Physiol Mol Biol Plants (February 2021) 27(2):213–222
https://doi.org/10.1007/s12298-021-00938-y

RESEARCH ARTICLE

Comprehensive analysis of coding sequence architecture features and gene expression in *Arachis duranensis*

Shuwei Dong[1] · Long Zhang[1] · Wenhui Pang[1] · Yongli Zhang[1] · Chang Wang[1] · Zhenyi Li[1] · Lichao Ma[1] · Wei Tang[1] · Guofeng Yang[1] · Hui Song[1]

Received: 13 December 2020 / Revised: 4 January 2021 / Accepted: 20 January 2021 / Published online: 18 February 2021

Abstract Coding sequence (CDS) architecture affects gene expression levels in organisms. Codon optimization can increase the gene expression level. Therefore, understanding codon usage patterns has important implications for research on genetic engineering and exogenous gene expression. To date, the codon usage patterns of many model plants have been analyzed. However, the relationship between CDS architecture and gene expression in *Arachis duranensis* remains poorly understood. According to the results of genome sequencing, *A. duranensis* has many resistant genes that can be used to improve the cultivated peanut. In this study, bioinformatic approaches were used to estimate *A. duranensis* CDS architectures, including frequency of the optimal codon (Fop), polypeptide length and GC contents at the first (GC1), second (GC2) and third (GC3) codon positions. In addition, *Arachis* RNA-seq datasets were downloaded from PeanutBase. The relationships between gene expression and CDS architecture were assessed both under normal growth as well as nematode and drought stress conditions. A total of 26 codons with high frequency were identified, which preferentially ended with A or T in *A. duranensis* CDSs under the above-mentioned three conditions. A similar CDS architecture was found in differentially expressed genes (DEGs) under nematode and drought stresses. The GC1 content differed between DEGs and non-differentially expressed genes (NDEGs) under both drought and nematode stresses. The expression levels of DEGs were affected by different CDS architectures compared with NDEGs under drought stress. In addition, no correlation was found between differential gene expression and CDS architecture neither under nematode nor under drought stress. These results aid the understanding of gene expression in *A. duranensis*.

Keywords *Arachis duranensis* · Coding sequence architecture · Codon usage bias · Drought stress · Gene expression · Nematode infection

Introduction

A growing body of evidence indicates that several features of the gene sequence architecture (such as synonymous codon usage, amino acid composition, polypeptide length, GC content and intron size) correlate with expression levels in prokaryotes and eukaryotes (Akashi 2001; Arunkumar et al. 2013; Chaney and Clark 2015; De La Torre et al. 2015; Duret 2000; Song et al. 2017a, 2017c; Williford and Demuth 2012). If a relationship exists between the abundance of aminoacyl-tRNA and the time required to occupy the acceptor site on a ribosome, then, genes with codons that correspond to the most abundant tRNAs may be translated faster than other genes in yeast (Arava 2003). In the *Chilodonella uncinata* genome, highly expressed genes are biased toward using optimal codons (Maurer-Alcalá and Katz 2016). In both *Gryllus bimaculatus* and *Oncopeltus fasciatus*, the expression levels were negatively correlated with amino acids with high size/complexity (S/C) scores (Whittle and Extavour 2015). However, in *Parhyale hawaiensis*, highly expressed genes had both large and small S/C values (Whittle and Extavour

Shuwei Dong and Long Zhang contributed equally to this work.

✉ Hui Song
biosonghui@outlook.com

[1] Grassland Agri-Husbandry Research Center, College of Grassland Science, Qingdao Agricultural University, Qingdao, China

2015). Short and GC-rich CDSs correlated positively with expression and optimal codon usage bias in four monocots, fifteen dicots and two mosses (Camiolo et al. 2015). In *Silene latifolia*, gene expression was positively correlated with the GC content at the third codon position (GC3), but highly negatively correlated with the intron GC content (Qiu et al. 2011). Genes with short intron lengths expressed more frequently than genes with long intron length, which were expressed at low levels in both *Caenorhabditis elegans* and *Homo sapiens* (Castillo-Davis et al. 2002).

It has been reported that codon usage bias, observed in genes, responds to different stressors (Quax et al. 2015). The suppression of circadian regulation at low temperature is caused by oscillator genes associated with the circadian clock, which contain rare codons (Xu et al. 2013). In yeast, stress can induce specific tRNA-modifying enzymes, either by DNA-damaging compounds or oxidative stress, which change codon usage (Begley et al. 2007; Chan et al. 2012). However, genes that respond to environmental adaption in *Ginkgo biloba* preferentially had codons that ended with G or C (He et al. 2016). In rice, compared with genes with low expression, genes with high expression had higher GC content, lower ENC (effective number of codon) value, different optimal codons and a bias for codons ending with GC under drought stress (Mohasses et al. 2020). Mohasses et al. (2020) proposed that codon optimization can increase gene expression under drought stress in rice.

With the development of genetic engineering, breeders using transgenic methods expect to increase both crop yield as well as biotic and abiotic resistance. Although these methods can shorten the time to obtain an aspired phenotype, contradiction exists among codon usage patterns between exogenous genes and the host genome. In this context, the expression of exogenous genes can be affected by codon usage. Several reasons may explain this result. The use of rare codons, but not optimal codons, may affect both the speed and accuracy of translation (Akashi 1994; Chaney and Clark 2015; Chu et al. 2014). In addition, the use of rare codons may cause protein misfolding and aggregation, resulting in decreased protein activity (Chaney and Clark 2015; Mitra et al. 2016). Codon optimization can help overcome the above-mentioned problems (Gustafsson et al. 2004; Quax et al. 2015). The first task is to identify the codon usage pattern in exogenous genes and host genomes. In summary, the disclosure of codon usage patterns at a genome-wide level can be considered as the basis for transgenic research.

Many studies on codon-bias and CDS architecture have been conducted using model species such as *Arabidopsis*, *Medicago truncatula*, *Populus* and rice (Ingvarsson 2007; Liu et al. 2015; Morton and Wright 2007; Song et al. 2018b). However, the codon usage patterns under normal growth and stress conditions have not been compared in *Arachis duranensis*, which is an ancestral species of the cultivated peanut, as well as an oil and protein crop (Bertioli et al. 2011; Kochert et al. 1996; Ramos et al. 2006; Seijo et al. 2007, 2004). The recent availability of the *A. duranensis* genome, its comprehensive tissue specific transcriptome characterization and transcriptome response to nematode and drought stresses has enabled the systematic analysis of codon usage patterns in this species (Bertioli et al. 2016; Clevenger et al. 2016; Dash et al. 2016). Using available genome sequences and RNA-seq datasets, this study investigated codon usage patterns and analyzed the relationships between the gene expression level and CDS architecture in *A. duranensis*. The CDS architecture includes the frequency of the optimal codon (Fop), polypeptide length and GC contents at the first (GC1), second (GC2) and third (GC3) codon positions.

Materials and methods

Sequence retrieval

To evaluate the codon usage patterns in *A. duranensis*, the CDSs of *A. duranensis* were obtained from PeanutBase (http://peanutbase.org/download) (Bertioli et al. 2016; Dash et al. 2016). To avoid biased analysis as a result by using partial sequences, the sequences were selected based on the following evaluation criteria (Song et al. 2017a, 2016): (1) CDS starting with ATG and ending with TAA, TAG, or TGA and (2) CDS without premature termination or ambiguous codons.

To assess the relationship between CDS architecture and gene expression l, the RNA-seq datasets were downloaded from the PeanutBase (https://peanutbase.org/external). The RNA-seq datasets were obtained for normal growth condition as well as in response to nematode and drought stresses. The RNA-seq datasets of the *A. hypogaea* cv. Tifrunner have been published before (Clevenger et al. 2016). The relevant RNA-seq data of root tissue were downloaded from PeanutBase. RNA-seq assembly used *A. duranensis* (AA genome) and *A. ipaensis* (BB genome) as reference genomes (Clevenger et al. 2016). The raw reads were obtained by a Illumina Hiseq 2500. The fragments per kilobase per million reads mapped (FPKM) were calculated using RSEM (Li and Dewey 2011).

Although a number of changes occurred on orthologous structures between the cultivated peanut and two wild forms of the peanut, high similarities were detected. The modal divergence between *A. duranensis* and subgenome A of *A. hypogaea* cv. Tifrunner was about 2.5 differences per 1,000 bp (Bertioli et al. 2020). The ratio between *A. ipaenesis* and subgenome B of *A. hypogaea* cv. Tifrunner was 2 differences per 10,000 bp (Bertioli et al. 2020). In addition, more than 98% DNA identity was identified between corresponding genes of *A. duranensis* and *A. ipaenesis* (Bertioli et al. 2019). In summary, these results indicated that gene expression levels are similar between *A. duranensis* (AA genome) and subgenome A and *A. ipaensis* (BB genome) and subgenome B under normal growth conditions because of similar gene sequences. The homoeolog expression result showed that *A. hypogaea* cv. Tifrunner had homoeolog expression balance in most tissues except for reproductive tissues, which showed a slightly stronger bias for subgenome B than for subgenome A (Bertioli et al. 2019). This result also indicated that homoeologs with similar gene structure had similar expression under normal growth conditions. In this study, the expression level of subgenome A from *A. hypogaea* cv. Tifrunner was used to represent corresponding gene expressions in *A. duranensis* under normal growth conditions. The gene expression level was transferred using Log_2 (FPKM) as standard expression level. The transform can decrease different expression levels between subgenome A and genome A (*A. duranensis*). The Log_2 (FPKM) values were assumed to represent the gene expression levels in root tissues under normal growth conditions in *A. duranensis*.

The differentially expressed genes (DEGs) in *A. duranensis* root tissue under drought and nematode stresses have been published before (https://peanutbase.org/gene_expression/atlas_drought and https://peanutbase.org/gene_expression/atlas_nematode), respectively (Brasileiro et al. 2015; Dash et al. 2016; Guimarães et al. 2015). All RNA-seq datasets were developed using Hi-seq 2000 (Brasileiro et al. 2015; Guimarães et al. 2015). Differential gene expression was obtained by a stress versus control comparison under drought treatment (Guimarães et al. 2015) and differential gene expression was estimated by a stress (3, 6 and 9 days after treatment) versus control comparison under nematode treatment (Brasileiro et al. 2015). These two papers were published in 2015, but the *A. duranensis* genome sequence was not available before 2016. These RNA-seq datasets were assembled using a de novo method before 2016. The authors re-assembled the RNA-seq using the *A. duranensis* genome as reference after 2016. The differential gene expression (Log_2 (FoldChange)) was uploaded on the PeanutBase. If the value of $\log_2$-FoldChange exceeded 2 or remained below -2 and if the adjusted-*p*-value (FDR) was less than 0.05, a gene was classified as differentially expressed using the edgeR package (Anders and Huber 2010).

Calculation of the codon index

To compare CDS architectures under normal growth condition as well as in response to nematode and drought stresses, the frequency of the optimal codon (Fop), polypeptide length, relative synonymous codon usage (RSCU), GC contents at the first (GC1), second (GC2) and third (GC3) codon positions were estimated as variables. Codon W (version 1.4, http://codonw.sourceforge.net) was used to calculate Fop, polypeptide length and RSCU. The GC1, GC2 and GC3 codon positions were estimated with an in-house Perl script. Fop was defined as the codon with the highest number of tRNA genes for its anticodon, among its synonymous codons (Lavner and Kotlar 2005). The RSCU value for a codon was defined as the observed frequency of the codon, divided by the expected frequency under the assumption of equal usage of the synonymous codons for an amino acid (Sharp and Li 1986; Sharp et al. 1986). The formulas for calculating Fop and RSCU are listed in the following (Lavner and Kotlar 2005; Sharp and Li 1986):

$$\mathrm{Fop}(g) = \frac{1}{N}\sum_{i}\mathrm{syn}(i)n_i(g) \qquad (1)$$

where $n_i(g)$ represents the count of codon i in gene g, N represents the total number of codons in g and $\mathrm{syn}(i)$ represents the degeneracy of the amino acid encoded by i.

$$\mathrm{RSCU}_{ij} = \frac{X_{ij}}{\frac{1}{n_i}}\sum_{j=1}^{n_i} X_{ij} \qquad (2)$$

where X_{ij} represents the number of occurrences of the jth codon for the i-th amino acid and n_i represents the number of alternative codons for the ith amino acid.

Correlation analysis

Spearman correlation was used to assess the correlation between two independent variables. In this study, these variables included gene expression level, differential gene expression, Fop, polypeptide length, GC1, GC2 and GC3. Spearman correlation analysis was used to assess the CDS architecture and how it affects gene expression level and differential gene expression. One-way ANOVA was used to analyze significance in the correlation analyses. For all statistical tests, a *p*-value below 0.05 was considered to indicate a significant difference. The JMP program was used to execute the spearman correlation analysis and ANOVA tests. Previous studies showed that the correlation coefficient is low between gene expression levels and CDS architecture (Ingvarsson 2007; Song et al. 2018c; Whittle and Extavour 2015). In this study, it was assumed that no correlation exists if the correlation coefficient was less than 0.1, despite significance of the correlation, based on previous studies (Ingvarsson 2007; Song et al. 2018c; Whittle and Extavour 2015).

Results

Similar CDS architecture in differentially expressed genes or non-differentially expressed genes under nematode and drought stresses in *A. duranensis*

A total of 32,725 *A. duranensis* CDSs were analyzed under normal growth conditions. A total of 528 and 1,113 CDS were identified as DEGs under nematode and drought stresses, respectively. The remaining 32,197 and 31,612 CDSs were identified as non-differentially expressed genes (NDEGs) under nematode and drought stresses, respectively. The average GC1 was 49.29%, followed by GC3 at 42.03% and GC2 at 40.07% in *A. duranensis* CDSs under normal growth conditions. The average GC content was 43.79% in *A. duranensis* CDSs under normal growth conditions. Therefore, the average AT content (56.21%) exceeded the average GC content in CDSs under normal growth conditions. Similar patterns were found for DEGs and NDEGs under normal growth conditions as well as nematode and drought stress conditions. The average GC contents of DEGs were 43.39% and 44.23% and the average GC contents of NDEGs were 43.80% and 43.78% under nematode and drought stress, respectively (Tables 1 and 2).

The average Fop value was almost identical between DEGs and NDEGs under nematode and drought stresses. The average polypeptide length differed for DEGs, but was almost identical for NDEGs under nematode and drought stresses. The average Fop and average polypeptide length were 0.38 and 356, respectively, in *A. duranensis* CDSs under the normal growth condition. In DEGs, the average Fop values were 0.38 and 0.39 and the polypeptide lengths were 356 and 450 under nematode and drought stresses, respectively (Fop: Mann–Whitney test, $p > 0.05$, polypeptide length: Mann–Whitney test, $p < 0.01$). In NDEGs, the average Fop were 0.38 and 0.38 and the polypeptide lengths were 369 and 366 under nematode and drought stresses, respectively (Mann–Whitney test, $p < 0.05$).

A RSCU value above 1 (and below 1) indicates that the codon usage frequency exceeded (or remained below) the expectation, respectively (Sharp and Li 1987). The results of the present study identified a similar RSCU pattern for all genes under normal growth conditions and both DEGs and NDEGs under nematode and drought stresses. Twenty-six codons with a RSCU value above 1 were found and the RSCU values of the remaining 35 codons were below 1 (Fig. 1). In addition, the 26 codons with high RSCU values could also be distinguished from other codons based on their sequence composition. These codons preferentially ended with A or T. The 35 codons with low RSCU values ended more often with C or G (Fig. 1).

Comparison of CDS architecture and gene expression between DEGs and NDEGs under nematode and drought stresses in *A. duranensis*

Despite the similarity of the CDS architectures between DEGs or NDEGs, differences were found between DEGs and NDEGs under nematode and drought stresses. Under nematode stress, GC1 of DEGs was lower than that of NDEGs (Table 1, Mann–Whitney test, $p < 0.01$). Under drought stress, both polypeptide length and GC3 of DEGs exceeded those of NDEGs (Table 2, Mann–Whitney test, $p < 0.01$). In addition, under drought stress GC1 was lower in DEGs than in NDEGs (Table 2, Mann–Whitney test, $p < 0.05$). In summary, a consistent change was found in the GC content between DEGs and NDEGs, but the change in other CDS architectures was inconsistent between DEGs and NDEGs under nematode and drought stresses. These results indicated that nematode-stress responsive genes had low expression under normal growth conditions.

Table 1 Comparison of CDS architecture between differentially expressed genes and non-differentially expressed genes in *Arachis duranensis* under nematode stress

	[a]DEG	[a]NDEG	*p* value
Fop	0.3836 ± 0.0484	0.3847 ± 0.0493	4.2436
Polypeptide length	356 ± 270	369 ± 320	0.7498
GC1	48.4302 ± 5.2275%	49.3023 ± 5.7875%	6.04E-6
GC2	39.9285 ± 6.2592%	40.0688 ± 6.3033%	0.1645
GC3	41.8068 ± 8.8414%	42.0335 ± 8.6655%	0.4084
GC	43.3885 ±5.1609%	43.8015 ± 5.1536%	0.0010

DEG: Differentially expressed gene; NDEG: Non-differentially expressed gene; Fop: frequency of the optimal codon; GC1: GC content at the first codon position; GC2: GC content at the second codon position; GC3: GC content at the third codon position; GC: the average of GC at three codon positions. Mann–Whitney test was used in statistical analysis

[a]Mean ± SD

Table 2 Comparison of CDS architecture between differentially expressed genes and non-differentially expressed genes in *Arachis duranensis* under drought stress

	[a]DEG	[a]NDEG	*p* value
Fop	0.3870 ± 0.0454	0.3846 ± 0.0494	0.1254
Polypeptide length	450 ± 351	366 ± 317	0.0000
GC1	49.0254 ± 4.8367%	49.2974 ± 5.8102%	0.0201
GC2	40.0496 ± 5.1786%	40.0671 ± 6.3386%	0.5854
GC3	43.6221 ± 8.5536%	41.9738 ± 8.6671%	2.16E-8
GC	44.2324 ± 4.4549%	43.7795 ± 5.1762%	0.0157

DEG: Differentially expressed gene; NDEG: Non-differentially expressed gene; Fop: frequency of the optimal codon; GC1: GC content at the first codon position; GC2: GC content at the second codon position; GC3: GC content at the third codon position; GC: the average of GC at three codon positions. Mann–Whitney test was used in statistical analysis

[a]Mean ± SD

To compare the expression levels between DEGs and NDEGs under nematode and drought stresses, the expression levels of DEGs and NDEGs were estimated in the root tissue under the normal growth conditions. Different gene expression patterns were detected between DEGs and NDEGs under nematode and drought stresses. Under nematode stress, the average expression level of NDEGs (0.33) exceeded that of DEGs (-0.89, Fig. 2, Mann–Whitney test, $p < 0.01$). These results indicated that the genes that respond to nematode infection had a low expression under normal growth conditions. Under drought stress, the average expression level of NDEGs (0.26) was lower than that of DEGs (1.57, Fig. 2, Mann–Whitney test, $p < 0.01$). These results indicated that drought-stress responsive genes were high expression under normal growth conditions.

Different correlations between gene expression levels under normal growth condition and CDS architectures in DEGs in response to nematode and drought stresses in *A. duranensis*

Correlation analysis was conducted between gene expression level and CDS architectures under normal growth conditions. The gene expression level of root tissue correlated positively with Fop, GC1 and GC3 (Table 3). This indicated that highly expressed genes had higher codon usage bias and preferentially used CDSs with higher GC1 and GC3 in root tissue. In addition, the same correlation analysis was also used to assess DEGs and NDEGs under nematode and drought stresses. The gene expression levels of DEGs and NDEGs were estimated in the root tissue under the normal growth condition. The expression levels of NDEGs correlated positively with Fop, GC1 and GC3 under nematode and drought stresses (Table 3). However, the expression levels of DEGs did not correlate with the CDS architecture under nematode stress and correlated positively with Fop, GC1, GC2 and GC3 under drought stress (Table 3). These results indicated that the expression levels of DEGs were affected by different CDS architectures under drought stress. Furthermore, the correlation between differential expression levels and CDS architecture was assessed. No correlation was found between differential expression level and CDS architecture under nematode and drought stresses (Table 3). This indicated that the CDS architecture did not affect the differential expression level in *A. duranensis*.

Discussion

This study investigated the response of the wild peanut *A. duranensis* to drought and nematode stresses, using the RNA-seq data and the sequenced *A. duranensis* genome as reference sequences. The analysis focused to find the influence of the CDS architecture on the gene expression level, in response to drought and nematode stresses. Many studies have reported codon usage biases in plants, animals, microorganisms and viruses (Behura and Severson 2012; Camiolo et al. 2015; Hershberg and Petrov 2009; Jia et al. 2015; Jiang et al. 2008; Li et al. 2016a, 2016b; RoyChoudhury and Mukherjee 2010; Song et al. 2017b). Natural selection and mutation pressure are typically considered as major causes of codon usage bias (Hershberg and Petrov 2008; Song et al. 2018b). In general, gene expression levels may increase by natural selection of codons, while codon bias exists because of non-random mutation pressures (Hershberg and Petrov 2008). The findings of the present study suggest natural selection as a major force for the codon usage bias in *A. duranensis* because highly expressed genes preferentially used optimal codons (represented by Fop). CDSs with GC-rich content tended to be highly expressed in *A. duranensis*. Such GC-rich gene sequences might waste much energy during their duplication and translation. However, Yang (2009) demonstrated that the time–cost hypothesis (rather than the energy-cost hypothesis) could provide a better

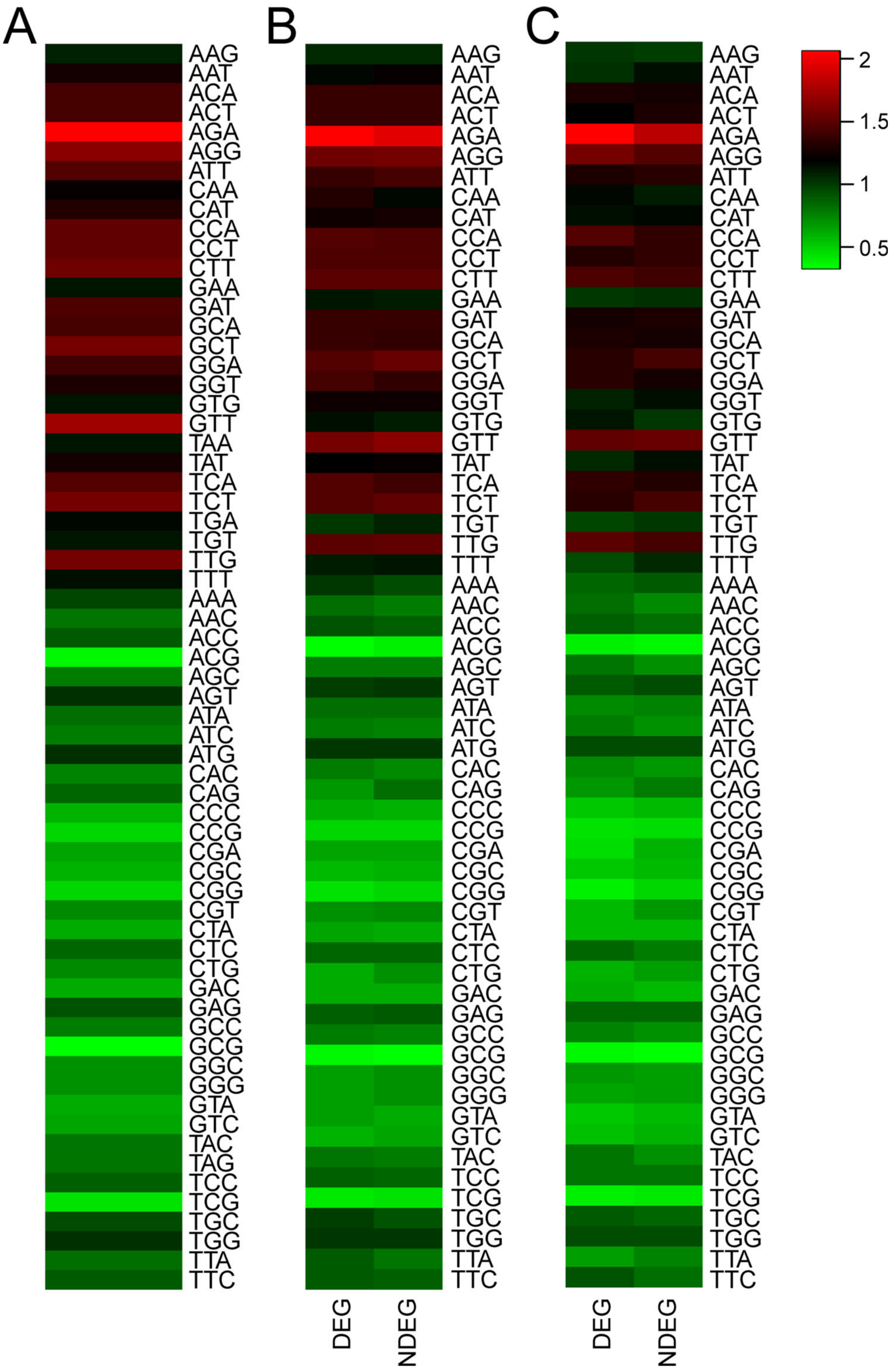

Fig. 1 Codon usage frequency based on relative synonymous codon usage values in *Arachis duranensis*. **A**: Codon usage frequency in all CDSs. **B**: Codon usage frequency between differentially expressed genes (DEG) and non-differentially expressed genes (NDEG) in response to nematode stress. **C**: Codon usage frequency between differentially expressed genes and non-differentially expressed genes in response to drought stress. The scale represents RSCU

interpretation of highly expressed genes. In addition, natural selection decreases the cost of biosynthesis and increases the speed of translation (Brandis and Hughes 2016; Ellegren and Parsch 2007).

This study showed that the GC contents at three codon sites follow the same trend (GC1 > GC3 > GC2) and the AT content is higher than their GC content both under normal growth and stress conditions. A previous study in eudicots indicated that the GC1 content was higher than the GC2 content and the GC3 content was similar or higher than the GC2 content (Li et al. 2016a). In addition, many studies reported that the AT content exceeded the GC

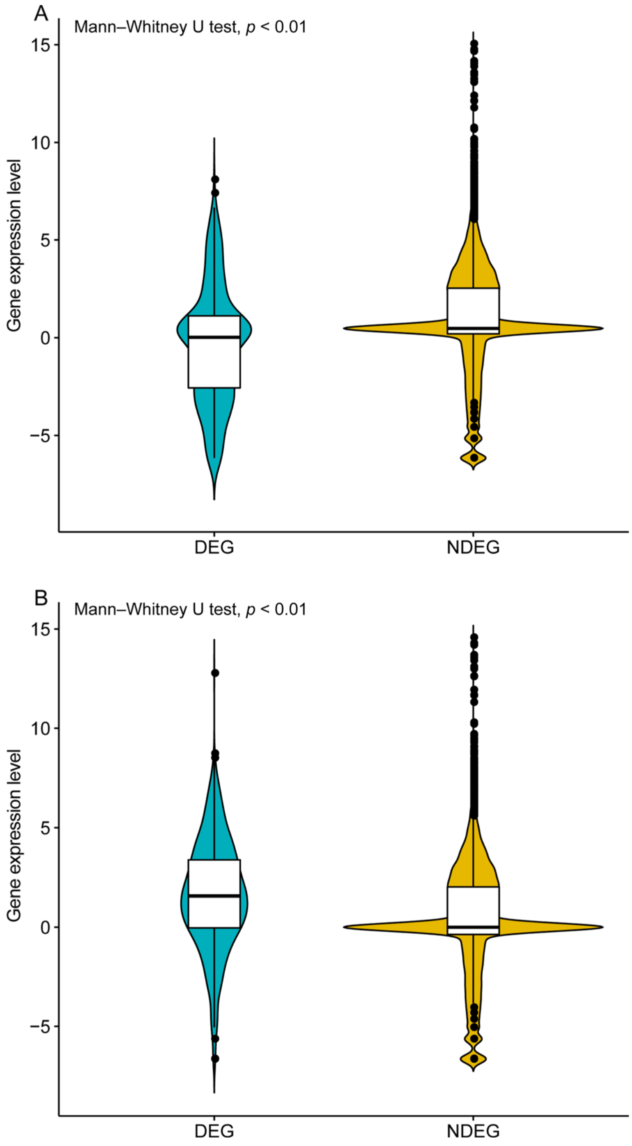

Fig. 2 Comparison of gene expression levels between differentially expressed genes (DEGs) and non-differentially expressed genes (NDEGs) in *Arachis duranensis*. **A**: Comparison of gene expression levels between DEGs and NDEGs in response to nematode stress. **B**: Comparison of gene expression levels between DEGs and NDEGs in response to drought stress

content in eudicots, while opposite patterns were found in Poaceae (Glémin et al. 2014; Kawabe and Miyashita 2003; Li et al. 2016a; Singh et al. 2016).

Previous studies have demonstrated different CDS architectures or codon usage biases between DEGs and NDEGs (Quax et al. 2015). For example, in *G. biloba*, He et al. (2016) reported that genes that responded to environmental adaption used codons ending with G or C. The present study showed that the GC1 content of NDEGs was higher than that of DEGs in *A. duranensis* under nematode and drought stresses. In addition, the RSCU values showed that high frequency codons preferentially ended with A or T in *A. duranensis* CDSs under nematode and drought stresses. Recently, Sidorenko et al. (2017) demonstrated that high GC content in *Arabidopsis* CDSs positively impacted transgene expression by decreasing the accumulation of small RNA and DNA methylation. In *Oryza sativa*, the GC3 content was negatively correlated with gene methylation (Elhaik et al. 2014). In the present study, in response to drought stress, the gene expression of DEGs was positively correlated with the GC content. In summary, these results showed that a reasonable increase in GC content, but not in GC1 content, may contribute to transgene expression when *A. duranensis* drought resistance genes are translated into other plants.

The gene expression pattern between DEGs and NDEGs differs in *A. duranensis*. The gene expression of DEGs in response to nematode stress is lower than that of NDEGs under normal growth condition. However, the gene expression of DEGs in response to drought stress is higher than that of NDEGs under normal growth condition. In a previous study, we found that the LRR-containing genes had low expression levels since they often acted as receptors of pathogen elicitors, which in turn triggered a cascade of defense responses that culminated in plant resistance (Song et al. 2018a). In this context, it is important to consider the environment from which *A. duranensis* originates: regions with low rainfall (699 mm/year) and with an average rainfall of approximately 1,050 mm/year (Leal-Bertioli et al. 2012). Therefore, *A. duranensis* has adapted to an area with erratic rainfall. In addition, the higher number and expression levels of *A. duranensis* genes in response to drought stress (in comparison to nematode infection) indicates the severe impact of drought on plants. Pathogen attack tends to trigger more specific and time-restricted responses.

The debate about the correlation between polypeptide length and gene expression level is ongoing. In this study, the gene expression levels did not correlate with polypeptide length for nematode- and drought-related DEGs and in all sequences under normal growth conditions. Previous studies identified different correlations between gene expression and sequence length. For example, no correlation was found between protein length and gene expression levels in *Arabidopsis* (Duret and Mouchiroud 1999). Highly expressed genes tend to encode short proteins in both *Populus* and rice (Ingvarsson 2007; Yang 2009). However, in both *Tribolium castaneum* and *Picea*, long protein-encoding genes were highly expressed (De La Torre et al. 2015; Williford and Demuth 2012).

The results of this study showed that the levels of differential gene expression did not correlate with the CDS architecture. However, the differential gene expression patterns of the non-Toll interleukin receptor, nucleotide-

Table 3 Correlation between gene expression levels and CDS architecture in *Arachis duranensis* roots

	Fop	Polypeptide length	GC1	GC2	GC3
Gene expression of root under normal growth condition	0.1469**	0.0792	0.1989**	0.0988	0.1188**
Expression of DEG under nematode stress	0.0537	− 0.0999	0.0794	0.0884	0.0914
Expression of NDEG under nematode stress	0.1487**	0.0819	0.2002**	0.099	0.1193**
DGE under nematode stress	0.0291	− 0.0923	0.0471	0.0551	− 0.0297
Expression of DEG under drought stress	0.2422**	− 0.0677	0.1897**	0.166**	0.2349**
Expression of NDEG under drought stress	0.1435**	0.0811	0.2009**	0.0974	0.1121**
DGE under drought stress	0.0138	0.0167	− 0.023	− 0.0657	0.006

DEG: Differentially expressed gene; NDEG: Non-differentially expressed gene; DGE: Differential gene expression; GC1: GC content at the first codon position; GC2: GC content at the second codon position; GC3: GC content at the third codon position

**Indicates significance at $p < 0.01$

binding site and leucine-rich repeat (nTNL) genes (all of which are involved in the response to pathogens (Dangl and Jones 2001)) correlated with the number of introns and the GC content in *Arabidopsis*, *Medicago*, soybean, *Populus* and rice (Nepal et al. 2017). These results indicated no correlation between any gene's ability to respond to environmental stimuli and CDS architecture in *A. duranensis*.

Conclusion

In the present study, the CDS architectures of DEGs and NDEGs were compared and the relationship between CDS architecture and gene expression levels was investigated. The GC1 content differed between DEGs and NDEGs under both drought and nematode stresses. No correlation was found between differential gene expression and CDS architecture, neither under nematode nor under drought stress. These results provide a theoretical foundation for transgene analysis. Codon optimization can be ignored when exogenous genes are transferred into the *A. duranensis* genome.

Acknowledgements The work was supported by the First Class Grassland Science Discipline Program of Shandong Province, China, the Natural Science Foundation of Shandong Province, China (ZR2019QC017, ZR2020MC172 and ZR2020QC185), Key Research and Development Program in Shandong Province (2019GNC106077), and Start-up Foundation for High Talents of Qingdao Agricultural University (No. 665/1120012, No. 663/1119038 and No. 663/1119040).

Compliance with ethical standards

Conflict of interest There is no conflict of interest.

References

Akashi H (1994) Synonymous codon usage in Drosophila melanogaster: nature selection and translation accuracy. Genetics 136(3):927–935

Akashi H (2001) Gene expression and molecular evolution. Curr Opin Genet Dev 11:660–666

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11(10):106–110

Arava Y (2003) Isolation of polysomal RNA for microarray analysis. Methods Mol Bio 224:79–87

Arunkumar R, Josephs EB, Williamson RJ, Wright SI (2013) Pollen-specific, but not sperm-specific, genes show stronger purifying selection and higher rates of positive selection than sporophytic genes in Capsella grandiflora. Mol Bio Evol 30(11):2475–2486

Begley U, Dyavaiah M, Patil A, Rooney JP, DiRenzo D, Young CM, Conklin DS, Zitomer RS, Begley TJ (2007) Trm9-catalyzed tRNA modifications link translation to the DNA damage response. Mol Cell 28:860–870

Behura SK, Severson DW (2012) Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. PLoS ONE 7(8):e43111

Bertioli DJ, Seijo G, Freitas FO, Valls JFM, Leal-Bertioli SCM, Moretzsohn MC (2011) An overview of peanut and its wild relatives. Plant Genet Resour Characteriz Utiliz 9(1):134–149

Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EKS, Liu X, Gao D, Clevenger J, Dash S, Ren L, Moretzsohn MC, Shirasawa K, Huang W, Vidigal B, Abernathy B, Chu Y, Burow MD, Varshney RK, Wang X, Zhang X, Barkley N, Guimarães PM, Isobe S, Guo B, Liao B, Stalker HT, Schmitz RJ, Scheffler BE, Leal-Bertioli SCM, Xun X, Jackson SA, Michelmore R, Ozias-Akins P (2016) The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. Nat Genet 48(4):438–446

Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G, Leal-Bertioli SCM, Ren L, Farmer AD, Pandey MK, Samoluk SS, Abernathy B, Agarwal G, Ballé-Taborda C, Cameron C, Campbell J, Chavarro C, Chitikineni A, Chu Y, Dash S, Eiaidouri M, Guo B, Huang W, Kim KD, Korani W, Lanciano S, Lui CG, Mirouze M, Moretzsohn MC, Pham M, Shin JH, Shirasawa K, Sinharoy S, Sreedasyam A, Weeks NT, Zhang X, Zheng Z, Sun Z, Froenicke L, Aiden EL, Michelmore R, Varshney RK, Holbrook CC, Cannon EKS, Scheffler BE, Grimwood J, Ozias-Akins P, Cannon SB, Jackson SA, Schmutz

J (2019) The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nat Genet 51:877–884

Bertioli DJ, Abernathy B, Seijo G, Clevenger J, Cannon SB (2020) Evaluating two different models of peanut's origin. Nat Genet 52:557–559

Brandis G, Hughes D (2016) The selective advantages of synonymous codon usage bias in Salmonella. PLoS Genet 12(3):e1005926

Brasileiro ACM, Morgante CV, Araujo ACG, Leal-Bertioli SCM, Silva AK, Martins ACQ, Vinson CC, Santos CMR, Bonfim O, Togawa RC, Saraiva MAP, Bertioli DJ, Guimaraes PM (2015) Transcriptome profiling of wild Arachis from water-limited environments uncovers drought tolerance candidate genes. Plant Mol Biol Rep 33:1876–1892

Camiolo S, Melito S, Porceddu A (2015) New insights into the interplay between codon bias determinants in plants. DNA Res 5:1–9

Castillo-Davis C, Mekhedov S, Hartl D, Koonin E, Kondrashov F (2002) Selection for short introns in highly expressed genes. Nat Genet 31:415–418

Chan CT, Pang YL, Deng W, Babu IR, Dyavaiah M, Begley TJ, Dedon PC (2012) Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. Nat Commun 3:937

Chaney J, Clark PL (2015) Roles for synonymous codon usage in protein biogenesis. Ann Rev Biophys 44:143–166

Chu D, Kazana E, Bellanger N, Singh T, Tuite MF, von der Haar T (2014) Translation elongation can control translation initiation on eukaryotic mRNA. EMBO J 33:21–34

Clevenger J, Chu Y, Scheffler B, Ozias-Akins P (2016) A developmental transcriptome map for allotetraploid Arachis hypogaea. Front Plant Sci 7:1446

Dangl JL, Jones JD (2001) Plant pathogens and integrated defence responses to infection. Nature 411(6839):826–833

Dash S, Cannon EKS, Kalberer SR, Farmer AD, Cannon SB (2016) PeanutBase and other bioinformatic resources for peanut. In: Stalker HT, Wilson RF (eds) Peanuts Genetics Processing and Utilization. AOCS Press, USA, pp 241–252

De La Torre AR, Lin YC, Van de Peer Y, Ingvarsson PK (2015) Genome-wide analysis reveals diverged pattern of codon bias, gene expression and rates of sequence evolution in Picea gene families. Genome Biol Evol 7(4):1002–1015

Duret L (2000) tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet 16:287–289

Duret L, Mouchiroud D (1999) Expression pattern and surprisingly, gene length shape codon usage in Caenorhabdits, Drosophila and Arabidopsis. Proc Natl Acad Sci USA 96(8):4482–4487

Elhaik E, Pellegrini M, Tatarinova TV (2014) Gene expression and nucleotide composition are associated with genic methylation level in Oryza sativa. BMC Bioinformatics 15:23

Ellegren H, Parsch J (2007) The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet 8:689–698

Glémin S, Clément Y, David J, Ressayre A (2014) GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. Trends Genet 30(7):263–270

Guimarães PM, Guimaraes LA, Morgante CV, Silva OB Jr, Araujo ACG, Martins ACQ, Saraiva MAP, Oliveira TN, Togawa RC, Leal-Bertioli SCM, Bertioli DJ, Brasileiro ACM (2015) Root transcriptome analysis of wild peanut reveals candidate genes for nematode resistance. PLoS ONE 10(10):e0140937

Gustafsson C, Govindarajan S, Minshull J (2004) Codon bias and heterologous protein expression. Trends Biochem Sci 22:346–353

He B, Dong H, Jiang C, Cao F, Tao S, Xu L (2016) Analysis of codon usage patterns in Ginkgo biloba reveals codon usage tendency from A/U-ending to G/C ending. Sci Rep 6:35927

Hershberg R, Petrov DA (2008) Selection on codon bias. Annu Rev Genet 42:287–299

Hershberg R, Petrov DA (2009) General rules for optimal codon choice. PLoS Genet 5(7):e1000556

Ingvarsson PK (2007) Gene expression and protein length influence codon usage and rates of sequence evolution in Populus tremula. Mol Bio Evol 24(3):836–844

Jia X, Liu S, Zheng H, Li B, Qi Q, Wei L, Zhao T, He J, Sun J (2015) Non-uniqueness of factors constraint on the codon usage in Bombyx mori. BMC Genomics 16:356

Jiang Y, Deng F, Wang H, Hu Z (2008) An extensive analysis on the global codon usage pattern of baculoviruses. Arch Virol 153:2273–2282

Kawabe A, Miyashita NT (2003) Patterns of codon usage bias in three dicot and four monocot plant species. Genes Genet Syst 78(5):343–352

Kochert G, Stalker H, Gimenes M, Galgaro M, Lopes C, Moore K (1996) RFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut, Arachis hypogaea (Leguminosae). Am J Bot 83(10):1282–1291

Lavner Y, Kotlar D (2005) Codon bias as a factor in regulating expression via translation rate in the human genome. Gene 345:127–138

Leal-Bertioli SCM, Bertioli DJ, Guimarães PM, Pereira TD, Galhardo I, Silva JP, Brasileiro ACM, Oliveira RS, Silva PÍT, Vadez V, Araujo ACG (2012) The effect tetraploidization of wild Arachis on leaf morphology and other drought-related traits. Environ Exp Bot 84:17–24

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinfo 12:323

Li N, Li Y, Zheng C, Huang J, Zhang S (2016a) Genome-wide comparative analysis of the codon usage patterns in plants. Genes Genom 38(8):723–731

Li X, Song H, Kuang Y, Chen S, Tian P, Li C, Nan Z (2016b) Genome-wide analysis of codon usage bias in Epichloë festucae. Int J Mol Sci 17(7):E1138

Liu Q, Hu H, Wang H (2015) Mutational bias is the driving force for shaping the synonymous codon usage pattern of alternatively spliced in rice (Oryza sativa L). Mol Genet Genomics 290(2):649–660

Maurer-Alcalá XX, Katz LA (2016) Nuclear architecture and patterns of molecular evolution are correlated in the Chilodonella uncinata. Genome Biol Evol 8(6):1634–1642

Mitra S, Ray SK, Banerjee R (2016) Synonymous codon influencing gene expression. Resear Rep Biochem 6:57–65

Mohasses FC, Solouki M, Ghareyazie B, Fahmideh L, Mohsenpour M (2020) Correlation between gene expression levels under drought stress and synonymous codon usage in rice plant by in-silico study. PLoS ONE 15(8):e0237334

Morton BR, Wright SI (2007) Selective constraints on codon usage of nuclear genes from Arabidopsis thaliana. Mol Bio Evol 24:122–129

Nepal MP, Andersen EJ, Neupane S, Benson BV (2017) Comparative genomics of non-TNL disease resistance genes from six plant species. Genes 8:249

Qiu S, Bergero R, Zeng K, Charlesworth D (2011) Patterns of codon usage bias in Silene latifolia. Mol Bio Evol 28(1):771–780

Quax TEF, Claassens NJ, Söll D, van der Oost J (2015) Codon bias as a means to fine-tune gene expression. Mol Cell 59:149–161

Ramos M, Fleming G, Chu Y, Akiyama Y, Gallo M, Ozias-Akins P (2006) Chromosomal and phylogenetic context for conglutin genes in Arachis based on genomic sequence. Mol Genet Genomics 275(6):578–592

RoyChoudhury S, Mukherjee D (2010) A detailed comparative analysis on the overall codon usage pattern in herpesviruses. Virus Res 148(1–2):31–43

Seijo J, Lavia G, Fernandez A, Krapovickas A, Ducasse D, Moscone E (2004) Physical mapping of the 5S and 18S–25S rRNA genes by FISH as evidence that Arachis duranensis and A ipaënsis are the wild diploid progenitors of A hypogaea (Leguminosae). Am J Bot 91(9):1294–1303

Seijo G, Lavia GI, Fernandez A, Krapovickas A, Ducasse DA, Bertioli DJ, Moscone EA (2007) Genomic relationships between the cultivated peanut (Arachis hypogaea, Leguminosae) and its close relatives revealed by double GISH. Am J Bot 94(12):1963–1971

Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28–38

Sharp PM, Li WH (1987) The codon adaption index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

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

Sidorenko LV, Lee T, Woosley A, Moskal WA, Bevan SA, Ann Owens Merlo P, Walsh TA, Wang X, Weaver S, Glancy TP, Wang P, Yang X, Sriram S, Meyers BC (2017) GC-rich coding sequences reduce transposon-like, small RNA-mediated transgene silencing. Nat Plants 3:875–884

Singh R, Ming R, Yu Q (2016) Comparative analysis of GC content variations in plant genomes. Trop Plant Biol 9(3):136–149

Song H, Wang P, Hou L, Zhao S, Zhao C, Xia H, Li P, Zhang Y, Bian X, Wang X (2016) Global analysis of WRKY genes and their response to dehydration and salt stress in soybean. Front Plant Sci 7:9

Song H, Gao H, Liu J, Tian P, Nan Z (2017a) Comprehensive analysis of correlations among codon usage bias, gene expression, and substitution rate in Arachis duranensis and Arachis ipaënsis orthologs. Sci Rep 7:14853

Song H, Liu J, Song Q, Zhang Q, Tian P, Nan Z (2017b) Comprehensive analysis of codon usage bias in seven Epichloë species and their peramine-coding genes. Front Microbiol 8:1419

Song H, Zhang Q, Tian P, Nan Z (2017c) Differential evolutionary patterns and expression levels between sex-specific and somatic tissue-specific genes in peanut. Sci Rep 7:9016

Song H, Guo Z, Chen T, Sun J, Yang G (2018a) Genome-wide identification of LRR-containing sequences and the response of these sequences to nematode infection in Arachis duranensis. BMC Plant Biol 18:279

Song H, Liu J, Chen T, Nan Z (2018b) Synonymous codon usage pattern in model legume Medicago truncatula. J Integr Agr 17(9):2074–2081

Song H, Sun J, Yang G (2018c) Comparative analysis of selection mode reveals different evolutionary rate and expression pattern in Arachis duranensis and Arachis ipaënsis duplicated genes. Plant Mol Biol 98(4–5):349–361

Whittle CA, Extavour CG (2015) Codon and amino acid usage are shaped by selection across divergent model organisms of the Pancrustacea. G3 Gene Genom Genet 5:2307–2321

Williford A, Demuth JP (2012) Gene expression levels are correlated with synonymous codon usage, amino acid composition, and gene architecture in the red flour beetle. Tribolium Castaneum Mol Bio Evol 29(12):3755–3766

Xu Y, Ma P, Shah P, Rokas A, Liu Y, Johnson CH (2013) Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. Nature 495:116–120

Yang H (2009) In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure. Biology Direct 4:45