

RESEARCH

Open Access

# Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles

Maya Botzman and Hanah Margalit\*

## Abstract

**Background:** It is widely acknowledged that synonymous codons are used unevenly among genes in a genome. In organisms under translational selection, genes encoding highly expressed proteins are enriched with specific codons. This phenomenon, termed codon usage bias, is common to many organisms and has been recognized as influencing cellular fitness. This suggests that the global extent of codon usage bias of an organism might be associated with its phenotypic traits.

**Results:** To test this hypothesis we used a simple measure for assessing the extent of codon bias of an organism, and applied it to hundreds of sequenced prokaryotes. Our analysis revealed a large variability in this measure: there are organisms showing very high degrees of codon usage bias and organisms exhibiting almost no differential use of synonymous codons among different genes. Remarkably, we found that the extent of codon usage bias corresponds to the lifestyle of the organism. Especially, organisms able to live in a wide range of habitats exhibit high extents of codon usage bias, consistent with their need to adapt efficiently to different environments. Pathogenic prokaryotes also demonstrate higher extents of codon usage bias than non-pathogenic prokaryotes, in accord with the multiple environments that many pathogens occupy. Our results show that the previously observed correlation between growth rate and metabolic variability is attributed to their individual associations with codon usage bias.

**Conclusions:** Our results suggest that the extent of codon usage bias of an organism plays a role in the adaptation of prokaryotes to their environments.

## Background

The genetic code is composed of triplets of four nucleotide types for 20 amino acids. This redundancy implies the use of synonymous codons - different codons encoding the same amino acid. Synonymous codons may differ in their frequency of occurrence among different genes within an organism, a phenomenon known as 'codon usage bias' [1]. It was demonstrated that many bacteria and yeast undergo translational selection, with highly expressed genes preferentially using codons assumed to be translated faster and/or more accurately by the ribosome [2,3]. Previous works suggested that these codons are the ones matching abundant tRNAs,

which are organism-specific [3-7]. Other works demonstrated additional factors affecting the frequencies of the synonymous codons in an organism, such as the genome GC content [8,9]. Thus, the preferred codons per amino acid may vary between different organisms, based on their tRNA repertoire and other factors.

While it was claimed that most prokaryotes undergo translational selection [10], it is conceivable that various organisms may differ in the extent of codon usage bias across their genes and in the forces determining it. For several organisms, such as *Escherichia coli* and *Saccharomyces cerevisiae*, a positive correlation between codon bias of genes and their protein levels was demonstrated (for example, [11,12]), suggesting that in those organisms translational selection is predominant. Other organisms, such as *Helicobacter pylori* [13], show almost no differential use of synonymous codons among

\* Correspondence: hanahm@ekmd.huji.ac.il  
Department of Microbiology and Molecular Genetics, Institute of Medical Research Israel-Canada, Faculty of Medicine, The Hebrew University of Jerusalem, POB 12272, Jerusalem 91120, Israel

different genes. While in the former organisms a mixture of mutation bias and natural selection underlies codon usage bias, in the latter organisms codon usage bias is mostly explained by mutation and very weak, if any, translational selection [14]. It was suggested by Andersson and Kurland [15] and recently substantiated by Kudla *et al.* [16] that selection towards highly adapted codons in highly expressed proteins has a global effect on the cell. By this convention, high expression of certain genes is mainly achieved by various mechanisms, such as regulation of transcription and/or translation initiation, and the use of well adapted codons in these genes guarantees efficient recycling of the ribosomes, resulting in an increase in cellular fitness. This might be reflected in a relatively enhanced translation of the whole proteome. It is thus intriguing to examine the association between the extent of codon usage bias of various organisms and their phenotypic traits towards understanding the environmental conditions where high extent of codon usage bias is advantageous. Of note, difference in the usage of synonymous codons that is uniform across all genes in a genome and probably arises from non-selective processes is not addressed here, but the different usage of codons among genes in a genome.

We regard an organism as 'biased' or 'unbiased' in association with its codon usage if the distribution of synonymous codons in its highly expressed genes differs from that in other genes in the genome. Several measures were proposed to estimate the extent of codon usage bias at an organism scale, enabling the classification of an organism as biased or unbiased [14,17-19]. Here we present such a measure based on the Codon Adaptation Index (CAI) [20] of individual genes. The CAI of a gene ranges between 0 and 1, with higher values indicating the use of more preferred codons. The extent of preference of each codon is determined by its frequency among the codons in genes encoding ribosomal proteins, using the latter as proxy for highly expressed genes. An organism-scale measure is obtained by computing  $CAI_{ave}$ , the average of the CAI values of all genes in a genome [17]. Since per definition highly expressed genes are assigned high CAI values, low or high  $CAI_{ave}$  values indicate whether there is a difference in codon usage between highly expressed genes and the rest of the genes in a genome. A low  $CAI_{ave}$  of an organism implies that there are many genes with low CAI values, and therefore preferred codons are assigned only to a small group of highly expressed genes. Accordingly, low  $CAI_{ave}$  values are indicative of biased organisms. A high  $CAI_{ave}$  of an organism implies that the CAI values of most genes are similar to those of genes encoding the ribosomal proteins, and therefore there is no differential use of synonymous codons among the genes encoded in that organism. Such organisms are

unbiased in regard to their codon usage. Another measure for the extent of codon usage bias of an organism is based on another measure of gene codon usage bias, the  $N_c$  (the effective number of codons) [21]. This gene measure ranges between 20 and 61, with lower values indicating the use of less codon types per amino acid along a protein-coding gene, which most often are the more preferred codons. To evaluate the extent of codon usage bias of an organism, a measure based on the difference between the average  $N_c$  values of the ribosomal genes and the average  $N_c$  values of the rest of the genes in the genome,  $N_{c,diff}$  is computed [17]. Organisms with high values of  $N_{c,diff}$  exhibit large extents of codon usage bias, and vice versa. The two genomic measures,  $CAI_{ave}$  and  $N_{c,diff}$ , are highly correlated (Figure S1 in Additional file 1; Pearson  $r = -0.91$ ,  $P \approx 0$ ,  $n = 1,169$ ).

Here we used these measures to characterize the extent of codon usage bias in 773 prokaryotic species representing 1,169 sequenced genomes. We demonstrate a wide range in the extent of codon usage bias among the various organisms, and trace the possible sources of this variation. Our results indicate that similarity in the extents of codon usage bias of different organisms reflects a similar ecological strategy they share.

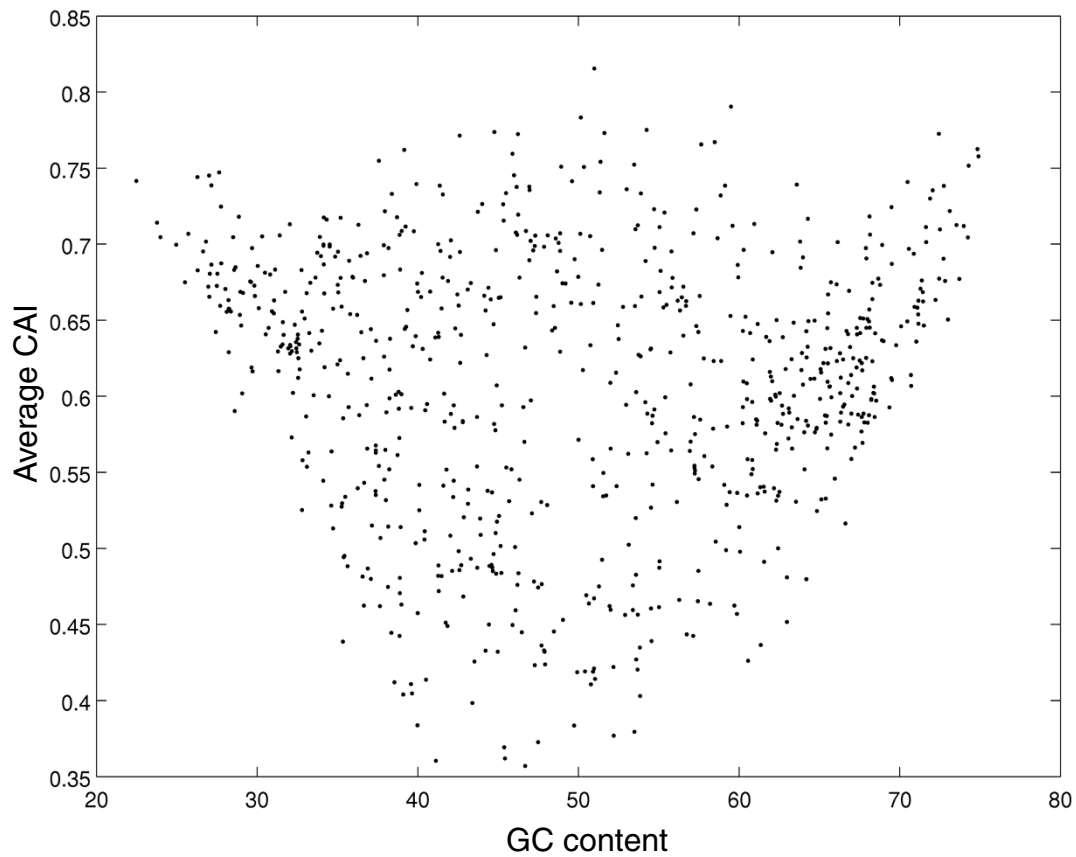
## Results

### Organisms differ in their extent of codon usage bias

We analyzed the extent of codon usage bias in 773 organisms: 699 bacteria and 74 archaea, representing 1,169 genomes (Materials and methods). Since the results for  $CAI_{ave}$  and  $N_{c,diff}$  are highly correlated, we present here the results for the first measure and in Additional file 1 the results for the second measure (Figures S2, S4, S5 and S6 in Additional file 1).

We first examined the association between the genomic values of  $CAI_{ave}$  and GC content (Figure 1), and observed that prokaryotes with extreme GC contents exhibit a narrower range of  $CAI_{ave}$  values than the other prokaryotes. Such organisms cannot achieve low  $CAI_{ave}$  values because their extreme GC content defines a limited repertoire of codons that are shared by both ribosomal genes and the rest of the genes, thus resulting in unbiased genomes (high  $CAI_{ave}$ ). To avoid bias in our conclusions due to the inclusion of organisms with extreme GC content, these organisms were excluded from further analyses. The next analyses were carried out for prokaryotes with GC contents between 35% and 65%, a total of 518 organisms.

As shown in Figure 2a, there is a wide distribution of  $CAI_{ave}$  values across the genomes, ranging between 0.35 (most biased organisms) and 0.82 (least biased organisms), with an average value of 0.59, median of 0.59 and standard deviation of 0.097. Another informative genomic measure is the standard deviation of CAI values of

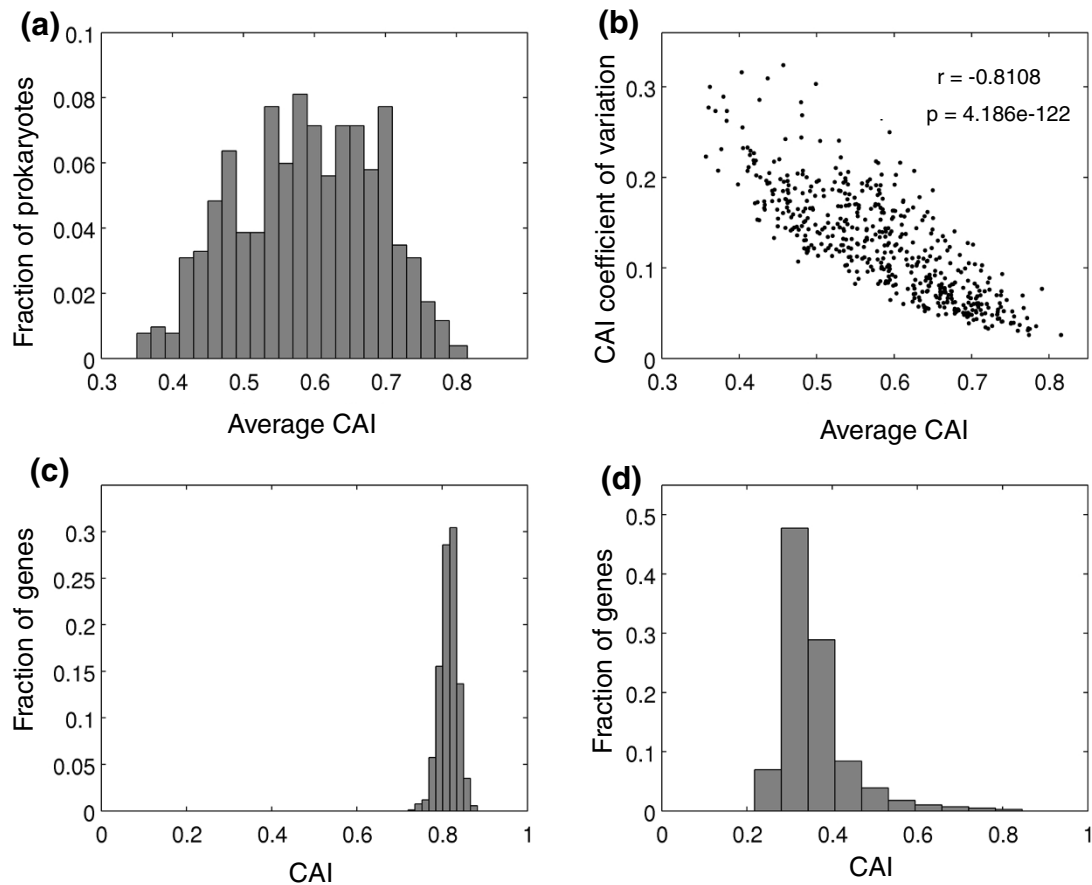


**Figure 1 Prokaryotes showing extreme GC content have relatively high  $CAI_{ave}$  values spanning a narrower range.** Scatter plot of GC content (x-axis) and  $CAI_{ave}$  (y-axis) of 773 organisms. Here and in all other figures  $CAI_{ave}$  is indicated as 'Average CAI'. CAI: Codon Adaptation Index; GC: genome content.

genes in a genome, which indicates the breadth of values used in calculating the genomic  $CAI_{ave}$  measures. To get an impression of this property across genomes we computed for each genome the coefficient of variation of CAI values of genes (standard deviation divided  $CAI_{ave}$ ). As expected, we find a strong inverse correlation between the  $CAI_{ave}$  values and their coefficients of variation (Figure 2b; Pearson  $r = -0.81$ ,  $P = 4.186E-122$ ,  $n = 518$ ), indicating that the distribution of CAI values of genes in biased organisms (low  $CAI_{ave}$ ) is broader than that of CAI values of genes in unbiased organisms.

To substantiate this result we investigated in each genome the difference in the usage of each codon between highly expressed genes and the rest of the genes encoded in the genome. For each organism we calculated the frequency of each codon out of the total codons encoding the corresponding amino acid, once in the ribosomal genes and once in the rest of the genes. We then calculated the average difference between codon frequencies in the ribosomal genes and in the

rest of the genes. Comparison of these average differences and  $CAI_{ave}$  values across genomes revealed a strong negative correlation (Pearson  $r = -0.8593$ ,  $P = 2.2639E-152$ ,  $n = 518$ ). This result reinforces our view that organisms with low  $CAI_{ave}$  are under stronger translational selection, resulting in differences between the codons in the highly expressed genes and the codons in the rest of the genes, while in organisms with high  $CAI_{ave}$  there is almost no differential use of codons between highly and weakly expressed genes. The latter unbiased organisms can be of two types: either the synonymous codons of both ribosomal and other genes are determined based on the GC content of the genome, or most genes use the preferred codons (that fit the abundant tRNAs or follow other rules that determine codon preference). To this end we looked at the association between  $CAI_{ave}$  and the average Nc of a genome in all 1,169 organisms. Low average Nc values imply that all genes in the genome use a specific set of codons and high average Nc values represent genomes where all



**Figure 2** Wide variation in the extents of codon usage bias among various organisms. **(a)** Distribution of  $CAI_{ave}$  values among 518 prokaryotes with GC content between 35% and 65%.  $CAI_{ave}$  is indicated as 'Average CAI'. **(b)** Negative correlation between organisms'  $CAI_{ave}$  values and their coefficients of variation. **(c)** Distribution of CAI values among the genes of *Syntrophothermus lipocalidus* (high  $CAI_{ave}$ , indicating low extent of codon usage bias). **(d)** Distribution of CAI values among the genes of *Vibrio vulnificus* (low  $CAI_{ave}$ , indicating high extent of codon usage bias). CAI: Codon Adaptation Index; GC: genome content.

genes use a mixture of codons. Indeed, we find among genomes with high  $CAI_{ave}$  these two types of trends (Figure S3a in Additional file 1). As to the biased genomes (with  $CAI_{ave}$  below the average of 0.59), most of them use relatively many codons across their genes (average Nc values above 40), but apparently they use specific codons in their highly expressed genes, represented by the ribosomal genes. In general, it seems that the average values of Nc are strongly determined by the GC content of the genome (Figure S3b in Additional file 1). Especially, as we noted above, in genomes with extreme GC content, where the possibilities of codon variation are severely restricted by the specific nucleotide repertoire, the average Nc values are low.

Organisms at the right tail of the  $CAI_{ave}$  distribution are the least biased. The prokaryote with the smallest extent of codon usage bias (highest  $CAI_{ave}$  value, 0.82)

is *Syntrophothermus lipocalidus*, a thermophilic, syntrophic, fatty-acid-oxidizing anaerobe that belongs to the Clostridia class [22] (Figure 2c). Other prokaryotes with almost as high  $CAI_{ave}$  values are quite diverse: *Geobacter metallireducens* ( $CAI_{ave}$  value of 0.79) is an anaerobic bacterium that uses iron oxides as the electron acceptor in the oxidation of organic compounds to carbon dioxide [23], *Nitrosococcus watsoni* ( $CAI_{ave}$  value of 0.78) is an aerobic marine bacterium, and *Coxiella burnetii* ( $CAI_{ave}$  value of 0.78) is a facultative, intracellular pathogenic bacterium that causes the Q fever.

Organisms at the left tail of the distribution are the most biased. The prokaryote that exhibited the greatest extent of codon usage bias (lowest  $CAI_{ave}$  value, 0.35) is *Vibrio vulnificus*, a human pathogen of the Gammaproteobacteria class (Figure 2d). Interestingly, other organisms that have such large extents of codon usage bias

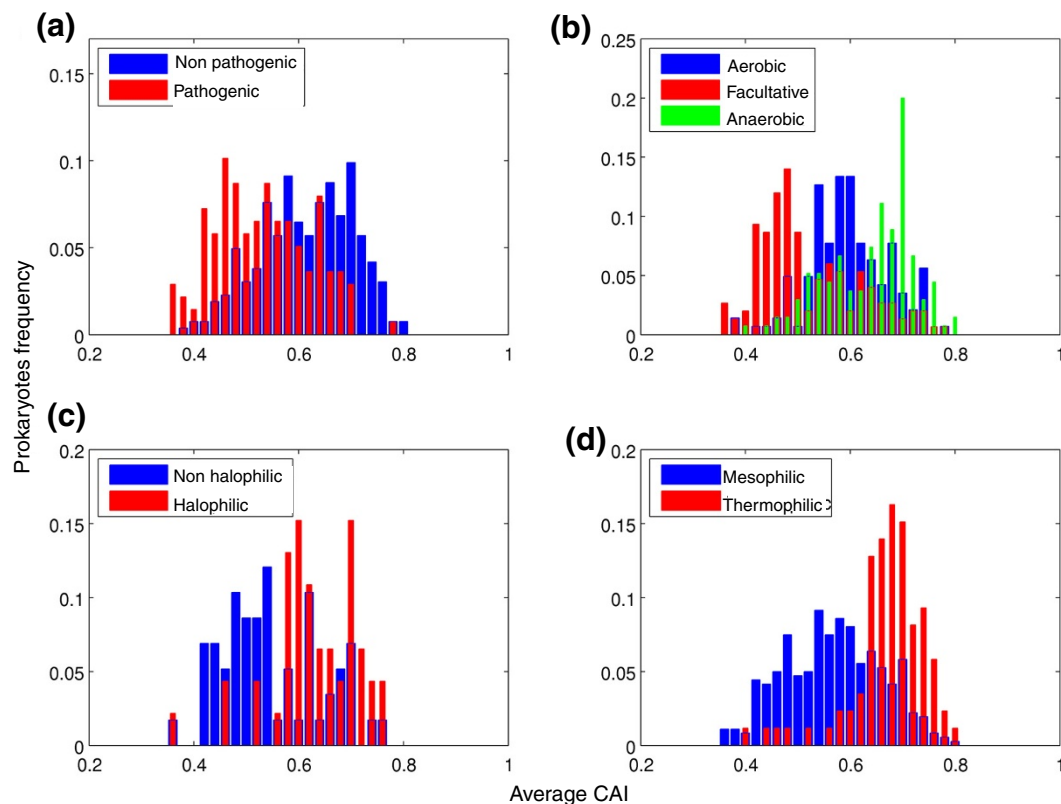
are also pathogenic: *Streptococcus suis* is a swine pathogen, *Vibrio cholera* causes cholera in humans and *Corynebacterium diphtheria* causes diphtheria, an upper respiratory tract illness (CAI<sub>ave</sub> values of 0.36, 0.37, and 0.38, respectively). This finding led us to examine the distribution of CAI<sub>ave</sub> values among pathogenic versus non-pathogenic prokaryotes (Figure 3a). As shown in Figure 3a, the distribution among pathogenic bacteria is biased to the left compared to non-pathogenic bacteria, with pathogenic prokaryotes having statistically significant lower CAI<sub>ave</sub> values ( $P = 1.26E-15$  by Mann-Whitney test), indicating they are more biased.

### Biased organisms can be classified by their phenotypic traits

We annotated each genome with its CAI<sub>ave</sub>, NC<sub>diff</sub> and several phenotypic traits, such as its oxygen requirement and range of growth temperatures (Additional file 2). This annotation system (Table 1) has enabled us to compare the distributions of the CAI<sub>ave</sub> values of

different groups of organisms, tagged according to a particular phenotypic trait.

This analysis (Figure 3) indicated that groups of prokaryotes classified by their oxygen requirement differ statistically significantly in the distributions of CAI<sub>ave</sub> values ( $P = 1.99E-23$  by Kruskal-Wallis test). Facultative organisms exhibited the largest extent of codon usage bias and anaerobic organisms showed the smallest values ( $P = 2.92E-12$ , Mann-Whitney test between facultative and aerobic;  $P = \sim 0$  between facultative and anaerobic;  $P = 3.14E-6$  between aerobic and anaerobic). Examining groups of prokaryotes that live in environments that differ in their salinity levels demonstrated statistically significant differences among them ( $P = 2.13E-5$  by Mann-Whitney test). Organisms that live in different temperature ranges showed statistically significant differences in their CAI<sub>ave</sub> values ( $P = 5.52E-21$  by Mann-Whitney test): thermophiles demonstrated statistically significantly higher CAI<sub>ave</sub> values than mesophiles. Intriguingly, we found that organisms living in multiple



**Figure 3** Prokaryotes exhibiting different environmental characteristics show different extents of codon usage bias. The 518 prokaryotes were classified according to various properties, and the distributions of CAI<sub>ave</sub> values of the different classes were compared. **(a)** Pathogenic prokaryotes show larger extents of codon usage bias than non-pathogenic prokaryotes. CAI<sub>ave</sub> is indicated as 'Average CAI'. **(b)** Facultative prokaryotes show the largest extent of codon usage bias and anaerobic prokaryotes show the smallest extent. **(c)** Prokaryotes that live in different salinity environments show a statistically significant difference in CAI<sub>ave</sub> values. **(d)** Thermophilic and hyperthermophilic prokaryotes show a smaller extent of codon usage bias than mesophilic prokaryotes. CAI: Codon Adaptation Index; GC: genome content.

**Table 1 Classification of organisms by phenotypic traits**

Environmental property	Number of prokaryotes
<b>Pathogenicity</b>	
Pathogenic	138
Non-pathogenic	263
Total	401
<b>Oxygen requirement</b>	
Aerobic	142
Anaerobic	135
Facultative	150
Microaerophilic	12
Total	439
<b>Salinity</b>	
Extreme halophilic	7
Mesophilic	19
Moderate halophilic	20
Non-halophilic	58
Total	104
<b>Temperature range</b>	
Hyperthermophilic	35
Mesophilic	361
Psychrophilic	12
Thermophilic	51
Total	459
<b>Habitat</b>	
Multiple	127
Specialized	81
Total	208

habitats have statistically significantly lower  $CAI_{ave}$  values than organisms living in specialized habitats (Figure 4;  $P = 6.3E-10$  by Mann-Whitney test). This result is consistent with the results presented above for the other phenotypic traits and generalizes them. Pathogenic bacteria often live in multiple environments outside and within their host, and facultative organisms live in environments with and without oxygen. On the other hand, thermophiles (found above to be less biased than mesophiles) are usually restricted to a specific environment with a specific temperature. The consistency between these results is also implied by the interdependence between these different phenotypic traits (as shown by  $\chi^2$  test; Table S2 in Additional file 1).

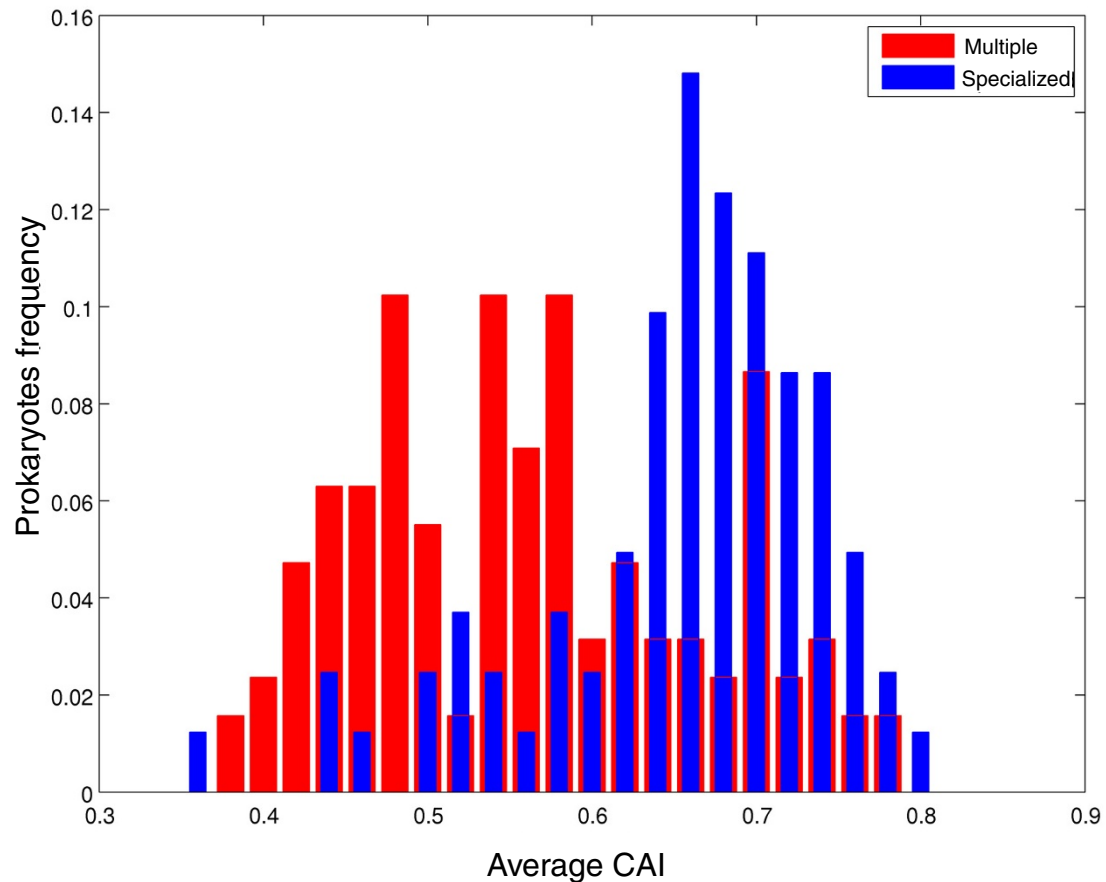
**Phenotypic traits rather than phylogenetic relatedness underlie the similarities in codon usage bias between organisms**

It is possible that biased organisms are evolutionarily related and their similar values of  $CAI_{ave}$  stem from their phylogenetic relatedness. To test this, we computed the correlation coefficient between the phylogenetic distance (Materials and methods) and difference in  $CAI_{ave}$  values of pairs of organisms. The pairwise

measures were computed for pairs of prokaryotes and pairs of archaea, and the correlation analysis was carried out for all pairs together (Figure 5). No correlation was found between the phylogenetic distances and the differences in  $CAI_{ave}$  (Pearson  $r = 0.078$ ), implying no noticeable influence of the phylogeny on the extent of codon usage bias. As shown in Figure 5, organisms with very small differences in their  $CAI_{ave}$  values can be distantly separated on the evolutionary tree. Of note, organisms that are extremely close on the phylogenetic tree do not exhibit differences in  $CAI_{ave}$  that are larger than 0.15. Thus, in these cases the similarity in  $CAI_{ave}$  values may stem from the close phylogenetic relatedness. However, beyond a certain (low) threshold, there is no dependence between the phylogenetic distance and extent of codon usage bias. These results strengthen our previous conclusion that the extent of codon usage bias is associated with the phenotypic traits of the organism.

**The empirical association between growth rate and metabolic variability is attributed to their individual associations with codon usage bias**

Previous studies showed that there is an association between codon bias and growth rate [17,24] and between growth rate and metabolic variability [25]. To verify that our result is not indirectly inferred from these two associations, we computed the correlation coefficient between pairs of properties ( $CAI_{ave}$ , growth rate, and type of habitat (multiple or specific)). This analysis included 82 organisms for which we had information on their growth rate and habitat type. We repeated this analysis twice. Once we simply computed the correlation coefficient of two variables, and in the second analysis we performed partial correlation, controlling for the third variable (Table 2a). This analysis demonstrates that our conclusion about an association between  $CAI_{ave}$  and the type of habitat is independent of the correlations with the growth rate, as the two correlation coefficients obtained in the two computations, with and without taking into account the growth rate, were very similar ( $r = 0.46$ ,  $P = 1.25E-5$  and  $r = 0.43$ ,  $P = 6.9E-5$ , respectively). Intriguingly, the correlation between growth rate and habitat type was shown to be highly dependent on  $CAI_{ave}$ . While the correlation between these two variables was found to be approximately 0.2 and nearly statistically significant ( $P = 0.07$ ), the partial correlation, controlling for  $CAI_{ave}$ , dropped substantially to 0.04 ( $P = 0.6$ ). Our results suggest that the empirical association observed between growth rate and metabolic variability can be attributed to their individual associations with codon usage bias. Of note, the correlation between  $CAI_{ave}$  and habitat type was the highest obtained and it is highly statistically significant. To verify that our conclusions are not affected by



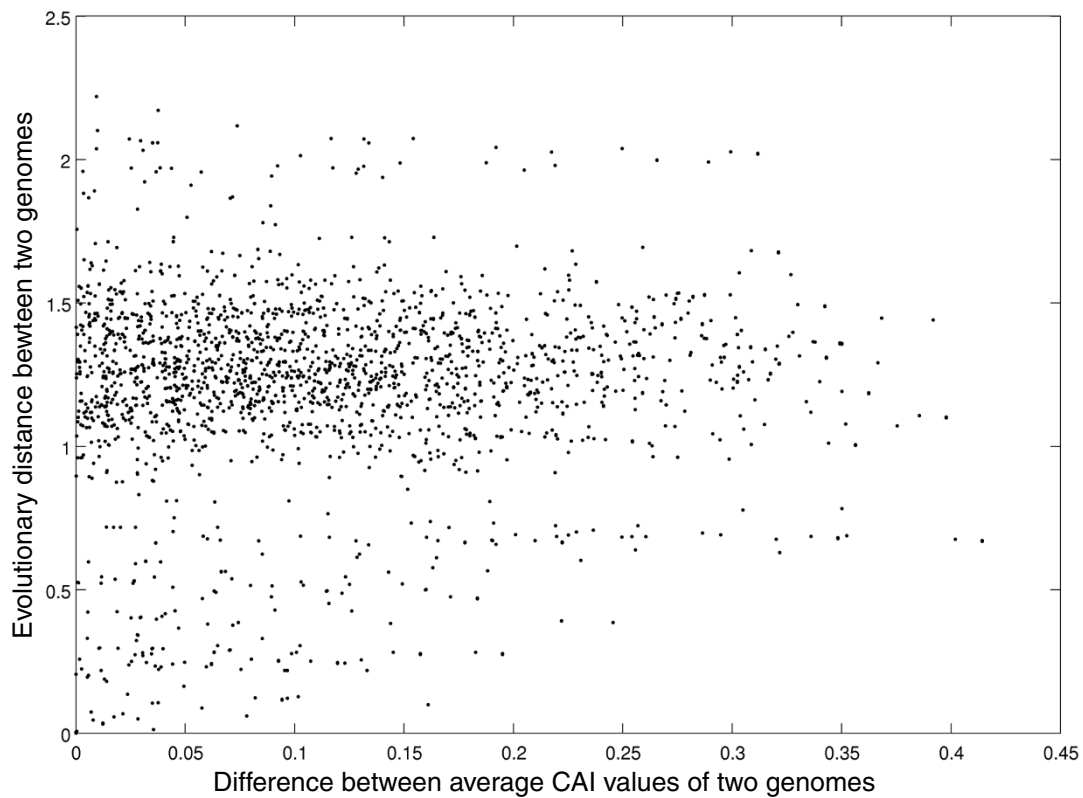
**Figure 4** The ability of an organism to live in multiple habitats is the feature most correlated with the extent of codon usage bias. The 518 prokaryotes were divided into two groups - organisms who each live in a specialized environment and organisms who each live in multiple habitats - and the  $CAI_{ave}$  values of the different groups were compared.  $CAI_{ave}$  is indicated as 'Average CAI'. CAI: Codon Adaptation Index; GC: genome content.

inclusion of closely related organisms, which might introduce redundancy in the data, we repeated the analysis with a subset of the 82 organisms, including 24 organisms that are phylogenetically remote (Materials and methods; Table 2, Analysis B). Using this dataset, the correlations between growth rate and both habitat type and codon bias were not statistically significant. The correlation and partial correlation between  $CAI_{ave}$  and habitat type were consistent with the results for the whole dataset.

### Discussion

It is widely acknowledged that synonymous codons are used unevenly among genes in a genome, with genes encoding highly expressed proteins being enriched with specific codons. It is still under debate whether the biased use of certain codons in highly expressed genes is one of the causes or the result of the high expression level. On

the one hand, it was shown that the use of certain codons affects directly the speed of translation [26] and its accuracy [27], and codon optimization is even used to elevate the levels of proteins expressed outside their original context [28,29]. On the other hand, Kudla *et al.* [16] showed that the variation in the levels of proteins translated from synthetic green fluorescent protein constructs, varying only at synonymous sites, was not correlated with the codon usage. They found that high expression was not associated with specific codons but with avoidance of secondary structure at the translation initiation site. This supports the proposition [15] that selection for well adapted codons in highly expressed genes does not affect directly the level of individual proteins, but provides a global benefit to the cell, as it assures efficient recycling of the ribosomes, which leads to an increase in cellular fitness. It should be noted that avoidance of secondary structure at the 5' end of the mRNA is only one



**Figure 5 No correlation between pairwise evolutionary distances across the phylogenetic tree and difference in CAI<sub>ave</sub> values of pairs of organisms.** Scatter plot of the difference in CAI<sub>ave</sub> values between each pair of two organisms in our data (x-axis) and their phylogenetic distances (y-axis). CAI<sub>ave</sub> is indicated as 'Average CAI'. The pairwise distances across the phylogenetic tree were based on the tree generated in [38] and computed as the path length between two organisms through the most recent common ancestor. CAI, Codon Adaptation Index.

mechanistic strategy that may lead to high levels of gene expression [30], and other mechanisms could underlie high levels of expression as well. These include high transcription level from strong promoters, high stability of the mRNA and/or efficient translation initiation by optimal Shine-Dalgarno sequences. Thus, different highly expressed genes might use well adopted codons to improve cellular fitness, independent of the molecular mechanism underlying their high expression. The

premise by which translation selection for preferred codons in highly expressed genes has a global effect motivated us to investigate its association with the phenotypic traits of a wide range of organisms. It should be noted that our conclusions are not affected by whether or not the suggested global effect is accompanied by local effects on the translation efficiency.

There have been various attempts to explain what makes certain codons preferred over others in highly

**Table 2 Correlations and partial correlations between growth rate, habitat type and global codon usage bias**

	Growth rate and habitat type		Growth rate and codon usage bias		Habitat type and codon usage bias	
	r	P	r	P	r	P
<b>Analysis A</b>						
Correlation	0.1995	0.0723	0.3451	0.0015	0.4619	1.2538E-5
Partial correlation	0.0482 <sup>a</sup>	0.6693 <sup>a</sup>	0.291 <sup>b</sup>	0.0084 <sup>b</sup>	0.4274 <sup>c</sup>	6.919E-5 <sup>c</sup>
<b>Analysis B</b>						
Correlation	0.0368	0.8645	0.2971	0.1586	0.4692	0.0207
Partial correlation	NA	NA	NA	NA	0.4802 <sup>c</sup>	0.0204 <sup>c</sup>

Analysis A was based on 82 organisms; analysis B was based on a subset of 24 organisms, which are phylogenetically remote from each other. Pearson correlation coefficients were computed. <sup>a</sup>Controlling for codon usage bias; <sup>b</sup>controlling for habitat type; <sup>c</sup>controlling for growth rate. NA, when the Pearson correlation coefficient was not statistically significant, partial correlation was not computed.



expressed genes, from correspondence to abundant tRNAs to physical considerations regarding the optimal stability of codon-anticodon interaction (summarized in [17]). Here we have not dealt with these various possible types of preferred codons but regarded them as the codons used by highly expressed genes in a genome, based on the premise that selection favored the most adapted codons in highly expressed genes. Hence, our analysis was based on measures that compare the codon usage of all genes to the codon usage of highly expressed genes, represented by the set of ribosomal genes in each genome ( $CAI_{ave}$  and  $N_{Cdiff}$ ). Such a comparative measure should provide us information on the selection forces that act on individual genes and on the whole genome. Indeed, we find a wide variation in these measures across genomes (Figure 2a; Figure S2 in Additional file 1), where some organisms are highly biased and others show only very slight, if any, difference between the codon usage in ribosomal genes and other genes. It is possible that the lack of selection implied for some of the unbiased genomes actually reflects their small population size [31], but in the absence of a reliable measure of effective population size in bacteria, we are unable to assess this further. In biased genomes selection acts only on genes that are highly expressed, to assure the overall translation mechanism to operate efficiently. Thus, our analysis is in line with previous observations [17] that there are organisms where translational selection is operational (biased genomes) and others where it is not (unbiased). In our study we used the set of ribosomal genes as a representative set of highly expressed genes. When data of gene expression in many prokaryotes become available it should be possible to extend our study by using sets of highly expressed genes based on their measured expression levels. It would be possible then to divide the organisms into those with high and low variation in gene expression, and to examine how the level of variation in gene expression is reflected in  $CAI_{ave}$  values.

We found that pathogenic prokaryotes have statistically significantly lower  $CAI_{ave}$  than non-pathogenic prokaryotes, but with a substantial overlap between the histograms of those two groups. This result might be surprising in view of previous studies that found that pathogenic lifestyle is linked to relaxation of selection [32-34], which should be linked to reduced codon bias. It is possible to reconcile the discrepancy by dividing the pathogenic and the non-pathogenic prokaryotes into two subtypes: some are capable of living in multiple environments and some stay mainly host-associated. When we compared pathogenic and non-pathogenic groups of host-associated prokaryotes there was no statistically significant difference in their  $CAI_{ave}$  values ( $P = 0.06$  by Mann-Whitney test), but when we compared

pathogenic and non-pathogenic groups of prokaryotes living in multiple habitats, the pathogenic prokaryotes showed statistically significantly lower  $CAI_{ave}$  values ( $P = 5.624E-7$  by Mann-Whitney test). Therefore, it seems that pathogenic host-associated microbes like *C. burnetti*, which is an intracellular pathogen with an extremely high  $CAI_{ave}$  value of 0.78, are not exposed to strong selection, while other pathogens that are able to live in multiple habitats are still under stronger selection than non-pathogenic prokaryotes living in multiple habitats when it comes to selection on codon usage.

Previous studies discussed the correspondence between ecology preferences and codon adaptation [35,36]. Our results suggest that organisms may adjust to metabolic variability by maintaining a high extent of codon usage bias (reflected by their low  $CAI_{ave}$  values). Previous studies analyzed the association between codon adaptation and growth rate and between growth rate and metabolic variability [14,17,24,25]. One study showed that most bacterial organisms choose one of two alternative ecological strategies: living in multiple habitats with a large extent of co-habitation or living in a specialized niche in which the co-habitation is limited. It was shown that growth rate is statistically significantly correlated with metabolic variability encountered by an organism, suggesting a universal principle by which metabolic flexibility is associated with a need to grow fast, perhaps because of the greater extent of competition [25]. Independently, Rocha demonstrated an association between bacteria with large extents of codon usage bias and fast growth, and also an association with the number of tRNA genes [17,24]. It was demonstrated that fast growing bacteria have more tRNA genes of fewer types, and suggested that the translation in those organisms depends on fast tRNA diffusion to the ribosome. That study proposed that co-evolution of the tRNA pool and the codon usage bias allows more efficient translation of highly expressed genes, and that the codon usage bias in highly expressed genes relative to the rest of the genome is predicted to be under stronger selection in fast growing organisms. Our results tie these two results together and suggest that translational selection towards most adapted codons in highly expressed genes is operational in organisms that live in variable environments, enabling them to efficiently address the metabolic variability and the competition.

## Conclusions

Codon usage bias in highly expressed genes was suggested to have a global effect on the cell, increasing cellular fitness. Here we perform the first large-scale study that examines the relationship between codon usage bias and the phenotypic traits of prokaryotic organisms. Our analysis revealed a large variation in the global

extent of codon bias among prokaryotic genomes, which is associated with their lifestyles. Especially, we discovered that organisms living in multiple habitats, including facultative organisms, mesophiles and pathogenic bacteria, exhibit high extents of codon usage bias, consistent with their need to adapt efficiently to different environments.

## Materials and methods

### Data organization

We retrieved 1,169 genome sequences from the NCBI Entrez Genome Project database [37]. For species that had sequenced genomes for more than one subspecies, we maintained one representative subspecies, which was chosen randomly. This resulted in a data set of 773 prokaryotes. From this list, only prokaryotes with GC content larger than 35% and smaller than 65% were included for further analyses, resulting in a data set of 518 prokaryotes.

### Computation of codon bias measures at an organism scale

Genome  $CAI_{ave}$ : for each organism we computed the average of CAI values over all genes in the genome [17]. The CAI of a gene was computed as described in [20].

$NC_{diff}$ : for each organism we calculated the Nc value for each gene in the genome, as described in [21]. We next computed the average Nc value of the genes encoding ribosomal proteins (*Nc (rib)*) and the average Nc value of the rest of the genome (*Nc (all)*) [17].  $NC_{diff}$  was obtained by:

$$\frac{Nc(all) - Nc(rib)}{Nc(all)}$$

### Comparison between groups of organisms

The prokaryotes were classified according to their environmental characteristics (oxygen requirement, salinity, temperature range and habitat) and also whether they are pathogenic, based on the documentation in the NCBI Entrez Genome Project database (detailed in Additional file 2). The number of organisms annotated with each property is detailed in Table 1. Comparisons between the value distributions of organism measures ( $CAI_{ave}$  or  $NC_{diff}$ ) were performed by Mann-Whitney or Kruskal-Wallis tests.

### Phylogenetic distances between organisms

The pairwise distances across the phylogenetic tree were based on the tree generated in [38] and computed as the path length between two organisms through the most recent common ancestor. We used 2,016 possible pairs of 64 bacteria and 78 possible pairs of 13 archaea in this analysis.

### Correlations and partial correlations between growth rate, habitat type and codon usage bias

This analysis included 82 prokaryotes, for which we had information for both their growth rates and habitat types. The growth rates were obtained from [24]. The prokaryotes' environmental annotations were obtained from the NCBI Entrez Genome Project database (multiple habitat-living organisms were annotated as 0, and specialized organisms were annotated as 1). The extents of codon usage bias were represented as  $CAI_{ave}$  values. We repeated the analysis using a subset of 24 organisms, which are all phylogenetically remote from each other, based on the phylogenetic tree used above [38].

### Additional material

#### Additional file 1: Additional figures and tables.

**Additional file 2: Table S1 - features and characteristics of the prokaryotes included in this study.** This file contains data and annotations of the organisms included in the analysis (nam, tax ID,  $CAI_{ave}$ , median CAI, coefficient of variation (of CAI),  $NC_{diff}$ , average Nc, environmental properties, if it is the representative subspecies of the species, super kingdom, GC content).

### Abbreviations

CAI: Codon Adaptation Index.

### Acknowledgements

This study was supported by the Israel Science Foundation administered by the Israeli Academy for Sciences and Humanities. We thank Ruth Hershberg for her critical comments on the manuscript. We are grateful to all our lab members for fruitful discussions.

### Authors' contributions

MB carried out all analyses, conceived the work, analyzed the data and wrote the paper. HM conceived the work, analyzed the data and wrote the paper. All authors have read and approved the manuscript for publication.

### Competing interests

The authors declare that they have no competing interests.

Received: 13 April 2011 Revised: 16 June 2011

Accepted: 27 October 2011 Published: 27 October 2011

### References

1. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A: **Codon catalog usage and the genome hypothesis.** *Nucleic Acids Res* 1980, **8**:r49-r62.
2. Gouy M, Gautier C: **Codon usage in bacteria: correlation with gene expressivity.** *Nucleic Acids Res* 1982, **10**:7055-7074.
3. Bennetzen JL, Hall BD: **Codon selection in yeast.** *J Biol Chem* 1982, **257**:3026-3031.
4. Post LE, Nomura M: **DNA sequences from the str operon of Escherichia coli.** *J Biol Chem* 1980, **255**:4660-4666.
5. Ikemura T: **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* 1981, **151**:389-409.
6. Ikemura T: **Codon usage and tRNA content in unicellular and multicellular organisms.** *Mol Biol Evol* 1985, **2**:13-34.
7. Dong H, Nilsson L, Kurland CG: **Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates.** *J Mol Biol* 1996, **260**:649-663.

8. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH: **Codon usage between genomes is constrained by genome-wide mutational processes.** *Proc Natl Acad Sci USA* 2004, **101**:3480-3485.
9. Hershberg R, Petrov DA: **General rules for optimal codon choice.** *PLoS Genet* 2009, **5**:e1000556.
10. Supek F, Skunca N, Repar J, Vlahovicek K, Smuc T: **Translational selection is ubiquitous in prokaryotes.** *PLoS Genet* 2010, **6**:e1001004.
11. Lithwick G, Margalit H: **Hierarchy of sequence-dependent features associated with prokaryotic translation.** *Genome Res* 2003, **13**:2665-2673.
12. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: **Global analysis of protein expression in yeast.** *Nature* 2003, **425**:737-741.
13. Lafay B, Atherton JC, Sharp PM: **Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*.** *Microbiology* 2000, **146**:851-860.
14. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE: **Variation in the strength of selected codon usage bias among bacteria.** *Nucleic Acids Res* 2005, **33**:1141-1153.
15. Andersson SG, Kurland CG: **Codon preferences in free-living microorganisms.** *Microbiol Rev* 1990, **54**:198-210.
16. Kudla G, Murray AW, Tollervey D, Plotkin JB: **Coding-sequence determinants of gene expression in *Escherichia coli*.** *Science* 2009, **324**:255-258.
17. Rocha EP: **Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization.** *Genome Res* 2004, **14**:2279-2286.
18. dos Reis M, Savva R, Wernisch L: **Solving the riddle of codon usage preferences: a test for translational selection.** *Nucleic Acids Res* 2004, **32**:5036-5044.
19. Suzuki H, Saito R, Tomita M: **Measure of synonymous codon usage diversity among genes in bacteria.** *BMC Bioinformatics* 2009, **10**:167.
20. Sharp PM, Li WH: **The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15**:1281-1295.
21. Wright F: **The 'effective number of codons' used in a gene.** *Gene* 1990, **87**:23-29.
22. Sekiguchi Y, Kamagata Y, Nakamura K, Ohashi A, Harada H: ***Syntrophothermus lipocalidus* gen. nov., sp. nov., a novel thermophilic, syntrophic, fatty-acid-oxidizing anaerobe which utilizes isobutyrate.** *Int J Syst Evol Microbiol* 2000, **50**:771-779.
23. Lovley DR, Phillips EJ: **Novel mode of microbial energy metabolism: organic carbon oxidation coupled to dissimilatory reduction of iron or manganese.** *Appl Environ Microbiol* 1988, **54**:1472-1480.
24. Vieira-Silva S, Rocha EP: **The systemic imprint of growth and its uses in ecological (meta)genomics.** *PLoS Genet* 2010, **6**:e1000808.
25. Freilich S, Kreimer A, Borenstein E, Yosef N, Sharan R, Gophna U, Ruppin E: **Metabolic-network-driven analysis of bacterial ecological strategies.** *Genome Biol* 2009, **10**:R61.
26. Tuller T, Carmi A, Vestsgian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y: **An evolutionarily conserved mechanism for controlling the efficiency of protein translation.** *Cell* 2010, **141**:344-354.
27. Akashi H: **Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy.** *Genetics* 1994, **136**:927-935.
28. Deng T: **Bacterial expression and purification of biologically active mouse c-Fos proteins by selective codon optimization.** *FEBS Lett* 1997, **409**:269-272.
29. Gustafsson C, Govindarajan S, Minshull J: **Codon bias and heterologous protein expression.** *Trends Biotechnol* 2004, **22**:346-353.
30. Gu W, Zhou T, Wilke CO: **A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes.** *PLoS Comput Biol* 2010, **6**:e1000664.
31. dos Reis M, Wernisch L: **Estimating translational selection in eukaryotic genomes.** *Mol Biol Evol* 2009, **26**:451-461.
32. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW, Gagneux S: **High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography.** *PLoS Biol* 2008, **6**:e311.
33. Hershberg R, Tang H, Petrov DA: **Reduced selection leads to accelerated gene loss in *Shigella*.** *Genome Biol* 2007, **8**:R164.
34. Ochman H, Moran NA: **Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis.** *Science* 2001, **292**:1096-1099.
35. Man O, Pilpel Y: **Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species.** *Nat Genet* 2007, **39**:415-421.
36. Jiang H, Guan W, Pinney D, Wang W, Gu Z: **Relaxation of yeast mitochondrial functions after whole-genome duplication.** *Genome Res* 2008, **18**:1466-1471.
37. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2007, **35**:D26-31.
38. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life.** *Science* 2006, **311**:1283-1287.

doi:10.1186/gb-2011-12-10-r109

**Cite this article as:** Botzman and Margalit: Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biology* 2011 **12**:R109.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

