ELSEVIER

Gene 261 (2000) 93–105

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes

Kaori Iida, Hiroshi Akashi*

Institute of Molecular Evolutionary Genetics, Department of Biology, 208 Mueller Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

Received 15 July 2000; received in revised form 15 September 2000; accepted 5 October 2000
Received by G. Bernardi

Abstract

Natural selection appears to discriminate among synonymous codons to enhance translational efficiency in a wide range of prokaryotes and eukaryotes. Codon bias is strongly related to gene expression levels in these species. In addition, between-gene variation in silent DNA divergence is inversely correlated with codon bias. However, in mammals, between-gene comparisons are complicated by distinctive nucleotide-content bias (isochores) throughout the genome. In this study, we attempted to identify translational selection by analyzing the DNA sequences of alternatively spliced genes in humans and in *Drosophila melanogaster*. Among codons in an alternatively spliced gene, those in constitutively expressed exons are translated more often than those in alternatively spliced exons. Thus, translational selection should act more strongly to bias codon usage and reduce silent divergence in constitutive than in alternative exons. By controlling for regional forces affecting base-composition evolution, this within-gene comparison makes it possible to detect codon selection at synonymous sites in mammals. We found that GC-ending codons are more abundant in constitutive than alternatively spliced exons in both *Drosophila* and humans. Contrary to our expectation, however, silent DNA divergence between mammalian species is *higher* in constitutive than in alternative exons. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Codon bias; Weak selection; Molecular evolution

1. Introduction

Natural selection appears to bias codon usage to enhance protein synthesis in *Escherichia coli*, *Saccharomyces cerevisiae* (reviewed in Andersson and Kurland, 1990; Sharp et al., 1993), *Drosophila melanogaster* (reviewed in Shields et al., 1988), and *Caenorhabditis elegans*, and *Arabidopsis thaliana* (Stenico et al., 1994; Chiapello et al., 1998; Duret and Mouchiroud, 1999). These species show positive correlations between synonymous codon bias and gene expression levels (Gouy and Gautier, 1982; Ikemura, 1985; Stenico et al., 1994; Duret and Mouchiroud, 1999). Furthermore, preferentially used codons tend to be recognized by abundant tRNAs in *E. coli* (Ikemura, 1981), *bacillus subtilis* (Kanaya et al., 1999) yeast (Ikemura, 1982), *D. melanogaster* (Moriyama and Powell, 1997), and *C. elegans* (Duret, 2000).

These patterns suggest both a role of natural selection at synonymous sites and a functional basis for fitness differences among synonymous codons. In *E. coli*, major tRNA-encoding codons are translated three- to six-fold faster than their synonymous counterparts (Sorensen et al., 1989). About 90% of energy production is used in the process of protein synthesis in *E. coli* (Tamarin, 1999), and major codons may save cellular energy and enhance translational efficiency (Ikemura, 1985). Favored codons may also enhance the accuracy of translation (Bulmer, 1988a; Akashi, 1994; Eyre-Walker, 1996). In *E. coli*, major codons can reduce the frequency of misincorporations approximately ten-fold over minor codons for the same amino acid (Precup and Parker, 1987). In addition, major codons may lower the energetic cost of proofreading (Bulmer, 1988a).

Patterns of codon usage and synonymous DNA evolution in *D. melanogaster*, *C. elegans*, and *A. thaliana* appear to be similar to those in *E. coli* and yeast. Among *D. melanogaster* genes, variation in GC content at synonymous sites does not correlate strongly with the base composition of introns (Kliman and Hey, 1994). In addition, the limited data on tRNA levels show a positive relationship between favored

Abbreviations: MH test, Mantel–Haenszel test; O/E, observed/expected; yn00, computer application from Yang and Nielson (2000)

* Corresponding author. Tel.: +1-814-865-5013; fax: +1-814-865-9131. *E-mail address:* akashi@psu.edu (H. Akashi).

0378-1119/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0378-1119(00)00482-0

codon usage and tRNA abundance (Moriyama and Powell, 1997). *Caenorhabditis elegans*, another invertebrate species, shows a positive correlation between codon usage and the number of tRNA genes (Duret, 2000). Finally, although relative expression levels can be specific to tissue and to developmental stage and thus difficult to quantify in multi-cellular organisms, evidence for higher codon usage bias in highly expressed genes appears consistent with selection pressure at the translation level in *D. melanogaster*, *C. elegans*, and *A. thaliana* (Shields et al., 1988; Duret and Mouchiroud, 1999).

Translational selection at synonymous sites in mammals remains equivocal. Fitness differences among synonymous codons are thought to be very small (Kimura, 1983; Li, 1987; Bulmer, 1988b; Hartl et al., 1994; Akashi, 1995; Akashi and Schaeffer, 1997), and large population sizes are required for such weak selection to overcome genetic drift (Fisher, 1930; Kimura, 1983; Ohta, 1992). Because *E. coli* and yeast presumably have large effective population sizes, small fitness differences among synonymous alternatives can result in high codon usage bias. *Drosophila melanogaster* probably has an effective population size intermediate between those of mammals and those of *E. coli* and yeast (Shields et al., 1988; Duret and Mouchiroud, 1999) and also shows codon selection. Synonymous sites in mammals, however, have been thought to evolve neutrally because of presumably small effective population sizes (Shields et al., 1988).

Testing for translational selection is complicated by base-composition heterogeneity within mammalian genomes. Mammalian chromosomes appear to be mosaics of long DNA segments called 'isochores' that have distinctive GC content and are usually over 300 kb (Bernardi et al. 1985; reviewed in Bernardi, 2000). In the human genome, GC content ranges from 30 to 60%, and five families of isochores have been identified: two GC-poor families (L1 and L2) representing 62% of the genome, and three GC-rich families (H1, H2, and H3) representing 22, 9, and 3%, respectively (Bernardi, 1993). The base composition of third positions within coding regions is strongly correlated with the base composition of introns and non-coding regions for a given gene (reviewed in Bernardi, 2000). Thus, a relationship between GC content and gene expression levels could result from a correlation between regional base composition and gene expression levels. To detect selection at synonymous sites, it is important to eliminate isochore effects.

Some evidence of selection on synonymous sites in mammals has been suggested (Cacciò et al., 1995; Mouchiroud et al., 1995; Zoubak et al., 1995; Alvarez-Valin et al., 1998; Eyre-Walker, 1999), but little evidence supports translational selection. Base-composition bias in mammals may be due to mutational bias or selection for regional base composition; translational selection for codon bias may be masked by such isochore effects. To identify translational selection, we examined alternatively spliced genes of humans and *D. melanogaster*. Alternatively spliced protein-coding genes provide an opportunity to examine exons that differ in translation levels but lie within an isochore. Codons within exons found in all isoforms expressed from a gene will be translated at higher levels than codons within exons found in a subset of transcripts. Since few genes cross isochore boundaries, alternatively spliced genes provide an opportunity to identify the effect of natural selection on synonymous sites; more highly expressed exons should have higher codon usage bias. Differences in base composition at silent sites between constitutive and alternatively spliced exons cannot be explained by differences in transcription rates or by region-specific forces affecting base composition.

2. Materials and methods

2.1. Codon families and major codons

'Codon families' or 'synonymous families' refer to groups of two to six codons that encode the same amino acid. Leucine and arginine codons are pooled into six-fold families, but serine codons are divided into a two- and a four-fold family so that, in a given synonymous family, all codons can mutate to all other codons through single-base synonymous changes. Candidates for 'major codons' have been identified as those that increase in frequency as a function of the level of codon bias of *D. melanogaster* genes (Sharp and Lloyd, 1993; Akashi, 1995). For humans, GC-ending codons in each codon family were classified as putative major codons. Although tRNA abundances have not been quantified in human tissues, there is experimental evidence that the expression level of human genes can be increased dramatically in mammalian cell lines by altering codon usage toward GC-ending synonymous alternatives (Kim et al., 1997; André et al., 1998).

2.2. Base-composition comparisons in alternatively spliced genes

We extracted over 500 human and 180 *D. melanogaster* DNA sequences of alternatively spliced genes from the GenBank database (Release 117.0). For each gene, we divided exons into two categories: exons common to all known isoforms are 'constitutive' exons and those found in a subset of isoforms are 'alternative' exons. Published data were used to identify the 5′ and 3′ ends of constitutive and alternative exons. Alternative exons translated in different reading frames among known isoforms were eliminated from the data set. Only genes for which the total length of both constitutive and alternative exons was at least 50 codons were analyzed. With these selection criteria, 77 genes for human and 33 genes for *D. melanogaster* were included in the analysis (Tables 1 and 2).

To determine whether translational selection affects synonymous codon usage, we constructed 2 × 2 contingency tables comparing the frequencies (within codon families) of

Table 1
Major codon usage in alternatively spliced *D. melanogaster* genes[a]

Gene symbol	Full name	Codons		*Z*	GenBank
		con (freq. major)	alt (freq. major)		
Abd-B	Abdominal B	235 (0.67)	221 (0.60)	2.12	U31961
alpha-Man-l	alpha Mannosidase l	441 (0.58)	380 (0.53)	1.14	X82640 X82641
arm	Armadillo	662 (0.58)	114 (0.50)	0.70	AF001213 X54468
br	broad	421 (0.74)	642 (0.67)	2.10	X54663 X54664 X54665 X54666
Btk29A	Btk family kinase at 29A	529 (0.56)	263 (0.55)	0.71	AB009841 AB009840
Cdic	Cytoplasmic dynein intermediate chain	139 (0.71)	455 (0.75)	− 0.98	AF070689 AF070690 AF070691 AF070692 AF070693 AF070694 AF070695 AF070696 AF070697 AF070698 AF070699
Cf2	Chorion factor 2	380 (0.64)	117 (0.53)	1.67	M97196
cora	coracle	312 (0.61)	1063 (0.58)	0.68	L27467 L27468 L27469
Ddc	Dopa decarboxylase	411 (0.64)	54 (0.33)	3.78	X04426
dec-1	defective chorion 1	854 (0.47)	767 (0.44)	0.77	M35889 M35888 M35887
dnc	dunce	590 (0.54)	143 (0.54)	0.08	X55167 M14982
Egfr	Epidermal growth factor receptor	1267 (0.59)	141 (0.65)	− 2.12	AF052754
Eip74EF	Ecdysone-induced protein 74EF	525 (0.59)	624 (0.62)	− 0.64	M37083 M37082
Fas2	Fasciclin2	711 (0.57)	205 (0.56)	0.42	M77166 M77165
Furl	Furin 1	816 (0.57)	642 (0.56)	0.79	L12372 L12375 L12376
lola	longitudinals lacking	442 (0.57)	441 (0.61)	− 1.52	U07607 U07606
M1c-k	Myosin light chain kinase	557 (0.62)	389 (0.56)	1.71	D089661 D89662 D89663
ninaC	neither inactivation nor afterpotential C	1033 (0.54)	453 (0.57)	− 1.49	M20231 M20230
Nrg	Neuroglian	1159 (0.57)	94 (0.37)	4.09	AF050085
para	paralytic	1782 (0.50)	97 (0.39)	1.97	M32078
ple	pale	458 (0.76)	70 (0.64)	2.19	U14395
pnt	pointed	384 (0.55)	535 (0.61)	− 0.96	X69167 X69166
Prm	Paramyosin	360 (0.88)	619 (0.74)	5.21	X62591 X58722
RecQ5	RecQ5	453 (0.42)	577 (0.32)	3.21	AF134239
sgg	shaggy	476 (0.50)	664 (0.57)	− 2.15	X70862 X70863 X70864

(continued overleaf)

Table 1 (*continued*)

Gene symbol	Full name	Codons		Z	GenBank
		con (freq. major)	alt (freq. major)		
Sh	Shaker	284 (0.27)	553 (0.54)	− 5.88	X07134 X07133 X07132 X07131
Su(var)3-9	Suppressor of variegation 3-9	79 (0.66)	869 (0.58)	0.90	X80070 X80069
svp	seven up	434 (0.69)	375 (0.60)	2.94	M28863 M28864
TBPH	TBPH	289 (0.72)	188 (0.72)	0.03	AB019705 AB019706
tkv	thickveins	450 (0.61)	69 (0.38)	3.62	L33784 L33785
ttk	tramtrack	277 (0.63)	861 (0.55)	2.05	Z11723 X71 627
tws	twins	392 (0.61)	54 (0.52)	− 0.14	L12544
zip	zipper	1813 (0.56)	81 (0.42)	3.33	U35816

[a] Gene symbols and full names are from FlyBase (1999). The numbers of codons examined in constitutive and alternative exons are given. The numbers in parentheses indicate the overall frequencies of major codons (pooled across synonymous families) in constitutive or alternative exons. The Z values for the Mantel-Haenszel test (see text) for data pooled across synonymous families are shown for each gene. The GenBank accession number(s) (Release 117.0) is given for each gene. Alternative splicing has been confirmed by comparing cDNA and genomic DNA sequence data (either through restriction map studies or by direct sequencing of genomic DNA) for all genes except the following: *cora*, *Fas2*, *lola*, *Prm*, *svp*, *ttk*.

putative major codons in constitutive and alternative exons. Data from constitutive and alternative exons are represented in the columns of the tables and the counts of major codon(s) and non-major codons from a synonymous family are represented in the rows of the tables (Table 3). Similar tables are constructed for each codon family within each gene. If synonymous sites are under translational selection, we predict an overall deviation in these contingency tables; the frequency of major codons should be higher in constitutive than in alternative exons. Note that codon usage is compared only within synonymous families; differences in amino acid composition between constitutive and alternative exons will not affect the analysis. We employed the Mantel–Haenszel (MH) procedure (Mantel and Haenszel, 1959; Mantel, 1963) to test for overall deviations across independent tables for a large number of genes and amino acids. This statistical test takes into account both the magnitude and direction of deviations within contingency tables and should be sensitive to consistent differences in codon usage between constitutive and alternative exons.

2.3. Silent divergence in alternatively spliced genes

We searched the GenBank database for mammalian orthologs of the 77 alternatively spliced human genes. We divided the exons into constitutive and alternative exons for each gene and aligned them with exons from orthologous human genes using CLUSTALX (Jeanmougin et al., 1998). Each alignment was checked manually to eliminate ambiguously aligned positions. We included for analysis only exons that are either constitutively expressed or alternatively spliced in both species. The same method and criteria as those described above for the within-species analysis were applied. However, because of the reduced sample size, we set the minimum codon number required for the analysis to 25 rather than 50. Twenty-six genes were analyzed (Table 4).

To calculate synonymous (d_S) and non-synonymous (d_N) divergence for constitutive and alternative exons for each gene, we employed the 'yn00' application (Yang and Nielsen, 2000) in the PAML package (Yang, 1997). This method takes into account both transition/transversion rate bias and base/codon frequency bias in calculating DNA divergence (Yang and Nielsen, 2000). Wilcoxon's signed-ranks test was performed to test for consistent differences in DNA divergence between constitutive and alternative exons (Tables 4 and 5).

2.4. CpG islands

CpG islands are regions of unmethylated CpG-rich sequences often located in gene promoter regions in mammals (Bird, 1987; Jones, 1999). Methylation of CpG islands may suppress the initiation of transcription (Jones, 1999). Methylated CpGs are thought to mutate to TpGs and CpAs at a high rate (Coulondre et al., 1978; Lindahl, 1982). Such regions often extend from 5′ flanking regions into the beginning, and sometimes the middle, of coding regions (Cross and Bird, 1995; Jones, 1999). Therefore, CpG islands could influence both base composition and silent divergence in coding regions. In order to eliminate this bias, we attempted to exclude such regions from human genes using two different methods. CpG islands have been identi-

Table 2
Major codon usage in alternatively spliced human genes[a]

Gene symbol	Name	Codons		*Z*	GenBank
		con (freq. major)	alt (freq. major)		
A	Protein A	108 (0.72)	458 (0.73)	− 0.09	U47924
ABCC3	ATP-binding cassette, sub-family C (CFTR/MRP), member 3	213 (0.79)	1060 (0.72)	2.21	AF085690 AF085691 AF085692
ABL1	v-abl Abelson murine leukemia viral oncogene homolog 1	1008 (0.71)	68 (0.60)	2.20	U07563 U07561
ACVR1B	Activin A receptor, type 1b (SKR2)	401 (0.62)	154 (0.55)	1.66	L31848 L10125 L10126
ADD2	Adducin 2 (beta)	212 (0.79)	397 (0.72)	1.78	S81079 S81083
ADRA1C	Adrenergic receptor alpha-1C	394 (0.78)	123 (0.58)	4.33	D32201 D32202 U03866
AF-6	Myeloid/lymphoid or mixed-lineage leukemia (trithorax (Drosophila) homolog)	1529 (0.48)	327 (0.57)	− 3.01	AB011399
AIRE	Autoimmune regulator	82 (0.84)	460 (0.75)	1.95	AB006684
AML1	Acute myeloid leukemia 1	219 (0.72)	243 (0.77)	− 0.88	D43969 D43968 D43967
ANK1	Ankyrin 1, erythrocytic	1603 (0.71)	195 (0.65)	1.24	X16609
APOER2	Apolipoprotein E receptor 2	745 (0.66)	167 (0.84)	− 4.77	D86407
APP	Amyloid beta (A4) precursor protein	640 (0.57)	70 (0.57)	0.01	D87675
ATBF1	AT-binding transcription factor 1	2722 (0.63)	886 (0.70)	− 4.29	L32832
ATP2A2	ATPase, Ca^{2+} transporting, cardiac muscle, slow twitch 2	894 (0.50)	46 (0.65)	− 2.05	M23114 M23115
ATP2B3	ATPase, Ca^{2+} transporting, plasma membrane 3	1053 (0.80)	146 (0.71)	1.47	U57971 U60414
BCL2	B-cell CLL/lymphoma 2	144 (0.83)	34 (0.47)	3.70	M13995 M13994
CACNA1C	Calcium channel, voltage-dependent, L type, alpha 1C subunit	669 (0.72)	1474 (0.75)	− 1.56	Z34822 L29534
CACNB1	Calcium channel, voltage-dependent, beta 1 subunit	426 (0.70)	233 (0.67)	1.10	M92303 M92302 M92301
CALCA	Calcitonin/calcitonin-related polypeptide, alpha	73 (0.75)	103 (0.63)	0.85	Xl5943 M26095
CAST	Calpastatin	238 (0.30)	53 (0.38)	− 0.26	M86251 Dl 6217
CD22	CD22 antigen	639 (0.66)	170 (0.66)	0.05	U62631
CD38	CD38 antigen (p45)	112 (0.65)	165 (0.50)	2.31	D84276 D84277
CD44	CD44 antigen	254 (0.54)	450 (0.41)	2.87	L05423 L05424
CD8B1	CD8 antigen, beta polypeptide 1 (p37)	163 (0.72)	76 (0.51)	3.30	X13444 X13445 X13446
CEACAM1	Carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)	178 (0.48)	290 (0.57)	− 2.26	D15202 M76742
CHN1	Chimerin (chimaerin) 1	263 (0.42)	234 (0.46)	− 0.86	Z22641 S75654
CLCN6	Chloride channel 6	196 (0.63)	54 (0.57)	0.34	X99473 X96391 X99474 X99475
CSF2RA	Granulocyte-macrophage colony stimulating factor 2 receptor, alpha	264 (0.54)	120 (0.52)	0.12	L29349 L29348
DAF	Decay-accelerating factor	322 (039)	53 (0.45)	− 0.71	M31516 M30142

(continued overleaf)

Table 2 (*continued*)

Gene symbol	Name	Codons		*Z*	GenBank
		con (freq. major)	alt (freq. major)		
DGKZ	Diacylglycerol kinase, zeta	848 (0.080)	289 (0.80)	− 0.51	U94905 U51477
DLK1	*Drosophila* delta-like 1	264 (0.83)	68 (0.85)	− 0.09	U15981 U15979
DSCR1	Down syndrome critical region gene 1	143 (0.56)	58 (0.72)	− 2.08	U85265 U85266 U85267
DUSP6	Dual specificity phosphatase 6	224 (0.79)	144 (0.73)	0.99	AB013382 AB013602
ED1	Ectodermal dysplasia	131 (0.77)	318 (0.51)	5.21	AF061194 AF061193 AF061192 AF061191 AF061190 AF061189 AF040628
ELN	Elastin	649 (0.40)	65 (0.54)	− 1.42	M36860 U93037
EPHB2	Ephrin receptor ephb2	897 (0.79)	68 (0.49)	5.27	L41939 AF025304
FCAR	Fc fragment of IgA receptor	76 (0.62)	194 (0.61)	− 0.25	U43677 U43774 X54150
FGFR2	Fibroblast growth factor receptor 2	299 (0.62)	523 (0.54)	2.17	U11814 M80634 AF097345
FLT3LG	Fms-related tyrosine kinase 3 ligand	115 (0.78)	75 (0.76)	0.35	U29874
FUT6	Fucosyltransferase 6 (alpha (1,3) fucosyltransferase)	239 (0.81)	100 (0.80)	0.47	U27334 U27332 U27331
GHRHR	Growth hormone-releasing hormone receptor	296 (0.73)	106 (0.68)	1.1 3	U17579
GLB1	Galactosidase, beta 1	492 (0.58)	124 (0.54)	0.22	M27507 M27508
HLA-G	HLA-G histocompatibility antigen, class I, G	139 (0.81)	167 (0.77)	1.01	M90683 M90684 M90686
IGF1	Insulin-like growth factor 1 (somatomedin C)	114 (0.72)	78 (0.47)	3.87	M12659 X56773
IL5RA	Interleukin 5 receptor, alpha	303 (0.41)	86 (0.44)	− 0.41	M96651 M96652
ITSN	Intersectin	1174 (0.41)	482 (0.65)	− 8.80	AF064244 AF064243
KCNAB1	Potassium voltage-gated channel, shaker-related subfamily, beta member 1	309 (0.45)	146 (0.58)	− 2.74	U33428 U16953
KL	Klotho	508 (0.71)	472 (0.54)	5.17	AB009667
KNG	Kininogen	394 (0.49)	261 (0.33)	3.72	M11437
LIMK2	LIM domain kinase 2	540 (0.73)	50 (0.64)	0.85	AC002073
MAN2A2	Mannosidase, Alpha, Class 2A, member 2	752 (0.74)	343 (0.72)	0.00	D55649 L28821
MICA	MHC class I polypeptide-related sequence	185 (0.61)	155 (0.43)	3.05	AF010446 AF010447 AF031469 U22963
MSH6	mutS (*E. coli*) homolog 6	1020 (0.42)	302 (0.31)	3.52	D89641 D89645
MYCL1	v-myc Avian myelocytomatosis viral oncogene homolog 1	158 (0.84)	235 (0.61)	4.73	M19720
NCAM1	Neural cell adhesion molecule 1	580 (0.67)	319 (0.65)	0.17	X16841 M22092 S71824
NRG1	Heregulin, alpha \945kf, RTNN2 p185-activator)	168 (0.48)	305 (0.83)	− 7.26	L12261 L12260

Table 2 (*continued*)

Gene symbol	Name	Codons		Z	GenBank
		con (freq. major)	alt (freq. major)		
PACE4	Paired basic amino acid cleaving system 4	293 (0.68)	769 (0.66)	0.49	AB001914
PAX8	Paired box gene 8	248 (0.73)	89 (0.72)	− 0.07	S77904 S77905 L19606
PCDH2	Protocadherin 2 (cadherin-like 2, pc43)	672 (0.62)	244 (0.68)	− 1.68	L11373 L11372 L11371
PDE4A	Phosphodiesterase 4A, camp-specific (dunce (*Drosophila*)-homolog phosphodiesterase E2)	259 (0.83)	645 (0.71)	3.42	L20965 AF069491
PPP2R3	Protein phosphatase 2 (formerly 2A), regulatory subunit B″	467 (0.41)	691 (0.36)	1.84	L12146 L07590
PRKCB1	Protein kinase C, beta 1	592 (0.64)	101 (0.50)	2.36	X07109 X06318
PS-PLA1	Phosphatidylserine-specific phospholipase Al	363 (0.59)	79 (0.51)	1.29	AF035269 AF035268
PTPRS	Protein tyrosine phosphatase, receptor, sigma	1439 (0.79)	399 (0.86)	− 2.93	U41725 U40317
RBP-MS	RNA-binding protein gene with multiple splicing	173 (0.51)	82 (0.52)	− 0.52	D084110 D84109 D84108 D84107
RTN2	Reticulon 2	196 (0.76)	331 (0.62)	1.84	AF004222 AF004223 AF004224
SHBG	Sex hormone-binding globulin	235 (0.60)	167 (0.68)	− 1.02	X16349 X16351
SHMT1	Serine hydroxymethyltransferase 1	391 (0.66)	77 (0.57)	1.77	L23928
Sox30	SOX30 protein	451 (0.63)	249 (0.45)	3.99	AB022441 AB022083
SYNJ1	Inositol 5-phosphatase (synaptojanin 1)	1240 (0.33)	270 (0.37)	− 1.56	AF009040 AF009039
TACR1	Tachykinin 1 receptor (substance P receptor, neurokinin 1 receptor)	289 (0.76)	94 (0.82)	− 1.72	M84426 M84425
TBXA2R	Thromboxane A2 receptor	259 (0.91)	54 (0.67)	4.75	U11271 AC005175
TCF7	Transcription factor 7 (T-cell specific, HMG-box)	220 (0.75)	73 (0.49)	3.39	Z47361 Z47362 Z47363 Z47364
THRA	Thyroid hormone receptor, alpha (avian erythroblastic leukemia viral (v-erb-a) oncogene homolog)	354 (0.73)	157 (0.70)	1.39	J03239 M24748
TNFRSF6	Tumor necrosis factor receptor superfamily, member 6	63 (0.49)	237 (0.38)	1.57	Z47993 Z47994 Z47995
TPO	Thyroid peroxidase	791 (0.72)	55 (0.69)	0.82	J02970 J02969
XE7	XE7	344 (0.87)	297 (0.89)	− 0.68	L03426

[a] Gene abbreviations are from the Human Gene Nomenclature Database (http://www.gene.ucl.ac.uk/nomenclature/) for all genes except: *A*, *PS-PLA1*, *RBP-MS*, *Sox30*, and *XE7*. The latter are given in the respective GenBank files. The numbers of codons examined in constitutive and alternative exons are shown. The overall frequencies of putative major codons in constitutive or alternative exons are shown in parentheses. The Z values for the Mantel–Haenszel test (see text) for data pooled across synonymous families are given for each gene. The GenBank accession number(s) (Release 117.0) is shown for each gene. Alternative splicing has been confirmed by comparing cDNA and genomic DNA sequence data (either through restriction map studies or by direct sequencing of genomic DNA) for all genes except the following: *ANK1*, *NRG1*, *PPP2R3*, and *PTPRS*.

Table 3
Contingency tables comparing codon usage in constitutive and alternative exons[a]

Gene symbol	Amino acid	Codons	con	alt
Btk29A	His	CAC (major)	11	2
		CAT (non-major)	8	7
alpha-Man-l	Lys	AAG (major)	20	10
		AAA (non-major)	6	9

[a] 2 × 2 Contingency tables comparing major codon usage within a synonymous family in constitutive and alternatively spliced exons in two *D. melanogaster* genes. The frequency of the major codon, CAC, among histidine codons is higher in constitutive than in alternatively spliced exons in the *Btk* gene. Similarly, AAG, the major codon for lysine, is found at higher frequencies in constitutive exons of the *alpha-Man-l* gene. For synonymous families with multiple major or non-major codons, the counts within each class were summed in the table cells.

fied as regions where the GC content is greater than 50% and the observed over expected frequency (O/E) of CpGs is greater than 0.6 (Gardiner-Garden and Frommer, 1987). Another method identifies CpG islands as regions where the frequency of CpG over GpC is greater than a set value (Bird, 1987; Eyre-Walker, 1999). We used values of 0.8 or 0.9 for this analysis. The frequencies of CpG and GpC were calculated for 100 bp windows sliding across a sequence at 1 bp intervals. For exons less than 100 bp, their lengths were used as the window size. The expected frequency of CpG within each window was calculated as the product of the frequencies of C and G nucleotides in the window.

We also attempted to minimize the effect of CpG islands in comparisons of silent divergence. CpG island densities are reduced in mouse genes relative to human orthologs and, in the few cases examined, mouse CpG island regions appear to be a subset of those found in humans (Antequera and Bird, 1993; Matsuo et al., 1993). Thus, we eliminated regions identified as CpG islands in humans in the between-species alignments. All sequences and data discussed below are available from the authors.

3. Results and discussion

3.1. Base-composition comparisons in alternatively spliced genes in D. melanogaster

In *D. melanogaster,* a number of lines of evidence support translational selection at silent sites (Shields et al., 1988; Sharp and Li, 1989; Kliman and Hey, 1993, 1994; Moriyama and Hartl, 1993; Akashi, 1994, 1995; Akashi and Schaeffer, 1997; Moriyama and Powell, 1997; Powell and Moriyama, 1997; Comeron et al., 1999; Duret and Mouchiroud, 1999). We first applied comparisons of constitutive and alternatively spliced exons to *D. melanogaster* genes in order to confirm the statistical power of our method. In *D. melanogaster*, candidates of major synonymous codons have been established in previous studies (Sharp and Lloyd, 1993; Akashi, 1995). These codons are mostly C-ending (Shields et al., 1988) and appear to correspond with tRNA concentrations (Moriyama and Powell, 1997). If our method has sufficient power, constitutive exons should encode major codons more frequently than alternative exons. As predicted, major codon usage is higher in constitutive exons than in alternative exons (MH test across genes, $Z = 4.24$, $P = 0.00001$; Tables 1 and 6).

The above result suggests that our method may have a sufficient power to detect translational selection. However, this comparison may be complicated by a decline in GC content along *D. melanogaster* genes; overall, GC content first increases for several hundred base pairs from translation initiation sites, then decreases gradually in the $5'$ to $3'$ direction (Kliman and Eyre-Walker, 1998). If an excess of constitutive exons is $5'$ relative to alternative exons, such a polarity in base composition could bias the comparison of constitutive and alternative exons. To eliminate this possibility, we restricted the comparison to constitutive exons that are $3'$ to alternative exons. Unfortunately, only 15 genes were left in the analysis, and the number of codons examined in alternative exons was reduced to less than one third of the original number; although the trend remained, there was no longer a statistically significant association between codon usage and constitutive exons (MH test across genes, $Z = 0.99$, $P = 0.16$; Table 6). More data will be required to determine whether base-composition polarity can explain the excess of major codons in constitutive exons.

3.2. Base-composition comparisons in alternatively spliced genes in humans

Although some patterns suggest selection at silent sites in mammals, the evidence for translational selection is not as strong as in *E. coli*, yeast, *C. elegans*, and *Drosophila*. To test for translational selection in humans, we performed the same analysis as described above on 77 human genes. The result was similar to those found in *Drosophila*; GC-ending codons are significantly more frequent in constitutive than in alternative exons (MH test across genes, $Z = 2.97$, $P = 0.0015$; Tables 2 and 6).

The analysis of codon usage in alternatively spliced genes in humans, however, is complicated by the existence of CpG islands. To remove the effects of CpG islands, we employed two methods. Evidence for higher codon bias in constitutive exons remained when we applied the CpG/GpC method. When the maximum ratio allowed in the data was 0.9, the codon bias remained higher in constitutive exons (MH test across genes, $Z = 4.25$, $P = 0.00001$), and the result remained significant (MH test across genes, $Z = 3.04$, $P = 0.0012$) when the ratio was set to 0.8. When the O/E method with a 100 bp window was applied, the result was not statistically significant (Table 6). However, Matsuo et al. (1993) suggested that an increased window size of 500 bp more accurately identifies CpG islands. Using the O/E

Table 4
DNA divergence in alternatively spliced genes between humans and non-human mammals[a]

Gene name	Non-human mammal	Constitutive			Alternative			d_N	d_S	GenBank
		codons	d_N	d_S	codons	d_N	d_S	con–alt	con–alt	
ABL1	*Mus musculus*	1057	0.060	0.659	70	0.013	0.306	0.047	0.353	M12263 M12264 M12265 M12266 J02995
AF-6	*Rattus norvegicus*	1573	0.028	0.492	141	0.127	0.340	− 0.099	0.152	U83231 U83230
ANK1	*M. musculus*	811	0.015	0.520	173	0.052	0.282	− 0.037	0.238	X69063 X69064 X69065 M84756
APP	*M. musculus*	694	0.015	0.516	71	0.042	0.167	− 0.027	0.348	M18373 M24397
BCL2	*M. musculus*	170	0.031	0.454	43	0.040	0.500	− 0.009	− 0.046	M16506 L31532
CEACAM1	*R. norvegicus*	71	0.292	0.480	50	0.213	0.176	0.079	0.304	J04963 X71122
CACNA1C	*M. musculus*	371	0.006	0.567	55	0.000	0.063	0.006	0.504	L01776
ATP2B3	*R. norvegicus*	1096	0.008	0.842	147	0.035	0.387	− 0.027	0.455	J05087 M96626
CALCA	*R. norvegicus*	63	0.133	0.338	100	0.047	0.337	0.086	0.002	L00109 L00110 L29188 M31027
CD44	*M. musculus*	242	0.088	0.870	229	0.227	0.257	− 0.139	0.613	X66081 X66082 X66083 X66084
ED1	*M. musculus*	132	0.070	0.350	243	0.008	0.248	0.062	0.102	AF004435 AF016628 AF016630 AF016331
GHRHR	*B. taurus*	318	0.085	0.506	35	0.180	0.860	− 0.095	− 0.353	AB022596 AB022597
ATP2A2	*R. norvegicus*	993	0.004	0.449	51	0.027	0.234	− 0.023	0.216	J04022 J04023
KCNAB1	*Oryctolagus cuniculus*	327	0.000	0.362	151	0.050	0.278	− 0.050	0.084	AF131934 AF131935
FGFR2	*M. musculus*	197	0.007	0.560	92	0.010	0.066	− 0.004	0.493	X55441 M63503
KNG	*B. taurus*	383	0.143	0.401	240	0.167	0.269	− 0.024	0.132	V01491
KL	*M. musculus*	527	0.055	0.562	474	0.111	0.620	− 0.055	− 0.058	AB005141 AB010088
LIMK2	*R. norvegicus*	599	0.024	0.740	52	0.160	0.441	− 0.136	0.299	AB005131 AB005132
MICA	*M. musculus*	41	0.179	0.635	87	0.138	0.258	0.041	0.377	AF010448 AF010449
NCAM1	*M. musculus*	577	0.027	0.705	170	0.038	0.538	− 0.011	0.167	X15049 X15050 X15051 X14526 X14527 X14402 X14403
PRKCB1	*R. norvegicus*	621	0.007	0.483	102	0.000	0.099	0.007	0.384	X04439 X04440
PTPRS	*M. musculus*	598	0.022	1.650	312	0.046	0.779	− 0.024	0.871	D28530 D28531
SHBG	*R. norvegicus*	200	0.187	0.390	37	0.145	0.745	0.042	− 0.354	M38759 M31179

(continued overleaf)

Table 4 (*continued*)

Gene name	Non-human mammal	Constitutive			Alternative			d_N	d_S	GenBank
		codons	d_N	d_S	codons	d_N	d_S	con–alt	con–alt	
SHMT1	*M. musculus*	259	0.029	0.991	37	0.102	0.596	− 0.074	0.395	X94478 X94479
SYNJ1	*R. norvegicus*	875	0.035	0.515	248	0.164	0.382	− 0.129	0.134	U45479
THRA	*R. norvegicus*	369	0.006	0.191	160	0.029	0.008	− 0.024	0.184	M18028 M31174

[a] The species of non-human mammals is shown in the second column and the GenBank accession number is given in the last column. The numbers of constitutive and alternative codons compared for each gene, and the synonymous, d_S, and non-synonymous, d_N, divergence, calculated according to Yang and Nielsen (2000) are shown. The differences in DNA divergence between constitutive and alternative exons are shown.

method with this window size gave a result that differed from the same method with a smaller window size; codon bias was higher in constitutive exons (MH test across genes, $Z = 2.90$, $P = 0.0019$; Table 6). Examination of genomic, rather than cDNA, sequence may allow more accurate detection of CpG islands. However, among the human genes examined, only four had genomic sequences available and satisfied the criteria of the minimum number of codons after removing CpG regions. In these four genes, the ratio of the frequencies of major codons in constitutive and alternative exons remained virtually identical to that obtained using cDNA sequences. The higher GC content at silent sites of constitutive exons appears to be robust to several methods of CpG island identification.

3.3. Divergence data

In comparisons between *E. coli* vs. *Salmonella typhimurium* (Sharp and Li, 1986, 1987) and among *Drosophila* species (Sharp and Li, 1989), silent divergence is inversely related to codon usage bias. If constitutive exons have higher GC content resulting from stronger selection, synonymous divergence should also be lower in constitutive than in alternative exons. Unfortunately, orthologs for the alternatively spliced genes of *D. melanogaster* are not available from other *Drosophila* species. A limited number of alternatively spliced genes could be analyzed between humans and other mammals. Contrary to our prediction, synonymous divergence between humans and non-human mammals was significantly higher in constitutive exons ($P < 0.005$). This difference remained significant when CpG regions were eliminated. Intriguingly, non-synonymous divergence was marginally significantly higher in alternative exons. Duret and Mouchiroud (2000) found lower nonsynonymous divergence in genes expressed in multiple tissues when compared to genes with more limited expression patterns. Our result shows a similar trend within genes; constitutively expressed exons have lower rates of nonsynonymous divergence. However, this result was no longer significant when CpG regions were eliminated (Table 5).

3.4. Multiple forces acting on base composition in mammalian coding regions

A number of factors may play a role in determining base composition within protein-coding regions. CpG islands may be a complicating factor in the analysis of human codon bias. In addition, some sequences that reside within an exon are important for alternative splice-site selection. Exonic splicing enhancers are usually purine-rich (reviewed in Wang et al., 1997) or A/C-rich (Coulter et al., 1997), and exonic splicing suppressors appear to be pyrimidine-rich (mostly C) although they have considerable sequence variation (Zheng et al. 1998). They are often found at the 5′ and 3′ ends, and sometimes in the middle, of an exon (Wang et al., 1997;

Table 5
DNA divergence in alternatively spliced human and non-human mammalian genes[a]

Method	Genes	Codons		P	
		con	alt	d_N	d_S
–	26	13164	3570	< 0.05 (alt > con)	< 0.005 (con > alt)
GpC O/E ≤0.6	19	5477	2109	n.s.	< 0.025 (con > alt)
GpC/CpG ≤0.8	21	5661	1921	n.s.	< 0.025 (con > alt)
GpC/CpG ≤0.9	21	8236	2330	< 0.05 (alt > con)	< 0.025 (con > alt)

[a] Summary of data from Table 5 for comparisons of DNA divergence between humans and non-human mammals in alternatively spliced genes. The total numbers of genes and codons compared are shown. Two-tailed probabilities from Wilcoxon's signed-ranks tests and the directions of the overall deviation are given in the last two columns (i.e., 'alt > con' refers to greater divergence in alternative than in constitutive exons). Results for full data (–) and restricted data sets are also shown (see text for methods).

Table 6
Codon usage comparisons in *D. melanogaster* and human genes[a]

Species	Method	Genes	Codons		Tables	Z	P
			con	alt			
D. melanogaster	–	33	19415	12820	605	4.24	0.000011
	$5'$ alt vs. $3'$ con	15	7256	3960	271	0.99	0.16
Human	–	77	36120	19000	1391	2.97	0.0015
	O/E CpG ≤0.6	44	12707	7467	775	0.68	0.25
	O/E CpG ≤ 0.6 (500 bp windows)	54	19874	10999	956	2.90	0.0019
	CpG/GpC ≤ 0.8	56	14617	9544	950	3.04	0.0012
	CpG/GpC ≤0.9	63	17913	11548	1082	4.25	0.000011

[a] Comparisons of major codon usage between constitutive and alternative exons are shown for *D. melanogaster* and human genes. Analyses for restricted data for each species are shown for a number of different methods (see text). –, Refers to the complete data. Genes, refers to the number of loci examined. The total number of codons examined under each method are shown for constitutive and alternative exons, and the number of 2 × 2 contingency tables in each analysis is shown. The *Z* values for the Mantel–Haenszel test and the one-tailed *P* values are given.

König et al., 1998; Muro et al., 1998) and may regulate the accessibility of different exons to the splicing machinery through the formation of secondary structures (Wang et al., 1997). Because very few such regions have been well-characterized in the genes examined, the effects of splicing enhancers/suppressors on our analysis are unclear.

Finally, DNA structural constraints may maintain certain dinucleotide contents (Karlin and Mrázek, 1996). However, the difference in our data between constitutive and alternative exons in a gene may be difficult to explain by this effect; it is unlikely that constitutive exons have more structural constraints than alternative exons in the same gene.

3.5. Translational selection in the human genome

Recent experiments show that altering codon usage to G- and C-ending codons can enhance the expression levels of genes in human cell lines (Kim et al., 1997; André et al., 1998). Although such results demonstrate biochemical variation caused by synonymous codon usage, evidence that natural selection has acted upon such variation in the human evolutionary lineage remains elusive. For some amino acids, Hatfield and Rice (1986) showed a correspondence between tRNA abundance in human and rabbit reticulocytes and the codon usage of alpha- and beta-globin mRNAs. In addition, Alvarez-Valin et al. (1998) showed a correlation between synonymous divergence and amino acid conservation within mammalian genes. The latter pattern is consistent with codon selection for translational accuracy (Akashi, 1994). However, Duret and Mouchiroud (2000) found no relationship between expression patterns and synonymous codon usage in human genes, and Smith and Hurst (1999) found no relationship between codon usage bias and synonymous divergence among mammals. It remains unclear whether selection at silent sites is less effective in mammals or whether such effects are masked by other factors such as isochores and CpG islands. Quantification of tRNA concentrations in a large number of tissues and developmental stages, in combination with analysis of gene expression and local base-composition data, may be necessary to resolve whether natural selection discriminates among synonymous codons to enhance protein synthesis in humans.

Acknowledgements

We thank Adam Eyre-Walker and an anonymous reviewer for many helpful suggestions. This research was supported by a grant from the Alfred P. Sloan Foundation.

References

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

Akashi, H., 1995. Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in Drosophila DNA. Genetics 139, 1067–1076.

Akashi, H., Schaeffer, S.W., 1997. Natural selection and the frequency distributions of 'silent' DNA polymorphism in *Drosophila*. Genetics 146, 295–307.

Andersson, S.G.E., Kurland, C.G., 1990. Codon preferences in free-living microorganisms. Microbiol. Rev. 54, 198–210.

André, S., Seed, B., Eberle, J., Schraut, W., Bültmann, A., Haas, J., 1998. Increased immune response elicited by DNA vaccination with a synthetic gp120 sequence with optimized codon usage. J. Virol. 72, 1497–1503.

Antequera, F., Bird, A., 1993. Number of CpG islands and genes in human and mouse. Proc. Natl. Acad. Sci. USA 90, 11995–11999.

Alvarez-Valin, F., Jabbari, K., Bernardi, G., 1998. Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. J. Mol. Evol. 46, 37–44.

Bernardi, G., 1993. The vertebrate genome: Isochores and evolution. Mol. Biol. Evol. 10, 186–204.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.

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

Bird, A.P., 1987. CpG islands as gene markers in the vertebrate nucleus. Trends Genet. 3, 342–347.

Bulmer, M., 1988a. Evolutionary aspects of protein synthesis. In: Harvey, P.H., Partridge, L. (Eds.). Oxford Survey Evolution Biology 5. Oxford University Press, Oxford, pp. 1–40.

Bulmer, M., 1988b. Are codon usage patterns in unicellular organisms determined by selection-mutation balance. J. Evol. Biol. 1, 15–26.

Cacciò, S., Zoubak, S., D'Onofrio, G., Bernardi, G., 1995. Nonrandom frequency patterns of synonymous substitutions in homologous mammalian genes. J. Mol. Evol. 40, 280–292.

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

Comeron, J.M., Kreitman, M., Aguade, M., 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics. 151, 239–249.

Coulondre, C., Miller, J.H., Farabough, P.J., Gilbert, W., 1978. Molecular bases of base substitution hotspots in *Escherichia coli*. Nature 274, 775–780.

Coulter, L.R., Landree, M.A., Cooper, T.A., 1997. Identification of a new class of exonic splicing enhancers by in vivo selection. Mol. Cell. Biol. 17, 2143–2150.

Cross, S.H., Bird, A.P., 1995. CpG islands and genes. Curr. Opin. Genet. Dev. 5, 309–314.

Duret, L., 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. 16, 287–289.

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.

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

Eyre-Walker, A., 1996. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol. Biol. Evol. 13, 864–872.

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.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection, Clarendon Press, Oxford.

FlyBase, 1999. The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Res. 27, 85–88 (http://flybase.bio.indiana.edu/).

Gardiner-Garden, M., Frommer, M., 1987. CpG islands in vertebrate genomes. J. Mol. Biol. 196, 261–282.

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

Hartl, D.L., Moriyama, E.N., Sawyer, S., 1994. Selection intensity for codon bias. Genetics 138, 227–234.

Hatfield, D., Rice, M., 1986. Aminoacyl-tRNA (anticodon): codon adaptation in human and rabbit reticulocyte. Biochem. Int. 13, 835–842.

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* translation system. J. Mol. Biol. 151, 389–409.

Ikemura, T., 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.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 13–34.

Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., Gibson, T.J., 1998. Multiple sequence alignment with Clustal X. Trends Biochem. Sci. 23, 403–405.

Jones, P.A., 1999. The DNA methylation paradox. Trends Genet. 15, 34–37.

Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T., 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.

Karlin, S., Mrázek, J., 1996. What drives codon choices in human genes? J. Mol. Biol. 262, 459–472.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Kim, C.H., Oh, Y., Lee, T.H., 1997. Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells. Gene 199, 293–301.

Kliman, R.M., Eyre-Walker, A., 1998. Patterns of base composition within the genes of *Drosophila melanogaster*. J. Mol. Evol. 46, 534–541.

Kliman, R.M., Hey, J., 1993. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. 10, 1239–1258.

Kliman, R.M., Hey, J., 1994. The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. Genetics 137, 1049–1056.

König, H., Ponta, H., Herrlich, P., 1998. Coupling of signal transduction to alternative pre-mRNA splicing by a composite splice regulator. EMBO J. 17, 2904–2913.

Li, W-H., 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. 24, 337–345.

Lindahl, T., 1982. DNA repair enzymes. Annu. Rev. Biochem. 51, 61–87.

Mantel, N., 1963. Chi-squared tests with one degree of freedom; extensions of the MANTEL-HAENSZEL procedure. J. Am. Stat. Assoc. 58, 690–700.

Mantel, N., Haenszel, W., 1959. Statistical aspects of the analysis of data from the retrospective analysis of disease. J. Natl. Cancer Inst. 22, 719.

Matsuo, K., Clay, O., Takahashi, T., Silke, J., Schaffner, W., 1993. Evidence for erosion of mouse CpG islands during mammalian evolution. Somat. Cell. Mol. Genet. 19, 543–555.

Moriyama, E.N., Hartl, D.L., 1993. Codon usage bias and base composition of nuclear genes in Drosophila. Genetics 134, 847–858.

Moriyama, E.N., Powell, J., 1997. Codon usage bias and tRNA abundance in *Drosophila*. J. Mol. Evol. 45, 514–523.

Mouchiroud, D., Gautier, C., Bernardi, G., 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J. Mol. Evol. 40, 107–113.

Muro, A.F., Iaconcig, A., Baralle, F.E., 1998. Regulation of the fibronectin EDA exon alternative splicing. Cooperative role of the exonic enhancer element and the 5′ splicing site. FEBS Lett. 437, 137–141.

Ohta, T., 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Sys. 23, 263–286.

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

Precup, J., Parker, J., 1987. Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. 262, 11351–11356.

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

Sharp, P.M., Li, W-H., 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4, 222–230.

Sharp, P.M., Li, W-H., 1989. On the rate of DNA sequence evolution in *Drosophila*. J. Mol. Biol. 28, 398–402.

Sharp, P.M., Lloyd, A.T., 1993. Codon usage. In: Maroni, G. (Ed.). An Atlas of *Drosophila* Genes: Sequences and Molecular Features. Oxford University Press, Oxford, pp. 378–397.

Sharp, P.M., Stenico, M., Peden, J.F., Lloyd, A.T., 1993. Codon usage:

mutational bias, translational selection, or both? Biochem. Soc. Trans. 21, 835–841.

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

Smith, N.G.C., Hurst, L.D., 1999. The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? Genetics 152, 661–673.

Sorensen, M.A., Higgins, D.G., Wright, F., 1989. Codon usage determines translation rate in *Escherichia coli*. J. Mol. Biol. 207, 365–377.

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

Tamarin, R.H., 1999. Principles of Genetics, sixth ed. McGraw-Hill, New York.

Wang, Y-C, Selvakumar, M., Helfman, D.M., 1997. Alternative Pre-mRNA Splicing. In: Krainer, A.R. (Ed.). Eukaryotic mRNA Processing. Oxford University Press, Oxford, pp. 242–279.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13, 555–556.

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

Zheng, Z-M., Huynen, M., Baker, C.C., 1998. A pyrimidine-rich exonic splicing suppressor binds multiple RNA splicing factors and inhibits spliceosome assembly. Proc. Natl. Acad. Sci. USA 95, 14088–14093.

Zoubak, S., D’Onofrio, G., Cacciò, S., Bernardi, G., Bernardi, G., 1995. Specific compositional patterns of synonymous positions in homologous mammalian genes. J. Mol. Evol. 40, 293–307.