Nucleic Acids Research, 2006, Vol. 34, No. 3 ***1015–1027***
doi:10.1093/nar/gkj488

tRNA properties help shape codon pair preferences in open reading frames

J. Ross Buchan, Lorna S. Aucott[1] and Ian Stansfield*

School of Medical Sciences, University of Aberdeen, Institute of Medical Sciences, Foresterhill,
Aberdeen AB25 2ZD, UK and [1]Department of Public Health, School of Medicine, University of Aberdeen,
Polwarth Building, Foresterhill, Aberdeen AB25 2ZD, UK

Received October 26, 2005; Revised December 16, 2005; Accepted January 18, 2006

ABSTRACT

Translation elongation is an accurate and rapid process, dependent upon efficient juxtaposition of tRNAs in the ribosomal A- and P-sites. Here, we sought evidence of A- and P-site tRNA interaction by examining bias in codon pair choice within open reading frames from a range of genomes. Three distinct and marked effects were revealed once codon and dipeptide biases had been subtracted. First, in the majority of genomes, codon pair preference is primarily determined by a tetranucleotide combination of the third nucleotide of the P-site codon, and all 3 nt of the A-site codon. Second, pairs of rare codons are generally under-used in eukaryotes, but over-used in prokaryotes. Third, the analysis revealed a highly significant effect of tRNA-mediated selection on codon pairing in unicellular eukaryotes, *Bacillus subtilis*, and the gamma proteobacteria. This was evident because in these organisms, synonymous codons decoded in the A-site by the same tRNA exhibit significantly similar P-site pairing preferences. Codon pair preference is thus influenced by the identity of A-site tRNAs, in combination with the P-site codon third nucleotide. Multivariate analysis identified conserved nucleotide positions within A-site tRNA sequences that modulate codon pair preferences. Structural features that regulate tRNA geometry within the ribosome may govern genomic codon pair patterns, driving enhanced translational fidelity and/or rate.

INTRODUCTION

For any living cell, the process by which the ribosome synthesizes proteins represents a keystone of cellular metabolism. The ribosome acts as the interpreter of the genetic code, and must be both an accurate and efficient translator of mRNA. Efficiency is important since the maximum rate of protein synthesis achievable by a cell defines the rate of cell division, and thus the competitiveness of unicellular species. Consistent with this, the cellular content of ribosomes and tRNAs are proportional to growth rate (1). The rate of amino acid incorporation during translation is both codon and growth condition-specific, and has been measured at between 7 and 20 incorporation events per second (2,3). Ribosomal accuracy is the second crucial property of the translation apparatus, since the activity of all enzymes and structural proteins depends upon the ribosome assembling a polypeptide chain with the correct amino acid sequence. This requirement is particularly important when translating proteins that are central to the expression, and propagation of, genetic information, such as subunits of the RNA polymerase and DNA polymerase complexes. Translational errors made during synthesis of the DNA replication apparatus can have knock-on consequences for the mutation rate in an organism (4). Various *in vitro* and *in vivo* estimates of the accuracy of amino acid misincorporation have revealed a global error frequency of between 1 error in 10 000 and 1 error in 500 000 codons translated (5–7). These estimates incorporate the summed error frequencies of transcription, tRNA charging and translation. The error and rate estimates together reflect a remarkably efficient translational machine.

Missense translation is only one of a range of errors that can occur during the process of translation. Ribosomal frameshifting and ribosome bypassing, the latter representing a ribosomal slide down the mRNA to an identical downstream A-site codon, are both contributors to a global error frequency (8,9). In addition, false recognition of sense codons as termination codons and of termination codons as sense can also occur (10,11). In a limited number of cases, off-pathway translational events such as these are employed by the cell to regulate gene expression at the level of translation. Regulation of gag and pol protein expression in human immunodeficiency virus (HIV) is achieved through a regulated −1 frameshift event

*To whom correspondence should be addressed. Tel: +44 1224 555806; Fax: +44 1224 555844; Email: i.stansfield@abdn.ac.uk

(12). Numerous other viruses and transposons similarly employ either +1 or −1 frameshifting to regulate their gene expression (8). Stop codon readthrough is employed by both viruses and cells as a mechanism to regulate the C-terminal extension of parent proteins with functional peptide sequences (11). Nevertheless, such 'recoding' events are the exception rather than the rule, and there is accumulating evidence that the sequence contexts that trigger such events are selected against. Heptanucleotide sequences known to trigger +1 frameshifting are significantly under-represented in the *Saccharomyces cerevisiae* open reading frame (ORF) set (13). Hexanucleotide sequences known to stimulate stop codon readthrough are likewise under-represented immediately downstream of stop codons (14). Selection is thus a powerful force that acts to eliminate accuracy-threatening sequences from ORFs.

If selection tends to eliminate frameshift and stop codon readthrough signals, is there any counterpart evidence that the accuracy, or rapidity, of sense codon decoding is enhanced through selection of optimal ORF sequences? It is known that codon usage within ORFs is subject to a high degree of bias (15). In many organisms, particularly fast-growing microbes, highly expressed genes make selective use of that subset of codons decoded by those most abundant isoacceptor tRNA species (16). There is some evidence that codon bias can contribute both to the rapidity and accuracy of ribosomal decoding events (17–19), although this 'rule' is not inviolate; those most heavily used codons are not always the most rapidly translated (20). Recent evidence also indicates a regulatory role for codon bias in the response of the translation apparatus to amino acid limitation (21,22). During starvation, those tRNAs for which demand (codon abundance in the transcriptome) is not matched by supply (tRNA abundance) will become exhausted more rapidly. In some cases, charged forms of abundant tRNAs, rather than minor isoacceptor forms, are predicted to be exhausted first. Codon bias therefore plays a central role in regulating translational output during amino acid starvation (21). Codon bias is thus increasingly understood as a versatile adaptation that enhances the fidelity, kinetics and starvation responses of the translation system.

Just as codon frequencies themselves are biased within ORFs, so codon pair frequencies are also non-random, termed codon pair bias. Based on a survey of a limited number of genes, the frequency with which codons are found juxtaposed was found to be non-random in *Escherichia coli* (23,24). Evidence of non-random associations remained even once codon bias and bias against specific amino acid pairings (dipeptide bias) were subtracted (23). A weak inverse correlation between codon pair bias and codon bias was reported (23). Those genes with high codon bias contain disproportionately high numbers of those codon pairs expected to be rare. These initial limited surveys of codon pair bias in *E.coli* have more recently been expanded to a genome-wide survey in both *E.coli* and yeast, confirming that codon pair frequencies are indeed highly biased (25,26), and that there are detectable differences in codon pair bias between high and low codon-biased gene sets (23,25). Consistent with a codon bias relationship, codon pair bias has a demonstrable effect on translational elongation rates *in vivo* (27,28). Other evidence suggests that misincorporation errors are highly dependent upon the context in which a codon lies, thus inferring that codon pairing can also influence translational fidelity (17,29,30). However, despite this body of evidence, there has been no detailed investigation of whether codon pairing is biased across all genomes, whether the rule-set that governs bias is the same in all genomes, and crucially, whether translation optimization might be a primary selective pressure driving codon pair bias.

In this research, a comprehensive survey of codon pair bias was undertaken across all ORFs in a range of 16 genomes. The study revealed that codon pair bias was a feature of every genome examined, and that conserved rules identify those nucleotide interactions between the A- and P-site codons that direct bias. Using a novel application of cluster analytical methods, we show that in the gamma proteobacteria, *Bacillus subtilis* and unicellular eukaryotes, A-site tRNA identity is a strong determinant of pairing preferences. The findings throw new light on the selective force exerted on ORF sequence by the translation apparatus. Specifically, in a number of species, multivariate analysis showed that structural properties of tRNAs appear to direct codon pair preferences within ORFs.

MATERIALS AND METHODS

Databases

Protein-encoding sequences derived from entire genomes were obtained in FASTA format from the following sources: all prokaryotic sequences were obtained from The Institute for Genomic Research (http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi). *Saccharomyces cerevisiae* sequence was obtained from the *Saccharomyces* Genome Database (ftp://genome-ftp.stanford.edu/pub/yeast). *Schizosaccharomyces pombe* sequence was obtained from the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/yeast/pombe/). *Caenorhabditis elegans, Drosophila melanogaster* and *Homo sapiens* sequence was obtained from the Sanger Institute using the Biomart sequence abstraction tool (http://www.ensembl.org/index.html). *Arabadopsis thaliana* sequence was obtained from the MIPS website (ftp://ftpmips.gsf.de/cress/). All tRNA sequences were obtained from The Genomic tRNA Database (31).

Perl programming

A series of novel programs was written using the Perl language (ActiveState Perl, version 5.005_3) and run in an MS-DOS environment. All programs written as part of this study are freely available on request from the authors. Formatted text output files generated by these programs were then analysed using Excel (Microsoft), the TIGR Multiexperiment Viewer [MeV (32)] and SIMCA-P (Umetrics) software.

Calculating and normalizing codon pair frequencies

Codon pair frequencies were calculated from data generated by the Perl program CODONCOUNT (this study). For each ORF in turn, the program determines the frequency of every codon and records the observed number (o) of all 3904 codon pairs possible ($61 \times 64 = 3904$; excluding stop codon:sense codon and stop:stop pairs). Additionally, for each ORF the expected number of a given codon pair (e_{ij}) was calculated using the measured codon frequencies within that ORF.

$$e_{ij} = \frac{c_i c_j N_p}{N_{\mathrm{tot}}^2}$$

where c_i is the number (*c*ount) of codons of type *i* within an ORF, N_{tot} is the total number of codons in that ORF and the number of codon pairs in that ORF is represented by $N_p = N_{\mathrm{tot}} - 1$.

Observed and expected codon pair counts calculated for each ORF were then summed to give tables of total observed and expected counts for the genome being analysed.

The effect of dipeptide bias on codon pairing was removed by normalizing the expected values of each codon pair, to generate e_{nor}:

$$e_{\mathrm{nor}\ ij} = \frac{\sum_{kl} [o \text{ codon pairs encoding dipeptide } kl]}{\sum_{kl} [e \text{ codon pairs encoding dipeptide } kl]} \times e_{ij}$$

(where *kl* is any dipeptide, *ij* is any codon pair, and e_{ij} is the expected codon pair count for a given individual *kl*-encoding codon pair).

This method of recording observed and expected codon pair values and normalizing for dipeptide bias has been described previously in a smaller scale study of *E.coli* codon pairing (23). In addition to analysing codon pair frequencies using this method in a range of genomes, codon pair frequencies were also analysed in codon biased, but artificially codon-order randomized genomes. These were generated using the Perl program RANDOM (this study).

Data analysis

Cluster analysis of codon pairing. The clustering program MeV (32) was employed to detect patterns of codon pair preference. Observed (*o*) and normalized expected (e_{nor}) codon pair counts were first converted into residual scores for each codon pair ($[o - e_{\mathrm{nor}}]/e_{\mathrm{nor}}$). Codon pair residual values were then analysed in MeV. Hierarchical average linkage clustering was used to group individual 5′ (P-site) and 3′ (A-site) codons from every codon pair according to patterns of similar codon pairing preference. This generated a 2D plot, organized with P-site and A-site codons on the horizontal and vertical axes, respectively. Bootstrap values with replacement, computed in MeV using 100 iterations, were employed to determine confidence in clustering patterns observed.

Analysing dinucleotide bias at codon–codon junctions. Dinucleotide bias at codon–codon junctions (cP3-cA1) acting as the sole influence on codon pair frequencies would be expected to uniformly select for, or against all codon pairs sharing a given class of cP3-cA1 dinucleotide. The degree of such selection was assessed by calculating the homogeneity of codon pair bias polarity within a given cP3-cA1 codon pair group. If all codon pairs within a particular cP3-cA1 group were uniformly over or under-represented, then bias polarity would be 100% homogeneous. In contrast, if there were equal numbers of over- and under-represented codon pairs in a particular cP3-cA1 group, then a 0% homogeneity of bias polarity would be recorded. The homogeneity index for a given dinucleotide ab (H_{ab}) is defined:

$$H_{\mathrm{ab}} = \frac{|N_+ - N_-| \times 100}{N_{\mathrm{tot}}}$$

(where N_+ is the total number of positive codon pair residual scores within a given cP3-cA1 dinucleotide pair type, N_- is the total number of negative codon pair residual scores within a given cP3-cA1 dinucleotide pair type, and N_{tot} is the total number of codon pairs of that cP3-cA1 dinucleotide type). The overall homogeneity index for a given genome is the mean of all 16 individual cP3-cA1 homogeneity indices.

Analysing nucleotide couples that span two codons to identify factors governing codon pair preference. The frequencies of nucleotide pair combinations that span two adjacent codons were determined. For all nucleotide pairs tested, nucleotide 1 was derived from the 5′ P-site codon (nucleotide cP1, cP2 or cP3) and nucleotide 2 was derived from the 3′ A-site codon (nucleotide cA1, cA2 or cA3). Nucleotide pair frequencies were calculated by first summing observed [*o*] and separately, normalized expected [e_{nor}] counts, for all codon pairs having an identical nucleotide couple (e.g. T-G) for a given combination type, e.g. cP2-cA3. This process was repeated for the eight other nucleotide pair permutations. The process generated 16 summed observed values, and 16 summed expected (e_{nor}) values for each of the 9 nt pair permutations that span P- and A-site codons. For each of these nucleotide pairs, the observed and expected values of each nucleotide couple were subject to Chi-squared analysis using 3503 degrees of freedom. This number is derived from the 3904 − 1 degrees of freedom representing the initial codon pair dataset, minus the 400 (20^2) degrees of freedom lost during dipeptide bias normalization. This Chi-squared analysis was used to determine those nucleotide pair combinations that exhibited a significant dinucleotide bias. Since multiple significance tests were carried out, a 0.1% significance level test was employed to reduce the possibility of falsely concluding that a given nucleotide couple frequency was significant.

Clustering of synonymous codons: significance assessment. A number of tRNA isoacceptors decode a pair of synonymous codons [hereafter referred to as mono-isoacceptor groups (MIGs)]. To determine if, during average linkage cluster analysis, MIG codons cluster on the A-site axis more often than expected, a Perl program PAIRSIM (this study) was employed to generate a random probability distribution for the association of two MIG codons in the A-site cluster analysis. PAIRSIM randomly clusters all 64 codons into sets of two, and then records the number of MIG codon sets that have been coupled by chance. In order to generate a probability distribution that did not under-estimate the numbers of MIG pairings regarded as significant, PAIRSIM actually couples codons after they have been first grouped into four sets of 16 codons on the basis of a shared first nucleotide identity. This increases the chance likelihood of MIG codon coupling in the random clustering exercise; the resulting probability distribution produced a more parsimonious estimate of the degree of MIG pairing considered significant. This modification of the program controlled for the fact that the first nucleotide of the A-site codon is typically a partially dominant factor in clustering A-site codons according to similar pairing preference, as revealed by codon pair cluster analyses. PAIRSIM was also tailored prior to use to reflect the unique complement of tRNA isoacceptors found in a given genome, since this influences the probability of MIG codon clustering at random. The probability distribution thus generated was used to determine the 2.5% confidence limit for the expected number of chance MIG codon associations at the base of the cluster tree.

Identifying the classes of codon pairs that are under- or overused. Codon pairs were ordered according to their expected frequency, and then grouped into 10 equal bins. The 0–10% bin contained codon pairs expected to be the rarest, and the 90–100% bin contained codon pairs expected to be most common. For each codon pair, the residual value ($o - e_{\mathrm{nor}}/e_{\mathrm{nor}}$) was calculated. Sense:stop codon pairs were ignored for this analysis. The mean residual value for codon pairs in each 10% bin was then calculated. The same procedure was also performed using a greater degree of smoothing by employing just two 50% bins.

Calculating average codon pair bias across entire genes. The program CODONPAIRINDEX (this study) calculates the average codon pair residual value across each gene in a genome. The program was tailored to incorporate the library of codon pair residual scores for the genome being analysed. For each gene, CODONPAIRINDEX records the totals of all codon pair types within the sequence, then looks up all the corresponding residual scores in the residual 'library', before calculating an average codon pair residual, or 'codon pair index' for that gene.

Calculation of codon adaptive index and codon bias index values. These were obtained by analysing the same sequences as analysed in CODONPAIRINDEX, using CodonW software (J.Peden; implemented at http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html).

Partial least squares (PLS) multivariate analysis. PLS multivariate analysis is a statistical method for relating two sets of data matrices x and y to one another (33). It uses regression techniques to model the associations of the two sets of variables in a predictive manner. Codon pairs and their residuals scores were tabulated using Excel (Microsoft) together with the sequences of the tRNAs that decode the codons in each pair. PLS analysis was implemented using the multivariate statistical package SIMCA-P (Umetrics). The output from this analysis was used to identify those tRNA nucleotide positions with a given identity that were good predictors of the codon pair residual value. In the PLS analysis, codon pair residual scores were assigned as the dependent variable y. The identity of nucleotides at all positions within the P- and A-site tRNAs (2 × 73 nucleotides), and in the codons themselves (6 nt), were assigned as qualitative independent x variables (i.e. potential predictors of the dependent variable). Four independent models were developed, each using 25% of the dataset defined by cP3 nucleotide identity. This division of the dataset greatly improved the quality of the models developed. Dataset analysis revealed this was because a given A-site tRNA nucleotide could often have a different polarity of effect on the residual value dependent upon the identity of the cP3 nucleotide. For each cP3-defined sub-dataset, automated modelling was performed in SIMCA-P. In most cases, this generated a two component model. Three of the twelve models (four per species) employed three components to explain the majority of variation. Each component represents a projection of the variation within the x point swarm (tRNA nucleotide identities) that maximizes prediction of the 1D y (codon pair residual value) dataset.

A PLS weights plot of $w_1 \times c_1$ against $w_2 \times c_2$, representing the first two components, was used to analyse the quantitative relationship between x and y variables. A weights plot reports the way in which x variables combine to form the 'latent variable' t score vectors, themselves the basis of the quantitative x–y relationship. The x weight matrix contains weight vectors w_a that describe how x variables are linearly combined to derive score vectors t_a. Similarly, weight vectors c_a describe how y variables are defined by a score vector u_a. Variable importance (VIP) analysis was performed using SIMCA-P, a computation of the influence of every x term in the model on the y variable (codon pair residual). Larger VIP values indicate a greater influence of a term x on the y variable. Nucleotide positions that scored greater than an arbitrary cut-off value of 2 in the SIMCA-P variable importance analyses were assigned as significant predictors (where a VIP value of $\geqslant 1$ is regarded as significant).

RESULTS

Codon pairing is biased in all genomes examined

Previous studies have established that codon pairing patterns in a limited range of species are non-random, but the underlying cause of this bias has not been established. In this study, multiple genomes were examined for evidence of codon pair bias, with the aim of establishing the basis of such bias. Species were chosen to reflect a diverse range of eubacteria and eukaryotes. Codon pair frequencies were measured for all 3904 codon pairs (omitting potential stop:stop and stop:sense codon pairs) in all ORFs for a variety of prokaryotic and eukaryotic genomes. Observed (o) codon pair counts were compared with summed expected counts (e) calculated 'locally' for each gene within a genome, thus negating the effect of codon bias on expected codon pair frequency (23). Having additionally normalized for the effects of dipeptide bias (Materials and Methods), Chi-square analysis was used to verify that codon pair bias in all genomes studied was indeed highly significant ($\sum\chi^2 > 100 \times$ standard deviations from the mean, Figure 1). As expected, codon pair bias in artificially randomized, but codon biased, *E.coli* and *S.cerevisiae* genomes was not significant (Figure 1).

Codon pair preference patterns can be revealed by cluster analysis

A statistical clustering method was used to group codons with similar pair preferences. By clustering codons with similar pair bias, rules governing pair preferences should be visible by eye. The degree of over- or under-use of each codon pair was therefore expressed as a residual value. Average linkage cluster analysis was then performed on pairing preferences of both the 5′ codon (that which would be located in the ribosomal P-site) and the 3′, A-site codon. Codons with similar pairing preference were clustered together on the two axes (A- and P-site) of a graphical display. This method facilitates data visualization of similarity of codon pair preference;

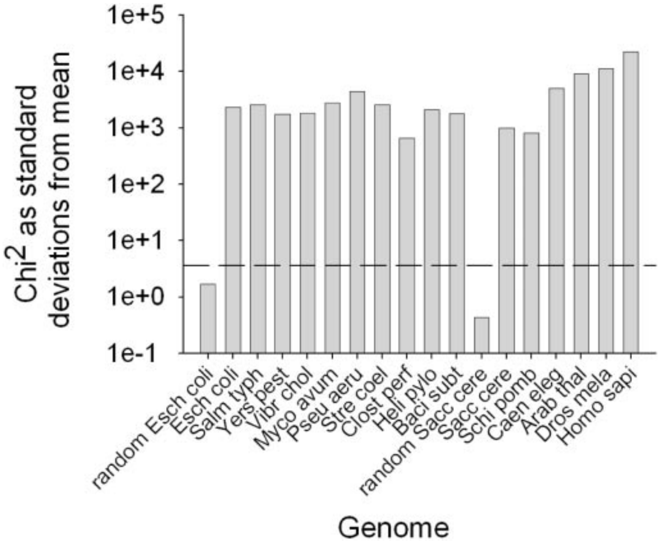


Figure 1. Codon pair bias is highly significant in all genomes. The statistical significance of codon pair bias (the difference between observed and expected codon pair counts) in the range of genomes tested was assessed using χ^2 analysis. On a semi-log plot, bars represent the number of standard deviations the $\sum\chi^2$ value lies from the mean. The dotted line indicates the number of standard deviations representing the 99.99% significance level. Species designations used comprise the first four letters of genus and species names, respectively.

those codons with similar pair preferences will be grouped by the hierarchical clustering at the extremities of the dendrogram (Figure 2). In the case of *E.coli* (Figure 2A), it was evident that P-site codons were strongly organized by the identity of their third nucleotide (XXN; 'cP3'). In contrast, A-site codons were clustered by the identity of their first nucleotide (NXX; 'cA1') albeit slightly less strongly.

Strikingly, for most prokaryotic genomes examined this pattern was repeated (*Yersinia pestis*, *Salmonella typhimurium*, *Vibrio cholerae*, *Pseudomonas aeruginosa*, *Helicobacter pylori* and *Bacillus subtilis*, Supplementary Figures S1–S6), although in a minority of prokaryotes the effect was less marked (*M.a.paratuberculosis*, *Clostridium perfringens* and *Streptomyces coelicolor*, Supplementary Figures S7–S9). In eukaryotes an even more regimented organization of P-site codons by cP3 nucleotide and identity, and A-site codons by cA1 nucleotide identity was observed (*S.cerevisiae*, *S.pombe*, *C.elegans*, *D.melanogaster*, *A.thaliana* and *H.sapiens*, Supplementary Figures S10–S15). The dominant cP3-cA1 trend revealed by this analysis demonstrates that codon pair bias examined on a genomic scale, across a large variety of organisms, is defined by a similar specific interplay between nucleotides flanking codon–codon junctions.

In addition to a cP3-cA1 influence on codon pairing, a careful examination of the *E.coli* cluster A-site clustering pattern reveals that other nucleotide positions in the codon (e.g. cA2 and cA3) affected pair preference, suggesting that complex selective forces drive codon pairing (Figure 2). In summary, codon pairing across all organisms studied is highly biased, and for the majority of organisms studied, an interaction based on nucleotide pair identity at positions cP3-cA1 within codon pairs is a key determinant of codon pair bias.

DNA dinucleotide bias is not a dominant force shaping codon pair bias

It is clear that cP3 and cA1 combine in some way to govern codon pair preference. This suggested that codon pairing could in theory be governed solely by a dinucleotide bias acting at codon junctions. In many organisms, DNA dinucleotides are not used with equal frequency in the genome. In higher eukaryotes, especially humans this effect is particularly marked for CG dinucleotide pairs, the targets of methylation (34). To test this, we grouped codon pairs sharing a common cP3-cA1 dinucleotide. An analysis was then performed of whether codon pair preference residual values were uniformly positive, or uniformly negative, within such groups. If DNA dinucleotide bias is a dominant force shaping codon pair bias, it would be predicted that such codon pair sets would exhibit a uniform polarity of bias, i.e. all codon pairs in that set should be either under-represented or over-represented. Alternatively, if dinucleotide bias were not a dominant selective force on codon pair bias, a set of codon pairs with an identical cP3-cA1 dinucleotide might comprise a mixture of over- and under-represented codons.

Codon pairs sharing a given cP3-cA1 dinucleotide (16 possible types) were grouped and the proportion of codon pairs either over- or under-represented (relative to the expected value e_{nor}), within each group was recorded. This was expressed as a homogeneity index (Materials and Methods). A value of 100% represents all codon pairs being either under- or over-represented, and 0% represents an equal split of under- or over-represented codon pairs. The homogeneity index was calculated for each organism (Figure 3). The results showed that even in the case of humans, simple dinucleotide bias alone cannot explain the patterns of codon pairing observed in all of the organisms studied. Most exhibit a mean homogeneity level of $\sim$35%, indicating a strong mixture of over- and under-represented codon pairs that nonetheless share identical dinucleotide pairs at their codon junctions. It is likely that DNA dinucleotide bias, acting at codon–codon junctions, might influence codon pairing to some degree, particularly in *H.sapiens*. However, in most prokaryotes and yeast, it is clearly not a dominant force.

Codon pair preference is governed by a tetranucleotide that spans adjacent codons. Careful inspection of the codon pair cluster analyses indicated that nucleotide couples other than cP3-cA1 were also important in directing codon pair patterns. To thoroughly address this, all nucleotide pair permutations spanning the two codons (e.g. cP1-cA1, cP2-cA3, etc.) were examined for bias using Chi-squared analysis (Materials and Methods).

The first finding was that in almost all cases, the interaction between cP3 and all three positions within the A-site codon is significantly biased (Figure 4). In agreement with previous data (Figures 2 and 3), this shows that while the interaction between cP3 and cA1 positions within codon pairs is an important factor acting upon codon pair selection, interactions between nucleotide positions cP3 and cA2/cA3 are also key contributors to the global patterns observed. This also supports the assertion that cP3-cA1 dinucleotide bias alone is not driving codon pair bias. Additionally, in most species cA3 is also a key codon position as it shows significant interplay with

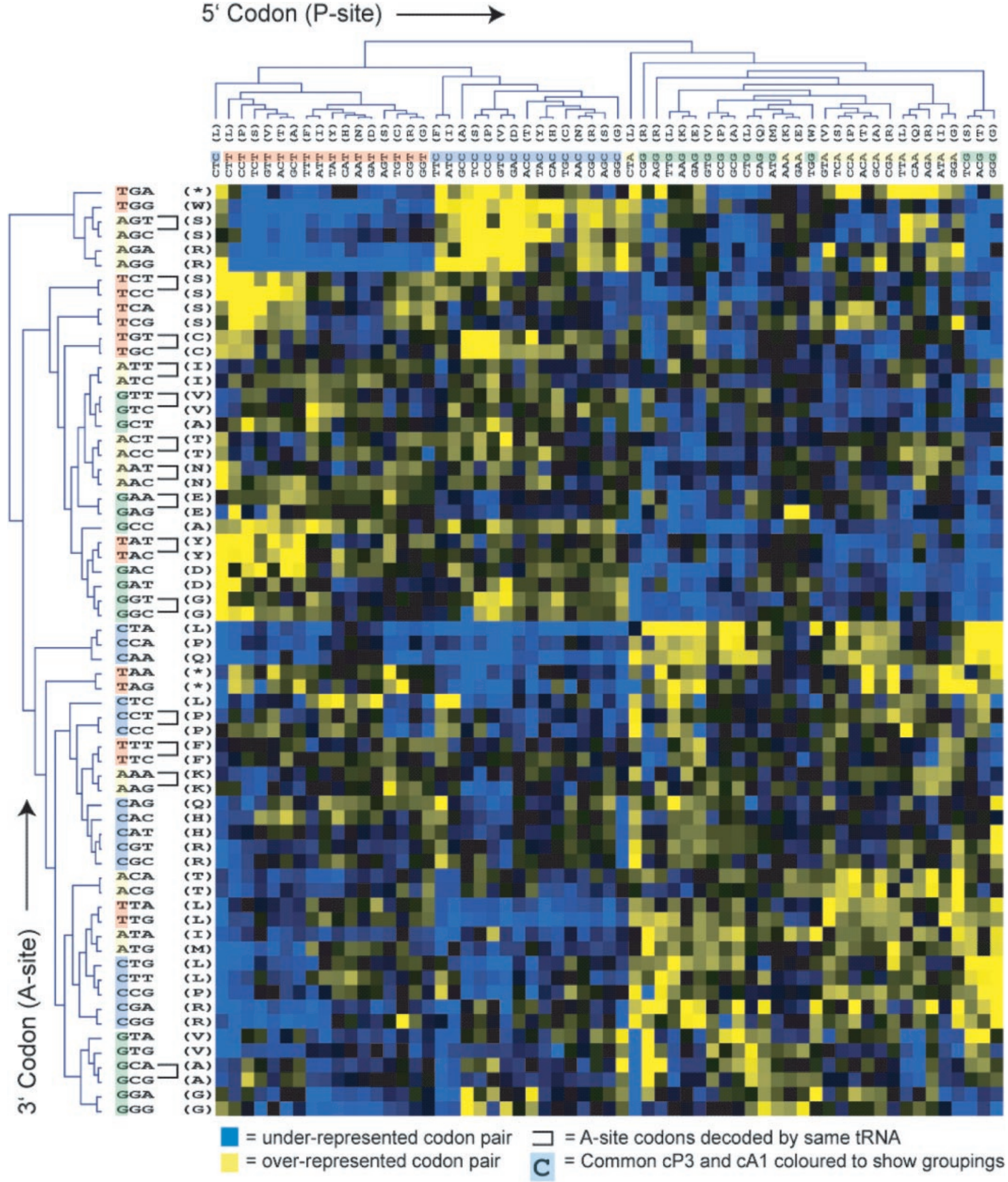


Figure 2. Codon pair residual values for *E.coli* were represented on a 61 by 64 colour grid 5′, P-site codons occupy the horizontal axis and 3′, A-site codons the vertical axis. Each colour pixel represents a codon pair residual value. Over-represented codon pairs are represented in yellow, under-represented values in blue. Colour intensity range represents the full span of residual values. Average linkage clustering of codon pair residual values was used to group codon pairs according to their similarity, producing a dendrogram on each axis. Clustering was carried out on the P-site codons based on their similarity of pair preferences for 3′, A-site codons, and *vice versa*. Where groups of two A-site codons decoded by a single isoacceptor tRNA (mono-isoacceptor groups; MIGs) are clustered at the extremities of the tree (i.e. most similar to each other), they are linked by ‘U’-shaped bars (see text for details).

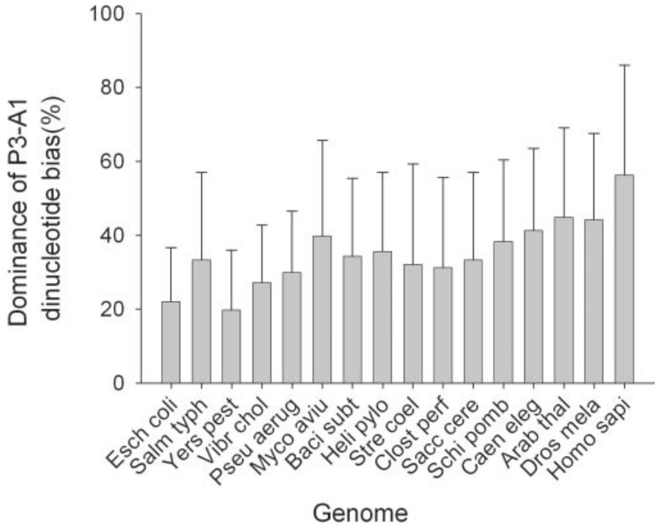
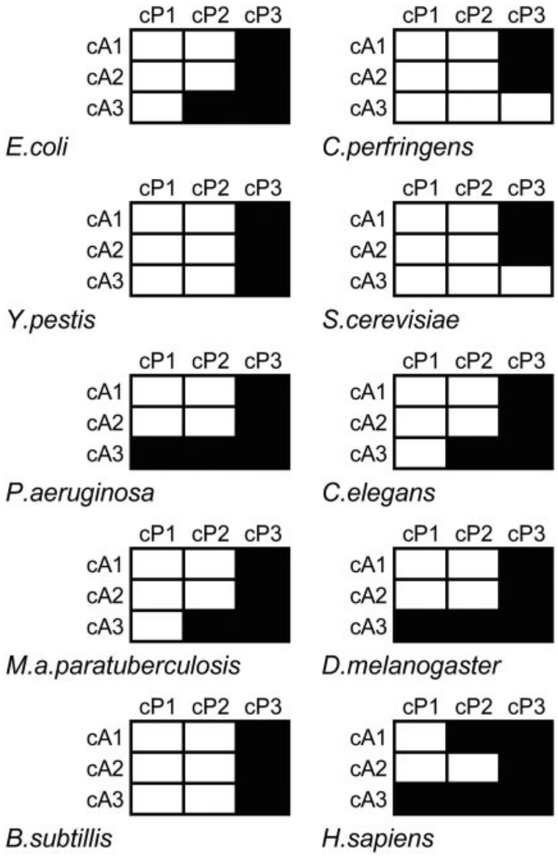


Figure 3. Dinucleotide bias at codon–codon junctions is not a dominant force shaping codon pair bias. Codon pair residuals in a range of genomes were grouped into 16 sets defined by the identity of the cP3-cA1 dinucleotide at the codon–codon junction. For each set, the ratio of under-represented: over-represented codon pairs was assessed and converted to an index representing the uniformity of residual value polarities for codon pairs sharing cP3-cA1 identity. The bar chart shows the average cP3-cA1 dinucleotide bias index for each genome. Error bars represent +/− 1 standard deviation (n = 16). Standard species designations were used (see Figure 1).

the cP2 nucleotide in all multi-cellular eukaryotes and all bacterial species analysed (except *Y.pestis, B.subtilis* and *C.perfringens*), and an interaction with the cP1 nucleotide in *P.aeruginosa, A.thaliana, D.melanogaster* and *H.sapiens* (Figure 4 and data not shown). This is clear evidence that a variety of statistically significant interactions between individual pairs of P- and A-site codon nucleotides underlie codon pair bias.

Codon pair bias is driven by tRNA identity in many microorganisms

Translation is already well known to exert a powerful selective force on ORF composition in the form of codon bias. It is also known that codon pair choice can affect translational rate and fidelity *in vivo* (27,28). However, there has been no direct demonstration to date that translation-based selection shapes codon pair bias. A careful inspection of the cluster analyses of *E.coli* (Figure 2) revealed that A-site codons decoded by the same tRNA (defined here as mono-isoacceptor groups; MIGs) are often found clustered together at the extremities of the cluster tree. The same was found for a number of other microorganisms [gamma proteobacteria and *B.subtilis* (Supplementary Figures S1–S6), *S.cerevisiae* and *S.pombe* (Supplementary Figures S10 and S11)]. For example, the AAC and AAT codons, encoding asparagine and decoded by a single tRNA, comprise a MIG found clustered at the extremity of the similarity tree in *E.coli* (Figure 2). In other words, among all 64 codons, sets of MIG codons are most similar to one another in terms of P-site pairing preference. In *E.coli*, 14 such examples of MIG clustering out of a total possible number of 20 are seen. Average bootstrap values of 80% for MIG codon associations indicate a high degree of confidence in these associations. In higher eukaryotes and some bacterial species (data not shown), very low numbers of MIG codons were clustered, implying that in those genomes translational selection of codon pair bias was not a dominant selective force. P-site MIG clustering was not seen in any organism.



E.coli

	cP1	cP2	cP3
cA1	□	□	■
cA2	□	□	■
cA3	□	■	■

C.perfringens

	cP1	cP2	cP3
cA1	□	□	■
cA2	□	□	■
cA3	□	□	□

Y.pestis

	cP1	cP2	cP3
cA1	□	□	■
cA2	□	□	■
cA3	□	□	■

S.cerevisiae

	cP1	cP2	cP3
cA1	□	□	■
cA2	□	□	■
cA3	□	□	□

P.aeruginosa

	cP1	cP2	cP3
cA1	□	□	■
cA2	□	□	■
cA3	■	■	■

C.elegans

	cP1	cP2	cP3
cA1	□	□	■
cA2	□	□	■
cA3	□	■	■

M.a.paratuberculosis

	cP1	cP2	cP3
cA1	□	□	■
cA2	□	□	■
cA3	□	■	■

D.melanogaster

	cP1	cP2	cP3
cA1	□	□	■
cA2	□	□	■
cA3	■	■	■

B.subtillis

	cP1	cP2	cP3
cA1	□	□	■
cA2	□	□	■
cA3	□	□	■

H.sapiens

	cP1	cP2	cP3
cA1	□	■	■
cA2	□	□	■
cA3	■	■	■

Figure 4. Codon pair preference is directed by combinations of nucleotides spanning adjacent codons. Codon pair residuals in a range of genomes were organized and grouped according to the identities of nucleotide couples composed of one P-site nucleotide, and one A-site nucleotide (e.g. cP1-cA1 or cP2-cA3). Within each of the nine dinucleotide-organized groups, observed and expected codon pair counts were used to calculate χ^2 values for all 16 nt pair combinations. These were summed and the significance of the $\sum\chi^2$ value recorded. For a range of organisms, black grid cells indicate which of the 9 nt couple frequencies differed significantly from that expected (P = 0.001).

To determine if the number of MIG codons found clustered was significant, numbers of MIG pairs were assessed using a probability distribution of chance codon clustering on the A-site axis (Materials and Methods). The analysis clearly showed that in all of the gamma proteobacteria, *B.subtilis* and the unicellular eukaryotes, association of the MIG codons on the A-site axis was significant at least at the 2.5% level (Figure 5A). This analysis demonstrates clearly that in combination with the P-site third nucleotide cP3, elements of the A-site tRNA structure must be strong determinants of codon pairing preference.

It is possible that MIG codon pairing could arise simply due to an mRNA-based effect of the first two nucleotides in the A-site interacting together with the cP3 nucleotide in the P-site, and not because of any property of the tRNA decoding the A-site codon. However, as evidenced by the cluster

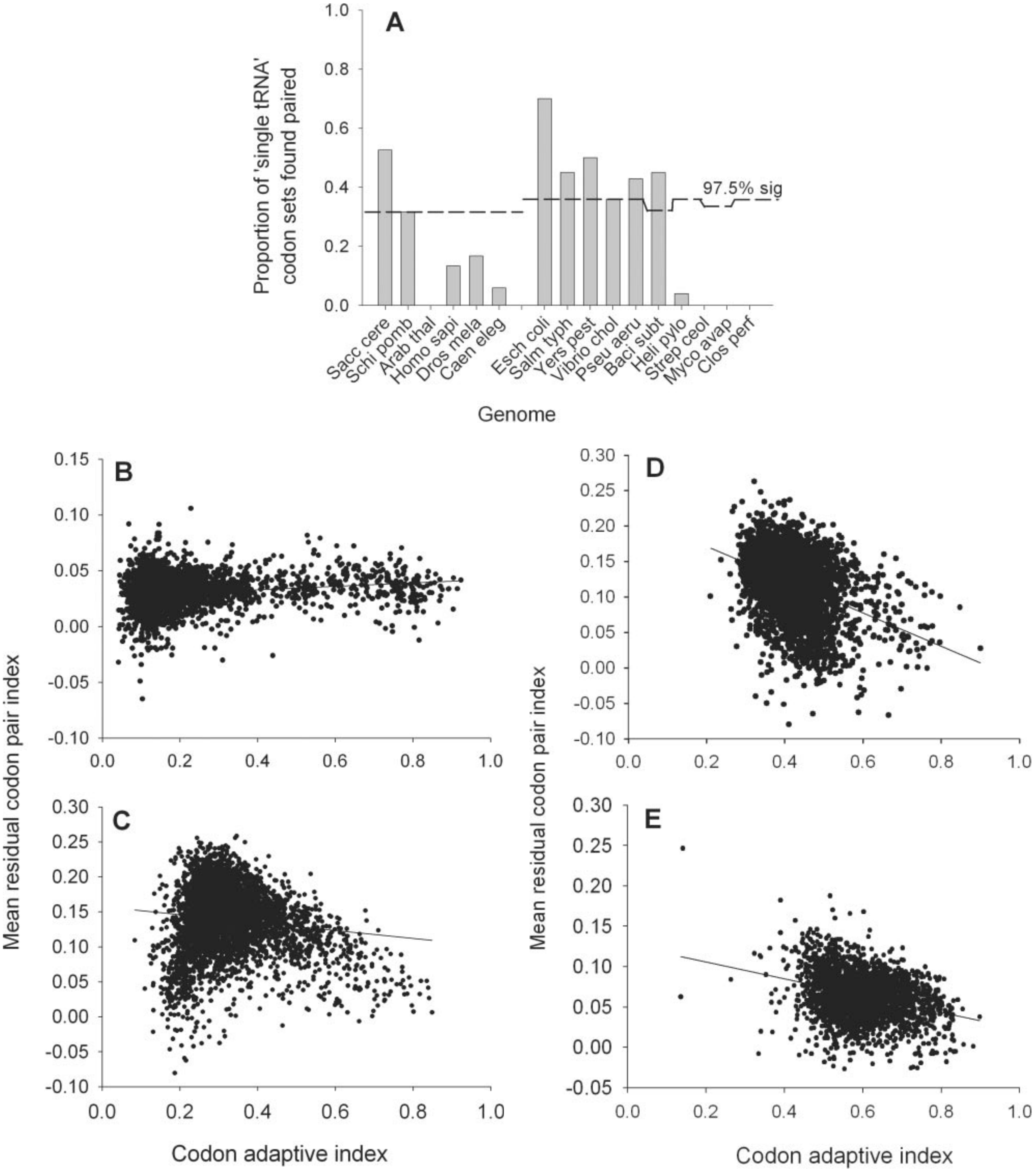


Figure 5. Codon pair preference is tRNA mediated in some genomes, but is poorly correlated with codon bias. The significance of A-site MIG codon clustering was statistically assessed (Materials and Methods). (**A**) For each organism, the proportion of all MIG codon groups (out of a total of between 20 and 25 depending on the tRNA isoacceptor complement for that species) found paired at tree extremities is represented in the bar chart. The dashed line represents the 2.5% confidence level for MIG codon associations assessed using a probability distribution of simulated pairings (Materials and Methods). (**B–E**) The mean codon pair index value was calculated for each ORF in a range of genomes (Materials and methods), and plotted against the codon adaptive index for that ORF. A linear regression line was fitted to the dot plot using the computer program SigmaPlot (Systat software Inc.). (B) *S.cerevisiae*; (C) *E.coli*; (D) *B.subtilis*; (E) *C.perfringens*.

analyses of both *E.coli*, this is obviously not the case. While MIG codons cluster in pairs on the A-site axis (Figure 2), they frequently exhibit widely different pairing patterns to other A-site codons which share the same first two nucleotides both in *E.coli* (e.g. compare GTT/C and GTA/G, GGT/C and GGA/G, ACT/C and ACA/ACG). The clear inference is that codon pairing is strongly influenced both by the identity of all three A-site codon nucleotides, and by some property of the A-site decoding tRNA. In all organisms where evidence for tRNA-mediated selection was obtained, the P-site axis was strongly organized by the cP3 nucleotide (Figure 2A and B). It is thus clear from analysis of organisms that demonstrate significant A-site MIG codon clustering that interplay between the cP3 nucleotide and the A-site tRNA is a governing influence on pair preference selection.

The codon pair bias of a gene correlates only weakly with its codon bias

While many studies have stated that overall codon bias correlates positively with gene expression (16,35–37), it is unclear how codon pair bias might affect gene expression. However, some studies in *E.coli* record codon pair bias to be negatively correlated with gene expression (23,25). In this study, the assumption that codon pair bias affects gene expression was re-examined on a genomic level for a variety of organisms by individually calculating the average codon pair bias for every ORF within a genome. These values were then plotted against the codon adaptive index (CAI) for each gene (Figure 5B–E). The genomes of *S.cerevisiae, E.coli* and *B.subtilis*, (Figure 5B–D, respectively) have been identified by this study as genomes within which codon pair bias is modulated by tRNA-based translational selection (Figure 5A). *C.perfringens* was chosen as a fast-growing microorganism that has a highly codon-biased genome (Figure 5) (38), but for which no translational selection of pair bias was detected in this study. However, for all the bacteria examined, only very weak negative correlations between codon pair bias and CAI were detected. A predominantly neutral relationship was detected for yeast (Figure 5B). Although the work of Gutman and Hatfield implied a strong negative correlation between codon pair index and gene expression in *E.coli*, this conclusion was limited by the availability of sequence at the time and by their chance selection of examined genes. In summary, in all genomes tested codon pair bias showed very weak correlation with codon bias.

Prokaryote genomes over-use, but eukaryote genomes generally under-use, pairs of rare codons

Codon pair bias was significant in all genomes examined (Figure 1), although the precise reasons underlying this bias had not been defined. A closer examination was thus made of which codon pairs were being over or under-used, based on their expected frequency. For each organism, codon pairs were ordered and binned (10 equal bins) according to their expected frequency. Within each bin, the mean residual value was determined and plotted (Figure 6A). The data were additionally smoothed by using two bins, each representing half the dataset (the 50% most rare, 50% commonest codon pairs; Figure 6B).

The analysis for prokaryotes such as *E.coli* and *M.avium paratuberculosis* shows that pairs of rare codons are over-used, whereas pairs of common codons are under-used (Figure 6A and data not shown). This trend is reversed in eukaryotes such as *H.sapiens* (Figure 6A). A notable exception to this is *D.melanogaster*, which like prokaryotes, also over-uses pairs of rare codons (Figure 6A). When the data for all organisms is compared, this time separated into two 50% bins, a clear demarcation between eukaryotes and prokaryotes is seen in terms of rare codon pair usage (Figure 6B); prokaryotes over-use such pairs whereas they are selected against (*D.melanogaster* excepted) in eukaryotes.

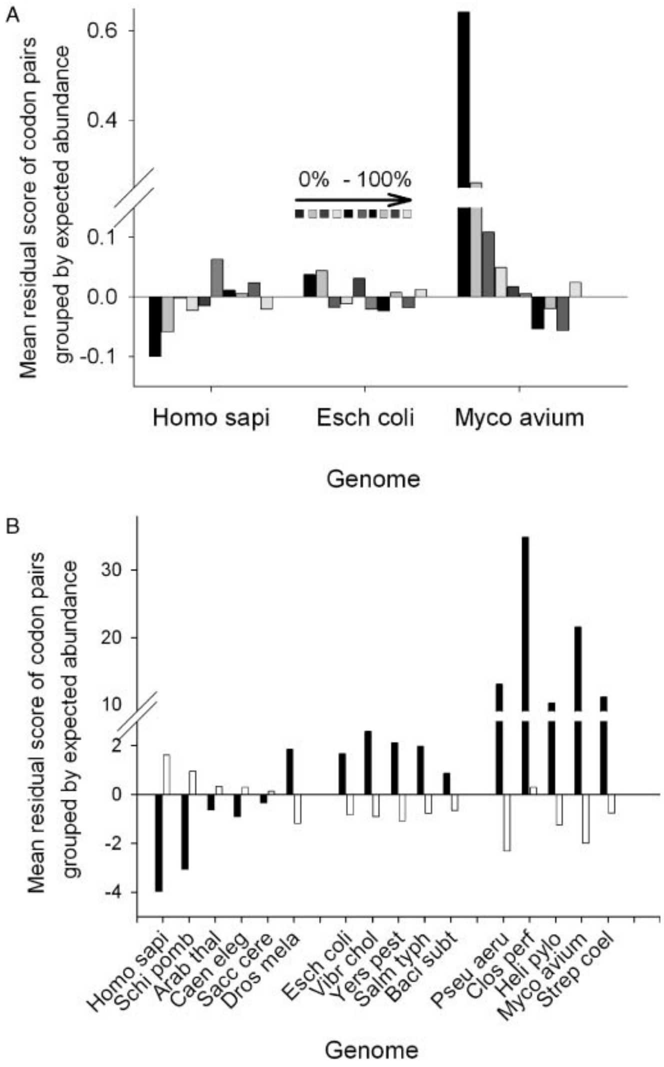


Figure 6. Prokaryote and eukaryote genomes are distinguished by distinct patterns of codon pair usage. For all genomes tested, the codon pair residual values ($o - e_{\mathrm{nor}}/e_{\mathrm{nor}}$) were tabulated, ordered by expected frequency, and separated into 10 bins. For each bin, the mean residual value was calculated. (**A**) Mean residual values plotted for each of the 10 bins (0–10% bin is the left-most bar in each group of ten). (**B**) Residuals were further smoothed into two equal bins before averaging, one bin containing the expected 50% least abundant codon pairs (black bars), the other the expected 50% most abundant codon pairs (white bars).

Multivariate analysis suggests A-site tRNA structural features impact upon codon pairing

Having shown that codon pairing in certain organisms is affected by translational selection via A-site tRNA identity (Figure 5A), it was important to identify features of the A-site tRNA that influence codon pairing preferences. Multivariate statistical analysis was employed to predict which nucleotide positions and identities within A-site tRNA sequence were strong determinants of the codon pair usage residual values. A PLS analysis (33) was conducted on a subset of organisms for which codon pair frequencies appeared to be related to tRNA identity (*E.coli*, *P.aeruginosa* and *B.subtilis*; Figure 5A). The residual score of every codon pair was treated as the dependent variable, and the identity of nucleotides at all positions within the P- and A-site tRNAs, together with the codon nucleotides themselves, were treated as qualitative independent variables (i.e. potential predictors of the residual value). Prior to analysis, codon pair and tRNA sequence datasets were split into four sections according to the cP3 nucleotide identity, and a model developed for each. This helped improve both the predictive power of the models generated and the percentage of explained residual variation, since cP3 nucleotide identity is undoubtedly a key contributor to codon pair preferences (Figures 2 and 4). For all three genomes analysed, average predictive power was 42% and average explained variation was 33%. Although seemingly low, these values are constrained by a modelling process based upon purely qualitative *x* variables and are nevertheless consistent with a good quality model using qualitative *x* data.

In all three organisms analysed in this way, nucleotide identity at many positions across the A-site tRNA, and at all three positions within the A-site codon was found to be an important contributor to codon pairing patterns. An example dataset is shown for illustrative purposes in the form of a weights plot (Materials and Methods) from *E.coli* where cP3 = G. The points are split between two plots for ease of interpretation (Figure 7A and B). On these plots, points representing tRNA and codon nucleotides that are close to the residual value plot position are associated with positive codon pair residual values. In contrast, those at the opposite end of a line bisecting the plot origin and residual value point are associated with negative codon pair residual values (Figure 7A and B). Points near the origin exert minimum influence on the residual value. Many positions on the A-site tRNA, depending on their nucleotide identity, contributed positively to codon pair residuals (Figure 7B). Clearly, the identity of a number of nucleotides located in different regions of the A-site tRNA is influential in governing codon pair preference. Nucleotide positions that were both positively and negatively influential on the residual value in *E.coli* were also important predictors in *P.aeruginosa* and *B.subtilis*, indicating a degree of conservation of the structural effect of tRNA sequence on codon pair preference. Influential nucleotide positions for all three organisms identified from the weights plots were mapped onto the cloverleaf structure of tRNA (Figure 7C). The analysis shows that the A-site tRNA anticodon nucleotides were, in the case of all three organisms, an important predictor of codon pairing (as would be expected). Also important in all three organisms were complementary positions within the acceptor stem helix

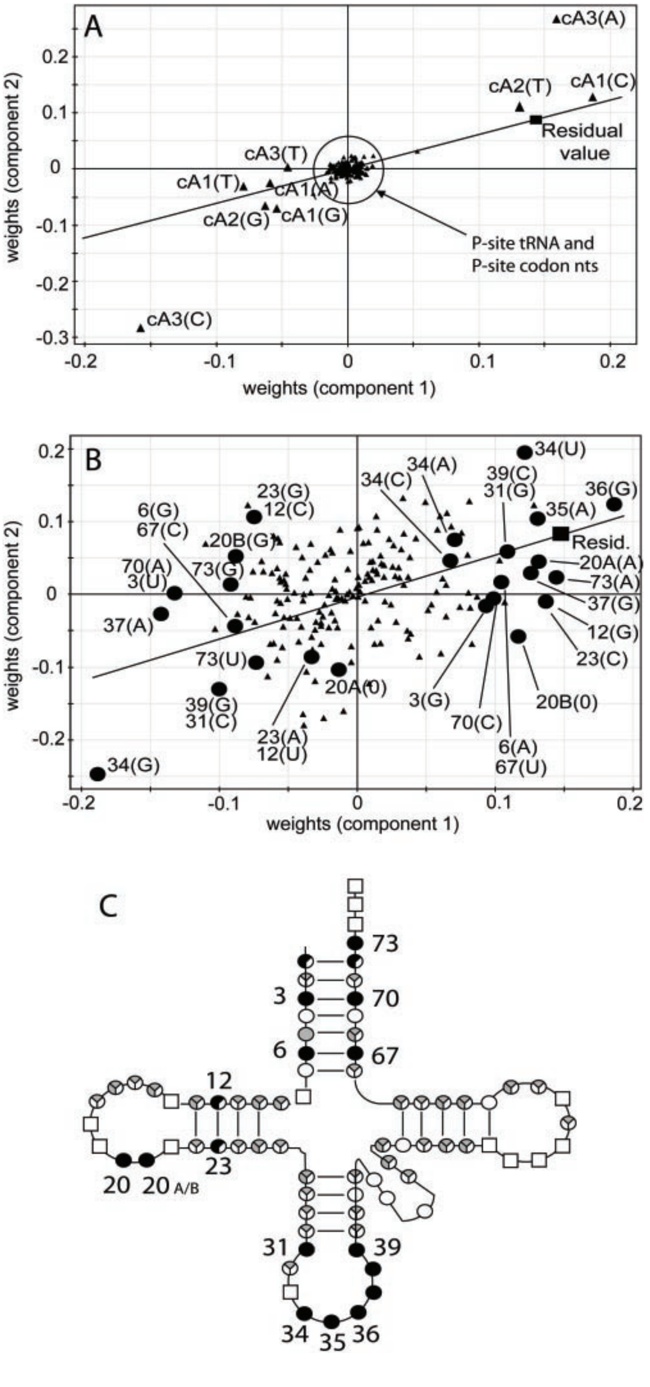


Figure 7. Multivariate analysis of the tRNA sequence influence on codon pairing preference. PLS analysis was used to identify nucleotides in P and A-site tRNAs that were good predictors of codon pair residual values. Representative data (data subset in which cP3 = G) from the *E.coli* analysis is presented, split between panels A and B for ease of interpretation. The weights plots (A and B) shown report the quantitative relationship between the *x* predictor variables (tRNA nucleotides) and *y* dependent variable (codon pair residual), plotted for the two components used to model the data. On these plots, tRNA and codon nucleotides plotted close to the residual value plot position (filled square) are typically those associated with over-represented codon pairs. Conversely, those at the opposite end of a line that bisects the plot origin and residual value point are typically those associated with under-represented codon pairs. (**A**) PLS weights plot showing P- and A-site codon nucleotides together with P-site tRNA nucleotides (filled triangles) plotted versus the codon pair residual value (filled square). (**B**) A-site tRNA nucleotides (filled triangles) plotted versus the codon pair residual value (filled square). Filled circle symbols represent key A-site tRNA nucleotide positions indicated by the PLS analysis to be significant predictors of the residual value across three bacterial species tested. Influential A-site tRNA nucleotide positions (filled circles) are labelled with the standard cloverleaf model nucleotide position, and the nucleotide identity. (**C**) A cloverleaf model of the basic tRNA structure, indicating those positions on the tRNA that were identified as good predictors of the codon pair residual value. Positions identified as important for residual prediction in *E.coli*, *P.aeruginosa* and *B.subtilis* are indicated by filled circles, those important in either two or just one of the three organisms, as 2/3 and 1/3 filled circles, respectively. Sector shadings indicate positions where >40% (black) or <40% (grey) of nucleotide identities at a given position were influential, averaged across the three species. Open squares represent invariant nucleotides.

(Figure 7C, nt 3–70, 6–67), and nucleotides immediately 3′ of the anticodon, which face the closely adjacent P-site tRNA anticodon within the ribosome (Figure 7C, nt 37–39).

Previously, anticodon nucleotides (cP3 in combination with cA1/cA2) were shown to be influential on codon pair residual values (Figure 4). It was therefore possible that the identification of other tRNA nucleotides (3, 6, 31, 38, 39 and 73) by the PLS analysis was a consequence of base identity of these nucleotides being tied to the base identity of (one of) the anticodon nucleotides. We therefore checked whether those nucleotides, indicated as important codon pair residual determinants (Figure 7, A3, G6, C31, G38, C39, C69, A73), were consistently associated with any single anticodon base type at positions 34, 35 or 36. None were tied in this way, confirming that these influential positions are modulating codon pair preference independently of the effect exerted by the anticodon itself (data not shown). This shows that the anticodon sequence is not the sole determinant of codon pair preference.

All positions on the P-site tRNA were located at the origin of the weights plot, apparently indicating a minimal influence on the residual value (Figure 7A). This was slightly surprising as P-site tRNA identity can affect the accuracy of A-site decoding events. For example, stop codon readthrough in *E.coli* is affected by P-site tRNA mutations within the 5′ half of the anticodon loop (29,39). The PLS modelling exercise was therefore repeated in *E.coli*, this time organizing the dataset into four sets grouped by cA1 nucleotide identity, rather than cP3 as previously. These results revealed that the P-site tRNA nucleotides were to some degree predictive of codon pair residual value. However, the explained variation (9.5%) and predictive power (16%) of the models generated was poor for three of the four cA1-grouped datasets, and the models could not therefore be used to unequivocally identify important P-site tRNA nucleotide positions. Other approaches will be required to address this question.

In conclusion, the identity of nucleotides at a number of positions across A-site tRNAs, help predict codon pair residual values grouped according to cP3 nucleotide identity. Nucleotide positions within the anticodon loop and acceptor stems of A-site tRNA are conserved predictors of codon pairing in *E.coli*, *P.aeruginosa* and *B.subtilis*, and identify potential mediators of tRNA-codon interactions between P and A-sites.

DISCUSSION

Codon pair usage within ORFs is known to be biased in *E.coli*, *S.cerevisiae* and *Candida albicans* (23,25,26). Additionally, codon context can affect both the speed and fidelity of ribosomal decoding (17,27,29,40). In this study, a wide-ranging survey was undertaken across a range of prokaryote and eukaryote species. For the first time, codon pair bias patterns were analysed to identify the selective forces driving codon pair bias. The research revealed that after removing the effects of dipeptide bias and codon bias from the dataset, codon pair bias was significant and universal, suggesting additional selective forces were at work (Figure 1). Analysis of the complex interacting forces driving codon pair selection revealed for the first time that clear rule-sets govern codon context in most of the genomes examined. In all eukaryotes, P-site codons were grouped by cP3 nucleotide identity, whereas A-site codons were grouped by cA1 nucleotide identity (Figure 2B and Supplementary Figures S10–S15). For most prokaryotes, including the gamma proteobacteria, the pattern was similar, although cA1 grouping for A-site codons was not as marked (Figure 2A and Supplementary Figures S1–S9).

Analysis of codon pair usage revealed that pairs of rare codons were as expected generally under-used in eukaryotes, but were in fact generally over-used in prokaryotes (Figure 6A and B). For reasons that are not clear, the only exception to this rule was *D.melanogaster*, which exhibited like prokaryotes over-used rare codon pairs (Figure 6B). It is known that the introduction of rare codons, or pairs of rare codons, into an ORF slows down translation of the mRNA (27). The over-representation of such pairs thus seems counter-intuitive; selection might be expected to favour codon pairs that permit rapid translation. Over-representation of rare codon pairs is possibly a mechanism to fine-tune translation rates across an mRNA, perhaps to aid correct folding of nascent polypeptide chains (41,42). However, studies of bacterial starvation responses could also explain why prokaryotes, without exception, over-use rare codon pairs. The RelE ribonuclease functions to cleave mRNAs within the A-site of stalled ribosomes (43). The ribosomal A-site is then bound by tmRNA, a tRNA-mRNA hybrid. So-called trans-translation of the tmRNA template adds a C-terminal tag to the partly completed nascent peptide, targeting the uncompleted protein for degradation (44,45). It is known that tmRNA activity in *E.coli* is stimulated by pairs of rare AGA arginine codons during logarithmic growth, but not by single, isolated AGA codons (46). Co-evolution of ORFs with the tmRNA starvation response system may thus explain the over-representation of rare codon pairs in the prokaryote genomes examined here, all of which do have a tmDNA gene (47).

The over-use of rare codon pairs thus shapes the weak negative correlation between mean codon pair index and codon adaptive index found for all three prokaryote genomes examined (Figure 5C–E). Highly expressed genes (e.g. ribosomal proteins, glycolytic enzymes) that are very biased in their codon usage should generally avoid over-represented codon pairs, since these are generally pairs of otherwise rare codons that are translated more slowly [(23,27), and this study; Figure 4C–E]. Conversely, because eukaryote genomes tend to under-use rare codon pairs while over-using pairs of common codons, a weak positive correlation with CAI would be expected, as indeed was found for *S.cerevisiae* (Figure 5B). The overall weak correlation between codon pair index and CAI in all four genomes tested, and its variable polarity, hints that selection optimizing translational fidelity, rather than gene expression level, may drive overall codon pair bias.

cP3-cA1 dinucleotide identity at codon–codon junctions could in theory be determined by a genome-wide preference for particular dinucleotides. In the human genome as a whole, the CG dinucleotide, a site for cytosine methylation, is present 5- to 10-fold less often than expected (34). However, we show here that other selective forces, in addition to a probable effect of DNA dinucleotide bias, combine to define codon pairing patterns. Dinucleotide bias at codon junctions, as a sole selective pressure, would generate either under-represented or over-represented codon pairs sharing cP3-cA1 dinucleotide identity. Consistent polarity of bias was not seen for groups of codon pairs sharing a particular cP3-cA1 dinucleotide (Figure 3). In addition to this, in all genomic ORF sets, the cP3-cA2 nucleotide couple (and in most cases cP3-cA3) was subject to significant bias (Figure 4). This strongly argues that although the cP3 nucleotide plays a cardinal role in determining pairing frequencies, it does so through an interaction with more than one A-site nucleotide. The cA3 position in the A-site codon is also often involved in interactions with cP2 nucleotide and sometimes the cP1 nucleotide (Figure 4). Overall then, codon pairing patterns result partly from a complex interplay between multiple nucleotide couples spanning the two adjacent codons.

Previous experimental data suggest that translation may act as a selective force on codon pair choice within ORFs for reasons of translational efficiency attained through optimal fit of tRNAs within the ribosomal A and P-sites (27,29). In this study, evidence was obtained for tRNA-mediated selection upon codon pairing in unicellular eukaryotes, *B.subtilis* and the gamma proteobacteria. In the A-site, mono-isoacceptor group (MIG) codons showed significantly similar patterns of P-site codon pairing preference (Figure 5A). This most likely reflects an important property of decoding A-site tRNAs, acting as a selective force on P-site codon pairing preference. Whilst A-site MIG codons cluster neatly, they typically exhibit very different P-site pairing preferences to other A-site codons sharing cA1-cA2 nucleotide identity (Figure 2A and B). Additionally, A-site codons sharing cA1 identity do not consistently cluster (Figure 2A and B). This argues strongly that codon pair bias has a component dictated by tRNA properties, rather than simply by codon properties. While the influence of tRNA properties on codon pair bias implies a possible effect of tRNA structure on either translational rate or fidelity, it is unclear why the other organisms tested, including *C.perfringens* [capable of very rapid growth and with a highly codon-biased genome (38)], do not apparently exhibit a tRNA-based selection of codon pair frequencies.

The role of A-site tRNA structure in driving codon pairing was further analysed using multivariate analysis of tRNA sequences. For *E.coli*, *P.aeruginosa* and *B.subtilis*, a variety of positions on A-site tRNA were identified as contributory predictors of codon pair pairing residuals, particularly those immediately 3′ of the anticodon (directly adjacent to the P-site anticodon), and complimentary nucleotide pairs in the acceptor stem. The crystal structure of the *Thermus thermophilus* ribosome complexed with mRNA and tRNAs in A-, P- and E-sites shows how tRNAs are juxtaposed (48). The 3′ side of the A-site anticodon loop and the A-site acceptor stem are those parts of the tRNA closest to the P-site tRNA. It is possible that nucleotide identity and/or chemical modifications at these points may influence tRNA positioning. Such interference could perhaps occur either during A-site tRNA delivery or during the peptidyl-transfer and translocation steps of the elongation cycle. It is known that chemical modifications of A-site tRNA nucleotides, particularly in the anticodon loop 3′ of the anticodon, can play a key role in reading frame maintenance (49). Changes to these nucleotide modifications can significantly alter rates of aminoacyl tRNA selection and increase +1 frameshifting (3,50,51). The codon pair bias data presented here also indicates a role for the 3′ side of the anticodon loop of the A-site tRNA in optimizing translation in some way.

It must be recognized that codon pair bias is subject to a variety of selective forces, which include codon bias, dipeptide bias and dinucleotide bias, and potentially, forces that exclude transcriptional regulatory signals from coding sequences. This work has additionally identified bias resulting from interplay between nucleotides of the codon–anticodon interaction in P- and A-sites, principally between cP3 and cA1/cA2. Furthermore, there is an additional influence on codon pair preference exerted by specific regions of the tRNA in many microorganisms. On a macro scale, the relative over-representation of rare and common codon pairs may be explicable by adaptations to amino acid starvation and/or nascent peptide folding. At the ‘micro’ level of specific nucleotide interactions within a given codon pair, it is less clear how selection operates. A central unifying theory is supported by some of the data in this work, showing a conserved interplay between P-site nucleotides and those of the A-site (Figure 4). However, while there is a clear indication of, for instance, cP3-cA2 bias, the identities of specific cP3-cA2 nucleotide couples selected for and against differ between species (data not shown). Clearly, species-specific factors influence codon pairing: for some species, there is evidence these factors are likely to include tRNA structural effects (Figure 7). Although speculative, it is possible that the interplay between P-site codon nucleotides and those of the A-site codon may influence the geometry of the 45° mRNA kink between P- and A-sites. This in turn could influence optimal juxtaposition of tRNAs during decoding and positioning of their acceptor stems in the peptidyl-transferase centre (48). Correlations between specific tRNA sequence elements and codon pair bias argue for a cardinal role for P-site wobble position codon–anticodon interaction, and the A-site anticodon loop and acceptor stem (Figure 7).

In summary, codon pair bias has been shown to be a universal phenomenon that results from a balance of competing selective forces. This study successfully identifies a number of selective forces operating on codon pair bias. The evidence is clear that in a number of organisms, translation optimization is one such force, strengthening evidence from other *in vivo* pilot studies that fidelity and translational step times are influenced by codon pairing. Experimental testing of our conclusions should further understanding of the role of codon pair preferences in gene expression.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Baukje De Roos (Rowett Research Institute, Aberdeen) for helpful discussions. This study was funded by a research studentship grant from the Biotechnology and Biological Sciences Research Council, and by a Tenovus Scotland Research Grant, both to I.S. Funding to pay the Open Access publication charges for this article was provided by the University of Aberdeen.

Conflict of interest statement. None declared.

REFERENCES

1. Maaloe,O. and Kjeldgaard,N.O. (1966) *Control of macromolecular synthesis*. W.A. Benjamin, Inc., NY.
2. Pedersen,S. (1984) *Escherichia coli* ribosomes translate *in vivo* with variable rate. *EMBO J.*, **3**, 2895–2898.
3. Kruger,M.K., Pedersen,S., Hagervall,T.G. and Sorensen,M.A. (1998) The modification of the wobble base of tRNAGlu modulates the translation rate of glutamic acid codons in vivo. *J. Mol. Biol.*, **284**, 621–631.
4. Boe,L. (1992) Translational errors as the cause of mutations in *Escherichia coli*. *Mol. Gen. Genet.*, **231**, 469–471.
5. Toth,M.J., Murgola,E.J. and Schimmel,P. (1988) Evidence for a unique first position codon-anticodon mismatch in vivo. *J. Mol. Biol.*, **201**, 451–454.
6. Loftfield,R.B. and Vanderjagt,D. (1972) The frequency of errors in protein biosynthesis. *Biochem. J.*, **128**, 1353–1356.
7. Stansfield,I., Jones,K.M., Herbert,P., Lewendon,A., Shaw,W.V. and Tuite,M.F. (1998) Missense translation errors in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **282**, 13–24.
8. Farabaugh,P.J. (1996) Programmed translational frameshifting. *Microbiol. Rev.*, **60**, 103–134.
9. Gallant,J.A. and Lindsley,D. (1998) Ribosomes can slide over and beyond 'hungry' codons, resuming protein chain elongation many nucleotides downstream. *Proc. Natl Acad. Sci. USA*, **95**, 13771–13776.
10. Jorgensen,F., Adamski,F.M., Tate,W.P. and Kurland,C.G. (1993) Release factor-dependent false stops are infrequent in *Escherichia coli*. *J. Mol. Biol.*, **230**, 41–50.
11. Bertram,G., Innes,S., Minella,O., Richardson,J. and Stansfield,I. (2001) Endless possibilities: translation termination and stop codon recognition. *Microbiology*, **147**, 255–269.
12. Jacks,T., Power,M.D., Masiarz,F.R., Luciw,P.A., Barr,P.J. and Varmus,H.E. (1988) Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*, **331**, 280–283.
13. Shah,A.A., Giddings,M.C., Parvaz,J.B., Gesteland,R.F., Atkins,J.F. and Ivanov,I.P. (2002) Computational identification of putative programmed translational frameshift sites. *Bioinformatics*, **18**, 1046–1053.
14. Williams,I., Richardson,J., Starkey,A. and Stansfield,I. (2004) Genome-wide prediction of stop codon readthrough during translation in the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **32**, 6605–6616.
15. Grantham,R., Gautier,C., Gouy,M., Mercier,R. and Pave,A. (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.*, **8**, r49–r62.
16. Berg,O.G. and Kurland,C.G. (1997) Growth rate-optimised tRNA abundance and codon usage. *J. Mol. Biol.*, **270**, 544–550.
17. Robinson,M., Lilley,R., Little,S., Emtage,J.S., Yarranton,G., Stephens,P., Millican,A., Eaton,M. and Humphreys,G. (1984) Codon usage can affect efficiency of translation of genes in *Escherichia coli*. *Nucleic Acids Res.*, **12**, 6663–6671.
18. Sorensen,M.A., Kurland,C.G. and Pedersen,S. (1989) Codon usage determines translation rate in *Escherichia coli*. *J. Mol. Biol.*, **207**, 365–377.
19. McNulty,D.E., Claffee,B.A., Huddleston,M.J. and Kane,J.F. (2003) Mistranslational errors associated with the rare arginine codon CGG in *Escherichia coli*. *Protein Expr. Purif.*, **27**, 365–374.
20. Curran,J.F. and Yarus,M. (1989) Rates of aminoacyl-tRNA selection at 29 sense codons *in vivo*. *J. Mol. Biol.*, **209**, 65–77.
21. Elf,J., Nilsson,D., Tenson,T. and Ehrenberg,M. (2003) Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science*, **300**, 1718–1722.
22. Dittmar,K.A., Sorensen,M.A., Elf,J., Ehrenberg,M. and Pan,T. (2005) Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Rep.*, **6**, 151–157.
23. Gutman,G.A. and Hatfield,G.W. (1989) Nonrandom utilization of codon pairs in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **86**, 3699–3703.
24. Yarus,M. and Folley,L.S. (1985) Sense codons are found in specific contexts. *J. Mol. Biol.*, **182**, 529–540.
25. Boycheva,S., Chkodrov,G. and Ivanov,I. (2003) Codon pairs in the genome of *Escherichia coli*. *Bioinformatics*, **19**, 987–998.
26. Moura,G., Pinheiro,M., Silva,R., Miranda,I., Afreixo,V., Dias,G., Freitas,A., Oliveira,J.L. and Santos,M.A. (2005) Comparative context analysis of codon pairs on an ORFeome scale. *Genome Biol.*, **6**, R28.
27. Irwin,B., Heck,J.D. and Hatfield,G.W. (1995) Codon pair utilization biases influence translational elongation step times. *J. Biol. Chem.*, **270**, 22801–22806.
28. Folley,L.S. and Yarus,M. (1989) Codon contexts from weakly expressed genes reduce expression *in vivo*. *J. Mol. Biol.*, **209**, 359–378.
29. Smith,D. and Yarus,M. (1989) tRNA-tRNA interactions within cellular ribosomes. *Proc. Natl Acad. Sci. USA*, **86**, 4397–4401.
30. Curran,J.F., Poole,E.S., Tate,W.P. and Gross,B.L. (1995) Selection of aminoacyl-tRNAs at sense codons: The size of the tRNA variable loop determines whether the immediate 3′ nucleotide to the codon has a context effect. *Nucleic Acids Res.*, **23**, 4104–4108.
31. 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.
32. Saeed,A.I., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M. *et al.* (2003) TM4: A free, open-source system for microarray data management and analysis. *BioTechniques*, **34**, 374–378.
33. Eriksson,L., Johansson,E., Kettameh-Wold,N. and Wold,S. (2001) In Eriksson,L. *et al.* (ed.) PLS. *Multi- and Megavariate Data Analysis: Principles and Applications*. Umetrics, Umea, Sweden, pp. 71–111.
34. Bird,A.P. (1995) Gene number, noise reduction and biological complexity. *Trends Genet.*, **11**, 94–100.
35. Sharp,P.M. and Li,W.H. (1987) The codon adaptation index: a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
36. Bennetzen,J.L. and Hall,B.D. (1982) Codon selection in yeast. *J. Biol. Chem.*, **257**, 3026–3031.
37. Duret,L. (2000) tRNA gene number and codon usage in the *C.elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.*, **16**, 287–289.
38. Sharp,P.M., Bailes,E., Grocock,R.J., Peden,J.F. and Sockett,R.E. (2005) Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.*, **33**, 1141–1153.
39. Raftery,L.A. and Yarus,M. (1987) Systematic alterations in the anticodon arm make tRNA(glu)-suoc a more efficient suppressor. *EMBO J.*, **6**, 1499–1506.
40. Miller,J.H. and Albertini,A.M. (1983) Effects of surrounding sequence on the suppression of nonsense codons. *J. Mol. Biol.*, **164**, 59–71.
41. Purvis,I.J., Bettany,A.J., Santiago,T.C., Coggins,J.R., Duncan,K., Eason,R. and Brown,A.J. (1987) The efficiency of folding of some proteins is increased by controlled rates of translation *in vivo*. A hypothesis. *J. Mol. Biol.*, **193**, 413–417.
42. Rothman,J.E. (1989) Polypeptide chain binding proteins: catalysts of protein folding and related processes in cells. *Cell*, **59**, 591–601.
43. Pedersen,K., Zavialov,A.V., Pavlov,M.Y., Elf,J., Gerdes,K. and Ehrenberg,M. (2003) The bacterial toxin RelE displays codon-specific cleavage of mRNAs in the ribosomal A site. *Cell*, **112**, 131–140.
44. Keiler,K.C., Waller,P.R. and Sauer,R.T. (1996) Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science*, **271**, 990–993.
45. Tu,G.F., Reid,G.E., Zhang,J.G., Moritz,R.L. and Simpson,R.J. (1995) C-terminal extension of truncated recombinant proteins in *Escherichia coli* with a 10Sa RNA decapeptide. *J. Biol. Chem.*, **270**, 9322–9326.
46. Roche,E.D. and Sauer,R.T. (1999) SsrA-mediated peptide tagging caused by rare codons and tRNA scarcity. *EMBO J.*, **18**, 4579–4589.
47. Zwieb,C., Gorodkin,J., Knudsen,B., Burks,J. and Wower,J. (2003) tmRDB (tmRNA database). *Nucleic Acids Res.*, **31**, 446–447.
48. Yusupov,M.M., Yusupova,G.Z., Baucom,A., Lieberman,K., Earnest,T.N., Cate,J.H. and Noller,H.F. (2001) Crystal structure of the ribosome at 5.5 A resolution. *Science*, **292**, 883–896.
49. Urbonavicius,J., Qian,Q., Durand,J.M., Hagervall,T.G. and Bjork,G.R. (2001) Improvement of reading frame maintenance is a common function for several tRNA modifications. *EMBO J.*, **20**, 4863–4873.
50. Urbonavicius,J., Stahl,G., Durand,J.M., Ben Salem,S.N., Qian,Q., Farabaugh,P.J. and Bjork,G.R. (2003) Transfer RNA modifications that alter +1 frameshifting in general fail to affect −1 frameshifting. *RNA*, **9**, 760–768.
51. Li,J., Esberg,B., Curran,J.F. and Bjork,G.R. (1997) Three modified nucleosides present in the anticodon stem and loop influence the *in vivo* aa-tRNA selection in a tRNA-dependent manner. *J. Mol. Biol.*, **271**, 209–221.