions is $<10^{-3}$ (Cadoret et al. 2008). Although the low sample size precludes a definitive conclusion, this result indicates that a significant proportion of *OriIS* predictions are genuine replication origins.

Conversely, we can ask what proportion of *OriExp* positions were detected in silico. We found that 11 N-domains intersect with ENCODE regions, and a total of 35 origins were experimentally detected within the intersection. Out of these 35 origins, as stated above, only two are close to *OriIS* predictions. The fraction of *OriExp* positions that can be detected through analyses of nucleotide composition asymmetry appears thus to be relatively low.

Colocalization between Origins of Replication and Transcriptional Promoters

We first analyzed the extent of the colocalization between origins of replication and gene promoters. First discovered in mitochondria and virus genomes (Baldacci and Bernardi 1982; Clayton 1991), the initiation of DNA replication at transcriptional promoters seems to be a frequent feature in eukaryotes (DePamphilis 1993).

As expected, we find considerable overlap between promoters and origins. For *OriExp*, 28% of the origins are less than 2 kb away from TSS, as defined by Ensembl and RefSeq annotations. This fraction is even higher (42%) for *OriIS*. We wanted to verify if these figures depart significantly from random expectations and to what extent the difference between the two data sets can be attributed to different characteristics, such as gene density, of the genomic regions where the origins were sampled. To do so, we developed a randomization procedure that constrains the regional distribution of origins to be identical to the one observed in the real data sets (Material and Methods). We find that the difference between *OriExp* and *OriIS* cannot be explained by their localization in different regions of the genome, on the contrary: the expected frequency of overlap between origins and promoters is lower for *OriIS* than for *OriExp*, whereas the opposite is true for the observed values (fig. 1). The colocalization between origins and promoters is significantly higher than random expectations (P value $< 10^{-3}$) for both data sets (fig. 1; supplementary table 1, Supplementary Material online).

In human, 50% of genes are associated with promoter sequences with unusually high content of CpG dinucleotides that escape DNA methylation: the so-called CpG islands (CGI) (Antequera and Bird 1993). It was previously reported that CGI can function not only as transcriptional promoters but also as origins of replication (Delgado et al. 1998). In agreement with these results, we found that 53% of *OriExp* and 65% of the *OriIS* are close to CGI (distance ≤ 2 kb). These proportions are significantly higher than expected by chance (randomization test, P value $< 10^{-3}$). The colocalization with CGI is almost perfect for origins that are associated with transcriptional promoters: the overlap reaches 94% for the experimental data set (P value 5×10^{-2}) and 99% for *OriIS* (P value 3×10^{-3}). Similarly, origins that are associated with CGI are more often associated with gene promoters: 49% and 64% for the two data sets (P value $< 10^{-3}$).

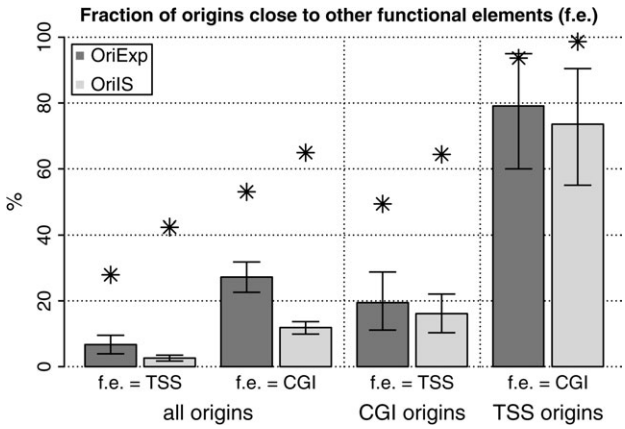

FIG. 1.—Distribution of origins of replication with respect to TSS and CGI. Origins are classified as TSS or CGI if they are at a distance ≤ 2 kb from a TSS or CGI, respectively. The positions of the TSS are taken from Ensembl annotations. The vertical bars represent the expected distribution (average and 2.5% and 97.5% quantiles) obtained through randomization. The star symbols represent the observed frequency (percentage) of the origin classes. Dark gray: experimental origins (*OriExp*). Light gray: in silico origins (*OriIS*). See also supplementary table 1 (Supplementary Material online).

Coorientation between Replication and Transcription

Initial analyses of gene distribution and orientation along the N-domains revealed that there is a tendency for coorientation between replication and transcription, especially in the vicinity of the putative origins (Huvet et al. 2007). The authors proposed that this mode of organization could be adaptive because it reduces the risk of deleterious frontal collisions between replication and transcription polymerases (Nomura and Morgan 1977; Brewer 1988).

We wanted to verify if the same tendency for coorientation can be detected using the experimental data set of origins of replication. To do so, we computed the fraction of transcribed nucleotides for which the sense strand for transcription corresponds to the leading strand for replication (fig. 2*A*, Materials and Methods); a fraction of 100% is expected if the coorientation between replication and transcription is perfect. Given that origins of replication are frequently associated with TSS (see above), we expect a leading strand fraction slightly higher than 50% because for genes that have an origin in the promoter region, the sense strand for transcription coincides with the leading strand for replication. To test the statistical significance of the observations while taking into account the colocalization between origins and promoters, we used a simulation procedure that constrains the number of promoter origins in each genomic region (Materials and Methods).

The leading strand fractions are strikingly different for the two data sets: 54.6% for *OriExp* and 90.8% for *OriIS* (fig. 2*B*; supplementary table 2, Supplementary Material online). For the experimental data set, the observed frequency is very

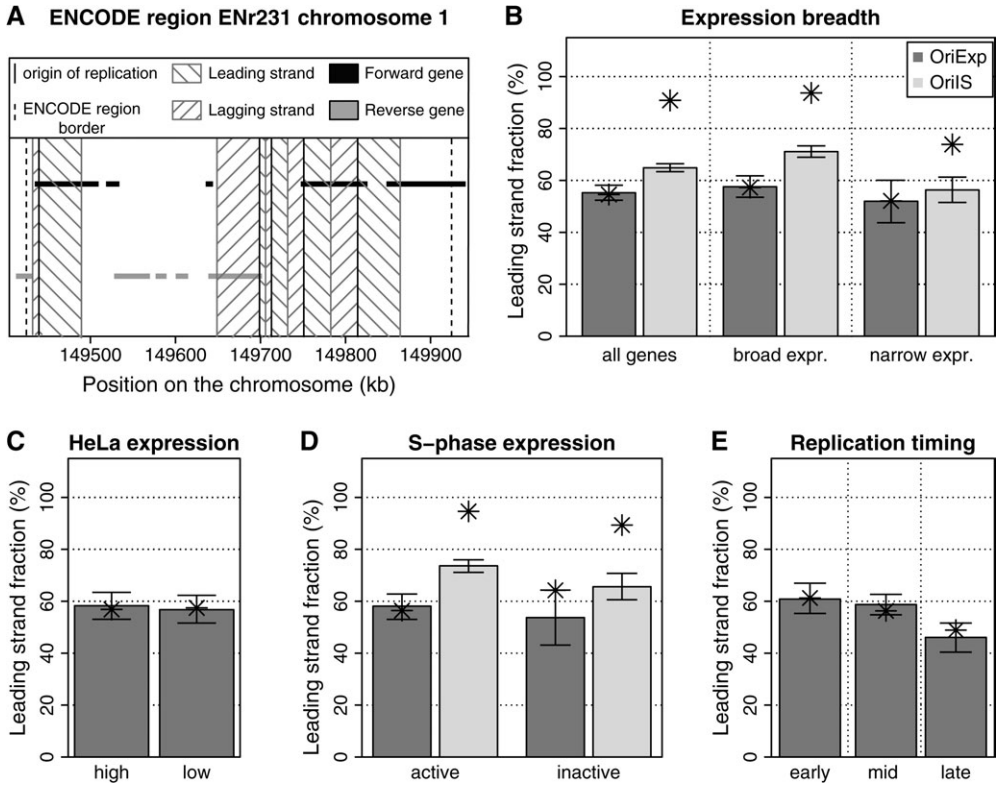

FIG. 2.—Coorientation between replication and transcription. (*A*) Example of leading and lagging strand definition for ENCODE region ENr231. (*B*) Fraction of transcribed nucleotides for which the sense strand for transcription is the leading strand for replication, as a function of the expression breadth (here defined with EST data). (*C*) The same, computed separately for genes with high and low levels of expression in HeLa cells. (*D*) The same, separately for genes estimated as active or inactive in the S-phase (only broadly expressed genes are considered for the two classes). (*E*) The same, separately for early-, mid-, and late-replicating regions. The vertical bars represent the expected distribution (average and 2.5% and 97.5% quantiles) obtained through randomization. The star symbols represent the observed frequency (percentage) of the origin classes. Dark gray: experimental origins (*OriExp*). Light gray: in silico origins (*OriIS*).

close to the average expected value, whereas for *OriIS*, the tendency for coorientation is highly significant (P value $< 10^{-3}$). The slight tendency for coorientation between replication and transcription observed for *OriExp* is therefore entirely explained by the colocalization between origins and promoter regions, whereas for *OriIS*, there is a clear orientation bias even for genes farther away from origins.

If the tendency for coorientation between replication and transcription were the result of a selective pressure to avoid frontal collisions between polymerases, when considering origins that are active in a given cell type, the coorientation should be stronger for genes that are expressed at high level in that cell type. As *OriExp* origins were determined in HeLa cells, we analyzed separately the orientation of genes expressed at high and low levels in this cell type (Materials and Methods). We find that the leading strand fraction is similar for the two expression classes (56.9% and 57.5%, respectively, χ^2 test, P value $= 0.72$). The observed values are not significantly higher than expected by chance (fig. 2*C*; supplementary table 2, Supplementary Material online).

It was previously noticed that the coorientation frequency for the in silico data set is higher for broadly expressed genes, that is, genes that are expressed in a wide number of tissues (Huvet et al. 2007). Using EST and SAGE expression data, we divided the genes in classes of broad and narrow expression (Materials and Methods). For *OriIS*, we confirm that the leading strand fraction is higher for broadly expressed genes (93.7% with EST data) than for genes with narrow expression (73.85% for EST data), but even for the latter class, it remains significantly higher than the random expectation (fig. 2*B*; supplementary table 2, Supplementary Material online). For *OriExp*, we also find a slight variation in the leading strand frequency with the expression breadth (57.3% for broad expression and 52.1% for narrow expression). However, for *OriExp*, the observed values are entirely explained by the colocalization between origins and promoters (fig. 2*B*). Similar results are obtained when using SAGE data to define the expression breadth (supplementary fig. 1 and supplementary table 2, Supplementary Material online).

As collisions between polymerases can only occur for genes that are transcribed during the S-phase of the cell cycle, when DNA replication occurs, we expect the tendency for coorientation to be present only for this class of genes. A recent paper analyzed variations in gene expression during the cell cycle for the tumoral cell line T98G (Litovchick et al. 2007). This analysis provided an estimation of genes that are likely to be activated or repressed at each phase of the cell cycle. We could therefore perform the above analysis separately for genes that are active or inactive during the S-phase. Noting that genes transcribed in the S-phase are often expressed in many tissues to remove this potential source of bias, we considered only broadly expressed genes. Surprisingly, we find that for *OriIS* the tendency for coorientation is not strongly affected by the S-phase expression status: the leading strand fraction is slightly higher for active than for inactive genes (94.6% and 89.3% with EST data), but both frequencies are significantly higher than expected (P value $< 10^{-3}$). Similar results are obtained using SAGE data (supplementary fig. 2 and supplementary table 2, Supplementary Material online). For *OriExp*, the observed values remain within the expected range for both active and inactive genes (56.6% and 64.2% with EST data).

Next, we wanted to verify whether the frequency of coorientation between replication and transcription might be affected by replication timing. Indeed, for *OriIS*, it was previously reported that many of the predicted origins are in relatively early-replicating regions of the genome (Huvet et al. 2007). Could it be that the high frequency of leading strand genes is specific to early-replicating zones? To test this hypothesis for *OriExp*, we used the high-resolution replication timing data provided by Karnani et al. (2007). This data set is particularly suited for the analysis of *OriExp* as it is specific to ENCODE regions and also determined for HeLa cells. Thus, we could analyze separately the leading strand proportion for genes in early-, mid-, and late-replicating regions (Materials and Methods). We find indeed that the leading strand frequency is slightly higher for early-replicating regions (61.25%) than for mid-replicating (56.3%) and for late-replicating regions (48.9%). This variation is, however, explained by different frequencies of origin–promoter colocalization in the three timing regions: the observed values are always within the range expected by chance. Moreover, even among early-replicating *OriExp*, the frequency of coorientation (61.5%) is much lower than the frequency observed for the entire *OriIS* data set (90.8%).

Variation in Expression Breadth for Neighbor Genes

Huvet et al. (2007) showed that the expression pattern of genes varies nonrandomly across the N-domains: broadly expressed genes are preferentially positioned near the putative origins of replication, whereas tissue-specific genes tend to occur near the center of the N-domains. Gene expression breadth thus appears to decrease with the distance from the origins of replication.

We wanted to verify if a similar variation in expression breadth is found for *OriExp*. To do so, for each gene present in ENCODE regions, we computed the minimum distance between its annotated TSS and origins of replication. Using EST data, we were able to estimate the expression breadth for 319 genes. We divided the genes into 10 classes of approximately equal size, according to the distance from origins, and we computed the average expression breadth for each class (fig. 3; supplementary table 3, Supplementary Material online). For *OriIS*, we did the same analysis by selecting genes that fitted into the 10 classes defined for *OriExp*. The variation in expression breadth as a function of the distance from the origins is presented in figure 3. For *OriIS*, we confirm that expression breadth decreases significantly with the distance from origins: a Kruskal–Wallis nonparametric test showed that the distance class has a significant effect on the average expression breadth (P value $< 2 \times 10^{-16}$). However, for *OriExp*, we find no evidence for such a relationship (P value 0.34, fig. 3; supplementary table 3, Supplementary Material online). Similar results were obtained when using SAGE data to define expression breadth (supplementary fig. 3 and table 3, Supplementary Material online). We must note that for *OriExp* the numbers

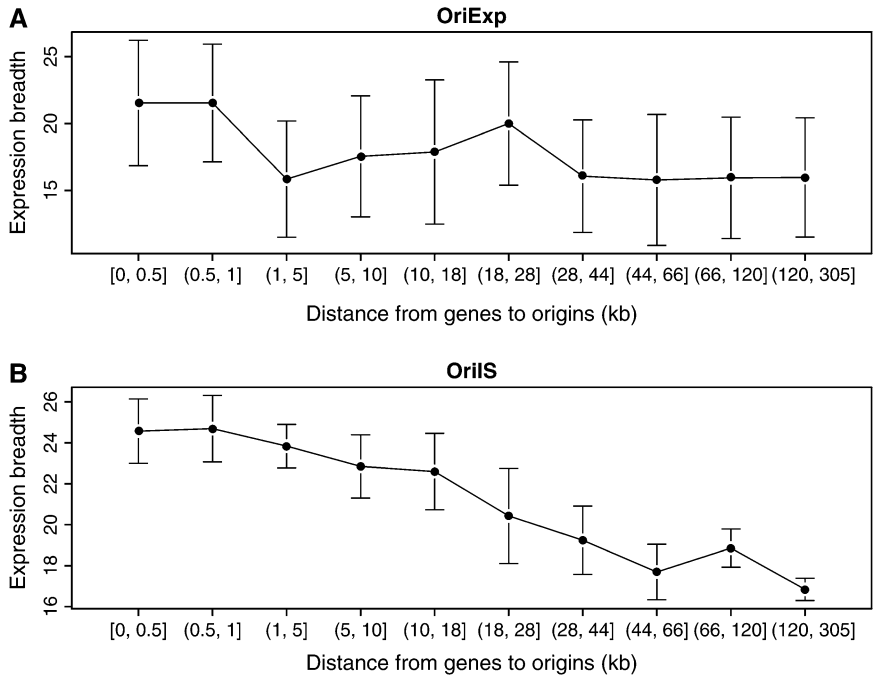

FIG. 3.—Variation in expression breadth as a function of the distance between genes and origins of replication. In abscissa, the distance class. Expression breadth (number of tissues) was estimated using EST data. The average expression breadth and the 95% confidence interval are represented for each distance class. See also supplementary table 3 (Supplementary Material online).

of genes in each class are small compared with *OriIS*; these results should therefore be taken with caution and are to be reassessed when larger data sets become available.

Evolutionary Conservation of Replication Origins

Clearly, origins of replication are among the most essential functional elements in a genome, and they are therefore expected to be conserved during evolution. Initial analyses of the experimentally determined origins of replication confirmed this expectation (Cadoret et al. 2008).

Here, we wanted to test if similar levels of sequence conservation are observed for the two data sets of origins of replication. To do so, we needed to define the start and end positions of the origins. For the origins determined by Cadoret et al. (2008), the borders of the origins are defined by the microarray hybridization method, but the positions of the computationally predicted origins are given as a single base pair, and we have no information regarding the size of the origin. To remove potential biases arising from this discrepancy, we homogenized the origin definition for the two data sets, by reducing the experimental origins to a single base pair, chosen as the center of the initially given segments. We then analyzed sequence conservation in 2 kb segments centered around the origin.

We defined the level of sequence conservation for origins of replication as the fraction covered by genomic regions that are highly conserved among mammals (Siepel et al. 2005). We observed that origins of replication are often found in proximity to other functional elements, such as TSS and CGI (see above). These sequences are expected to show high levels of sequence conservation because they contain regulatory elements required for transcription. To minimize the effect of this confounding factor, we divided the origins of replication into three classes according to their distance from TSS and CGI. As there are very few origins that are close to TSS but not to CGI, we removed these origins from the analysis.

As expected, the proximity to TSS and CGI increases the level of sequence conservation: origins of replication that are close to both TSS and CGI are the most conserved (fig. 4 and supplementary table 4, Supplementary Material online). We found that the experimentally determined origins display significantly higher levels of sequence conservation than expected by chance, independently of the distance to TSS and CGI. It is therefore probable that specific sequence elements, necessary for origin function (and not uniquely promoter function), are conserved even at wide evolutionary scales. Surprisingly, the computationally predicted origins generally show lower sequence conservation than the experimental ones and only differ significantly from random expectations when they are close to CGI or TSS. As the level of conservation for *OriIS* predictions that are far (>2 kb away) from both TSS and CGI is not higher than expected by chance, it is possible that for this particular class of predictions the detection method lacks the precision needed to correctly identify all functional elements.

Nucleotide Composition Asymmetry

The replication mechanism is asymmetric: the leading strand is replicated continuously, whereas the lagging strand is synthesized in a fragmented manner. This can lead to different mutation patterns for the leading and lagging strands: in Bacteria, the two strands have highly asymmetric nucleotide composition (Lobry 1996a), and a similar

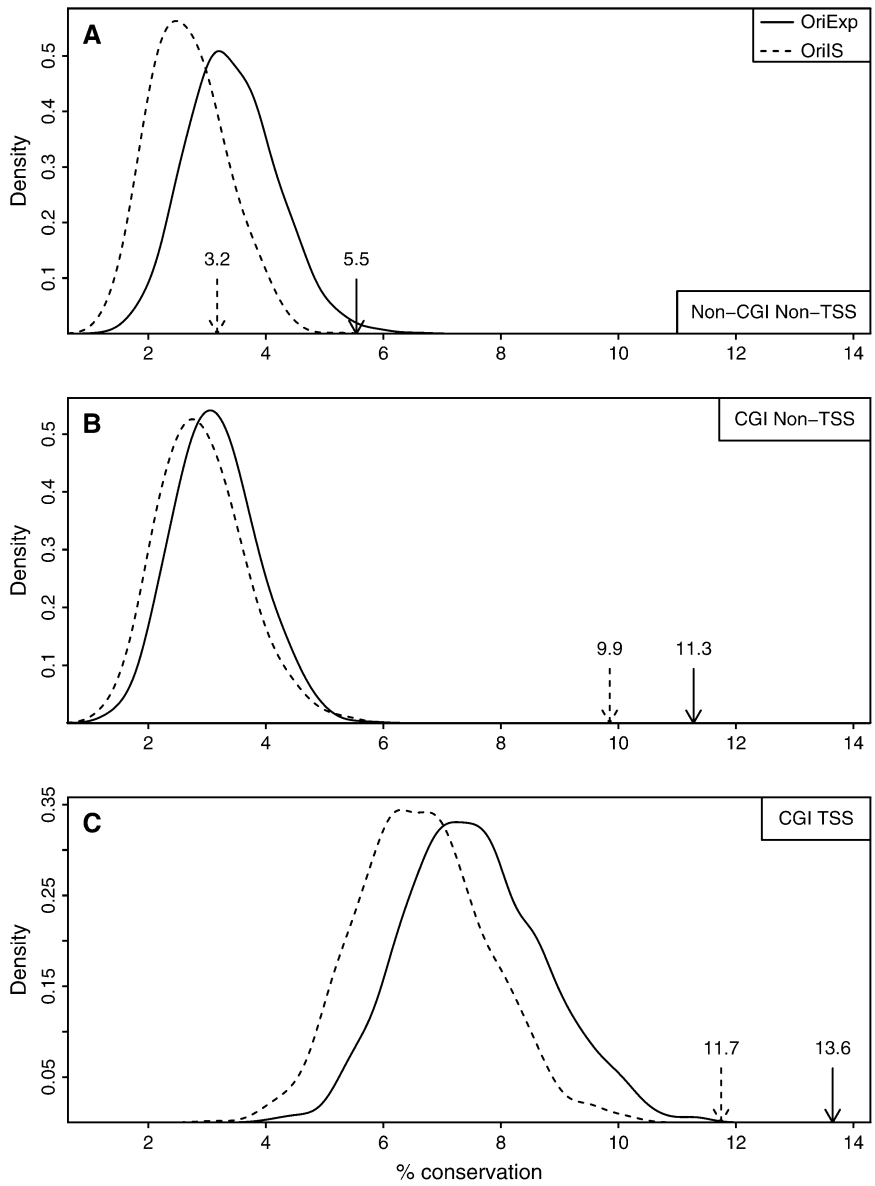

FIG. 4.—Sequence conservation for origins of replication, measured as the fraction covered by highly conserved sequences among mammalian genomes (Siepel et al. 2005). The curves represent the expected distribution obtained from 1,000 randomized data sets, the arrows represent the observed values. Solid line: *OriExp*, dashed line: *OriIS*. (*A*) Origins more than 2 kb away from CGI and annotated TSS. (*B*) Origins less than 2 kb away from CGI and more than 2 kb away from annotated TSS. (*C*) Origins less than 2 kb away from both CGI and annotated TSS.

pattern of asymmetry has been observed around six replication origins in human (Touchon et al. 2005). This property is at the basis of the in silico detection method.

We wanted to test whether experimental origins display the same pattern of nucleotide composition as do in silico predicted origins. To do so, we considered regions that could unambiguously be assigned as leading or lagging strand, separately for intergenic regions and introns (Materials and Methods). We then computed the global skew measure for leading and lagging strand segments defined by *OriExp* and *OriIS* origins. As expected, we find that leading and lagging strand segments defined by *OriIS* origins are characterized by opposite skew values for both intergenic regions and introns (fig. 5, supplementary fig. 4 and table 5, Supplementary Material online). For *OriExp*, the expected pattern is found only for intergenic regions: leading strand segments have positive skew values, whereas lagging strand segments have negative skews. For introns, the difference in skew between the two replication strands is negligible. The average difference between leading and lagging strands is much smaller for *OriExp* than for *OriIS*: $\Delta S \approx 0.48\%$ and 8.0%, respectively. The weakness of the composition asymmetry observed around experimental origins could explain why only a small proportion of *OriExp* origins were detected in silico.

Discussion

The goal of our analysis was to evaluate the association between DNA replication and human genome

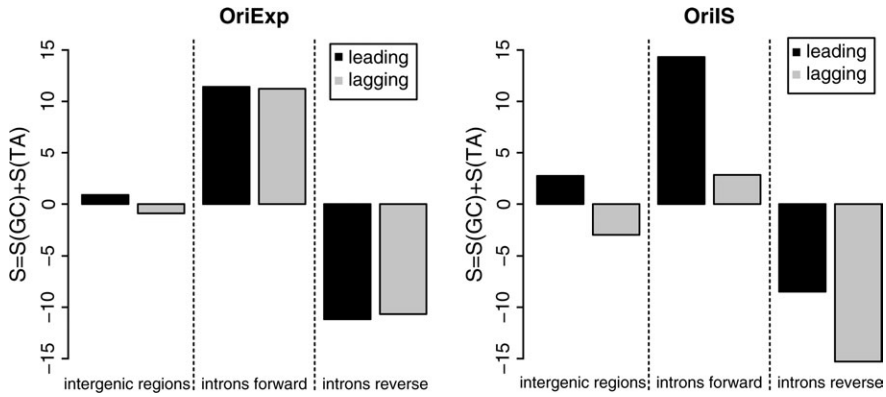

FIG. 5.—Global skew measure ($S = S_{GC} + S_{TA}$) computed on the leading and lagging strand for replication for *OriExp* and *OriIS*. The vertical bars represent the average value for each class of regions: intergenic, introns (forward and reverse orientations). Black: leading strand composition. Gray: lagging strand composition.

organization. That such an association exists is confirmed by our observation that origins of replication are not randomly distributed in human genome: we showed that there are significant tendencies for colocalization between origins, transcriptional promoters, and CGI. However, our analysis of gene distribution around experimentally determined origins does not bring support for an important structural feature that was previously inferred based on replication domains predicted in silico: the coorientation between replication and transcription. For *OriIS*, the tendency for coorientation is very strong: 90.8% of transcribed nucleotides are on the leading strand for replication. This leading strand fraction is significantly higher than expected by chance, even when taking into account the proportion of genes that have an origin in their promoter region. Thus, the transcription orientation bias appears to be present at a relatively wide scale and not just in the immediate proximity of the predicted origins. The situation is strikingly different for *OriExp*: the leading strand fraction is much lower than for *OriIS* (54.6%), and this slight tendency for coorientation is entirely explained by the colocalization between origins and promoter regions. There are several possible explanations for this discrepancy between *OriExp* and *OriIS*, discussed below.

Cell-Type Specificity of Origins of Replication

The prediction method used by Huvet et al. (2007) relies on the detection of regions with skewed nucleotide composition. This method can, therefore, only determine origins of replication that are active in the germ line or in early embryogenesis before the differentiation between germ line and somatic cells, as mutations that occur in the soma are not evolutionarily relevant and cannot generate the observed composition bias. On the other hand, the experimental origins data set determined by Cadoret et al. (2008) has been obtained for HeLa cell cultures. Could the cell-type specificity of origins of replication be the cause of the discrepancy observed between the two data sets?

The hypothesis proposed by Huvet et al. (2007) to explain the high frequency of genes coded on the leading strand was the existence of a selective pressure to avoid deleterious head-on collisions between polymerases. If this hypothesis is valid, a tendency for coorientation between replication and transcription is expected even in somatic tissues. However, for *OriExp* origins, we find no evidence for gene orientation bias, even when considering specifically genes that are expressed at high levels in HeLa cells. Moreover, under the polymerase collision hypothesis, broadly expressed genes are expected to be associated with constitutive origins, ensuring that the coorientation between replication and transcription is effective in all tissues. Whereas for *OriIS*, this prediction is in agreement with the data, for *OriExp*, we find no significant tendency for coorientation even when considering genes expressed in a wide number of tissues. Finally, this hypothesis allows us to make one more prediction: genes that are not expressed during the S-phase of the cell cycle, when the DNA sequence is replicated, are not expected to be preferentially coded on the leading strand. However, our results show that for *OriIS* the tendency for coorientation between replication and transcription is strong even for situations where polymerase collisions are not likely to occur. These latter results need to be taken with caution as they were based on expression data for one particular somatic cell line (T89G). Nevertheless, if they are confirmed for other cell types, we can infer that the gene strand bias observed for *OriIS* cannot be caused by the need to avoid frontal polymerase collisions.

If, despite these considerations, the gene orientation bias were indeed caused by a selective pressure against frontal polymerase collisions, its absence for *OriExp* can only be explained by assuming that the effects of such collisions are strongly deleterious only in the germ line or during early development. However, as we currently lack information on the consequences or even the occurrence of polymerase collisions in human, this reasoning is purely speculative.

Are In Silico Predictions Representative of All Germ Line Origins?

When analyzing the intersection between ENCODE regions and N-domains, we found that only a small proportion (2/35) of experimentally determined origins are also detected computationally. This result was perhaps expected, as Huvet et al. (2007) noted that density of origins predicted in silico is

much lower than previous estimations made for somatic cells. However, the magnitude of the difference between *OriExp* and *OriIS* raises the question of the sensitivity of the in silico prediction method. Indeed, this discrepancy is not likely to be explained only by the germ line/soma distinction: a priori, there is no reason to imagine that the number of origins active in the germ line should be more than 15 times smaller than in somatic cells. It was argued that the large interorigin distance predicted in silico is compatible with observations made for premeiotic replication (Huvet et al. 2007). The duration of DNA replication is indeed longer in meiosis than in mitosis for all organisms studied so far (Strich 2004). However, we must note that a longer premeiotic S-phase does not necessarily imply that fewer origins are used: in yeast, the same origins are used during mitotic and meiosis replication (Collins and Newlon 1994), only less efficiently in the latter (Heichinger et al. 2006). Moreover, the premeiotic S-phase represents only one particular step in the germ line, preceded by many mitotic divisions.

In addition, a clear indication of sensitivity of the computational method comes from the fact that in silico prediction was possible for only 28% of the genome. Thus, we can reasonably assume that the in silico method detected only a small subset of all germ line origins. One possible explanation for the disagreement between *OriExp* and *OriIS* follows from this reasoning: the subset detected in silico may not be representative for all origins active in the germ line.

Efficiency of Origin Activation

It has long been known that the efficiency of activation of eukaryotic origins of replication is highly variable: "strong" origins are active during a high proportion of all S-phases in a given cell population, whereas "weak" origins initiate replication sporadically (Fangman and Brewer 1991). Could the discrepancy between *OriExp* and *OriIS* be explained by different origin strength? Indeed, it can be imagined that the biased gene distribution around replication origins is correlated positively with the efficiency of origin activation. It is also plausible that strong replication origins are more easily identified in silico because the nucleotide composition asymmetry should be more pronounced than for weak origins.

However, we must note that the experimental data set is likely to contain a significant proportion of strong origins. Cadoret et al. (2008) performed a quantitative analysis for 29 randomly sampled origins in the data set and compared the intensity of their detection signal with the one obtained for one previously known origin ("c-myc," also identified in silico). Out of these 29 origins, 16 presented a signal intensity comparable with or higher than that of c-myc. This analysis strongly suggests that differential origin activity cannot be the only cause for disagreement with the results presented by Huvet et al. (2007).

Replication Timing

Huvet et al. (2007) observed that, for a number of cases, the predicted origins correspond to regions of the genome that replicate relatively early in the S-phase. Moreover, replication timing is known to be correlated to several structural features of mammalian genomes, such as density of genes and CGI or GC content; the association between early replication and gene expression is also well documented (Woodfine et al. 2004; Karnani et al. 2007; Farkash-Amar et al. 2008).

We must therefore ask if replication timing could be a confounding factor for the pattern of genome organization observed around *OriIS* origins. Using recent, high-resolution timing data for ENCODE regions (Karnani et al. 2007), we were able to conclude that the frequency of coorientation between replication and transcription is not significantly affected by the replication timing: even when considering only origins activated early in the S-phase, the percentage of genes coded on the leading strand is not higher than expected by chance. Therefore, experimental data do not support the existence of an association between replication timing and transcription orientation. Due to the small size of the data set, we were unable to test whether replication timing has an influence on the variation in expression breadth with the distance from origins; further studies are therefore needed, when whole-genome data for origins of replication become available.

Are Computational Predictions Genuine Origins?

The association between replication and transcription is clearly stronger for the *OriIS* data set than for *OriExp*. We showed that the extent of the colocalization between origins of replication and transcriptional promoters is more important for *OriIS* than for *OriExp*, although the opposite is expected by chance, given the different characteristics of the genomic regions where the origins were sampled. Moreover, in the proximity of in silico predicted origins, 90.8% of all transcribed nucleotides are on the leading strand for replication, whereas no strong orientation bias is observed for *OriExp*. Finally, for *OriIS*, gene expression breadth decreases with the distance from the predicted origins, whereas for *OriExp*, this pattern is not found. One possible interpretation of the observations made for *OriIS*, as proposed by Huvet et al. (2007), is that replication is a determinant of gene organization in the human genome. An alternative explanation, and one that can elucidate the discrepancies between *OriIS* and *OriExp*, is that the presence of transcription may enhance (or in some cases even mislead) the computational detection of replication origins.

The computational method used by Huvet et al. (2007) relies on the identification of regions with linearly decreasing skew, starting from positive values in 5′ and ending at negative values in 3′. The identification of this nucleotide composition pattern with a replication domain is reasonably justified by the proposed model of fixed replication initiation and random termination (Touchon et al. 2005; Huvet et al. 2007). However, we must note that searching for this nucleotide composition pattern appears to result in an important fraction of false positives: on shuffled chromosomes obtained after randomly permuting genes and intergenic regions, the number of N-domains reaches 23% of the number detected on the actual chromosomes (Huvet et al. 2007).

Considerations on the specificity of the detection method put aside, we argue that this pattern of asymmetric nucleotide composition can also be caused by transcription

and not just by replication. Indeed, it is well known that transcription and replication produce similar composition biases (Green et al. 2003; Touchon et al. 2003), and the separation of these two sources of nucleotide composition asymmetry is often difficult. Moreover, an analysis at the whole-genome scale previously showed that gene expression breadth is positively correlated with the nucleotide skew (Duret 2002). Thus, the variation in gene expression breadth along the N-domains observed by Huvet et al. (2007) may in fact be a direct cause of the linearly decreasing skew pattern, instead of an organizational feature dictated by DNA replication. A sensible counterargument to this reasoning is the observation that the linearly decreasing skew was also observed on regions annotated as intergenic and that transcription is unlikely to be the cause of asymmetric composition for these regions (Touchon et al. 2005; Huvet et al. 2007). However, recent findings suggest that most of the human genome is transcribed, even regions previously annotated as intergenic (The ENCODE Project Consortium 2007). Another possible counterargument is the fact that the linearly decreasing skew is also found along a single gene. Nevertheless, transcription can cause this observation: It has been recently shown that in human cells transcription initiates at most gene promoters, whereas full transcript elongation occurs in a smaller fraction of the genes (Guenther et al. 2007). The per-base transcription rate is therefore likely to decrease from the TSS to the termination site, which can also induce a decreasing skew along the gene length.

Under this alternative scenario, the borders of the N-domains may correspond, at least in a number of cases, to transcriptional promoters rather than bona fide origins of replication. This hypothesis is in agreement with the colocalization between predicted origins and TSS as well as with the strong gene orientation bias found for *OriIS*. We also remark that the theory initially proposed by Huvet et al. (2007) to explain the decrease in expression breadth with the distance from the borders of the N-domains remains plausible: if *OriIS* predictions are strong transcriptional promoters, they may be associated with an open chromatin structure in most tissues, and this conformation may partially extend to neighbor genes, affecting their expression pattern (Semon and Duret 2006).

Finally, the hypothesis that in silico predictions may correspond to transcriptional promoters is perfectly compatible with the fact that a significant proportion of *OriIS* positions are true replication origins. Indeed, we noted that there is a significant overlap between the two data sets: out of the seven predicted origins found in ENCODE regions, two were confirmed experimentally. If the positions of the seven origins were randomly drawn from ENCODE regions, such a colocalization is expected with a P value $< 10^{-3}$. However, we also observed that the two confirmed origins were in close proximity to annotated TSS and to CGI. In ENCODE regions, 27.8% of CGI promoters coincide with *OriExp* origins. If the seven *OriIS* positions were drawn among CGI promoters, an overlap with *OriExp* for more than two out of seven origins is expected with a P value of 0.5 (supplementary fig. 5, Supplementary Material online). Although more data are needed to draw a clear conclusion, the colocalization with CGI promoters might be a confounding factor for the overlap between *OriExp* and *OriIS*.

No Evidence for Replication-Related Genome Organization in Human

In silico analyses of nucleotide skews have undoubtedly provided valuable information regarding human genome architecture. Nevertheless, we believe that, due to potential confounding factors cited above, using computational predictions as a substitute for experimentally determined origins of replication might provide only a partial image of human genome organization. Experimental analyses of replication initiation, at the genome-wide level and in different cell types, are still necessary to fully understand the impact of this fundamental mechanism on human genome structure.

In the model of genome architecture proposed by Huvet et al. (2007), DNA replication plays a substantial part. Having evaluated independently the validity of this model, we are also in favor of a significant association between DNA replication and genome organization but much weaker than previously proposed. In confirmation of previous findings, we showed that the distribution of replication origins in the human genome is not random with respect to functional elements such as CGI and transcriptional promoters. However, the strong relationship between replication and transcription described by Huvet et al. (2007), and similar to the one encountered in bacteria (Rocha 2004), is not supported by our data. Notably, there is no evidence for a selective pressure to avoid collisions between replication and transcription machineries. As far as the effect of DNA replication on genome structure is concerned, we must for now conclude that what is true for *Escherichia coli* is probably not true for human and probably not true for the elephant either.

Supplementary Material

Supplementary figures 1–5 and tables 1–5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Marie Sémon for her help with expression data retrieval and analysis and Subhashini Sadasivam for advice on the Affymetrix data set. This work was supported by Agence Nationale de la Recherche (GIP ANR JC05_49162) and by the Centre National de la Recherche Scientifique. We thank the IN2P3 Computing Center for providing the computer resources.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Antequera F, Bird A. 1993. Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci USA. 90:11995–11999.

Ardell DH, Kirsebom LA. 2005. The genomic pattern of tDNA operon expression in E. coli. PLoS Comput Biol. 1:e12.
Baldacci G, Bernardi G. 1982. Replication origins are associated with transcription initiation sequences in the mitochondrial genome of yeast. EMBO J. 1:987–994.
Benson D, Karsch-Mizrachi I, Ostell DLJ, Wheeler DL. 2008. GenBank. Nucleic Acids Res. 36:D25–D30.
Boon K, Osorio E, Greenhut SF, et al. (11 co-authors). 2002. An anatomy of normal and malignant gene expression. Proc Natl Acad Sci USA. 99:11287–11292.
Brewer BJ. 1988. When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. Cell. 53:679–686.
Cadoret J-C, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H, 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.
Clayton D. 1991. Replication and transcription of vertebrate mitochondrial DNA. Annu Rev Cell Biol. 7:453–478.
Collins I, Newlon CS. 1994. Chromosomal DNA replication initiates at the same origins in meiosis and mitosis. Mol Cell Biol. 14:3524–3534.
Couturier E, Rocha E. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol. 59: 1506–1518.
Delgado S, Gomez M, Bird A, Antequera F. 1998. Initiation of DNA replication at CpG islands in mammalian chromosomes. EMBO J. 17:2426–2435.
DePamphilis M. 1993. Eukaryotic DNA replication: anatomy of an origin. Annu Rev Biochem. 62:29–63.
Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 12:640–649.
Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30:207–210.
Fangman WL, Brewer BJ. 1991. Activation of replication origins within yeast chromosomes. Annu Rev Cell Biol. 7:375–402.
Farkash-Amar S, Lipson D, Polten A, Goren A, Helmstetter C, Yakhini Z, Simon I. 2008. Global organization of replication time zones of the mouse genome. Genome Res. 18:1562–1570.
Giardine B, Riemer C, Hardison R, et al. (13 co-authors). 2005. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 15:1451–1455.
Gierlik A, Kowalczuk M, Mackiewicz P, Dudek MR, Cebrat S. 2000. Is there replication-associated mutational pressure in the *Saccharomyces cerevisiae* genome? J Theor Biol. 202: 305–314.
Green P, Ewing B, Miller W, Thomas P, Green E. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nature Genet. 33:514–517.
Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. Cell. 130:77–88.
Heichinger C, Penkett CJ, Bohler J, Nurse P. 2006. Genome-wide characterization of fission yeast DNA replication origins. EMBO J. 25:5171–5179.
Hubbard TJP, Aken BL, Beal K, et al. (58 co-authors). 2007. Ensembl 2007. Nucleic Acids Res. 35:D610–D617.
Huberman JA, Riggs AD. 1968. On the mechanism of DNA replication in mammalian chromosomes. J Mol Biol. 32: 327–334.
Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton Carafa Y, Arneodo A, Thermes C. 2007. Human gene organization driven by the coordination of replication and transcription. Genome Res. 17:1278–1285.
Karnani N, Taylor C, Malhotra A, Dutta A. 2007. Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. Genome Res. 17:865–876.
Karolchik D, Baertsch R, Diekhans M, et al. (13 co-authors). 2003. The UCSC Genome Browser Database. Nucleic Acids Res. 31:51–54.
Litovchick L, Sadasivam S, Florens L, Zhu X, et al. (10 co-authors). 2007. Evolutionarily conserved multisubunit RBL2/p130 and E2F4 protein complex represses human cell cycle-dependent genes in quiescence. Mol Cell. 26:539–551.
Lobry J. 1996a. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol. 13:660–665.
Lobry J. 1996b. Origin of replication of *Mycoplasma genitalium*. Science. 272:745–746.
McLean MJ, Wolfe KH, Devine KM. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol. 47:691–696.
Mrazek J, Karlin S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci USA. 95:3720–3725.
Nieduszynski CA, Knox Y, Donaldson AD. 2006. Genome-wide identification of replication origins in yeast by comparative genomics. Genes Dev. 20:1874–1879.
Nomura M, Morgan EA. 1977. Genetics of bacterial ribosomes. Ann Rev Genet. 11:297–347.
Ponger L, Duret L, Mouchiroud D. 2001. Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res. 11:1854–1860.
Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:D61–D65.
Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, Conway A, Lockhart DJ, Davis RW, Brewer BJ, Fangman WL. 2001. Replication dynamics of the yeast genome. Science. 294:115–121.
Rocha EPC. 2004. The replication-related organization of bacterial genomes. Microbiology. 150:1609–1627.
Rocha EPC, Danchin A. 2003. Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res. 31:6570–6577.
Scotto L, Narayan G, Nandula SV, et al. (11 co-authors). 2008. Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. Gene Chromosomes Cancer. 47:755–765.
Semon M, Duret L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol. 23:1715–1723.
Siepel A, Bejerano G, Pedersen JS, et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.
Smit AFA, Hubley R, Green P. 1996–2004. RepeatMasker Open-3.0 [Internet]. [Accessed October 2008]. Available from: http://www.repeatmasker.org
Sousa C, de Lorenzo V, Cebolla A. 1997. Modulation of gene expression through chromosomal positioning in *Escherichia coli*. Microbiol. 143:2071–2078.
Strich R. 2004. Meiotic DNA replication. Curr Top Dev Biol. 61:29–60.
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.
Touchon M, Nicolay S, Arneodo A, d'Aubenton Carafa Y, Thermes C. 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. FEBS Lett. 555: 579–582.

Touchon M, Nicolay S, Audit B, Brodie of Brodie E-B, d'Aubenton Carafa Y, Arneodo A, Thermes C. 2005. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. Proc Natl Acad Sci USA. 102:9836–9841.

Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, Young BD, Debernardi S, Mott R, Dunham I, Carter NP. 2004. Replication timing of the human genome. Hum Mol Genet. 13:191–202.

Wyrick JJ, Aparicio JG, Chen T, Barnett JD, Jennings EG, Young RA, Bell SP, Aparicio OM. 2001. Genome-wide distribution of ORC and MCM proteins in *S. cerevisiae*: high-resolution mapping of replication origins. Science. 294: 2357–2361.

Zhu J, Newlon CS, Huberman JA. 1992. Localization of a DNA replication origin and termination zone on chromosome III of *Saccharomyces cerevisiae*. Mol Cell Biol. 12:4733–4741.

Aoife McLysaght, Associate Editor

Accepted December 22, 2008