DNA Replication Timing Data Corroborate *In Silico* Human Replication Origin Predictions

B. Audit,[1] S. Nicolay,[1,*] M. Huvet,[2] M. Touchon,[3,4] Y. d'Aubenton-Carafa,[2] C. Thermes,[2] and A. Arneodo[1]

[1]*Laboratoire Joliot-Curie and Laboratoire de Physique, ENS-Lyon, CNRS, 46 Allée d'Italie, 69364 Lyon Cedex 07, France*
[2]*Centre de Génétique Moléculaire, CNRS, Allée de la Terrasse, 91198 Gif-sur-Yvette, France*
[3]*Génétique des Génomes Bactériens, Institut Pasteur, CNRS, Paris, France*
[4]*Atelier de Bioinformatique, Université Pierre et Marie Curie - Paris 6, Paris, France*

(Received 31 May 2007; published 12 December 2007)

We develop a wavelet-based multiscale pattern recognition methodology to disentangle the replication- from the transcription-associated compositional strand asymmetries observed in the human genome. Comparing replication skew profiles to recent high-resolution replication timing data reveals that most of the putative replication origins that border the so-identified replication domains are replicated earlier than their surroundings whereas the central regions replicate late in the *S* phase. We discuss the implications of this first experimental confirmation of these replication origin predictions that are likely to be early replicating and active in most tissues.

DOI: 10.1103/PhysRevLett.99.248102 PACS numbers: 87.14.Gg, 87.10.+e, 87.16.Nr, 87.80.Tq

DNA replication is a crucial cellular process responsible for robust and accurate transmission of genetic information through successive cell generations. Small circular genomes of bacteria and viruses replicate as a single replication unit via the bidirectional progression along the chromosome of two replication forks originating from a single origin of replication (*ori*) and that ultimately collide at a unique terminus [1]. In larger genomes, initiation takes place at multiple *oris* up to several thousands in higher eukaryotic cells [2]. Unlike in *Saccharomyces cerevisiae* where *oris* are well annotated thanks to the existence of a consensus autonomously replicating sequence [3], there is no clear consensus sequence in multicellular organisms where initiation may occur at multiple sites distributed over thousand of base pairs [2]. To explain this lack of sequence specificity, epigenetic mechanisms at the level of the chromatin organization have been speculated as taking part in the spatial and temporal control of the initiation of replication in relation with gene expression [4]. To which extent these mechanisms can account for the experimental lack of well identified *oris* in multicellular eukaryotes, namely, around 20 in metazoa and only about 10 in human [2], is the subject of increasing current interest. In that context, the recent observation [5] that *oris* so far identified in the human genome correspond to large amplitude upward jumps in the noisy skew (nucleotide compositional strand asymmetry) profiles, has led to the detection of more than 1000 similar upward jumps as good candidates for putative replication initiation zones. Here, our goal is to make a first step towards the experimental confirmation of these putative *oris* by mapping them on recent high-resolution replication timing data [6] and by checking that these regions replicate earlier than their surrounding.

During the duplication of eukaryotic genomes, that occurs during the *S* phase of the cell cycle, the different *oris* are not all activated simultaneously [2,6]. Recent technical developments in genomic clone microarrays have led to a novel way of detecting the temporal order of DNA replication [6]. The arrays are used to estimate *replication timing ratios*, i.e., ratios between the average amount of DNA in the *S* phase at a locus along the genome and the usual amount of DNA present in the *G*1 phase for that locus. These ratios should vary between 2 (throughout the *S* phase, the amount of DNA for the earliest replicating regions is twice the amount during *G*1 phase) and 1 (the latest replicating regions are not duplicated until the end of *S* phase). This approach has been successfully used to generate genome-wide maps of replication timing for *S. cerevisiae* [7], *Drosophila melanogaster* [8] and human [9]. Very recently, two new analyses of human chromosome 6 [6(a)] and 22 [6(b)] have improved replication timing resolution from 1 Mbp down to ~100 kbp using arrays of overlapping tile path clones. Chromosome 22 being rather atypical in gene and GC contents, here we will focus on timing data from chromosome 6 [6(a)] which is more representative of the whole human genome. Data for clones completely included in another clone will be removed after checking for timing ratio value consistency leaving 1648 data points. The timing ratio value at each point will be chosen as the median over the 4 closest data points to remove noisy fluctuations resulting from clone heterogeneity (clone length 100 ± 51 kbp and distance between successive clone midpoints 104 ± 89 kbp), so that the spatial resolution is rather inhomogeneous ~300 kbp. Note that using asynchronous cells also results in some smoothing of the data, possibly masking local maxima.

In this study, we aim to measure correlations between replication timing ratios and nucleotide compositional skew. To provide a convincing experimental test for the set of putative *oris* identified in a previous work [5], we need as a prerequisite to extract the contribution to the compositional skew specific to replication. We will take advantage of the mammalian replicon model introduced in

Ref. [5] to develop a wavelet-based multiscale pattern recognition methodology that will allow us to disentangle replication- from transcription-associated strand asymmetries. This model with well-positioned *oris* and randomly distributed terminations, imposes serrated factory roof profiles for the nucleotide compositional strand asymmetry $S = (G - C)/(G + C) + (T - A)/(T + A)$. But, the overall observed skew S also contains some contribution induced by transcription that generates steplike blocks corresponding to sense (+) and antisense (−) genes [10]. Hence, when superposing the replication serrated and transcription steplike skew profiles, one gets the following theoretical skew profile in a replication domain:

$$S(x') = S_R(x') + S_T(x') = -2\delta(x' - 1/2) + \sum_{\text{gene}} c_g \chi_g(x'), \tag{1}$$

where position x' within the domain has been rescaled between 0 and 1, $\delta > 0$ is the replication bias, χ_g is the characteristic function for the g^{th} gene (1 when x' points within the gene and 0 elsewhere) and c_g is its transcriptional bias calculated on the Watson strand (likely to be positive for + genes and negative for − genes). The objective is thus to detect human replication domains by delineating, in the noisy S profile obtained at 1 kbp resolution [Fig. 1(a)], all chromosomal loci where S is well fitted by the theoretical skew profile (1). Since mammalian replicon size is known to be highly variable from a few hundred kbps up to several Mbps [2(a)], one thus needs a tool adapted to multiscale pattern recognition like the mathematical microscope provided by the continuous wavelet transform (WT) [11]. The WT is a space-scale expansion of a signal or S profile in terms of wavelets $\psi_{b,a}(x) = \frac{1}{\sqrt{a}}\psi(\frac{x-b}{a})$ that are constructed from a single function, the analyzing wavelet ψ, by means of translations (b) and dilations ($a > 0$):

$$T_\psi[S](b, a) = \int S(x)\psi_{b,a}(x)dx. \tag{2}$$

The wavelet coefficient $T_\psi[S](b, a)$ quantifies to which extent, around position b over a distance a, S has a similar shape as the analyzing wavelet ψ. In other words, by looking for the maxima of $T_\psi[S](b, a)$ over the space-scale half plane, the WT can be used as a multiscale shape detector. Thus, using an adapted analyzing wavelet constituted by a linearly decreasing segment between two upward jumps ($\psi(x) = -(x - 1/2)$ for $x \in [0, 1]$ and 0 elsewhere), the WT will provide a very efficient segmentation strategy of the human genome into candidate replication domains where the skew S displays a characteristic factory roof pattern [Fig. 1(a)]. In order to enforce strong compatibility with the mammalian replicon model [5], we will only retain the domains the most likely to be bordered by putative *oris*, namely, those that are delimited by upward jumps corresponding to a transition from a negative S value $< -3\%$ to a positive S value $> +3\%$. Also, for each domain so-identified, we will use a least-square fitting procedure to estimate the replication bias δ, and each of the gene transcription bias c_g. The resulting χ^2 value will then be used to select the candidate domains where the noisy S profile is well described by Eq. (1). As illustrated in Fig. 1 for a fragment of human chromosome 6 that contains 4 adjacent replication domains [Fig. 1(a)], this method provides a very efficient way of disentangling the steplike transcription skew component [Fig. 1(b)] from the serrated component induced by replication [Fig. 1(c)]. Applying this procedure to the 22 human autosomes, we delineated 678 replication domains of mean length $\langle L \rangle = 1.2 \pm 0.6$ Mbp, spanning 28.3% of the genome and predicted 1060 *oris*. Chromosome 6 of interest here, contains 54 such domains bordered by 83 putative *oris* among which 25 are common to two adjacent domains.

In Fig. 1(c), on top of the replication skew profile S_R, are reported for comparison the high-resolution timing ratio t_r

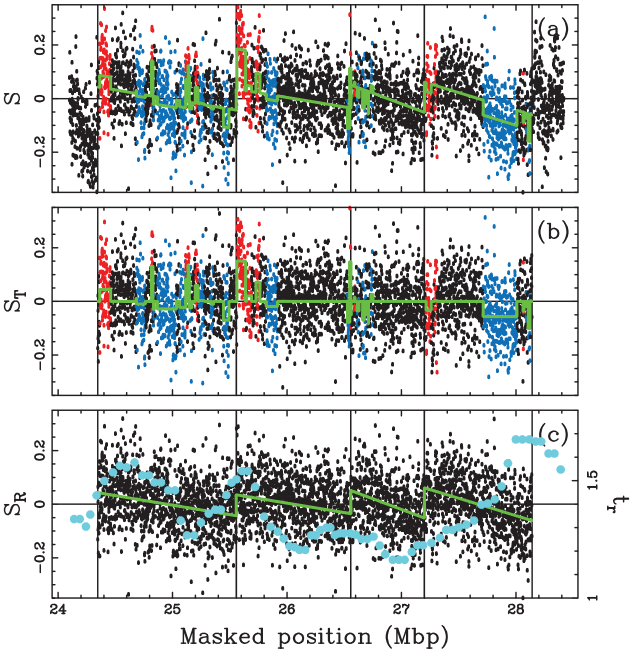

FIG. 1 (color). (a) Skew profile S of a 4.3 Mbp repeat-masked fragment of human chromosome 6; each point corresponds to a 1 kbp window: red, sense (+) genes; blue, antisense (−) genes; black, intergenic regions (the color was defined by majority rule); the estimated skew profile [Eq. (1)] is shown in green; vertical lines correspond to the locations of 5 putative *oris* that delimit 4 adjacent domains identified by the wavelet-based methodology. (b) Transcription-associated skew S_T obtained by subtracting the estimated replication-associated profile [green lines in (c)] from the original S profile in (a); the estimated transcription steplike profile [second term on the right-hand side (rhs) of Eq. (1)] is shown in green. (c) Replication-associated skew S_R obtained by subtracting the estimated transcription steplike profile [green lines in (b)] from the original S profile in (a); the estimated replication serrated profile [first term in the rhs of Eq. (1)] is shown in green; the light-blue dots correspond to high-resolution t_r data.

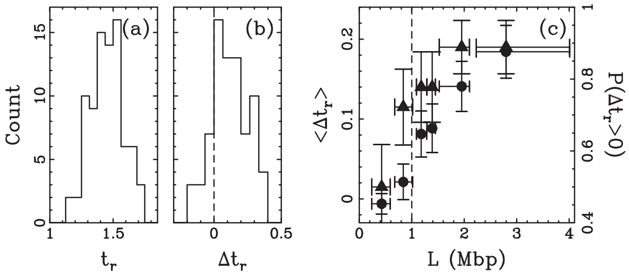

FIG. 2. (a) Histogram $N(t_r)$ of the replication timing ratio t_r at the 83 putative *oris*. (b) Histogram $N(\Delta t_r)$ of the difference Δt_r in replication timing ratio between the borders and the center of the 38 replication domains of length $L \geq 1$ Mbp. (c) Percentage of border-center pairs such that $\Delta t_r > 0$ (▲) and mean value $\langle \Delta t_r \rangle$ (●) versus the domain length L; domains have been ordered by length and binned into 6 groups of border-center pairs.

data from Ref. [6(a)]. The histogram of t_r values obtained at the 83 putative *ori* locations is shown in Fig. 2(a). It displays a maximum at $t_r \simeq \langle t_r \rangle \simeq 1.5$ and confirms what is observed in Fig. 1(c), namely, that a majority of the predicted *oris* are rather early replicating with $t_r \gtrsim 1.4$. This contrasts with the rather low t_r ($\simeq 1.2$) values observed in domain central regions [Fig. 1(c)]. In Fig. 2(c) are reported the results of a statistical analysis of the difference Δt_r of replication timing ratios between borders and domain centers versus the native length L of the domain. The average $\langle \Delta t_r \rangle$ steadily increases from 0 to 0.2 for the largest domains. Actually, more that 75% of the border-center pairs are such that $\Delta t_r > 0$ when the total domain length $L \geq 1$ Mbp, i.e., when the border-center distance is significantly larger than the timing data resolution. The overall histogram of Δt_r values obtained for these 76 border-center pairs is shown is Fig. 2(b). It mainly expands over positive Δt_r values with a right tail reaching $\Delta t_r = 0.4$ as a confirmation that while a majority of putative *oris* at the borders of the replication domains replicate rather early in the S phase, the central part of these domains systematically corresponds to regions that replicate quite late.

But there is an even more striking feature in the replication timing profile in Fig. 1(c): 4 among the 5 predicted *oris* correspond, relatively to the experimental resolution, to local maxima of the t_r profile. As shown in Fig. 3(a), the average t_r profile around the 83 putative *oris* decreases regularly on both sides of the *oris* over a few (4–6) hundreds kbp confirming statistically that domain borders replicate earlier than their left and right surroundings which is consistent with these regions being true *oris* mostly active early in S phase. In fact, when averaging over the top 20 *oris* with a well defined local maximum in the t_r profile, $\langle t_r \rangle$ displays a faster decrease on both sides of the *ori* and a higher maximum value ~1.55 corresponding to the earliest replicating *oris*. On the opposite, when averaging t_r profiles over the top 10 late replicating *oris*, we get, as expected, a rather flat mean profile ($t_r \sim 1.2$) [Fig. 3(a)]. Interestingly, these *oris* are located in rather wide regions of very low GC content (≲34%, not shown) correlating with chromosomal G banding patterns predominantly composed of GC-poor isochores [12]. This illustrates how the statistical contribution of rather flat profiles observed around late replicating *oris* may significantly affect the overall mean t_r profile. Individual inspection of the 38 replication domains with $L \geq 1$ Mbp shows that, in those domains that are bordered by early-replicating *oris* ($t_r \gtrsim 1.4$–1.5), the replication timing ratio t_r and the absolute value of the replication skew $|S_R|$ turn out to be strongly correlated. This is quantified in Fig. 3(b) by the histogram of the Pearson's correlation coefficient values that is clearly shifted towards positive values with a maximum at ~0.4. As illustrated in Fig. 4 for 5 of those replication domains, when using the phase portrait reconstruction technique from dynamical systems theory [13], one reveals a remarkable noisy cyclelike behavior when plotting t_r versus S_R [Figs. 4(a)–4(e)] and identifying the position from the domain 5' extremity to time. If, like in Fig. 4(d), one includes the left upward jump in S_R in the domain, one starts observing a fast variation of S_R from negative (~ −0.08) to positive (~ +0.08) values while $t_r \sim 1.5$ remains almost unchanged. Then, when entering the domain, S_R starts decreasing rather linearly while t_r decreases slowly leading to a somehow semicircular behavior of the trajectory in the (t_r, S_R) phase space; a minimum is actually reached in the domain central region where S_R (≃ 0) switches from positive to negative values and t_r takes its smallest value (~ 1.2–1.25), before increasing again while S_R keeps decreasing to negative values until the 3' extremity of the domain is reached. Let us emphasize that this noisy cyclelike behavior of relaxational

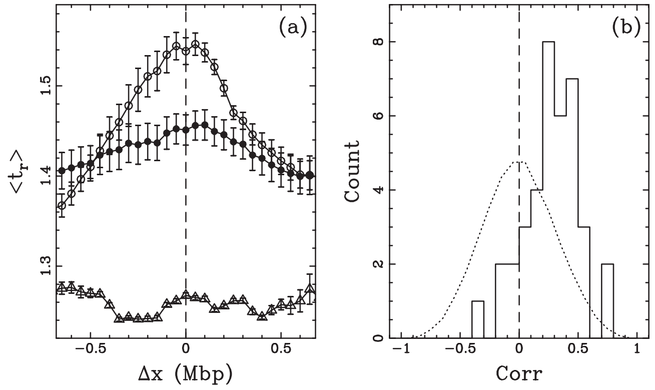

FIG. 3. (a) Average replication timing ratio (±SEM) determined around the 83 putative *oris* (●), 20 *oris* with well defined local maxima (○) and 10 late replicating *oris* (△). Δx is the native distance to the origins in Mbp units. (b) Histogram of the Pearson's correlation coefficient values between t_r and the absolute value of S_R over the 38 predicted domains of length $L \geq 1$ Mbp. The dotted line corresponds to the expected histogram computed with the correlation coefficients between t_r and $|S|$ profiles over independent windows randomly positioned along chromosome 6 and with the same length distribution as the 38 detected domains.

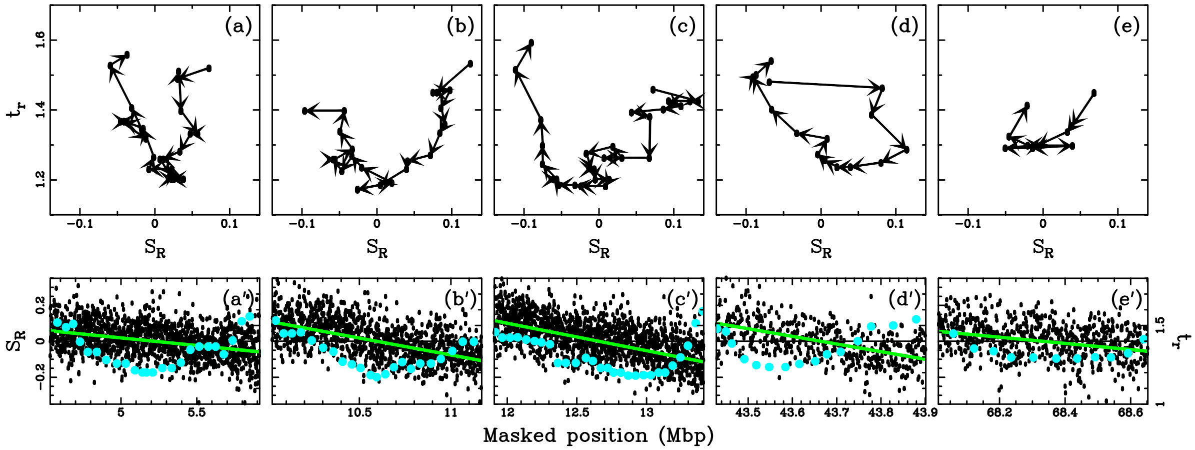

FIG. 4 (color online). Correlation between the replication-associated compositional asymmetry S_R and the replication timing ratio t_r along 5 replication domains bordered by rather early-replicating putative *oris*. (a)–(e) Phase portrait representation of t_r vs S_R; the arrows points in the 5'-3' direction. (a')–(e') Corresponding S_R (black points) and t_r (blue dots) profiles.

character is robustly observed in 16(/38) replication domains that are bordered by early-replicating putative *oris* [Figs. 4(a)–4(e)]. In 11(/38) domains the cycle is deformed due to the late replication of one of the borders.

To summarize, we have developed an adapted wavelet-based multiscale methodology to disentangle the replication from the transcription-associated nucleotide compositional skew. We have found that more than one quarter of the human genome is constituted by replication domains bordered by putative *oris*. Our analysis of recent experimental high-resolution replication timing data for human chromosome 6 confirms that these domain borders are certainly *oris* mostly active in the early *S* phase, whereas the central regions replicate more likely in the late *S* phase. This first experimental verification of *in silico ori* predictions is even more convincing when considering that, on top of the limitations due to experimental resolution, putative *oris* predictions concern only the *oris* that are well positioned and active in germ line cells which does not guarantee that they are also active in somatic cells and, in particular, in the lymphoblastoid cells used for the replication timing experiments [6(a)]. Reciprocally, there is no guarantee that the *oris* that are active in this particular cell line are also active in the germ line. In that respect, the results reported in this work strongly suggest that the predicted *oris* are functional, early-replicating initiation zones likely active in most tissues.

We thank O. Hyrien for helpful discussions. This work was supported by ACI IMPBio 2004, the Agence Nationale de la Recherche (No. NT05-3_41825), the PAI Tournesol and the Conseil Regional de l'Ile de France (M. T.).

*Permanent address: Institut de Mathématique, Université de Liège, Grande Traverse 12, 4000 Liège, Belgium.

[1] F. Jacob, S. Brenner, and F. Cuzin, Cold Spring Harbor Symposia on Quantitative Biology **28**, 329 (1963).
[2] (a) R. Berezney, D. D. Dubey, and J. A. Huberman, Chromosoma **108**, 471 (2000); (b) D. M. Gilbert, Science **294**, 96 (2001); (c) S. P. Bell and A. Dutta, Annu. Rev. Biochem. **71**, 333 (2002); (d) S. A. Gerbi and A. K. Bielinsky, Curr. Opin. Genet. Dev. **12**, 243 (2002).
[3] C. S. Newlon and J. F. Theis, Curr. Opin. Genet. Dev. **3**, 752 (1993).
[4] J. A. Bogan, D. A. Natale, and M. L. Depamphilis, J. Cell. Physiol. **184**, 139 (2000); M. Méchali, Nat. Rev. Genet. **2**, 640 (2001); D. M. Gilbert, Nat. Rev. Mol. Cell Biol. **5**, 848 (2004).
[5] M. Touchon *et al.*, Proc. Natl. Acad. Sci. U.S.A. **102**, 9836 (2005); E. B. Brodie of Brodie *et al.*, Phys. Rev. Lett. **94**, 248103 (2005).
[6] (a) K. Woodfine *et al.*, Cell Cycle **4**, 172 (2005); (b) E. J. White *et al.*, Proc. Natl. Acad. Sci. U.S.A. **101**, 17 771 (2004).
[7] M. K. Raghuraman *et al.*, Science **294**, 115 (2001).
[8] D. Schübeler *et al.*, Nat. Genet. **32**, 438 (2002).
[9] Y. Watanabe *et al.*, Human Molecular Genetics **11**, 13 (2002); K. Woodfine *et al.*, Human Molecular Genetics **13**, 191 (2004).
[10] M. Touchon *et al.*, FEBS Lett. **555**, 579 (2003); M. Touchon *et al.*, Nucleic Acids Res. **32**, 4969 (2004); S. Nicolay *et al.*, Phys. Rev. E **75**, 032902 (2007).
[11] A. Arneodo, G. Grasseau, and M. Holschneider, Phys. Rev. Lett. **61**, 2281 (1988); A. Arneodo *et al.*, in *Wavelets and Applications*, edited by Y. Meyer (Springer, Berlin, 1992), p. 286.
[12] (a) C. Schmegner *et al.*, Cytogenet. Genome Res. **116**, 167 (2007); (b) M. Costantini *et al.*, Chromosoma **116**, 29 (2007).
[13] P. Bergé, Y. Pomeau, and C. Vidal, *Order within Chaos* (Wiley, New York, 1986); *Chaos II*, edited by B.-L. Hao (World Scientific, Singapore, 1990).