Letter

Human gene organization driven by the coordination of replication and transcription

Maxime Huvet,[1] Samuel Nicolay,[2,5] Marie Touchon,[1,3,4] Benjamin Audit,[2] Yves d'Aubenton-Carafa,[1] Alain Arneodo,[2] and Claude Thermes[1,6]

[1]*Centre de Génétique Moléculaire (CNRS), 91198 Gif-sur-Yvette, France;* [2]*Laboratoire Joliot Curie et Laboratoire de Physique, Ecole Normale Supérieure de Lyon, CNRS, 69364 Lyon, France;*
[3]*Génétique des Génomes Bactériens, CNRS URA2171, Institut Pasteur, 75015 Paris, France;*
[4]*Atelier de Bioinformatique, Université Pierre et Marie Curie-Paris 6, 75005 Paris, France*

In this work, we investigated a large-scale organization of the human genes with respect to putative replication origins. We developed an appropriate multiscale method to analyze the nucleotide compositional skew along the genome and found that in more than one-quarter of the genome, the skew profile presents characteristic patterns consisting of successions of N-shaped structures, designated here N-domains, bordered by putative replication origins. Our analysis of recent experimental timing data confirmed that, in a number of cases, domain borders coincide with replication initiation zones active in the early S phase, whereas the central regions replicate in the late S phase. Around the putative origins, genes are abundant and broadly expressed, and their transcription is co-oriented with replication fork progression. These features weaken progressively with the distance from putative replication origins. At the center of domains, genes are rare and expressed in few tissues. We propose that this specific organization could result from the constraints of accommodating the replication and transcription initiation processes at chromatin level, and reducing head-on collisions between the two machineries. Our findings provide a new model of gene organization in the human genome, which integrates transcription, replication, and chromatin structure as coordinated determinants of genome architecture.

[Supplemental material is available online at www.genome.org.]

It has long been known that genes are nonrandomly distributed in eukaryote genomes (gene-dense regions alternating with gene deserts) (Mouchiroud et al. 1991; Zoubak et al. 1996). Over the past few years, complete genome sequences have confirmed this striking nonrandomness in several species (Lander et al. 2001; Hurst et al. 2004). In the human genome, statistical studies have shown that highly expressed genes have a tendency to form clusters (Caron et al. 2001; Versteeg et al. 2003). However, it has also been reported that these clusters, in fact, result from the clustering of genes coexpressed in a large number of tissues (housekeeping genes); although individual expression rates may vary from tissue to tissue, the overall expression pattern remains similar (Lercher et al. 2002). Several hypotheses have been advanced to explain the formation and/or maintenance of this organization. On the one hand, short-range regulatory mechanisms can be responsible for small clusters of coexpressed genes. A gene might be turned on solely because of its proximity to signals regulating neighboring genes (Cajiao et al. 2004). On the other hand, long-range mechanisms could maintain large-size clusters of coexpressed genes, and it has often been argued that the chromatin structure could play such a role. When chromatin is in open conformation during gene transcription, this conformation can extend to neighboring genes. The presence in a region of a high proportion of genes active in most tissues would keep chromatin in an open structure in most cell types, thus leading to the observed coexpressed gene clusters (Spellman and Rubin 2002; Gilbert et al. 2004; Hurst et al. 2004; Sproul et al. 2005). Comparative analysis of clusters of coexpressed genes in human and mouse indicate that they are found together in both species more often than expected by chance, suggesting that clustering could result from natural selection (Singer et al. 2005). This process could contribute to a high degree of organization of human chromosomes. However, these observations were challenged by a recent study showing that most clusters of coexpressed genes seem to contain only two to three genes, the number of clusters only slightly exceeding the number expected by chance, which limits their impact on global genome organization (Sémon and Duret 2006). Moreover, these clusters seem to be held together by short-range effects resulting from promoters sharing common regulatory elements or transcriptional read-through.

Here, we address the question of gene organization with respect to replication. Previous studies have shown that the human genome displays nucleotide compositional strand asymmetries that probably result from asymmetric mutation and repair processes associated with replication and transcription (Green et al. 2003; Touchon et al. 2003, 2004, 2005; Brodie of Brodie et al. 2005). Genome-wide analyses of these skews allowed us to predict a large number of putative human replication origins (Brodie of Brodie et al. 2005; Touchon et al. 2005). These analyses also suggested that genes in the immediate vicinity of these putative origins displayed a characteristic organization, which prompted us to extend this analysis to a larger scale. We devised a new methodology to identify large genome domains bordered by putative replication origins and found that, within these domains, genes are highly ordered according to their breadth of expres-

[5]Present address: Département de Mathématique, Université de Liège, 12 Grande Traverse, 4000 Liège, Belgium.
[6]Corresponding author.
E-mail thermes@cgm.cnrs-gif.fr; fax 33-1-69-82-38-77.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6533407.

17:1278–1285 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

sion, orientation, and position. The data allowed us to propose a model suggesting that replication and transcription are essential and coordinated determinants of gene organization, leading to a new vision of the human genome architecture.

Results

In order to predict putative replication origins, in previous studies we explored the large-scale behavior of nucleotide strand compositional asymmetries defined as the skew $S = (G - C)/(G + C) + (T - A)/(T + A)$ (i.e., the relative excess of G over C and T over A) (Brodie of Brodie et al. 2005; Touchon et al. 2005). Among the well-known, experimentally determined human replication origins (very few due to experimental difficulties), most (six of nine) are associated with a specific property of the skew *S*: as it crosses a replication origin, the sign of *S* changes abruptly, producing a sharp upward transition of the *S* profile. This property allowed us to identify ~1000 putative replication origins in the human genome. Remarkably, successive transitions are connected to each other by DNA segments in which the *S* values decrease in the 5′ to 3′ direction, thereby displaying a characteristic serrated pattern reminiscent of factory roofs (Fig. 1A). We propose as a working model that this N-like shape results from the superimposition of two patterns. One decreases steadily in the 5′ to 3′ direction and would be attributable to replication initiating at two fixed adjacent origins (associated with two upward transitions) and terminating during the successive germline cell divisions at various positions randomly dispersed between these origins (Fig. 1B,C). The other pattern would result from transcription-associated strand asymmetries that generate step-like blocks corresponding to (+) and (−) genes (Touchon et al. 2003, 2004) (Fig. 1D). When the two profiles are superimposed, this leads to the factory-roof pattern (Fig. 1E). We define as an "N-domain" any DNA segment for which the *S* profile displays the characteristic factory-roof pattern. This model implies that no other fixed replication origin active in germ-line cells can be located within the N-domains (since any additional origin would disrupt the N-shape of the domains). We set out to detect these domains in the human genome and to study gene organization within these domains.

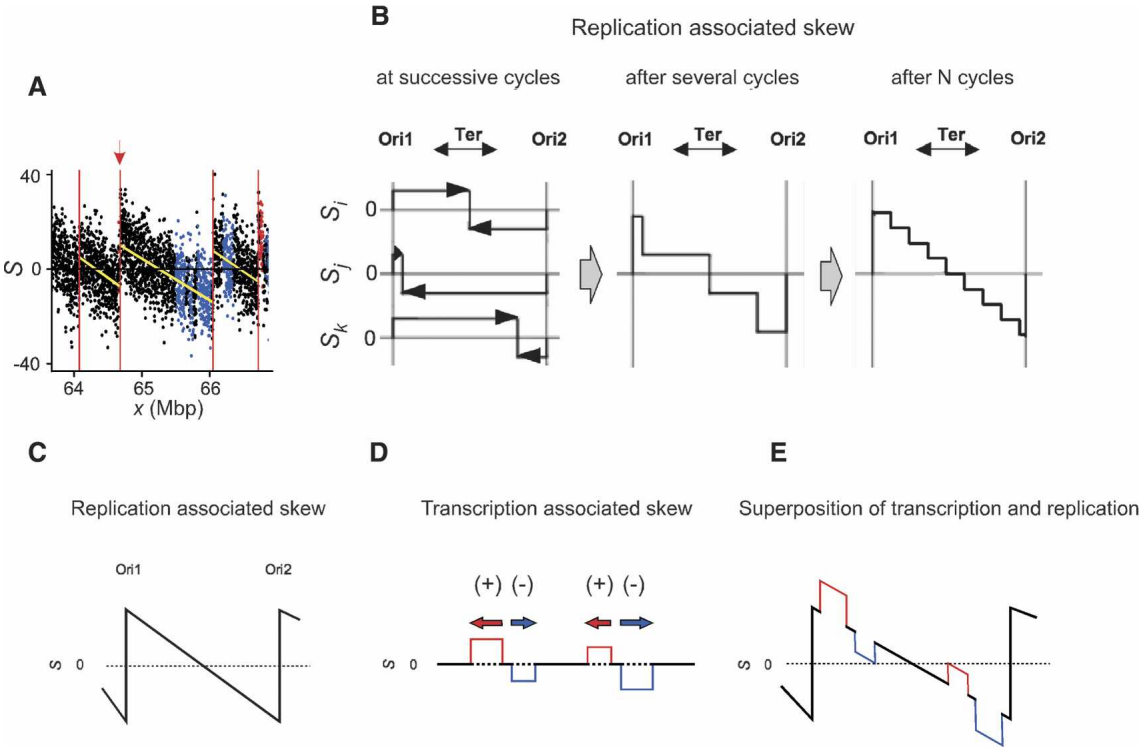

Figure 1. Factory-roof pattern of the skew profile. (*A*) Skew (*S*) profile around an experimentally identified replication origin. The skew is computed along a DNA fragment containing the experimentally determined replication origin associated with the *MYC* gene (Vassilev and Johnson 1990) (red arrow). *S* is computed in 1-kbp adjacent windows of masked sequences; (red) + genes (coding strand identical to the Watson strand); (blue) − genes (opposite direction); (black) intergenic regions (the color of each point is defined by the majority rule). In abscissa, the position on the sequence; in ordinate, the skew, *S*, in percent. (Red vertical lines) Putative replication origins associated with upward transitions of the *S* profile. (*B–E*) Working model of the factory-roof pattern of the *S* profile. We propose that this pattern results from the superimposition, in germ-line cells, of strand asymmetries associated with replication and transcription. (*B*) Model of the replication-associated skew profiles corresponding to two fixed putative adjacent replication origins, Ori1 and Ori2, and to a replication termination site (Ter) occurring with equal probability between Ori1 and Ori2 (adapted from Touchon et al. 2005). Upward or downward jumps of the *S* profile correspond to the origin and termination positions, respectively. (*Left*) Three elementary skew profiles, S_i, S_j, and S_k, are associated with three successive replication cycles and display three different Ter positions. (*Middle*) Superimposition of the S_i, S_j, and S_k profiles. (*Right*) Superimposition of a large number of elementary skew profiles, ultimately leading to a pattern decreasing linearly in the 5′ to 3′ direction; note that reverse complementation of the sequence leaves the factory roof structure intact. (*C*) Final replication-associated skew profile. (*D*) Transcription-associated skew profile showing positive step-like blocks at + gene positions and negative step-like blocks at − gene positions. (*E*) Superimposition of the replication- and transcription-associated skew profiles producing the final factory-roof pattern that defines the N-domains.

Detecting the N-domains

To extract the N-domains from the noisy *S* profile of the genome, we developed an adapted wavelet-based multiscale methodology to identify segments of variable length and position displaying a factory-roof pattern (Methods; Supplemental Figs. S1, S2). According to the model, the selection involves (1) searching for segments that decrease between two large upward jumps, and (2) retaining those containing both intergenic regions with a linearly decreasing *S* profile (possibly induced by replication) and genes associated with step-like blocks (possibly induced by transcription) superimposed over this linearly decreasing profile. This amounts to disentangling the components of the skew attributed to replication and to transcription (Methods; Supplemental Fig. S3). When applied to the human genome, the method detected 678 N-domains bordered by 1060 putative replication origins. These domains are evenly distributed in most chromosomes, spanning 28.3% of the genome with a mean length $L = 1.2 \pm 0.6$ Mbp (Fig. 2A; Methods; Supplemental Fig. S4).

During the selection process, a number of candidate structures were examined that were not finally retained as N-domains since they display some departure from symmetry of the skew with respect to the center of the domain. However, these structures, which span approximately another 30% of the genome, can be considered as N-domain-like structures, and do display a type of gene organization reminiscent of that observed in the bona-fide N-domains (described below). In most of the remaining genome regions, two types of *S* profile were observed. The first type, observed in regions with high gene density, small gene size, and high GC content (Lander et al. 2001) displayed a high density of large upward and downward jumps (they span ~20% of the genome). These complex *S* profiles hampered the detection of the N-domains. For example, both small domain density and small chromosome coverage were observed in chromosome 19, which contains a high proportion of gene-rich and GC-rich regions (Supplemental Fig. S5c,d). The second type, observed in gene-poor regions with low GC content did not display large upward jumps, but rather flat patterns, suggesting that replication origins are not fixed. These regions span ~15% of the genome and correspond to gene deserts (Lander et al. 2001; Ovcharenko et al. 2005).

Analysis of the *S* profile of the N-domains

The mean *S* profile of the selected N-domains decreases steadily between opposed values to form a jagged pattern with rather symmetrical left (5′) and right (3′) halves (the mean *S* values at the 5′ and 3′ extremities are $6.8 \pm 0.2\%$ and $-7.1 \pm 0.2\%$, respectively) (Fig. 2B). Between these extreme values, the mean *S* profile decreases fairly linearly (Fig. 2C) and accordingly, the slope of the domains varies approximately as $-1/L$ (hyperbolic curve in Fig. 2E, orange dots). On average, genes and intergenic regions both display linear *S* profiles (Fig. 2D) that parallel the profiles of the corresponding domains (Fig. 2E). The fact that the *S* values for gene sequences were larger than those for intergenic sequences (Fig. 2D) reflects the contribution of transcription. These results strongly support our hypothesis that the skew profile corresponds to the superimposition of replication- and transcription-associated profiles (Fig. 1C–E).

To what extent could the specific *S* profile of the N-domains be expected to result from chance? We first examined the human *S* profile, looking for structures presenting an inverted factory-roof pattern, i.e., two downward jumps separated by a steadily increasing skew. We adapted our method to detect such structures (the method is the same as that described above apart from the analyzing wavelet; Supplemental Fig. S1b). When this method was applied to human autosomes, it detected no more than 27 inverted structures (vs. 678 N-domains) spanning only 0.6% of the genome. N-domains therefore very significantly out-

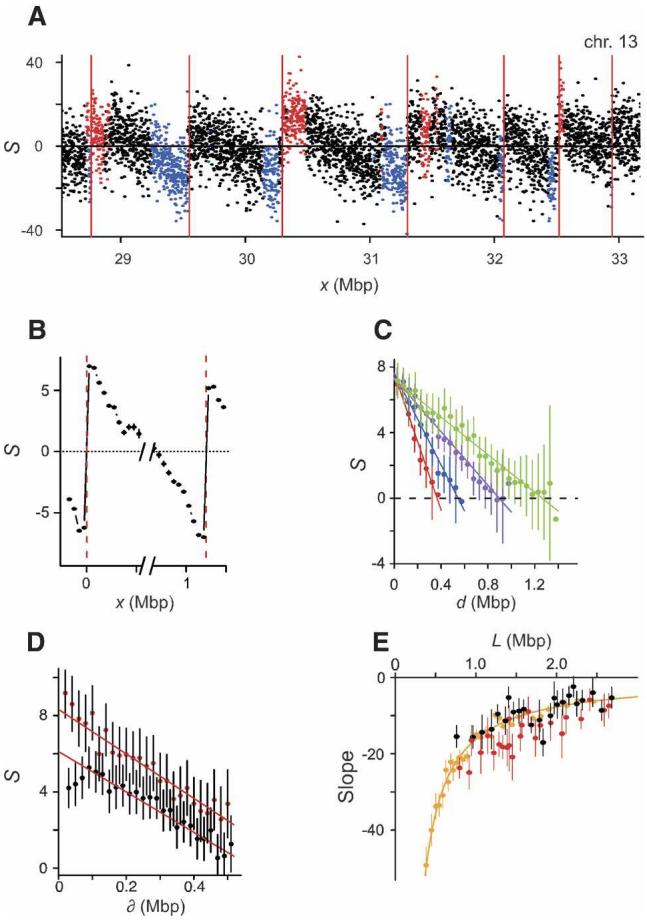

Figure 2. Properties of the N-domains detected in the human genome. (*A*) Examples of N-domains detected in the chromosome 13. *S* values are computed in 1-kbp windows (without repeats); (red) + genes; (blue) − genes; (black) intergenic regions; the N-domain borders are indicated by red vertical lines. In abscissa, the window position is in megabase pairs; in ordinate, the skew, *S*, in percent. (*B*) Mean *S* profile of the N-domains. The mean *S* values are computed along the N-domains of length $L \leq 1.2$ Mbp. In abscissa, the region used for analysis extends from the extremity to the center of each domain. In ordinate, the mean skew, *S*, in percent ± SEM. (*C*) Mean *S* profile of the half-domains for $L < 0.75$ Mbp (red), $0.75 < L < 1.2$ Mbp (blue), $1.2 < L < 2$ Mbp (purple), and $L > 2$ Mbp (green). The sequences of the 3′ halves of the domains are reverse-complemented and analyzed together with the 5′ halves. (*D*) Mean skew profile of + genes located in 5′ half-domains analyzed together with − genes (reverse-complemented) located in 3′ halves (red) and intergenic regions (black) (both larger than 400 kbp, and situated in domains with $L > 1$ Mbp). In abscissa, the distance ∂ to the 5′ end of genes or intergenic regions. (*E*) Mean slope of the domains versus their length *L*; domains are ranked by *L* values and grouped by bins of 20 domains; in ordinate, the mean (±SEM) of the slopes in percent/megabase pair (orange); the orange hyperbolic curve is obtained by a linear regression fit of $-1/\text{slope}$ versus *L* (Supplemental Fig. S5f). In red, the genes with a length >400 kbp are ranked by length of their domain, and grouped by constant bins; the mean slope is computed for each bin. The same is true for the intergenic regions (>400 kbp) (black). In abscissa, the mean length of the corresponding domains.

number inverted structures ($P < 10^{-15}$). Secondly, we looked for N-domains in sequences obtained after shuffling the order of genes and intergenic regions (Methods), and found that they were significantly less frequent than in the native sequences ($P < 10^{-15}$). This observation also provides the first indication that the existence of N-domains does indeed reflect some specific gene organization.

Replication timing profile of the N-domains

Using a high-resolution replication timing map of human chromosome 6 (Woodfine et al. 2005), we determined the timing profile along the corresponding N-domains identified by our method. On average, this profile displays maxima at positions corresponding to the domain extremities, and decreases regularly on both sides, revealing that (1) a significant number of domain extremities correspond to early replicating sequences, (2) they are replicated earlier than their surroundings, and (3) the central region of large N-domains replicate late in the S phase (Fig. 3). These results provide experimental evidence that at the degree of resolution of the timing map, a number of N-domain extremities correspond to bona-fide replication initiation zones that are active rather early in the S phase.

Gene organization in the N-domains

Gene shuffling experiments revealed an underlying gene organization in the N-domains (see above). In order to decipher this organization, we analyzed the gene patterns. Most putative origins (domain borders) are intergenic (77%) and located near a gene promoter more often than would be expected by chance (Supplemental Fig. S6a,b). The N-domains contain approximately equal numbers of genes oriented in each direction (1511 + genes and 1507 − genes). Gene distributions in the 5′ halves of domains contain more + genes than − genes, regardless of the total number of genes located in the half-domains (Supplemental Fig. S6c). Symmetrically, the 3′ halves contain more − genes than + genes (Supplemental Fig. S6d). A total of 32.7% of half-domains contain one gene, and 50.9% contain more than one gene. For convenience, + genes in the 5′ halves and − genes in the 3′ halves are defined as R+ genes (Fig. 4A): their transcription

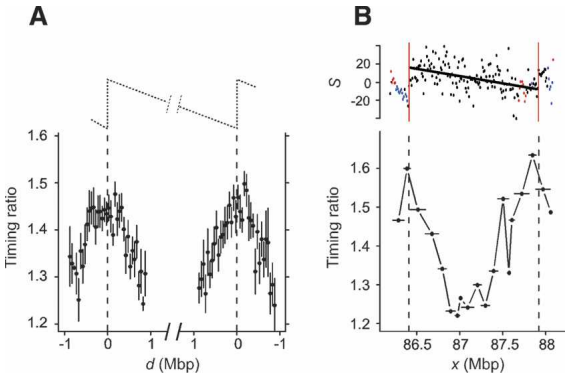

Figure 3. Replication timing profile of the N-domains. (*A*) Average replication timing values (±SEM) determined around the extremities of the domains located in chromosome 6; in abscissa, the distance to the indicated 5′ (*left*) or 3′ (*right*) closest domain extremity; in ordinate, the mean timing ratio value; data are retrieved from Woodfine et al. (2005). (*B*) Example of replication timing profile along a complete N-domain. Horizontal bars indicate the DNA probes (~94 kb) used in the microarray experiments (Woodfine et al. 2005).

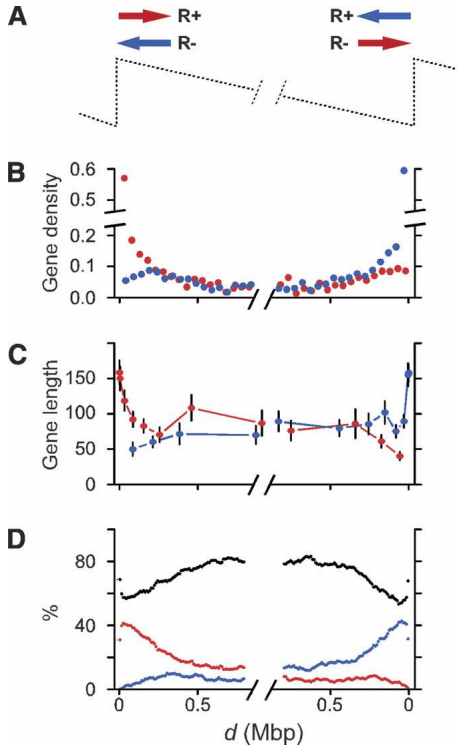

Figure 4. Analysis of the genes located in the N-domains. (*A*) Arrows indicate the R+ orientation i.e., the same orientation as the most frequent direction of putative replication fork progression; R− orientation (opposed direction); (red) + genes; (blue) − genes. (*B*) Gene density. The density is defined as the number of 5′ ends (for + genes) or of 3′ ends (for − genes) in 50-kbp adjacent windows, divided by the number of corresponding domains. In abscissa, the distance, *d*, in megabase pairs, to the closest domain extremity. (*C*) Mean gene length. Genes are ranked by their distance, *d*, from the closest domain extremity, grouped by sets of 150 genes, and the mean length (kilobase pairs) is computed for each set. (*D*) Relative number of base pairs transcribed in the + direction (red), − direction (blue), and nontranscribed (black) determined in 10-kbp adjacent sequence windows.

is, in most cases, oriented in the same direction as the putative replication fork progression (genes transcribed in the opposite direction are defined as R− genes). The 678 N-domains contain significantly more R+ genes (2041) than R− genes (977) ($\chi^2 = 375$, $P < 10^{-15}$, Supplemental Table S2a). Within 50 kbp of putative replication origins, the mean density of R+ genes is 8.2 times greater than that of R− genes. This asymmetry weakens progressively with the distance from the putative origins, up to ~250 kbp (Fig. 4b). A similar asymmetric pattern is observed when the domains containing duplicated genes are eliminated from the analysis, whereas control domains obtained after randomization of domain positions (Methods) present similar R+ and R− gene density distributions (Supplemental Fig. S7a–a″). The mean length of the R+ genes near the putative origins is significantly greater (~160 kbp) than that of the R− genes (~50 kbp); however, both tend toward similar values (~70 kbp) at the center of the domain (Fig. 4C). A similar pattern is observed after eliminating duplicated genes, whereas, in contrast, the control domains display fairly constant gene length (Supplemental Fig. S7b–b″). Within 50 kbp of the putative origins, the ratio between the numbers of base pairs transcribed in the R+ and R− directions is 23.7; this ratio falls to ~1 at the domain centers (Fig. 4D). A similar pattern is observed after eliminating duplicated genes;

this ratio is constant in the control domains (Supplemental Fig. S7c–c″). This strong transcriptional polarity could be mainly attributable to the preferential R+ orientation of the first gene (closest to the extremity). However, polarity is still observed for half-domains harboring various gene numbers even after the first gene has been eliminated (Supplemental Fig. S8).

Gene expression in the N-domains

We analyzed the breadth of expression, N_t (number of tissues in which a gene is expressed), of genes located within the N-domains. We found that it significantly decreases from the extremities to the center, regardless of whether it is measured by EST, SAGE, or microarray data ($\chi^2 = 29$, $P = 10^{-8}$ for EST data). The distribution is symmetrical in the 5′ and 3′ half-domains (Fig. 5A,B). Significantly decreasing mean N_t values are also observed after eliminating duplicated genes, whereas they remain constant within the control domains obtained after randomizing the domain positions (Methods; Supplemental Fig. S7d-f″). The distribution of N_t values (determined using ESTs) displays a bimodal pattern for the genes located in the domains (Fig. 5C), with one mode (peak at $N_t < 5$) corresponding to the genes expressed in only a few tissues, and a second mode (a wide bump centered at $N_t \sim 15$) corresponding to widely expressed genes. It is noteworthy that this distribution is similar to that found for the complete set of human genes (Supplemental Fig. S9d). Genes located near the putative replication origins tend to be widely expressed (Fig. 5D), whereas those located far from them are mostly tissue specific (Fig. 5E). We checked that the decrease in both N_t values and gene length L, from the N-domain border to its center (Fig. 4C), did not reflect a correlation between these factors: no correlation was observed between gene length and expression breadth measured using EST, SAGE, or microarray data (Supplemental Fig. S9a–c). In addition, no significant correlation was observed between the transcription rate of a gene and its position within an N-domain, whether or not duplicated genes are eliminated from the analysis (data not shown).

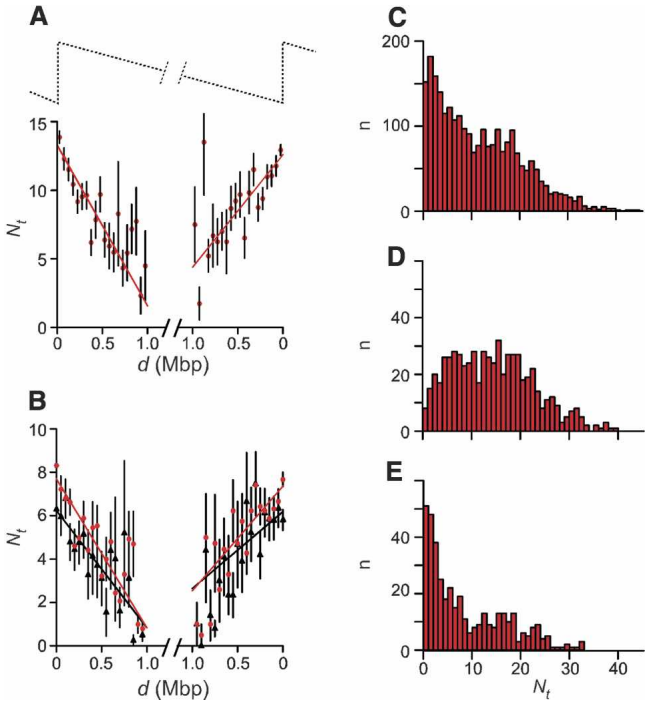

Figure 5. Expression breadth, N_t, of the genes located in the N-domains. (*A*) Mean expression breadth calculated using EST data (red). In abscissa, the distance, *d*, in megabase pairs, to the closest domain extremity. (*B*) Same as in *A* with SAGE data (red) and microarray data (black). (*C*) Histogram of the expression breadth (determined with EST data) of the genes located in the domains. (*D*) Histogram of the expression breadth of the genes with an extremity (5′ for R+ genes, 3′ for R− genes) located at distance *d* from the putative replication origins where $d < 5\%$ of the length of the half-domain. (*E*) Same as in *D*, but with $70\% < d < 100\%$.

Discussion

This study shows that some features of human genome organization can be unraveled by examining the properties of the nucleotide compositional skew. The *S* profile exhibits a highly significant number of occurrences of so-called N-domains, specific structures consisting of two sharp upward transitions connected by a downward-sloping segment. These large structures are recognizable along all chromosomes. They are unambiguously detected by our methodology in more than one-quarter of the genome. Could these structures be generated solely by transcription? Transcription generates strand asymmetries along gene sequences, leading to step-like blocks in the *S* profile (Green et al. 2003; Touchon et al. 2003, 2004). Premature termination of RNA polymerase elongation could occur during transcription, leading to *S* profiles that decrease along the gene sequence. However, it would be unlikely to produce linear downward profiles (termination at random positions would generate exponentially decreasing *S* profiles). Moreover, this cannot account for the shift from positive to negative *S* values observed along gene profiles (Supplemental Fig. S4c–f), since transcription always generates positive *S* values on the coding strand (and negative ones on the noncoding strand) (Green et al. 2003; Touchon et al. 2003, 2004). Recent studies have revealed the existence of complex networks of unannotated transcripts (Cheng et al. 2005; Kapranov et al. 2005). Superimposition of sense and antisense transcription could also generate decreasing *S* profiles, but it is unlikely that these transcripts would display the specific organization required to produce linear profiles that are, moreover, parallel in genes and intergenic regions. In addition, most of these unannotated transcripts are weakly expressed (Cheng et al. 2005), so that their transcription would not generate significant compositional skew. These data are compliant with our hypothesis that the factory-roof pattern is produced by the superimposition of a jagged profile, resulting from replication, over a crenellated profile resulting from transcription (Fig. 1B–E).

The replication timing profile of the N-domains shows that, on average, the extremities replicate earlier than the neighboring regions, which is consistent with these regions being true replication origins, active early in the *S* phase. This profile was established using replication timing data obtained from lymphoblastoid cells (Woodfine et al. 2005), suggesting that a number of putative replication origins detected by our method (i.e., active in germ-line cells) are also active in these cells in the early *S* phase. We therefore propose that the putative replication origins detected by our approach are, at least in part, early, well-positioned replication origins active in most cell types. This proposition is supported by earlier studies of replication timing in various cell types that suggested some conservation of timing between tissues (White et al. 2004). It is also consistent with

previous analyses of the *S* profile, showing that most well-known, experimentally determined replication origins coincide with sharp upward transitions of the *S* profile (Brodie of Brodie et al. 2005; Touchon et al. 2005), indicating that these origins, which were all identified in somatic cells, are also likely to be active in germ-line cells.

According to our model, replication units would be better described by N-domains, as defined in this study, than by the usual replicons. Indeed, the fixed terminators of the replicons would not be suitable for describing the putative variable termination sites within the N-domains. The length of the N-domains matches the large, ~1 Mbp-long replicons (Yurov and Liapunova 1977; Berezney et al. 2000) rather than the usually 50–300 kbp-long replicons (Edenberg and Huberman 1975), and is consistent with the large replication units observed in meiotic chromosomes (Callan 1972). In somatic cells, additional origins may be activated within the N-domains, thus leading to the commonly observed shorter replication units.

We then asked whether these N-domains correspond to a specific gene organization pattern. Most putative replication origins located at domain extremities are intergenic and located close to promoters of widely expressed genes (housekeeping genes) oriented toward the domain center. Gene density, breadth of expression, and transcription polarity all tend to decrease progressively from the extremities of the domain toward its center. In the central region, genes are few in number, tissue specific, and have no preferential orientation (Fig. 6). We propose that coordination between replication and transcription is the key to this complex architecture. The putative replication origins would mostly be active early in the *S* phase in most tissues. Their activity could result from particular genomic context involving transcription-factor binding sites and/or from the transcription of their neighboring housekeeping genes. This activity could also be associated with an open chromatin structure, permissive to early replication and gene expression in most tissues (Gilbert et al. 2004; Hurst et al. 2004; Chakalova et al. 2005; Sproul et al. 2005). This open conformation could extend along the first gene, possibly promoting the expression of further genes. This effect would progressively weaken with the distance from the putative replication origin, leading to the observed decrease in expression breadth. This model is consistent with a number of data showing that in metazoans, ORC and RNA polymerase II colocalize at transcriptional promoter regions (MacAlpine et al. 2004), and that replication origins are determined by epigenetic information such as transcription-factor binding sites and/or transcription (Lin et al. 2003; Danis et al. 2004; Ghosh et al. 2004; DePamphilis 2005). It is also consistent with studies in *Drosophila* and humans that report correlation between early replication timing and increased probability of expression (Schubeler et al. 2002; MacAlpine et al. 2004; White et al. 2004; Jeon et al. 2005; Woodfine et al. 2005). The data we report here provide the first demonstration of quantitative relationships in the human genome between gene expression, orientation, and distance from putative replication origins.

Near the putative origins bordering the N-domains, transcription is preferentially oriented in the same direction as replication fork progression. We propose that this co-orientation would reduce head-on collisions between the replication and transcription machineries, which could induce deleterious recombination events either directly or via stalling of the replication fork (Deshpande and Newlon 1996; Takeuchi et al. 2003). In bacteria, co-orientation of transcription and replication has been observed for essential genes, and has been associated with a reduction in head-on collisions between DNA and RNA polymerases (Rocha and Danchin 2003). Recent results support the hypothesis that co-orientation bias of replication and transcription in *Bacillus subtilis* results from deleterious effects on replication caused by head-on transcription (Wang et al. 2007). It is noteworthy that in human N-domains such co-orientation usually occurs in widely expressed genes located near putative replication origins. Near domain centers, head-on collisions may occur in 50% of replication cycles, regardless of the transcription orientation, since there is no preferential orientation of the replication fork progression in these regions. However, in most cell types, there should be few head-on collisions, due to the low density and expression breadth of the corresponding genes. Selective pressure to reduce head-on collisions may thus have contributed to the simultaneous and coordinated organization of gene orientation and expression breadth along the N-domains (Fig. 6).

The data presented here strongly suggest the existence in the human genome of regions bordered by putative early replication origins in which gene position, orientation, and expression breadth present a high level of organization, possibly mediated by the chromatin structure. This allows us to propose a model of gene order that relates transcription and replication as coordinated determinants of genome organization.

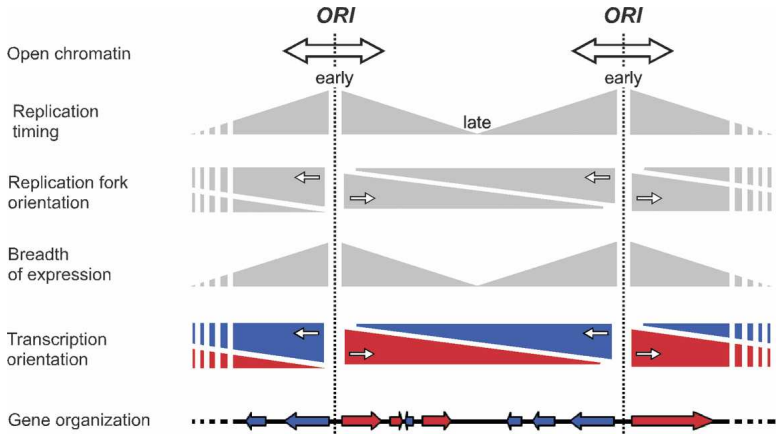

Figure 6. Model of gene organization coordinated by replication and transcription. Two successive putative replication origins (ORI) delineate a replication N-domain. (Open chromatin) Arrows illustrate an open chromatin state at replication origin position; (replication timing) the triangles figure the replication timing values along the N-domain. Replication fork orientation: the triangles indicate the proportion of replication forks progressing from each extremity to the other extremity along the domain (during the successive cell cycles, replication terminates at random sites within the domain). Breadth of expression is maximum near the replication origins and decreases toward the domain center (gray triangles). Transcription orientation: it is preferentially co-oriented with the replication fork progression; the colored triangles indicate the proportion of base pairs along the domain transcribed in the + direction (red) and – direction (blue). Gene organization: red (resp. blue) arrows indicate + (resp. −) genes in the domains.

Methods

Sequence and expression data

Sequence and annotation data were retrieved from the Genome Browser of the University of California Santa Cruz (UCSC, hg17). To obtain gene sequences, we used the RefSeq annotation (containing only protein-coding transcripts). When two genes presenting the same orientation overlap, the largest gene was retained. For the detection process of N-domains, sequences masked with REPEATMASKER were retrieved from the UCSC browser to avoid the biases intrinsic to repeated elements. In all other analyses, sequences were not masked. The skew, *S,* was computed in nonoverlapping, 1-kbp windows. EST, SAGE, and microarray data were provided by M. Sémon and L. Duret (Sémon et al. 2005). Among the 3018 genes located in the N-domains, EST (Expressed Sequence Tags) data were available for 2514 genes in 50 normal tissues, SAGE (Serial Analysis of Gene Expression) data were available for 2668 genes in 22 normal tissues, and microarray data were available for 1276 genes in 22 normal tissues.

Detection of N-domains using the wavelet transform

Using the wavelet transform (WT) as multi-scale shape detector, we search at every sequence position for segments of variable length presenting a factory-roof skew pattern (space-scale analysis) (Supplemental Section S1). In the first step, we used as the analyzing wavelet the function, Ξ, constituted by a linearly decreasing segment between two upward jumps (Supplemental Fig. S1), and computed the WT of the strand compositional asymmetry *S* measured in 1-kbp windows (Supplemental Fig. S2). The space-scale locations of significant maximum values in this two-dimensional decomposition (red areas in Supplemental Fig. S2b) indicate the middle position (spatial location, abscissa in Supplemental Fig. S2b) of candidate N-domains, the size of which is shown by the scale location (ordinate in Supplemental Fig. S2b). In order to avoid false positives, we then checked that there was indeed a well-defined upward jump at each domain extremity (Supplemental Fig. S2b). Because the mean value of the analyzing wavelet was zero, the WT decomposition was insensitive to (global) asymmetry offset. Hence, in order to enforce strong compatibility with the working model of replication (Fig. 1B–E), we retained from the set of candidate domains obtained at the previous step, only those where the two upward jumps corresponded to a transition from a negative *S* value $< -3\%$ to a positive *S* value $> +3\%$. In the second step, we disentangled two components of the skew profile possibly associated with replication and transcription. Ignoring transcription bias, the asymmetry profile S_R in one N-domain can be expressed as follows:

$$S_R(t) = -2\delta \times (t - \tfrac{1}{2}), \quad (1)$$

where position *t* within the domain has been rescaled between 0 and 1, and $\delta > 0$ is the replication bias. If we now take into account the contribution of transcription S_T to the bias in a gene-containing domain, the asymmetry profiles can be written as:

$$S(t) = S_R(t) + S_T(t) = -2\delta \times (t - \tfrac{1}{2}) + \sum_{gene} c_g \times \chi_g(t), \quad (2)$$

where χ_g is the characteristic function for the g^{th} gene (1 when there are *t* 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). For each domain identified in the previous step, we used a least-square fitting procedure to estimate the replication bias, δ, and each value of the gene transcription bias, c_g. The resulting χ^2 value was used to select the domains where the *S* noisy profile is well described by Equation 2. As illustrated in Supplemental Figure S3 and Supplemental Table S1 for a fragment of human chromosome 6 that contains three adjacent N-domains (Supplemental Fig. S3a), this method provides a very efficient way of disentangling the step-like component of strand asymmetry associated with transcription (Supplemental Fig. S3b) from the jagged component associated with replication (Supplemental Fig. S3c).

Applying this procedure to the 22 human autosomes, we detected 678 N-domains and predicted 1060 putative origins of replication (in 296 cases, the right origin of a domain is also the left origin of the following domain). Examples of such N-domains are illustrated in Figure 2A and Supplemental Fig. S4. The domain length ranges between ~300 kbp and ~2.8 Mbp, with an average domain density of 0.22 ± 0.07 domain/Mbp (Supplemental Fig. S5a). The distribution of the domain GC content (39.8 ± 4.1%) is narrower than that of the whole genome (41.0 ± 5.1%) indicating some degree of under-representation of regions presenting high GC contents (Supplemental Fig. S5b). The mean values of several characteristics decrease as the GC content increases: the density of N-domains, the proportion of the chromosome length covered by the domains, and the domain length (Supplemental Fig. S5c–e).

Randomization of gene order

In order to compare the N-domains detected in human chromosomes to those detected in sequences obtained after randomization of gene order, we randomly permuted genes and intergenic regions without any change in orientation (genes and intergenic regions alternate in the shuffled chromosomes). The process was performed 100 times using chromosome 3 (this chromosome has domain properties representative of those of the whole genome, Supplemental Fig. S5c,d). The process detected, on average, 12.8 control domains per shuffled chromosome, compared with 56 putative replication domains detected in native chromosome 3. Control domains have a mean length of 1.0 ± 0.5 Mbp, a density of 0.065 ± 0.02 domain/Mbp (to be compared with 0.22 ± 0.07 N-domain/Mbp in native sequences), and correspond to only 6.9% of the shuffled sequences.

Randomization of N-domain positions

In this control test, we studied the gene characteristics in DNA segments chosen at random along the chromosome sequences. These segments are considered as control domains. In each chromosome, the length and number of these control domains are equal to those of the previously detected N-domains in the corresponding chromosome. This operation was repeated 10 times on the 22 autosomes (leading to 6780 control domains).

Detection of duplicated genes

Genes displaying a high level of sequence identity were identified using BLASTP. Two genes were considered duplicates if they presented *E*-values of <0.2 (Lercher et al. 2002) and a number of identical amino acids >30% of the shortest protein length (Li et al. 2001). This led to the identification of 322 genes displaying a duplicated gene in the same domain, among the 3018 genes contained in all the domains; 94 domains contained at least two duplicated genes.

Statistics

To assess correlations, Pearson's correlation coefficients were computed. To evaluate the statistical significance of the decreasing N_t pattern observed along N-domains, we used the following

procedure. A linear fit of the N_t profile was performed in each half-domain containing more than one gene. The numbers of fits with negative and positive slopes were compared with the corresponding numbers obtained with the control domains (randomization of N-domain positions, see previous section).

The positions of the N-domains and of the inverted N-domains are available as Supplemental material.

Acknowledgments

We thank M. Sémon and L. Duret for providing the expression data for human genes, S. Camier, L. Duquenne, and M. Ghosh for their careful reading of the manuscript, and O. Hyrien and B. Michel for helpful discussions. This work was supported by the Centre National de la Recherche Scientifique (CNRS), the Agence Nationale de la Recherche (NT05-3_41825), the ACI IMPBIO 2004, the French Ministère de l'Education et de la Recherche, and the PAI Tournesol. B.A. acknowledges support from the European Commission Marie Curie action (MERG-CT-2004-511923).

References

Berezney, R., Dubey, D.D., and Huberman, J.A. 2000. Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. *Chromosoma* **108:** 471–484.

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.

Cajiao, I., Zhang, A., Yoo, E.J., Cooke, N.E., and Liebhaber, S.A. 2004. Bystander gene activation by a locus control region. *EMBO J.* **23:** 3854–3863.

Callan, H.G. 1972. Replication of DNA in the chromosomes of eukaryotes. *Proc. R. Soc. Lond.* **181:** 19–41.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., et al. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289–1292.

Chakalova, L., Debrand, E., Mitchell, J.A., Osborne, C.S., and Fraser, P. 2005. Replication and transcription: Shaping the landscape of the genome. *Nat. Rev. Genet.* **6:** 669–677.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–1154.

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

DePamphilis, M.L. 2005. Cell cycle dependent regulation of the origin recognition complex. *Cell Cycle* **4:** 70–79.

Deshpande, A.M. and Newlon, C.S. 1996. DNA replication fork pause sites dependent on transcription. *Science* **272:** 1030–1033.

Edenberg, H.J. and Huberman, J.A. 1975. Eukaryotic chromosome replication. *Annu. Rev. Genet.* **9:** 245–284.

Ghosh, M., Liu, G., Randall, G., Bevington, J., and Leffak, M. 2004. Transcription factor binding and induced transcription alter chromosomal c-myc replicator activity. *Mol. Cell. Biol.* **24:** 10193–10207.

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.

Green, P., Ewing, B., Miller, W., Thomas, P.J., and Green, E.D. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33:** 514–517.

Hurst, L.D., Pal, C., and Lercher, M.J. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5:** 299–310.

Jeon, Y., Bekiranov, S., Karnani, N., Kapranov, P., Ghosh, S., Macalpine, D., Lee, C., Hwang, D.S., Gingeras, T.R., and Dutta, A. 2005. Temporal profile of replication of human chromosomes. *Proc. Natl. Acad. Sci.* **102:** 6419–6424.

Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S., and Gingeras, T.R. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15:** 987–997.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lercher, M.J., Urrutia, A.O., and Hurst, L.D. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31:** 180–183.

Li, W.H., Gu, Z., Wang, H., and Nekrutenko, A. 2001. Evolutionary analyses of the human genome. *Nature* **409:** 847–849.

Lin, C.M., Fu, H., Martinovsky, M., Bouhassira, E., and Aladjem, M.I. 2003. Dynamic alterations of replication timing in mammalian cells. *Curr. Biol.* **13:** 1019–1028.

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.

Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, G., Gautier, C., and Bernardi, G. 1991. The distribution of genes in the human genome. *Gene* **100:** 181–187.

Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W., and Stubbs, L. 2005. Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15:** 137–145.

Rocha, E.P. and Danchin, A. 2003. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat. Genet.* **34:** 377–378.

Schubeler, D., Scalzo, D., Kooperberg, C., van Steensel, 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.

Sémon, M. and Duret, L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol. Biol. Evol.* **23:** 1715–1723.

Sémon, M., Mouchiroud, D., and Duret, L. 2005. Relationship between gene expression and GC-content in mammals: Statistical significance and biological relevance. *Hum. Mol. Genet.* **14:** 421–427.

Singer, G.A., Lloyd, A.T., Huminiecki, L.B., and Wolfe, K.H. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol. Biol. Evol.* **22:** 767–775.

Spellman, P.T. and Rubin, G.M. 2002. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* **1:** 5. doi: 10.1186/1475-4924-1-5.

Sproul, D., Gilbert, N., and Bickmore, W.A. 2005. The role of chromatin structure in regulating the expression of clustered genes. *Nat. Rev. Genet.* **6:** 775–781.

Takeuchi, Y., Horiuchi, T., and Kobayashi, T. 2003. Transcription-dependent recombination and the role of fork collision in yeast rDNA. *Genes & Dev.* **17:** 1497–1506.

Touchon, M., Nicolay, S., Arneodo, A., d'Aubenton-Carafa, Y., and Thermes, C. 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Lett.* **555:** 579–582.

Touchon, M., Arneodo, A., d'Aubenton-Carafa, Y., and Thermes, C. 2004. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Res.* **32:** 4969–4978.

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.

Vassilev, L. and Johnson, E.M. 1990. An initiation zone of chromosomal DNA replication located upstream of the c-myc gene in proliferating HeLa cells. *Mol. Cell. Biol.* **10:** 4899–4904.

Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J., and van Kampen, A.H. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13:** 1998–2004.

Wang, J.D., Berkmen, M.B., and Grossman, A.D. 2007. Genome-wide coorientation of replication and transcription reduces adverse effects on replication in *Bacillus subtilis*. *Proc. Natl. Acad. Sci.* **104:** 5608–5613.

White, E.J., Emanuelsson, O., Scalzo, D., Royce, T., Kosak, S., Oakeley, E.J., Weissman, S., Gerstein, M., Groudine, M., Snyder, M., et al. 2004. DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. *Proc. Natl. Acad. Sci.* **101:** 17771–17776.

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.

Yurov, Y.B. and Liapunova, N.A. 1977. The units of DNA replication in the mammalian chromosomes: evidence for a large size of replication units. *Chromosoma* **60:** 253–267.

Zoubak, S., Clay, O., and Bernardi, G. 1996. The gene distribution of the human genome. *Gene* **174:** 95–102.

Received March 22, 2007; accepted in revised form June 10, 2007.