Translational Selection on Codon Usage in *Xenopus laevis*

Héctor Musto,† Stéphane Cruveiller,* Giuseppe D'Onofrio,* Héctor Romero,*†‡ and Giorgio Bernardi**

*Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Naples, Italy; †Laboratorio de Organización y Evolución del Genoma, Sección Bioquímica, Facultad de Ciencias, Montevideo, Uruguay; and ‡Departamento de Genética, Facultad de Medicina, Montevideo, Uruguay

A correspondence analysis of codon usage in *Xenopus laevis* revealed that the first axis is strongly correlated with the base composition at third codon positions. The second axis discriminates between putatively highly expressed genes and the other coding sequences, with expression levels being confirmed by the analysis of Expressed sequence tag frequencies. The comparison of codon usage of the sequences displaying the extreme values on the second axis indicates that several codons are statistically more frequent among the highly expressed (mainly housekeeping) genes. Translational selection appears, therefore, to influence synonymous codon usage in *Xenopus*.

Introduction

With very few exceptions, the genetic code is the same in all living organisms. All amino acids except for methionine and tryptophan are encoded by more than one codon. Synonymous codons are, however, not randomly used (Grantham et al. 1980), and the factors governing synonymous codon preferences are not the same in different organisms.

Indeed, in the unicellular organisms *Escherichia coli* and *Saccharomyces cerevisiae,* synonymous codon choices appear to be positively correlated with the relative abundances of tRNAs, with the correlation being very strong for highly expressed genes (Ikemura 1981, 1982, 1985; Bennetzen and Hall 1982; Gouy and Gautier 1982; Sharp and Li 1986; Bulmer 1988, 1991; Kanaya et al. 1999; for reviews, see Sharp and Matassi 1994; Sharp et al. 1995; Akashi and Eyre-Walker 1998).

In multicellular organisms, different patterns can be found. For example, in *Caenorhabditis elegans* and *Drosophila melanogaster,* which are characterized by extensive variation in codon usage, the factors governing the choices have been attributed to an equilibrium between mutational biases and translational selection (Shields et al. 1988; Sharp and Li 1989; Moriyama and Gojobori 1992; Carulli et al. 1993; Akashi 1994, 1997; Stenico, Lloyd, and Sharp 1994; Moriyama and Powell 1997; Powell and Moriyama 1997). Translational selection at silent sites has also been reported to be the main factor shaping codon usage in *Zea mays* (Fennoy and Bailey-Serres 1993) and *Arabidopsis thaliana* (Chiapello et al. 1998).

Compositionally compartmentalized genomes, like those of vertebrates and, in particular, those of warm-blooded vertebrates, show multiple codon usages. The compositional properties of those genomes and, more precisely, the compositional correlations existing between coding sequences (and their different codon positions) and isochores (see Bernardi [2000] for a review) affect codon usage. The situation is strikingly different for genes located in GC-poor and GC-rich isochores (Bernardi and Bernardi 1985; Bernardi et al. 1985; D'Onofrio et al. 1991; Cruveiller, D'Onofrio, and Bernardi 2000). This point is best illustrated by the example of α- and β-globin genes, which show very different codon usages because they are located in isochores with very different levels of GC, in spite of both being very highly expressed and at nearly equimolar amounts in the same cells (Bernardi et al. 1985).

Expectedly, therefore, when applied to mammalian sequences, multivariate statistical analysis reveals a single major trend that is strongly correlated with the GC level at third codon positions (GC3) of each gene. Moreover, the first axis does not discriminate aspects of gene function such as regulation during development, tissue specificity, constitutive expression, intracellular localization of the protein product, etc. (Sharp et al. 1988, 1995; Sharp and Matassi 1994).

Along another line, no correlation was found between the rate of synonymous substitutions (K_s) and either the expression level or the tissue specificity of genes in a mouse/rat comparison (Wolfe and Sharp 1993). The conclusion that expression levels do not influence the codon usage pattern in mammals was also drawn by analyzing expressed sequence tags (ESTs) in different tissues (Duret and Mouchiroud 2000).

Since GC3-rich genes represent roughly half of human genes (see Bernardi [2000] for a review), and since the multivariate analysis was carried out on human genes regardless of their GC3 levels (Sharp et al. 1988), one might think, however, that even if a translational selection effect exists in mammals, it could be swamped out by the much stronger compositional constraints. We decided, therefore, to apply multivariate analysis to the coding sequences of *Xenopus laevis,* which are characterized by a much narrower GC3 distribution, with very scarce high GC3 values.

Materials and Methods

Complete coding sequences (CDS) from *X. laevis* were retrieved from GenBank (release 114.0, December 1999) using ACNUC (Gouy et al. 1985). Redundancies were removed by CLEANUP (Grillo et al. 1996). The

Key words: codon usage, isochores, translational selection, vertebrates, correspondence analysis.

Address for correspondence and reprints: Giorgio Bernardi, Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121, Naples, Italy. E-mail: bernardi@alpha.szn.it.

Mol. Biol. Evol. 18(9):1703–1707. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

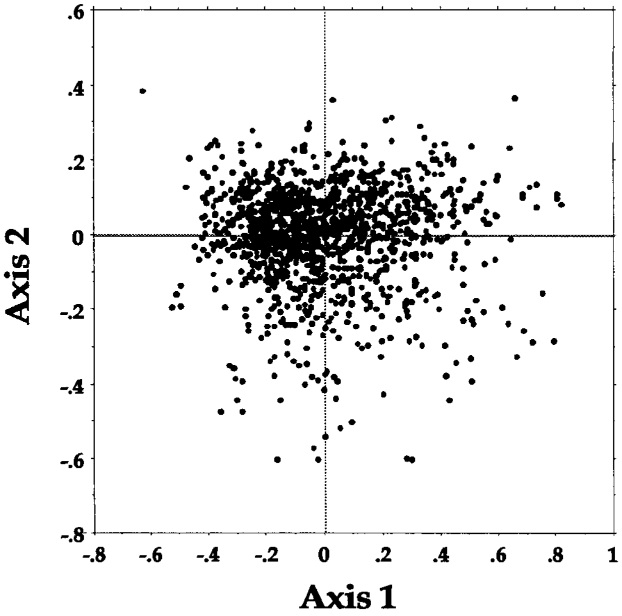

FIG. 1.—Distribution of *Xenopus* genes on the plane defined by the two main axes of the correspondence analysis.

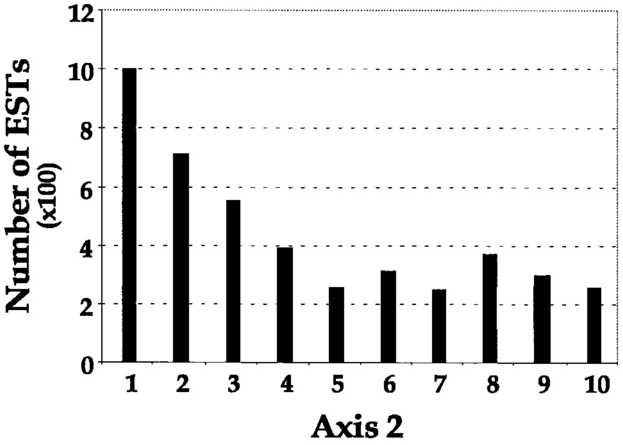

FIG. 2.—Histogram of the distribution of expressed sequence tags (ESTs) along axis 2. The axis was divided into 10 parts, each of them containing an equal number of genes. In each part, the total number of ESTs was calculated.

final data set included 1,303 genes. Codon usage, correspondence analysis COA, GC3 (the frequency of codons ending in C or G, excluding Met, Trp, and stop codons), the ''effective number of codons'' (N_c) (Wright 1990), the relative synonymous codon usage (RSCU) (Sharp and Li 1986), and the codon adaptation index (CAI) (Sharp and Li 1987) were calculated using the program CodonW 1.3 (J. Peden; http://molbiol.ox.ac.uk/Win95.codonW.zip.) Expression levels were estimated by retrieving from the TIGR database the ESTs (http://www.tigr.org/tdb/xgi/) that matched our CDS sample.

Results and Discussion

The mean GC3 content of the nuclear genes of *Xenopus* is 48.8%, with a standard deviation of 11%. The distribution of GC3 is much narrower than that for human genes and does not show the abundant GC3-rich values of human sequences (see Bernardi 2000). Indeed, genes with GC3 values higher than 60% represent only 19% of all genes investigated in the *Xenopus* genome, whereas they represent 55% in the data set reference from the human genome. The N_c values (N_c is a measure of the bias in codon usage of the genes, and usually highly expressed sequences display lower values compared with lowly expressed sequences) show, however, a relatively broad range, from 30.8 to 61.0 (not shown). These features suggest that there is some variation in codon usage among the sequences. In order to understand the causes of this variation, we conducted a COA of the RSCU values for all of the genes available from *Xenopus*. The position of each sequence on the plane defined by the first two axes is displayed in figure 1.

The proportions of the total variance accounted for by these principal axes were 20.3% and 6.5%, respectively. The analysis therefore detects a single major source of variation, which is strongly correlated with the GC3 level of each gene (R = 0.98). This axis does not discriminate any other biological feature of the genes, such as gene size, housekeeping or tissue-specific pattern of expression, or intracellular localization of the protein product. Moreover, no correlation was found (R = 0.047; NS) with gene expression levels, that is, with the EST frequencies, even excluding from the analysis those genes with undetected expression levels. Similar results were previously reported for a human data set by Sharp et al. (1988).

When genes were sorted according to their positions along the second axis, a significant correlation was found with the pyrimidine (Y) content of the genes at the third codon positions ($R = -0.37$; $P < 0.0001$). A striking result was, however, that constitutively highly expressed housekeeping genes, such as ribosomal proteins, histones, elongation factors, tubulins, and several enzymes from the intermediary metabolism, were clustered in the top 10% of the distribution. For example, there are 24 sequenced genes coding for ribosomal proteins, and 63% of them are placed in the first third of the distribution along axis 2. Furthermore, highly expressed, tissue-specific coding sequences, such as several actins, α- and β-globin, troponin, cytokeratin, etc., were also located in the same group. Regulatory sequences such as zinc finger proteins, oncogenes, homeobox genes, growth factors, etc. were located at the other 10% end of the distribution, which did not comprise any highly expressed housekeeping sequences. Therefore, it seems clear that axis 2 of COA is related to the expression level of each gene.

In order to confirm this interpretation and to obtain an approximate quantitative estimation of the expression levels of the 1,303 genes studied in this paper, we counted the number of matching ESTs for each sequence and their distribution along axis 2. The result of this analysis is shown in figure 2. In spite of the biased nature of the libraries and the fact that 45% of the genes did not match any EST, the general pattern clearly confirms that there is a gradient of expression from the left (where the majority of ribosomal proteins and other highly expressed housekeeping genes are placed) to the right of the distribution along axis 2. Furthermore, it should be stressed that a significant correlation holds between the position of each sequence on the second axis and the number of corresponding ESTs ($R = -0.23$; $P < 0.0001$). This quantitative analysis demonstrates that the sequences with most negative values along axis 2 are more highly expressed. In other words, axis 2 does indeed discriminate expression levels.

Our next step was to calculate the CAI value for each gene taking as a reference only the genes coding for ribosomal proteins, which are very highly expressed. In this case, we found that the CAI values were strongly correlated with the positions of the genes along the second axis after removal of the ribosomal proteins ($R = -0.62$; $P < 0.0001$). A very similar result ($R = -0.53$; $P < 0.0001$) was obtained when the reference set comprised the genes that matched more than 20 ESTs. Finally, the CAI values are significantly correlated with the corresponding numbers of matching ESTs ($R = 0.21$; $P < 0.0001$). These results confirm that axis 2 is strongly correlated with the expression level of each sequence in *Xenopus*. Interestingly, the mean N_c value of the genes located at the 5% end of the second axis, where putatively highly expressed sequences were placed, was 46.5, while at the other end it was 52.2, as expected (see above).

Similar results with regard to codon usage, using multivariate analyses, have been widely reported for unicellular species. They have usually been interpreted in terms of natural selection acting at the level of translation (for reviews, see Sharp and Matassi 1994; Sharp et al. 1995; Akashi and Eyre-Walker 1998). Remarkably, similar results were found not only among microorganisms, but also in multicellular species, such as *C. elegans* (Stenico, Lloyd, and Sharp 1994).

It should be stressed, however, that there are two main differences between the nematode results and the *Xenopus* results. These differences concern the source of variation which discriminates expression levels, and the amount of variation which is accounted for by the axis correlated with expression levels. Indeed, while in the nematode it is the first axis which is correlated with expression (and which, by definition, accounts for the majority of the variance), in *Xenopus* the axis related to that feature is the second, which accounts for a lower proportion of the total variability in codon usage. Accordingly, the differences appear to be more quantitative than qualitative, and hence we conclude that in *Xenopus*, translational selection indeed influences synonymous codon usage, even if it does so to a lesser extent than in *C. elegans.*

Our final step was to identify the translationally preferred codons in *Xenopus*. The codon usage patterns of the sequences displaying the extreme values at both ends of the second axis (100 genes each) were compared, and the differences were tested with a χ^2 test. The result of this analysis (table 1) shows that there are 22 putative preferred codons corresponding to 17 amino acids (the only amino acid with no preferred codon is Tyr), and 50% of the codons are T-ending. Among stop

Table 1
Putative Preferred Codons in *Xenopus laevis*

		Highly Expressed		Lowly Expressed	
Amino Acid	Codon	RSCU	*N*	RSCU	*N*
Phe	TTT	0.89	449	1.09	665
	TTC	1.11	556	0.91	554
Tyr........	TAT	0.96	393	0.94	487
	TAC	1.04	422	1.06	553
His........	**CAT**	0.96	299	0.80	419
	CAC	1.04	322	1.20	626
Asn	AAT	0.87	467	0.95	651
	AAC	1.13	608	1.05	714
Asp	**GAT**	1.16	946	1.00	810
	GAC	0.84	682	1.00	814
Cys	TGT	0.82	169	0.98	433
	TGC	1.18	245	1.02	455
Gln	CAA	0.60	349	0.73	553
	CAG	1.40	806	1.27	962
Lys	AAA	0.78	894	1.09	1,078
	AAG	1.22	1,393	0.91	907
Glu	GAA	0.98	1,001	1.07	1,155
	GAG	1.02	1,034	0.93	1,001
Val........	**GTT**	1.17	580	0.79	339
	GTC	0.95	469	0.90	389
	GTA	0.58	289	0.74	318
	GTG	1.30	641	1.57	678
Pro........	**CCT**	1.58	669	1.02	455
	CCC	1.00	424	1.07	478
	CCA	1.29	549	1.35	603
	CCG	0.13	56	0.55	245
Thr........	**ACT**	1.38	570	0.96	422
	ACC	1.39	575	1.09	477
	ACA	1.13	469	1.29	565
	ACG	0.11	44	0.67	294
Ala........	**GCT**	1.76	1,127	0.93	420
	GCC	1.23	784	1.25	567
	GCA	0.85	545	1.26	570
	GCG	0.16	101	0.57	259
Gly	**GGT**	1.57	1,087	0.47	239
	GGC	0.95	659	1.04	525
	GGA	1.11	769	1.19	601
	GGG	0.37	258	1.29	648
Leu	TTA	0.28	110	0.85	380
	TTG	0.93	364	0.88	395
	CTT	1.23	478	0.89	399
	CTC	0.90	350	1.00	447
	CTA	0.42	163	0.55	247
	CTG	2.24	872	1.83	821
Ser........	**TCT**	1.67	532	1.02	460
	TCC	1.46	465	1.06	480
	TCA	0.71	225	1.01	455
	TCG	0.13	41	0.52	234
	AGT	0.87	277	0.90	407
	AGC	1.17	373	1.49	671
Arg	**CGT**	2.18	589	0.17	47
	CGC	1.15	310	0.49	140
	CGA	0.37	100	0.70	198
	CGG	0.26	70	0.96	273
	AGA	1.15	310	2.00	567
	AGG	0.89	241	1.69	479
Ile	**ATT**	1.34	708	1.00	478
	ATC	1.39	736	1.13	538
	ATA	0.27	143	0.87	412
Met	ATG	1.00	763	1.00	814
Trp........	TGG	1.00	255	1.00	368
TER.......	**TAA**	2.25	75	0.99	33
	TAG	0.36	12	0.60	20
	TGA	0.39	13	1.41	47

NOTE.—RSCU is the relative synonymous codon usage (Sharp and Li 1986), and *N* is the count of each codon in each group of genes. Codons in bold or underlined are statistically more frequent in the highly expressed genes ($P < 0.01$ or $P < 0.05$, respectively).

codons, TAA is by far the most frequently used in highly expressed sequences, while an opposite trend was found for TAG. Remarkably, 82% of the preferred triplets are Y-ending, a point which explains the negative correlation previously described between the positions of sequences along the second axis and the corresponding Y levels in the third codon positions.

Some general rules emerge from the analysis of the preferred codons: (1) for all quartets (including those belonging to sextets), the T-ending codons are always preferred in highly expressed genes; among these codons, if there is a second favored triplet, it is C-ending; (2) the G-ending codons are the chosen ones for the purine-ending duets; (3) the duets of sextets are never significantly incremented among the highly expressed sequences; (4) the A-ending codons are never preferred; and (5) the NCG codons are always the least frequent, and their usage is further decreased among highly expressed genes.

Conclusions

Codon usage in *Xenopus*, although determined mainly by compositional constraints, is influenced by translational selection. Indeed, a COA on RSCU values detects a trend, independent of GC3, that discriminates putatively highly expressed (mainly housekeeping, but also tissue-specific) genes from the other sequences. This is confirmed by the significant correlation found between the position of each gene along the second axis and the number of matching ESTs and CAI values.

Acknowledgments

We thank the referees for their very useful suggestions. S.C. thanks the European Community for contract ERBFMRXCT980221 in the frame of the Training and Mobility of Researchers Programme.

LITERATURE CITED

Akashi, H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics **136**:927–935.

———. 1997. Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. Gene **205**:269–278.

Akashi, H., and A. Eyre-Walker. 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. **8**: 688–693.

Bennetzen, J. L., and B. D. Hall. 1982. Codon selection in yeast. J. Biol. Chem. **257**:3026–3031.

Bernardi, G. 1995. The human genome: organization and evolutionary history. Annu. Rev. Genet. **29**:445–476.

———. 2000. Isochores and the evolutionary genomics of vertebrates. Gene **241**:3–17.

Bernardi, G., and G. Bernardi. 1985. Codon usage and genome composition. J. Mol. Evol. **22**:363–365

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier. 1985. The mosaic genome of warm-blooded vertebrates. Science **228**:953–958.

Bulmer, M. 1988. Codon usage and intragenic position. J. Theor. Biol. **133**:67–71.

———. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics **129**:897–907.

Carulli, J. P., D. E. Krane, D. L. Hartl, and H. Ochman. 1993. Compositional heterogeneity and patterns of molecular evolution in the Drosophila genome. Genetics **134**: 837–845.

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

Cruveiller, S., G. D'Onofrio, and G. Bernardi. 2000. The compositional transition between the genomes of cold- and warm-blooded vertebrates: codon frequencies in orthologous genes. Gene **261**:71–83.

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

Duret, L., and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**:68–74.

Fennoy, S. L., and J. Bailey-Serres. 1993. Synonymous codon usage in Zea mays L. nuclear genes is varied by levels of C and G-ending codons. Nucleic Acids Res. **21**:5294–5300.

Gouy, M., and C. Gautier. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10**: 7055–7074.

Gouy, M., C. Gautier, M. Attimonelli, C. Lanave, and G. Di Paola. 1985. ACNUC–a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput. Appl. Biosci. **1**:167–172.

Grantham R, C. Gautier, M. Gouy, R. Mercier, and A. Pave. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. **8**:r49–r62.

Grillo, G., M. Attimonelli, S. Liuni, and G. Pesole. 1996. CLEANUP: a fast computer programme for removing redundancies from nucleotide sequence databank. CABIOS **12**:1–8.

Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151**:389–409.

———. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J. Mol. Biol. **158**:573–597.

———. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.

Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238**:143–155.

Moriyama, E. N., and T. Gojobori. 1992. Rates of synonymous substitution and base composition of nuclear genes in Drosophila. Genetics **130**:855–864.

Moriyama, E. N., and J. R. Powell. 1997. Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. J. Mol. Evol. **45**:378–391.

Powell, J. R., and E. N. Moriyama. 1997. Evolution of codon usage bias in Drosophila. Proc. Natl. Acad. Sci. USA **94**: 7784–7790.

Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi, and J. F. Peden, 1995. DNA sequence evolution: the sounds of silence. Philos. Trans. R. Soc. Lond. B Biol. Sci. **349**:241–247.

Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe, and F. Wright. 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res. **16**:8207–8211.

Sharp, P. M., and W. H. Li. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24**:28–38.

———. 1987. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**:1281–1295.

———. 1989. On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28**:398–402.

Sharp, P. M., and G. Matassi. 1994. Codon usage and genome evolution. Curr. Opin. Genet. Dev. **4**:851–860.

Shields, D. C., P. M. Sharp, D. G. Higgins, and F. Wright. 1988. 'Silent' sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5**:704–716.

Stenico, M., A. T. Lloyd, and P. M. Sharp. 1994. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. Nucleic Acids Res. **22**:2437–2446.

Wolfe, K. H., and P. M. Sharp. 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. **37**:441–456.

Wright, F. 1990. The 'effective number of codons' used in a gene. Gene **87**:23–29.

Adam Eyre-Walker, reviewing editor

Accepted May 11, 2001