Published online September 24, 2004

5036–5044 *Nucleic Acids Research, 2004, Vol. 32, No. 17*
doi:10.1093/nar/gkh834

Solving the riddle of codon usage preferences: a test for translational selection

Mario dos Reis*, Renos Savva and Lorenz Wernisch

School of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Received June 2, 2004; Revised and Accepted August 31, 2004

ABSTRACT

Translational selection is responsible for the unequal usage of synonymous codons in protein coding genes in a wide variety of organisms. It is one of the most subtle and pervasive forces of molecular evolution, yet, establishing the underlying causes for its idiosyncratic behaviour across living kingdoms has proven elusive to researchers over the past 20 years. In this study, a statistical model for measuring translational selection in any given genome is developed, and the test is applied to 126 fully sequenced genomes, ranging from archaea to eukaryotes. It is shown that tRNA gene redundancy and genome size are interacting forces that ultimately determine the action of translational selection, and that an optimal genome size exists for which this kind of selection is maximal. Accordingly, genome size also presents upper and lower boundaries beyond which selection on codon usage is not possible. We propose a model where the coevolution of genome size and tRNA genes explains the observed patterns in translational selection in all living organisms. This model finally unifies our understanding of codon usage across prokaryotes and eukaryotes. *Helicobacter pylori*, *Saccharomyces cerevisiae* and *Homo sapiens* are codon usage paradigms that can be better understood under the proposed model.

INTRODUCTION

The controversial ideas of Kimura (1) and of King and Jukes (2) on neutral evolution led some early detractors to postulate that usage of synonymous codons in protein coding genes is not necessarily random and that codon composition could be biased towards codons that would match the tRNA pool of the host organism (3). This prediction was partially confirmed when the first genes were sequenced, in particular for the ribosomal genes of *Escherichia coli* (4). Soon after, Grantham *et al*. (5,6) compiled codon usage tables for all sequenced genes at that time, and proposed that each genome has a particular codon usage signature that reflects particular evolutionary forces acting within that genome. This genome hypothesis was soon adopted, and extensive studies were performed in yeast and other microorganisms (7,8) that confirmed this view. However, it was not until Ikemura (9,10) performed his elegant experiments on *E.coli* that the hypothesis of codon adaptation to the tRNA pool was confirmed, giving a plausible explanation for the presence of codon usage bias in highly expressed genes and leading to rapid development in this area of research. These early discoveries led to the formulation of a yeast–*E.coli* paradigm, establishing that highly expressed genes use a subset of optimal codons in accordance with their respective major isoacceptor tRNA levels. The evolutionary force responsible for this was coined *translational selection*.

However, when attempts were made to extend the findings from unicellular microorganisms to higher eukaryotes, the situation became confusing. Organisms such as humans seem to have their codon usage determined solely by genomic GC content or isochore composition, while others such as the fly or worm, seem to present an intermediate degree of selection partly determining their codon usage. A mutation–selection balance theory of synonymous codon usage was developed to explain these observations (11,12,13).

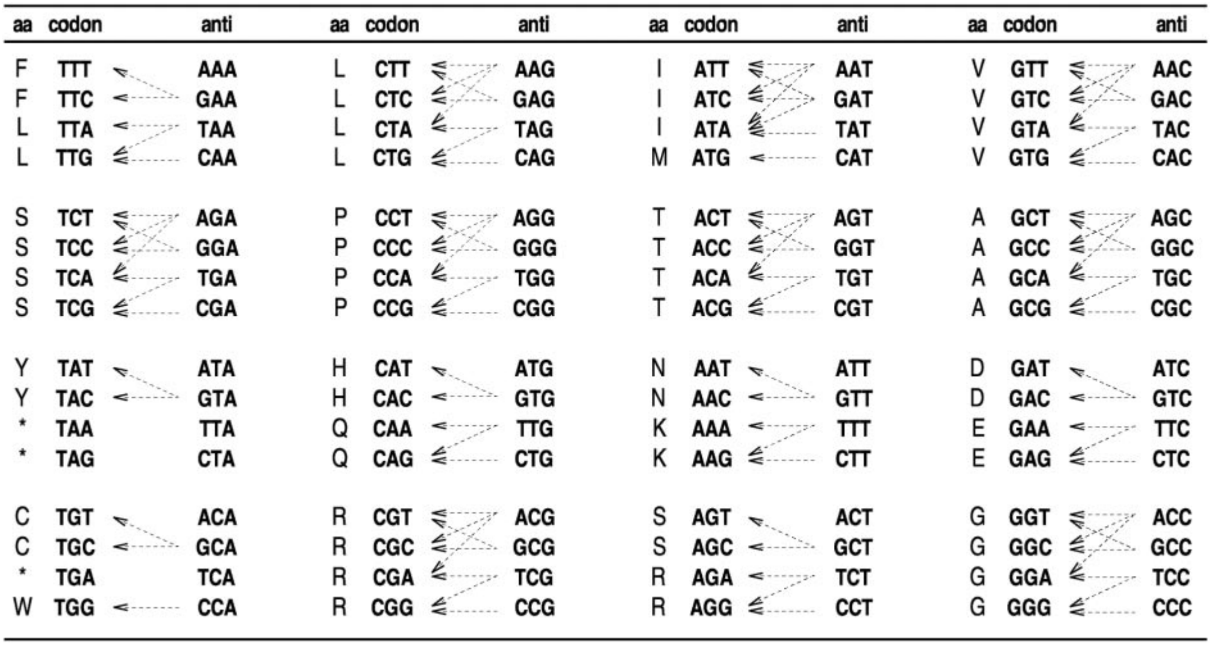
Although there have been detailed studies on particular organisms and attempts to find trends in larger groups (14,15,16,17), those studies have failed to explain the reasons for the disparity in the action of selection on codon usage across different genomes, giving rise to the notion that the study of codon usage is one of the most controversial areas of molecular evolution (18). After all, why do certain organisms such as *Helicobacter pylori* or humans (19,20) present no evidence of translational selection while others such as *E.coli* or worms show a marked codon bias due to selection (21,10)? The problem is aggravated since there seems to be no formal measure available for estimating translational selection on a whole genome basis. Such a measure would help in the investigation of the factors shaping translational selection and the trends underlying its idiosyncratic behaviour across living kingdoms could be better understood. The 20-year-old puzzle of the high diversity in codon usage across genomes might then be settled. In this paper, a statistical test for translational selection is developed and applied to a large subset of fully sequenced genomes. As a result, a unified framework is presented for the understanding of translational selection and codon usage trends in prokaryotic and eukaryotic genomes.

MATERIALS AND METHODS

The tRNA adaptation index

The tRNA adaptation index (tAI) (22) is a measure of the tRNA usage by coding sequences inspired by the codon adaptation

*To whom correspondence should be addressed. Tel: +44 20 7631 6869; Fax: +44 20 7631 6803; Email: m.reis@mail.cryst.bbk.ac.uk



aa	codon	anti	aa	codon	anti	aa	codon	anti	aa	codon	anti
F	TTT	AAA	L	CTT	AAG	I	ATT	AAT	V	GTT	AAC
F	TTC	GAA	L	CTC	GAG	I	ATC	GAT	V	GTC	GAC
L	TTA	TAA	L	CTA	TAG	I	ATA	TAT	V	GTA	TAC
L	TTG	CAA	L	CTG	CAG	M	ATG	CAT	V	GTG	CAC
S	TCT	AGA	P	CCT	AGG	T	ACT	AGT	A	GCT	AGC
S	TCC	GGA	P	CCC	GGG	T	ACC	GGT	A	GCC	GGC
S	TCA	TGA	P	CCA	TGG	T	ACA	TGT	A	GCA	TGC
S	TCG	CGA	P	CCG	CGG	T	ACG	CGT	A	GCG	CGC
Y	TAT	ATA	H	CAT	ATG	N	AAT	ATT	D	GAT	ATC
Y	TAC	GTA	H	CAC	GTG	N	AAC	GTT	D	GAC	GTC
*	TAA	TTA	Q	CAA	TTG	K	AAA	TTT	E	GAA	TTC
*	TAG	CTA	Q	CAG	CTG	K	AAG	CTT	E	GAG	CTC
C	TGT	ACA	R	CGT	ACG	S	AGT	ACT	G	GGT	ACC
C	TGC	GCA	R	CGC	GCG	S	AGC	GCT	G	GGC	GCC
*	TGA	TCA	R	CGA	TCG	R	AGA	TCT	G	GGA	TCC
W	TGG	CCA	R	CGG	CCG	R	AGG	CCT	G	GGG	CCC

Figure 1. Genetic code and general codon–anticodon recognition rules for tRNA genes. This table simply summarizes all the theoretically possible interactions between the coding codons and the extant tRNA sequences in the organisms analysed in this work. The interested reader is advised to refer to the literature (47,46) for a detailed description of codon–anticodon pairings.

index of Sharp and Li (23). In order to develop this index, we take advantage of the fact that the tRNA gene copy number across some genomes has a high and positive correlation with tRNA abundance within the cell (9,24,14,25) and with codon preferences in such genomes. Since tRNA abundance may be thought of as the driving force for translational selection, it is reasonable to speculate that measuring the tRNA usage of a gene may provide an indirect way for detecting translational selection according to how well the gene in question is adapted to the tRNA gene pool. In order to calculate this index, the absolute adaptiveness value W_i for each codon i is defined as

$$W_i = \sum_{j=1}^{n_i} \left(1 - s_{ij}\right) \mathrm{tGCN}_{ij}, \qquad 1$$

where n_i is the number of tRNA isoacceptors that recognize the ith codon, tGCN_{ij} is the gene copy number of the jth tRNA that recognizes the ith codon, and s_{ij} is a selective constraint on the efficiency of the codon–anticodon coupling. To build a table of W_i values, it is best to sort the codons as shown in Figure 1 and to analyse the way in which each tRNA recognizes its particular codons. It can be seen that the 64 codons that comprise the genetic code can be clustered into groups of four elements, which reflect the natural way in which tRNAs recognize them. Based on Figure 1, a simple set of formulae for calculating all W_i values can easily be drawn taking into account Crick's wobble rules (26) for codon–anticodon pairing (Table 1). From the W_i values the *relative adaptiveness value* w_i of a codon is obtained as

$$w_i = \begin{cases} W_i/W_{\max} & \text{if } W_i \neq 0 \\ w_{\text{mean}} & \text{else} \end{cases}, \qquad 2$$

Table 1. Formulas for calculating *W*s according to Crick's wobble rules (26)

n	Anticodon	Codon	*W*
i	INN	NNU	$(1 - s_{I:U})\mathrm{tGCN}_i + (1 - s_{G:U})\mathrm{tGCN}_{i+1}$
$i + 1$	GNN	NNC	$(1 - s_{G:C})\mathrm{tGCN}_{i+1} + (1 - s_{I:C})\mathrm{tGCN}_i$
$i + 2$	UNN	NNA	$(1 - s_{U:A})\mathrm{tGCN}_{i+2} + (1 - s_{I:A})\mathrm{tGCN}_i$
$i + 3$	CNN	NNG	$(1 - s_{C:G})\mathrm{tGCN}_{i+3} + (1 - s_{U:G})\mathrm{tGCN}_{\mathrm{i+2}}$

I, inosine. The interested reader should refer to the literature (46,47) for a detailed description of nucleoside modifications and codon–anticodon pairings.

where $W_{\max}$ is the maximum W_i value and w_{mean} is the geometric mean of all w_i with $W_i \neq 0$. The *tRNA adaptation index* tAI_g of a gene g is defined as the geometric mean of the relative adaptiveness values of its codons

$$\mathrm{tAI}_g = \left(\prod_{k=1}^{l_g} w_{i_{kg}} \right)^{1/l_g}, \qquad 3$$

where i_{kg} is the codon defined by the kth triplet in gene g and l_g is the length of the gene in codons (except the stop codon). Consequently, tAI_g estimates the amount of adaptation of a gene g to its genomic tRNA pool.

The relationship of tAI to Nc

The effective number of codons (Nc) is a measure that quantifies the departure of a gene from the random usage of synonymous codons (27). Nc is related to the amount of entropy in the codon usage of a sequence. It reaches is maximal value (61) when all codons are used equally and its minimal value (20) when only one codon is used per amino acid. Since the effect of selection is a reduction of the entropy of codon usage in a sequence, Nc provides a reliable way of testing this effect. Since the silent GC content x_g of a gene g affects its Nc value, an equation to approximate this relationship under the hypothesis of no selection was proposed by Wright (27):

$$\mathrm{Nc}_g = f\left(x_g\right), \qquad 4$$

where

$$f\left(x_g\right) = 2 + x_g + \frac{29}{x_g^2 + \left(1 - x_g\right)^2}. \qquad 5$$

This model can easily be extended to account for the selection effect φ_g on the codon usage of gene g and for uncontrollable, random factors ε_g

$$\mathrm{Nc}_g = f\left(x_g\right) - \varphi_g + \varepsilon_g, \qquad 6$$

since the effect of selection is to reduce the value of Nc, it has been assigned a negative value in Equation 6. The random element ε_g simply represents sources of variation on Nc that cannot be accounted for by selection or silent GC content alone, and that cannot be controlled in this study. Since in practical terms we can only calculate Nc_g and $f(x_g)$ for every gene, terms φ_g and ε_g in Equation 6 are confounded, and cannot be estimated independently. Consequently, the amount of selection acting on the codon usage of gene g cannot be estimated directly, but the confounded factor, $\psi_g = \varphi_g - \varepsilon_g$ can be estimated as

$$\psi_g = f\left(x_g\right) - \mathrm{Nc}_g. \qquad 7$$

We can use tAI_g and ψ_g to estimate the degree of co-adaptation between the codon usage of a set of genes and the genomic tRNA pool. If a sample of genes from a given genome G is obtained, then the vectors $\mathrm{tAI}_G = (\mathrm{tAI}_g)$ and $\Psi_G = (\psi_g)$ can be calculated. The correlation S between tAI_G and Ψ_G measures this co-adaptation. In fact, the squared correlation coefficient S^2 is the proportion of the variance in codon bias explained by tRNA adaptation that cannot be explained by GC content variation (x_g) or other factors (ε_g) alone. The correlation S is a convenient indicator of the amount of selection due to tRNA adaptation since it is a single number between -1 and 1. It can be seen that the stronger the action of selection, the higher the correlation coefficient. If this test is applied to a representative set of genes in any given organism, a measure of the intensity of translational selection that has acted upon the evolution of its genome will be obtained.

The statistical significance of S can be assessed by a permutation test. The method consists in permuting the assignment of w_i values to their respective codons. The permuted set is then used to calculate tAI and S and the process is repeated iteratively until a sufficiently large sample of S-values is generated to estimate its probability distribution under the assumption of no selection. P-values for the significance of the naturally observed S-values can be obtained from this re-sampling distribution.

Non-parametric regression of *S*

In order to understand how genome size and tRNA gene copy number contribute to explaining selection on codon usage in genomes, a non-parametric regression of S on these variables was performed. A Gaussian processes model (28) was applied to the data since it provides a very flexible and powerful way to analyse data with unusual properties such as highly correlated predictors (such as genome size and tRNA gene number in this case). Gaussian processes are specified by a covariance structure that determines how response values at neighbouring points influence each other. A fully Bayesian treatment of the parameters describing the covariance structure is possible by a Monte Carlo Markov Chain (MCMC) algorithm. Random surfaces are generated according to how well they fit the observed data points, and a consensus regression surface is then obtained by averaging.

Although the smoothness of the surface is essentially derived from the data, it can be influenced by setting a prior on the scale parameter. We intentionally chose a prior that encouraged a smooth appearance of the regression surface. An advantage of the Bayesian approach is that parameters of interest can be given a probabilistic interpretation. For example, we were interested in locating the point of maximum translational selection in the regression surface. A density plot of the most likely location of this point can be easily obtained from the MCMC samples of surfaces by evaluating their maxima.

A logarithmic transformation was used to scale tRNA gene numbers and genome sizes to appropriate values in the regression analysis. Also, a generalized additive model (GAM) (29) was fitted to the data as an independent regression model to confirm the quality of the analysis. Both models agree reasonably well, but the Gaussian model was preferred due to its robustness, so the GAM analysis is not discussed any further.

Genomic sequences

In this work, 126 genomes were analysed for the presence or absence of translational selection. All protein coding sequences from bacterial genomes and yeast species (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe)* were retrieved from NCBI (ftp://ftp.ncbi.nih.gov/genomes/), sequences from *Homo sapiens* and *Arabidopsis thaliana* were retrieved from RefSeq (http://www.ncbi.nih.gov/RefSeq/), *Mus musculus* sequences from the Mammalian Gene Collection (http://mgc.nci.nih.gov), *Plasmodium falciparum* sequences from PlasmoDB (http://plasmodb.org), *Neurospora crassa* sequences from (http://www.broad.mit.edu/annotation/fungi/neurospora), *Drosophila melanogaster* sequences were obtained from flybase (http://flybase.bio.indiana.edu/) and *Caenorhabditis elegans* sequences from wormbase (http://www.wormbase.org). A detailed list of all organisms studied and their accession numbers is provided as supporting material. Genomic tRNA information was obtained from the Genomic tRNA Database (http://lowelab.ucsc.edu/GtRNAdb/), and from scanning individual genomes with tRNAscan-SE (30).

Software

Statistical analyses were done using the R software for statistical computing (http://www.r-project.org), and the Flexible Bayesian Modeling program suite (R. Neal, http://www.cs.toronto.edu/~radford/fbm.software.html). A modified version of CodonW (J. Penden, unpublished data; http://www.molbiol.ox.ac.uk/cu/) was used to estimate Nc and GC content of the sequences analysed; the modification eliminates the restriction on Nc values bigger than 61. All software developed specifically for this work is publicly available (http://people.cryst.bbk.ac.uk/~fdosr01/tAI/).

RESULTS

Choosing appropriate *s*-values for calculating tAI

One of the most challenging issues when computing tAI is the selection of a meaningful set of s_{ij}-values (Equation 1). Since tRNA usage should be maximal for highly expressed genes, it would be natural to find the set of s_{ij}-values that maximize the correlation between expression levels and tAI values for any given organism. In this study, microarray data from yeast were obtained (31) and used to optimize these values. A set of highly expressed genes was selected using the criteria chosen previously (32), and tAI was calculated for every one of them, assuming the initial s_{ij}-values as shown in Table 2. The correlation between the obtained tAI values for each gene and its corresponding expression level was then calculated iteratively using an implementation of the Nelder and Mead algorithm (R package) until the optimal set of s_{ij}-values that maximized this correlation (here, $R_{\text{final}} = 0.71$) was obtained (Table 2). A similar method was used with *E.coli* microarray data (33) to obtain an appropriate *s*-value for the recognition of AUA by LAT in prokaryotic genomes (Table 2).

Optimization of Wright's Nc

A puzzling issue in the study by Wright on Nc (27) is the fact that it is not explained how Equation 5 ($f(x_g)$), was obtained. Although this formula fits some experimental data, it seems to be biased towards higher Nc values than would be expected. Since an unbiased estimator is needed in Equation 7 to calculate S, its accuracy needs to be reviewed. In order to achieve this goal, each *E.coli* K-12 open reading frame (ORF) was simulated according to the following rules: (i) the amino acid composition should remain intact, (ii) the codon that codes for any given amino acid will be chosen randomly according to a silent set GC content for its gene, (iii) the silent GC content of any given gene will be chosen randomly from a uniform distribution. In order to obtain a sufficiently large dataset, each gene was simulated three times. The result of this simulation is a set of genes whose codon usage is solely determined by their GC content. As can be seen in Figure 2, the equation suggested by Wright has certainly deviated not only from the simulated results but also from real data. The problem was then finding a better fit for the simulated data utilizing the method of minimum squares. For simplicity, it was assumed that the structure of Equation 5 is right but that the constants present in it are inaccurate; based on this, Equation 5 was redefined as

$$f_1\left(x_g\right) = a + x_g + \frac{b}{x_g^2 + \left(c - x_g\right)^2}, \qquad \mathbf{8}$$

where a, b and c are constants with unknown values. The problem was reduced to finding the set of values for these constants that minimize the sum of squares of the residuals for the fitted curve:

$$\text{Sum of squares} = \sum_{g=1}^{n}\left(f_1\left(x_g\right) - \text{Nc}_g\right)^2, \qquad \mathbf{9}$$

where n is the total number of simulated sequences. Notice that in this case $\text{E}[f_1(x_g) - \text{Nc}_g] = 0$ as expected, since the effect of selection has effectively been eliminated by the simulation process. The actual minimization of Equation 9 was done using an implementation of the Nelder and Mead algorithm (R software). The obtained parameter values are shown in Table 3. In order to exclude any possible 'organism' effect on the estimation of a, b and c, the simulation experiment and the optimization process were repeated again using human genes (Table 3). We have found that Equation 8 is robust

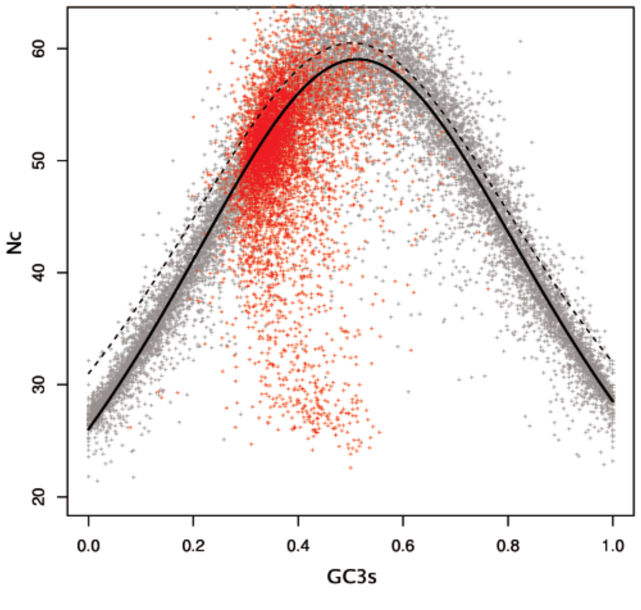


Figure 2. Nc-plot for yeast and simulated *E.coli* K12 genes. Grey points, simulated *E.coli* K12 genes; red points, actual yeast genes; dashed line, Wright's proposed function (Equation 5); bold line, the function proposed herein (Equation 8) with optimized parameters.

Table 2. Optimized *s*-values

s	Fixed	*s*	Initial	Final
$s_{\text{I:U}}$	0.0	$s_{\text{G:U}}$	0.50	0.41
$s_{\text{G:C}}$	0.0	$s_{\text{I:C}}$	0.50	0.28
$s_{\text{U:A}}$	0.0	$s_{\text{I:A}}$	0.25	0.9999
$s_{\text{C:G}}$	0.0	$s_{\text{U:G}}$	0.50	0.68
—	—	$s_{\text{L:A}}$	0.50	0.89

L, lysidine.

Table 3. Optimized parameters for Equation 8

par	*E.coli* K12	*H.sapiens*	Suggested
a	−6.459	−6.650	−6.0
b	34.01	34.43	34.0
c	1.023	1.028	1.025

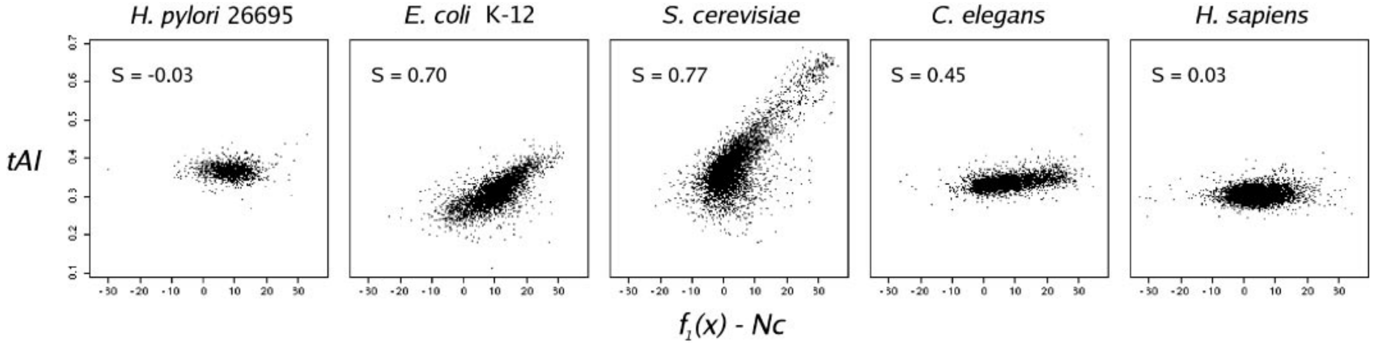


Figure 3. tAI versus $f_1(x)$–Nc for five organisms for which codon usage has been well studied.

for broad values of *a* and *b*, but is very sensitive to small deviations of *c*.

Distribution of translational selection across prokaryotic and eukaryotic genomes

To illustrate how the *S*-test can be used, *S*-values were calculated for five organisms in which codon usage has been well studied: *H.pylori*, *E.coli*, *S.cerevisiae*, *Caenorhabditis elegans* and *H.sapiens* (19,9,7,21,20). As expected (Figure 3), *S.cerevisiae* and *E.coli* show the highest *S*-values, *C.elegans* shows a moderate *S*-value, while *H.pylori* and *H.sapiens* show no sign of translational selection acting on their genomes.

A total of 126 genomes were tested for translational selection. The organisms analysed ranged in genome size from 0.58 Mb (*Mycoplasma genitalium*) up to ~3000 Mb (*H. sapiens*), with total tRNA gene copy numbers ranging from 29 (*Mycoplasma pulmonis*) to 620 (*A.thaliana*). We found that the presence or absence of translational selection is independent of the kingdom being considered; both eukaryotes and prokaryotes presented organisms whose codon usage is determined mainly by selection or mainly by mutational processes. *S*-values ranged from −0.28 (*Halobacterium* sp.) up to 0.82 (*S.pombe*). In total, 36 genomes have values of *S* statistically different from zero ($P < 0.05$). A complete table with all the estimated *S*-values for each genome and their statistical significance is available as supporting material.

The non-parametric regression of *S* on genome size and tRNA gene copy number explains ~60% of the variation observed in *S*-values. Genome size and tRNA gene copy number interact to form a landscape (Figure 4a) that determines where selection is operative. This landscape can be represented as a thermal image (Figure 4b) that shows the hot regions of selection activity. This landscape shows a conspicuous hot spot where the activity of selection on silent site evolution is maximal, and cooler, marginal regions of little selection. Small bacterial genomes (such as *H.pylori* or *Borrelia burgdorferi*) and big eukaryotic genomes (like those of *H.sapiens* or *M.musculus*) fall in these marginal regions. The yeast genomes fall within the hot spot region. This thermal image is a pan-genomic picture that depicts for the first time where translational selection is operative. It can be used to predict the presence of translational selection in any given genome, provided that genomic size and the number of resident tRNA genes are known. The maximum observed in the regression surface is highly stable as indicated by the MCMC simulation (Figure 4c).

DISCUSSION

The classic approach to test for natural selection at the molecular level has been through alignments of nucleotide sequences and the estimation of the number of synonymous and non-synonymous substitutions that have occurred during the evolution of those sequences. Since codon usage is a particular characteristic of every genome, usually there is not enough polymorphism data available of sequences within genomes to perform this kind of analysis (13), so indirect approaches such as measuring different sorts of codon bias indexes have been the norm in these types of study. Since these indirect approaches do not measure selection itself, most studies embracing larger groups of organisms have focused on the effect of nucleotide composition (such as GC content) on codon usage, and have paid scant attention to the problem of translational selection (16,17).

Presented in this paper is our development of the first consistent method to test for translational selection in any given genome, with some very interesting findings. This provides a framework for partially understanding the original observations of Grantham *et al*. (5) and their genome hypothesis. Furthermore, for the first time, the importance of considering the roles of tRNA gene redundancy and genome size in the mutation–selection balance theory of codon usage is emphasized.

Genome size and tRNA gene redundancy are interacting forces that determine the action of natural selection on codon usage bias

Perhaps the most important contribution of this work to the understanding of codon usage is the unexpected finding that genome size and tRNA gene redundancy interact to determine the action of natural selection on codon usage in all living organisms. Our findings suggest that an optimal combination of these factors exists, for which the action of translational selection is maximal. It is now possible to readily predict the presence of translational selection in any given genome given that we know its size and the number of tRNA genes it contains. Perhaps Paramecia will provide a good model system to test these predictions. For example, *Paramecium aurelia* has a genome size of ~190 Mb, while *P.caudatum* has a genome

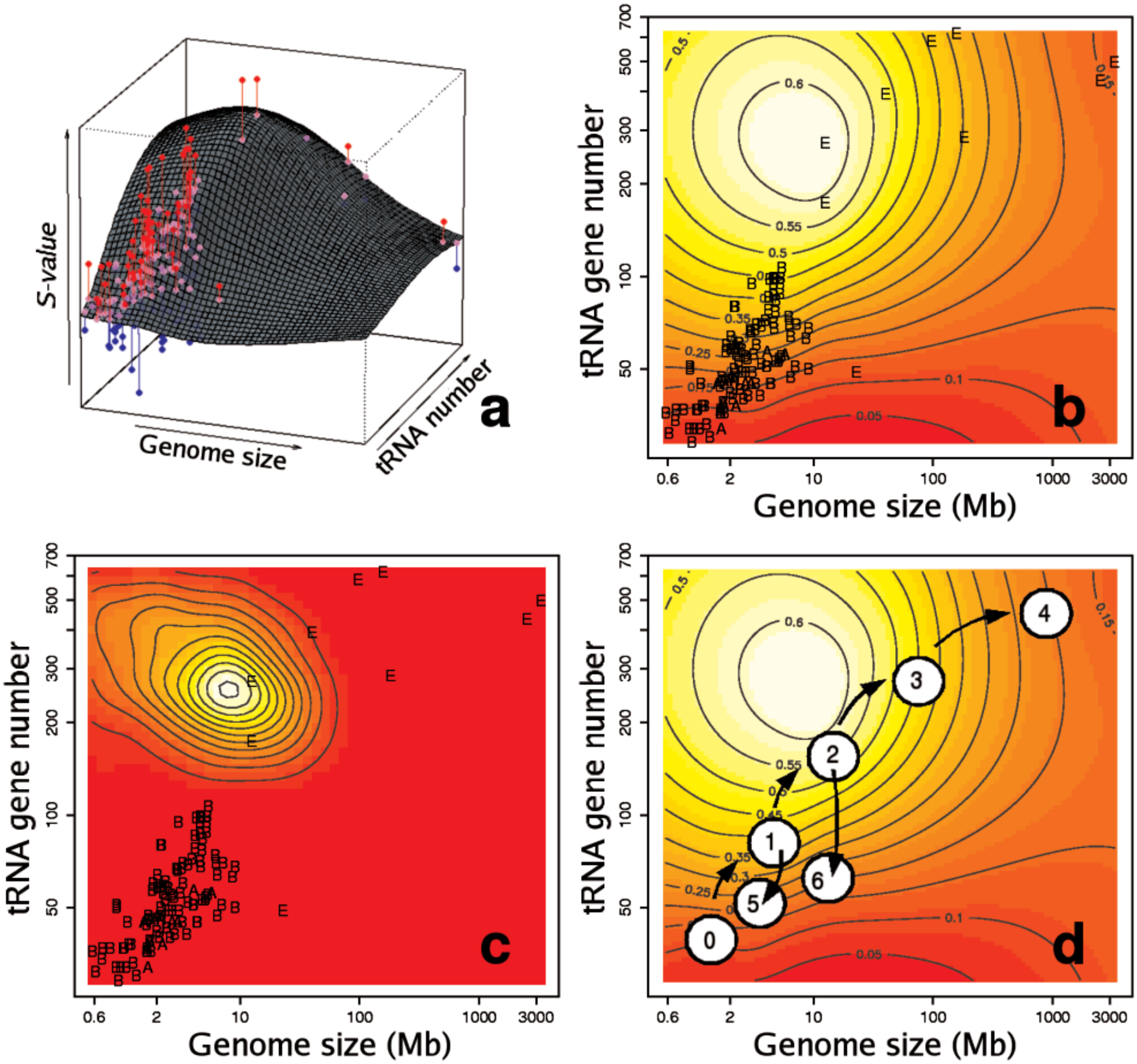


Figure 4. Action of natural selection on codon usage in the genomic landscape. (**a**) Fitted regression surface of *S*-values to tRNA gene number and genome size. Pink dots, predicted *S*-values for every organism; red dots, organisms with *S*-values higher than predicted by the model; blue, organisms with S-values lower than predicted. Vertical lines join each observed data point to its predicted value. (**b**) Thermal image (contour plot) of the same regression surface. The hottest (highest) *S*-values are shown in white, while the cooler (lowest) values are in red. E, Eukaryota; B, Eubacteria; and A, Archaea. The contour lines reflect the estimated *S*-value for a particular region. (**c**) Estimated probability density function for the presence of a maximum in the regression analysis. (**d**) Hypothetical evolution of codon usage optimization. A small genome sized ancestor (0), suffered a series of genome expansions (1–4). During this evolutionary process, the phylogeny would move into, and then out of the selective hot-spot. The process can be reverted at any time if selection for genome size or tRNA set reduction is present, such as in non-free living organisms e.g. *H.pylori* (5) or *P.falciparum* (6).

size of $\sim$8900 Mb, so, our model predicts that the former will have a codon usage bias similar to that of fly, i.e. with selection moderately shaping codon usage in highly expressed genes, while the latter should have a codon usage resembling humans, i.e. with no evidence of translational selection.

The lack of tRNA gene redundancy is responsible for the absence of translationally selected codons in small genomes

The adaptation of codon usage to the genomic tRNA gene pool is a well-known phenomenon in various organisms where translational selection is known to be present. In fact, some authors have discussed how the redundancy in the gene number of certain tRNA isoacceptors matches the frequencies of the preferred set of codons in yeast and worm (24,25). However, what does not appear to have been discussed so far, is how the lack of duplicated tRNA genes might explain the absence of translationally selected codons in bacteria with small genomes, although the idea has been framed previously: Kanaya *et al*. (14) made a detailed analysis of codon usage and tRNA abundance in 18 unicellular microorganisms, calculating a *Z*-index of codon usage for protein genes in the genomes studied. It was found that the relative bias in Z-values between ribosomal and non-ribosomal proteins was proportional to the total number of tRNA genes in the genome; furthermore, these data suggest that such bias is effectively zero for genomes with low tRNA gene numbers. The conclusion presented was that '. . . codon usage in most bacteria, if not all, is constrained by translation efficiency'; however, the aforementioned data appear to suggest the contrary. A vivid example of this is presented in the genome of *H.pylori*, where the absence of translationally selected codons is well documented (19); *H.pylori* presents only 36 tRNA genes with only one tRNA species presenting two copies, it is this lack of tRNA gene duplication that determines the absence of translational selection in that organism. It can be argued that it is the need for translational optimization and hence codon usage that shapes the tRNA pool of organisms (34). However, we contend that selection favouring small genome size implies an overall reduction of the 'redundancy' of the whole genome, i.e. reduction of duplicated genes of any kind (tRNA, rRNA, protein genes, repetitive elements, etc.), and it is this kind of selective force, not codon usage itself, that shapes tRNA redundancy.

Genome size presents upper and lower boundaries beyond which selection on synonymous codon usage is not possible

Since genome size and genomic tRNA number are highly correlated ($R = 0.85$), it seems logical to think that both factors co-evolve in a concerted way. The evolution of genome size has largely been related to the evolution of repetitive DNA (35), so the mechanisms that explain the increase in copy number of selfish genes within genomes might be taken into account to explain the evolution of tRNA genes. In fact, the association between tRNAs and transposable elements is well documented in eukaryotic genomes (36,37), and the evolution of tRNA genes themselves have been described as a repetitive process (38). Since a particular genome size perhaps limits the maximal number of tRNA genes it contains, it is this factor which imposes limits to the action of selection on codon usage. As discussed above, it becomes clear why small genomes lack translationally selected codon usage bias; however, why large genomes also show similar behaviour is puzzling. A possible explanation might be related to the fact that the length of the cell cycle is positively correlated with genome size (39), i.e. the larger the genome, the longer the cell division cycles. After all, *E.coli* cells might divide thousands of times in a few hours, while human cells in culture may only divide a few times. Translational optimization for expression of large proteins might be advantageous to *E.coli*, although of little value to human cells. Another possible explanation might exist if the effective population size (Ne) of an organism is correlated to its genome size. Whether this should be so is unclear, but it is well known that a substantially large Ne is needed (11) for selection to be effective in shaping codon usage.

The findings presented in this paper permit the following conjecture of how codon usage optimization might have evolved. First a hypothetical ancestor (Figure 4d) with a small genome and a reduced set of tRNA genes suffered a series of genome expansions that led to an increased set of tRNA genes. As successive expansions took place, the redundancy of the tRNA set increased and selective pressure for codon optimization started to be operative. From this, the first medium genome sized bacterial genomes originated, similar to *E.coli*. Further expansions might have produced the first eukaryotic genomes such as yeast, where codon optimization is highly developed. As genome size increased further, other ecological variables progressively hindered the action of selection on codon usage, generating the large modern genomes such as those of mammals. Selection for reduction of genome size or tRNA redundancy in certain non-free living organisms would invert the process. This conjectural model might be used as a plausible framework onto which research into codon usage may be devised.

Limitations of the model

The statistical model presented in this study failed to explain $\sim$40% of the variation observed in S-values in the genomes studied. The reasons for this are many, and include some of a technical, and others of a biological nature. The technical limitations are statistical in their character, and relate to the choice of regression model and the unusual structure of the data. tRNA gene numbers and genome size are highly correlated, so ample areas of 'genomic landscape' do not present data points, therefore, the predictions of the model in the upper left, and lower right corner of Figure 4 are extrapolations that might not necessarily reflect the true behaviour of organisms in these regions. Another unusual characteristic of the data is the excessive oversampling of small genomes, which are easier and cheaper to sequence; this certainly affects the model in the sense that densely packed areas of data contribute more heavily to the shape of the regression surface than areas with more scattered data. The choice of a particular regression analysis is also a problem. Classical parametric analyses such as polynomial regressions tend to over-fit the data as the degree of the polynomial is increased, and they also tend to produce surfaces that vary wildly in areas where the data is poorly represented. On the other hand, Gaussian processes are robust against highly correlated predictors, discontinuous data structures, or biased samples. The final regression surface seen in Figure 4 is an average of all the regression surfaces that explain the data 'equally' well, so the wild fluctuations of classical regression surfaces over the areas where the data is poorly represented is averaged out. Furthermore, the inclusion of simulation and Bayesian analysis allows us to test the reliability of the regression model, and obtain probability values for the features observed in it. If the model predicts a maximum in an area where the data is poorly represented, this maximum will have a very low probability because simulated surfaces in that area will tend to produce random maxima and minima; only regions that consistently produce the same feature are reliable. However, Gaussian processes also present limitations; this is a non-parametric model, and hence we lack a meaningful parametric equation to describe the data, and using it to predict accurate S-values for new organisms with values outside the range of the current data is not necessarily appropriate.

The model also failed to account for some biological variables. For example, secondary structure effects on codon preferences, or context-dependent mutational biases were not explicitly taken into account, and are grouped together in the error term of Equation 6. Another important variable that was considered in preliminary analysis was the silent

GC content; this variable can improve the predictive power of the model to ~70% but its inclusion does not change the shape of the regression surface (data not shown). GC content and codon usage have been widely discussed in the literature, and it is known that highly biased mutational patterns can restrict codon selection in certain organisms. Some bacteria with extreme values of GC content, present low *S*-values despite their intermediate genome sizes and tRNA numbers, e.g. *Clostridium tetani* (GC = 0.29, *S* = 0.03), species of the genus *Streptomyces* (GC > 0.70, *S* < −0.03) or *B.burgdorferi* (GC = 0.29, S = 0.11) where asymmetrical replication is the major source of codon usage variation (40), and where the presence of translational selection has been debated (41). We also ignored the fact that the proportion of tRNA isoaceptors can vary as a function of growth rate, tissue type, etc. and it has been suggested that genes accommodate their codon usage according to their particular 'tRNA environment'(42); in this work, the overall genomic tRNA content was used to calculate tAI and this might not always be appropriate. Also particular taxonomic groups or organisms with similar forms of life may show similar codon trends; e.g. thermophilic bacteria have been shown to have the same codon preferences despite their large variations in overall GC content (43).

Perhaps the best way to exemplify the limitations of the model is by analysing the case of *Bacillus subtilis*. This organism presents an *S*-value of −0.01, however, it has been reported that translational selection is operative in its genome (44). Previous analysis on codon usage in this organism (45) showed that its genome can be divided into three gene classes according to their codon composition. Class II, which represents only 4.6% of ORFs, comprises highly expressed genes that show biased codon patterns that can be explained through codon optimization to match the tRNA pool of this organism (14). Classes I (82.3%) and III (13.1%), which comprise the rest of the genome, show codon patterns determined by amelioration and mutation-random drift equilibrium; this is in contrast with the genome of *E.coli* K-12 where most of the genome seems to be under translational selection (22). It is evident that the selection effect on class II is hindered by classes I and III in our analysis when we consider the whole genome. Therefore, a small *S*-value for a whole genome means that translational selection might be negligible at a genomic scale, but it can nonetheless have a strong effect on smaller scales, such as particular gene sets.

This model unifies our understanding of selection on codon usage in prokaryotic and eukaryotic genomes

The findings presented in this paper coherently unify our understanding of the action of natural selection on codon usage in prokaryotes and eukaryotes. They show that there are indeed conspicuous trends that explain the different roles of selection and mutational biases across all living organisms. It is now possible to trace the enigma of codon usage down to the evolution of tRNA genes and genome size and organization. We think two lines of research need to be pursued in order to disentangle the codon usage riddle completely.

(i) Explaining the poorly understood evolution of tRNA genes and their role in fitness of an organism: do they propagate within genomes in a selfish manner hence determining codon usage, or do they indeed co-evolve with codon usage itself (34)?

(ii) Studies of a more ecological nature are needed to understand why eukaryotes tend to accumulate larger genomes, and what variables correlated to genome size can explain the lack of translational selection in the larger genomes. Work still needs to be done in this area, but having a test for translational selection will certainly help in tackling these questions.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

M.d.R. is currently being supported by a research studentship awarded by the Biosciences and Biotechnology Research Council, UK. Thanks to Dr Sean Eddy who kindly provided the mouse tRNA table.

REFERENCES

1. Kimura,M. (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.
2. King,J.L. and Jukes,T.H. (1969) Non-Darwinian evolution. *Science*, **164**, 788–798.
3. Clarke,B. (1970) Darwinian evolution of proteins. *Science*, **168**, 1009–1011.
4. Post,L.E. and Nomura,M. (1980) DNA sequences from the *str* operon of *Escherichia coli*. *J. Biol. Chem.*, **255**, 4660–4665.
5. Grantham,R., Gautier,C., Guoy,M., Mercier,R. and Pave,A. (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.*, **8**, r49–r62.
6. Grantham,R., Gautier,C. and Guoy,M. (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res.*, **8**, 1893–1912.
7. Bennetzen,J.L. and Hall,B.D. (1982) Codon selection in yeast. *J. Biol. Chem.*, **257**, 3026–3031.
8. Guoy,M. and Gautier,C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055–7074.
9. Ikemura,T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.*, **146**, 1–21.
10. Ikemura,T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.*, **151**, 389–409.
11. Bulmer,M. (1991) The selection–mutation–drift theory of synonymous codon usage bias. *Genetics*, **129**, 897–907.
12. Sharp,P.M., Stenico,M., Penden,J.F. and Lloyd,A.T. (1993) Codon usage: mutational bias, translational selection or both? *Biochem. Soc. Trans.*, **21**, 835–841.
13. Sharp,P.M., Averof,M., Lloyd,A.T., Matassi,G. and Penden,J.F. (1995) DNA sequence evolution: the sounds of silence. *Phil. Trans. Biol. Sci.*, **349**, 241–247.
14. Kanaya,S., Yamada,Y., Kudo,Y. and 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.
15. Kanaya,S., Yamada,Y., Kinouchi,M., Kudo,Y. and Ikemura,T. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translational efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.*, **53**, 290–298.
16. Knight,R.D., Freeland,S.J. and Landweber,L.F. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.*, **2**, research0010.1-0010.13.

17. Chen,S.L., Lee,W., Hottes,A.K., Shapiro,L. and MacAdams,H.H. (2004) Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl Acad. Sci. USA*, **101**, 3480–3485.
18. Brookfield,J.F.Y. (2003) Genome evolution. In Balding,D.J., Bishop,M. and Cannings,C. (eds), *Handbook of Statistical Genetics*. Wiley, Chichester, pp. 255–281.
19. Lafay,B., Atherton,J.C. and Sharp,P.M. (2000) Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology*, **146**, 851–860.
20. Urrutia,A.O. and Hurst,L.D. (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, **159**, 1191–1199.
21. Stenico,M., Lloyd,A.T. and Sharp,P.M. (1994) Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res.*, **22**, 2437–2446.
22. dos Reis,M., Wernisch,L. and Savva,R. (2003) Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res.*, **31**, 6976–6985.
23. Sharp,P.M. and Li,W.-H. (1986) The codon adaptation index: a measure of directional synonymous codon usage, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
24. Percudani,R., Pavesi,A. and Ottonello,S. (1997) Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **268**, 322–330.
25. 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 in Genetics*, **16**, 287–289.
26. Crick,F.H. (1966) Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, **19**, 548–555.
27. Wright,F. (1990) The 'effective number of codons' used in a gene. *Gene*, **87**, 23–29.
28. Neal,R.M. (1998) Regression and classification using Gaussian process priors. In Bernardo,J.M. *et al*. (eds), *Bayesian Statistics 6*. Oxford University Press, Oxford, pp. 475–501.
29. Hastie,T.J. and Tibshirani,R.J. (1990) *Generalized Additive Models*. Chapman and Hall, London.
30. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
31. Holstage,F.C.P., Jennings,E.G., Wyrick,J.J., Lee,T.I., Hengartner,C.J., Green,M.R., Golub,T.R., Lander,E.S. and Young,R.A. (1999) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.
32. Coghlan,A. and Wolfe,K.H. (2000) Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast*, **16**, 1131–1145.
33. Bernstain,J.A., Khodursky,A.B., Lin,P.-H., Lin-Chao,S. and Cohen,S.N. (2002) Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl Acad. Sci. USA*, **99**, 9697–9702.
34. Bulmer,M. (1987) Coevolution of codon usage and transfer RNA abundance. *Nature*, **325**, 728–730.
35. Petrov,D.A. (2001) Evolution of genome size: new approaches to a new problem. *Trends Genet.*, **17**, 23–28.
36. Lawrence,C.B., MacDonnel,D.P. and Ramsey,W.J. (1985) Analysis of repetitive sequence elements containing tRNA-like sequences. *Nucleic Acids Res.*, **13**, 4239–4252.
37. MGSC (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
38. Jukes,T.H. and Holmquist,R. (1972) Evolution of transfer RNA molecules as a repetitive process. *Biochem. Biophys. Res. Commun.*, **49**, 212–126.
39. Cavalier-Smith,T. (1985) Introduction: the evolutionary significance of genome size. In Cavalier-Smith,T. (ed.), *The Evolution of Genome Size*. Wiley, NY, pp. 1–36.
40. McInerney,J.O. (1998) Replication and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc. Natl Acad. Sci. USA*, **95**, 10698–10703.
41. Perrière,G. and Thioulouse,J. (2002) Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res.*, **30**, 4548–4555.
42. Kurland,C.G. (1993) Major codon preferences: theme and variations. *Biochem. Soc. Trans.*, **21**, 841–846.
43. Lynn,D.J., Singer,G.A.C. and Hickey,D.A. (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.*, **30**, 4272–4277.
44. Moszer,I., Rocha,E.P.C. and Danchin,A. (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr. Opin. Microbiol.*, **2**, 524–528.
45. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessières,P., Bolotin,A. and Borchert,S. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.
46. Watanabe,K. and Osawa,S. (1995) tRNA sequences and variation in the genetic code. In Söll,D. and RajBhandary,U. (eds), *tRNA: Structure, Biosynthesis and Function*. AMS Press, Washington, DC, pp. 225–250.
47. Yokoyama,S. and Nishimura,S. (1995) Modified nucleosides and codon recognition. In Söll,D. and RajBhandary,U. (eds), *tRNA: Structure, Biosynthesis and Function*. AMS Press, Washington, DC, pp. 207–223.