OXFORD
UNIVERSITY PRESS
DNA Research

New insights into the interplay between codon bias determinants in plants

Journal:	*DNA Research*
Manuscript ID:	DNAR-2015-092
Manuscript Type:	Full Paper
Date Submitted by the Author:	30-Jun-2015
Complete List of Authors:	Camiolo, Salvatore; Università di Sassari, Dipartimento di Agraria Melito, Sara; Università di Sassari, Dipartimento di Agraria Porceddu, Andrea; University of Sassari, AGRARIA
Keywords:	Codon bias, Plant genomics, translational selection, Mutational bias

New insights into the interplay between codon bias determinants in plants

S. Camiolo, S. Melito, A. Porceddu*

Università degli Studi di Sassari, Dipartimento di Agraria, SACEG, Sassari

Corresponding author:

A. Porceddu

e-mail: aporceddu@uniss.it

Running title: Codon bias in plants

Supplemental material is available online

Abstract

Codon bias is the non-random use of synonymous codons, a phenomenon has been observed in species as diverse as bacteria, plants and mammals. The preferential use of particular synonymous codons may reflect neutral mechanisms (e.g. mutational bias, G|C-based gene conversion, genetic drift) and/or selection for mRNA stability, translational efficiency and accuracy. The extent to which these different factors influence codon usage is unknown, so we dissected the contribution of mutational bias and selection towards codon bias in genes from 17 eudicots and 4 monocots. We analysed the frequency of mononucleotides, dinucleotides and trinucleotides, and investigated whether the compositional genomic background could account for the observed codon usage profiles. Neutral forces such as mutational pressure and G|C-based gene conversion appeared to underlie most of the observed codon bias, although there was also evidence for the selection of optimal translational efficiency and mRNA folding. Our data confirmed the compositional differences between monocots and dicots, with the former featuring in general a lower background compositional bias but a higher overall codon bias.

Keywords: Codon bias, mutational bias, translational selection, plant genetics.

Introduction

The genetic code is redundant, with most amino acids encoded by two or more synonymous codons [1–4]. The non-random use of synonymous codons is known as codon bias, and it may reflect several underlying factors including mutational bias in the genome and translational selection. The possibility that mutational bias affects codon usage has led to the neutralist model, in which codon identity is mainly determined by nucleotide substitution patterns in the genome. In contrast, the possibility of translational selection has led to the selective model, in which the choice of synonymous codons reflects tRNA abundance [5] to optimize the efficiency [4] and accuracy [6] of translation. These models are not mutually exclusive, i.e. the choice among synonymous codons may reflect a balance between selective and mutational pressures [7].

Species-dependent differences in codon usage are well known [8], but recent studies have identified variations within species that must also be addressed by the neutralist and selective models. For example, the abundance of tRNA may vary during development and in response to external stimuli [9], suggesting that codon bias may represent an adaptive response to tRNA levels that differ among plant tissues [10]. Nucleotide substitution patterns are also unequally distributed in the genome [11], e.g. there are large homogeneous blocks G|C-rich sequences known as isochores in the genomes of warm-blooded vertebrates [12,13]. Likewise, differences in nucleotide substitution patterns have been observed in rice (*Oryza sativa*), with genes expressed in the roots being predominantly G|C-rich and genes expressed in seeds and leaves being predominantly A|T-rich [14].

The compositional context can also influence synonymous codon selection, a phenomenon known as context-dependent codon bias (CCDB). In mammals, bacteria and plants, the first nucleotide after each codon drives synonymous codon choice, because several dinucleotide sequences such as CG, GA and TA are underrepresented [15–17]. In plant genomes, there is also a general bias in the use of specific dinucleotides and trinucleotides in different genomic regions [18].

The composition of coding sequences is determined by a complex series of interacting factors, so it is difficult to identify the relative impact of different components. A model to determine the influence of nucleotide substitution patterns on codon bias has been developed by building a new set of sequences in which the third codon position in the coding sequence is replaced with a random nucleotide from the neighbouring intergenic region [19]. Such intergenic corrected coding sequences (ICCSs) retain the same amino acid sequence while mirroring the background nucleotide substitution pattern of the genome. Comparing the codon bias between the original coding sequence (CS) and ICCS datasets can therefore highlight the influence of the background nucleotide substitution pattern on the coding sequence composition [19].

Although the selective model has been studied in several plant species, the impact of background composition on gene structure has been largely overlooked. Codon bias in *Arabidopsis thaliana* (Arabidopsis) tends to be associated with the composition of the 3' flanking region in both strongly and weakly expressed genes [20], although the impact of selection on both classes of genes has also been recognized. Here we used the Hershberg and Petrov approach [19] in order to determine the effect of background composition on the codon bias of 21 plant species while also accounting for bias in the frequencies of dinucleotides and trinucleotides. We discuss in detail the impact of these multiple factors and others influencing codon bias in plants.

Materials and methods

Sequence data

The genomic sequences and annotation data for 21 plant species were downloaded from Phytozome (http://www.phytozome.net)[21] . We analysed the sequences of 4 monocots (*Brachypodium distachyon* (BD), *Oryza sativa* (OS), *Sorgum bicolour* (SB) and *Zea mays* (ZM)), 15 dicots (*Arabidopsis lyrata* (AL), *A. thaliana* (AT), *Brassica rapa* (BR), *Citrus clementine* (CC), *C. sinensis* (CS), *Eucalyptus grandis* (EG), *Phaseolus vulgaris* (FV), *Glycine max* (GM), *Linum* *usitatissimum* (LU)*, Medicago truncatula* (MT)*, Populus trichocarpa* (PopT)*, Solanum lycopersicum* (SL)*, Solanum tuberosum* (ST)*, Theluginella halophile* (TH) and *Vitis vinifer*a (VV)) and two mosses (*Selaginella moellendorffii* (SM)*, Physcomitrella patens* (PP)) .

We used gff2sequence [22] to identify coding sequences, proteins, introns and intergenic sequences. Coding sequences featuring non-canonical bases (other than A, C, G or T), missing stop codons or incomplete triplets were excluded. Finally the longest splicing variant was chosen when multiple transcripts representing the same gene were annotated. Monocot coding sequences were divided into three subsets for analysis: (1) the entire genome, (2) high-G|C sequences (GC content > 60%, HGC) and (3) low-G|C sequences (GC content ≤ 60%, LGC).

Expression data

Gene expression data for Arabidopsis and rice were downloaded from the Plexdb database (http://www.plexdb.org/)[23]. We chose the expression atlases representing Arabidopsis dataset AT40 (http://www.plexdb.org/modules/PD_browse/experiment_browser.php?experiment=AT40) and rice dataset OS5 (http://www.plexdb.org/modules/PD_browse/experiment_browser.php?experiment=OS5). All the expression data were RMA normalized. An expression value was calculated for each gene by averaging the replicates within each experiment and then computing a mean value over all the experiments in which the corresponding gene was expressed [24].

Intergenic controlled coding sequences.

For each gene, the first 100 two-fold and four-fold degenerate codons were used to create the coding sequence for analysis. Transcripts with fewer degenerate codons were excluded. The ICCSs were generated to estimate the influence of the background composition on coding sequence codon bias. Upstream and downstream sequences were extracted from the leading strand and joined together to form a set of concatenated intergenic sequences (CISs) and those shorter than 50 nucleotides were excluded at this stage.

Four different background controls were used to generate ICCS datasets, beginning with the mononucleotide composition as originally used in the Hershberg and Petrov method [19]. Briefly, a subsequence of 100 consecutive base pairs was randomly selected from the CIS, and the third base of each codon in the coding sequence was replaced with a nucleotide from this CIS subsequence (Figure 1a). This yielded a new dataset called monoICCS. The intergenic dinucleotide composition was used to generate a second class of ICCS (dinuICCS) by choosing a random subsequence of 200 consecutive base pairs from the CIS and picking the second base of each coding sequence codon randomly from within that subsequence. The adjacent base was then selected as the third codon positon in the ICCS (Figure 1b). If the CIS was shorter than 100 bp it was excluded from the monoICCS and if it was shorter than 200 bp it was excluded from the dinuICCS. Genes were also excluded from further analysis if any base in the CIS occurred fewer than four times.

Intergenic trinucleotide controlled coding sequences (trinuICCS) were produced by randomly picking the first dinucleotide for each coding sequence codon from within the CIS and selecting the adjacent nucleotide as the third base for the ICCS (Figure 1c). Finally, CDCB was estimated by randomly selecting an interrupted dinucleotide comprising the second base of each coding sequence codon and the first base of the subsequent triplet. The intervening nucleotide was then selected as the third codon base in the context-dependent codon bias intergenic controlled coding sequence (cdcbICCS). Genes in the trinuICCS and cdICCS datasets were excluded from further analysis if the corresponding dinucleotide appeared fewer than four times in the CIS.

In order to maintain the coding sequence amino acid structure, two-fold degenerate codons ending in A|G were used to create an ICCS ending in A when the corresponding intergenic position was A or T, and otherwise the ICCS ended in G. Similarly, two-fold degenerate codons ending in T|C were used to create an ICCS featuring T at the third codon position when the corresponding CIS nucleotide was A or T, and otherwise the ICCS ended in C.

Codon bias measurements

The effective number of codons (Nc)[25] was used to estimate the overall codon bias for each gene in the coding sequence and ICCS datasets. The Nc index generated higher values for less-biased genes (theoretical values between 21 and 64). To explore the contribution of each individual codon, relative synonymous codon usage (RSCU) values were computed for informative two-fold and four-fold degenerate codons (i.e. excluding methionine, tryptophan and stop codons). RSCU values were calculated as the ratio of the observed and expected codon frequencies, i.e. the random use of all codons within a specific degenerate family[26] .

Signature of selection

For the 44 codons with two-fold or four-fold degeneracy, differences between the average RSCU values in the CS and ICCS datasets were calculated to highlight overrepresentation and underrepresentation in the coding sequences. For each codon *cod*, the deviation from background was calculated as follows:

$$\Delta RSCU^{cod} = \frac{RSCU_{CS}^{cod} - RSCU_{ICCS}^{cod}}{Deg^{cod}}$$

where Deg^{cod} is the degeneracy of the codon. The significance of highlighted differences was tested using a paired Wilcoxon test. The same statistical analysis was applied to highlight differences in the effective number of codons between the CS and ICCS datasets.

The association between the RSCU values of the CS and ICCS datasets and the gene expression levels was investigated in Arabidopsis and rice by sorting genes on the basis of their expression levels into 20 bins containing the same number of genes. The bin rank was then plotted against the average CS|ICCS RSCU value within the bin using either the intergenic or intron portion for the construction of the ICCS dataset. In the latter case, all introns within the same gene were concatenated. In rice, this analysis was also carried out separately on the HGC and LGC gene sets.

Results

Codon bias in the intergenic corrected coding sequences

Mutational bias is one of the main forces affecting synonymous codon choice in bacteria [19], plants [27] and humans [28]. In theory, if no additional forces act on the coding sequences, the third base of each codon should reflect the background nucleotide frequency in the genome. However, the direction and strength of the codon bias should be investigated in the local genomic context to correct for compositional pattern heterogeneity. Previous studies have revealed the non-random distribution of the four nucleotides in several eukaryotic genomes [29] including the presence of compositionally homogenous isochores in mammals and birds [12,13]. Plant genomes also comprise a mosaic of compositionally homogenous segments although the overall compositional heterogeneity is much less extreme than the sequences found in mammals [12]. The non-random usage of specific dinucleotides [30] and trinucleotides [18] should also be considered during the analysis of local mutational bias to determine the impact on the overall compositional pattern.

We used the composition of intergenic regions as a proxy for background bias according to the Hershberg and Petrov method [19]. The portion of codon bias caused by background composition was measured by generating several ICCS datasets taking into account the mononucleotide, dinucleotide, trinucleotide and context-dependent intergenic composition of each plant species. This allowed us to compare the actual structure of the coding sequences with the hypothetical structure based on compositional features of the flanking intergenic regions. We found that codon bias in the coding sequence, and to a lesser extent in the ICCSs, fluctuated along the chromosomes of several species (Figure S1) highlighting the presence of position-specific compositional patterns. We calculated spatial autocorrelations between the chromosome location and the RSCU values in Arabidopsis and rice to determine whether codon bias was conserved among clustered genes. Significant spatial autocorrelation was observed for several codons in the CS datasets but for only a few codons in the ICCS datasets, although the trinuICCS dataset was an exception (Figure S2). This result mirrors the more uniform composition of isochores in plants [29] and contrasts with equivalent results generated by analysing human genes, where the isochore structure is more heterogeneous (Figure S2).

Next we investigated the direction of mutational bias by analysing the codon bias among all the ICCS datasets. If the four nucleotides are distributed randomly in the background, then there should be no codon bias in any of the gene sequences. However, we observed the opposite trend, i.e. Nc values compatible with a multilevel background compositional bias. Interestingly, the lowest Nc values were found in the dinuICCS dataset, revealing the strongest bias for all species (Figure 2). Whereas the legumes (*P. vulgaris*, *M. truncatula* and *G. max*) showed the highest ICCS codon bias at all levels, the opposite trend was observed for the monocots (rice, *Z. mays*, *S. bicolor* and *B. distachyon*).

Nc values provide a snapshot of overall codon bias within genes but do not reveal the specific contributions of each codon. ICCS RSCU values were therefore calculated for all codons with two-fold or four-fold degeneracy (Figure 3). The frequency of specific mononucleotides, dinucleotides and trinucleotides in the intergenic regions was associated, by construction, with the RSCU values of the codons containing them (e.g. overrepresentation of the dinucleotide CA would be associated with the higher RSCU values for codon GCA in the dinuICCS dataset). We observed a significant overrepresentation of codons ending in A|T in the monoICCS dataset, particularly in *V. vinifera*, *Solanum* spp. and the legumes. This highlighted the high degree of A|T-enrichment within the intergenic regions in all the eudicot species we analysed, and the lower degree of enrichment in the monocots and *S. moellendorffii*.

A similar picture emerged from the analysis of the dinuICCS and trinuICCS RSCU values although several compositional signatures also emerged. The codons ending in A|T generally showed higher RSCU values at the expense of those ending in G|C, but the extent of the bias was variable. Indeed, codons ending in CG, AG and AC tended to be suppressed more strongly in the dinuICCS dataset, e.g. these dinucleotides were underrepresented in the intergenic regions of all the plants. But codons ending in CA showed the opposite trend. As previously observed for the monoICCS dataset, the codon bias in the monocot dinuICCS and trinuICCS datasets was weaker than the corresponding eudicot datasets. Only marginal differences were observed between the dinuICCS and trinuICCS datasets, indicating that the intergenic trinucleotide bias is predominantly caused by bias in the frequency of the underlying dinucleotides.

Codon bias in the coding sequences

If synonymous codon choice solely reflects the distribution of nucleotides in the genomic background, there should be no differences in Nc value between the CS and ICCS datasets. However, our results revealed significant differences (Wilkoxon paired sample test) between the datasets, although the direction and extent of divergence differed among the species we investigated and divergence was particularly evident between eudicots and monocots. Although the coding sequences were more biased than the monoICCS dataset in monocots, only small differences were observed in dicots (Figure 2). Interestingly, the coding sequences were even less biased than the corresponding ICCSs in some species, with the trend most noticeable when comparing the CS and dinuICCS datasets. These data suggest that additional forces shape the coding sequences and oppose the mutational bias, e.g. resulting in G|C enrichment despite background A|T enrichment.

Differences in RSCU values between the CS and ICCS datasets allowed us to focus on codons whose frequency cannot be explained by background bias alone. Comparisons between the CS dataset and the four ICCS datasets constructed using alternative approaches led to similar results (Figures 4 and S3). All codons ending in A|T were less frequent in the coding sequence, with GTA suffering the most suppression. However, several codons ending in G|C were overrepresented in the coding sequence, with certain species-dependent exceptions. Codons CCC and GGG were suppressed in many species, together with codons CCG, GCG and ACG. The underrepresentation of GTA was more striking when comparing the CS and monoICCS datasets and less pronounced when comparing the CS and dinuICCS datasets, suggesting that bias against GTA in the coding sequences in part reflects the general suppression of the TA dinucleotide in the intergenic regions (Figure 3). Similarly, the underrepresentation of CCG, GCG and ACG observed when comparing the CS and monoICCS datasets was not so apparent when comparing the CS and dinuICCS datasets, suggesting it reflects the general suppression of the CG dinucleotide in the intergenic regions (Figure 3).

Some differences between the CS and the ICCS datasets were also taxon-dependent, e.g. the preference for G|C at the third codon position was more apparent in monocots than eudicots, particularly *M. truncatula, V. vinifera* and *Solanum* spp., where there was little evidence for preference.

Correlation between codon bias and gene expression in Arabidopsis and rice

Codon bias in the species we investigated was not fully explained by background compositional differences so we investigated the influence of gene expression on synonymous codon usage in the CS and ICCS datasets. Codons that are translated more rapidly or accurately should be preferentially found in the coding sequences of genes, particularly those expressed at high levels. We chose Arabidopsis and rice as representative dicot and monocot species and assigned genes to

20 bins based on expression levels, and then mapped the RSCU values of the CS and the ICCS datasets onto these bins (Figure 5). This strategy should not only reveal selection but also the impact of selection on the coding and noncoding regions. For example, a codon whose frequency is positively associated with expression level may be under selection to optimize translation efficiency and/or accuracy. However, the same positive association should also be observed in the ICCS datasets, e.g. genes with comparable expression levels should share the sequence composition of the intergenic sequences. Both scenarios rely on events that are not mutually exclusive, but this method can highlight the participation of additional forces in shaping the compositional pattern of the coding sequences. Indeed if RSCU values in the CS and ICCS databases differ regardless of the expression level, then forces other than the translational selection are likely to be responsible.

We found that two-fold degenerate codons ending in G|C are more frequent in the coding sequences of rice and Arabidopsis, although in the latter case the aspartic acid codon GAC was an exception. Although previous studies have shown that optimal codons tend to end in G|C [31,32] this is not solely dependent on translational selection because a positive association between the RSCU value and expression was evident only at extreme expression levels. There was no association between expression level and the RSCU values of the ICCSs, suggesting that expression-dependent variation in codon bias does not reflect differences in the background composition of the Arabidopsis genome.

A more complex picture emerged from the analysis of rice genes. The correlation between RSCU values and gene expression was non-monotonic for codons with two-fold degeneracy, although as in Arabidopsis the codons ending in G|C were still used on average more frequently in the coding sequences. However an initial reduction in the frequency of such codons was complemented by an increase in RSCU values at higher expression levels. If we exclude the effect of the background composition, which again does not vary with the expression level, such non-monotonic trends may reflect the coexistence of contrasting forces whose strength could be dependent on the expression level. Alternatively, the different expression bins in rice may be populated with genes that do not experience the same selective pressure. Indeed, it is well known that monocots have two classes of genes that differ in GC content. For this reason, the above analysis was repeated focusing specifically on LGC and HGC genes. The LGC genes accounted for most of the dataset and mirrored the overall trends discussed above, but the HGC genes showed a general positive association between the expression level and the RSCU values of two-fold degenerate codons ending in G|C.

Weaker trends were observed in Arabidopsis when the four-fold degenerate codons were analysed both in terms of deviation from the background and variation with expression. Indeed, the frequency of many codons mirrored the background composition of the entire set of bins, so that the curves generated by the CS and ICCS datasets could be superimposed. However, there was a positive association between the RSCU values and expression levels of codons ending in C and the opposite trend was observed for codons ending in A. Interestingly, the frequency of several codons ending in T was positively associated with the expression level, particularly the alanine codon CGT.

In rice, four-fold degenerate codons ending in G|C were always used more frequently in the coding sequence than the ICCS whereas the opposite trend was observed for codons ending in A|T. Furthermore, codons ending in G|C were generally used less frequently in strongly expressed genes, whereas there was a positive association between codons ending in T and expression level. As previously reported in monocots, such trends are often non-monotonic, either disappearing or changing direction when HGC and LGC genes are analysed separately (Figure S4).

The positive association between the RSCU value and expression level of codons ending in T is reminiscent of transcription-associated mutational bias (TAMB), a well-known repair system that increases bias towards G|T rather than C|A in the pre-mRNA sequence [33]. For this reason, we repeated our analysis by constructing a new set of ICCSs using intron sequences as a proxy for the background composition. If codons ending in T become more frequent in strongly expressed genes due to TAMB, the same trend should be observed in intron-corrected coding sequences. Our results revealed no such association (Figure S6), indicating that TAMB is not responsible for the positive association between gene expression level and the frequency of codons ending in T.

Discussion

Background composition

Several ICCS datasets were generated to investigate whether codon bias in the coding sequence reflected the genomic background nucleotide composition of the plants included in this study. Four ICCS datasets constructed using different strategies were analysed to determine the effective number of codons and relative synonymous codon usage. A strong dinucleotide compositional bias was evident in all the species, as shown by the higher levels of codon bias (i.e. lower Nc values) in the dinuICCS dataset compared to the others (Figure 2). The RSCU values of the monoICCS dataset revealed the overrepresentation of A|T compared to G|C although to a different extent in each species. The legumes and solanaceous species showed the greatest difference in the representation of A|T and G|C, whereas there was a weaker distinction in the monocot species and the brassicas showed intermediate values. G|C to A|T mutations may be more frequent in Arabidopsis, leading to AT enrichment in the genome. The deamination of methylated cytosine residues at CG dinucleotide motifs, and the UV-induced mutagenesis of dipyrimidines (CC and TC) may explain this phenomenon[34]. Our data support this model in all the plants included in this investigation. We found that codons ending in CG were among the most underrepresented in the dinuICCS dataset, whereas codons ending in CA (the reverse complement of the C deamination product in CG) were among the most overrepresented, as recently reported in Arabidopsis and rice [30]. Codons ending in CC and TC were also suppressed in the dinuICCS dataset, confirming the underrepresentation of these dinucleotides in the non-coding sequences. Several factors may account for the observed differences in the background nucleotide composition of plants. In monocots, the vertical leaf orientation, protective basal sheath, and concealed apical meristem make the interception of solar radiation less efficient [35] thus reducing the prevalence of AT-enrichment induced by UV light. Furthermore, A|T pairs contain seven nitrogen atoms compared to the eight present in G|C pairs, which may drive A|T-enrichment in non-cultivated plants [36]. However, the cluster analysis of ICCS RSCU values revealed a pattern that is consistent with the genomic composition of plants solely depending on their phylogenesis (Figure S5).

Differences between the CS and ICCS datasets

Paired Wilcoxon tests generally revealed significant differences in codon usage between the CS and ICCS datasets in terms of both the Nc and RSCU values. The observed differences were similar in magnitude when comparing the CS dataset with all four methods for the construction of ICCS datasets, despite the clear dinucleotide signature of the background composition. This validates the Hershberg and Petrov method for the identification of optimal codons even though it only considers the mononucleotide composition of the background sequences. Indeed, the optimal codon datasets calculated using this method could be almost precisely superimposed over datasets generated using the other three methods (Figure S7).

The CS|ICCS comparisons suggested there was enrichment for codons entirely composed of G and C (hereafter described as GC3 codons) particularly in monocots, although among codons ending in G|C the frequency of the complementary codons GGG and CCC was close to the genomic background. This suggests there may be selection against codons that promote the formation of complex mRNA tertiary structures, and the prevalence of this phenomenon in monocots with their higher overall G|C content may emphasize such an underlying mechanism.

Codons ACG, CCG and GCG were marginally overrepresented in the coding sequences of some species and underrepresented in others. This supports the observation that the CG dinucleotide is suppressed, in part due to deleterious methylation/deamination events. However, this is actually a genomic tendency and the effect is not seen when comparing the CS and dinuICCS datasets (Figure 4).

Finally, the general suppression of codon GTA was one of the most conserved features among the plant species we investigated. This is not surprising because the codon ends with dinucleotide TA, which is known to be suppressed in the coding sequences of several plant species [18,30] possibly to discourage insertion events that target the TA dinucleotide [37], to reduce the likelihood of mutations leading to stop codons, and to prevent attacks by TA-specific RNases [38]. The less striking divergence between the CS and dinuICCS datasets in terms of GTA preference suggests that TA suppression is also a general genomic signature (Figure 3).

Mutational bias in coding sequences

The background nucleotide pattern alone cannot explain the observed codon bias in the coding sequences, so additional forces must be involved (although the signature of mutational pressure may still be observed at sites that are more loosely constrained in some plant species). Indeed, modestly frequent amino acids (or underrepresented amino acids in the proteins encoded by strongly expressed genes) may be under weaker selection, and for this reason their codons may better tolerate changes driven by mutational bias. For example, histidine and cysteine are among the less abundant amino acids in the Arabidopsis proteome and there is a significant negative correlation between their frequency and the expression level of the corresponding genes (Table S1). The frequency of codons CAC (His) and TGC (Cys) in Arabidopsis is similar in the CS and ICCS datasets (Figure 5). The codons GAC (Asp), GGC (Gly) and GCC (Ala) also revealed biases that mirrored the background composition. Interestingly, these codons feature the generic sequence CNG which may represent the core of the primitive genetic code [39]. Such a trend may therefore reflect the more prolonged effect of mutational bias on the most ancient codons. In contrast, there appeared to be little mutational bias in monocots, where other factors increase the frequency of CG3 codons.

Evidence for biased gene conversion in plants

The increasing G|C content of plant genomes may have been driven by G|C-biased gene conversion (gBGC) during double-strand break repair followed by a recombination event. This phenomenon is accompanied by the correction of eventual mismatches between two paired DNA strands featuring high sequence similarity, with such a correction being biased toward the placement of either G or C. By definition, gBGC is associated with the recombination rate along the genome and has contributed to isochore formation in mammals and birds. For this reason, a positive correlation between the G|C content and the recombination rate provides evidence that supports gBGC, and such an association has been observed in grasses but not Arabidopsis. Nevertheless, the association between recombination rate and G|C content may be perturbed (or possibly lost) when the evolutionary history of a species features a large number of genomic rearrangements, as is the case for Arabidopsis and the eudicots in general [40]. Although gBGC should not be restricted to genes, and differences in codon bias between the CS and ICCS datasets would therefore not highlight gBGC events, mutations caused by gBGC are more likely to be fixed in the coding sequence because codons ending in G|C are known to mirror the most abundant tRNAs in most plant species. There is also evidence that recombination in plant genomes occurs mainly in genes [41].

Our data provided several lines of evidence supporting the occurrence of gBGC in plants. First, higher standardized ΔRSCU values were observed for codons with two-fold degeneracy ending in G|C compared to those with four-fold degeneracy, which indicates the occurrence of gBGC events featuring a more diluted effect on codons with four-fold degeneracy. Moreover, if gBGC rather than selection is considered to be the main source of GC3 enrichment, such an effect should be evident in all the transcripts regardless of the expression level. As shown in Figure 5, this was the case for the majority of the codons and expression bins in both Arabidopsis and rice, with the latter featuring wider divergence from the background in accordance with previous reports of gBGC in grasses [42].

Translational selection and mRNA stability

Genes were assigned to 20 expression bins whose average RSCU values were plotted for the CS dataset and all four ICCS datasets for Arabidopsis and rice, representing the eudicots and monocots respectively (Figure 5). ICCS RSCU values were not associated with the expression level suggesting that co-expression clustering within the genome of these two species cannot be detected using this analytical approach.

In Arabidopsis, we observed a positive correlation between expression level and the frequency of optimal codons [31,32] ending in G|C mainly for genes assigned to the last few expression bins, underlining the marginal effect of translational selection in this species. A more cumbersome scenario emerged in rice, where non-monotonic trends were observed for codons ending in G|C, i.e. a reduction in frequency in the first expression bins changing to an increase in frequency in the last few. This behaviour may be typical of G|C-rich monocot genomes and may indicate a compromise between the advantage of using optimal codons and the avoidance of tightly-packed mRNA tertiary structures. Indeed a significant negative correlation was found between the expression level and both the coding sequence length ($r = –0.05$, $p < 0.0001$) and the G|C content ($r = –0.15$, $p < 0.0001$) of rice genes. Taken together, these data suggest that weakly expressed genes are longer and have a lower content of CG3 codons than strongly expressed genes. Longer transcripts are more likely to form strongly-packed mRNA tertiary structures if they are enriched in optimal codons ending in G|C, so the accumulation of such codons may be counterselected in genes expressed at low or moderate levels. Such a trend would diminish in shorter genes expressed at high levels. This hypothesis was confirmed by analysing HGC and LGC rice genes separately. We found that short

HGC genes were generally characterized by a monotonic positive association between the expression level and the frequency of optimal codons ending in G|C.

Additional factors

Factors other than mutational bias, selection and gBGC that shape codon bias include TAMB, a well-known DNA repair system that influences the composition of primary transcripts (both exons and introns) by increasing the prevalence of G|T over C|A. Despite this general GC3 enrichment, we observed the enrichment of four-fold degenerate codons ending in T, suggesting that TAMB may affect codon bias in Arabidopsis and rice, so we repeated our analysis using introns rather than intergenic DNA as an index of background composition. As shown in Figure S7, we found no differences in the frequency of codons ending in T when we used introns rather than intergenic DNA, suggesting the overrepresentation of such codons is a general genomic trend that probably reflects non-localized effects such as the preference for nitrogen saving base pairs in non-cultivated species (the A|T and A|U base pairs contain only seven nitrogen atoms, compared to eight in the G|C base pair). In contrast, crop species are provided with ample nitrogen and phosphorous which removes such constraints. However, the convergence of these trends between the brassicas and the nitrogen-fixing legumes, and between the monocots and the ancestral species *S. moellendorffii* (Figure S5) suggests that the overrepresentation of codons ending in T may reflect the phylogenesis of the species rather than the efficiency of nitrogen utilization.

Figure 1: Method used for the construction of the mononucleotide ICCS (monoICCS), dinucleotide ICCS (dinuICCS) and trinucleotide ICCS (trinuICCS) datasets.

Figure 2: Heat map showing the average Nc values for the CS and ICCS datasets.

Figure 3: RSCU values calculated for the monoICCS, dinuICCS and trinuICCS datasets (only codons ending in G|C with two-fold degeneracy are shown).

Figure 4: Calculation of ΔRSCU values as the standardized differences between the CS RSCU and (a) the monoICCS and (b) the dinuICCS RSCUs. Black crosses in white squares show insignificant differences between the datasets.

Figure 5: RSCU values of (a) Arabidopsis and (b) rice genes after splitting the datasets into 20 expression bins. Intergenic sequences were used for the construction of the ICCS datasets. Colour codes: blue = CS, red = monoICCS, green = dinuICCS, yellow = trinuICCS, brown = cdcbICCS.

References

1. Chamary, J. V, Parmley, J. L., and Hurst, L. D. 2006, Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.*, **7**, 98–108.

2. Duret, L. 2002, Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.*, **12**, 640–9.

3. Plotkin, J. B., Dushoff, J., Desai, M. M., and Fraser, H. B. 2006, Codon usage and selection on proteins. *J. Mol. Evol.*, **63**, 635–53.

4. Plotkin, J. B., and Kudla, G. 2011, Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.*, **12**, 32–42.

5. Ikemura, T. 1981, Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J. Mol. Biol.*, **151**, 389–409.

6. Akashi, H. 1994, Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics*, **136**, 927–35.

7. Bulmer, M. 1991, The Selection-Mutation-Drift Theory of Synonymous Codon Usage. *Genetics*, **129**, 897–907.

8. Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H., and Wright, F. 1988, Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. *Nucleic Acids Res.*, **16**, 8207–11.

9. Najafabadi, H. S., Goodarzi, H., and Salavati, R. 2009, Universal function-specificity of codon usage. *Nucleic Acids Res.*, **37**, 7014–23.

10. Camiolo, S., Farina, L., and Porceddu, A. 2012, The relation of codon bias to tissue-specific gene expression in Arabidopsis thaliana. *Genetics*, **192**, 641–9.

11. Wong, G. K.-S., Wang, J., Tao, L., et al. 2002, Compositional gradients in Gramineae genes. *Genome Res.*, **12**, 851–6.

12. Bernardi, G., Olofsson, B., Filipski, J., et al. 1985, The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953–8.

13. Carels, N., and Bernardi, G. 2000, Two classes of genes in plants. *Genetics*, **154**, 1819–25.

14. Liu, Q. 2012, Mutational bias and translational selection shaping the codon usage pattern of tissue-specific genes in rice. Robinson-Rechavi, M., (ed.), . *PLoS One*, **7**, e48295.

15. Berg, O. G., and Silva, P. J. 1997, Codon bias in Escherichia coli: the influence of codon context on mutation and selection. *Nucleic Acids Res.*, **25**, 1397–404.

16. De Amicis, F., and Marchetti, S. 2000, Intercodon dinucleotides affect codon choice in plant genes. *Nucleic Acids Res.*, **28**, 3339–45.

17. Karlin, S., and Mrázek, J. 1996, What drives codon choices in human genes? *J. Mol. Biol.*, **262**, 459–72.

18. Porceddu, A., and Camiolo, S. 2011, Spatial analyses of mono, di and trinucleotide trends in plant genes. *PLoS One*, **6**, e22855.

19. Hershberg, R., and Petrov, D. A. 2009, General rules for optimal codon choice. *PLoS Genet.*, **5**, e1000556.

20. Morton, B. R., and Wright, S. I. 2007, Selective constraints on codon usage of nuclear genes from Arabidopsis thaliana. *Mol. Biol. Evol.*, **24**, 122–9.

21. Goodstein, D. M., Shu, S., Howson, R., et al. 2012, Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–86.

22. Camiolo, S., and Porceddu, A. 2013, gff2sequence, a new user friendly tool for the generation of genomic sequences. *BioData Min.*, **6**, 15.

23. Dash, S., Van Hemert, J., Hong, L., Wise, R. P., and Dickerson, J. A. 2012, PLEXdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Res.*, **40**, D1194–201.

24. Camiolo, S., Rau, D., and Porceddu, A. 2009, Mutational biases and selective forces shaping the structure of Arabidopsis genes. *PLoS One*, **4**, e6356.

25. Wright, F. 1990, The “effective number of codons” used in a gene. *Gene*, **87**, 23–9.

26. Sharp, P. M., and Li, W. H. 1986, An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.*, **24**, 28–38.

27. Palidwor, G. A., Perkins, T. J., and Xia, X. 2010, A general model of codon bias due to GC mutational bias. *PLoS One*, **5**, e13431.

28. Sémon, M., Lobry, J. R., and Duret, L. 2006, No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Mol. Biol. Evol.*, **23**, 523–9.

29. Nekrutenko, A., and Li, W. H. 2000, Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.*, **10**, 1986–95.

30. Karlin, S. 1998, Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.*, **1**, 598–610.

31. Wright, S. I., Yau, C. B. K., Looseley, M., and Meyers, B. C. 2004, Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. *Mol. Biol. Evol.*, **21**, 1719–26.

32. Porceddu, A., Zenoni, S., and Camiolo, S. 2013, The signatures of selection for translational accuracy in plant genes. *Genome Biol. Evol.*, **5**, 1117–26.

33. Green, P., Ewing, B., Miller, W., Thomas, P. J., and Green, E. D. 2003, Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.*, **33**, 514–7.

34. Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., et al. 2010, The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science*, **327**, 92–4.

35. Cline, M. G., and Salisbury, F. B. 1966, Effects of ultra-violet alone and simulated solar ultra-violet radiation on the leaves of higher plants. *Nature*, **211**, 484–6.

36. Günther, T., Lampei, C., and Schmid, K. J. 2013, Mutational bias and gene conversion affect the intraspecific nitrogen stoichiometry of the Arabidopsis thaliana transcriptome. *Mol. Biol. Evol.*, **30**, 561–8.

37. Mashkova, T. D., Oparina, N. Y., Lacroix, M. H., et al. 2001, Structural rearrangements and insertions of dispersed elements in pericentromeric alpha satellites occur preferably at kinkable DNA sites. *J. Mol. Biol.*, **305**, 33–48.

38. Beutler, E., Gelbart, T., Han, J. H., Koziol, J. A., and Beutler, B. 1989, Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proc. Natl. Acad. Sci. U. S. A.*, **86**, 192–6.

39. Ikehara, K., Omori, Y., Arai, R., and Hirose, A. 2002, A novel theory on the origin of the genetic code: a GNC-SNS hypothesis. *J. Mol. Evol.*, **54**, 530–8.

40. Salse, J., Abrouk, M., Bolot, S., et al. 2009, Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 14908–13.

41. Rafalski, A., and Morgante, M. 2004, Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends Genet.*, **20**, 103–11.

42. Serres-Giardi, L., Belkhir, K., David, J., and Glémin, S. 2012, Patterns and evolution of nucleotide landscapes in seed plants. *Plant Cell*, **24**, 1379–97.

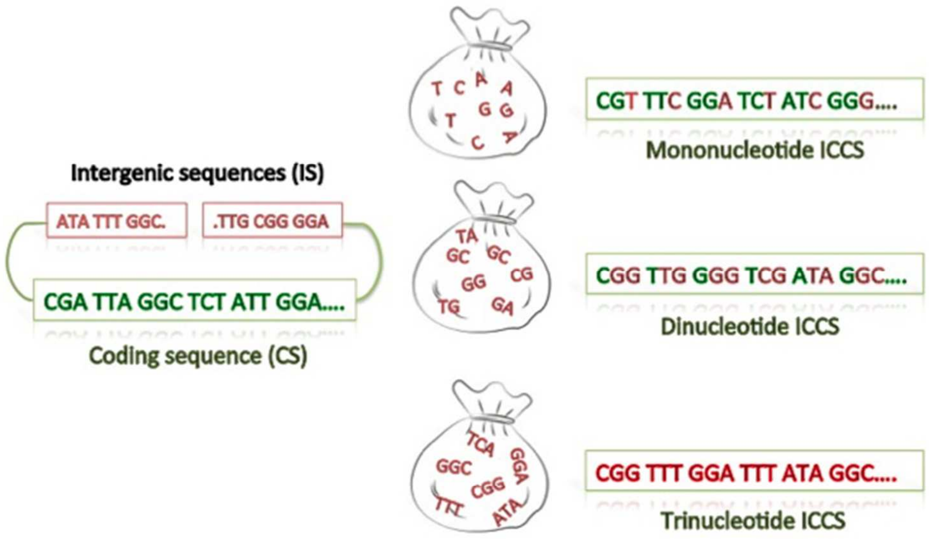

Figure 1: Method used for the construction of the mononucleotide ICCS (monoICCS), dinucleotide ICCS (dinuICCS) and trinucleotide ICCS (trinuICCS) datasets.
243x182mm (72 x 72 DPI)

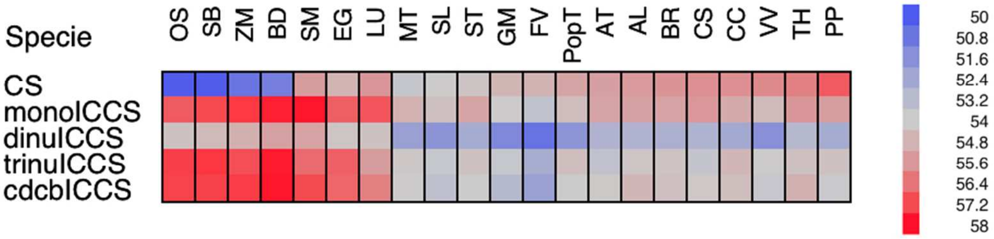

Figure 2: Heat map showing the average Nc values for the CS and ICCS datasets.
323x82mm (72 x 72 DPI)

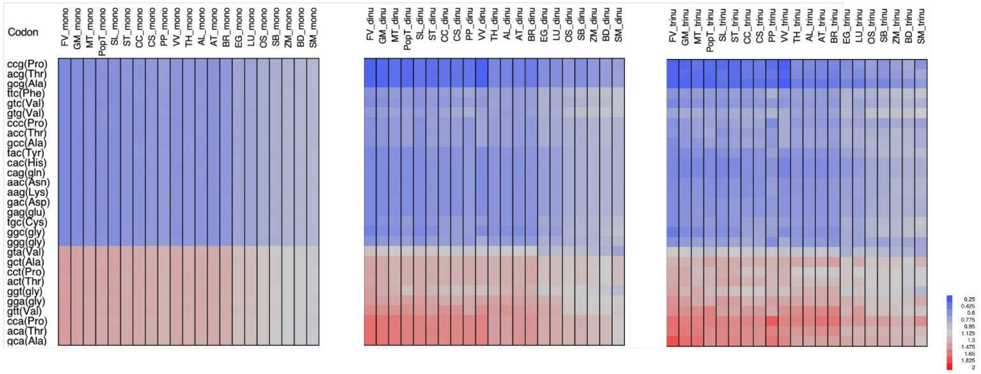

Figure 3: RSCU values calculated for the monoICCS, dinuICCS and trinuICCS datasets (only codons ending in G|C with two-fold degeneracy are shown).
799x310mm (150 x 150 DPI)

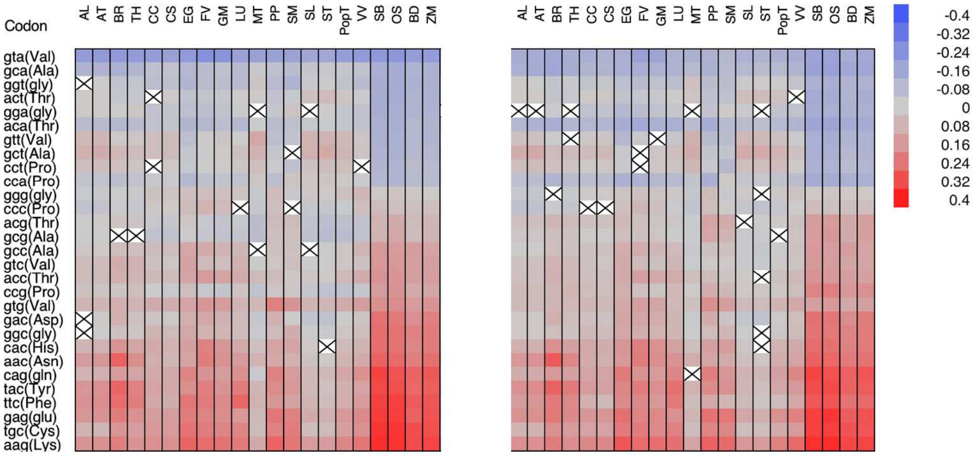

Figure 4: Calculation of ΔRSCU values as the standardized differences between the CS RSCU and (a) the monoICCS and (b) the dinuICCS RSCUs. Black crosses in white squares show insignificant differences between the datasets.
610x278mm (72 x 72 DPI)

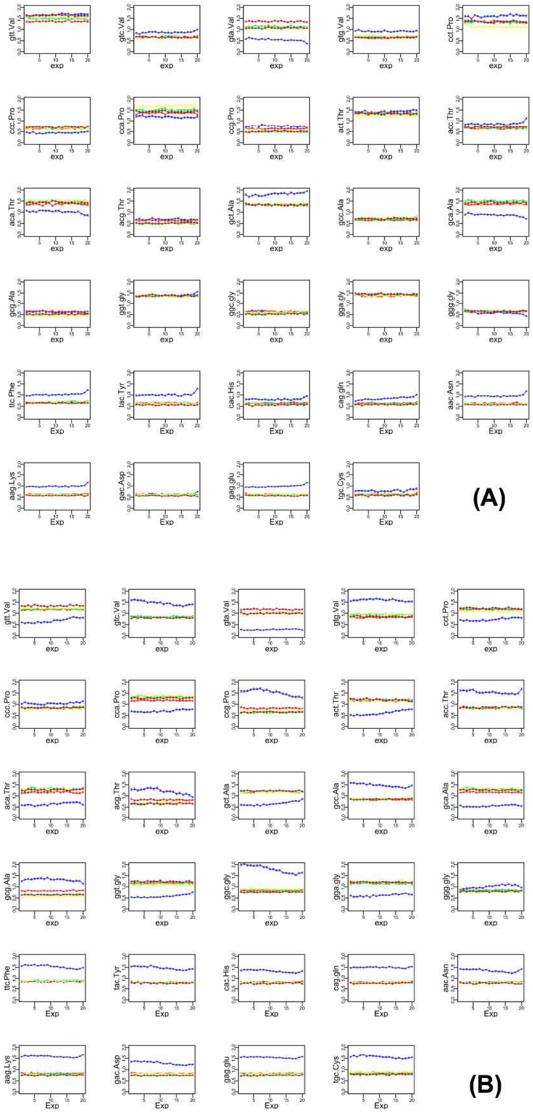

Figure 5: RSCU values of (a) Arabidopsis and (b) rice genes after splitting the datasets into 20 expression bins. Intergenic sequences were used for the construction of the ICCS datasets. Colour codes: blue = CS, red = monoICCS, green = dinuICCS, yellow = trinuICCS, brown = cdcbICCS.
199x399mm (300 x 300 DPI)