Transcription-coupled TA and GC strand asymmetries in the human genome

M. Touchon[a], S. Nicolay[b], A. Arneodo[b], Y. d'Aubenton-Carafa[a], C. Thermes[a,*]

[a] *Centre de Génétique Moléculaire, CNRS, Allée de la Terrasse, 91198 Gif-sur-Yvette, France*
[b] *Laboratoire de Physique, Ecole Normale Supérieure de Lyon, 69364 Lyon, France*

Received 22 September 2003; accepted 14 October 2003

First published online 24 November 2003

Edited by Lev Kisselev

Abstract **Analysis of the whole set of human genes reveals that most of them present TA and GC skews, that these biases are correlated to each other and are specific to gene sequences, exhibiting sharp transitions between transcribed and non-transcribed regions. The GC asymmetries cannot be explained solely by a model previously proposed for (G+T) skew based on transitions measured in a small set of human genes. We propose that the GC skew results from additional transcription-coupled mutation process that would include transversions. During evolution, both processes acting on a large majority of genes in germline cells would have produced these transcription-coupled strand asymmetries.**

Key words: Skewness; Strand asymmetry; Mutation bias; Human genome

1. Introduction

Under no strand bias conditions, the following equalities A = T and G = C should be observed in each DNA strand [1]. These deviations from intrastrand equimolarities have been extensively studied in prokaryotic, organelle and viral genomes and they have been used to detect the origin of replication in a number of eubacterial genomes [2–4]. In these genomes the leading strand is relatively enriched in G over C and T over A in positions under weak selective pressure (intergenic regions and third codon position). These asymmetries have been attributed to various mechanisms. In some models based on replication, the leading and lagging strands are subject to different mutational and repair pressures resulting in asymmetric nucleotide compositions [4]. Other models emphasize C to T deamination and transcription-coupled repair in the coding strand. Transcription increases single-strand deamination [5] and favors transcription-coupled repair, leading to pronounced strand asymmetries [6]. In eukaryotes, studies of strand compositional asymmetry have led to contrasting observations. Unlike eubacterial genomes, the yeast genome did not present strand asymmetry except in subtelomeric regions of chromosomes [7]. In primates, a comparative study of the β-globin replication origin did not support the existence of a replication-coupled mutational bias [8]. In human, excess of T was observed in a set of gene introns [9] and some large-scale asymmetry was observed in human sequences but they were attributed to replication [10]. Recently, a comparative analysis of mammalian sequences demonstrated a transcription-coupled excess of G+T over A+C in the coding strand [11]. In contrast to the substitution biases observed in bacteria presenting an excess of C→T transitions, these asymmetries are characterized by an excess of purine transitions and a deficit of pyrimidine transitions. These might be a by-product of the transcription-coupled repair mechanism [12] acting on uncorrected substitution errors during replication. Here, we investigated human sequences for TA and GC strand asymmetries and for their possible association with transcription. In most gene sequences we observed an excess of T over A and of G over C, specifically in transcribed regions, reflecting transcriptional-coupled mutational processes in the germ line. Our results further indicate that strand biases do not uniquely result from transitions producing (G+T)/(A+C) asymmetries, but that the TA and GC skews likely result from different mechanisms also involving transversions.

2. Materials and methods

2.1. Sequences

Human intron sequences were downloaded from RefGene (April 2003) at UCSC. When several genes presented identical exonic sequences, only the longest one was retained; repeated elements were removed with RepeatMasker. The introns of each gene were taken as a single sequence; introns without repeats were also taken as a single sequence; to avoid the effects of selective pressure on splice site sequences, 30 nt were removed at both intron extremities. When the resulting intron sequences were shorter than 100 bp, they were not considered for the analysis, leading to 14 854 intron-containing genes (Fig. 1).

2.2. Determination of the nucleotide composition at equilibrium

The calculation was based on the values of the neutral substitution rates measured in a previous study of repeated elements in the human genome [13]; we used the substitution rates that correspond to a G+C content equal to 43%, close to the 45% G+C content of the intronic sequences studied here (Fig. 4). Among these rate values, the transition rates t_i were modified as follows:

$$r_i = t_i\,(Srepeats_i/Suntr_i).$$

where $Srepeats_i$ are the transition rates measured in transcribed repeats and $Suntr_i$ are those measured in untranscribed regions [11].

The calculation of the nucleotide composition at equilibrium was then performed with these values as described in [14] and led to the following values: $S_{TA} = 4.7\%$ and $S_{GC} = 7.8\%$. In order to obtain values at equilibrium close from those observed here in intron sequences (without repeats), we could not only modify the transition rates but transversion rates had also to be altered. In order to reduce the GC skew, the value of the G→C mutation rate was enhanced from 2.9%

*Corresponding author. Fax: (33)-1-69 82 38 77.
E-mail address: thermes@cgm.cnrs-gif.fr (C. Thermes).

0014-5793/03/$22.00
doi:10.1016/S0014-5793(03)01306-1

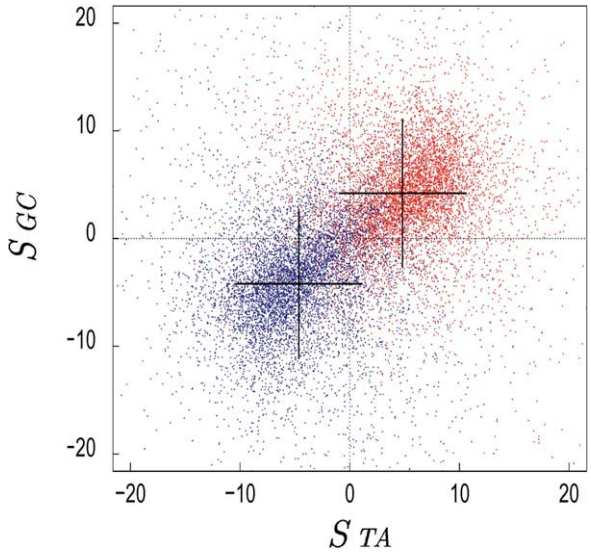

Fig. 1. TA and GC strand asymmetries in human introns. Scatter diagram of S_{TA} and S_{GC} skews of intronic regions. Each point corresponds to one of the 14 854 intron-containing genes; repeated elements are removed from the analysis (Section 2); red points, genes with (+) orientation (same orientation as the Watson strand, 7508 genes) ($\overline{S}_{TA} = 4.81\%$, $\overline{S}_{GC} = 4.20\%$); blue points, genes with (−) orientation (7346 genes) ($\overline{S}_{TA} = -4.64\%$, $\overline{S}_{GC} = -4.20\%$); black crosses represent the standard deviations of the distributions.

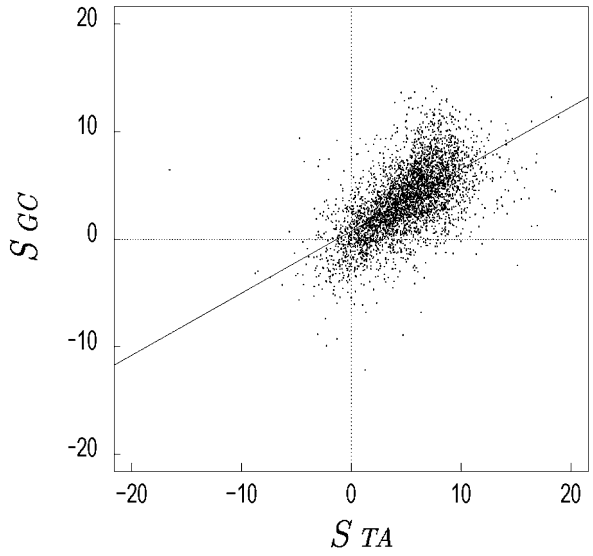

Fig. 2. Correlation between TA and GC strand asymmetries. Scatter diagram of S_{TA} and S_{GC} skews determined in the coding strand for intronic regions without repeats; each point corresponds to a gene for which the total length of intronic regions is $l > 25$ kb (7797 genes); Pearson's $r = 0.61$ (the slope of the regression line is 0.58).

to 3.36% (simultaneously the C→T transition rate was reduced from 14.4% to 13.7%) leading to final skews (Fig. 4).

3. Results and discussion

3.1. Strand asymmetries in gene sequences

We examined human sequences for nucleotide compositional strand asymmetries and in particular for the deviations from intrastrand parity rules T = A and G = C calculated as $S_{TA} = (T\text{-}A)/(T+A)$ and $S_{GC} = (G\text{-}C)/(G+C)$. These skews were calculated for intron sequences which can be considered as weakly selected sequences. For each gene, all introns were concatenated and considered as a unique sequence (Section 2). The distributions of the TA and GC skews presented positive mean values when the genes were transcribed in the same direction as the Watson strand (+), $\overline{S}_{TA} = 4.72 \pm 0.07\%$ and $\overline{S}_{GC} = 2.97 \pm 0.07\%$, and nearly opposed values when the genes were transcribed in the opposed direction (−), $\overline{S}_{TA} = -4.56 \pm 0.07\%$ and $\overline{S}_{GC} = -3.05 \pm 0.07\%$ (mean values were calculated for all genes). When repeated sequences were removed from the analysis (Section 2) the TA biases were not strongly altered but the modulus of the GC biases was significantly enhanced (Fig. 1). This was due to repeated elements that constitute 43% of intronic sequences and that present small GC bias (Table 1). When examined on the coding strand, the mean values for all intron sequences (without repeats) presented significant excess of T over A and almost equal excess of G over C (Table 1).

We searched for possible correlation between S_{TA} and S_{GC} biases measured in intronic sequences (without repeats). When all genes were considered, small correlation could be observed ($r = 0.09$). However, the values of the skews for small genes presented strong stochastic noise. When these genes were not considered, S_{TA} and S_{GC} presented larger correlation. For example, for genes with intron length $l > 10$ kb (l is the total length of intron sequences without repeats) the correlation coefficient was $r = 0.45$; for $l > 25$ kb, we obtained $r = 0.61$ (Fig. 2). We also observed that S_{TA} and S_{GC} presented weak correlation with the intronic GC content (Table 1) even when only large genes were considered (data not shown). Similarly, weak correlation was observed between the TA and GC skews on one hand and the sequence length on the other hand (Table 1). This property remained true when only large genes were analyzed (data not shown).

3.2. TA and GC biases are associated to transcribed regions

TA and GC asymmetries were examined along the genome sequence (in 1 kb windows) in order to compare their values

Table 1
Strand asymmetries in human introns

	l kb	(G+C)%	S_{TA}	S_{GC}	r_{TA}	r_{GC}
Introns	54.8	45.8	4.64 (±0.05)	3.01 (±0.05)	0.006 −0.063	−0.057 0.073
Introns (− repeats)	31.2	45.0	4.73 (±0.05)	4.20 (±0.06)	−0.023 −0.038	−0.062 −0.005
Intronic repeats	25.2	45.8	4.41 (±0.09)	1.24 (±0.08)	0.035 −0.071	−0.006 0.050

l, mean value of the total length of the sequences indicated in the first column for all examined genes (Section 2); G+C, mean value of the GC content of the same sequences; S_{TA} and S_{GC} are given in percent ± S.E.M.; for each type of sequences, r_{TA} (r_{GC}) is the Pearson's correlation coefficient between S_{TA} (S_{GC}) and the length (first line) or GC content (second line) of the indicated sequences.

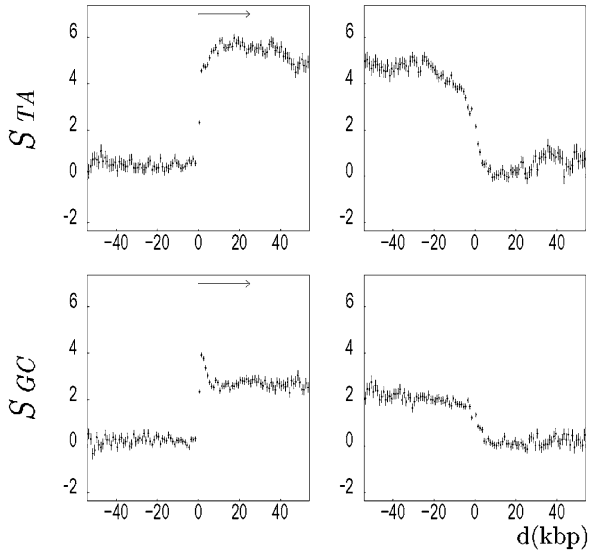

Fig. 3. TA and GC skew profiles in the regions surrounding 5′ and 3′ gene extremities. The values of S_{TA} and S_{GC} were calculated in 1 kb windows starting from each gene extremity in both directions. In abscissa is figured the distance (d) of each 1 kb window to the indicated gene extremity; zero values of abscissa correspond to 5′- (left panels) or 3′- (right panels) gene extremities. In ordinate is reported the mean value of the skews for all 1 kb windows at the corresponding abscissa. Error bars represent the standard error of the means; arrows indicate the transcription start site.

in transcribed regions to those in the neighboring intergenic sequences. The mean values of the biases for all genes were reported as function of the distance to the 5′- or 3′-end (Fig. 3). At the 5′ gene extremities, sharp transitions of the skews were observed from about zero values to values in transcribed regions ranging between 4 and 6% for $\overline{S}_{TA}$ and between 2 and 4% for $\overline{S}_{GC}$. Interestingly, the GC skew profile presented a 2–3 kb wide peak starting at the transcription start site and decreasing in the direction of transcription. This could result from abortive transcripts leading to increased transcription rate near the gene 5′-end, but the length of such transcripts is much shorter than 2–3 kb [15]. In addition, the fact that the TA skew profile did not present a similar peak made this possibility unlikely. Alternatively, this peak could be associated to particular sequence elements situated at the gene 5′-end, like CpG islands [16] or to some unknown subclass of genes with highly GC-biased 5′-extremities. At the gene 3′-extremities the TA and GC skews also presented transitions from significantly large values in transcribed regions to very small values in untranscribed regions. Conversely to the steep transitions observed at 5′-ends, the 3′-end profiles exhibited a smooth transition pattern extending on 10–15 kb and including regions downstream of the 3′-end. This could reflect the fact that transcription continues to some extent downstream of the polyadenylation site which generally corresponds to the gene 3′-extremity annotated in the databank. When averaged over numerous genes, the strand bias would then progressively decrease downstream of the gene 3′ end. Such pattern of the skew profiles might also be due to the fact that a number of genes present several polyadenylation sites that can be used differently in various cell types [17]. In pluricellular organisms, mutations responsible for the observed biases are expected to occur in germline cells. It could happen that gene 3′-ends annotated in the databank differ from the polyA sites effectively used in germline cells. Such differences would then lead to some broadening of the skew profile.

3.3. A model for the TA and GC biases

TA and GC biases were specifically observed in transcribed sequences indicating that each of them clearly resulted from transcription-coupled processes acting in germline cells. This observation was reinforced by the correlation between S_{TA} and S_{GC} (Fig. 2). Indeed, according to such hypothesis S_{TA} and S_{GC} were expected to increase simultaneously with transcription. How many genes presented biased sequences? When the observed biases were compared to those expected for random sequences with same length and same (T+A) composition, we observed that 64% of genes presented significant TA bias ($P < 10^{-2}$). When only larger genes were analyzed, (intron length $l > 10$ kb) this proportion increased to 82%; for $l > 25$ kb, it raised to 86%. These results indicate that in germline cells a large majority of genes are expressed.

A recent study showed a transcription-coupled excess of purine transitions and a deficit of pyrimidine transitions in a small set of human genes [11]. To examine if these transition rates might explain the strand asymmetries measured here in the whole set of human genes, we performed numerical calculation of the composition at equilibrium of a DNA sequence (given the substitution rates). We supposed that transcription altered the transition rates to generate strand asymmetries as proposed by [11] but that transversion rates were not altered and remained identical to those in the non-transcribed regions [13]. We then obtained $S_{TA} = 4.7\%$ and $S_{GC} = 7.8\%$ (Section 2, Fig. 4). This value of S_{TA} was similar to our observations. However, S_{GC} was much larger than that observed here. Several reasons might explain this discrepancy. First, the excess of GC bias might result from non-stationarity with respect to the transcription-coupled mutational process. Indeed, compositional non-equilibrium has been observed in the human genome [13,18]. This hypothesis would suggest that the processes responsible for the TA and GC skews present significant differences although each of them is coupled to transcription. A second possibility would be that the transcription-coupled transition rates vary along the genome as previously observed for mutation rates in untran-

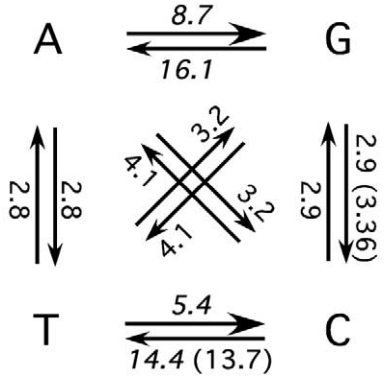

Fig. 4. Substitution rates used in the calculation of nucleotide composition at equilibrium (Section 2). The values (%) were first taken from [13] (fig. 27); the transition rates were then modified according to their values in transcribed regions, given in [11]; resulting transition rates are shown in italics; these rates led to: $S_{TA} = 4.7\%$ and $S_{GC} = 7.8\%$. When substitution rates G→C and C→T were replaced by the values shown in parentheses, we obtained the final values $S_{TA} = 4.8\%$ and $S_{GC} = 4.3\%$ similar to the observed skews.

scribed regions [13,18]. In this case, the transition rates measured in [11] in transcribed regions of a 1.5 Mb segment would only partially reflect these rates at the genome scale considered here. Along this line, we examined if some modification of these transition rates in the calculation process could lead to our measured values of the skews. However, modifying the transition rates could not lead to smaller values of S_{GC} without modifying S_{TA}, except if the transversion rates were also modified. We thus supposed the existence of an additional transcription-coupled process leading to transversions. Such process is still unknown in eukaryotes. However, it has been observed in some bacterial genomes that, when transitions are not considered, GC transversions lead to the largest strand asymmetry [19]. We thus supposed that GC transversions might also produce strand asymmetry in eukaryotes. In order to lower the value of the GC skew obtained in the calculation, we increased the G→C mutation rate, which finally led to calculated values of the skews similar to our observed values (Fig. 4). These observations sustain a modification of the model proposed by [11]. For these authors the observed transition rates would result from transcription-coupled repair [12] acting on mismatched base pairs due to replication errors, leading to their observation of (G+T)/(A+C) asymmetry. We propose that in parallel to this process, additional transcription-coupled mutation and/or repair mechanism leading to G→C transversions would have also acted to finally produce the GC skew observed here. This justifies that the TA and GC skews are analyzed separately, and not considered as resulting from a unique process leading to (G+T)/(A+C) asymmetry. During evolution, both processes would have been active on a vast majority of genes in germline cells.

The absence of asymmetries in intergenic regions does not exclude the possibility of additional replication-associated biases. Such biases would present opposed signs on leading and lagging strands and would cancel each other in our analyses as a result of multiple unknown replication origins [20]. However, replication-coupled biases present small values compared those measured in gene sequences as already observed for the β-globin replication origin [8] and confirmed by our examination of S_{TA} and S_{GC} profiles in intergenic sequences (data not shown). This further indicates that transcription-coupled TA and GC biases constitute the major strand compositional asymmetries in the human genome.

Acknowledgements: This work was supported by the Action Bioinformatique Inter EPST 2002, by the Centre National de la Recherche Scientifique and by the French Ministère de la Jeunesse, de l'Education et de la Recherche. We thank Eduardo Rocha for invaluable discussions about this manuscript.

References

[1] Lobry, J.R. (1995) J. Mol. Evol. 40, 326–330.
[2] Grigoriev, A. (1998) Nucleic Acids Res. 26, 2286–2290.
[3] Mrazek, J. and Karlin, S. (1998) Proc. Natl. Acad. Sci. USA 95, 3720–3725.
[4] Frank, A.C. and Lobry, J.R. (1999) Gene 238, 65–77.
[5] Beletskii, A. and Bhagwat, A.S. (1998) Biol. Chem. 379, 549–551.
[6] Francino, M.P. and Ochman, H. (1997) Trends Genet. 13, 240–245.
[7] Gierlik, A., Kowalczuk, M., Mackiewicz, P., Dudek, M.R. and Cebrat, S. (2000) J. Theor. Biol. 202, 305–314.
[8] Francino, M.P. and Ochman, H. (2000) Mol. Biol. Evol. 17, 416–422.
[9] Duret, L. (2002) Curr. Opin. Genet. Dev. 12, 640–649.
[10] Shioiri, C. and Takahata, N. (2001) J. Mol. Evol. 53, 364–376.
[11] Green, P., Ewing, B., Miller, W., Thomas, P.J. and Green, E.D. (2003) Nat. Genet. 33, 514–517.
[12] Svejstrup, J.Q. (2002) Nat. Rev. Mol. Cell. Biol. 3, 21–29.
[13] Lander, E.S. et al. (2001) Nature 409, 860–921.
[14] Sueoka, N. (1995) J. Mol. Evol. 40, 318–325.
[15] Dvir, A. (2002) Biochim. Biophys. Acta 1577, 208–223.
[16] Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) Genomics 13, 1095–1107.
[17] Beaudoing, E. and Gautheret, D. (2001) Genome Res. 11, 1520–1526.
[18] Smith, N.G., Webster, M.T. and Ellegren, H. (2002) Genome Res. 12, 1350–1356.
[19] Rocha, E.P. and Danchin, A. (2001) Mol. Biol. Evol. 18, 1789–1799.
[20] Berezney, R., Dubey, D.D. and Huberman, J.A. (2000) Chromosoma 108, 471–484.