Journal of Integrative Plant Biology 2007, **49** (9): 1393–1399

Evidence that Natural Selection is the Primary Cause of the Guanine-cytosine Content Variation in Rice Genes

Xiaoli Shi[1,2*], Xiyin Wang[2,3*,], Zhe Li[2], Qihui Zhu[2], Ji Yang[4], Song Ge[1***] and Jingchu Luo[2***]**

([1]*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, the Chinese Academy of Sciences*, Beijing 100093, China;
[2]*College of Life Sciences, National Laboratory of Plant Genetic Engineering and Protein Engineering, Center of Bioinformatics, Peking University*, Beijing 100871, China;
[3]*College of Mathematics, Hebei Polytechnic University,* Tangshan 063009, China;
[4]*Center for Evolutionary Biology, School of Life Sciences, Fudan University*, Shanghai 200433, China)

Abstract

Cereal genes are classified into two distinct classes according to the guanine-cytosine (GC) content at the third codon sites (GC_3). Natural selection and mutation bias have been proposed to affect the GC content. However, there has been controversy about the cause of GC variation. Here, we characterized the GC content of 1 092 paralogs and other single-copy genes in the duplicated chromosomal regions of the rice genome (ssp. *indica*) and classified the paralogs into GC_3-rich and GC_3-poor groups. By referring to out-group sequences from *Arabidopsis* and maize, we confirmed that the average synonymous substitution rate of the GC_3-rich genes is significantly lower than that of the GC_3-poor genes. Furthermore, we explored the other possible factors corresponding to the GC variation including the length of coding sequences, the number of exons in each gene, the number of genes in each family, the location of genes on chromosomes and the protein functions. Consequently, we propose that natural selection rather than mutation bias was the primary cause of the GC variation.

Key words: guanine-cytosine content; mutation bias; natural selection; paralogs; synonymous substitution rate; two gene classes.

Shi X, Wang X, Li Z, Zhu Q, Yang J, Ge S, Luo J (2007). Evidence that natural selection is the primary cause of the guanine-cytosine content variation in rice genes. *J. Integr. Plant Biol.* **49**(9), 1393–1399.

Available online at www.blackwell-synergy.com/links/toc/jipb, www.jipb.net

As one of the most accessible characteristics of genes, guanine-cytosine (GC) content and its variation have been widely discussed since the 1970s (Filipski et al. 1973). Generally, there is often extreme heterogeneity in GC content among genes, especially at the third codon site (GC_3). Synonymous codons ending with G and C are highly favored in Drosophila (Powell and Moriyama 1997), and in rice (Wang et al. 2002), whereas in humans, some amino acids are often coded by codons ending with G and C, but others with A and T (Lander et al. 2001). Akashi and Schaeffer (1997) proposed that natural selection preferred the synonymous substitutions producing codons to increase translational accuracy. This is supported by the observation that frequently used codons tend to have more copies of tRNA genes (Wang et al. 2002), and the corresponding tRNAs are in higher abundance in the cell (Moriyama and Powell 1997). However, the variation in base composition can also arise from mutation bias (Eyre-Walker 1997; Hurst and Williams 2000; Eyre-Walker and Hurst 2001). For example, when the nucleotides G and C are preferred, mutation bias can occur for preferentially mis-incorporating G and C into DNA chains during

Received 15 Jan. 2007 Accepted 28 Apr. 2007
Supported by the State Key Basic Research and Development Plan of China (2003CB715900), and the National Natural Science Foundation of China (90408015, 30121003 and 30430030).

*These authors contributed equally to this work.

**Present address: Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602, USA

***Authors for correspondence.
Jingchu Luo
Tel: +86 (0)10 6275 7281;
Fax: +86 (0)10 62759001;
E-mail: <luojc@pku.edu.cn>.
Song Ge
Tel: +86 (0)10 6283 6097;
Fax: +86 (0)10 6259 0843;
E-mail: <gesong@ibcas.ac.cn>.

doi: 10.1111/j.1672-9072.2007.00532.x

the process of DNA replication, or for preferentially changing the mismatched bases towards G and C during DNA repairing (Brown and Jiricny 1988).

Compared with dicot genes, cereal genes have an elevation in GC content, and are classified into two distinct classes (Carels and Bernardi 2000). The two classes of genes are supposed to be functionally different, implying the possible effect of natural selection. Instead of dividing cereal genes into two classes, Wong et al. (2002) suggested that the elevation in GC content had a negative gradient along the direction of transcription. They further attributed the GC elevation to transcription-related mutation bias or translation-related selection. Recently, Wang et al. (2004) analyzed 7 886 rice genes and compared them with the *Arabidopsis* genes. They affirmed the existence of two gene classes and proposed that mutation bias rather than natural selection was the primary cause of two classes of genes in cereals.

Here, we made our analysis on rice genes in large duplicated chromosomal segments, which accounted for more than half of the rice genome. We systematically analyzed collinear genes in duplicated regions produced by two large-scale duplication events, one ancient genomic and one relatively recent segmental duplication (Wang et al. 2005). With the help of outgroup sequences from *Arabidopsis* and maize, we characterized the pattern of GC content as well as nucleotide substitution bias and nucleotide substitution rate in paralogous gene pairs. Moreover, we made a genome-context analysis on the GC content in the coding sequences, introns and intergenic regions. These analyses provided us with a valuable opportunity to trace the cause of the formation of two gene classes in cereals.

Results

We computed GC content at all three codon sites (GC_1, GC_2, GC_3), as well as in intron (GC_{intron}) and intergenic (GC_{inter}) regions. Our results showed that both GC_1 and GC_2 had uni-modal distributions, ranging from 0.40 to 0.80 with a peak at 0.55 (GC_1) and from 0.25 to 0.65 with a peak at 0.43 (GC_2). Different from GC_1 and GC_2, GC_3 had a bi-modal distribution spreading from 0.27 to 1.00 (Figure 1). This extreme compositional heterogeneity among three codon sites supported the former observation (Carels and Bernardi 2000). The obvious bi-modality of GC_3 suggested the existence of two distinct classes of rice genes. Since the sister paralogs had an approximate GC_3 composition ($t = 1.319$, $P = 0.187$), two classes of genes existed in inter-paralogs rather than intra-paralogs. We carried out a hierarchical average clustering analysis to partition all paralogs into two classes at $GC_3 = 0.75$, resulting in 469 GC_3-rich and 623 GC_3-poor pairs. Further analysis revealed that the GC_3 bimodality existed in all duplicated chromosomal regions throughout the genome. We also calculated the GC distribution in intron (GC_{intron}) and intergenic (GC_{inter}) regions and found that GC_{intron} had a uni-modal distribution ranging from 0.20 to 0.80 with a peak at 0.36, and GC_{inter} had a uni-modal distribution spreading roughly in the same range as GC_{intron}.

We calculated the correlation between the third codon site and the other sites. GC_3 was positively correlated to GC_1 ($rho = 0.576$, $P < 2.2 \times 10^{-16}$) and GC_2 ($rho = 0.473$, $P = 2.2 \times 10^{-16}$). GC_3 was also positively correlated to GC_{intron} but the correlation was much weaker ($rho = 0.291$, $P = 0.284$). GC_3 was not correlated to GC_{inter} ($rho = 0.007$, $P < 2.2 \times 10^{-16}$).

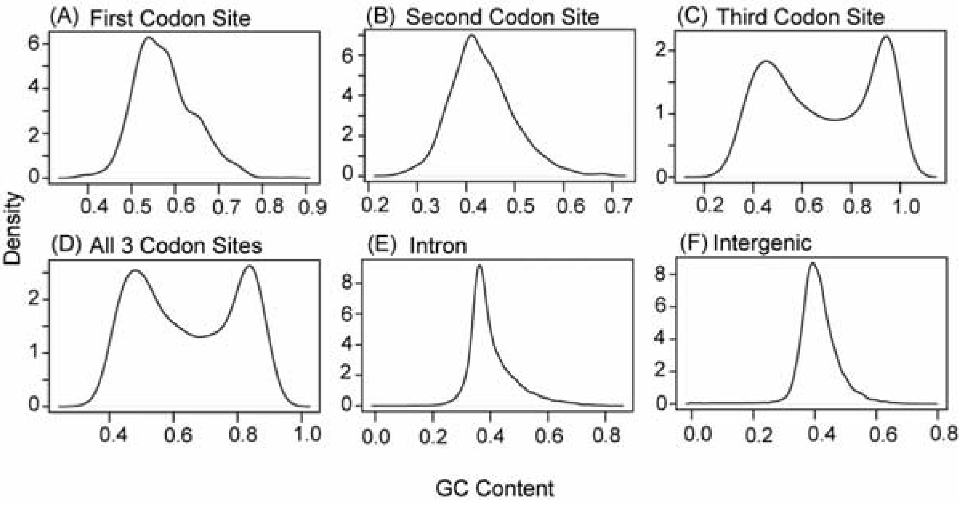

Figure 1. Guanine-cytosine (GC) content at three codon sites and in intron and intergenic regions.

Table 1. The spatial autocorrelation tests of guanine-cytosine (GC) content at the third codon sites (GC_3)

Block	Sister segment A		Sister segment B	
	I	*P*-value	I	*P*-value
1	0.16	**0.020**	0.01	0.460
2	0.04	0.188	0.04	0.227
3	0.05	0.132	0.10	**0.015**
4	0.05	0.313	0.07	0.192
5	0.12	**0.032**	0.17	**0.001**
6	0.01	0.434	0.15	**0.005**
7	0.10	0.052	0.15	**0.018**
8	0.06	0.114	0.14	**0.001**
9	−0.14	0.861	0.13	0.086
10	−0.01	0.542	0.03	0.340

Bold face indicates the significant values.

Table 2. Functional categories of the rice paralogs

	Binding activity	Catalytic activity	Transporter	Antioxidant	Structural activity	Signal transducer	Obsolete molecule	Enzyme regulator	Transcription regulator	Motor activity
$GC_3 > 0.75$	154	116	27	5	21	5	2	1	1	0
	(136.42)	(136.18)	(29.58)	(2.88)	(11.91)	(4.95)	(7.40)	(2.46)	(0.82)	(0.41)
$GC_3 < 0.75$	178	213	45	2	8	7	16	5	1	1
	(195.58)	(193.82)	(42.41)	(4.12)	(17.08)	(7.07)	(10.6)	(3.53)	(1.18)	(0.59)

Observed number and expected number (in brackets) of genes in each category. GC_3, third codon sites.

These observations implied the co-variation of the GC contents at three codon sites and in the intron region.

We checked several factors supposed to be related to GC_3, such as the physical location, the length of coding sequence, the exon number, the protein functions and the size of the gene family.

First, GC_3 was not obviously related to the physical location on the chromosome (Table 1). The spatial autocorrelation coefficients were calculated for the genes in 20 duplicated regions separately. A total of 13 segments showed no correlation between GC_3 and the physical location of the genes indicating that there was no clustering tendency of GC_3-rich genes.

Second, we found that GC_3 was negatively correlated to the length of the coding sequences (Spearman correlation coefficient $rho = -0.35$, $P < 2.2 \times 10^{-16}$). Also, our study revealed a much higher negative correlation between GC_3 and the exon number ($rho = -0.622\,3$, $P < 2.2 \times 10^{-16}$). That is, the shorter a gene was and the fewer exons it contained the higher GC_3 content it had. These observations were also previously reported in cereals (Carels and Bernardi 2000; Wang et al. 2004).

Third, we characterized the amino acid composition in two classes of genes and found that the composition was rather different ($\chi^2 = 6\,525.01$, $P < 2.2 \times 10^{-16}$), which was also observed by Wang et al. (2004).

Fourth, by checking the function of the paralogs in the gene ontology database (http://www.godatabase.org), we discovered that the two classes of genes differed significantly in function ($\chi^2 = 31.42$, $P = 2.8 \times 10^{-5}$). The GC_3-rich genes were more likely to participate in binding and structural molecular activity, whereas the GC_3-poor genes were more likely to engage in catalytic and obsolete molecular activity (Table 2).

Furthermore, we evaluated whether the genes in large families had rich GC composition, and found that GC_3 was independent of the size of gene families. A total of 3 704 gene families contained two or more genes, and the largest family included 253 genes. We checked whether the GC_3 content was correlated to the gene family size at the following three levels: the number of genes in the family, and the number of family members on the same chromosome, and the number of family members within the neighboring 100 genes. The Spearman correlation coefficients of these three tests were all smaller than 0.14, indicating little correlation between GC_3 and the gene family size.

Nucleotide substitution rate

Using the NG (Nei and Gojobori 1986) and YN (Yang and Nielsen 2000) methods, we computed the synonymous substitution rate (Ks). The correlations between Ks calculated by these two methods and GC_3 were inconsistent. To explore the true relationship between synonymous substitution rate and GC_3, we computed the average synonymous substitution number per third codon site (ASR) for the two classes of genes, respectively. Our results strongly supported the negative correlation between the synonymous substitution rate and GC_3, which was consistent with the results obtained by the NG method. We considered that the positive correlation obtained by the YN method was a methodological artifact. The detailed calculations and results regarding ASRs for two classes of genes had been published previously (Shi et al. 2006).

Discussion

Mutation bias, natural selection and BGC

It has been hypothesized that base composition variation could be a consequence of the following processes: mutation bias (Filipski et al. 1973; Sueoka 1988; Wolfe et al. 1989), natural selection (Bernardi 1986; Hughes and Yeager 1997; Eyre-Walker 1999), biased gene conversion (BGC) (Holmquist 1992; Eyre-Walker 1993), or a combination of them (Eyre-Walker and Hurst 2001). They act at one or several stages of the central dogma. Mutation bias could occur during DNA replication, repairing and transcription. It has been reported that during DNA replication, the concentrations of free nucleotides affect base composition (Meuth 1989). For example, when the G and C are rich in nucleic plasma, they are supposed to be mis-incorporated more likely into the replicates. The nucleotide concentrations are supposed to vary during the cell cycle and different parts of DNA might duplicate at different stages of the cell cycle (McCormick et al. 1983; Leeds et al. 1985). If this hypothesis is true, it is envisioned that a mosaic pattern of base composition would be discovered in chromosomes. Mutation bias could also occur for biased DNA repairing. The repairing of mismatched base pairs is hypothesized to be biased to certain nucleotides (e.g. G and C, and the bias could vary along chromosomes) (Brown and Jiricny 1988; Boulikas 1992). According to this hypothesis, a mosaic pattern of base composition in a chromosome should be detected. DNA repairing could happen during DNA replication or gene transcription (Eyre-Walker and Hurst 2001; Wong et al. 2002). The transcription-related mutation bias might result in a base composition change in the intragenic but not in the intergenic region.

Natural selection might also change base composition in a gene. Generally, selection is associated with the change of the amino acid composition. Bernardi and colleagues (D'Onofrio et al. 1991; Carels and Bernardi 2000) found that proteins produced by GC-rich and GC-poor genes were different in amino acid composition. They argued that GC-rich and GC-poor genes tended to be functionally different, the former acting as housekeeping genes and the latter as tissue-specific genes (Bickmore and Craig 1997; Chiapello et al. 1998). Moreover, natural selection is also supposed to be responsible for the base composition in synonymous codon sites (Sharp and Li 1987; Duret and Mouchiroud 1999). Akashi and Schaeffer (1997) proposed that selection favored synonymous substitutions to the preferred codons and restrained synonymous substitutions to the non-preferred codons, which would increase the translation accuracy. Although selection is often related to translation, it might work during the other stages of the central dogma.

Biased gene conversion was supposed to occur through the formation of heteroduplex DNA during the process of homologous recombination (Eyre-Walker and Hurst 2001). BGC is often grouped together with natural selection for it is considered to be equivalent to weak directional selection (Nagylaki 1983; Eyre-Walker and Hurst 2001). It is further suggested that BGC is more likely to occur between members in gene families with large size and result in the increase of GC content (Galtier 2003; Wang et al. 2004).

Possible causes of the formation of two gene classes

As for two classes of genes in rice and other cereals, Carels and Bernardi (2000) related them to biological function, implying the possibility of the effect of the natural selection. However, Wong et al. (2002) did not classify the rice genes they analyzed into two classes and argued that the increase in GC content was in a negative gradient along the transcriptional orientation. They further attributed the base composition change to transcription-related mutation bias and translation-related natural selection. In a recent study, Wang et al. (2004) attempted to reconcile former findings by exploring the base and amino acid compositions in GC_3-rich and GC_3-poor rice genes. They proposed that mutation bias at the nucleotide level should be primarily responsible to the elevation in GC content because they found that (i) the changes in GC content were the most significant at the third codon site, and (ii) there was a clear increase in substitution rate at the synonymous sites for the GC-rich genes (i.e. GC_3 was positively correlated to Ks according to the YN method). Therefore, three groups came to inconsistent conclusions on the cause of two classes of genes.

Here, we systematically checked the substitution bias and the factors related to GC content. Our results suggested that the two classes of rice genes were primarily caused by natural selection rather than mutation bias or BGC (Table 3). First, we rejected the possibility of mutation bias during DNA replication for the reason that GC_3 is unrelated to GC content in the intergenic region and there is no obvious clustering tendency of GC_3-rich or GC_3-poor genes.

Second, we found that there might be mutation bias at the transcription level implied by the observation that GC_3 is weakly correlated to GC content in introns. However, this transcription-related mutation bias accounts for only a small fraction of the variation in GC_3, for the correlation coefficient is rather small ($rho = 0.291$) (i.e. only 9% of the variation in GC_3 could be resulted from factors related to GC content in introns). This type of

Table 3. Observations indicating possible causes of guanine-cytosine (GC) content variation

Observations	DNA replication	DNA transcription	RNA process	Translation and post-translation	Conclusion
No evidence of correlation of GC_3 to physical location	×				Negate mutation bias during replication
GC_3 is only weakly correlated to GC in the intron (*rho* = 0.291)		√			Support mutation bias during transcription
GC_3 is independent of GC content in intergenic region (*rho* = 0.007)	×				Negate mutation bias during replication
GC_3 is negatively correlated to the exon number and coding sequence length (*rho* = −0.62, −0.35)			√		Support selection during RNA processing
GC_3-rich and GC_3-poor genes are functionally divergent				√	Support selection
GC_3-rich and GC_3-poor genes are different in amino acid composition				√	Support selection
GC_3-rich and GC_3-poor genes are independent of size of gene family					Negate BGC

BGC, biased gene conversion; GC_3, third codon sites.

mutation bias is supposed to occur during transcription-coupled DNA repair (Thoma 1999; Svejstrup 2002), as previously reported (Wong et al. 2002). Wang et al. (2004) proposed that mutation bias at the nucleotide level was the primary cause of GC content variation based on their observation that the differences in GC content were greatest at the third codon site. However, many researchers (Sharp and Li 1987; Duret and Mouchiroud 1999) proposed that selection was associated with the change of synonymous codon sites.

Third, natural selection during RNA processing might considerably affect GC_3. This is indicated by the significant negative correlation between GC_3 and the exon number (*rho* = −0.62) and the coding sequence length (*rho* = −0.35) from our analysis. Wang et al. (2004) also observed these possible selective constraints related to RNA splicing. The less exons a gene has, or the shorter its coding sequence is, the higher GC_3 it contains. This implies strong directional selection in base composition for genes with less exons.

Fourth, we found that GC_3 was negatively correlated to the synonymous substitution rate. We used both methods of YN and NG, and obtained incongruent results (i.e. a positive correlation by the YN method but a negative one by the NG method). This inconsistency has been debated for years and is supposed to arise methodologically (Bielawski et al. 2000; Dunn et al. 2001; Bierne and Eyre-Walker 2003). Therefore, we further evaluated the correlation with the help of simultaneously duplicated orthologs and their out-group sequences and found that GC_3-rich genes evolved with a lower synonymous substitution rate. The positive correlation supposed by the YN method might be caused by the hypotheses that the codon/nucleotide composition is constant. The negative correlation between GC_3 and the synonymous substitution rate, and a strong correlation between gene expression and codon usage bias are consistent with greater selective pressure on synonymous sites with high codon bias (Dunn et al. 2001). Wang et al. (2004) reached a conclusion that mutation bias was the primary cause of two rice gene classes by the finding that GC_3 was in positive correlation to the Ks calculated by the YN method. However, we found statistically significant differences in ASRs for two gene classes that strongly supported negative correlation between the synonymous substitution rate and GC_3. Therefore, the positive correlation found in Wang et al. (2004) might be an artifact of the method and cannot be the evidence of biased mutation.

Fifth, the effect of natural selection was also supported by the findings that two classes of genes were functionally divergent and had different amino acid compositions in our analysis as well as was revealed by Carels and Bernardi (2000) and Wang et al. (2004). Natural selection might act during translation, supported by the finding that the highest GC content appears near the translation initiation site (Wong et al. 2002).

Finally, the possibility of BGC was rejected because GC_3 is not related to the size of the gene family by our observations and those of Wang et al. (2004).

The pattern of GC content at three codon sites might be produced by the mixed action of two types of natural selection. Although GC_1 and GC_2 do not have obvious bi-modal distributions like GC_3, they are in considerable positive correlation to

GC_3 (*rho* = 0.576 or 0.473). This suggests that the coordinated change in GC content at all three codon sites implies a type of selection acting on all codon sites. Another type of selection affects the fixation of the mutation in amino acids, therefore, acting mainly on the first and second codon sites. The mixture of the two types of selection leads to the different patterns at three codon sites.

Materials and Methods

Rice paralogs

We extracted the rice genome sequences from RISE (rise.genomics.org.cn) and predicted rice genes using the BGF (Bejing Gene Finding) program (Li et al. 2005). The homologous genes were defined according to a criterion of BLASTP (Altschul et al. 1997) score > 100. With the gene homology information as input, duplicated chromosomal regions and paralogous genes in colinearity were detected by a program named ColinearScan we developed (Wang et al. 2006) A total of 1 738 paralogous genes in collinearity from ten duplicated blocks in the rice genome were analyzed. Nine of them were produced by a whole genome duplication 70 million years ago (MYA) before the divergence of the cereals but after the divergence of cereals and *Arabidopsis*, and one was produced by a segmental duplication that occurred only 5 MYA after the divergence of the cereals (Wang et al. 2005).

We extracted 1 738 rice paralogs and searched for their outgroup homologs in *Arabidopsis* and maize through GenBank and excluded those pairs without out-group or cDNA evidence. This eventually resulted in a dataset of 1 092 homologous gene triplets of rice paralogs and corresponding outgroups.

Synonymous substitution rate

We computed the synonymous substitution rate with the NG and YN methods implemented in Phylogenetic Analysis by Maximum Likelihood (PAML) package (Yang 1997). To tackle the problem of incongruent results obtained by these two methods, we developed a new approach to assess the average synonymous substitution number per third codon site (ASR) (i.e., the ratio of observed synonymous substitutions to matched triplets with four-fold degenerate last common ancestor (LCA) codons) (Shi et al. 2006).

Spatial correlation of GC_3

To examine whether GC_3 content is correlated to the chromosomal location, we carried out a spatial autocorrelation test involving all genes in the duplicated segments. For each of the 20 duplicated chromosomal segments, we calculated the spatial autocorrelation coefficient using the following statistic, a modification of Moran's *I* (Matassi et al. 1999):

$$I = \frac{\sum_{i=1}^{n-1} (x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{n-1} (x_i - \bar{x})^2} \sim N\left(0, \left(\frac{4}{n-1}\right)^2\right),$$

where n is the number of paralogous genes, x_i is the GC_3 content for the i-th gene in the chromosomal segment, and $\bar{x}$ is the mean of GC_3 contents.

Gene family size and GC_3

To assess the relationship between gene family size and GC_3, we clustered all genes in the rice genome into groups using BlastClust (Wolfsberg et al. 2001). These groups were taken as putative gene families. Based on this, we checked whether the gene family size was correlated to the GC_3.

Acknowledgements

We thank Ziheng Yang for his help in the calculation of synonymous substitution patterns and ASRs.

References

Akashi H, Schaeffer SW (1997). Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. *Genetics* **146**, 295–307.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Bernardi G (1986). Compositional constraints and genome evolution. *J. Mol. Evol.* **24**, 1–11.

Bickmore W, Craig J (1997). *Chromosome Bands: Patterns in the Genome*. Springer-Verlag, Heidelberg, New York.

Bielawski JP, Dunn KA, Yang Z (2000). Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. *Genetics* **156**, 1299–1308.

Bierne N, Eyre-Walker A (2003). The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: Implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics* **165**, 1587–1597.

Boulikas T (1992). Evolutionary consequences of nonrandom damage and repair of chromatin domains. *J. Mol. Evol.* **35**, 156–180.

Brown TC, Jiricny J (1988). Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54**, 705–711.

Carels N, Bernardi G (2000). Two classes of genes in plants. *Genetics* **154**, 1819–1825.

Chiapello H, Lisacek F, Caboche M, Henaut A (1998). Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene* **209**, GC1–GC38.

D'Onofrio G, Mouchiroud D, Aissani B, Gautier C, Bernardi G (1991). Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* **32**, 504–510.

Dunn KA, Bielawski JP, Yang Z (2001). Substitution rates in Drosophila nuclear genes: Implications for translational selection. *Genetics* **157**, 295–305.

Duret L, Mouchiroud D (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **96**, 4482–4487.

Eyre-Walker A (1993). Recombination and mammalian genome evolution. *Proc. R. Soc. Lond. B. Biol. Sci.* **252**, 237–243.

Eyre-Walker A (1997). Differentiating between selection and mutation bias. *Genetics* **147**, 1983–1987.

Eyre-Walker A (1999). Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics* **152**, 675–683.

Eyre-Walker A, Hurst LD (2001). The evolution of isochores. *Nat. Rev. Genet.* **2**, 549–555.

Filipski J, Thiery JP, Bernardi G (1973). An analysis of the bovine genome by Cs_2SO_4-Ag density gradient centrifugation. *J. Mol. Biol.* **80**, 177–197.

Galtier N (2003). Gene conversion drives GC content evolution in mammalian histones. *Trends Genet.* **19**, 65–68.

Holmquist GP (1992). Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.* **51**, 17–37.

Hughes AL, Yeager M (1997). Comparative evolutionary rates of introns and exons in murine rodents. *J. Mol. Evol.* **45**, 125–130.

Hurst LD, Williams EJ (2000). Covariation of GC content and the silent site substitution rate in rodents: Implications for methodology and for the evolution of isochores. *Gene* **261**, 107–114.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Leeds JM, Slabaugh MB, Mathews CK (1985). DNA precursor pools and ribonucleotide reductase activity: Distribution between the nucleus and cytoplasm of mammalian cells. *Mol. Cell Biol.* **5**, 3443–3450.

Matassi G, Sharp PM, Gautier C (1999). Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**, 786–791.

McCormick PJ, Danhauser LL, Rustum YM, Bertram JS (1983). Changes in ribo- and deoxyribonucleoside triphosphate pools within the cell cycle of a synchronized mouse fibroblast cell line. *Biochim. Biophys. Acta* **756**, 36–40.

Meuth M (1989). The molecular basis of mutations induced by deoxyribonucleoside triphosphate pool imbalances in mammalian cells. *Exp. Cell Res.* **181**, 305–316.

Moriyama EN, Powell JR (1997). Codon usage bias and tRNA abundance in Drosophila. *J. Mol. Evol.* **45**, 514–523.

Nagylaki T (1983). Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. USA* **80**, 6278–6281.

Nei M, Gojobori T (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426.

Powell JR, Moriyama EN (1997). Evolution of codon usage bias in Drosophila. *Proc. Natl. Acad. Sci. USA* **94**, 7784–7790.

Sharp PM, Li WH (1987). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* **4**, 222–230.

Shi X, Wang X, Li Z, Zhu Q, Tang W, Ge S et al. (2006). Nucleotide substitution pattern in rice paralogues: Implication for negative correlation between the synonymous substitution rate and codon usage bias. *Gene* **376**, 199–206.

Sueoka N (1988). Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**, 2653–2657.

Svejstrup JQ (2002). Mechanisms of transcription-coupled DNA repair. *Nat. Rev. Mol. Cell Biol.* **3**, 21–29.

Thoma F (1999). Light and dark in chromatin repair: Repair of UV-induced DNA lesions by photolyase and nucleotide excision repair. *EMBO J.* **18**, 6585–6598.

Wang XY, Shi XL, Hao BL (2002). The transfer RNA genes in *Oryza sativa* L. ssp. *indica*. *Sci. China* **45**, 504–511.

Wang HC, Singer GA, Hickey DA (2004). Mutational bias affects protein evolution in flowering plants. *Mol. Biol. Evol.* **21**, 90–96.

Wang X, Shi X, Hao B, Ge S, Luo J (2005). Duplication and DNA segmental loss in the rice genome: Implications for diploidization. *New Phytol.* **165**, 937–946.

Wang X, Shi X, Li Z, Zhu Q, Kong L, Tang W et al. (2006). Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinformatics* **7**, 447.

Wolfe KH, Sharp PM, Li WH (1989). Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283–285.

Wolfsberg T, McEntyre J, Schuler G (2001). Guide to the draft human genome. *Nature* **409**, 824–826.

Wong GK, Wang J, Tao L, Tan J, Zhang J, Passey DA et al. (2002). Compositional gradients in Gramineae genes. *Genome Res.* **12**, 851–856.

Yang Z (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556.

Yang Z, Nielsen R (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43.

(Handling editor: Yang Zhong)