Codon usage between genomes is constrained by genome-wide mutational processes

Swaine L. Chen*, William Lee†, Alison K. Hottes, Lucy Shapiro, and Harley H. McAdams

Department of Developmental Biology, Stanford University School of Medicine, Beckman Center, B300, Stanford, CA 94304

Contributed by Lucy Shapiro, December 9, 2003

Analysis of genome-wide codon bias shows that only two parameters effectively differentiate the genome-wide codon bias of 100 eubacterial and archaeal organisms. The first parameter correlates with genome GC content, and the second parameter correlates with context-dependent nucleotide bias. Both of these parameters may be calculated from intergenic sequences. Therefore, genome-wide codon bias in eubacteria and archaea may be predicted from intergenic sequences that are not translated. When these two parameters are calculated for genes from nonmammalian eukaryotic organisms, genes from the same organism again have similar values, and genome-wide codon bias may also be predicted from intergenic sequences. In mammals, genes from the same organism are similar only in the second parameter, because GC content varies widely among isochores. Our results suggest that, in general, genome-wide codon bias is determined primarily by mutational processes that act throughout the genome, and only secondarily by selective forces acting on translated sequences.

Translation of mRNA to protein is universal, and the genetic code describing how the 64 nucleotide triplets (codons) specify 20 amino acids is nearly universal (1). Grantham's genome hypothesis proposes that each species systematically uses certain synonymous codons (codons that code for the same amino acid) in coding sequences (2–4), in other words, that each species has a distinct codon bias. Many studies have since confirmed that, at least in prokaryotes, selective forces acting at the level of translation maintain biased codon usage (5–7). The realization that selection may act on gene sequences in the absence of amino acid changes has had profound implications for the study of the molecular evolution of genes. In particular, analysis of codon bias has helped establish that horizontal gene transfer is a major evolutionary force (8–10).

What causes differences in codon bias and why? Does codon bias exist (*i*) because it is necessary for efficient and accurate protein expression or (*ii*) because codons, as DNA sequences, are subject to mutational pressures acting on all the DNA sequences in a given organism? Explanation *i* is generally termed a selective or selectionist explanation for codon bias. In contrast, explanation *ii* is referred to as a neutral or mutational explanation. Variation in codon bias among genes from the same organism has been shown to depend on many parameters, including expression level (4, 5, 11), amino acid composition (12–15), gene length (16, 17), mRNA structure (18–20), and protein level noise considerations (21). In most of these cases, evidence exists that selection at different steps during protein expression shapes codon bias. In addition, global forces differentiate the codon bias of genes between different organisms: species-specific codon bias is strongly correlated with overall genome percentage GC content (22, 23), genes from organisms with similar phylogeny or with similar tRNA content have similar codon bias (22), and an organism's optimal growth temperature influences the codon bias of its genes (24). Most of these global forces are thought to be mutational, acting on all DNA sequences, although it has also been argued that growth temperature exerts a selective force on mRNA structure (25) and codon bias (24). Although both selection and mutation are clearly important for establishing codon bias, the relative importance of selection and mutation has been difficult to define in general.

With the recent availability of many complete genome sequences, it has become possible to directly analyze the determinants of genome-wide codon bias. Studying genome-wide codon bias allows us to focus on global forces shaping codon bias. In this article, we examine the relative importance of mutation versus selection in shaping genome-wide codon bias.

Because each of the 20 amino acids, on average, is encoded by three synonymous codons, the space of possible patterns of codon usage is very large. Here we analyze variation in codon bias of genes in archaeal and eubacterial organisms. We introduce a method to quantitatively separate within-genome from between-genome variation in codon bias, and, to analyze global forces shaping codon bias, we focus on between-genome variation. We find, consistent with others, that GC content variation is the most important parameter differentiating codon bias between different organisms. A key finding in our analysis is that a combination of nearest-neighbor nucleotide biases is the next most important parameter differentiating codon bias between different organisms. We demonstrate that genome-wide codon bias in prokaryotic genomes may be predicted with surprising accuracy by using only intergenic sequence statistics, which are unaffected by selective forces acting during protein expression. Furthermore, we find that the codon bias of genes from several nonmammalian eukaryotes is also characterized by genome GC content and nearest-neighbor nucleotide biases. We conclude that genome-wide codon bias can be well characterized by only two parameters, which are determined predominantly by genome-wide mutational forces rather than by coding-region-specific selective forces in all three domains of life.

Materials and Methods

Data Sources. All genome sequences are from GenBank (ftp://ftp.ncbi.nih.gov). A list of genome sequences used is in *Data Set 1*, which is published as supporting information on the PNAS web site. Some eukaryotic sequences were taken from the RefSeq project (www.ncbi.nlm.nih.gov/RefSeq) on June 5, 2003; only sequences marked "provisional," "reviewed," or "validated" were used. Genome sequence processing was done with PERL (www.perl.com) with the genome-tools (http://genome-tools.sourceforge.net) (26) and PDL packages (http://pdl.perl.org) by using ad hoc scripts on DEBIAN GNU/LINUX 3.0 (www.debian.org and www.gnu.org). Growth temperature data were taken from the recommended culture conditions of American Type Culture Collection (www.atcc.org) or from ref. 27.

Using the Singular Valve Decomposition (SVD) to Find a Basis for the Space of Codon Vectors. Similar to ref. 22, we represent the codon bias of a gene, i, with a codon vector, $\boldsymbol{c_i}$, with components $c_{i,m(w)}$,

Abbreviation: SVD, singular valve decomposition.

*To whom correspondence should be addressed. E-mail: slchen@stanford.edu.

†Present address: Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305.

www.pnas.org/cgi/doi/10.1073/pnas.0307827100

Table 1. Frequently used symbols

m	Index variable for amino acids
g	Index variable for genomes
$m(w)$	wth codon for amino acid m
N	Number of genes selected from each genome for SVD (400)
G	Number of prokaryotic genomes used in study (100)
$\boldsymbol{c_i}$	Codon vector for gene i
$\boldsymbol{c_i^g}$	Codon vector for gene i from genome g
$\boldsymbol{\bar{c}^g}$	Genome-wide codon bias for genome g; mean of the codon vectors for *all* genes in g, not just the N used in the SVD
$\boldsymbol{\bar{c}}$	Mean codon vector for the NG selected genes
$\boldsymbol{\tilde{c}_i}$	Mean-centered codon vector for gene i
$\boldsymbol{v_j}$	jth eigencodon
$v_{m(w),j}$	Component of $\boldsymbol{v_j}$ representing codon $m(w)$
σ_j	jth singular value; global scale factor representing the variance of the selected NG genes in the direction of $\boldsymbol{v_j}$
$u_{i,j}$	Amount of $\boldsymbol{v_j}$, scaled by σ_j, in $\boldsymbol{\tilde{c}_i}$
$\bar{u}_j$	Mean usage of $\sigma_j\boldsymbol{v_j}$ among the mean-centered codon vectors for all genes in all genomes
$\bar{u}_j^g$	Mean of $u_{i,j}$ taken over all genes i from genome g; mean usage of $\sigma_j\boldsymbol{v_j}$ among the mean-centered codon vectors for all genes from genome g
$\mathrm{var}(u_j^g)$	Variance in $u_{i,j}$ among all genes from genome g; variance in usage of $\sigma_j\boldsymbol{v_j}$ among the mean-centered codon vectors for all genes from genome g
$\mathrm{var}(u_j)$	Overall variance in $u_{i,j}$ for all genes, i, among all genomes, g
$\mathrm{var}(u_j)^{\mathrm{within}}$	Within-genome variance in $\mathrm{var}(u_j)$
$\mathrm{var}(u_j)^{\mathrm{between}}$	Between-genome variance in $\mathrm{var}(u_j)$
$\boldsymbol{d_g}$	Vector of intergenic bias parameters for genome g

where $c_{i,m(w)}$ is the codon frequency of the $m(w)$th codon (the wth codon for amino acid m), normalized for amino acid content (notation used throughout is listed in Table 1). Each $c_{i,m(w)}$ is calculated as $c_{i,m(w)} = f_{i,m(w)}/\Sigma_{w=1}^{M(m)} f_{i,m(w)}$, where $f_{i,m(w)}$ is the number of times the $m(w)$th codon is used in gene i, and $M(m)$ is the number of synonymous codons that code for amino acid m. The denominator in the calculation of $c_{i,m(w)}$ normalizes for amino acid content in that the sum of the components $c_{i,m(w)}$, which code for the same amino acid, add to 1 regardless of how many times that amino acid is coded for in the gene. Start codons and stop codons are excluded from the calculations of $c_{i,m(w)}$. After excluding stop codons, $\boldsymbol{c_i}$ is a 61-dimensional vector. When the genome, g, from which a gene, i, was taken is relevant, it is denoted with a superscript, as in $\boldsymbol{c_i^g}$. Because different genomes contain different numbers of genes, we randomly selected $N = 400$ genes from each of the $G = 100$ genomes so that each genome had equal weight in the SVD. Altogether, $NG = 40{,}000$ genes were selected. We define the mean codon vector for the genes in the study as the mean codon vector of these 40,000 randomly selected genes, $\boldsymbol{\bar{c}} = (1/NG)\Sigma_{i=1}^{NG} \boldsymbol{c_i}$. Let $\boldsymbol{\tilde{c}_i^g} = \boldsymbol{c_i^g} - \boldsymbol{\bar{c}}$ be the mean-centered codon vector for gene i in genome g. Define the matrix

$$C = [\boldsymbol{\tilde{c}_1^1}, \ldots, \boldsymbol{\tilde{c}_N^1}, \boldsymbol{\tilde{c}_{N+1}^2}, \ldots, \boldsymbol{\tilde{c}_{2N}^2}, \ldots, \boldsymbol{\tilde{c}_{N(G-1)+1}^G}, \ldots, \boldsymbol{\tilde{c}_{NG}^G}]^T, \quad [1]$$

where each row is a mean-centered codon vector and the superscript T indicates the transpose of the matrix. Only the 400 randomly selected genes from each genome are used in C.

We used a thin SVD (28) to decompose C into $C = USV^T$, where $U^TU = I$, $V^TV = I$, $S = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{61})$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{61} \geq 0$. The matrix V and the singular values σ_j are given in Table 3, which is published as supporting information on the PNAS web site. Fig. 1*a* shows a plot of the σ_j for $j = 1, \ldots, 61$. Because of the normalization for amino acid content, C is not full rank; thus, $\sigma_{42} = \sigma_{43} = \ldots = \sigma_{61} = 0$ and can be excluded from further consideration. By using the SVD, the mean-centered codon vector for gene i can be written as $\boldsymbol{\tilde{c}_i} = \Sigma_{j=1}^{41} u_{i,j}\sigma_j\boldsymbol{v_j}$, where $u_{i,j}$ is the component in the ith row and jth column of U and $\boldsymbol{v_j}$ is the jth column of V. In other words, each mean-centered codon vector is a weighted sum of the columns of V. More specifically, each column of V, $\boldsymbol{v_j}$, is scaled by two different weights. The first weight, σ_j, is the jth singular value and can be thought of as a global scale factor. The larger the value of σ_j, the more the codon vectors vary in the direction of $\boldsymbol{v_j}$. The second scalar weight, $u_{i,j}$, is a gene-specific weight that describes how much $\boldsymbol{v_j}$, scaled globally by σ_j, contributes to $\boldsymbol{\tilde{c}_i}$. We refer to the columns of V, $\{\boldsymbol{v_1}, \ldots, \boldsymbol{v_{41}}\}$, as eigencodons. Table 2 shows values for $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$. The mean normalized usage of eigencodon $\boldsymbol{v_j}$, i.e., the arithmetic mean of $u_{i,j}$ over all genes (not just the subset used in the SVD) is denoted as $\bar{u}_j$ and the corresponding variance is denoted as $\mathrm{var}(u_j)$. The mean and variance of eigencodon $\boldsymbol{v_j}$ usage for all genes (not just the subset used in the SVD) within a single genome g are denoted as $\bar{u}_j^g$ and $\mathrm{var}(u_j^g)$, respectively.

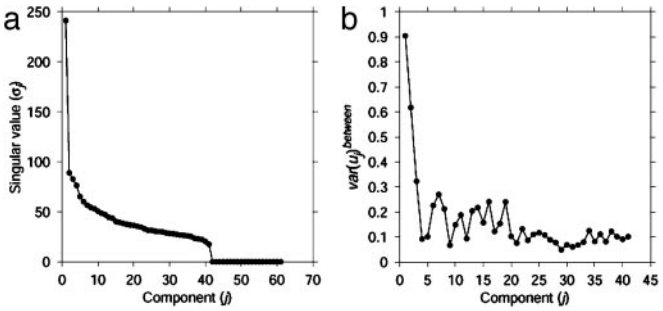

Fig. 1. (*a*) Scree plot of singular values. Singular values (σ_j) were obtained from a SVD of 400 genes from each of 100 genomes. (*b*) Contribution of $\mathrm{var}(u_j)^{\mathrm{between}}$ (between-genome variance) to overall variance. Overall variance is scaled to 1 in each dimension. The rest of the overall variance is due to $\mathrm{var}(u_j)^{\mathrm{within}}$ (within-genome variance). In only two dimensions, $j = 1$ and 2, is $\mathrm{var}(u_j)^{\mathrm{between}}$ the major source of variance.

Separation of Within-Genome from Between-Genome Variation. $\mathrm{Var}(u_j)$ can be decomposed into two parts: (*i*) variance present within individual genomes [within-genome variance, $\mathrm{var}(u_j)^{\mathrm{within}}$] and (*ii*) mean values within genomes that vary from the overall mean [between-genome variance, or $\mathrm{var}(u_j)^{\mathrm{between}}$]. This can be expressed as:

$$\mathrm{var}(u_j) = \frac{\sum_{g=1}^{G} \mathrm{var}(u_j^g)}{G} + \frac{\sum_{g=1}^{G} (\bar{u}_j^g - \bar{u}_j)^2}{G} \quad [2]$$

$$= \mathrm{var}(u_j)^{\mathrm{within}} + \mathrm{var}(u_j)^{\mathrm{between}}, \quad [3]$$

where G is the number of genomes considered. For convenience, $\mathrm{var}(u_j)^{\mathrm{within}}$ and $\mathrm{var}(u_j)^{\mathrm{between}}$ will hereafter refer to a fraction of $\mathrm{var}(u_j)$ [obtained by dividing both sides of the equation above by $\mathrm{var}(u_j)$], such that $\mathrm{var}(u_j)^{\mathrm{within}} + \mathrm{var}(u_j)^{\mathrm{between}} = 1$.

Context-Dependent Intergenic Nucleotide Biases (Intergenic Bias). Context-dependent intergenic nucleotide biases were calculated by using a fixed, second-order Markov model to analyze all intergenic sequences for each of the 100 genomes examined in this study. The frequency of each nucleotide was calculated for

EVOLUTION

Table 2. Codon vectors v_1 and v_2

	Codon	v_1	v_2		Codon	v_1	v_2		Codon	v_1	v_2
Ala	GCA	−0.088	0.077	Gly	GGA	−0.099	0.176	Pro	CCA	−0.107	0.091
	GCC	0.118	−0.056		GGC	0.150	−0.178		CCC	0.076	0.030
	GCG	0.078	−0.093		GGG	0.014	0.022		CCG	0.139	−0.161
	GCT	−0.108	0.071		GGT	−0.065	−0.022		CCT	−0.109	0.040
Arg	AGA	−0.139	0.263	His	CAC	0.165	0.294	Ser	AGC	0.067	−0.023
	AGG	0.004	0.240		CAT	−0.165	−0.292		AGT	−0.062	0.002
	CGA	−0.018	−0.045	Ile	ATA	−0.070	0.297		TCA	−0.063	0.057
	CGC	0.131	−0.231		ATC	0.200	−0.155		TCC	0.064	0.006
	CGG	0.053	−0.053		ATT	−0.130	−0.138		TCG	0.073	−0.068
	CGT	−0.031	−0.176	Leu	CTA	−0.025	0.042		TCT	−0.079	0.026
Asn	AAC	0.187	0.134		CTC	0.077	0.050	Thr	ACA	−0.106	0.084
	AAT	−0.187	−0.132		CTG	0.153	−0.108		ACC	0.144	−0.100
Asp	GAC	0.174	0.064		CTT	−0.031	0.050		ACG	0.067	−0.065
	GAT	−0.174	−0.062		TTA	−0.160	0.012		ACT	−0.105	0.080
Cys	TGC	0.196	−0.162		TTG	−0.014	−0.045	Trp	TGG	0.000	0.000
	TGT	−0.196	0.161	Lys	AAA	−0.181	−0.079	Tyr	TAC	0.167	0.240
Gln	CAA	−0.216	−0.077		AAG	0.181	0.089		TAT	−0.167	−0.244
	CAG	0.217	0.051	Met	ATG	0.000	0.000	Val	GTA	−0.087	0.080
Glu	GAA	−0.133	−0.107	Phe	TTC	0.212	0.064		GTC	0.099	−0.066
	GAG	0.133	0.109		TTT	−0.211	−0.070		GTG	0.098	−0.113
									GTT	−0.110	0.100

each possible combination of nucleotides immediately 5′ and immediately 3′. Taking all intergenic sequences 5′-$N_1N_2N_3$-3′, where N_1 and N_3 are fixed, we calculated the fraction in which N_2 is G, A, T, or C, which we denote as $p(N_2|N_1, N_3)$. All such intergenic three-nucleotide sequences were included in the calculation, except for the first three and last three nucleotides of each intergenic region. Because $\Sigma_{N_2=\mathrm{G,A,T,C}}\, p(N_2|N_1, N_3) = 1$ for all 16 pairs of N_1 and N_3, 64 parameters exist of which $64 - 16 = 48$ are linearly independent. In total, this set of 64 nearest-neighbor nucleotide bias parameters calculated from the intergenic sequences of genome g is denoted as $\boldsymbol{d_g}$. For a given organism, g, we will refer to the set of parameters $\boldsymbol{d_g}$ as the intergenic bias of that organism.

Least-squares techniques (29) were used to model the average usage of eigencodon $\boldsymbol{v_2}$ in all genes in genome g, denoted $\bar{u}_2^g$, as a function of $\boldsymbol{d_g}$, the intergenic bias of genome g. We let $\boldsymbol{f} = (\bar{u}_2^g)$ be a vector with G components (one for each genome), and let $\bar{f}$ be the mean of the components of $\boldsymbol{f}$. We further let $D = [\boldsymbol{d_1} \ldots \boldsymbol{d_G}]^T$ be a matrix with intergenic bias parameters as its rows and let $\tilde{D}$ be a version of D with every column centered about zero. By using a thin SVD, we decomposed $\tilde{D}$ into $\tilde{D} = YTX^T$, where $Y^TY = I, X^TX = I$, and $T = \mathrm{diag}(t_1, \ldots, t_{64})$, where $t_1 \geq t_2 \geq t_3 \geq \cdots \geq t_{64} \geq 0$. The matrix Y and the singular values t_j are given in Table 4, which is published as supporting information on the PNAS web site. Because $\tilde{D}$ has rank 48, $t_{49} = \cdots = t_{64} = 0$, and the first 48 columns of Y form an orthogonal basis for the range of $\tilde{D}$. In a least-squares model, $\boldsymbol{f}$ is approximated as $\bar{f} + \Sigma_{i=1}^{48} \boldsymbol{f}^T\boldsymbol{y_i}\boldsymbol{y_i}$. The larger the $(\boldsymbol{f}^T\boldsymbol{y_i})^2$, the greater the amount of variance in $\boldsymbol{f}$ that can be explained by $\boldsymbol{y_i}$. In our case, the $\boldsymbol{y_i}$ corresponding to larger singular values, in general, explained more variance than those corresponding to smaller singular values, with $\boldsymbol{y_2}$, $\boldsymbol{y_3}$, and $\boldsymbol{y_8}$ being most critical to the model. To avoid overfitting, we model $\boldsymbol{f}$ using $\boldsymbol{y_i}$ for $i = 1, \ldots, 8$.

To determine the quality of the resulting fit, we tested the ability of randomized versions of D to explain $\boldsymbol{f}$ by using models of the same complexity. Specifically, we permuted the entries of D and renormalized the values so that each row satisfied the constraints of a set of intergenic bias parameters. Then D was centered as before and a least-squares fit to $\boldsymbol{f}$ was generated by using the directions corresponding to the eight largest singular values of the centered matrix. The randomization and fitting procedure was repeated 10,000 times. For both the real and randomized data, the quality of the fit was taken to be the fraction of the variance in $\boldsymbol{f}$ explained by the model (the R^2 statistic).

Results

Codon Bias Varies Between Genomes Primarily Along Two Dimensions. Using a SVD as described above, we decomposed the space of possible codon vectors into 41 orthogonal directions $\{\boldsymbol{v_1}, \ldots, \boldsymbol{v_{41}}\}$, referred to as eigencodons. The eigencodons are ordered so that gene to gene usage varies most in the direction of eigencodon $\boldsymbol{v_1}$ and least in the $\boldsymbol{v_{41}}$ direction. Every codon vector can be represented uniquely as a linear combination of the 41 eigencodons. The fraction of var(u_j) due to between-genome variance [var(u_j)$^{\mathrm{between}}$] is plotted for each eigencodon, $\boldsymbol{v_j}$, in Fig. 1*b*. For $j = 1$ and 2, between-genome variance accounts for the majority of var(u_j) (90% and 62%); for all other $\boldsymbol{v_j}$ values, between-genome variance accounts for much less of var(u_j) (5–32%). This means that var(u_1^g) and var(u_2^g), the within-genome variances, are relatively small for all genomes; thus, for most genes, i, in a given genome, g, $u_{i,1}^g$ and $u_{i,2}^g$ are close to $\bar{u}_1^g$ and $\bar{u}_2^g$, and $\bar{u}_1^g$ and $\bar{u}_2^g$ tend to differ for different genomes. In other words, $\bar{u}_1^g$ and $\bar{u}_2^g$ are characteristic values for each genome g because var(u_1^g) and var(u_2^g) are small for each genome g. On the other hand, usage of all other eigencodons, $\boldsymbol{v_j}$, for $j = 3, \ldots, 41$, varies little between genomes compared with its variance among the genes within a genome. Therefore, differences in codon bias between genomes can be reasonably modeled by using only two parameters, average usage of $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ ($\bar{u}_1^g$ and $\bar{u}_2^g$, respectively). Inclusion of additional eigencodons adds little discriminatory power.

Genome GC Content Correlates with $\bar{u}_1^g$. Each component of each eigencodon represents a codon, $m(w)$. Simple inspection of the components of $\boldsymbol{v_1}$ suggests that $\boldsymbol{v_1}$ is related to gene GC content. Nearly all codons ending in G or C contribute positively to $\boldsymbol{v_1}$ (positive $v_{m(w),1}$) and those ending in A or T contribute negatively (Table 2). Plotting $\bar{u}_1^g$ versus genome GC content (Fig. 2*a*) shows a strong positive correlation ($R^2 = 0.961$). This correlation also holds for individual genes; plotting $u_{i,1}$ versus gene GC

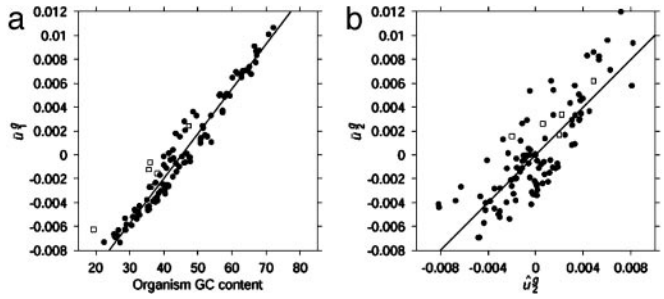

Fig. 2. (*a*) Plot of $\bar{u}_1^g$ versus genome GC content for each organism. Usage of the first eigencodon correlates with genome GC content ($R^2 = 0.961$). (*b*) Plot of $\bar{u}_2^g$ versus intergenic bias. The second eigencodon correlates with a model constructed as a linear combination of intergenic bias parameters ($R^2 = 0.669$). In both plots, open boxes are data points for *A. thaliana*, *C. elegans*, *E. cuniculi*, *P. falciparum*, *S. cerevisiae*, and *S. pombe*.

content (data not shown) gives a squared correlation coefficient of $R^2 = 0.895$.

Intergenic Context-Dependent Nucleotide Biases Correlate with $\bar{u}_2^g$. We created a 64-parameter model, $\boldsymbol{d_g}$ (referred to as intergenic bias), from each genome's intergenic regions, which describes nucleotide biases that depend on the identity of immediately adjacent bases (see *Materials and Methods*). The nearest-neighbor nucleotide biases found in the intergenic regions are in all cases (except those where the context specifies a stop codon) positively correlated with biases in the third codon position of genes from the same organism, when some constraints of the genetic code are corrected for by fixing the first and second codon positions and the first codon position of the following codon (data not shown). This finding is not surprising because it has been shown that dinucleotide biases influence codon bias in several organisms (30, 31). To quantify whether differences in $\bar{u}_2^g$ could be explained by intergenic bias, we first constructed a matrix, $D = [\boldsymbol{d_1} \ldots \boldsymbol{d_G}]^T$, whose rows were the intergenic bias parameters $\boldsymbol{d_g}$. We used a SVD to find the directions of largest variance in D and then used least-squares techniques to model $\bar{u}_2^g$ as a linear combination of the eight directions of largest variance (the approximation of $\bar{u}_2^g$ is denoted $\hat{u}_2^g$). As shown in Fig. 2*b*, the resulting model (referred to as the $\hat{u}_2$ model) explained 66.9% of the variance in $\bar{u}_2^g$. As a control, we also estimated $\bar{u}_2^g$ by using 10,000 randomized versions of D. In no case did the resulting models explain >30.3% of the variance in $\bar{u}_2^g$ (see *Materials and Methods* for details).

Eukaryotic Genomes Have Characteristic Values for $\bar{u}_2^g$. Several eukaryotic species, namely *Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster*, and *Saccharomyces cerevisiae*, have been noted to have a "prokaryotic-like" pattern of codon bias (7, 32, 33), in that they obey the genome hypothesis. Others, such as humans and other mammals, do not; GC content varies greatly between regions of the mammalian genome, which are termed isochores (34). Because GC content influences codon bias (35), genes from different isochores have distinct patterns of codon bias.

As expected, when we expressed the codon bias of *A. thaliana, C. elegans, D. melanogaster*, and *S. cerevisiae* genes in terms of the eigencodon basis generated from the SVD of prokaryotic codon vectors, we found that the var(u_j^g) for all eigencodons, $\boldsymbol{v_j}$, is similar to that in prokaryotic organisms. Namely, var(u_1^g) and var(u_2^g) are small, whereas var(u_j^g) for $j = 3, \ldots, 41$ are large (Fig. 5, which is published as supporting information on the PNAS web site, and Fig. 3). Expressing the codon bias of *Danio rerio, Encephalitozoon cuniculi, Plasmodium falciparum*, and *Schizosaccharomyces pombe* genes in terms of eigencodons also

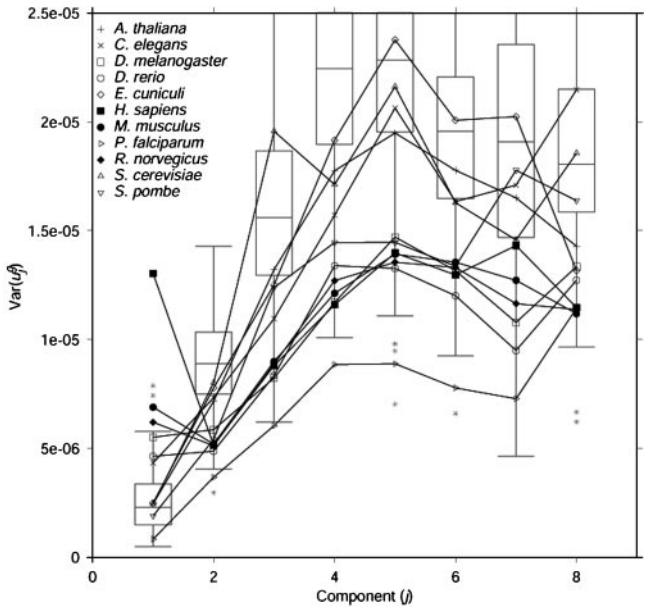

Fig. 3. Eukaryotic genomes have low variance in usage of the second eigencodon. Expanded view of box and whisker plots of var(u_j^g) for $j = 1, \ldots, 8$ for all prokaryotic genomes *g*, with values for eukaryotic genomes superimposed. A full diagram can be found in Fig. 5. Box and whisker plots are drawn in gray. Asterisks indicate outlying prokaryotic values. Values for eukaryotic organisms are drawn individually with symbols as indicated in the upper left corner. Compared with prokaryotic genomes, many eukaryotic genomes have large variance in the usage of eigencodon $\boldsymbol{v_1}$ but relatively small variance in usage of eigencodon $\boldsymbol{v_2}$. In general, variance is smaller for eukaryotic genomes than for prokaryotic genomes because eukaryotic genes tend to be longer than prokaryotic genes and hence provide less noisy samples of codon bias. Considering only long prokaryotic genes does not change the results qualitatively (see Figs. 7–9, which are published as supporting information on the PNAS web site).

produced the same pattern. Therefore, $\bar{u}_1^g$ and $\bar{u}_2^g$ are also characteristic values in these eukaryotic organisms.

Human genes, on the other hand, have high var($u_1^{H.\ sapiens}$) (i.e., they vary greatly in usage of $\boldsymbol{v_1}$), as expected from the differences in GC content between isochores. Rat and mouse genes also have somewhat high var(u_1^g), although much smaller than var($u_1^{H.\ sapiens}$). This finding is consistent with the observation that isochores in humans vary more in GC content than those in rodents (36). Interestingly, var(u_2^g) was small for all three mammals examined, similar to var(u_2^g) for prokaryotic genomes. Therefore, although $\bar{u}_1^g$ is not characteristic of all mammalian genomes, $\bar{u}_2^g$ is still characteristic of each of these three mammalian genomes.

Codon Usage in Prokaryotes Can Be Estimated from Intergenic Sequences. For any given genome, *g*, we define genome-wide codon bias ($\bar{\boldsymbol{c}}^{\boldsymbol{g}}$) as the mean codon vector for all the genes in *g* (not just the subset used in the SVD). Given the intergenic sequences of any prokaryote, we estimate that organism's genome-wide codon bias in the following manner. First, we calculate the GC content of the intergenic sequences, which is highly correlated with the overall GC content (35) and therefore $\bar{u}_1^g$, allowing us to estimate $\bar{u}_1^g$ (denoted $\hat{u}_1^g$) by using the following equation: $\hat{u}_1^g = 0.000359 \cdot (\text{intergenic GC content}) - 0.0143$, where GC content is measured in percent. Also from the intergenic sequences, we can calculate intergenic bias parameters, $\boldsymbol{d_g}$. From $\boldsymbol{d_g}$ and the $\hat{u}_2$ model, we can compute $\hat{u}_2^g$, an estimate for $\bar{u}_2^g$. Then, predicted genome-wide codon bias can be approximated as $\hat{\boldsymbol{c}}^{\boldsymbol{g}} = \hat{u}_1^g \boldsymbol{v_1} + \hat{u}_2^g \boldsymbol{v_2} + \bar{\boldsymbol{c}}$, where $\bar{\boldsymbol{c}}$ is the average codon vector for the genes used in the SVD. This method predicts the genome-wide codon bias

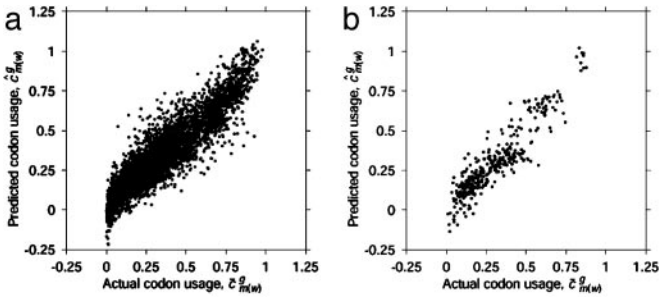

Fig. 4. Graph of components of predicted genome-wide codon bias vector, $\hat{\boldsymbol{c}}^{\boldsymbol{g}}$, based on intergenic nucleotide sequences versus components of actual genome-wide codon bias vector, $\bar{\boldsymbol{c}}^{\boldsymbol{g}}$. Each point in the plot represents a ($\bar{c}^g_{m(w)}$, $\hat{c}^g_{m(w)}$) coordinate pair for some organism g and some codon $m(w)$. $\bar{c}^g_{m(w)}$ is a component of $\bar{\boldsymbol{c}}^{\boldsymbol{g}}$, and $\hat{c}^g_{m(w)}$ is a component of $\hat{\boldsymbol{c}}^{\boldsymbol{g}}$. Different organisms and codons are not differentiated in these plots. Stop codons (TAA, TAG, and TGA) and the single codons for methionine (ATG) and tryptophan (TGG) were excluded. (*a*) Prokaryotes. Overall $R^2 = 0.858$. Average for individual genomes is $R^2 = 0.840$. (*b*) Data for the following eukaryotes: *A. thaliana, C. elegans, E. cuniculi, P. falciparum, S. cerevisiae*, and *S. pombe*. Overall $R^2 = 0.847$. R^2 values for the individual genomes are given in the text.

of any individual prokaryotic genome with an average R^2 of 0.840. As shown in Fig. 4*a*, the components of $\bar{\boldsymbol{c}}^{\boldsymbol{g}}$ correlate with the corresponding components of $\hat{\boldsymbol{c}}^{\boldsymbol{g}}$; in other words, the average usage of individual codons among all genes within a genome correlates very well with the usage predicted based on intergenic sequence statistics.

Prokaryotic Parameters Can Be Used to Effectively Predict Eukaryotic Genome-Wide Codon Bias from Eukaryotic Intergenic Sequences. We then tested whether $\bar{u}^g_1$ and $\bar{u}^g_2$ for eukaryotic organisms correlated with organism GC content and intergenic bias, respectively. Not surprisingly, as shown by the open boxes in Fig. 2*a*, GC content for *A. thaliana, C. elegans, E. cuniculi, P. falciparum, S. cerevisiae*, and *S. pombe* has a correlation with $\bar{u}^g_1$ similar to what it has for prokaryotic organisms. As shown by the open boxes in Fig. 2*b*, using the $\hat{u}_2$ model with eukaryotic intergenic bias parameters results in values close to the regression line for prokaryotic data. Based on these results, we can also predict genome-wide codon bias quite well in these eukaryotes based only on their intergenic sequences by using the relationships between intergenic GC content and $\bar{u}^g_1$ and $\bar{u}^g_2$ calculated from prokaryotic sequences (Fig. 4*b*). The squared correlation coefficients (R^2) for the individual organisms were the following: *A. thaliana*, 0.789; *C. elegans*, 0.753; *E. cuniculi*, 0.717; *P. falciparum*, 0.932; *S. cerevisiae*, 0.892; and *S. pombe*, 0.915.

Discussion

Using a SVD, we defined 41 eigencodons, $\boldsymbol{v_1}, \ldots, \boldsymbol{v_{41}}$. Linear combinations of these eigencodons can completely describe the codon bias of any gene. By quantitatively decomposing variation in codon bias into a term for variation within genomes and a term for variation between genomes, we show that between-genome variation accounts for most of the variation in the usage of only two of the 41 eigencodons, $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$. In other words, genes from the same genome are typically more similar to each other in their usage of $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ than genes from different genomes. Genes from each prokaryotic organism, g, examined thus use characteristic amounts of $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$, denoted $\bar{u}^g_1$ and $\bar{u}^g_2$, respectively. Because (*i*) codon bias varies more in the direction of $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ than in any other direction, and (*ii*) between-genome variation [$\mathrm{var}(u_j)^{\mathrm{between}}$] is large in these two directions, $\bar{u}^g_1$ and $\bar{u}^g_2$ (usage of $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$) are the most important (i.e., most necessary) parameters with respect to our eigencodon basis for describing codon bias in archaeal and eubacterial organisms. In addition, because (*i*) codon bias varies less in the directions of $\boldsymbol{v_3}, \ldots, \boldsymbol{v_{41}}$ (i.e., $\sigma_3, \ldots, \sigma_{41}$ are small) and (*ii*) between-genome variation in $\bar{u}^g_j$ for $j = 3, \ldots, 41$, is small, $\bar{u}^g_1$ and $\bar{u}^g_2$ are also largely sufficient for describing genome-wide codon bias. The limitation to only two parameters is somewhat surprising because differences between genomes would be expected to be caused by many global differences between organisms, such as in the replication and transcription machineries, repair systems, and physical and chemical environments.

Genome GC content, which is determined by directional mutation pressure (ref. 37; although see ref. 38), correlates closely with $\bar{u}^g_1$; thus, $\bar{u}^g_1$ is likely also specified by directional mutation pressure. In this article, we show that $\bar{u}^g_2$ is also determined by mutational pressures acting throughout the genome. Referred to as intergenic bias, the mutational pressures correlated with $\bar{u}^g_2$ depend on adjacent (nearest-neighbor) nucleotide context. Because $\bar{u}^g_1$ and $\bar{u}^g_2$ are determined by mutational pressures, they may be predicted from parameters calculated from intergenic sequences. Intergenic sequences can therefore be used in a two-parameter model to predict genome-wide codon bias in eubacterial and archaeal genomes. This two-parameter model is also accurate in predicting codon bias from intergenic sequences in most eukaryotic genomes, confirming that codon bias is largely determined by mutational processes unrelated to protein expression in all three domains of life.

Two observations argue that mutational processes are mostly responsible for differences in codon bias between genomes in the direction of $\boldsymbol{v_2}$. First, usage of $\boldsymbol{v_2}$ varies little among genes within the same genome regardless of effective population size. The effective population size of *Escherichia coli* has been estimated at 10^8 to 10^9 (39), whereas that for mammals such as humans is $\sim 10^4$ (40). Estimates for the difference in selection coefficients for different synonymous codons range from 10^{-9} (39) to 10^{-5} (41). Despite the large range, all estimates are consistent with the notion that selection on synonymous codons may be operative in *E. coli* ($N_e s \geq 1$, i.e., effective population size times selective coefficient is large) but not in *Homo sapiens* ($N_e s \leq 0.1$) (5, 7, 42). However, although population size varies over more than four orders of magnitude, $\mathrm{var}(u^g_2)$ for all genomes studied varies only over a five-fold range. Because *H. sapiens* has a small effective population size compared with *E. coli*, one would expect, given the same difference in selective coefficients for different synonymous codons, that selectively maintained codon usage would vary more in *H. sapiens.* In fact, $\mathrm{var}(u_2^{H.\ sapiens})$ is actually half the value of $\mathrm{var}(u_2^{E.\ coli})$. Thus, the small value of $\mathrm{var}(u^g_2)$ for all organisms is difficult to explain by using selection. Second, intergenic nucleotide biases explain more than two-thirds of the variation in genome-wide average usage of $\boldsymbol{v_2}$ in all archaeal, all eubacterial, and most eukaryotic organisms examined. Thus, most of the variation in $\bar{u}^g_2$ can be explained by properties of sequences in the same genome that are never translated. In principle, some selective process acting during protein expression may cause these correlated changes in intergenic sequences; however, a simpler explanation is that mutational processes affect all the DNA within a given organism. These mutational processes result in correlations between intergenic sequence nucleotide biases and codon bias. Taking GC content as a special case of nucleotide biases (where adjacent nucleotide context is ignored), the preceding statement reduces to the statement that directional mutation pressure influences codon bias by causing qualitatively similar changes in GC content in all DNA within a given organism.

It might be that selection maintains the small value of $\mathrm{var}(u^g_2)$ in *E. coli* and other bacteria, whereas mutation maintains it in *H. sapiens* and other mammals. More generally, it might be that different mechanisms besides mutation are responsible for maintaining the small value of $\mathrm{var}(u^g_2)$ in different organisms. This

possibility cannot be completely excluded. However, because of the good correlation of $\bar{u}_2^g$ with parameters calculated from intergenic sequences and because usage of v_2 measures usage of all codons to some extent, we prefer the simpler explanation that mutation is the primary force maintaining small values of $\bar{u}_2^g$ in all organisms.

Usage of v_2 is also correlated with organism optimal growth temperature (Fig. 6, which is published as supporting information on the PNAS web site). Organisms with higher optimal growth temperature tend to have higher values of $\bar{u}_2^g$. This result is in agreement with the results of others who note that the second factor in a principal-components analysis or correspondence analysis correlates with the organism's optimal growth temperature (24, 43). The results of ref. 24 demonstrate that selection related to elevated growth temperature plays a role in establishing codon bias in thermophilic organisms, which may be related to the tendency for thermophilic organisms to systematically load RNA sequences with purines (25). However, our results emphasize the importance of mutational (not related to protein expression) forces in determining global trends in codon bias. Specifically, because selection on codon bias or mRNA structure during protein expression cannot explain the correlation we observe with patterns of nearest-neighbor nucleotide bias in intergenic sequences, we conclude that mutational pressures are primarily responsible for the differences in usage of v_2 between genomes, as discussed above. The role of selection is instead appropriately ascribed to generating the relatively smaller variation in usage of v_2 among highly expressed ribosomal genes and other genes within the same genome (24) and to a minor role in determining overall genome-wide codon bias.

In agreement with our interpretation that mutation is primarily responsible for $\bar{u}_2^g$, other studies of the effect of high growth temperature on DNA sequences also point toward a mutational effect on DNA sequences in general and codon bias in particular. A linear combination of dinucleotide abundances calculated over entire genomes correlates well with optimal growth temperature for one mesophilic and several thermophilic archaeal organisms (30). The same result is obtained when coding sequences and intergenic sequences are analyzed separately. Because context-dependent nucleotide biases also influence codon bias (30, 31), one would therefore expect growth temperature to correlate with a mutational effect on codon bias. Furthermore, recent work has shown that patterns of codon bias across many different organisms, both thermophilic and mesophilic, can be explained by a single mutational model dependent on position-specific nucleotide parameters (23). Finally, elevated temperatures result in markedly elevated rates of DNA damage (44), but GC content does not correlate with optimal growth temperature (30), suggesting a role for biases that are not captured by simple GC content.

Based on our results and those of others, we propose the following interpretation of observed codon bias and the genome hypothesis (in the special case of mammals, the following general statements may not apply to codon bias changes related to isochores). The genome-wide codon bias of each organism is set primarily by mutational forces, which create a point about which the codon bias of individual genes in that organism are clustered. The codon bias of individual genes or subsets of genes is additionally perturbed from the genome-wide average codon bias by selective and other mutational forces acting during translation, but this effect is relatively much smaller. Therefore, in all three domains of life, the "systems of codon usage" referred to by Grantham (which we have called genome-wide codon bias) are coarsely set by mutational pressures and precisely modified by selective pressures.

This work was supported by National Institutes of Health Grants 2T32GM07365 to the Medical Scientist Training Program (to S.L.C.), GM51426 (to S.L.C. and L.S.), and HG00044 (to A.K.H.), Department of Energy Grant DE-FG03-01ER63219 (to W.L., A.K.H., L.S., and H.H.M.), and Defense Advanced Research Projects Agency Defense Sciences Office Grant MDA972-00-1-0032 (to W.L., L.S., and H.H.M.).

1. Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. (1992) *Microbiol. Rev.* **56,** 229–264.
2. Grantham, R. (1980) *Trends Biochem. Sci.* **5,** 327–331.
3. Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pave, A. (1980) *Nucleic Acids Res.* **8,** r49–r62.
4. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. (1981) *Nucleic Acids Res.* **9,** r43–r74.
5. Ikemura, T. (1985) *Mol. Biol. Evol.* **2,** 13–34.
6. Sharp, P. M. & Li, W. H. (1987) *Mol. Biol. Evol.* **4,** 222–230.
7. Sharp, P. M., Stenico, M., Peden, J. F. & Lloyd, A. T. (1993) *Biochem. Soc. Trans.* **21,** 835–841.
8. Doolittle, W. F. (1998) *Trends Genet.* **14,** 307–311.
9. Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000) *Nature* **405,** 299–304.
10. Woese, C. R. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8742–8747.
11. Gouy, M. & Gautier, C. (1982) *Nucleic Acids Res.* **10,** 7055–7074.
12. D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C. & Bernardi, G. (1991) *J. Mol. Evol.* **32,** 504–510.
13. Collins, D. W. & Jukes, T. H. (1993) *J. Mol. Evol.* **36,** 201–213.
14. Lobry, J. R. & Gautier, C. (1994) *Nucleic Acids Res.* **22,** 3174–3180.
15. D'Onofrio, G., Jabbari, K., Musto, H. & Bernardi, G. (1999) *Gene* **238,** 3–14.
16. Akashi, H. (1994) *Genetics* **136,** 927–935.
17. Eyre-Walker, A. (1996) *Mol. Biol. Evol.* **13,** 864–872.
18. Hasegawa, M., Yasunaga, T. & Miyata, T. (1979) *Nucleic Acids Res.* **7,** 2073–2079.
19. Gambari, R., Nastruzzi, C. & Barbieri, R. (1990) *Biomed. Biochim. Acta* **49,** S88–S93.
20. Huynen, M. A., Konings, D. A. & Hogeweg, P. (1992) *J. Mol. Evol.* **34,** 280–291.
21. Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. (2003) *Nature* **422,** 633–637.
22. Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., Mori, H. & Ikemura, T. (2001) *Gene* **276,** 89–99.
23. Knight, R. D., Freeland, S. J. & Landweber, L. F. (2001) *Genome Biol.* **2,** RESEARCH0010.
24. Lynn, D. J., Singer, G. A. & Hickey, D. A. (2002) *Nucleic Acids Res.* **30,** 4272–4277.
25. Lao, P. J. & Forsdyke, D. R. (2000) *Genome Res.* **10,** 228–236.
26. Lee, W. & Chen, S. L. (2002) *Biotechniques* **33,** 1334–1341.
27. Vieille, C. & Zeikus, G. J. (2001) *Microbiol. Mol. Biol. Rev.* **65,** 1–43.
28. Golub, G. H. & Van Loan, C. F. (1996) in *Matrix Computations* (Johns Hopkins Univ. Press, Baltimore), 3rd Ed., pp. 48–86.
29. Weisberg, S. (1985) *Applied Linear Regression*, Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics (Wiley, New York), 2nd Ed.
30. Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., *et al.* (2000) *Proc. Natl. Acad. Sci. USA* **97,** 14257–14262.
31. Karlin, S., Campbell, A. M. & Mrazek, J. (1998) *Annu. Rev. Genet.* **32,** 185–225.
32. Moriyama, E. N. & Powell, J. R. (1997) *J. Mol. Evol.* **45,** 514–523.
33. Duret, L. & Mouchiroud, D. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 4482–4487.
34. Bernardi, G. (2000) *Gene* **241,** 3–17.
35. Muto, A. & Osawa, S. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 166–169.
36. Mouchiroud, D., Gautier, C. & Bernardi, G. (1988) *J. Mol. Evol.* **27,** 311–320.
37. Sueoka, N. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2653–2657.
38. Naya, H., Romero, H., Zavala, A., Alvarez, B. & Musto, H. (2002) *J. Mol. Evol.* **55,** 260–264.
39. Hartl, D. L., Moriyama, E. N. & Sawyer, S. A. (1994) *Genetics* **138,** 227–234.
40. Shields, D. C., Sharp, P. M., Higgins, D. G. & Wright, F. (1988) *Mol. Biol. Evol.* **5,** 704–716.
41. Bulmer, M. (1991) *Genetics* **129,** 897–907.
42. Bulmer, M. (1987) *Nature* **325,** 728–730.
43. Lobry, J. R. & Chessel, D. (2003) *J. Appl. Genet.* **44,** 235–261.
44. Lindahl, T. (1993) *Nature* **362,** 709–715.