The Signatures of Selection for Translational Accuracy in Plant Genes

Andrea Porceddu[1,*], Sara Zenoni[2], and Salvatore Camiolo[1]

[1]Dipartimento di Agraria (Sezione di Agronomia e Coltivazione Erbacee Genetica-SACEG) Università degli studi di Sassari, Italy

[2]Dipartimento di Biotecnologie, Università degli studi di Verona, Italy

*Corresponding author: E-mail: aporceddu@uniss.it.

Accepted: May 16, 2013

Abstract

Little is known about the natural selection of synonymous codons within the coding sequences of plant genes. We analyzed the distribution of synonymous codons within plant coding sequences and found that preferred codons tend to encode the more conserved and functionally important residues of plant proteins. This was consistent among several synonymous codon families and applied to genes with different expression profiles and functions. Most of the randomly chosen alternative sets of codons scored weaker associations than the actual sets of preferred codons, suggesting that codon position within plant genes and codon usage bias have coevolved to maximize translational accuracy. All these findings are consistent with the mistranslation-induced protein misfolding theory, which predicts the natural selection of highly preferred codons more frequently at sites where translation errors could compromise protein folding or functionality. Our results will provide an important insight in future studies of protein folding, molecular evolution, and transgene design for optimal expression.

Key words: coding sequences evolution, codon bias, constrained sites.

Introduction

The efficiency of mRNA translation is controlled at multiple levels by different mechanisms. A major part of the controls occurs at the time of ribosome recruitment, which culminates in the formation of the initiation complex (Ingolia et al. 2009). Chain elongation is then controlled by ribosomal intrinsic factors, mRNA secondary structure (Gray and Hentze 1994), and the adaptation of the coding sequence to the cellular availability of tRNA pools (Sharp 1987; dos Reis et al. 2004). A proposed strategy for compositional adaptation is the preferential use of synonymous codons that correspond to the most abundant tRNAs (Sharp 1987). In practice, if there is a relationship between the abundance of aminoacyl tRNAs and the time taken to occupy the acceptor site on the ribosome, then codons corresponding to abundant tRNAs could be translocated faster than other synonymous codons (Arava 2003). Codons recognized by abundant tRNAs are thus likely to have a high translational efficiency and would be preferred in highly expressed genes, whereas those recognized by rare tRNAs would cause translational bottlenecks. Indeed, the observation that in species as diverse as *Drosophila*, *Escherichia coli*, and humans, the preferred synonymous codons are preferentially found in highly expressed genes lent support to such hypothesis (Percudani 1999; dos Reis et al. 2004; Wright et al. 2004; Pál et al. 2006). Furthermore, synthetic coding sequences in which rare codons have been deliberately mutated into their preferred synonymous counterparts are expressed at higher levels than the wild-type sequences (DeRisi et al. 1997; Arava 2003; Tuller et al. 2010, 2011).

Because aminoacyl tRNAs compete at the ribosome acceptor site until the correct one is stably installed, abundant tRNAs (and the corresponding codons) may therefore be associated with a lower rate of mistranslation errors (Marais and Duret 2001; Akashi 2003; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2009). This is particularly important under conditions requiring high levels of protein synthesis and accumulation. It has been estimated that missense errors in translation occur every 10^3–10^4 codons (Parker 1989; Kramer and Farabaugh 2007). Taking the average error rate of 5×10^{-4} and an average polypeptide length of 400 amino acids, this means approximately 18% of all proteins are expected to contain at least one missense substitution that may cause misfolding or loss of function (Drummond and

Wilke 2008, 2009). The total effect of mistranslation could be even stronger because misfolded proteins may expose normally buried residues that seek the nonpolar surfaces of other misfolded proteins, promoting cytotoxicity through protein aggregation (Bucciantini et al. 2002).

A conceptual framework to identify the signatures of selection for translational accuracy was originally proposed by Akashi (1994). Residues necessary for correct protein folding and/or function should be evolutionarily constrained, and their position can therefore be inferred by the alignment of orthologs (Mirny and Shakhnovich 1999; Schueler-Furman and Baker 2003). Other criteria for the identification of such important sites include solvent accessibility (Zhou et al. 2009) and localization within functional domains (Akashi 1994, 1998; Zhou et al. 2009). If natural selection biases codon usage to enhance the accuracy of translation then synonymous codons corresponding to the most abundant tRNA should be favored at functionally constrained sites than at the less constrained ones (Akashi 1994). An intrinsic difficulty in these studies is that other intragenic patterns of codon usage may partially overlap and therefore obscure any pattern generated by selection for translational accuracy (Wong et al. 2002; Niimura et al. 2003; Qin et al. 2004; Tuller et al. 2010). For example, Tuller et al. (2010) demonstrated that the proximal coding regions of genes from diverse species are rich in nonpreferred synonymous codons, probably reflecting an adaptive feature that ensures optimal ribosome density along the transcript and optimizes translation speed by minimizing the risk of collisions between ribosomes. Other patterns can be species dependent or conserved in several species but species dependent in terms of extent in the coding sequence (Wong et al. 2002; Niimura et al. 2003; Qin et al. 2004; Porceddu and Camiolo 2011). For example, Wong et al. (2002) have described the variability of intragenic patterns of codon usage bias among different plant species. The codon usage bias increases along the direction of translation in monocots genes, whereas it appears rather constant in dicots genes (Wong et al. 2002).

We analyzed the coding sequences of several plant species to determine whether preferred codons are distributed differently within the evolutionarily constrained and variable regions of plant genes, whether the association is similar for all amino acids, whether it is dependent on the gene expression profile, and whether preferred codons and evolutionarily constrained sites have coevolved to maximize translation efficiency. We found that the intragenic arrangements of preferred codons are associated with the evolutionary constraints of protein sites regardless of gene expression or function and that the set of preferred codons was more frequently associated with evolutionarily constrained sites than randomly chosen codon sets, suggesting an interaction between conserved sites and preferred codons to enhance translation efficiency. Our data, therefore, demonstrate that the codon composition of plant genes is affected by selection for translational accuracy.

Materials and Methods

Sequence Data, Protein Functional Annotation, and Gene Families

Coding sequences and corresponding polypeptide sequences were downloaded from Phytozome (http://www.phytozome.net; Goodstein et al. 2012). The annotations of *Arabidopsis thaliana* functional/structural domains were downloaded from The Arabidopsis Information Resource (ftp://ftp.arabidopsis.org/home/tair/proteins/Domains/). Pairs of orthologous genes were identified using Inparanoid with standard settings (Ostlund et al. 2010). Only genes with unique orthologs in the other species were considered (one-to-one orthologous).

Analysis of Gene Expression

Gene expression profiles were measured by expression breadth (taking into account the number of tissues in which each gene is expressed) and expression level (based on mean microarray hybridization signals measured for the probe sets corresponding to each locus). We employed the *A. thaliana* gene expression data set (AT40) from Schmid et al. (2005) retrieved from PlexDB (www.plexdb.org). Each probe set was hybridized three times in a given experiment, and genes were considered to be expressed when the corresponding probe set was detected significantly in all replicates and the hybridization signal was never below 100 (technical threshold). When these conditions were met, the hybridization signals in the three replicates were averaged to obtain the final expression level. For rice, we elaborated the microarray data from Jain et al. (2007) (http://www.plexdb.org/modules/PD_browse/experiment_browser.php?experiment=OS5). Additional data sets and procedures used to identify the preferred codon list are provided as supplementary material, Supplementary Material online. For the analysis of the effect of gene expression level on translational accuracy, we analyzed genes belonging to different classes of either breadth (EB) or level of expression (EL). Three classes of EB were considered: high ($EB > 0.9$), intermediate ($0.4 < EB < 0.6$), and low ($EB < 0.15$). For the EL, the data set, sorted based on DS-I, was divided in 10 percentiles, and genes were classified as highly expressed (percentiles 8–10), low expressed (percentiles 1–2), and intermediately expressed (percentiles 5–6).

Identification of Preferred Codon Sets

A synonymous codon was described as preferred if it was used at a significantly higher frequency in strongly expressed compared with weakly expressed genes. The high and low gene expression groups were identified by considering the top and bottom 10% of genes in the data set ranked by expression level (percentile method). More information on the percentile method and alternatives for the identification of preferred codons are provided in the supplementary material, Supplementary Material online. The preferentiality of each

codon was calculated as $Codon_{pref} = [n_{high}/(N_{high} - n_{high})]/[n_{low}/(N_{low} - n_{low})]$, where n_{high} and n_{low} are the observed number of codons in the high and low gene expression groups, respectively, and N_{high} and N_{low} are the observed numbers of the corresponding amino acids in the high and low gene expression groups, respectively (Zhou et al. 2009).

Statistical Test of Association

Orthologous protein sequences were aligned using MUSCLE with standard settings (Edgar 2004). We considered all sites with conserved amino acids in each orthologs as evolutionarily constrained.

The list of *A. thaliana* preferred synonymous codons used in our analysis is presented in supplementary table S1, Supplementary Material online. The list of rice preferred codons (supplementary table S2, Supplementary Material online) was defined by considering separately the data sets of high-GC and low-GC genes. Loci were attributed to the high-GC data set if the total GC content was more than 65% and to the low-GC data set if it was less than 60% (Guo et al. 2007).

Contingency tables (2×2) were constructed for each amino acid by considering the frequency of preferred and nonpreferred codons encoding evolutionarily conserved and variant residues in each aligned protein. For global association analysis, the tables relating to all genes and all amino acids were combined using the Mantel–Haesenzl (MH) *Z* statistic as suggested by Akashi (1994).

Nonsynonymous mutations may change an optimal codon to nonoptimal one. In theory, this could produce the conditions for the presence of a high frequency of nonoptimal codons at unconstrained sites and, hence, influence the estimations of selection for translational accuracy. To control for this effect, the associations were recalculated taking into account nonconserved amino acids for which all intermediate states in all reconstructed pathways (between the extant codons) showed the same silent bases and the same favored silent base(s).

Because the MH procedure, using contingency tables whose sum of all four entries is less than 2, yields undefined results, these tables were excluded from further analysis (Zhou et al. 2009). The analysis of selection for translational accuracy was carried out using Sephora software (Camiolo et al., submitted). Association analysis relating to each amino acid was carried out as above but using contingency tables relating to individual amino acids.

Similarly, association analysis relating to each individual codons was carried by analyzing contingency tables in which each codon within a synonymous family was treated as preferred and other synonymous codons as nonpreferred. Finally, alternative lists of preferred codons were constructed by treating codons randomly extracted from each family of synonymous codons as preferred. The number of codons extracted from each family was always equal to the number of preferred synonymous codons in the actual set of preferred codons

Results

Preferred Codons Are Favored at Evolutionarily Constrained Sites

We assessed the relationship between preferred codons and evolutionarily constrained sites in proteins from two model plant species: the dicot *A. thaliana* and the monocot *Oryza sativa* (rice).

The positions of evolutionarily constrained sites in *A. thaliana* proteins were inferred by looking at conserved residues in alignments with rice orthologs. For each amino acid, we constructed a 2×2 contingency table to summarize the usage of preferred and nonpreferred synonymous codons in either conserved or nonconserved residues within each alignment. A global test was obtained by combining the tables across all amino acids and all genes using the MH procedure (Akashi 1994). In this design, an MH odds ratio (MH_{OR}) greater than 1 signifies that preferred synonymous codons are used more frequently to encode the amino acids at evolutionarily constrained sites compared with variable sites. The MH test on all amino acids and all genes in the data set was significant ($P < 0.001$) with an MH_{OR} of 1.06 (table 1). The degree of association remained qualitatively unchanged when *A. thaliana* and *A. lyrata* orthologous alignments were used to infer the positions of constrained and variable residues ($MH_{OR} = 1.08$, $P < 0.001$; table 1).

The positions of conserved sites in rice proteins were inferred by alignments with *A. thaliana* orthologs. Carels and Bernardi (2000) demonstrated that the distribution of the GC content of rice coding sequences is bimodal and that the two classes of genes also differ for several structural parameters. Accordingly, Mukhopadhyay et al. (2008) suggested that GC-rich and GC-poor rice coding sequences may experience different selective pressures, suggesting these genes should be analyzed separately. On the basis of these considerations, we partitioned the rice coding sequences into two subdata sets: the first containing sequences with a GC content more than 65% (hereafter the high-GC data set) and the other containing sequences with a GC content less than 60% (low-GC data set). The two subdata sets were analyzed separately using the lists of preferred codons calculated from expression data relative to genes with either high- or low-GC contents ($Pref_{High}$ and $Pref_{Low}$). We also analyzed the whole data set of rice genes with *Arabidopsis* orthologs, assuming that codons common to the two lists were preferred ($Pref_{Both}$). The tests on the whole data set ($Pref_{Both}$ $MH_{OR} = 1.09$; $Z = 18.08$***) and on the low-GC data set ($Pref_{Low}$ $MH_{OR} = 1.12$; $Z = 26.51$***) showed that preferred codons tended to encode evolutionarily constrained residues. The high-GC data set also gave an MH_{OR} value greater than unity although the *Z* score was lower ($Pref_{High}$ $MH_{OR} = 1.06$;

Table 1
Global Test for Translational Accuracy in Plant Genes

Orthologous from	Alignments with	Whole Coding Sequence		Lacking First 100 Residues	
		Odds Ratio	*Z*	Odds Ratio	*Z*
Arabidopsis thaliana	*Oryza sativa*	1.06	22.36***	1.07	25.77***
A. thaliana	*A. lyrata*	1.08	18.45***	1.09	18.87***
O. sativa$^{3(both)}$	*A. thaliana*	1.09	18.08***	1.12	21.2***
O. sativa$^{(both)}$	*Brachypodium distachyon*	1.12	21.02***	1.13	20.51***
O. sativa$^{(low\ GC)}$	*A. thaliana*	1.12	26.51***	1.08	14.71***
O. sativa$^{(low\ GC)}$	*B. distachyon*	1.11	22.27***	1.06	12.30***
O. sativa$^{(high\ GC)}$	*A. thaliana*	1.06	4.02***	1.04	2.35**
O. sativa$^{(high\ GC)}$	*B. distachyon*	1.06	6.22***	1.03	2.77**

NOTE.—An odds ratio greater than 1 dictates the preferential usage of preferred codons to encode evolutionarily constrained residues. The positions of evolutionarily constrained residues in *A. thaliana* proteins were identified from alignments between *A. thaliana* and *O. sativa* or *A. lyrata* orthologs. Evolutionarily conserved residues in rice proteins were identified from alignments between *O. sativa* and *A thaliana* or *B. distachyon* orthologs.

**$P < 0.01$.

***$P < 0.001$.

$Z = 4.02$***). These results were confirmed by analogous tests with orthologs from rice and *Brachypodium distachyon* (table 1).

Lipman and Wilbur (1985) suggested an alternative explanation for the association between codon usage and conservation of amino acids. When preferred silent base differs between synonymous families, then replacement mutations may change a codon from preferred to unpreferred. For example, the replacement from GTC to GGC in *A. thaliana* will change the preferred codon for Valine to an unpreferred codon for Glycine. Such type of event could add to mutation pressure away from preferred codons at unconstrained sites (Akashi 1994). To test whether such an effect could account for the observed associations, we limited our analysis to unconstrained sites that have retained the same silent bases and the same favored base(s) in all reconstructed pathways between extant codons (Akashi 1994). Although the sample sizes were reduced, the significance of the associations reported remained highly significant: The odds ratio calculated from alignments between *A. thaliana* and *A. lyrata* orthologous was 1.17 with a *Z* score of 16.23 ($P < 0.001$), whereas the tests calculated on alignments between rice and *B. distachyon* gave an odds ratio of 1.18 ($Z = 23.88$; $P < 0.001$) and 1.45 ($Z = 8.23$; $P < 0.001$) for rice low-GC and high-GC data sets, respectively (table 2).

It is important to recognize that this type of analysis is dependent on the list of preferred codons, which may in turn reflect the specific gene expression data set used to identify them. To determine the impact of this factor on our results, we deduced several lists of optimal codons by employing different data sets of gene expression for both rice and *Arabidopsis*. Despite the slight differences between these lists (see supplementary tables S1 and S2, Supplementary Material online), MH_{OR} was always significantly greater than unity (data not shown).

Several authors have reported nonrandom codon usage in the proximal coding sequences of many genes (Niimura et al. 2003; Qin et al. 2004; Porceddu and Camiolo 2011; Tuller et al. 2011). To rule out any influence of such patterns on the association between codon usage and evolutionarily constrained residues, we reanalyzed the associations after excluding from the calculations the first 100 N-terminal residues in the protein alignments. In all cases, the level of association and its significance remained almost unchanged (table 1). The picture was confirmed also when only unconstrained sites involving amino acids encoded by codons with the same silent bases and the same favored base(s) were considered (table 2).

We next determined whether the signature of selection for translational accuracy could also be detected using tests based on the functional annotation of *Arabidopsis* proteins. These annotations predict the position of structurally and/or functionally critical sites and thus are expected to be evolutionary constrained. We compiled 2×2 contingency tables considering codon counts either within or outside protein domains identified based on annotations generated using the ProDom (Servant et al. 2002) or ScanProsite resources (De Castro et al. 2006). The positive and significant association between preferred codons and residues included in protein domains was confirmed in both analyses (table 3).

Selection for Translational Accuracy Is Consistent across Several Synonymous Codon Families and across Genes with Different Expression Profiles

We determined whether the association between the use of preferred codons and evolutionarily conserved residues was consistent across all amino acids by carrying out a separate analysis for each amino acid. We found that 11 of

Table 2

Global Test for Translational Accuracy in Plant Genes after Controlling for the Lipman-Wilbur Effect

Orthologous from	Alignments with	Whole Coding Sequence		Lacking First 100 Residues	
		Odds Ratio	*Z*	Odds Ratio	*Z*
Arabidopsis thaliana	*A. lyrata*	1.17	16.23***	1.08	15.94***
Oryza sativa$^{(low\ GC)}$	*Brachypodium distachyon*	1.18	23.88***	1.11	13.52***
O. sativa$^{(high\ GC)}$	*B. distachyon*	1.45	8.23***	1.29	4.5***

Note.—An odds ratio greater than 1 dictates the preferential usage of preferred codons to encode evolutionarily constrained residues. Only unconstrained sites involving amino acids encoded by codons with the same silent bases and the same favored base(s) were considered. The positions of evolutionarily constrained and unconstrained residues in *A. thaliana* proteins were identified from alignments between *A. thaliana* and *A. lyrata* orthologs. Evolutionarily conserved residues in rice proteins were identified from alignments between *O. sativa* and *B. distachyon* orthologs.

***$P < 0.001$.

Table 3

Selection for Translational Accuracy for Residues Included in the Functional Domains of *Arabidopsis thaliana* Proteins

Algorithm	All Genes	
	Odds Ratio	*Z*
ProDrom	1.106	5.18***
PatternScan	1.05	7.44***

***$P < 0.001$.

Table 4

Signatures of Translational Selection Are Consistent for Most Amino Acids in Plant Proteins

Residue	*Arabidopsis*		Rice (Low GC)		Rice (High GC)	
	Odds Ratio	*Z*	Odds Ratio	*Z*	Odds Ratio	*Z*
Ala	1.05	5.97***	1.24	16.09***	1.03	0.85
Cys	1.20	9.64***	ND	ND	1.43	1.37
Asp	1.01	1.40	ND	ND	ND	ND
Glu	1.14	16.21***	1.17	10.70***	1.65	4.70***
Phe	1.03	3.18**	ND	ND	ND	ND
Gly	0.93	6.54***	1.04	2.20*	0.75	2.84*$^{(*)}$
His	1.05	3.35**$^{(*)}$	ND	ND	1.19	1.18
Iso	0.99	1.23	0.92	4.68***	1.30	2.46*
Leu	1.13	16.72***	1.21	12.93***	1.05	1.37
Asn	1.11	9.72***	ND	ND	2.14	5.63***
Pro	0.92	4.86***	1.12	6.10***	1.16	2.82*$^{(*)}$
Gln	1.31	23.11***	1.24	10.69***	1.69	3.47**$^{(*)}$
Arg	1.05	4.19***	1.109	5.76***	1.01	0.230
Ser	0.95	4.29***	ND	ND	1.18	3.88**$^{(*)}$
Thr	0.93	7.65***	0.984	0.58	0.96	0.70
Val	1.00	0.14	1.09	6.20***	0.92	2.05$^{(*)}$
Tyr	1.13	8.72***	ND	ND	0.69	1.80
Lys	1.11	12.42***	1.16	9.37***	1.36	2.31$^{(*)}$

Note.—ND, not determined. Significance level in parentheses disappears after Benjamini–Hochberg correction for multiple testing (Benjamini and Hochberg 1996).

*$P < 0.05$.

**$P < 0.01$.

***$P < 0.001$.

Table 5

Signatures of Selection for *Arabidopsis thaliana* Genes with Different Expression Profiles (Breadth and Level of Expression)

	Odds Ratio	*Z*
Expression breadth low	1.06	6.36***
Expression breadth intermediate	1.06	5.52***
Expression breadth high	1.05	14.54**
Expression level low	1.06	11.075**
Expression level intermediate	1.05	7.32***
Expression level high	1.04	6.75***

**$P < 0.01$.

***$P < 0.001$.

the 18 degenerate amino acids showed a positive association ($MH_{OR} > 1$ and $P < 0.001$) between preferred codon use and constrained residues in *A. thaliana* proteins (table 3). In rice, we found that 8 out of 11 amino acids showed a significant ($P < 0.01$) association between preferred codon use and constrained residues in the low-GC data set, whereas 13 of 16 amino acids showed significant association in the high-GC data set, and only four were significant at $P < 0.01$ (table 4).

Previous reports indicated that the gene expression patterns may affect selection for translational efficiency. For example, Wright et al. (2004) demonstrated that coding sequence adaptation correlates with the expression level of *Arabidopsis* genes and provided evidence that the rate of both synonymous and nonsynonymous substitutions is inversely correlated to the expression breadth of a gene. To determine whether we could detect signatures of selection for translational accuracy in genes with different expression profiles, we reanalyzed groups of *Arabidopsis* genes with different expression profiles in terms of expression breadth and expression level. Interestingly, the association between preferred codons and evolutionarily conserved residues was confirmed for all classes of genes regardless of expression breadth or expression level (table 5).

Some Nonpreferred Codons Are Favored at Evolutionarily Constrained Sites

Although some amino acids showed indications for the selection of preferred codons, others did not indicating that in some cases the preferred codons are not favored at evolutionarily constrained sites. We thus gained further insight on the relationship between codon's tendency to be preferentially used in high expressed genes and the same codon's tendency to be used to encode constrained sites within proteins. Two odds ratios were defined for each of the 59 codons (table 6). The first odds ratio, hereafter referred as codon preferentiality, measured whether a codon was preferentially used in highly expressed genes than in low expressed ones compared with other codons encoding the same amino acid (Zhou et al. 2009). The second odds ratio (MH odds ratio) measures the tendency of a given codon to be preferentially used to encode constrained sites compared with other synonymous codons. The two odds ratio were positively and significantly correlated in both *Arabidopsis* ($\rho = 0.5$ $P < 0.01$) and rice data sets ($\rho = 0.42$, $P < 0.01$ for the rice low-GC data set; $\rho = 0.34$, $P < 0.01$ for the rice high-GC data set). Interestingly, we identified cases of nonpreferred codons (odds preferentiality < 1) showing significant positive association with evolutionarily conserved sites (odds preferentiality > 1). For example, in *A. thaliana* genes, the ACA codon showed the highest odds ratio among synonymous codons ($MH_{OR} = 1.15$, $P < 0.05$) although the odds ratio of codon preferentiality was 0.73. Similar cases were noticed in rice although with lower significance, for example, TCA in the low-GC data set (table 6).

The Set of Preferred Codons and the Codon Position within Plant Genes Interact to Enhance Translational Accuracy

The results presented earlier suggest that alternative codon sets (rather than the set of preferred codons) may better explain the codon composition at evolutionarily constrained sites in plant proteins, questioning whether such associations have arisen by chance. To test this hypothesis, preferred and nonpreferred synonymous codons were shuffled to generate 2×10^6 random codon combinations that were individually used to calculate MH_{OR} using aligned *Arabidopsis* and rice orthologs. Interestingly, the results revealed that only 1.5% of the randomly defined lists of preferred codons in *A. thaliana* proteins achieved an MH_{OR} value higher than the combination containing only the preferred synonymous codons. In rice, the result depended on which GC data set was considered. For the high-GC data set, up to 42.3% of the randomly defined lists of preferred codons achieved an MH_{OR} value higher than the list of preferred codons, but this proportion fell to 1.7% when the low-GC data set was used.

Discussion

We have shown that selection for translational accuracy affects codon usage in plant genes. The signatures representing this form of selection were identified as positive associations between preferred codons and genic regions encoding evolutionarily constrained sites in proteins, assuming that preferred codons should be beneficial at sites where substitutions can be most harmful (Akashi 1994; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2009). These associations did not occur by chance during evolution. Indeed, the preferred codon sets in both *A. thaliana* and rice (low-GC data set) produced higher association scores than the majority of randomly generated sets. These finding indicate that both the set of preferred codons and the codon position within genes interact to enhance translational accuracy.

Because all the tests used in this study considered intragenic patterns of codon bias, the conclusions should not be affected by regional mutational biases or selection mediated by the level of gene expression (Akashi 1994). However, other intragenic patterns of codon biases cannot be assumed a priori to have no influence on our results. Codon bias is stronger in the proximal part of the coding sequence in many species, including plants (Qin et al. 2004; Wong et al. 2002; Niimura et al. 2003; Tuller et al. 2010). Tuller et al. (2010) have suggested that such a phenomenon could function as a ramp to achieve the optimal ribosome density on the mRNA.

We controlled for this effect by excluding the first 100 N-terminal amino acids from the calculation, and in both *A. thaliana* and rice, the association between optimal codons and evolutionarily constrained residues was confirmed. Other patterns of intragenic codon bias may reflect the impact of selection for mRNA folding (Mukhopadhyay et al. 2008). Carels and Bernardi (2000) have demonstrated that the GC content of rice coding sequences is bimodal and that GC-rich genes are shorter and have fewer introns than GC-poor ones. Mukhopadhyay et al. (2008) proposed that GC-rich genes have a codon bias that is influenced predominantly by selection for mRNA folding, whereas the codon bias of GC-poor genes would depend more on translational selection. On the basis of these considerations, we analyzed separately the two classes of rice genes and found a significant association between preferred codons and evolutionarily constrained residues in GC-poor genes, although the same association was present albeit with lower significance in the GC-rich data set. However, the association observed in the high-GC data set should be interpreted with caution. Up to 42.3% of the 2×10^6 randomly defined sets of preferred codons were associated with constrained residues and achieved higher scores than the actual set of preferred codons. It is, therefore, possible that the compositional properties of high-GC rice genes are strongly influenced by

Table 6

Codon Preferences and Translational Accuracy in *Arabidopsis thaliana* and Rice Genes (Low-GC Data Set)

Amino Acid	Codon	*A. thaliana*		*Oryza sativa* (Low GC)	
		Preferentiality (High vs. Low)	Accuracy (Akashi Test)	Preferentiality (High vs. Low)	Accuracy (Akashi Test)
Ala	GCA	0.73***	0.97	0.88**	1.24**
	GCC	1.23***	0.99	0.86*	0.79**
	GCG	0.94	0.94	0.95	0.66***
	GCT	1.19***	1.05*	1.24***	1.25**
Cys	TGC	1.13*	1.20**	1.09	1.16*
	TGT	0.88*	0.82**	0.92	0.86*
Asp	GAC	1.35***	1.01	1.01	0.98
	GAT	0.74***	0.99	0.99	1.02
Glu	GAA	0.71***	0.87***	0.88**	0.85**
	GAG	1.41***	1.14***	1.13**	1.17**
Phe	TTC	1.60***	1.03	1.14*	1.08
	TTT	0.62***	0.96	0.87*	0.92
Gly	GGA	1.06*	1.05*	0.93	1.19**
	GGC	0.79***	0.93*	0.90	0.78**
	GGG	0.73***	1.08*	0.97	1.03
	GGT	1.27***	0.93*	1.18***	1.04
His	CAC	1.57***	1.05	0.91	0.97
	CAT	0.64***	0.95	1.10	1.03
Ile	ATA	0.50***	1.05*	0.80***	1.02
	ATC	1.58***	0.98	1.07	1.07
	ATT	1.06*	0.97	1.14*	0.92
Leu	CTA	0.78***	1.03	0.86*	1.05
	CTC	1.37***	1.05*	0.93	0.91*
	CTG	0.92*	1.07*	0.96	1.00
	CTT	1.20***	1.12**	1.29***	1.22**
	TTA	0.61***	0.84***	0.83**	0.89*
	TTG	1.03	0.90**	0.99	0.91*
Asn	AAC	1.56***	1.11**	1.02	1.17*
	AAT	0.64***	0.89**	0.98	0.85*
Pro	CCA	0.96	0.99	0.96	1.19**
	CCC	1.08	0.92*	0.92	0.89*
	CCG	0.91*	0.98	0.81*	0.69**
	CCT	1.05	1.04	1.16**	1.12*
Gln	CAA	0.78***	0.76***	0.82***	0.81**
	CAG	1.29***	1.31***	1.22***	1.24**
Arg	AGA	0.79***	0.84***	0.81***	0.90*
	AGG	1.22***	1.05	1.01	0.90*
	CGA	0.75***	1.05	0.97	1.26*
	CGC	1.13*	1.07	1.04	0.93
	CGG	0.64***	1.11*	0.94	1.02
	CGT	1.64***	1.09*	1.46***	1.33**
Ser	AGC	1.03	1.00	0.99	0.84**
	AGT	0.87***	0.86***	0.98	0.79**
	TCA	0.86***	1.11**	0.95	1.19**
	TCC	1.19***	0.95	1.01	1.09*
	TCG	1.06	1.04	1.24***	1.03
	TCT	1.09***	1.00	1.00	1.05

(continued)

Table 6 Continued

Amino Acid	Codon	*A. thaliana*		*Oryza sativa* (Low GC)	
		Preferentiality (High vs. Low)	Accuracy (Akashi Test)	Preferentiality (High vs. Low)	Accuracy (Akashi Test)
Thr	ACA	0.73***	1.15**	0.95	1.14*
	ACC	1.48***	0.88**	1.01	0.89*
	ACG	0.77***	0.90*	0.87***	0.94
	ACT	1.14***	1.00	1.09	0.99
Val	GTA	0.54***	0.83***	0.80***	0.86*
	GTC	1.29***	1.00	1.00	1.02
	GTG	1.06*	1.08*	0.92	0.98
	GTT	1.11***	1.03	1.21***	1.10*
Tyr	TAC	1.76***	1.13**	1.08	1.12
	TAT	0.57***	0.88**	0.93	0.89
Lys	AAA	0.70***	0.89**	0.90*	0.86**
	AAG	1.42***	1.11**	1.11*	1.16**

NOTE.—Columns list the odds ratios for preferential synonymous codon usage in the 10% genes with the highest and lowest expression levels, based on AT40 (Schmid et al. 2005) and OS5 (Jain et al. 2007 for *Arabidopsis* and rice).

***$P < 0.001$; **$P < 0.01$; *$P < 0.05$ (after Benjamini–Hochberg correction for multiple testing [Benjamini and Hochberg 1996]).

additional selective forces that may obscure the signatures of selection for translational accuracy.

We next analyzed whether the strength of selection could vary depending on the amino acid. In both, *Arabidopsis* and rice, we found evidence of amino acids apparently not affected by selection for translational accuracy. Association analysis carried out separately for each codon indicated that preferential usage of nonpreferred codons at evolutionarily constrained sites was the main factor accounting for such observations. These data suggest that fidelity and speed of translation may not be coincident in all cases. It is important to note that we defined as optimal codons those codons that are significantly more used in highly expressed than in lowly expressed genes. Such strategy is expected to identify codons that are translated with an high speed (Zhou et al. 2009). Lee et al. (2010) have suggested that this method could be inappropriate if a specific speed accuracy trade-off exists. For those cases, faster codons could not be the most accurate and vice versa codons that are identified as nonoptimal may be more used at constrained sites if they are translated with high accuracy. In *E. coli,* Lee et al. (2010) have proposed that the codon for Valine that is most rapidly translated is not the most accurate for Val in this species. An additional explanation could be related to proteome-specific features. Some optimal codon may be chosen for rapid rather than accurate translation, because, for example, slow folding regions are particularly susceptible to misfolding in case of ribosome stalls. To cite an example, aggregation-prone regions in *E. coli* are associated to slow folding rate (Lee et al. 2010). These regions would be prone to dysfunctional intermolecular interactions if not adequately protected by the folding process. In this case, selection could promote fastly translated rather than most accurated codons because the formers could better prevent ribosome stalls and consequent prolonged exposure of this region in a not protected state. Whether a relationship between speed and accuracy could be dependent on specific amino acids is not clear. All our attempts to find explanations based on biochemical issues such as polarity or volume did not provide convincing results. Another, as yet untested possibility is that the modification of certain tRNAs in plants can alter their accuracy but not speed or vice versa.

Drummond and Wilke (2008) analyzed the selection of translational accuracy in several unicellular and multicellular organisms. In most cases, the odd scores were higher than those that we observed in plants. This may reflect the low rate of adaptive evolution for some species (Bustamante et al. 2002; Slotte et al. 2010) or the low effective population size of plant species and the small impact of these effects on the fitness of individuals, thus reducing the magnitude of the responses detected in plant genes. The evolutionary distance and evolutionary context of the species we investigated could also affect the outcome. Stoletzki and Eyre-Walker (2007) discussed the importance of species choice, suggesting that when there is a high rate of adaptive amino acid substitutions, and most of the amino acid substitutions are not due to random genetic drift, then conserved residues may not provide a good indication of whether a site is evolutionarily constrained or not. Although we did not perform direct tests for such an effect, we did consider species that had diverged to different degrees and that had different evolutionary histories, and these generated similar results. Other confounding factors may include the definition of codon preferentiality, the approach and material used to measure gene expression, and the size of the gene data set. The preferred codons we used

were defined as those showing significantly more use in strongly expressed genes compared with weekly expressed ones and were determined based on genome-wide surveys of gene expression (Akashi 1994; Drummond and Wilke 2008; Zhou et al. 2009). Codon preferentiality based on independent analysis of different gene expression atlases produced similar results, as did the association tests carried out with each codon list.

Finally, we cannot rule out that the apparent lower association between preferred codons and evolutionarily constrained sites in plant proteins could subtend to a lower selection for translational accuracy of plants. It has recently been shown that mistranslation increases under certain stress conditions in mammals, leading to the increased misincorporation of methionine residues (Netzer et al. 2009). Proteome-wide mistranslation has been shown to increase the fitness of certain organisms under particular environmental conditions (Moura et al. 2009). We could therefore speculate that lower translational fidelity in plants may be advantageous to generate flexible proteomes, which may help them to adapt under adverse environmental conditions. The dedicated compositional analysis of stress-regulated genes will provide insight into this hypothesis.

Supplementary Material

Supplementary tables S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Prof. D.A. Drummond and Prof. Claus Wilke for helpful suggestions and Prof. Mario Pezzotti for critical reading of the manuscript. The project was funded by the Università di Sassari.

Literature Cited

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136:927–935.
Akashi H. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev. 8:688–693.
Akashi H. 2003. Translational selection and yeast proteome evolution. Genetics 164:1291–1303.
Arava Y. 2003. Isolation of polysomal RNA for microarray analysis. Methods Mol Biol. 224:79–87.
Benjamini Y, Hochberg Y. 1996. Controlling for false discovery rate: practical and powerful approach to multiple testing. J R Statist Soc. 57(1):289–300.
Bucciantini M, et al. 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature 416: 507–511.
Bustamante CD, et al. 2002. The cost of inbreeding in *Arabidopsis*. Nature 416(4):531–534.
Carels N, Bernardi G. 2000. Two classes of genes in plants. Genetics 154: 1819–1825.
De Castro E, et al. 2006. ScanProsite: detection of PROSITE signatures matches and ProRule associated functional and structural residues in proteins. Nucleic Acid Res. 34(Web server issue):W362–W365.
DeRisi JL, Iyer VR, Brown PO. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680–686.
dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 32: 5036–5044.
Drummond DA, Wilke CO. 2008. Mistranslation –induced protein misfolding as a dominant constraint in coding sequence evolution. Cell 134: 341–352.
Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet. 10:715–724.
Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity 5:113.
Goodstein DM, et al. 2012. Phytozome a comparative platform for green plant genomics. Nucleic Acid Res. 40:D1178–D1186.
Gray NK, Hentze MW. 1994. Regulation of protein synthesis by mRNA structure. Mol Biol Rep. 19:195–200.
Guo X, Bao J, Fan L. 2007. Evidence of selectively driven codon usage in rice: implications for GC content evolution of Graminae genes. FEBS Lett. 58:1015–1021.
Ingolia NT, Ghaemmaghami S, Newman JRS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324:218–223.
Jain M, et al. 2007. F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. Plant Physiol. 143(4):1467–1483.
Kramer EB, Farabaugh PJ. 2007. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. RNA 13: 87–96.
Lee Y, Zhou T, Tartaglia G, Vendruscolo M, Wilke C. 2010. Translationally optimal codons associate with aggregation-prone sites in proteins. Proteomics 10:4163–4171.
Lipman DJ, Wilbur WG. 1985. Interaction of silent and replacement changes in eukaryotic coding sequences. J Mol Evol. 21:166–167.
Marais G, Duret L. 2001. Synonymous codon usage, accuracy of translation and gene length in *Caenorhabditis elegans*. J Mol Evol. 52(3): 275–280.
Mirny LA, Shakhnovich E. 1999. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol. 291:177–196.
Moura GR, Carreto LC, Santos M. 2009. Genetic code ambiguity an unexpected source of proteome innovation and phenotypic diversity. Curr Opin Microbiol. 12:631–637.
Mukhopadhyay P, Basak S, Ghosh GTC. 2008. Differential selective constraints shaping codon usage pattern of housekeeping and tissue-specific homologous genes of rice and *Arabidopsis*. DNA Res. 15:347–356.
Netzer N, et al. 2009. Innate immune and chemically triggered oxidative stress modifies translational fidelity. Nature 462:522–526.
Niimura Y, Terabe M, Gojobori T, Miura K. 2003. Comparative analyses at the gene terminal portions in seven eukaryotes genomes. Nucleic Acids Res. 31:5195–5201.
Ostlund G, et al. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38:D196–D203.
Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.
Parker J. 1989. Errors and alternatives in reading the universal genetic code. Microbiol Rev. 53:273–298.
Percudani R. 1999. Selection at the wobble position of codons read by the same tRNA in *Saccharomyces cerevisiae*. Mol Biol Evol. 16: 1752–1762.

Porceddu A, Camiolo S. 2011. Spatial analyses of mono, di and trinucleotide trends in plant genes. PLoS One 6:e22855.
Qin H, Wu WB, Comeron JM, Kreitman M, Li WH. 2004. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics 168:2245–2260.
Schmid M, et al. 2005. A gene expression map of *Arabidopsis thaliana* development. Nat Genet. 37:501–506.
Schueler-Furman O, Baker D. 2003. Conserved residue clustering and protein structure prediction. Proteins 52:225–235.
Servant F, et al. 2002. ProDom: automated clustering of homologous domains. Brief Bioinform. 3:246–251.
Sharp PM. 1987. The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.
Slotte T, Foxe JPM, Hazzouri JM, Wright S. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Mol Biol Evol. 27(8):1813–1821.
Stoletzki N, Eyre Walker A. 2007. Synonymous codon usage in *Escherichia coli*. Selection for translational accuracy. Mol Biol Evol. 24(2):374–381.
Tuller T, et al. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141:344–354.
Tuller T, et al. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol. 12:R110.
Wong GKS, et al. 2002. Compositional gradients in graminae genes. Genome Res. 12:851–856.
Wright SI, Yau CBK, Looseley M, Meyers B. 2004. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Mol Biol Evol. 21:1719–1726.
Zhou T, Mason W, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol. 26:1571–1580.

Associate editor: Brandon Gaut