Annals of Botany **120**: 893–909, 2017
doi:10.1093/aob/mcx112, available online at www.academic.oup.com/aob

ANNALS OF BOTANY Founded 1887

RESEARCH IN CONTEXT

Codon usage and codon pair patterns in non-grass monocot genomes

Purabi Mazumdar[1], RofinaYasmin Binti Othman[1,2], Katharina Mebus[1], N. Ramakrishnan[3] and Jennifer Ann Harikrishna[1,2*]

[1]Centre for Research in Biotechnology for Agriculture, University of Malaya, 50603 Kuala Lumpur, Malaysia, [2]Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia and [3]Electrical and Computer System Engineering, School of Engineering, Monash University Malaysia, 47500 Bandar Sunway, Malaysia
[*]For correspondence. E-mail: jennihari@um.edu.my

Received: 17 January 2017 Returned for revision: 22 March 2017 Editorial decision: 9 August 2017 Accepted: 19 September 2017

- **Background and Aims** Studies on codon usage in monocots have focused on grasses, and observed patterns of this taxon were generalized to all monocot species. Here, non-grass monocot species were analysed to investigate the differences between grass and non-grass monocots.
- **Methods** First, studies of codon usage in monocots were reviewed. The current information was then extended regarding codon usage, as well as codon-pair context bias, using four completely sequenced non-grass monocot genomes (*Musa acuminata*, *Musa balbisiana*, *Phoenix dactylifera* and *Spirodela polyrhiza*) for which comparable transcriptome datasets are available. Measurements were taken regarding relative synonymous codon usage, effective number of codons, derived optimal codon and GC content and then the relationships investigated to infer the underlying evolutionary forces.
- **Key Results** The research identified optimal codons, rare codons and preferred codon-pair context in the non-grass monocot species studied. In contrast to the bimodal distribution of GC_3 (GC content in third codon position) in grasses, non-grass monocots showed a unimodal distribution. Disproportionate use of G and C (and of A and T) in two- and four-codon amino acids detected in the analysis rules out the mutational bias hypothesis as an explanation of genomic variation in GC content. There was found to be a positive relationship between CAI (codon adaptation index; predicts the level of expression of a gene) and GC_3. In addition, a strong correlation was observed between coding and genomic GC content and negative correlation of GC_3 with gene length, indicating a strong impact of GC-biased gene conversion (gBGC) in shaping codon usage and nucleotide composition in non-grass monocots.
- **Conclusion** Optimal codons in these non-grass monocots show a preference for G/C in the third codon position. These results support the concept that codon usage and nucleotide composition in non-grass monocots are mainly driven by gBGC.

Key words: Codon usage, codon-pair context, GC content, GC_3 distribution, non-grass monocots, banana, date palm.

INTRODUCTION

Codon usage bias (also known as codon bias) is the selective use of nucleotide triplets (codons) to encode specific amino acid sequences in the protein coding genes of a species. Every amino acid in a sequence can be encoded by one (in the case of methionine and tryptophan) to six different codons. The frequencies with which synonymous codons are used to encode an amino acid vary between organisms and sometimes even within the same organism (Hershberg and Petrov, 2008; Liu and Xue, 2005).

Three major hypotheses have been proposed to describe the codon usage bias in living systems: mutational bias (MB hypothesis), selection on codon usage (SCU) and GC-biased gene conversion (gBGC). To explain the codon usages that are produced by point mutations, contextual biases in the point mutation rates or biases in DNA repair, the MB hypothesis proposes that differences in mutation rates across species result in non-random variation of synonymous codon usage and an increase in GC content (Plotkin and Kudla, 2011). However, under the MB hypothesis, the increase in GC content in recombination hotspots via fixation of GC alleles at polymorphic sites cannot be explained as all classes of mutations have an equal probability of fixation (reviewed by Duret and Galtier, 2009). To explain the increase in probability of fixation of GC alleles, SCU was proposed as an alternative. The SCU hypothesis suggests natural selection as a contributor to the increase of GC content to improve translational efficiency (Akashi, 2003), arguing that synonymous mutations influence the fitness of an organism and can therefore be promoted or repressed through evolution. However, the presence of GC heterogeneity observed in pseudogenes and non-coding regions cannot be explained by SCU (Gossmann *et al.*, 2010). Furthermore, significant fitness advantages between individuals differing in only a single base pair are unlikely to be common. This, however, is required for selection to act (Mugal *et al.*, 2015).

After several years of debate around the MB and SCU hypotheses, a third hypothesis, gBGC, was put forward to explain codon usage and GC heterogeneity. The gBGC hypothesis suggests GC-biased gene conversion to shape codon usage bias in a manner dependent upon the local recombination rate, which favours fixation of G and C alleles over A and T at polymorphic sites. As a consequence of this association between GC and recombination, local GC content increases rapidly in genomic hotspots of recombination (Spencer, 2006) while genome-wide GC content increases rapidly in species with higher recombination rates (Figuet *et al.*, 2014; Weber *et al.*, 2014). Codon usage bias was recorded to be highest in areas of intermediate levels of recombination for GC-ending optimal codons (Harrison and Charlesworth, 2010). Even though the gBGC process is distinct from natural selection, it affects the probability of fixation of alleles in patterns similar to selection (reviewed by Duret and Galtier, 2009; Ratnakumar *et al.*, 2010) and interferes with selection by promoting the fixation of deleterious alleles (Galtier *et al.*, 2009; Necşulea *et al.*, 2011). Hence, gBGC is considered a neutral process as it does not rely on the fitness effect of alleles of the individuals. For the last 10 years a plethora of evidence has been accumulated for gBGC as a major evolutionary force affecting codon usage and base composition in humans (reviewed by Duret and Galtier, 2009, Glémin *et al.*, 2015), yeast (Lesecque *et al.*, 2013) and plants (Muyle *et al.*, 2011; Pessia *et al.*, 2012; Wu *et al.*, 2015; Rodgers-Melnick *et al.*, 2016; Clément *et al.*, 2017).

Apart from codon usage, codon context bias is also important for understanding the preferences in the sequence of a pair of codons within an organism. Codon-pair context bias is linked more to decoding accuracy than to translational speed (Moura *et al.*, 2005). As different species have varied abundance of tRNA isoacceptors for each codon family, codon-pair context sequences impact the translational accuracy of genes (Moura *et al.*, 2007).

Plant species are reported to show a wide diversity in terms of their gene expression, physiology and stress response in varied environmental conditions (De La Torre *et al.*, 2015). Hence, knowledge of the codon usage and codon-pair context patterns of plants and underlying evolutionary forces will be useful to understand the molecular mechanism of environmental adaptation and biological diversity of each species.

A number of comparative analyses have examined the codon usage of genes within and between the two major groups of flowering plants, dicots and monocots (Kawabe and Miyashita, 2003; Mukhopadhyay *et al.*, 2008; Liu, 2012; Liu *et al.*, 2015). Still, comparatively little is known about codon usage and codon-pair context in monocot plant species, which can be subclassified into grass monocots (Poaceae) and non-grass monocots (e.g. Orchidaceae, Musaceae, Arecaceae, Zingiberaceae and Liliaceae), as shown in Fig. 1. Furthermore, most studies on monocots focus on grass monocots and cannot explain the extent and pattern of codon usage or the forces that may be affecting codon usage in monocots in general.

This review outlines the methods available for the quantification of codon usage and presents the current state of knowledge of codon usage in monocots. Furthermore, we extend this information on codon usage, optimal and rare codons, GC content, GC_3 (GC content in third codon position) distribution, codon-pair context patterns and shaping factors in non-grass monocot genomes by analysing four completely sequenced genomes with publicly available comparable transcriptome datasets of two major banana species, *Musa acuminata* and *Musa balbisiana*, the date palm, *Phoenix dactylifera*, and the small aquatic plant, duckweed, *Spirodela polyrhiza.* The analysis suggests that gBGC affects the nucleotide composition and codon usage in non-grass monocots. This information on optimal and rare codons, factors affecting codon usage and codon-pair context bias will facilitate further research in molecular phylogenetics and genomics, as well as in the development of varieties tolerant of various biotic and abiotic stresses among the non-grass monocots.

Tools and indices for assessing codon usage

Investigation of codon usage in an organism requires quantification of codon usage patterns within the genes and genomes of the species under study. Codon usage values can be determined based on one or more indices (described in the following sections and summarized in Table 1) for different genes within and across the species.

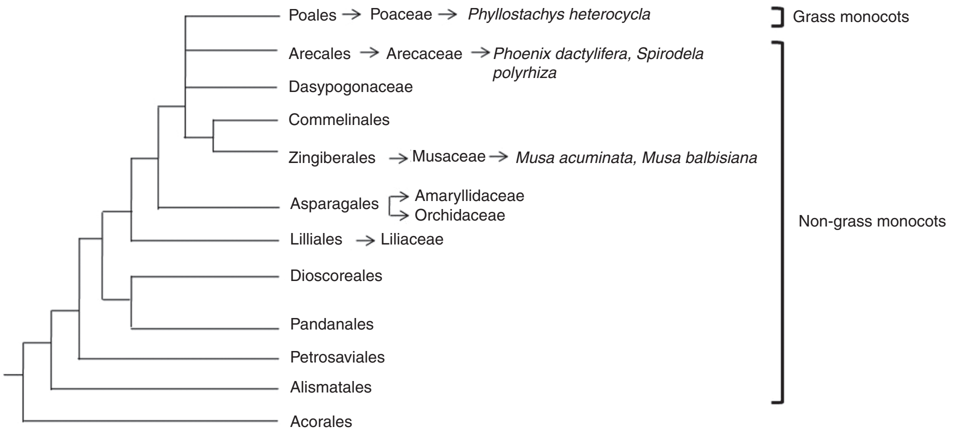
Poales → Poaceae → *Phyllostachys heterocycla*
Grass monocots
Arecales → Arecaceae → *Phoenix dactylifera, Spirodela polyrhiza*
Dasypogonaceae
Commelinales
Zingiberales → Musaceae → *Musa acuminata, Musa balbisiana*
Asparagales → Amaryllidaceae
Orchidaceae
Lilliales → Liliaceae
Non-grass monocots
Dioscoreales
Pandanales
Petrosaviales
Alismatales
Acorales

FIG. 1. The species used in the current analysis of codon usage in non-grass monocots (source of the tree; Hertweck *et al.*, 2015).

TABLE 1. *Indices for codon usage analysis*

Index	Principle	Mathematical formula	Scores	Reference
P2 index	Quantifies the proportion of codons that conform to the intermediate strength of the codon–anticodon interaction energy	$P2 = \frac{WWC + SSU}{WWY + SSY}$ where W = A or U, S = G or C, Y = C or U and A/C/G/U are the nucleotide composition of codon triplets	Under uniform codon usage P2 is equal to 0.5. A value more than 0.5 indicates strong codon bias and less than 0.5 indicates no codon bias	Gouy and Gautier (1982)
Relative synonymous codon usage (RSCU)	Calculated as the ratio of the observed frequency of a codon to the expected frequency of that codon, assuming uniform codon usage	$RSCU = \frac{g_{ij}}{ni} n_i$ $\sum_j g_{ij}$ where g_{ij} is the observed number of *i*th codon for the *j*th amino acid, which has n_i kinds of synonymous codons	The synonymous codons with RSCU values of 1 indicates no codon usage bias for that amino acid and the codons are chosen equally or randomly. RSCU values above 1 indicate positive codon usage bias and RSCU values below 1 indicate negative codon usage bias	Sharp *et al.* (1986)
Effective number of codons (ENC)	Measures how far the codon usage of a gene stays from the equal usage of synonymous codons	$ENC = 2 + \frac{9}{\bar{F}_2} + \frac{1}{\bar{F}_3} + \frac{5}{\bar{F}_4} + \frac{3}{\bar{F}_6}$ where $\bar{F}_k$ (*k* = 2, 3, 4, and 6) is the average of $\bar{F}_k$ values for *k*-fold degenerate amino acids and can be estimated by the formula: $F_k = \frac{mS - 1}{m - 1}$ where *m* is total number of codons for that amino acid and $S = \sum_{i=1}^{k} \left(\frac{m_i}{m}\right)^2$ where m_i is the number of occurrences of the *i*th codon for this amino acid	ENC values range between 20 and 61; a value of 20 indicates an extremely biased gene that uses only one codon for each amino acid, while a value of 61 indicates an unbiased gene	Ikemura (1981)
Chi-squared index	Calculates the divergence of the observed data from the values that would be expected under the null hypothesis of no association between observed and expected data	$\chi^2 = \sum_{i=1}^{18} \sum_{j=1}^{f_i} \left[(e_{i\_} 0_{ij})^2 / e_i\right]$ where o_{ij} is the number of occurrences of the *j*th codon for the *i*th amino acid, ei is the expected usage of the *j*th codon under conditions of equal synonymous codon usage, and f_i is the degeneracy of the codons for *i*th amino acid	The probability (*P*) of codon occurrence for a particular amino acid is estimated from the chi-square distribution (upper tail) based on calculated χ^2 value. If the value of *P* is less than 0.05, then the codon cannot occur for that particular amino acid and hence the null hypothesis cannot be accepted	Shields *et al.* (1988)
Frequency of optimal codons (F_{op})	Calculated as the ratio of the frequency of optimal codons in a gene to the total number of synonymous codons based on a specific reference gene set	$Fop = \frac{N_{opt}}{N_{tot}}$ where N_{opt} = number of optimal codons and N_{tot} = number of synonymous codons	Fop values range between 0 and 1.0: a value of 0 indicates that there is no optimal codon, and a value of 1.0 indicates that a gene is entirely composed of optimal codons	Ikemura (1985)
Codon adaptation index (CAI)	Quantifies the geometric mean of the relative adaptiveness for each codon with respect to the codon usage of a reference set of highly expressed genes and is calculated based on RSCU values	$CAI = \exp\left(1/L \sum_{i=1}^{L} \log(\omega_i(l))\right)$ $\omega_i = \frac{f_{ij}}{f_{xj}}$ where ω_i is the relative adaptiveness of codon i, fij is the frequency of codon i encoding amino acid *j*, and L is the length of the gene	CAI values range between 0 and 1: a value of 0 indicates random codon usage and low expression level of the gene, whereas a value of 1 suggests extreme codon bias and potentially high expression level of the gene	Sharp and Li (1987)
Codon bias index (CBI)	Measure of codon usage bias based on the codon usage of a specific reference set of genes	$CBI = \frac{N_{opt} - N_{ran}}{N_{tot} - N_{ran}}$ where N_{opt} = number of optimal codons, N_{ran} = number of optimal codons and N_{tot} = number of synonymous codons	CBI values range between 0 and 1: a value of 0 indicates random codon usage, whereas a value of 1 suggests extreme codon bias	Bennetzen and Hall (1982)

Methods that use codon usage indices summarize data into useful limited variables, facilitating comparisons of codon usage among species (Shields *et al.*, 1988). Indices are categorized into two groups: one group of codon usage indices measures the codon usage bias based on the total discrepancy between the expected codon usage (assuming no bias) and the actual codon usage, while a second group quantifies bias based on closeness to preferred codons by comparing the codon usage of a test gene to a reference set of genes. The principles for calculation and scoring are summarized in Table 1. The theories and mathematical formulas for the calculation of these indices are described in detail by Behura and Severson (2013). These indices, either alone or in combination, have been adopted in several online tools that are commonly used to measure codon usage (Table 2).

The P2 index was one of the earliest indices developed using the principle of distance between expected and actual codon use. It predicts the proportion of preferred codons based on codon–anticodon binding strength between an mRNA and the tRNAs. The P2 index has largely been replaced by other indices in most recent publications, with use of RSCU (relative synonymous codon usage), ENC (effective number of codons) and CAI (codon adaptation index) indices being widely reported. The RSCU measures codon bias based on non-random usage of codons for a specific amino acid in a coding sequence. As the RSCU value is independent of amino acid composition, this is very useful for the comparison of genes that differ in their length and amino acid composition, and hence this index is frequently used (Mukhopadhyay *et al.*, 2008; Sablok *et al.*, 2011; He *et al.*, 2016; Ning *et al.*, 2016). ENC values are based on a principle similar to that underlying RCSU; they are also relatively independent of amino acid composition. Comparison of ENC values with GC content yields information regarding patterns in codon usage across the genome. Recent studies have used ENC to determine codon usage bias in *Ginkgo biloba* (He *et al.*, 2016) and *Haberlea rhodopensis* (Ivanova *et al.*, 2017). Fop (frequency of optimal codons), CAI and CBI (codon bias index) measure codon usage bias according to the codon usage of a reference set of genes, which may be composed of a gene class, such as ribosomal genes, highly expressed genes or a single gene (Table 1). As Fop, CAI and CBI indices use specific reference gene sets, they provide a measure of bias towards particular codons that appear to be translationally optimal in that species (Behura and Severson, 2013). Among these indices, use of CAI is most widely reported, for example in codon usage analysis of *Oenothera* (Nair *et al.*, 2014) and *G. biloba* (He *et al.*, 2016), while Fop has been applied in studies for *Picea* (De La Torre *et al.*, 2015) and *Medicago truncatula* (Song *et al.*, 2015). As CAI values are determined from reference sets of highly expressed genes, which may be different among species, the relative fitness values of the species will also be different. Hence, the CAI values for all genes for a particular species can be compared within the species, but cannot be compared to those for other species. There are no recent reports of codon analysis in plants using chi-squared, CBI and P2. Early methods for visualizing codon usage data involved tabulating codon usage values from a pool of different sets of genes (Greenacre, 1984). Such analyses relied heavily on the grouping of genes, and these groups were formulated based on pairwise comparisons. However, analyses that can be performed using simple tabulation are limited and, with the emergence of increasingly large

TABLE 2. *Online computational tools for codon analysis*

Online tools	Application	Web link
ACUA (Automated Codon Usage Analysis Software)	Performs statistical profiling of codon usage in high-throughput sequence data	http://www.bioinsilico.com/acua
ANACONDA	Performs comparative analysis of codon context patterns of genomes based on Pearson's statistic	http://bioinformatics.ua.pt/software/anaconda/
CAIcal SERVER	Runs several computations in relation to codon preference and the codon adaptation of nucleic acid sequences to host organisms	http://genomes.urv.es/CAIcal
Codon Adaptation Index (CAI) Calculator 2	Measures codon bias within and across genomes	http://userpages.umbc.edu/~wug1/codon/cai/cais.php
Codon Explorer	A wrapper program of G-language REST/SOAP web services calculates codon usage visualizing genomic information and predicts gene expression levels from codon usage bias	http://www.g-language.org/gembassy/
Codon O	Measures synonymous codon usage bias within and across genomes	http://sysbio.cvm.msstate.edu/CodonO/
Codon W	Performs correspondence analysis of codon, analyses amino acid usage and measures standard indices of codon usage	http://codonw.sourceforge.net/
E-CAI Calculator	Determines codon adaptation index of nucleic acid sequences	http://genomes.urv.es/CAIcal/E-CAI/
GCUA (Graphical Codon Usage Analyser)	Displays codon usage frequency values or relative adaptiveness values	http://gcua.schoedl.de/
GCUA (General Codon Usage Analysis)	Uses multivariate analysis to estimate the variation in codon usage amongst genes	http://bioinf.nuim.ie/gcua/
Gene to codon usage	Creates a codon usage table from a DNA sequence, counting the frequency of all codons in the specified sequence	http://www.entelechon.com/2008/10/gene-to-codon-usage/
JCAT (Java Codon Adaptation Tool)	Analyses nucleic acid and protein sequences by avoiding rho-independent terminators, prokaryotic ribosome binding sites and restriction sites	http://www.jcat.de/
MEGA	Measures codon frequencies and using relative synonymous codon usage statistics	http://www.megasoftware.net/
Optimizer	Suggests optimized codon usage of a DNA sequence for increasing expression level	http://genomes.urv.es/OPTIMIZER/
Rare codon analysis	Analyses codon preference input coding sequences for use prior to heterologous expression	http://www.genscript.com/cgi-bin/tools/rare_codon_analysis
RCDI (Relative Codon Deoptimization Index)	Measures the codon deoptimization by comparing the codon usage of a gene with the codon usage of a reference genome	http://genomes.urv.cat/CAIcal/RCDI/
RSCUNET (Relative Synonymous Codon Usage Neural Network)	Analyses codon usage variation based on a self-organizing map neural network algorithm	http://bioinf.nuigalway.ie/RescueNet

sets of sequence data, it became cumbersome to tabulate the large sets of codon usage data. To address this issue in analytical approaches to codon analysis, tools that adapt multivariate analysis methods such as correspondence analysis (COA) and cluster analysis (Gouy and Gautier, 1982), either alone or in combination, have been adopted for visualizing codon usage pattern in several online tools (Table 2).

Cluster analysis is the statistical visualization method that has been most widely used to analyse codon usage in plant genomes. In this method, the frequencies of codon occurrences are calculated using codon usage indices such as RSCU and CAI. Euclidean measurements of distance between the genes and the codon frequencies are estimated and the values plotted to visually indicate the association of genes to a particular codon and their similarities with other genes (Perrière and Thioulouse, 2002). Examples of cluster analysis applied to plant species include studies of the codon usage of protein coding sequences in *G. biloba* (He *et al.*, 2016), herbaceous peony (*Paeonia lactiflora*) (Wu *et al.*, 2015), citrus species (Xu *et al.*, 2013), the chloroplast genome of *Oenothera* (Nair *et al.*, 2014) and the mitochondrial genomes of wheat (*Triticum aestivum*), maize (*Zea mays*), *Arabidopsis thaliana*, tobacco (*Nicotiana tabacum*), *Physcomitrella patens* and *Marchantia polymorpha* (Zhou and Li, 2009). Cluster analyses have also been used to classify and categorize codon usage in various plant genomes (Ma *et al.*, 2015; Priya *et al.*, 2016), but in a few cases, the method did not accurately reflect the phylogenetic relationships among plants (Nair *et al.*, 2014; You *et al.*, 2015). With the advent of genome-scale sequencing projects and greater computational power, improved tools are being developed to cope with the size and complexity of the data. While several sets of data are available for plant species, the lack of high-quality transcriptome data that are comparable in depth and biological context limits the value and integrity of analyses; as with any bioinformatic approach, it is critical to identify high-quality and comparable datasets for meaningful outcomes.

Codon usage patterns in monocots

Among monocots, several major food crops are members of the Poaceae (grass monocots) and have consequently been well studied, including complete genome sequencing, so it is not surprising that most of the codon usage studies in monocots have focused on this family. Campbell and Gowri (1990) were the first to illustrate the differential usage of codons in monocot species and found that the third codon nucleotide position shows bias toward A/U in dicot species and toward G/C in monocot species, concurrent with the higher GC content of monocot genomes. A bias for G/C-ending codons has also been reported for wheat, barley, maize (Kawabe and Miyashita, 2003) and rice (Wang and Hickey, 2007; Mukhopadhyay *et al.*, 2008). The higher numbers of G/C-ending codons were proposed as supporting evidence of relatively strong mutational bias in monocot species (Kawabe and Miyashita, 2003; Wang and Hickey, 2007; Mukhopadhyay *et al.*, 2008).

Among all plants, monocots have been reported to be the most heterogenous in terms of their GC content, which contrasts with the homogeneous GC content of the genomes of dicot species (Carels and Bernardi, 2000; Wang and Hickey, 2007; Serres-Giardi *et al.*, 2012; Tatarinova *et al.*, 2013). This GC heterogeneity is similar to that found in the genomes of warm-blooded vertebrates, which contain patches of GC-rich and AT-rich regions known as isochores (Duret and Galtier, 2009). These isochore structures have been reported to affect both coding and non-coding sequences of genomes, while particularly GC-rich regions have been observed to contain more genes with shorter introns (Eyre-Waker and Hurst, 2001). As GC_3 is relatively independent of the protein coding potential, it is considered a marker for GC richness in a genome (Tatarinova *et al.*, 2013). Analysis of GC_3 distribution showed a tendency for genes to fall into two classes (bimodality) in grass monocots, unlike the single class of genes (unimodal distributions) in dicot genomes (Campbell and Gowri, 1990; Tatarinova *et al.*, 2013; Clément *et al.*, 2015). Genes in the grass monocots (rice) could be classified into the two groups based on GC_3 values below or above 80 %. In the rice genome, genes with GC_3 of 80 % or more showed a tendency to be mono-exonic or intron-poor and have stronger or more variable expression levels (Tatarinova *et al.*, 2013). Based on a study of GC_3 in seven grass species together with banana and oil palm, Clément et al. (2015) suggested that a bimodal distribution may be an ancestral feature of monocots that is preserved to different extents in different lineages.

To explain this notable GC heterogeneity in monocots, initially MB (Wang and Roossinck, 2006; Wang and Hickey, 2007) was proposed as a major underlying evolutionary force shaping codon use, based on the comparative analysis of Arabidopsis and rice genomes. However, mutation can hardly explain recombination-associated segregation distortion that favours GC over AT alleles and consequently high GC content in high-recombination regions, which was revealed by analyses of several varieties of rice (*Oryza rufipogon*, *O. sativa* subsp. *japonica* and *O. sativa* subsp. *indica*, *O. barthii* and *O. meridionali*) polymorphism datasets (Muyle *et al.*, 2011). Alternatively, SCU was suggested as a determinant of codon usage and heterogeneous GC content in rice (Guo *et al.*, 2007; Mukhopadhyay *et al.*, 2008). An explanation has been put forward that GC heterogeneity could be related to the regulation of gene expression, because the local GC content is associated with the distribution of genes and chromosomal architecture (reviewed by Mugal *et al.*, 2015). However, SCU cannot explain GC heterogeneity in non-coding regions found in rice (Muyle *et al.*, 2011). With more genome sequences being available for monocot species, gBGC appears to be the more suitable hypothesis to explain the evolutionary forces determining overall GC heterogeneity and codon usage in monocots (Muyle *et al.*, 2011; Serres-Giardi *et al.*, 2012; Camiolo *et al.*, 2015; Clément *et al.*, 2017).

In addition to studies of codon usage in the nuclear genome, a few studies report codon usages in monocot chloroplast genomes. Codon usage bias in the chloroplast genomes of the grass monocots *Brachypodium distachyon*, *Triticum aestivum* and *Hordeum vulgare* showed optimal codons to mainly have A/T in the third codon position, unlike their respective codons in nuclear genomes (Sablok *et al.*, 2011; Zhang *et al.*, 2012). The major underlying force was proposed to be MB with the additional influence of natural selection, hydrophobicity, aromaticity and gene length.

Very little information is available on codon usage of non-grass monocots, and most of the reports relate to chloroplast genomes, rather than nuclear-encoded gene sequences, which mainly focused on preferred codon in third codon position and preferred amino acid using small datasets (Table 3). These studies cannot capture the full range of variability in this group and thus may affect the reliability of the findings for making general conclusions. So far in non-grass monocot species, a detailed study analysed codon usage and quantified the magnitude of evolutionary forces including oil palm (*Elaeis guineensis*) and banana (*Musa acuminata*) (Clément *et al.*, 2017). The analysis showed that nucleotide composition of the non-grass monocots is far from mutation-drift equilibrium and that gBGC is a more widespread and stronger process than selection (Clément *et al.*, 2017).

Motivated by the lack of comprehensive codon usage information among non-grass monocot genomes and aided by the growth of transcriptome data in publicly available databases, we investigated the patterns of codon usage within members of the non-grass monocots by analysing coding sequences (CDS) of the two major banana species, *Musa acuminata* (A genome) (Martin *et al.*, 2016) and *Musa balbisiana* (B genome) (Davey *et al.*, 2013), the date palm, *Phoenix dactylifera* (Al-Dous *et al.*, 2011), and the small aquatic plant, duckweed, *Spirodela polyrhiza* (Wang *et al.*, 2014) (Fig. 1). In addition, we used leaf transcript data for the moss bamboo, *Phyllostachys heterocycla* var. *pubescens*, as the representative grass monocot outgroup (Peng *et al.*, 2013).

MATERIALS AND METHODS

Sequence data

To compare codon usage metrics between equivalent CDS data for each species, we selected transcriptome datasets with similar sequencing depth and from equivalent tissues (leaf was the most available across species) and for species with completed and available genome sequences. From their respective genome sequence databases, CDS of the two major banana species, *Musa acuminata* (A genome) (Martin *et al.*, 2016; http://banana-genome-hub.southgreen.fr/download) and *Musa balbisiana* (B genome) (Davey *et al.*, 2013; http://banana-genome-hub.southgreen.fr/organism/Musa/balbisiana), the date palm, *Pho. dactylifera* (Al-Dous *et al.*, 2011; https://qatar-weill.cornell.edu/research-labs-and-programs/date-palm-research-program/date-palm-draft-sequence), and the small aquatic plant, duckweed, *Spirodela polyrhiza* (Wang *et al.*, 2014; https://www.waksman.rutgers.edu/spirodela/genome), were retrieved. In addition, the CDS of the moss bamboo, *Phy. heterocycla* var. *pubescens*, was used as the representative grass monocot outgroup (Peng *et al.*, 2013; http://202.127.18.221/bamboo/down.php). All the CDS used in the analyses were derived from young non-treated leaves. After predicting open reading frames (ORFs) using Transdecoder v3.01 (Haas and Papanicolaou, 2013), a total set of 41 390 CDS from *M. acuminata*, 24 250 CDS from *M. balbisiana*, 11 932 CDS from *Pho. dactylifera*, 15 151 CDS from *S. polyrhiza* and 27 420 CDS from *P. heterocycle* was obtained, which were used for further analysis using Anaconda

TABLE 3. *Summary of codons usage studies in non-grass monocots*

Parameters	Orchidaceae		Arecaceae	Arecaceae	Liliaceae		Musaceae		Amaryllidaceae
Species	*Phalaenopsis aphrodite*	*Oncidium* Gower Ramsey	*Phoenix dactylifera*	*Elaeis guineensis*	*Colchicum autumnale*	*Gloriosa superba*	*Musa acuminata*		*Allium cepa*
Genome type	Chloroplast	Chloroplast	Chloroplast	Nuclear	Chloroplast	Chloroplast	Nuclear	Chloroplast	43.50 %
GC content (%)		37.00		48–50.8					
AT content (%)		62.68							
Codon preference at 3rd codon position	A or T	A or T			A or T		G or C	A or T	G or C
Main factor for shaping codon usage		Mutational bias		GC-biased gene conversion			GC-biased gene conversion		
Total codons representing all the protein-coding genes			22 950	22 950					
Most preferred stop codon				TGA					
Most frequent amino acids			Isoleucine		Isoleucine	Isoleucine		Isoleucine, leucine	
Least frequent amino acids			Cysteine					Cysteine	
Reference(s)	Chang *et al.* (2006)	Xu *et al.* (2011)	Yang *et al.* (2010)	Jouannic *et al.* (2005), Nakamura *et al.* (2000), Low *et al.* (2008), Ho *et al.* (2007), Clément *et al.* (2017)	Nguyen *et al.* (2015)	Nguyen *et al.* (2015)	Clément *et al.* (2017)	D'Hont *et al.* (2012), Martin *et al.* (2013)	Kuhl *et al.* (2004)

2 software (a software package developed by the University of Aveiro Bioinformatics research group, Portugal) (Moura *et al.*, 2005). CDS shorter than 300 base pairs were excluded from the analysis. For each transcriptome, the sequences with more than one stop codon, lacking a start or stop codon, lacking complete reading frames and/or with undetermined nucleotides (N) were discarded by using filters in the Anaconda 2 software (Moura *et al.*, 2005). Hypothetical and duplicated genes were identified using BLAST and were discarded from the datasets used for codon analysis.

To perform correlation analysis between GC content of CDS with genome sequences for each species, BLASTn of CDS was performed against the genomes of *M. acuminata* (http://banana-genome-hub.southgreen.fr/download), *M. balbisiana* (http://banana-genome-hub.southgreen.fr/organism/Musa/balbisiana) and *Pho. dactylifera* (https://qatar-weill.cornell.edu/research-labs-and-programs/date-palm-research-program/date-palm-draft-sequence). Matches with an e-value <10e-5 and at least 2 kb flanking region both 5′ and 3′ to the CDS were extracted as gene sequences, resulting in non-redundant sets of 2282 genes from *M. acuminata*, 9742 genes from *M. balbisiana* and 5701 genes from *Pho. dactylifera*. The GC content (%) of genes [along with untranslated region (UTR)] and of flanking regions was calculated separately.

Codon usage analyses

RSCU values were estimated using the following formula from Sharp and Li (1987) using Anaconda 2 software (Moura *et al.*, 2005):

$$RSCU = \frac{g_{ij}}{\sum_{j}^{ni} g_{ij}} n_i$$

where g_{ij} is the observed number of the *i*th codon for the *j*th amino acid, which has n_i kinds of synonymous codons.

The RSCU values obtained from the analysis are listed in conventional amino acid codon order and highlighted on a scale from green (indicating the most rarely used codons) to red (indicating the most frequently used codons). An RSCU value of 1.0 indicates no codon usage bias while values above or below 1.0 indicate codons that are utilized less or more frequently (preferred) than expected.

CAI values were obtained using the Codon W program (version 1.4.2; http://codonw.sourceforge.net). The genes falling within the top and bottom 5 % of CAI values were considered as the high and low datasets. RSCU values of these two datasets were compared using the chi-square contingency test. Those codons whose usage frequency was significantly higher ($P < 0.01$) in highly expressed genes compared to those genes with low levels of expression were identified as optimal codons (Sablok *et al.*, 2011).

The nucleotide compositions of the third codon position in the full set of ORFs (A_3, U_3, C_3 and G_3) were calculated for each species using the Codon W program. The parity rule 2 (PR2) bias was detected based on AU bias [$A_3/(A_3+U_3)$] as the *y*-axis and GC-bias [$G_3/(G_3+C_3)$] as the *x*-axis at the third codon position of the two- and four-codon amino acids of entire CDS (Sueoka, 1995).

GC content of the genome, transcriptome and the GC frequency in the three codon positions of individual coding sequences (GC_1, GC_2 and GC_3) were measured for all species using Anaconda 2 software (Moura *et al.*, 2005), and GC_{12} was calculated as the average of GC_1 and GC_2.

Frequency distributions of GC_3 (all CDS with start and stop codon) were plotted according to the method suggested by Carels and Bernardi (2000). Neutrality plots of GC_{12} against GC_3 were analysed based on the method suggested by Sueoka (1988).

Effective numbers of codons (ENC) were calculated by following the formula of Wright (1990), using Anaconda 2 software (Moura *et al.*, 2005):

$$ENC = 2 + \frac{9}{\bar{F}_2} + \frac{1}{\bar{F}_3} + \frac{5}{\bar{F}_4} + \frac{3}{\bar{F}_6}$$

where the $\bar{F}$ value denotes the probability that two randomly chosen codons for an amino acid with two codons are identical.

ENC values were plotted against GC_3 values. The expected ENC values were calculated using the following formula:

$$ENCexp = 2 + S + \left({}^{29}\!/_{S2 + (1 - S2)} \right)$$

where *S* is the frequency of GC_3s (Wright, 1990). To estimate the difference between the observed and the expected ENC values for all CDS, frequency distributions of (ENCexp – ENCobs)/ENCexp were plotted. A paired Student's *t*-test ($P < 0.001$) was performed to test for significance using SPSS 16.0 (http://www.spss.com/).

COA (Greenacre, 1984) was performed with CodonW using the RSCU values to compare the intra-genomic variation of 59 informative codons, partitioned along 59 orthogonal axes with 41 degrees of freedom. Correlation analyses (Pearson correlation, two-tailed) was performed for GC content and codon usage indices with SPSS 16.0 (http://www.spss.com/).

Codon-pair context analysis

Codon-pair context quantifications of the transcriptome datasets were performed using Anaconda 2 (Moura *et al.*, 2005). The Anaconda 2 software package uses transcriptome data to generate a frequency table consisting of codon-pair contexts using a ribosome simulation algorithm. The resulting data are interrogated via a chi-square test. The results are then transferred to a visualization module to generate a codon-pair context map.

RESULTS

Preferred, optimal and rare codons in non-grass monocots

To analyse the codon usage patterns in non-grass monocots, RSCU for *M. acuminata*, *M. balbisiana*, *Pho. dactylifera*, *S. polyrhiza* and *Phy. heterocycla* was calculated (Supplementary Data Table S1). *M. acuminata* and *M. balbisiana* showed an approximately equal number of preferred AT- and GC-ending codons (Table 4). By contrast, *S. polyrhiza* and *Phy. heterocycla* showed higher GC-ending codon as preferred codons. In addition, non-grass monocot species showed avoidance of XUA codons [AUA (ILE), UUA (LEU), CUA (LEU) and GUA (VAL)] in all species (Supplementary Data Table

TABLE 4. *Preferred codons, optimal codons and rare codons in non-grass monocots*

Codon type	Codon 3rd base	*M. acuminata*	*M. balbisiana*	*Pho. dactylifera*	*S. polyrhiza*	*Phy. heterocycla*
RSCU > 1	G/C/A/U	34	33	30	29	33
	G/C	16	17	12	26	21
	A/U	18	16	18	3	12
	A	6	6	6	1	4
	U	12	10	12	2	8
	G	8	7	7	11	8
	C	8	10	5	15	13
Optimal	G/C/A/U	27	27	27	27	27
	G/C	27	27	27	27	27
	A/U	0	0	0	0	1
	A	0	1	0	1	1
	U	0	0	0	0	0
	G	11	11	10	11	11
	C	16	16	16	16	16
Rare codon	G/C/A/U	1	1	0	3	4
	U	0	0	0	1	3
	A	1	1	0	2	1

S1). This avoidance may be related to the use of UA in stop codons and/or the presence of UA-selective ribonucleases, as suggested by Beulter *et al.* (1989). All optimal codons for the studied species showed G/C at the third codon position (shown as '*' in Table 4; Supplementary Data Table S2A–E). These codons are optimal for highly expressed genes (based on CAI) and do not reflect the corresponding tRNA pool. The most preferred stop codon for *M. acuminata*, *M. balbisiana* and*Phy. heterocycla* was UGA, whereas *Pho. dactylifera* and *S. polyrhiza* preferred UGA and UAG. Furthermore, rare codons (RSCU < 0.10) in non-grass monocots showed A/U at the third codon position similar to *Phy. heterocycla* (shown as '–' in Table 4; Supplementary Data Table S2A–E).

CG dinucleotide suppression

Analysis of the XCG/XCC ratio based on RSCU values from the present analysis in non-grass monocot species showed values of 0.62 (*M. acuminata*), 0.64 (*M. balbisiana*), 0.57 (*Pho. dactylifera*) and 0.69 (*S. polyrhiza*) (Supplementary Data Table S1). This indicates moderate CG dinucleotide suppression in non-grass monocots. Compared to established non-grass species, *Phy. heterocycla* exhibited a higher XCG/XCC ratio, 0.82.

PR2-bias plot analysis

PR2-bias plot analysis was carried out to investigate the effect of mutation and selection pressure on codon usage bias. If only mutational bias shapes the codon usage bias, then G and C (A and U) should be used proportionally among the two- and four-fold codon amino acids. On the other hand, if natural selection dominates, it would not necessarily cause proportional use of G and C (A and U) (Wright, 1990; Sueoka, 1995). The results showed that G and C (A and U) were not used proportionally in non-grass monocot genomes (Supplementary Data Fig. S1A–D) as was also observed for *Phy. heterocycla* (Supplementary Data Fig. S1E), which indicates mutational force is not the major force in determining codon usage in monocots.

GC pattern in non-grass monocots

Total GC content of the genome and transcriptome and the GC (pair) content in the three codon positions of coding sequences were examined for the four non-grass species and grass monocot *Phy. heterocycla* (Fig. 2). A higher genomic GC content was observed in *Phy. heterocycla* (43.9 %) compared to the non-grass monocots. Among the non-grass monocots, *S. polyrhiza* showed a higher genomic GC value (41.96 %) and transcriptomic GC content (56.16 %) compared to other species. Total GC content analysis of non-grass monocots revealed a higher transcriptomic GC compared to genomic GC in all species. Furthermore, the non-grass monocot species *M. acuminata*, *M. balbisiana* and *Pho. dactylifera* demonstrated their highest GC contents at the first codon position, followed by GC in the third and the second codon positions ($GC_1 > GC_3 > GC_2$) (Fig. 2). Analysis of the pattern of GC distribution in the grass monocot used in this study (*Phy. heterocycla*) showed that it shared a similar pattern for GC content in the three codon positions ($GC_1 > GC_3 > GC_2$) as previously reported in grasses (Wong *et al.*, 2002), but this was not the case for most of the non-grass monocots; three of the four non-grass monocots in our study shared a pattern of $GC_1 > GC_3 > GC_2$ (Fig. 2) as previously reported for other non-grass monocots onion (Kuhl *et al.*, 2004) and *Lycoris longituba* (Cui *et al.*, 2004) as well as for the dicot Arabidopsis (Kuhl *et al.*, 2004).

Variation of GC_3 was further analysed in monocots (Fig. 3), which revealed a unimodal distribution of GC_3 for all studied non-grass monocot species and bimodal distribution for the grass monocot *Phy. heterocycla.* Compared to *M. acuminata* and *M. balbisiana*, *Pho. dactylifera* and *S. polyrhiza* showed a clear unimodal distribution.

Association between ENC and GC_3 in non-grass monocots

The influence of GC_3 content on the codon usage of non-grass monocots was further examined by calculating the ENC (Wright, 1990) for each gene and plotting this against the GC_3 composition of each individual gene (Fig. 4A–D). The CDS of

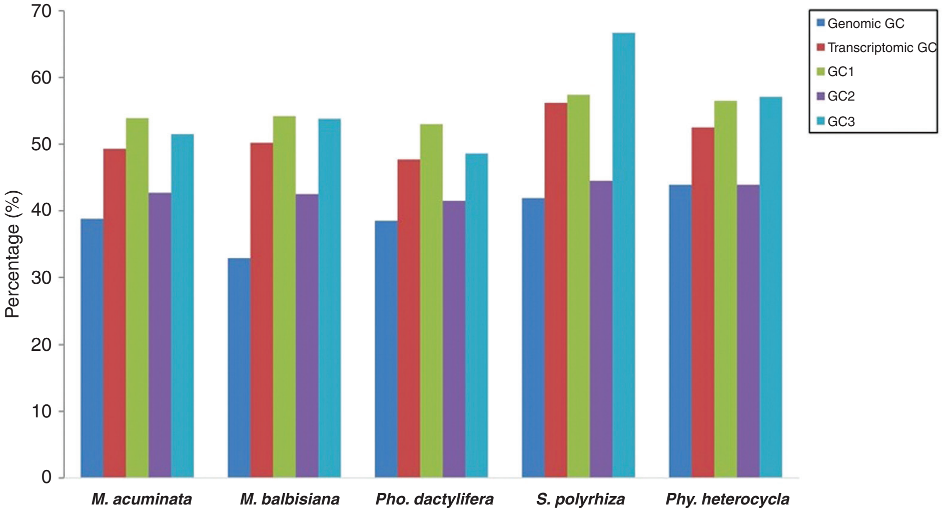

FIG. 2. Variation of GC content in *M. acuminata*, *M. balbisiana*, *Pho. dactylifera*, *S. polyrhiza* and *Phy. heterocycla*

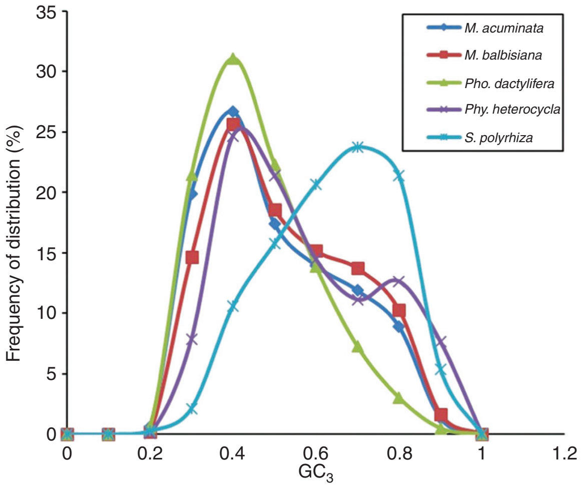

FIG. 3. Distributions of GC_3 content in *M. acuminata*, *M. balbisiana*, *Pho. dactylifera*, *S. polyrhiza* and *Phy. heterocycla*

non-grass monocot species showed a wide distribution of GC_3 from 20 to 80 % (0.2–0.8 in Fig. 4A–D), similar to *Phy. heterocycla* (Fig. 4E).

Analysis of the ENC values of individual genes within each of the non-grass monocot species showed variability of bias, with values of 20.00–61.00 for *M. acuminata*, 23.26–61.00 for *M. balbisiana*, 22.38–61.00 for *Pho. dactylifera* and 21.78–61.00 for *S. polyrhiza*. This indicates the presence of genes with a high bias to no bias within non-grass monocot genomes, as was also observed for the grass monocot *Phy. heterocycla* (20.86–61.00). ENC values were plotted against GC_3 to determine whether the difference in ENC is related to the difference in GC_3 content (Fig. 4). In general, the ENC plot is a parabolic curve in which genes are represented as points and where the expected curve shows the positions of genes under no selection that are only subject to GC_3 codon compositional constraint (Wright, 1990). Such comparison of the actual distribution of genes with expected distribution explains the presence of an influence of forces other than compositional constraint (Wright, 1990). The ENC plot of all the species under study showed, although some genes lay on the expected curve, large numbers of CDS with low ENC values lying below the expected curve. The presence of points far below the curve indicates these genes have additional codon usage bias that is independent of GC_3s (Wright, 1990).

A significant negative correlation of GC_3 with ENC values was observed for non-grass monocot species. This negative correlation indicates that genes with higher GC_3 values and lower ENC values had strong codon usage bias. Liu and Xue (2005) also reported a similar negative correlation of GC_3 with ENC values of CDS in four grass monocot species (rice, maize, barley and wheat). To estimate the difference between observed and expected ENC value of CDS we calculated (ENCexp – ENCobs)/ENCexp ratio (Supplementary Data Fig. S2A–E). Comparison of observed ENCs with expected ENC using paired Student's *t*-tests ($P < 0.001$) revealed significant differences.

COA and correlations between codon usage, nucleotide composition, CAI, ENC, hydrophobicity, aromaticity and gene length in non-grass monocots

To further understand the variation in shaping codon usage bias in coding sequences of non-grass monocots, COA was performed (Fig. 5A–D). In the COA, a series of orthogonal axes were generated that reflect the trends of variation in codon usage bias based on RSCU values of the CDS (Greenacre, 1984). The distance between CDS on this plot is a reflection of their dissimilarity in RSCU with respect to the two axes. The first main dimensional coordinates, axis 1, explained 36.87 % (*M. acuminata*), 38.67 % (*M. balbisiana*), 23.28 % (*Pho. dactylifera*) and 36.84 % (*S. polyrhiza*) of overall codon usage variation in the studied non-grass monocot species. Axis 2 accounted for 4.72 % (*M. acuminata*), 4.53 % (*M. balbisiana*), 5.64 % (*Pho. dactylifera)* and 4.73 % (*S. polyrhiza*). The remaining axes 3 and 4 each represented even smaller amounts of the variance. Axis 3

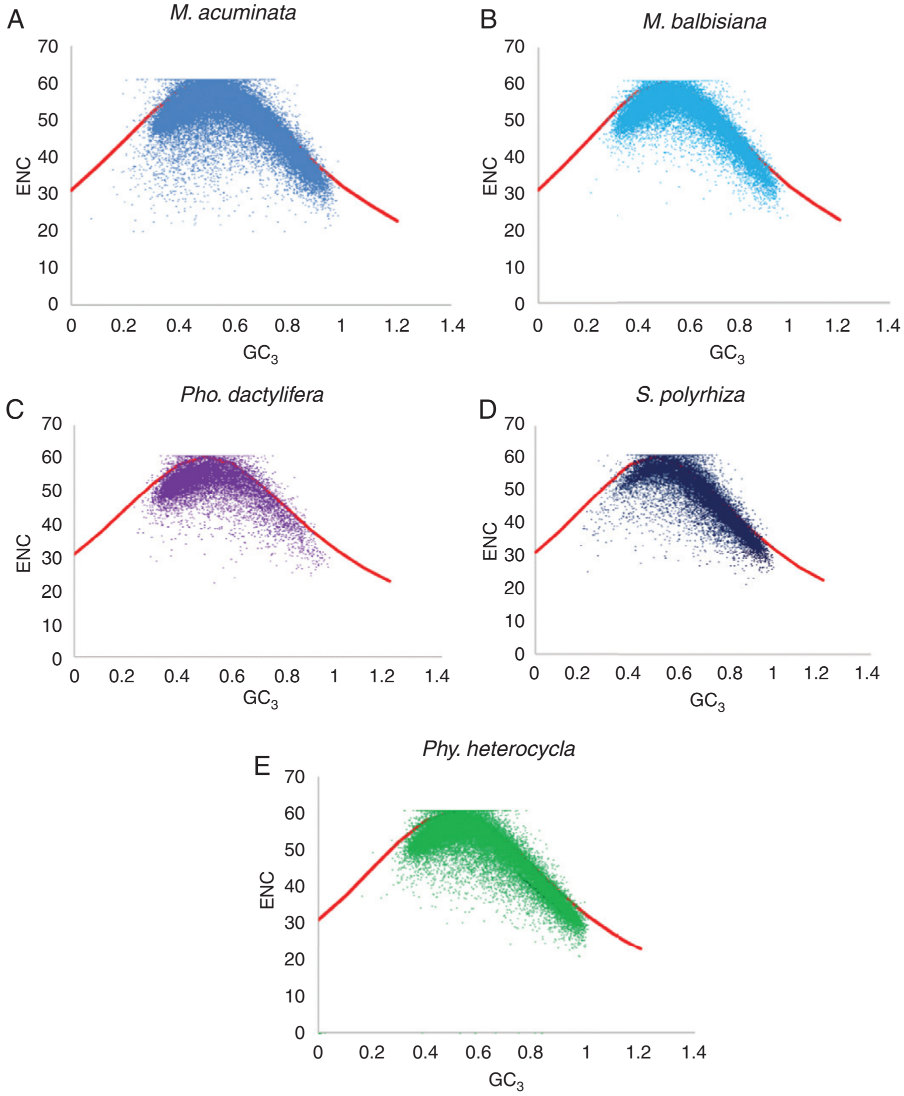

Fig. 4. Relationship between GC_3 and effective numbers of codon (ENC plot) for *M. acuminata*, *M. balbisiana*, *Pho. dactylifera*, *S. polyrhiza*, and *Phy. heterocycla.* ENC values were plotted against GC_3 values following the method of Wright (1990). The continuous curve depicts the expected curve between ENCs and GC_3. (A) *M. acuminata*, (B) *M. balbisiana*, (C) *Pho. dactylifera*, (D) *S. polyrhiza* and (E) *Phy. heterocycla.*

showed 3.99 % (*M. acuminata*), 3.45 % (*M. balbisiana*), 4.34 % (*Pho. dactylifera*) and 3.66 % (*S. polyrhiza*) of the variation, and axis 4 3.14 % (*M. acuminata*), 3.19 % (*M. balbisiana*), 3.88 % (*Pho. dactylifera*) and 2.68 % (*S. polyrhiza*). The grass monocot (*Phy. heterocycla*) also showed a similar trend (axis 1: 37.99 %, axis 2: 4.67 %, axis 3: 3.50 %, axis 4: 3.09 %) (Fig. 5E). Thus, the first axis reflects the major factor that explains the differences in codon usage among studied species. To elucidate the effects of the GC content on codon usage bias, CDS with different GC contents were labelled with different colours. Genes with GC of 60 % were plotted in grey, while genes with a GC less than 45 % were plotted in blue (Fig. 5). Orange dots indicate genes with a GC content between 45 and 60 %. The plot showed a clear separation of high- and low-GC genes along the primary axis.

Although Fig. 5 presents a general relationship between the nucleotide content of genes and their position on the first axis of the COA, it does not give a statistical measure of this relationship. To do this, the correlation between the GC content, codon usage indices of individual CDS and their location on the primary axis of the COA was measured (Table 5). Axis 1 showed significant correlations between GC content and position on the first axis (Table 5). The gene positions on axis 1 showed strong and significant correlation with GC_3 value. ENC values of *M. acuminata*, *M. balbisiana* and *Pho. dactylifera* showed a significant negative correlation, whereas *S. polyrhiza* showed a positive correlation with ENC (Table 5). In addition, correlation analysis between ENC and GC_3 showed ENC values were negatively correlated with GC_3 values for all species. Strong correlation between axis 1 and CAI and between GC_3 and CAI was observed. A positive correlation between hydropathy and axis 1 was found for *M. acuminate*, *M. balbisiana*, *Pho. dactylifera* and *Phy. heterocycla*, but a negative correlation for *S. polyrhiza*. A weak negative correlation was observed

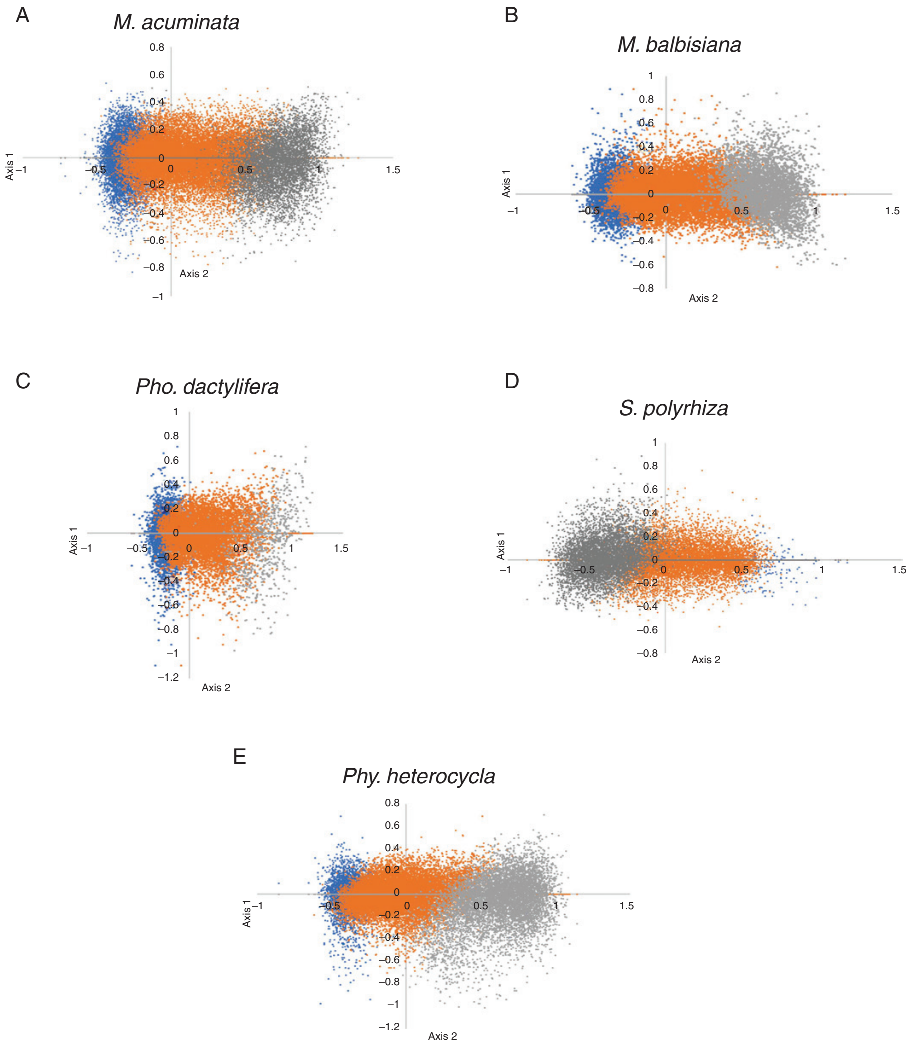

FIG. 5. Correspondence analysis of coding sequences in (A) *M. acuminata*, (B) *M. balbisiana*, (C) *Pho. dactylifera*, (D) *S. polyrhiza* and (E) *Phy. heterocycla*. Grey: genes with GC of 60 %; blue: genes with GC < 45 %; orange: genes with GC between 45 and 60 %.

for aromaticity and axis 1 for *M. acuminate*, *M. balbisiana* and *Phy. heterocycla.* A strong positive correlation was found between GC_3 and GC content in the other two codon positions (GC_1 and GC_2) for all species. Gene length showed a negative correlation with axis 1 and a negative correlation with GC_3 in all studied monocot species.

The relationship between coding sequence GC_3 content and genomic GC content was further examined in three non-grass monocot species, *M. acuminata*, *M. balbisiana* and *Pho. dactylifera*. *S. polyrhiza* and *Phy. heterocycla* were excluded from the analysis due to a lack of suitable data. A strong and highly significant positive correlation was observed between coding GC_3 content and genomic GC content (Fig. 6) for all three species. However, correlation analysis between GC_3 and flanking GC content showed significant but low correlation coefficients for *M. acuminata* ($r = 0.065$, $P < 0.1$), *M. balbisiana* ($r = 0.063$, $P < 0.1$) and *Pho. dactylifera* ($r = 0.013$, $P < 0.1$).

Variation in codon-pair context in non-grass monocots

Codon-pair context maps generated for *M. acuminata*, *M. balbisiana*, *Pho. dactylifera*, *S. polyrhiza* and *Phy. heterocycla* transcriptomes (Fig. 7A–E) revealed variations and similarities among these non-grass monocots. The region highlighted by two parallel diagonal lines for each dataset in Fig. 7 represents the codon-pair contexts for the codons with identical or similar sequences. The codon pairs that fall within the diagonal area indicate a degree of tandem codon duplications; a characteristic of codons in eukaryotic genomes is an identical codon on their 3′ side (Moura *et al.*, 2005). In the present monocot genome study, codon-pair context sequences GAG-GAG, AAG-AAG and GAU-GAU were observed to be the most frequently used (Supplementary Data Table S3). The homogeneous codon-pair contexts NNN-GAA, NNN-AAG and NNN-GAG were also observed to be frequently used codon-pairs for all species examined in this study (Supplementary Data Table S3).

TABLE 5. *Correlation of codon usage, nucleotide composition, ENC, CAI, hydropathy, aromaticity and gene length in non-grass monocots*

Species	Parameters	Total GC	GC_1	GC_2	GC_3	ENC	CAI	Hydropathy	Aromaticity	Gene length
M. acuminata	Axis 1	0.966**	0.606**	0.484**	0.988**	−0.625**	0.973**	0.173**	−0.006	−0.329**
	GC_3	0.937**	0.544**	0.384**		−0.615***	0.973**			−0.332**
	GC_{12}				0.562**					−0.323**
	Total GC									
M. balbisiana	Axis1	0.970**	0.634**	0.527**	0.990**	−0.673**	0.976**	0.173**	−0.013*	−0.302**
	GC_3	0.944**	0.575**	0.452**		−0.665**	0.979**			−.308**
	Total GC									−0.302**
	GC_{12}				0.604**					
Pho. dactylifera	Axis1	0.928**	0.424**	0.328**	.980**	−0.334**	0.966**	0.112**	0.021*	−0.262**
	GC_3	0.892**	0.351**	0.235**		−0.341**	0.960**			−0.066**
	Total GC									−0.249**
	GC_{12}				0.3655**					
S. polyrhiza	Axis1	−0.953**	−0.588**	−0.483**	−0.992**	0.842**	0.976**	−0.169**	0.006	−0.214**
	GC_3	0.922**	0.534**	0.406**		−0.840**	0.973**			−0.213**
	Total GC									−0.218**
	GC_{12}				0.550**					
Phy. heterocycla	Axis1	0.958**	0.661**	0.509**	0.989**	−0.813**	0.980**	0.180**	−0.059**	−0.334**
	GC_3	0.931**	0.603 **	0.435**		−0.808**	0.976**			−0.332**
	Total GC									−0.353**
	GC_{12}				0.592**					

* Correlation is significant at the 0.05 level; ** correlation is significant at the 0.01 level.

Condon-pair context results indicate highest similarity between *M. acuminata* and *M. balbisiana* among all of the non-grass monocot species analysed. To identify the differences in preferred codon-pair context between the two *Musa* species, a differential codon-pair context map (DCM) was constructed by overlapping the complete codon-pair context maps (Fig. 7F). A colour scale based on gradation of yellow was used for the differential display. Common features are indicated in black and significant differences are represented in yellow. Despite the two species being closely related, they showed significant differences.

DISCUSSION

The increase in publicly available large transcriptome datasets for non-grass monocot species has presented an opportunity for determining codon usage and codon bias in this important but so far neglected group of plants. However, the sourcing of materials for transcriptomic analysis from different tissues, which furthermore may have been exposed to different conditions, results in the creation of datasets that are not equivalent and therefore not optimal for comparison. In the present study, four non-grass monocots (*M. acuminata*,

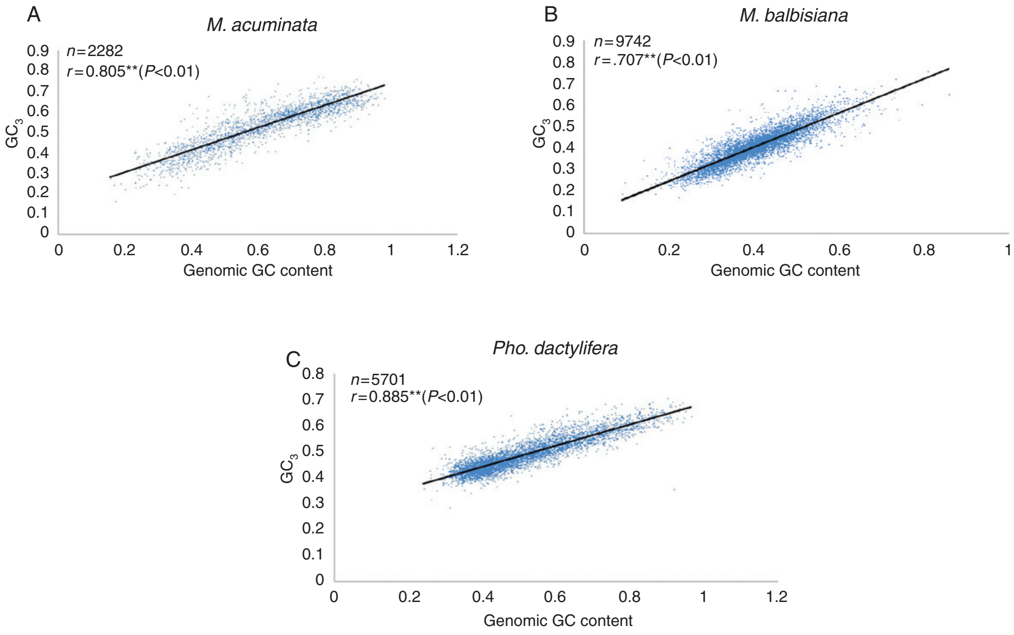

FIG. 6. Correlation analysis between GC content in third codon position and genomic GC content for (A) *M. acuminata*, (B) *M. balbisiana* and (C) *Pho. dactylifera.*

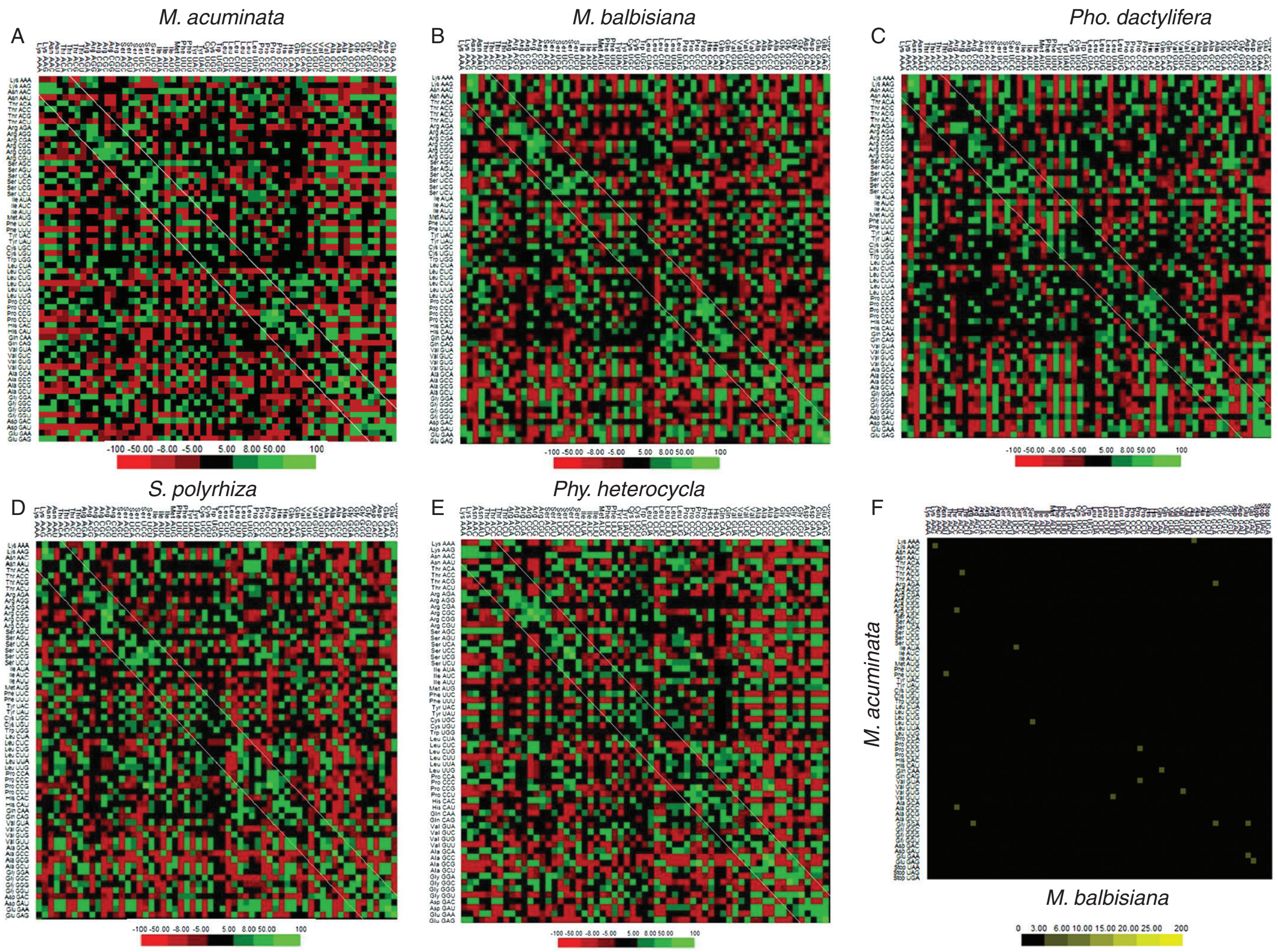

FIG. 7. Codon-pair context patterns for (A) *M. acuminata*, (B) *M. balbisiana*, (C) *Pho. dactylifera*, (D) *S. polyrhiza* and (E) *Phy. heterocycla.* Green represents preferred codon-pair contexts and red represents rejected codon-pair contexts. Values that are not statistically significant are represented in black. The regions between the two parallel diagonal lines represent the codon-pair contexts that are more frequently used compared to other contexts in the species under study. (F) Differential display map for comparative analysis of codon-pair context between *M. acuminata* and *M. balbisiana*. A colour scale based on gradation of yellow was used for the differential display. Common features are indicated in black and the differences are represented in yellow.

M. balbisiana, *Pho. dactylifera*, *S. polyrhiza*) and one grass monocot (*Phy. heterocycla*) were selected, for which comparable datasets were publicly available: The transcriptome datasets used in the present study were constructed from young non-treated leaves for each species and represented similar relative genome coverage, to provide as unbiased a comparison as was possible, which we believe adds value to our analysis. Whether the data were able to support each of the three current hypothesis, MB, SCU and gBGC, for the mechanisms which determine codon usage bias in non-grass monocots was also tested. Analysis was made of codon usage, optimal and rare codons, GC content, GC_3 distribution, codon-pair context patterns and shaping factors in four non-grass monocots and one grass monocot genome.

To place the results in the context of previous analysis of codon usage in plants, the literature was reviewed, with a focus on monocot species. There is a consensus that GC content, which varies widely across plant genomes, is one of the most important factors in determining codon usage in a genome. The codon usage bias corresponds to the AT- or GC-rich content of a genome, while the third nucleotide position of a codon reflects genome base composition of an organism (Wu *et al.*, 2015). The overall GC contents reported in monocot groups (55–59 %) are much higher than for dicot species (42–44 %) (Kawabe and Miyashita, 2003; Garg *et al.*, 2011; Singh *et al.*, 2016). GC values among non-grass monocot genomes (47–56 %; Fig. 2) were found overall to be higher than those reported for dicot species, the variation within this group probably reflecting differences in mating systems between species.

While previous reports have suggested that the high GC content of monocot genomes is a consequence of MB within monocot genomes (Wang and Roossinck, 2006; Wang and Hickey, 2007), the MB hypothesis appears not to explain nucleotide composition or codon usage of non-grass monocots. According to the MB hypothesis, high GC content is a consequence of a positive feedback loop to reduce cytosine deamination and favouring GC enrichment in monocots (Fryxell and Zuckerkandl, 2000). However, to the best of our knowledge, all mutations reported in eukaryotic species show an AT bias (Dillon *et al.*, 2015), including rice (Muyle *et al.*, 2011) and the non-grass monocots banana and oil palm (Clément *et al.*, 2017). Similarly, mutations of methylated cytosine, a cause of GC enrichment in grass monocots

(Ossowski *et al.*, 2010; Alonso *et al.*, 2015), was also found to be unlikely as methylation tends to increase the mutation bias toward AT (as methylated CpGs are usually highly mutable toward TpG) (Nachman and Crowell, 2000). Furthermore, methylation levels reported in monocots including rice and secale (Kalinka *et al.*, 2017) are much higher compared to the dicot species poplar and *Arabidopsis* (Feng *et al*., 2010). Therefore, if MB was the major evolutionary force in shaping codon usage in monocots, one would expect monocots to have a lower GC content than dicots, which is not the case. Finally, as all classes of mutations have an equal probability of fixation, a proportionate usage between G and C and between A and U would be expected. However, analysis of the rice genome showed disproportionate usage between G and C and between A and U in the third codon position and higher probability of fixation of GC over AT bases (Muyle *et al.*, 2011). Analysis of polymorphism data in banana and oil palm has also shown similar results (Clément *et al.*, 2017). A similar disproportionate usage between G and C and between A and U in the third codon position in non-grass monocots was observed here (Supplementary Data Fig. S1). Hence, together with the present data, the MB hypothesis appears not to be suitable to explain nucleotide composition and codon usage in non-grass monocots. Further availability of polymorphism data on non-grass monocots will be required to completely rule out MB as the major influence in shaping nucleotide composition in non-grass monocots.

Whether SCU can better explain the data was next examined, and it was found that while some criteria can fit the model, these could also result from gBGC and that SCU is also unable to explain the GC_3 heterogeneity seen in non-grass monocots. According to SCU, high GC content is the consequence of selective pressures to ensure translational efficiency of genes, which is supported by the fact that most of the monocot species have more GC-ending codons in optimal codons (Muyle *et al.*, 2011). In the presemt data analysis, a strong correlation between GC_3 we observed content and codon usage variation was observed, and between GC_3 content and CAI (predicts the level of expression of a gene) values (Fig. 5A–D; Table 5). In addition, all optimal codons were found to have GC in the third codon position (Supplementary Data Table S2), which appears to be an influence of SCU (Table 5). However, the optimal codons presented in the current study show a bias for highly expressed genes (based on CAI) and do not reflect the corresponding tRNA pool (Supplementary Data Table S2). Hence, this cannot be concluded to be indicative of SCU. Beside this, our analysis on GC_3 distribution between CDS of non-grass monocot species showed a remarkable GC content heterogeneity, with values ranging from 20 to 80 % (Figs 3 and 4A–D), contrary to the narrower distribution of GC_3 (20–60 %) reported in dicot species (Kawabe and Miyashita, 2003; Mukhopadhyay *et al.*, 2008) and not compatible with the SCU hypothesis. The heterogenic genome composition observed in non-grass monocots is similar to the genome organization reported in mammals (reviewed by Duret and Galtier, 2009) and the grass monocots rice, maize, barley and wheat (Kawabe and Miyashita, 2003; Tatarinova *et al.*, 2010; Šmarda and Bureš, 2012).

Overall, we found the gBGC hypothesis, where high GC content is the result of bias of GC over AT bases during mismatch repair during meiotic recombination (reviewed by Duret and Galtier, 2009), to best explain the codon usage in non-grass monocots. Experimental evidence of gBGC has been found in the grass monocots *Secale cereal*, *Aegilops speltoides*, *Triticum urartu*, *Triticum monococcu* (Haudry *et al.*, 2008), rice (Muyle *et al.*, 2011) and maize (Rodgers-Melnick *et al.*, 2016) based on recombination data of effective populations, which showed a strong correlation between recombination rate and GC content of those genomes. The heterogeneous GC_3 distribution in monocots, which is incongruous with the SCU hypothesis, can easily be explained as a consequence of gBGC. An initial study on grass genomes explored the presence of a strong negative 5′–3′ GC content gradient along genes (Wong *et al.*, 2002). This strong 5′–3′ GC decreasing pattern was also reported for both exonic and intronic regions of genes (Zhu *et al.*, 2009) and was suggested to be a consequence of recombination hotspots around transcription start sites (Hellsten *et al.*, 2013). Over time, this has resulted in genes which range from short mono-exonic and GC-rich to those that are longer and relatively GC-poor (Ressayre *et al.*, 2015). Hence, grass monocots with a strong recombination and gBGC gradient clearly show a bimodal distribution, representing short and long genes (Glémin *et al.*, 2014). By contrast, in non-grass monocots even where the gene composition is similar, a weaker gradient results in a unimodal distribution (Clément *et al.*, 2015). Our analysis on GC_3 distribution also showed a clear bimodal distribution for the grass species (*Phy. heterocycla*) and unimodal distribution in non-grass monocots (Fig. 3), in agreement with the gBGC hypothesis. Among non-grass monocots, there is still a lack of direct evidence of gBGC due to the unavailability of recombination data. Nevertheless, Clément et al. (2017) quantified the magnitude of gBGC based on correlative approaches between neutrality and selection indices to disentangle the processes of SCU and gBGC in several plant species, including the non-grass monocots banana and oil palm, finding a stronger intensity of gBGC in shaping nucleotide composition and codon usage in non-grass monocots. They proposed that comparisons of GC content in coding regions with introns or non-coding regions would be helpful to confirm these findings. The present data analysis showed a strong positive correlation between GC_3 and genomic GC content in the three studied non-grass monocot species (Fig. 6). In addition, low but significant correlation coefficients between GC_3 and flanking GC content were found, similar to that reported for other grass monocots including maize, rice, sorghum (Tatarinova *et al.*, 2010; Glémin *et al.*, 2014) and the non-grass monocot banana (Glémin *et al.*, 2014) in which the GC content is mainly shaped by gBGC. These results demonstrate that the GC bias we report in non-grass monocots is not restricted to third-codon positions but affects surrounding sites as well. Furthermore, consistent with an influence of gBGC, GC-rich CDS were shorter (Table 5), similar to that reported for the grass monocots rice (Wang and Hickey, 2007, Muyle *et al.*, 2011) and maize (Sundararajan *et al.*, 2016).

Taken together, the analyses of non-grass monocots indicate that gBGC is the main driving force in this group of plants. Quantification of differences in the occurrence, intensity and patterns of gBGC could explain the variations at all positions across non-grass monocot genomes. Further availability of recombination data and genome data well-annotated for coding and non-coding sequences will help to elucidate the impact of gBGC in non-grass monocots.

Apart from codon usage, codon-pair context also has a significant impact on the translational selection of genes (Moura *et al.*, 2007) due to its crucial role in suppressing mutations (Kopelowitz *et al.*, 1992; Irwin *et al.*, 1995). Prior to the current study, there were no reports on codon-pair contexts in monocots so here we can report that, as for other phyla, significant differences in codon-pair context patterns were found between two *Musa* species (*M. acuminata* and *M. balbisiana*) (Fig. 7F). The data also support the view of Moura *et al.* (2007) that codon-pair context maps are species-specific and can be used as a fingerprint for a specific species. Among the codon-pair contexts for the non-grass monocots, a preference was observed for homogeneous codon-pairs (Fig. 7A–D, Supplementary Data Table S3). There are no previous reports on homogeneous codon pairs other than in tea and mustard (Paul and Chakraborty, 2016) and insects (Diptera and Hymenoptera) (Behura and Severson, 2012), but this may be worth further investigation as the use of homogeneous codon pairs has been proposed as a tactic to save energy during translation: the energy required to carry out the translation of homogeneous codon-pair contexts will be lower as fewer tRNA species need to be synthesized (Moura *et al.*, 2011). However, it has also been suggested that the continuous use of the same set of tRNAs could limit the translational speed of highly expressed proteins and narrow the amino acid variability of proteins (Irwin *et al.*, 1995). At the current time, there are limited accessible comparable sets of gene expression data for relative analysis of tRNA and mRNA from plant genomes. The increase in plant transcriptome studies and anticipated improved databases to share assembled and annotated plant transcriptomes with associated expression data should make it feasible to provide supporting data for these hypotheses in the near future.

CONCLUSIONS

Codon usage bias is largely influenced by nucleotide composition across the genome. Knowledge about active evolutionary forces, codon usage and codon-pair context in plants is a prerequisite to deciphering the molecular mechanisms underlying survival strategies and functional adaptation of this kingdom of sessile organisms. While early studies of plant genomes led to MB and SCU being proposed as the major evolutionary forces in shaping codon usage in the plant genome, gBGC is now regarded as the prevalent and robust evolutionary process at play. Turning our attention to codon usage in non-grass monocots, our analysis of CDS based on leaf transcriptomic data for four completely sequenced transcriptomes (*M. acuminata*, *M. balbisiana*, *Pho. dactylifera* and *S. polyrhiza*) showed optimal codons of non-grass monocots have a strong preference to end in G or C and, unlike grass monocots, the non-grass monocots in our study showed a unimodal GC_3 distribution. Based on our observation of a positive relationship between CAI (predicts the level of expression of a gene) and GC_3 as well as a positive relationship between GC_3 in codons and GC content of non-coding DNA, we conclude that codon usage and nucleotide composition in non-grass monocots are mainly influenced by gBGC. To quantify the intensity and patterns of gBGC in non-grass monocot genomes will require additional research, including comprehensive recombination and methylation maps at the gene scale. Such data can be generated either by direct sequencing of meiosis products (parents and progenies) (Yang *et al.*, 2012) or by indirect population genomic methods (Muyle *et al.*, 2011).

SUPPLEMENTARY DATA

Supplementary data are available online at www.aob.oxfordjournals.org and consist of the following. Figure S1: parity rule 2-bias plot $[A_3/(A_3+T_3)]$ against $G_3/(G_3+U_3)]$ for *M. acuminata*, *M. balbisiana*, *Pho. dactylifera*, *S. polyrhiza* and *Phy. heterocycla*. Figure S2: frequency distribution of effective number of codons ratio for *M. acuminata*, *M. balbisiana*, *Pho. dactylifera, S. polyrhiza* and *Phy. heterocycla*. Table S1: relative synonymous codon usage values of *M. acuminata*, *M. balbisiana*, *Pho. dactylifera*, *S. polyrhiza* and *Phy. heterocycla.* Table S2: RSCU values of highly expressed and lowly expressed genes for *M. acuminata*, *M. balbisiana*, *Pho. dactylifera*, *S. polyrhiza* and *Phy. heterocycla*. Table S3: codon-pair frequency in *M. acuminata*, *M. balbisiana*, *Pho. dactylifera*, *S. polyrhiza* and *Phy. heterocycla.*

ACKNOWLEDGEMENTS

PM, KM and JAH are partially supported by CEBAR Research University grant (RU006-2017 and RU015-2016).

LITERATURE CITED

Al-Dous E, **George B**, **Al-Mahmoud ME**, ***et al*. 2011**. *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature Biotechnology* **29**: 521–527.

Akashi H. 2003. Translational selection and yeast proteome evolution. *Genetics* **164**: 1291–1303.

Alonso C, **Pérez R**, **Bazaga P**, **Herrera CM. 2015**. Global DNA cytosine methylation as an evolving trait: phylogenetic signal and correlated evolution with genome size in angiosperms. *Frontiers in Genetics* **6**: 4.

Behura SK, **Severson DW. 2012**. Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. *PLoS ONE* **8**: 43111.

Behura SK, **Severson DW. 2013**. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. *Biological Review of Cambridge Philosophical Society* **88**: 49–61.

Bennetzen JL, **Hall BD. 1982**. Codon selection in yeast. *Journal of Biological Chemistry* **257**: 3026–3031.

Beulter E, **Gelbart T**, **Han J**, **Koziol JA**, **Beutler B. 1989**. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proceedings of the National Academy of Sciences of the United States of America* **86**: 192–196.

Campbell WH, **Gowri G. 1990**. Codon usage in higher plants, green algae and cyanobacteria. *Plant Physiology* **92**: 1–11.

Camiolo S, **Melito S**, **Porceddu A. 2015**. New insights into the interplay between codon bias determinants in plants. *DNA Research* **5**: 27.

Carels N, **Bernardi G. 2000**. Two classes of genes in plants. *Genetics* **154**: 1819–1825.

Chang C-C, **Lin H-C**, **Lin I-P**, ***et al*. 2006**. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Molecular Biology and Evolution* **23**: 279–291.

Clément Y, **Fustier MA**, **Nabholz B**, **Glémin S. 2015**. The bimodal distribution of genic GC content is ancestral to monocot species. *Genome Biology and Evolution* **1**: 336–348.

Clément Y, **Sarah G**, **Holtz Y**, ***et al*. 2017**. Evolutionary forces affecting synonymous variations in plant genomes. *PLoS Genetics* **13**: 1006799.

Cui Y, **Zhang X**, **Zhou Y**, ***et al***. **2004**. Identification and expression analysis of EST-based genes in the bud of *Lycoris longituba*. *Genomics, Proteomics & Bioinformatics* **2**: 43–46.

D'Hont A, **Denoeud F**, **Aury JM**, ***et al***. **2012**. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**: 213–217.

Davey MW, **Gudimella R**, **Harikrishna JA**, **Lee WS**, **Khalid N**, **Keulemans W. 2013**. A draft *Musa balbisiana* genome sequence for molecular genetics in polypoid, inter- and intra-specific Musa hybrids. *BMC Genomics* **14**: 683.

Duret L, **Galtier N. 2009**. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics* **10**: 285–311.

De La Torre AR, **Lin YC**, **Van de Peer Y**, **Ingvarsson PK. 2015**. Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in picea gene families. *Genome Biology and Evolution* **7**: 1002–1015.

Dillon MM, **Sung W**, **Lynch M**, **Cooper VS. 2015**. The rate and molecular spectrum of spontaneous mutations in the GC-rich multichromosome genome of *Burkholderia cenocepacia*. *Genetics* **200**: 935–946.

Eyre-Walker A, **Hurst LD. 2001**. The evolution of isochores. *Nature Reviews Genetics* **2**: 549–555.

Figuet E, **Ballenghien M**, **Romiguier J**, **Galtier N. 2014**. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biology and Evolution* **7**: 240–250.

Fryxell KJ, **Zuckerkandl E. 2000**. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Molecular Biology and Evolution* **17**: 1371–83.

Garg R, **Patel RK**, **Jhanwar S**, ***et al***. **2011**. Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. *Plant Physiology* **156**: 1661–1678.

Galtier N, **Duret L**, **Glémin S**, **Ranwez V. 2009**. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in Genetics* **25**: 1–5.

Glémin S, **Clément Y**, **David J**, **Ressayre A. 2014**. GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends in Genetics* **30**: 263–270.

Glémin S, **Arndt PF**, **Messer PW**, **Petrov D**, **Galtier N**, **Duret L. 2015**. Quantification of GC-biased gene conversion in the human genome. *Genome Research* **25**: 1215–1228.

Greenacre MJ. 1984. *Theory and applications of correspondence analysis*. London: Academic Press.

Gouy M, **Gautier C. 1982**. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research* **10**: 7055–7074.

Gossmann TI, **Song BH**, **Windsor AJ**, ***et al***. **2010**. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution* **27**: 1822–1832.

Guo X, **Bao J**, **Fan L. 2007**. Evidence of selectively driven codon usage in rice: implications for GC content evolution of Gramineae genes. *FEBS Letters* **581**: 1015–1021.

Haas BJ, **Papanicolaou A**, **Yassour M**, ***et al.*** **2013**. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature protocols* **8**: 1494–1512.

Harrison RJ, **Charlesworth B. 2010**. Biased gene conversion affects patterns of codon usage and amino acid usage in the *Saccharomyces sensu stricto* group of yeasts. *Molecular Biology and Evolution* **28**: 117–129.

Haudry A, **Cenci A**, **Guilhaumon C**, ***et al***. **2008**. Mating system and recombination affect molecular evolution in four Triticeae species. *Genetics Research* **90**: 97–109.

He B, **Dong H**, **Jiang C**, **Cao F**, **Tao S**, **Xu LA. 2016**. Analysis of codon usage patterns in *Ginkgo biloba* reveals codon usage tendency from A/U-ending to G/C-ending. *Scientific Reports* **6**: 35927.

Hellsten U, **Wright KM**, **Jenkins J**, ***et al***. **2013**. Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. *Proceedings of the National Academy of Sciences USA* **110**: 19478–19482.

Hertweck KL, **Kinney MS**, **Stuart SA**, ***et al***. **2015**. Phylogenetics, divergence times and diversification from three genomic partitions in monocots. *Botanical Journal of the Linnean Society* **178**: 375–393.

Hershberg R, **Petrov DA. 2008**. Selection on codon bias. *Annual Review of Genetics* **42**: 287–299.

Ho C-L, **Kwan Y-Y**, **Choi M-C**, ***et al***. **2007**. Analysis and functional annotation of expressed sequence tags (ESTs) from multiple tissues of oil palm (*Elaeis guineensis* Jacq.). *BMC Genomics* **8**: 381.

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* system. *Journal of Molecular Biology* **151**: 389–409.

Ikemura T. 1985. Codon usage and transfer-RNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2**: 13–34.

Irwin B, **Heck JD**, **Hatfield GW. 1995**. Codon pair utilization biases influence translational elongation step times. *The Journal of Biological Chemistry* **270**: 22801–22806.

Ivanova Z, **Sablok G**, **Daskalova E**, ***et al***. **2017**. Chloroplast genome analysis of resurrection tertiary relict *Haberlea rhodopensis* highlights genes important for desiccation stress response. *Frontiers in Plant Science* **8**: 204.

Jouannic S, **Argout X**, **Lechauve F**, ***et al***. **2005**. Analysis of expressed sequence tags from oil palm (*Elaeis guineensis). FEBS Letters* **579**: 2709–2714.

Kalinka A, **Achrem M**, **Poter P. 2017**. The DNA methylation level against the background of the genome size and t-heterochromatin content in some species of the genus *Secale* L. *PeerJ* **5**: 2889.

Kawabe A, **Miyashita NT. 2003**. Pattern of codon usage bias in three dicot and four monocot plant species. *Genes& Genetic Systems* **78**: 343–352.

Kopelowitz J, **Hampe C**, **Goldman R**, **Reches M**, **Engelberg-Kulka H. 1992**. Influence of codon context on UGA suppression and read through. *Journal of Molecular Biology* **225**: 261–269.

Kuhl JC, **Cheung F**, **Yuan Q**, ***et al***. **2004**. A unique set of 11,008 onion expressed sequence tags reveals expressed sequence and genomic differences between the monocot orders Asparagales and Poales. *The Plant Cell* **16**: 114–125.

Lesecque Y, **Mouchiroud D**, **Duret L. 2013**. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Molecular Biology and Evolution* **30**: 1409–1419.

Liu Q. 2012. Mutational bias and translational selection shaping the codon usage pattern of tissue-specific genes in rice. *PLoS ONE* **7**: 10.

Liu Q, **Xue Q. 2005**. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *Journal of Genetics* **84**: 1.

Liu Q, **Hu H**, **Wang H. 2015**. Mutational bias is the driving force for shaping the synonymous codon usage pattern of alternatively spliced genes in rice (*Oryza sativa* L.). *Molecular Genetics and Genomics* **290**: 649–660.

Low E-TL, **Alias H**, **Boon S-H**, ***et al***. **2008**. Oil palm (*Elaeis guineensis* Jacq.) tissue culture ESTs: Identifying genes associated with callogenesis and embryogenesis. *BMC Plant Biology* **8**: 62.

Ma QP, **Li C**, **Wang J**, **Wang Y**, **Ding ZT. 2015**. Analysis of synonymous codon usage in FAD7 genes from different plant species. *Genetics and Molecular Research* **14**: 1414–1422.

Martin G, **Baurens F-C**, **Cardi C**, **Aury J-M**, **D'Hont A. 2013**. The complete chloroplast genome of banana (*Musa acuminata*, Zingiberales): insight into plastid monocotyledon evolution. *PLoS ONE* **8**: 67350.

Martin G, **Baurens F-C**, **Droc G**, ***et al***. **2016**. Improvement of the banana '*Musa acuminata*' reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics* **17**: 243.

Moura G, **Pinheiro M**, **Silva R**, ***et al***. **2005**. Comparative context analysis of codon pairs on ORFeome scale. *Genome Biology* **6**: 28.

Moura G, **Pinheiro M**, **Arrais J**, ***et al***. **2007**. Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLoS ONE* **2**: 847.

Moura GR, **Pinheiro M**, **Freitas A**, ***et al***. **2011**. Species-specific codon context rules unveil non-neutrality effects of synonymous mutations. *PLoS ONE* **6**: 26817.

Mugal CF, **Arndt PF**, **Holm L**, **Ellegren H. 2015**. Evolutionary consequences of DNA methylation on the GC content in vertebrate genomes. *G3: Genes, Genomes, Genetics* **5**: 441–447.

Mukhopadhyay P, **Basak S**, **Ghosh TC. 2008**. Differential selective constraints shaping codon usage pattern of housekeeping and tissue-specific homologous genes of rice and Arabidopsis. *DNA Research* **15**: 347–356.

Muyle A, **Serres-Giardi L**, **Ressayre A**, **Escobar J**, **Glémin S. 2011**. GC-biased gene conversion and selection affect GC content in the *Oryza genus* (rice). *Molecular Biology and Evolution* **28**: 2695–2706.

Nachman MW, **Crowell SL. 2000**. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**: 297–304.

Nair RR, **Raveendran NT**, **Dirisala VR**, ***et al***. **2014**. Mutational pressure drives evolution of synonymous codon usage in genetically distinct *Oenothera* plastomes. *Iranian Journal of Biotechnology* **12**: 58–72.

Necşulea A, **Popa A**, **Cooper DN**, ***et al*. 2011**. Meiotic recombination favors the spreading of deleterious mutations in human populations. *Human Mutation* **32**: 198–206.

Nguyen PA, **Kim JS**, **Kim J-H. 2015**. The complete chloroplast genome of colchicine plants (*Colchicum autumnale* L. and *Gloriosa superba* L.) and its application for identifying the genus. *Planta* **242**: 223–237.

Ning LI, **Sun MH**, **Jiang ZS**, **Shu HR**, **Zhang SZ. 2016**. Genome-wide analysis of the synonymous codon usage patterns in apple. *Journal of Integrative Agriculture* **15**: 983–991.

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

Paul P, **Chakraborty S. 2016**. Codon usage bias analysis for the coding sequences of *Camellia sinensis* and *Brassica campestris. African Journal of Biotechnology* **15**: 236–251.

Peng Z, **Lu Y**, **Li L**, **Zhao Q**, ***et al*. 2013**. The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). *Nature Genetics* **45**: 456–461.

Perrière G, **Thioulouse J. 2002**. Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Research* **30**: 4548–4555.

Pessia E, **Popa A**, **Mousset S**, **Rezvoy C**, **Duret L**, **Marais GA. 2012**. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biology and Evolution* **4**: 675–82.

Plotkin JB, **Kudla G. 2011**. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics* **12**: 32–42.

Priya R, **Dass FP**, **Siva R. 2016**. Gene expression prediction and hierarchical clustering analysis of plant CCD genes. *Plant Molecular Biology Report* **34**: 618–627.

Ratnakumar A, **Mousset S**, **Glémin S**, ***et al*. 2010**. Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **365**: 2571–2580.

Ressayre A, **Glemin S**, **Montalent P**, **Serre-Giardi L**, **Dillmann C**, **Joets J. 2015**. Introns structure patterns of variation in nucleotide composition in *Arabidopsis thaliana* and rice protein-coding genes. *Genome Biology and Evolution* **7**: 2913–2928.

Rodgers-Melnick E, **Vera DL**, **Bass HW**, **Buckler ES. 2016**. Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences USA* **113**: 3177–3184.

Singh R, **Ming R**, **Yu Q. 2016**. Comparative analysis of GC content variations in plant genomes. *Tropical Plant Biology* **1**: 14.

Sablok G, **Nayak KC**, **Vazquez F**, **Tatarinova TV. 2011**. Synonymous codon usage, GC3, and evolutionary patterns across plastomes of three pooid model species: emerging grass genome models for monocots. *Molecular Biotechnology* **49**: 116–128.

Serres-Giardi L, **Belkhir K**, **David J**, **Glémin S. 2012**. Patterns and evolution of nucleotide landscapes in seed plants. *The Plant Cell* **24**: 1379–1397.

Sharp PM, **Li WH. 1987**. The rate of synonymous substitution in Enterobacterial genes is inversely related to codon usage bias. *Molecular Biology and Evolution* **4**: 222–230.

Sharp PM, **Tuohy TM**, **Mosurski KR. 1986**. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* **14**: 5125–5143.

Shields DC, **Sharp PM**, **Higgins DG**, **Wright F. 1988**. 'Silent' sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution* **5**: 704–716.

Šmarda P, **Bureš P. 2012**. The variation of base composition in plant genomes. *In Plant Genome Diversity Springer Vienna* **1**: 209–235.

Song H, **Wang P**, **Ma D**, **Xia H**, **Zhao C**, **Zhang Y**, **Zhao S. 2015**. Analysis of codon usage bias in *Medicago truncatula* WRKY transcription factors. *Journal of Agricultural Biotechnology* **23**: 203–212.

Spencer CCA. 2006. Human polymorphism around recombination hotspots. *Biochemical Society Transactions* **34**: 535–536.

Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences USA* **85**: 2653–2657.

Sueoka N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *Journal of Molecular Evolution* **40**: 318–325.

Sundararajan A, **Dukowic-Schulze S**, **Kwicklis M**, ***et al*. 2016**. Gene evolutionary trajectories and GC patterns driven by recombination in *Zea mays*. *Frontiers in Plant Science* **7**: 1433.

Tatarinova TV, **Alexandrov NN**, **Bouck JB**, **Feldmann KA. 2010**. GC_3 biology in corn, rice, sorghum and other grasses. *BMC Genomics* **16**: 308.

Tatarinova TV, **Elhaik E**, **Pellegrini M. 2013**. Cross-species analysis of genic GC3 content and DNA methylation patterns. *Genome Biology & Evolution* **5**: 1443–1456.

Wang H-C, **Hickey DA. 2007**. Rapid divergence of codon usage patterns within the rice genome. *BMC Evolutionary Biology* **7**: 1–10.

Wang L, **Roossinck MJ. 2006**. Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants. *Plant Molecular Biology* **61**: 699–710.

Wang W, **Haberer G**, **Gundlach H**, ***et al*. 2014**. The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nature Communications* **5**: 3311.

Weber CC, **Boussau B**, **Romiguier J**, **Jarvis ED**, **Ellegren H. 2014**. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biology* **15**: 549.

Wong GK, **Wang J**, **Tao L**, ***et al*. 2002**. Compositional gradients in Gramineae genes. *Genome Research* **12**: 851–856.

Wright F. 1990. The effective number of codons used in a gene. *Gene* **87**: 23–29.

Wu Y, **Zhao D**, **Tao J. 2015**. Analysis of codon usage patterns in herbaceous peony (*Paeonia lactiflora* Pall.) based on transcriptome data. *Genes* **6**: 1125–1139.

Xu C, **Cai X**, **Chen Q**, **Zhou H**, **Cai Y**, **Ben A. 2011**. Factors affecting synonymous codon usage bias in chloroplast genome of *Oncidium* Gower Ramsey. *Evolutionary Bioinformatics* **7**: 271–8.

Xu C, **Dong J**, **Tong C**, **Gong X**, **Wen Q**, **Zhuge Q. 2013**. Analysis of synonymous codon usage patterns in seven different citrus species. *Evolutionary Bioinformatics* **9**: 215–228.

Yang M, **Zhang X**, **Liu G**, ***et al*. 2010**. The Complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). *PLoS ONE* **5**: 12762.

Yang S, **Yuan Y**, **Wang L**, ***et al*. 2012**. Great majority of recombination events in Arabidopsis are gene conversion events. *Proceedings of the National Academy of Sciences USA* **109**: 20992–20997.

You E, **Wang Y**, **Ding ZT**, **Zhang XF**, **Pan LL**, **Zheng C. 2015**. Codon usage bias analysis for the spermidine synthase gene from *Camellia sinensis* (L.) O. Kuntze. *Genetics and Molecular Research* **14**: 7368–7376.

Zhang Y, **Nie X**, **Jia X**, ***et al*. 2012**. Analysis of codon usage patterns of the chloroplast genomes in the Poaceae family. *Australian Journal of Botany* **60**: 461–470.

Zhou M, **Li X. 2009**. Analysis of synonymous codon usage patterns in different plant mitochondrial genomes. *Molecular Biology Report* **36**: 2039–2046.

Zhu L, **Zhang Y**, **Zhang W**, **Yang S**, **Chen JQ**, **Tian D. 2009**. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics* **10**: 47.