Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Biochemical and Biophysical Research Communications 344 (2006) 1258–1262

BBRC

www.elsevier.com/locate/ybbrc

Replication-associated strand asymmetries in vertebrate genomes and implications for replicon size, DNA replication origin, and termination

Wen-Ru Hou, Hai-Fang Wang, Deng-Ke Niu *

Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences,
Beijing Normal University, Beijing 100875, China

Received 29 March 2006
Available online 24 April 2006

Abstract

Strand compositional asymmetry has been observed in prokaryotes and used in predicting prokaryotic DNA replication origins and termini. However, it was not found in eukaryotic genomes by the same methods. We propose that transcription-associated strand asymmetries mask the replication-associated ones. By analyzing the nucleotide composition of intergenic sequences larger than 50 kb by cumulative skew diagrams (CSD), we found replication-associated strand asymmetry in vertebrate genomes. Furthermore, we found that the most common replicon sizes in vertebrates are 50–100 kb, and show evidence that the replication origin and termination regions of vertebrate genomes range from a discrete site to a broad zone.

Keywords: Strand asymmetry; DNA replication origin and terminus; Replicon size; Cumulative skew diagram; Vertebrate genomes

Preferences for G over C and for T over A in the leading strand were observed in many bacterial, viral, and organelle genomes [1–7]. This strand asymmetry has been widely used to predict bacterial DNA replication origin regions [8–10]. In *Saccharomyces cerevisiae*, some evidence indicates that the lagging strand DNA replication errors are preferentially repaired [11,12]. However, the few studies of strand asymmetry in eukaryotic genomes have yielded inconsistent results. Early studies described local asymmetric substitution patterns in the β-globin region of six primate genomes [13], but a later study based on updated knowledge of the replication origin position within the β-globin region did not show the existence of strand asymmetric mutations [14].

Some methods, e.g., cumulative skew diagrams (CSD) [4], DNA walks [15], and *Z*-curves [16,17], can show the strand asymmetries of prokaryotic or archaeal genomes without reference to the experimentally identified replication origin positions. In fact, they can be used to predict the positions of the replication origin and terminus [4,15–17]. Unfortunately, such methods have not led to conclusive results for eukaryote genomes. This has often been attributed to the fact that a random choice from an excess of potential replicons is made in each replication cycle [5,18,19].

By analyzing the intergenic and transcribed sequences flanking the limited number of experimentally identified replication origins, replication-associated strand compositional asymmetries were recently revealed in mammalian genomes [20,21].

Our previous study [22] suggested that replication-associated strand asymmetries in eukaryotes, if they exist, are generally much weaker than the transcription-associated asymmetries that have been well demonstrated in eukaryotes [23–26]. Recent evidence indicates that transcription-related mutational pressure, together with DNA replication-associated strand asymmetric orientation of genes, may be the main source of replication-associated strand compositional asymmetry in bacterial genomes [27]. We suspect that the replication-associated strand

* Corresponding author. Fax: +86 10 58807721.
E-mail addresses: dkniu@bnu.edu.cn, dengkeniu@hotmail.com (D.-K. Niu).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.04.039

asymmetries may be partially masked by transcription-associated asymmetries. Here we reveal the replication-associated strand asymmetries in eukaryotic genomes by analyzing intergenic sequences (IGSs).

Materials and methods

To clearly reveal replication-associated strand asymmetries, it is better to select large sequences that are expected to comprise at least one replicon. As eukaryotic replicons are generally believed to be 40–100 kb in length [28], we analyzed IGSs larger than 50 kb by CSD. All the annotated eukaryotic genomes were downloaded from the NCBI GenBank database (ftp://ftp.ncbi.nih.gov). We found adequate number of large IGSs only from vertebrate genomes: *Danio rerio* (NCBI build 1 version 1), *Gallus gallus* (NCBI build 1 version 1), *Mus musculus* (NCBI build 32), *Rattus norvegicus* (NCBI build 2), *Canis familiaris* (NCBI build 1 version 1), *Pan troglodytes* (NCBI build 1 version 1), and *Homo sapiens* (NCBI build 34 version 3). In previous study, excluding repetitive elements did not change the results [24]; there seem to be very few recently inserted repetitive elements that have not yet reached compositional equilibrium. To accurately estimate replicon size, repetitive elements were not excluded from this study.

Strand compositional asymmetries in IGSs were analyzed by CSD [4]. The skew values [(C − G)/(C + G) and (A − T)/(A + T)] of adjacent 100 bp windows (1 kb windows give similar results) along the DNA sequence were consecutively added together and plotted. Both (C − G)/(C + G) and (G − C)/(C + G) should reveal asymmetries, but the former is expected to have the same sign as (A − T)/(A + T) [1,3,23,25,26], so for convenience, we use (C − G)/(C + G) rather than the more conventional (G − C)/(C + G). In CSDs [4], the origin and terminus of replication can be sensitively detected as the loci corresponding to distinct global extrema of a curve. The size of a replicon is twice the length between two loci that correspond to adjacent distinct global extrema. Only when AT skew and CG skew are positively correlated ($P < 0.05$), the CSD was used for further analysis.

It should be noted that the first two authors of this paper manually identified CSD shape and estimated the replicon size by eye. Whilst this may have introduced some arbitrariness or error, we do not believe these would be significant enough to weaken the general conclusions.

Results and discussion

Strand asymmetries in large eukaryotic intergenic sequences

Replication-associated strand asymmetry in bacterial genomes is characterized by two distinct global extrema in the V/inverted-V-shaped CSD [4]. As each eukaryotic chromosome contains many replicons [28], if replication-associated strand asymmetries exist, large IGSs and large genes would have V- or multiple-connected-V-shaped (abbreviated as V-shaped below) cumulative skew diagrams, each with one or several distinct global extrema. For an imaginary sequence with no changes other than a constant rate of replication-associated strand-specific substitutions, the diagram would consist of straight lines except for switches at global extrema [4]. In 24975 large IGSs parsed out from the seven genomes, there are a few with CSDs that looked like schematic diagrams for replication-associated strand asymmetries (Fig. 1 and Supplementary figures 1 and 2).

More commonly, the asymmetry is disturbed by other phenomena, like transcription (of coding or noncoding RNA), chromosome rearrangements, recent horizontal gene transfer or transposition [1,2,4,25,29,30]. The diagrams are expected to be smoother, rougher, or more disordered depending on the extent of the disturbances. In the mammalian and bird IGSs analyzed (Table 1), about two-thirds have disordered CSDs (Supplementary figure 3) and about one-third have V-shaped CSDs (Supplementary figures 1 and 2). A large proportion of fish IGSs have V-shaped CSDs (Table 1). In all the vertebrates analyzed, most of the V-shaped CSDs have only one distinct global extremum (Supplementary figure 1); only a few IGSs have multiple-connected-V-shaped CSDs (Supplementary figure 2), i.e., with two or more distinct global extrema (Table 1).

Much evidence suggests that transcription does occur outside the boundaries of known genes ([31,32], and references therein). Transcription of specific segments may partially contribute to the disordered CSDs observed in more than half of the sequences annotated as IGSs. The stronger effects of gene or noncoding sequence transcription (compared with DNA replication) may be the reason why traditional approaches failed to reveal replication-associated strand asymmetries in eukaryotes [4,5,18,19].

The V-shaped CSDs in large vertebrate IGSs reinforce the recent reports of replication-associated strand asymmetries in mammalian genomes [20,21]. However, the replicon sizes, and the origins and termini of replication revealed in our diagrams are different from those described previously [20,21].

Replicon sizes in vertebrate genomes

Animal replicons were generally believed to be about 100 kb in length, with variations of >10-fold within a genome [28]. However, it has been suggested that the average size of mammalian replicons may be up to 500 kb [33]. The most recent analyses based on strand compositional asymmetries indicated that mammalian replicon sizes may be about 1–2 Mb [20,21].

For bacterial genomes, cumulative AT skew diagrams are not as canonical as cumulative GC skew diagrams [4]. Compared with cumulative CG skew diagrams, the cumulative AT skew diagrams in vertebrate sequences are smooth, featureless curves (Supplementary figures 1 and 2). So we estimated replicon sizes by the positions of distinct global extrema in the cumulative AT skew diagrams of large IGSs. For example, in the CSD of the IGSs between the chicken genes *LOC427554* and *LOC415802* (Fig. 1), the maximum value of cumulative skew is at 214.5 kb and the minimum value at 597.6 kb. So the size of this replicon is estimated to be about 766 kb. Similarly, the replicon near gene *FLJ35863* and gene *HKR1* in the human genome is estimated to be about 90 kb (Fig. 1).

As shown in Fig. 2, our results support the classic view of replicon sizes [28]. In zebrafish and all the mammalian species we studied, the most common replicon sizes are between 50 and 100 kb. Large replicons of more than 1 Mb make up only a small fraction of the total in each

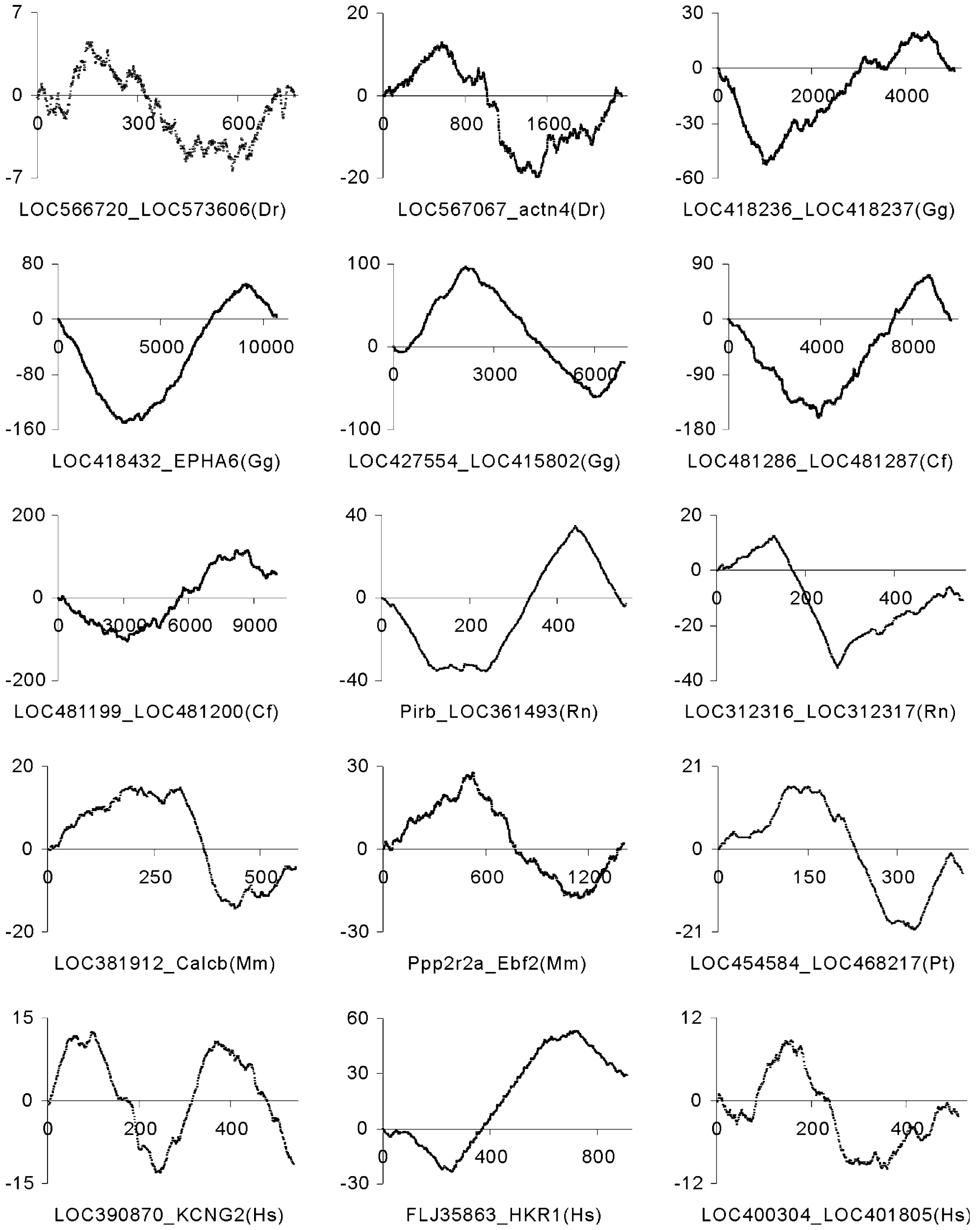

Fig. 1. Cumulative AT skew diagrams of some large vertebrate intergenic sequences (IGSs). The *X* axis represents the sequence length in 100 bp units and the *Y* axis represents the cumulative skews. The intergenic sequences are named by the genes flanking them. Species name abbreviations (in parentheses after the gene name): Gg, *Gallus gallus*; Cf, *Canis familiaris*; Rn, *Rattus norvegicus*; Mm, *Mus musculus*; Pt, *Pan troglodytes*; Hs, *Homo sapiens*.

Table 1
Percentage of large vertebrate intergenic sequences (IGSs) with different types of cumulative AT skew diagrams (CSDs)

Species	Number of IGSs studied[a]	Disordered CSDs	V-shaped CSDs with one distinct global extremum	V-shaped CSDs having two or more distinct global extrema
Danio rerio	1870	0.411	0.504	0.085
Gallus gallus	1695	0.763	0.224	0.014
Canis familiaris	3588	0.758	0.222	0.021
Mus musculus	4466	0.620	0.344	0.036
Rattus norvegicus	5336	0.689	0.272	0.039
Pan troglodytes	3672	0.756	0.227	0.018
Homo sapiens	4348	0.652	0.326	0.023

[a] We only counted the IGSs with positively correlated AT skew and CG skew ($P < 0.05$).

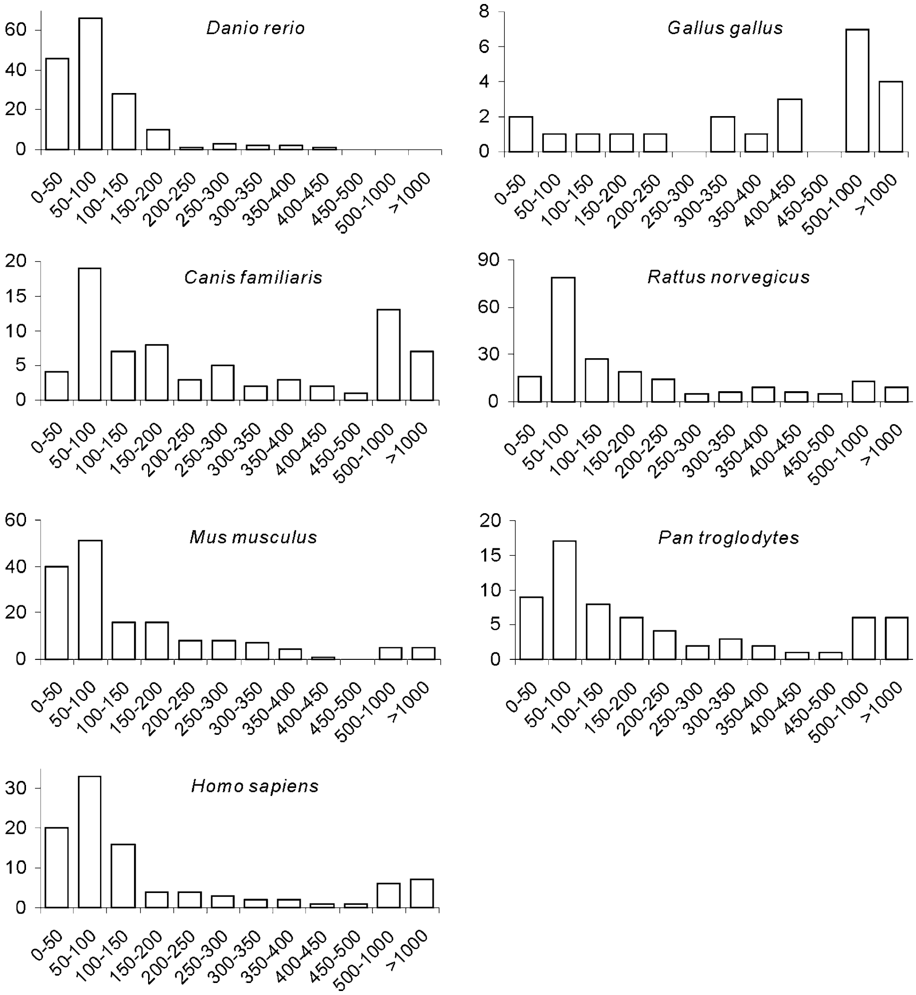

Fig. 2. The replicon size distribution of vertebrate genomes. The *X* axis shows replicon size in 1 kb units while the *Y* axis represents the number of replicons observed.

mammalian species. The replicon sizes estimated for chicken are slightly different from those of other species (Fig. 2). However, as only a limited number of replicons were analyzed, it is not possible to conclude whether DNA replication in birds differs slightly from that of other vertebrates.

It should be noted that the IGSs with disordered CSDs (Supplementary figure 3) may be combination of replicons with different sizes. But for a validating estimation of replicon sizes, we can only consider the clear pattern, i.e., one IGS with one replicon size, which gives canonical CSD.

Replication origin and terminus: random or fixed

In the *Z*-curves of some archaea, there are two peaks: one sharp and one broad. Taking this and other evidence into account, Zhang and Zhang [16] suggested that the sharp peak corresponds to the replication origin, and therefore the broad peak corresponds to the replication terminus. Similar jagged changes in strand compositional asymmetries were observed near some termini of mammalian DNA replication [20]. The hypothesis of random termination of DNA replication was proposed [20]. The replication of the *Escherichia coli* genome terminates at specific terminator sequences [34]. Consistent with such a hypothesis, there is a sharp peak in the CSD corresponding to the replication terminus (see the figure of reference [4]). Conversely, if vertebrate DNA replication can initiate at broad zones of potential sites rather than at a single discrete site, the corresponding peaks in CSDs are expected to be broad and jagged.

Following the rationale of G over C and for T over A in the leading strand in bacterial genomes [1–7], we can distinguish leading strand segment and lagging strand segment in each IGS by CSD. Together with the 5′ to 3′ direction of DNA replication, we predict replication origin corresponds to the global maximum while the terminus corresponds to the global minimum of the cumulative skew.

It is very easy to find broad peaks and sharp peaks in both replication origins and termini (Fig. 1 and Supplementary figures 1 and 2). But the IGS length would affect the visual appearance of the peaks. A large IGS may have sharp peaks because we have to compact its CSD into the same size with other IGS. Similarly, a short IGS may have broad peaks because its CSD was expanded.

Examples are the chicken IGS between *LOC481286* and *LOC481287* and the chimpanzee IGS between *LOC454584* and *LOC468217* (Fig. 1). Still, we can easily find sharp peaks in short IGSs (e.g., those less than 100 kb) and broad peaks in large IGSs (e.g., those more than 1000 kb), indicating that both replication origin and termini of vertebrate genomes vary in a spectrum ranging from a discrete site to a broad zone. These results are consistent with data on the few experimentally identified mammalian replication origins, which cover a spectrum ranging from tightly circumscribed to extremely broad initiation zones [35].

Acknowledgments

We thank Wei Liu, Kui Lin, and Da-Yong Zhang for help and discussion. This work was supported by the National Natural Science Foundation of China (30270695) and Beijing Normal University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.04.039.

References

[1] A.C. Frank, J.R. Lobry, Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms, Gene 238 (1999) 65–77.

[2] M.P. Francino, H. Ochman, Strand asymmetries in DNA evolution, Trends Genet. 13 (1997) 240–245.

[3] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria, Mol. Biol. Evol. 13 (1996) 660–665.

[4] A. Grigoriev, Analyzing genomes with cumulative skew diagrams, Nucleic Acids Res. 26 (1998) 2286–2290.

[5] J. Mrazek, S. Karlin, Strand compositional asymmetry in bacterial and large viral genomes, Proc. Natl. Acad. Sci. USA 95 (1998) 3720–3725.

[6] E.P.C. Rocha, A. Danchin, A. Viari, Universal replication biases in bacteria, Mol. Microbiol. 32 (1999) 11–16.

[7] J. Lobry, N. Sueoka, Asymmetric directional mutation pressures in bacteria, Genome Biol. 3 (2002), research0058.

[8] P. Mackiewicz, J. Zakrzewska-Czerwinska, A. Zawilak, M.R. Dudek, S. Cebrat, Where does bacterial replication start? Rules for predicting the *oriC* region, Nucleic Acids Res. 32 (2004) 3781–3791.

[9] J.R. Lobry, Origin of replication of *Mycoplasma genitalium*, Science 272 (1996) 745–746.

[10] L. Li, J.P. Bannantine, Q. Zhang, A. Amonsin, B.J. May, D. Alt, N. Banerji, S. Kanjilal, V. Kapur, The complete genome sequence of *Mycobacterium avium* subspecies *paratuberculosis*, Proc. Natl. Acad. Sci. USA 102 (2005) 12344–12349.

[11] Y.I. Pavlov, I.M. Mian, T.A. Kunkel, Evidence for preferential mismatch repair of lagging strand DNA replication errors in yeast, Curr. Biol. 13 (2003) 744–748.

[12] Y.I. Pavlov, C.S. Newlon, T.A. Kunkel, Yeast origins establish a strand bias for replicational mutagenesis, Mol. Cell. 10 (2002) 207–213.

[13] C.I. Wu, N. Maeda, Inequality in mutation-rates of the 2 strands of DNA, Nature 327 (1987) 169–170.

[14] M.P. Francino, H. Ochman, Strand symmetry around the β-globin origin of replication in primates, Mol. Biol. Evol. 17 (2000) 416–422.

[15] J.R. Lobry, A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria, Biochimie 78 (1996) 323–326.

[16] R. Zhang, C.T. Zhang, Single replication origin of the archaeon *Methanosarcina mazei* revealed by the *Z* curve method, Biochem. Biophys. Res. Commun. 297 (2002) 396–400.

[17] C.T. Zhang, R. Zhang, H.Y. Ou, The *Z* curve database: a graphic representation of genome sequences, Bioinformatics 19 (2003) 593–599.

[18] S. Karlin, A.M. Campbell, J. Mrazek, Comparative DNA analysis across diverse genomes, Annu. Rev. Genet. 32 (1998) 185–225.

[19] A. Gierlik, M. Kowalczuk, P. Mackiewicz, M.R. Dudek, S. Cebrat, Is there replication-associated mutational pressure in the *Saccharomyces cerevisiae* genome? J. Theor. Biol. 202 (2000) 305–314.

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

[21] E.-B. Brodie of Brodie, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, From DNA sequence analysis to modeling replication in the human genome, Phys. Rev. Lett. 94 (2005) 248103.

[22] D.K. Niu, K. Lin, D.-Y. Zhang, Strand compositional asymmetries of nuclear DNA in eukaryotes, J. Mol. Evol. 57 (2003) 325–334.

[23] P. Green, B. Ewing, W. Miller, P.J. Thomas, NISC Comparative Sequencing Program, E.D. Green, Transcription-associated mutational asymmetry in mammalian evolution, Nat. Genet. 33 (2003) 514–517.

[24] J. Majewski, Dependence of mutational asymmetry on gene-expression levels in the human genome, Am. J. Hum. Genet. 73 (2003) 688–692.

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

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

[27] C. Nikolaou, Y. Almirantis, A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species, Nucleic Acids Res. 33 (2005) 6816–6822.

[28] B. Lewin, Gene VIII, Pearson Prentice Hall, Upper Saddle River, New Jersey, 2004.

[29] J.M. Freeman, T.N. Plasterer, T.F. Smith, S.C. Mohr, Patterns of genome organization in bacteria, Science 279 (1998) 1827a.

[30] A. Grigoriev, Graphical genome comparison—rearrangements and replication origin of *Helicobacter pylori*, Trends Genet. 16 (2000) 376–378.

[31] J.-M. Claverie, Fewer genes, more noncoding RNA, Science 309 (2005) 1529–1530.

[32] J.M. Johnson, S. Edwards, D. Shoemaker, E.E. Schadt, Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments, Trends Genet. 21 (2005) 93–102.

[33] R. Berezney, D.D. Dubey, J.A. Huberman, Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci, Chromosoma 108 (2000) 471–484.

[34] T.A. Brown, Genomes, second ed., John Wiley, New York, 2002.

[35] Y.J. Machida, J.L. Hamlin, A. Dutta, Right place, right time, and only once: replication initiation in metazoans, Cell 123 (2005) 13–24.