7014–7023 *Nucleic Acids Research, 2009, Vol. 37, No. 21* *Published online 22 September 2009*
doi:10.1093/nar/gkp792

Universal function-specificity of codon usage

Hamed Shateri Najafabadi[1,2], Hani Goodarzi[3] and Reza Salavati[1,2,4,*]

[1]Institute of Parasitology, McGill University, 21111 Lakeshore Road, Ste. Anne de Bellevue, Montreal, Quebec, H9X3V9, [2]McGill Centre for Bioinformatics, McGill University, Duff Medical Building, 3775 University Street, Montreal, Quebec, H3A2B4, Canada, [3]Department of Molecular Biology and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton NJ 08544, USA and [4]Department of Biochemistry, McGill University, McIntyre Medical Building, 3655 Promenade Sir William Osler, Montreal, Quebec, H3G1Y6, Canada

Received August 16, 2009; Revised September 2, 2009; Accepted September 8, 2009

ABSTRACT

Synonymous codon usage has long been known as a factor that affects average expression level of proteins in fast-growing microorganisms, but neither its role in dynamic changes of expression in response to environmental changes nor selective factors shaping it in the genomes of higher eukaryotes have been fully understood. Here, we propose that codon usage is ubiquitously selected to synchronize the translation efficiency with the dynamic alteration of protein expression in response to environmental and physiological changes. Our analysis reveals that codon usage is universally correlated with gene function, suggesting its potential contribution to synchronized regulation of genes with similar functions. We directly show that coexpressed genes have similar synonymous codon usages within the genomes of human, yeast, *Caenorhabditis elegans* and *Escherichia coli*. We also demonstrate that perturbing the codon usage directly affects the level or even direction of changes in protein expression in response to environmental stimuli. Perturbing tRNA composition also has tangible phenotypic effects on the cell. By showing that codon usage is universally function-specific, our results expand, to almost all organisms, the notion that cells may need to dynamically alter their intracellular tRNA composition in order to adapt to their new environment or physiological role.

INTRODUCTION

Genome-wide analysis of gene expression has been extensively used to study the mechanisms underlying the dynamic regulation of gene expression. Simultaneous changes in transcript levels across different conditions (i.e. environmental as well as spacio-temporal and genetic variables) have been widely used as a proxy for identifying sets of coregulated genes, thus revealing the common regulatory elements underlying these observed correlations (1,2). These regulatory elements can act at both transcriptional and post-transcriptional levels including mechanisms such as localization and stability of the transcript and enhancement or suppression of translation (3,4). Most studies have focused on the non-coding genome for finding such regulatory elements. While coding sequences have also been shown to contain elements that contribute to regulation of expression for a minority of genes (5–7), the relationship between the dynamic regulation of expression and the sequence of coding regions is not considered widespread among and within organisms.

A novel relationship between coding sequence and dynamic regulation of protein expression can be readily hypothesized from the observed variations in patterns of isoacceptor tRNA abundance and tRNA charging in different conditions and tissues (8–10). For example, during amino acid starvation, unlike common tRNA species, the charged level of certain isoacceptor tRNAs cognate to rare codons remains high (8,11). Interestingly, these rare codons are used in higher frequencies among genes involved in amino acid biosynthesis. Therefore, the high charged level of their cognate tRNA species can boost the amino acid biosynthesis pathway by supporting the high expression level of its enzymes (11). Methylation of the wobble base of a tRNA in yeast has also been shown to affect its codon preference, enhancing levels of certain proteins (12).

Congruent with the observed tissue-specificity of tRNA composition (10), Plotkin *et al.* (13) report the presence of a tissue-specific codon usage in human genes, although Sémon *et al.* (14) reject their hypothesis using a different statistical analysis and a richer database of tissue-specific genes. However, many other studies indirectly suggest the

*To whom correspondence should be addressed. Tel: +1 514 398 7721; Fax: +1 514 398 7857; Email: reza.salavati@mcgill.ca

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

act of selection on synonymous codons in human genome (15); these models propose translational efficiency (16–18), mRNA stability (19–21) and splicing control (22) as mechanisms underlying such selection.

Despite all these studies, factors shaping codon usage of genes in many organisms, including human, are still not completely understood. For example, no significant evidence for the presence of selection on codon usage was found in 30% of the bacteria that were examined using a population genetics-based model (23). A recent study has profoundly added to this complication by reporting that it is not the codon usage, but rather RNA structure that affects expression level of a protein (24).

More than 30 years ago, Garel (25) reported that the tRNA composition of silk gland in silkworm changes during the development of this organ in order to accommodate for the high rate of synthesis of fibroin which is rich in glycine, alanine, serine and tyrosine. In other words, silk gland cells try to synchronize the translation efficiency of fibroin with its required amount at each time by providing the tRNAs that carry these four amino acids. Matching tRNA composition with coding sequence may extend beyond amino acid usage of the proteins and include synonymous codon usage as well. Here, based on several rigorous statistical analyses of coding sequences from almost all available genomes, we suggest function-specificity of synonymous codon usage in a wide range of organisms. This implies that functional adaptation of tRNA content (25) may be a universal mechanism for synchronizing the translation efficiency with the dynamic, function-specific alteration of protein expression. In other words, rather than having a single set of optimal codons, organisms could harbor different sets that change depending on environmental conditions or physiological roles and are related to the functions that are most expressed at each of these conditions. We show that this hypothesis best explains the synonymous codon usage of genes across all domains of life. It also explains our recent observation that in three different organisms, *Saccharomyces cerevisiae*, *E. coli* and *Plasmodium falciparum*, genes whose products interact either physically or functionally use similar codons (26). Although not as comprehensive as our computational analysis, we also provide limited experimental data showing that differences in codon usage or variations in the tRNA content of the cell can result in varied responses to environmental changes, in terms of regulation of protein expression and cell phenotype.

MATERIALS AND METHODS

Analysis of correlation between codon usage and function/expression pattern

Normalized frequency of each codon in each gene (f_c) was calculated as the usage of that codon divided by the usage of the amino acid it codes for. For each gene, we calculated the normalized frequencies of codons of an amino acid only if the usage of that amino acid was higher than a defined cutoff, in order to remove noisy measurements. The distance of normalized frequencies of each codon in each pair of genes was calculated, and the relationship between this distance value and likelihood of functional linkage/coexpression was assessed by Pearson correlation coefficient. Recently duplicated genes have high sequence identity and, thus, similar codon usages that may not necessarily reflect act of selection on gene sequence. In addition, these genes tend to cross-hybridize on expression arrays and appear as coexpressed genes. Therefore, to avoid biased analysis, paralogous genes were removed from each genome prior to calculating Pearson correlation coefficients. This was done by sequentially selecting two paralogous genes and randomly removing one of them, until the remaining genes contained no paralogs in the dataset (2). Non-random distribution of f_c in each functional cluster or cluster of coregulated genes was assessed by mutual information of f_c across the genes of that cluster, with high mutual information values indicating non-random distribution and low mutual information values indicating random distribution.

Evaluating the effect of codon usage on the pattern of translation efficiency

The effect of codon usage on protein expression profile was assessed by measuring the amount of expression of two *lacZ* variants under 16 different conditions in yeast cells. One of these two *lacZ* variants was the genomic *lacZ* from *E. coli* K12-MG1665, and the other variant was a synthetic gene with the same protein sequence but extensively different codon usage. Both of these variants were inserted in pBridge (Clontech), and cloned in AH109 yeast strain (Clontech), thus having the same upstream and downstream sequences. The expression of each variant of *lacZ* in each growth condition was measured by a β-galactosidase assay using ortho-nitrophenyl-β-galactoside (ONPG) as substrate. The rate of conversion of ONPG to ortho-nitrophenol was measured by absorbance at 405 nm. This rate, normalized for cell density (estimated by OD_{600}), was considered as the expression level of *lacZ*. Each of the experiments was performed in hexaplicate.

Evaluating the ability of tRNA composition in conferring new phenotypes

A tRNA library was constructed with random combinations of 25 *E. coli* tRNA genes cloned into pBAD18 in *E. coli* host. This library was exposed to different stress conditions, including 6.7 μg/ml kanamycin, 0.5 μg/ml tetracycline, 2.0 μg/ml chloramphenicol, 5.5% ethanol, pH 4.5 and 7.5. The enrichment and/or depletion of particular constructs were assessed through cloning site amplification of the pool of plasmids and visualization on agarose gels. Competition assays were performed by mixing equal volumes of *E. coli* cells carrying tRNA-containing plasmids in a Δ*lacZ* background and $lacZ^+$ *E. coli* cells carrying empty plasmids, inoculating the stress-delivering culture with this mixture, and counting the red and white colony forming units (CFUs) on MacConkey plates after incubating overnight. Selection index R was defined as the ratio of logarithm of growth

of tRNA-containing cells over logarithm of growth of cells carrying empty plasmids.

See 'Methods' section in Supplementary Data for detailed description of mathematical and experimental procedures.

Software packages developed and used in this work are available online at http://webpages.mcgill.ca/staff/Group2/rsalav/web/ICodPack.zip.

RESULTS

A universal correlation between codon usage and function

We examined the relationship between codon usage and function in 785 organisms (including 72 eukaryotes, 661 bacteria and 52 archaea), the sequences and functions of whose genes were retrieved from Kyoto Encyclopedia of Genes and Genomes—KEGG (27). Since paralogous duplicates usually have the same function as well as similar synonymous codon usages, their presence might result in over-estimating the similarity of codon usage among proteins of similar function. Therefore, duplicates were removed within each genome using nucleotide BLAST as described before (2). In this work, we used a relatively large *E*-value cutoff of 0.001 to make sure that all duplicates were removed and the results are unbiased. For each gene, the usages of synonymous codons were calculated and normalized over the usages of their corresponding amino acids, here referred to as f_c. We applied suitable filters to reduce random fluctuations and obtain a robust measure of synonymous codon usage (see 'Methods' section in Supplementary Data). The distance of a pair of genes i and j regarding the usage of codon c was calculated as $d_{ij}(c) = |f_{c,i} - f_{c,j}|$ (26). If genes with similar functions use similar frequencies of synonymous codons, we shall expect a negative correlation between $d_{ij}(c)$ and the likelihood of sharing a biological function; i.e. the more dissimilar the synonymous codon usages of two genes, the less likely they participate in the same pathway. We observed significantly negative linear correlations between d and likelihood of functional linkage in almost all the examined genomes (Supplementary Table S1), indicating a universal pattern in which genes of similar functions have similar usages of synonymous codons. Our set of genomes covered all taxonomic domains, although, in particular, negative correlations were highly significant in eukaryotes (Figure 1).

These results indicate that our previously observed pattern (26) in which functionally interacting proteins use similar codon usages is not restricted to a few organisms; rather, it is a universal characteristic of the genomes across all domains. We will investigate the cause of this pattern in the next section.

Genes with similar expression patterns have similar synonymous codon usages

In the classic view on the relationship between codon usage and protein expression, a constant set of optimal codons is assumed for an organism over different life stages and conditions. This model implies that genes with a particular codon usage should have a translation efficiency that remains constant across different conditions. Assuming that this constant translation efficiency is selected based on the overall expression level of each gene (28), genes with similar codon usages should have similar 'average' expression levels, but not necessarily similar expression 'patterns'. An alternative, unexplored hypothesis can be that the set of optimal codons is not constant, and changes from one condition to another. In this case, the translation efficiencies of genes that have similar codon usages do not remain constant, but change in a synchronized manner in response to the changes of the set of optimal codons. Thus, it is reasonable to assume that such genes would have similar expression 'patterns'.

Knowing that genes with similar functions have similar expression patterns [(29) and the references within], the observed similarity between codon usages of functionally linked proteins led us to reevaluate the two abovementioned hypotheses. We tested these hypotheses on four divergent organisms with available genome-wide expression profiles, human, yeast, *E. coli* and *C. elegans* (30–33). In each organism, clusters of coexpressed genes (i.e. genes with similar expression patterns) were analyzed in the same way as we analyzed the clusters of functionally linked genes in the previous section. Similarly, we also clustered the genes in each organism according to their average expression level, and performed the same analysis. Figure 2 shows that in all of the tested genomes, codon usage has the strongest correlation with expression pattern rather than average expression level, corroborating the 'variable set of optimal codons' hypothesis. Strikingly, this correlation is most obvious for the human genome, where most of the correlations are between -0.80 and -0.90 (Supplementary Table S2) and, with only one exception (correlation between CGU and coexpression), all of them are highly significant ($P \leq 1 \times 10^{-4}$).

We further analyzed the codon usage of each coexpression cluster in human, following the same methodology that has been used before for finding informative regulatory elements (2). We calculated the mutual information of the usage of each codon for each expression profile, and assessed whether the observed mutual information was significantly higher than expected by chance (see 'Methods' section in Supplementary Data for details). A high mutual information value signifies a non-random usage of the corresponding codon among genes within the corresponding coexpression cluster. Figure 3 shows that, in many different coexpression clusters, synonymous codons are used non-randomly; in other words, specific frequencies of synonymous codons are preferred for each expression pattern, resulting in similar synonymous codon usages among the genes that are coregulated. Moreover, this non-random distribution of synonymous codon usage is not merely a result of similar GC content as reported before (14). This is particularly obvious in coexpression clusters whose genes occur in genomic isochores with different GC contents but still show significantly similar synonymous codon usages (the lower half of Figure 3). It should be noted that non-random usage of a codon can be due to either preference for using that codon or preference for not using it (both

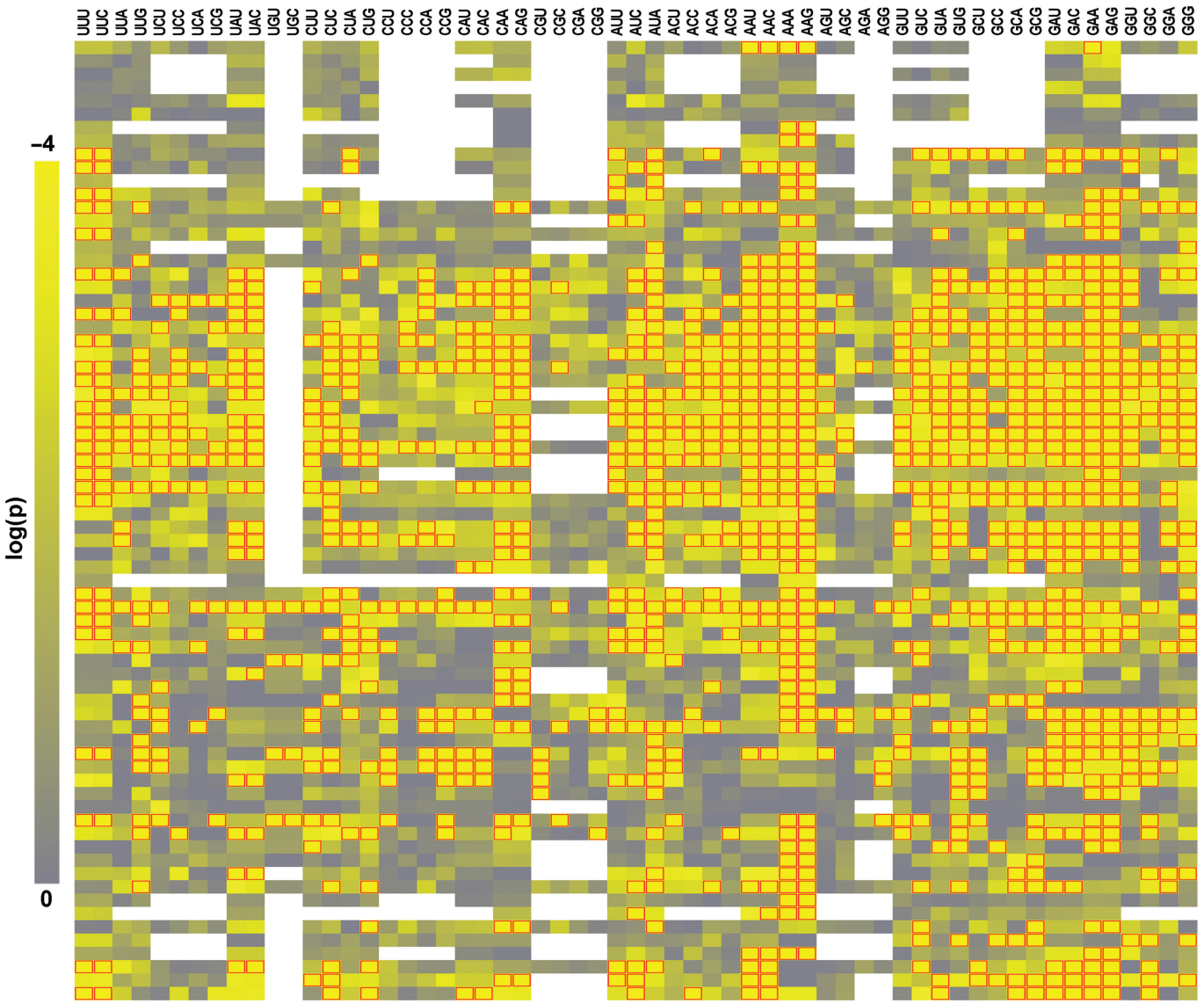

Figure 1. A heat map illustrating the significance of the negative correlations between *d* and likelihood of functional linkage in 72 eukaryotes and 59 codons. Each row represents one organism in the same order as the top 72 rows of Table S1, while each column represents one codon. Stop codons, AUG and UGG are omitted. The *P*-values of the correlations are shown by a color gradient on log scale (left bar), with yellow color standing for small *P*-values. Significant correlations ($P \leq 1 \times 10^{-4}$) are highlighted by red frames, indicating that the corresponding codons are used similarly among the proteins that share the same function. The expected value of false discovery rate (FDR) is $<4 \times 10^{-4}$. White regions stand for cases in which the correlation coefficient could not be calculated due to lack of enough functional linkages. Refer to Supplementary Table S1 for the correlation between codon usage and function in all the 785 organisms studied in this work.

resulting in high mutual information values). An example is shown in Supplementary Figure S1, where some coexpression clusters are over-represented among genes with high frequencies of codon UUU, while some other clusters are over-represented among genes that have low frequencies of UUU. Clustering the human genes based on their 'average' expression level instead of expression pattern, we performed the same analysis and found no significant mutual information values. We also found no significant mutual information between expression pattern and the usage of any amino acid, indicating that the non-randomness of codon usage among coexpressed genes is independent of the amino acid context.

As a complementary method, we also clustered human genes just based on their synonymous codon usage (using 59 codons), and examined whether different coexpression groups show non-random distribution among these clusters. It is shown in Supplementary Figure S1 that many coexpression groups show significantly non-random distribution; each coexpression group is specifically enriched within certain codon usage clusters, while significantly under-represented in other clusters. A similar analysis on *S. cerevisiae* also reveals coexpressed genes with significantly similar synonymous codon usages. Interestingly, these coexpression clusters do not always consist of the most abundant genes. Instead, there

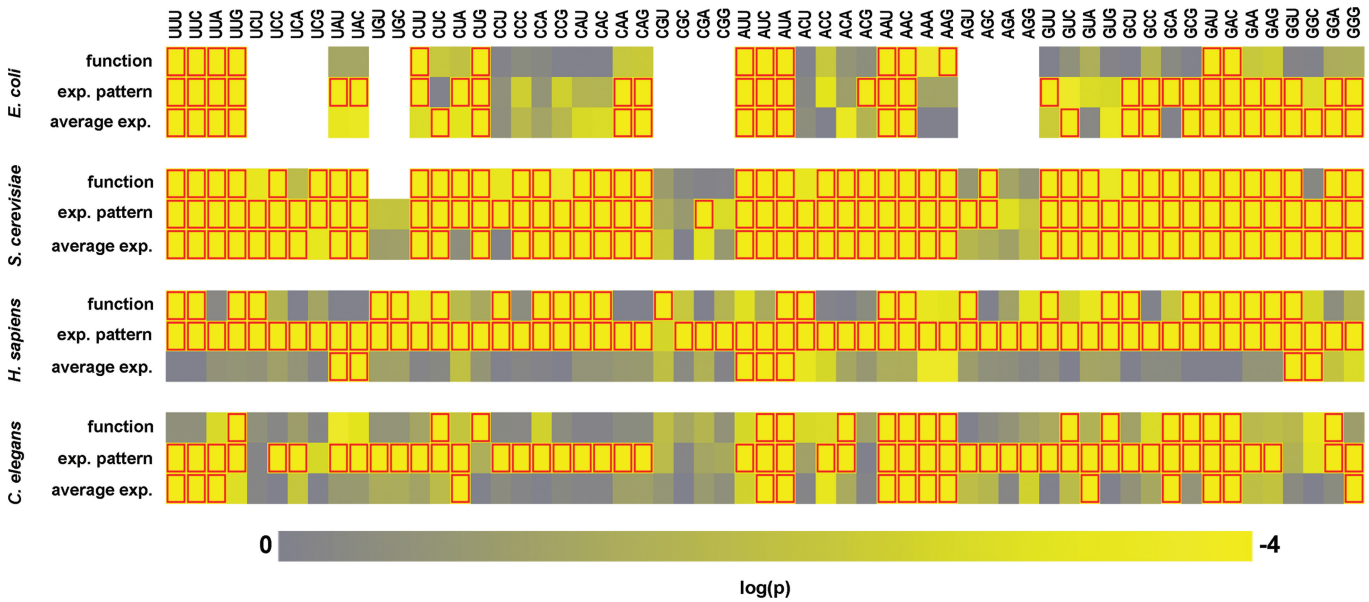

Figure 2. The significance of correlation between codon usage and clusters of genes according to different properties. Genes were clustered according to either function, expression profile (resulting in clusters of coexpressed genes) or average gene expression level (resulting in clusters of genes with similar average expression levels). Functional clusters were obtained from KEGG pathway database (27). Coexpression clusters for *S. cerevisiae*, *Homo sapiens* and *C. elegans* were derived from (2). For *E. coli*, expression profiles of the genes were obtained from (32) and were clustered using Iclust (37). Average gene expression levels were obtained by averaging the expression profile of each gene, except for *S. cerevisiae* where a previously reported reference mRNA level dataset was used (38). The correlation between *d* and the likelihood of occurrence in the same cluster was assessed for each property in each organism for all codons, excluding stop codons, AUG and UGG. Significantly negative correlations are indicated by light-red frames ($P \leq 1 \times 10^{-4}$). Refer to Supplementary Table S2 for the values associated with this figure.

exist many low-abundance coexpressed genes that show significant similarities regarding the usage of several synonymous codons (Supplementary Figure S2), although it is not as noticeable as in human.

Difference in codon usage directly affects regulation of protein expression

We examined whether modification of the codon usage of a gene can change the response of this gene to environmental conditions *in vivo*. To this end, we constructed a modified version of *lacZ* from *E. coli* K12-MG1655, in which the codon usage was changed considerably while keeping the original protein sequence (see 'Methods' section in Supplementary Data). The original *lacZ* [GenBank:U00096, region 362455–365529] and the modified *lacZ* [GenBank:FJ839685] were cloned in yeast, and the expression pattern of LacZ was assessed in 16 different growth and stress conditions, using a quantitative β-galactosidase assay. Although the CAI values of these two genes were different (0.649 for modified *lacZ* compared to 0.213 for original *lacZ*), their average expression levels were not significantly different (paired Student's *t*-test score 0.86, $P < 0.25$). However, as we expected, there were several conditions in which the protein expression was significantly different between the two variants. Particularly, three conditions yielded significantly higher galactosidase activities of the modified *lacZ* compared to the original *lacZ*, while two conditions yielded significantly higher activities of the original *lacZ* (Figure 4).

We propose that codon usage affects the regulation of protein expression by linking it to the regulation of tRNA composition in the cell. In other words, as in different conditions different proteins are required, tRNA composition of the cell may change accordingly in order to accommodate the changing demands for synthesis of new proteins. The response of genes to the new tRNA composition depends on their codon usages; hence, the translation efficiencies of genes with different codon usages change differently, causing the observed difference between the patterns of expression of the two *lacZ* variants. It has to be emphasized that this is a very likely, yet indirect conclusion from our experiment and we did not measure the tRNA content of yeast in the examined 16 conditions. In the next section, we hypothesize that not only the tRNA content may change according to the expression demands of the cell, but also we can change the cell phenotype by engineering the tRNA content.

Changes in tRNA abundance confer adaptative capacity

The results of the previous analyses and experiments suggest indirectly that the tRNA composition of the cell may follow its expression demands. But is tRNA composition also able to push the expression profile of the cell to a different state in order to cause a particular phenotype? We examined this by looking for phenotypic changes that might occur in the cell as a result of perturbation of its tRNA content.

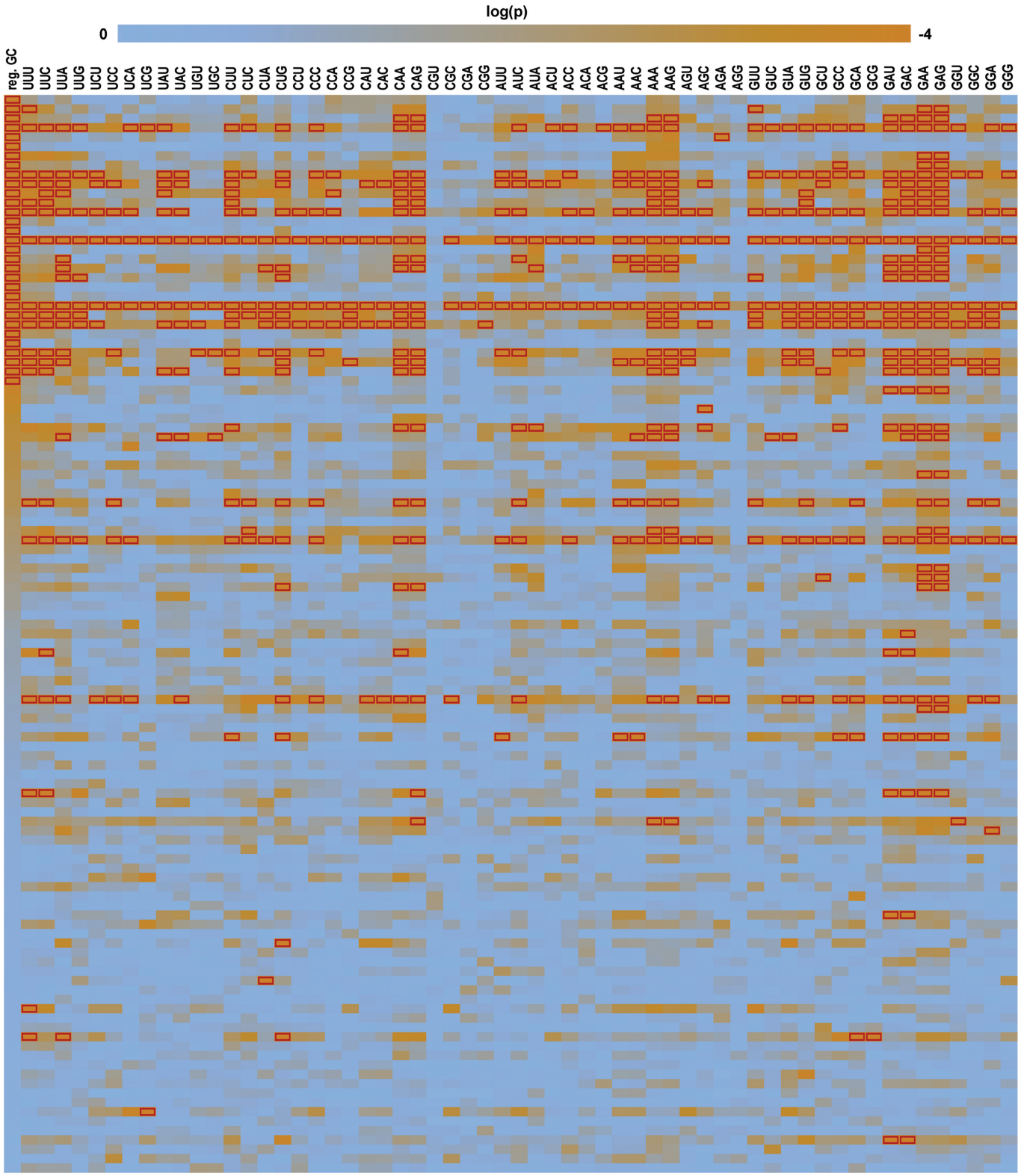

Figure 3. Mutual information of synonymous codon usage in human coexpression clusters is significantly higher than expected from a random distribution. Each row represents a coexpression cluster, while each column, except for the first one from left, represents a codon. Stop codons, AUG and UGG are omitted. Significant mutual information (MI) values are shown by red frames ($P \leq 1 \times 10^{-4}$). Therefore, a red frame around a square indicates that the genes within the corresponding coexpressoin cluster use similar frequencies of the corresponding codon. The expected value of FDR is 1×10^{-3}. MI of regional GC content in each coexpression cluster was assessed similarly (shown in the first column from left). Regional GC content was calculated as the GC content of the 50 kb genomic region surrounding each gene, similar to (14). Coexpression clusters are sorted according to the descending order of the MI of regional GC content; thus, the upper rows represent clusters whose genes have significantly similar regional GC contents, while the lower rows correspond to clusters whose genes occur in different GC contexts.

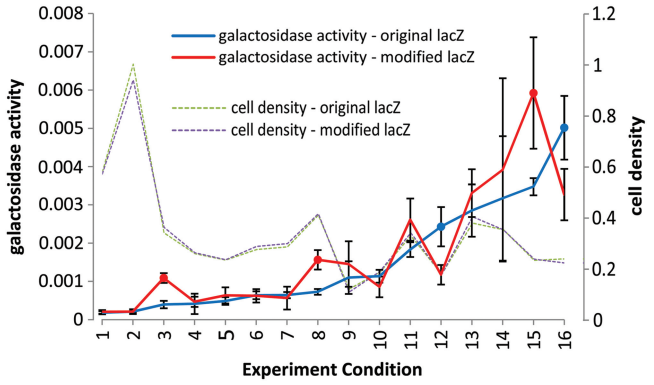

Figure 4. Expression pattern of *lacZ* gene differs as a result of codon usage modification. Two different *lacZ* sequences, i.e. the original gene from *E. coli* K12-MG1665 and a variant with modified codon usage, were expressed in yeast in different conditions, and galactosidase activity was measured. The two yeast strains carrying the two *lacZ* variants did not show any significant differences in their growth pattern (estimated by OD_{600}). However, the galactosidase activity was significantly different between the two strains in several conditions. Galactosidase activity was measured as the rate of increase in OD_{405} in a reaction containing cell extract and ONPG substrate, normalized for cell densitity (see Supplementary Methods). The unit of galactosidase activity in this figure is $OD_{405} \times s^{-1} \times (OD_{600})^{-1}$. Blue circles indicate higher expression of original *lacZ*, while red circles indicate higher expression of modified *lacZ* ($P < 0.05$ with Bonferroni correction for 16 experiments). Experiment conditions: (i) YPDA: 37°C; (ii) YPDA: 30°C; (iii) DTT shock: 30°C; (iv) 2% sucrose: 37°C; (v) DTT shock: 37°C; (vi) 2% ethanol: 37°C; (vii) 2% glucose: 37°C; (viii) 2% glucose: 30°C; (ix) hyper-osmotic shock: 37°C; (x) SD: 37°C; (xi) 2% ethanol: 30°C; (xii) steady 1 M sorbitol: 37°C; (xiii) 2% sucrose: 30°C; (xiv) SD: 30°C; (xv) hyper-osmotic shock: 30°C; (xvi) steady 1 M sorbitol: 30°C. Each experiment was performed in hexaplicate; the standard deviations are depicted by error bars. Quantile and median normalization of expression values do not change the results (data not shown).

Briefly, we constructed a plasmid library in pBAD18 backbone, each carrying a random selection of 25 tRNA genes from *E. coli* (see 'Methods' section in Supplementary Data). From each tRNA gene, a plasmid could have zero, one, or multiple copies. Each copy could be oriented randomly in forward or reverse direction. The rational for this approach was that those sequences cloned in the forward orientation result in an overabundance of the tRNA, while those cloned in reverse most likely decrease the tRNA concentration through double-strand formation with the tRNA transcribed in the cell. This library was transformed into *E. coli*, and the pool of transformed cells was grown under different environmental stresses. The initial frequency of constructs was visualized through cloning site amplification (Supplementary Figure S3). After one or two rounds of selection, the plasmid population was visualized to see whether certain constructs were enriched upon selection. We found two selection conditions in which particular plasmids had highly significant adaptive consequences in a short time-scale (~10 generations): sub-inhibitory concentrations of kanamycin (6.7 μg/ml) and tetracycline (0.5 μg/ml). In the case of kanamycin-containing medium, the enriched plasmid (named pBAD-tKAN) contained one copy of *glyT* tRNA gene in forward direction and one copy of *serW* tRNA gene in reverse direction. On the other hand, growth in the presence of tetracycline results in the enrichment of a plasmid containing *ileX* tRNA gene in forward direction, designated pBAD-tTET (in each case, 10 clones were randomly selected for sequence analysis; see 'Methods' section in Supplementary Data). Repeating the experiment on the library resulted in selection of the same plasmids, indicating that the observed enrichment is selective and is not due to drift. We also confirmed that the selection of these particular tRNA isoacceptors was not a result of bias in the original library; all tRNA isoacceptors of Gly, Ser and Ile were represented in the library in both forward and reverse directions at different combinations (see 'Methods' section in Supplementary Data and Supplementary Figure S3). This shows that, for example, in the presence of tetracycline, *ileX* has a significant fitness advantage over *ileT* since only *ileX* was selected.

Using a competition assay, we observed that in kanamycin-containing medium, *E. coli* cells freshly transformed with pBAD-tKAN have a selection index of about 1.5 over wildtype *E. coli* (MG1655) carrying an empty pBAD18 plasmid ($P < 0.025$; for definition of selection index, see 'Methods' section). pBAD-tKAN confers no growth advantage over empty pBAD18 in tetracycline-containing medium (negative control). Similarly, pBAD-tTET-carrying cells with clean genomic background have a selection index of ~2 in tetracycline-containing medium ($P < 0.001$), and no growth advantage in kanamycin-containing medium, indicating that each plasmid specifically increases the fitness in the medium at which it is selected.

DISCUSSION

We showed that there is a universal correlation between codon usage and gene function, and that this correlation is even more obvious if we consider, instead of function, expression pattern as the basis for clustering the genes within each genome. The best hypothesis that can explain this observation as well as the results of our experiments is that the tRNA composition follows the expression demands of the cell. In other words, if in a particular condition a set of proteins with a particular function are needed and thus are expressed at high levels, tRNA composition changes accordingly to provide the required material. Since this adaptation would best work if in each condition the expressed genes had similar codon usages, a universal function-specificity has emerged in the codon usage within each genome.

The new hypothesis postulates that genes with similar expression patterns, even though having different average expression levels, should have similar codon usages. This is most obvious in organisms with complex developmental and physiological circuits such as human and *C. elegans* (34), in which there is a very strong correlation between codon usage and expression pattern but almost no correlation between codon usage and average expression level (Figure 2). In the case of simpler, fast-growing organisms such as yeast and *E. coli*, it is more difficult to discriminate between our new hypothesis and the conventional view,

since there is a high correlation between average expression level and expression pattern: in these organisms, genes that have similar expression patterns usually show similar average expression levels as well. It is reasonable to think that in microorganisms with high growth rate, both the overall expression level of proteins and the dynamic pattern of expression may contribute to shaping the coding sequence. This is supported by the observation that in many cases the correlation coefficient of codon usage with expression pattern is still significant after correcting for the confounding effect of average expression level and vice versa (Supplementary Figure S4). However, the effect of expression pattern seems to be more profound than average expression level.

There have been previous reports suggesting that, via regulation of tRNA activity, genes with certain codon usages may be regulated in particular conditions (12,35). In this work, we showed that codon usage may have a wider effect on the response of a protein to environmental stimuli: in five out of 16 environmental conditions that we examined, a change in codon usage alters the extent and sometimes even the direction of LacZ response. Since both *lacZ* variants that we used had the same regulatory sequences in their upstream and downstream regions, the simplest explanation for the differences in their expression patterns is a difference in translation efficiency: while in some conditions the original *lacZ* showed greater translation efficiency, in some other conditions the modified *lacZ* exhibited greater translation efficiency. This corroborates the 'variable optimal codon set' hypothesis. This is not however the only explanation; for example, the translation efficiency of *LacZ* may be affected by, in addition to codon usage, the structure of its mRNA. To examine the latter case, we studied the free folding energy for the critical regions of mRNAs of the two *lacZ* variants, and found no significant differences (Supplementary Figure S5). Our results are congruent with a recent report that the folding energy affects overall expression level (24): indeed the average expression levels of the two *lacZ* variants were similar; rather, the expression patterns were different. To our knowledge, there is no known mechanism based on which mRNA structure, without involvement of regulated *trans*-acting factors, could affect the expression pattern.

We also showed that changes in tRNA composition may bring about significant adaptive consequences, such as higher resistance to particular antibiotics. This means that changes in tRNA composition results in tangible phenotypic effects, thus suggesting the possibility that tRNA composition not only follows the expression demands of the cell, but may also change the expression profile of the cell on its own.

The antibiotics that we examined suppress cell growth by inhibition of translation. Thus, it might be argued that the plasmids pBAD-tKAN and pBAD-tTET confer resistance to these antibiotics by overexpression of tRNAs and, hence, generally enhancing translation. However, this scenario seems unlikely due to the nature of the tRNAs that these plasmids carry: both *glyT* and *ileX* encode tRNAs that recognize rare codons in *E. coli* (36) (GGA/G and AUA, respectively). This is while the overall rate of translation would be enhanced much more efficiently if tRNAs that could recognize abundant codons were overexpressed. In fact, overexpression of *glyT* and *ileX* has no direct effect on translation efficiency of many highly demanded genes in *E. coli*, as these genes lack the cognate codons of these tRNAs. Furthermore, selection of the reverse complement of *serW* cannot be explained by enhancement of translation: the reverse complement of *serW* is assumed to inhibit Seryl-tRNA 5 through direct binding. Specificity of pBAD-tKAN and pBAD-tTET in conferring resistance towards only kanamycin and tetracycline, respectively, and not vice versa, adds to the above reasons to conclude that the mechanism of action of these constructs is not through a general enhancement of translation.

It is interesting to see that the combination of *glyT* in forward direction ($glyT_f$) and *serW* in reverse direction ($serW_r$), and not $glyT_f$ or $serW_r$ alone, was selected in the presence of kanamycin. This indicates that the combination of these two is much stronger in conferring kanamycin resistance than any of them alone. Assuming that such cooperative action of tRNAs can have beneficial effects on cell fitness in other situations as well, a regulatory network that manages and synchronizes the activity of different tRNAs can be readily hypothesized. This also suggests that usages of different codons may have coevolved.

Although in this experiment we tested the effect of tRNA concentration on phenotype, cells may potentially change the activity profiles of tRNAs in ways other than changing the concentration as well. For example, enzymatic modification of tRNA has been shown to change the codon preference (12). This mechanism especially seems possible since even for amino acids that have only two codons there is an expression pattern-specific codon usage (Figures 2 and 3). In most organisms, there is only one kind of tRNA for each of the amino acids with two U/C-ending codons. It is thus not possible to change which codon is preferred by varying the concentration of this tRNA, as this tRNA recognizes both of the codons. However, enzymatic modification of tRNA can result in a change in its preference towards its two cognate codons (12), which can describe expression pattern-specificity of codon usage for two-codon amino acids. This indicates that variation of tRNA composition may extend beyond concentration and may include variation of tRNA activity by means such as enzymatic modification as well.

The above experiments suggest a novel approach for modulating functions with applications in biology and biotechnology: we showed that perturbations to a single tRNA gene may change the survival of the cell in certain conditions. It can also be anticipated that certain metabolic pathways will be enhanced by overexpression of tRNAs that recognize the specific codons of those pathways. Also, pathway engineering may benefit from more careful design of coding sequences regarding their codon usage.

Last but not least, the above observations lead to a novel method for prediction of gene expression profiles and gene functions. We employed a naïve Bayesian network to construct a classifier that is able to predict

expression profile or the function of a gene solely based on its codon usage. Using this classifier, we were able to achieve a high sensitivity and specificity in predicting many human coexpression clusters. Furthermore, applying the same classifier on functional groups instead of coexpression clusters showed that some functions can be reliably predicted based on synonymous codon usage (Supplementary Figure S6). We anticipate that this method can considerably enhance homology-independent annotation of genes, especially for genomes whose genes are not conserved in well-studied model organisms (see Supplementary Figure S6 for a few examples in *Trypanosoma brucei*, the causative agent of human African trypanosomiasis).

ACCESSION NUMBER

FJ839685.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Canadian Institutes of Health Research (CIHR) grant #94445 and Leaders Opportunity Fund (LOF) grant #12573 to R.S. Research at the Institute of Parasitology is supported by the Centre for Host-Parasite Interactions and Le Fonds quebecois de la recherche sur la nature et les technologies (FQRNT), Quebec. Lloyd Carr-Harris Fellowship to H.S.N. Funding for open access charge: Canadian Institutes of Health Research (CIHR) (grant #94445).

Conflict of interest statement. None declared.

REFERENCES

1. Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.
2. Elemento,O., Slonim,N. and Tavazoie,S. (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell*, **28**, 337–350.
3. dos Santos,G., Simmonds,A.J. and Krause,H.M. (2008) A stem-loop structure in the wingless transcript defines a consensus motif for apical RNA transport. *Development*, **135**, 133–143.
4. Haile,S. and Papadopoulou,B. (2007) Developmental regulation of gene expression in trypanosomatid parasitic protozoa. *Curr. Opin. Microbiol.*, **10**, 569–577.
5. Merriam,L.C. and Chess,A. (2007) cis-Regulatory elements within the odorant receptor coding region. *Cell*, **131**, 844–846.
6. Lin,X., Parsels,L.A., Voeller,D.M., Allegra,C.J., Maley,G.F., Maley,F. and Chu,E. (2000) Characterization of a cis-acting regulatory element in the protein coding region of thymidylate synthase mRNA. *Nucleic Acids Res.*, **28**, 1381–1389.
7. Wenz,P., Schwank,S., Hoja,U. and Schuller,H.J. (2001) A downstream regulatory element located within the coding sequence mediates autoregulated expression of the yeast fatty acid synthase gene FAS2 by the FAS1 gene product. *Nucleic Acids Res.*, **29**, 4625–4632.
8. Sorensen,M.A., Elf,J., Bouakaz,E., Tenson,T., Sanyal,S., Bjork,G.R. and Ehrenberg,M. (2005) Over expression of a tRNA(Leu) isoacceptor changes charging pattern of leucine tRNAs and reveals new codon reading. *J. Mol Biol.*, **354**, 16–24.
9. 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.
10. Dittmar,K.A., Goodenbour,J.M. and Pan,T. (2006) Tissue-specific differences in human transfer RNA expression. *PLoS Genet.*, **2**, e221.
11. Elf,J., Nilsson,D., Tenson,T. and Ehrenberg,M. (2003) Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science*, **300**, 1718–1722.
12. Begley,U., Dyavaiah,M., Patil,A., Rooney,J.P., DiRenzo,D., Young,C.M., Conklin,D.S., Zitomer,R.S. and Begley,T.J. (2007) Trm9-catalyzed tRNA modifications link translation to the DNA damage response. *Mol. cell*, **28**, 860–870.
13. Plotkin,J.B., Robins,H. and Levine,A.J. (2004) Tissue-specific codon usage and the expression of human genes. *Proc. Natl Acad. Sci. USA*, **101**, 12588–12591.
14. Semon,M., Lobry,J.R. and Duret,L. (2006) No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Mol. Biol. Evol.*, **23**, 523–529.
15. Chamary,J.V., Parmley,J.L. and Hurst,L.D. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Rev.*, **7**, 98–108.
16. Lavner,Y. and Kotlar,D. (2005) Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene*, **345**, 127–138.
17. Comeron,J.M. (2004) Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics*, **167**, 1293–1304.
18. Kotlar,D. and Lavner,Y. (2006) The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids. *BMC Genomics*, **7**, 67.
19. Chamary,J.V. and Hurst,L.D. (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.*, **6**, R75.
20. Duan,J., Wainwright,M.S., Comeron,J.M., Saitou,N., Sanders,A.R., Gelernter,J. and Gejman,P.V. (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Human Mol. Genet.*, **12**, 205–216.
21. Capon,F., Allen,M.H., Ameen,M., Burden,A.D., Tillman,D., Barker,J.N. and Trembath,R.C. (2004) A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Human Mol. Genet.*, **13**, 2361–2368.
22. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev.*, **3**, 285–298.
23. 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.
24. Kudla,G., Murray,A.W., Tollervey,D. and Plotkin,J.B. (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, **324**, 255–258.
25. Garel,J.P. (1974) Functional adaptation of tRNA population. *J. Theor. Biol.*, **43**, 211–225.
26. Najafabadi,H.S. and Salavati,R. (2008) Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biol.*, **9**, R87.
27. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
28. 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.
29. van Noort,V., Snel,B. and Huynen,M.A. (2003) Predicting gene function by conserved co-expression. *Trends Genet.*, **19**, 238–242.
30. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
31. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

32. Faith,J.J., Hayete,B., Thaden,J.T., Mogno,I., Wierzbowski,J., Cottarel,G., Kasif,S., Collins,J.J. and Gardner,T.S. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.
33. Kim,S.K., Lund,J., Kiraly,M., Duke,K., Jiang,M., Stuart,J.M., Eizinger,A., Wylie,B.N. and Davidson,G.S. (2001) A gene expression map for Caenorhabditis elegans. *Science*, **293**, 2087–2092.
34. Dunn,R.K. and Kingston,R.E. (2007) Gene regulation in the postgenomic era: technology takes the wheel. *Mol. Cell*, **28**, 708–714.
35. Kuhar,I., van Putten,J.P., Zgur-Bertok,D., Gaastra,W. and Jordi,B.J. (2001) Codon-usage based regulation of colicin K synthesis by the stress alarmone ppGpp. *Mol. Microbiol.*, **41**, 207–216.
36. Baca,A.M. and Hol,W.G. (2000) Overcoming codon bias: a method for high-level overexpression of Plasmodium and other AT-rich parasite genes in Escherichia coli. *Int. J. Parasitol.*, **30**, 113–118.
37. Slonim,N., Atwal,G.S., Tkacik,G. and Bialek,W. (2005) Information-based clustering. *Proc. Natl Acad. Sci. USA*, **102**, 18297–18302.
38. Jansen,R., Bussemaker,H.J. and Gerstein,M. (2003) Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res.*, **31**, 2242–2251.