Comparative analysis of codon usage patterns in chloroplast genomes of five Miscanthus species and related species

Jiajing Sheng [1] , **Xuan She** [2] , **Xiaoyu Liu** [1] , **Jia Wang** [3] , **Zhongli Hu** Corresp. [2]

[1] Nantong University, School of Life Sciences, Jiangsu Key Laboratory of Neuroregeneration, Co-innovation Center of Neuroregeneration, Nantong, China, □□
[2] Wuhan University, Wuhan, China, □□
[3] Anhui University of Science and Technology, Huainan, China

Corresponding Author: Zhongli Hu
Email address: huzhongli@whu.edu.cn

Miscanthus is not only a perennial fiber biomass crop, but also valuable breeding resource for its low-nutrient requirements, photosynthetic efficiency and strong adaptability to environment. In the present study, the codon usage patterns of five different *Miscanthus* plants and other two related species were systematically analyzed. The results indicated that the cp genomes of the seven representative species were preference to A/T bases and A/T-ending codons. In addition, 21 common high-frequency codons and 4-11 optimal codons were detected in the seven chloroplast genomes. The results of ENc-plot, PR2-plot and neutrality analysis revealed the codon usage patterns of the seven chloroplast genomes are influenced by multiple factors, in which nature selection is the main influencing factor. Comparative analysis of the codon usage frequencies between the seven representative species and four model organisms suggested that *Arabidopsis thaliana*, *Populus trichocarpa* and *Saccharomyces cerevisiae* could be considered as preferential appropriate exogenous expression receptors. These results might not only provide important reference information for evolutionary analysis, but also shed light on the way to improve the expression efficiency of exogenous gene in transgenic research based on codon optimization.

Comparative analysis of codon usage patterns in cp genomes of five Miscanthus species and related species

Jiajing Sheng12, Xuan She2, Xiaoyu Liu1, Jia Wang3, Zhongli Hu2*

1 School of Life Sciences, Jiangsu Key Laboratory of Neuroregeneration, Co-innovation Center of Neuroregeneration, Nantong University, Nantong, 226019, PR China

2 State Key Laboratory of Hybrid Rice, Lotus Engineering Research Center of Hubei Province, College of Life Sciences, Wuhan University, Wuhan, 430072, PR China

3 Key Laboratory of Industrial Dust Prevention and Control & Occupational Safety and Health of the Ministry of Education, Anhui University of Science and Technology, Huainan, 232001 China

Corresponding Author:
Zhongli Hu
State Key Laboratory of Hybrid Rice, Lotus Engineering Research Center of Hubei Province,
College of Life Sciences, Wuhan University, Wuhan, 430072, PR China
Email address: huzhongli@whu.edu.cn

Abstract

Miscanthus is not only a perennial fiber biomass crop, but also valuable breeding resource for its low-nutrient requirements, photosynthetic efficiency and strong adaptability to environment. In the present study, the codon usage patterns of five different *Miscanthus* plants and other two related species were systematically analyzed. The results indicated that the cp genomes of the seven representative species were preference to A/T bases and A/T-ending codons. In addition, 21 common high-frequency codons and 4-11 optimal codons were detected in the seven chloroplast genomes. The results of ENc-plot, PR2-plot and neutrality analysis revealed the codon usage patterns of the seven chloroplast genomes are influenced by multiple factors, in which nature selection is the main influencing factor. Comparative analysis of the codon usage frequencies between the seven representative species and four model organisms suggested that *Arabidopsis thaliana*, *Populus trichocarpa* and *Saccharomyces cerevisiae* could be considered as preferential appropriate exogenous expression receptors. These results might not only provide important reference information for evolutionary analysis, but also shed light on the way to improve the expression efficiency of exogenous gene in transgenic research based on codon optimization.

Introduction

Miscanthus Andersson (Poaceae) are C4 photosynthetic plants, which have been widely investigated as potential second-generation bio-energy crops (Barling et al. 2013). The genus Miscanthus includes approximately 20 species, which could be classified into Miscanthus clades and Triarrhena clades (Ge et al. 2017). China is the biological diversity center of *Miscanthus* species, of which *Miscanthus lutarioriparius L.Liou* (*M. lutarioriparius*), *Miscanthus sinensis Andersson* (*M. sinensis*), *Miscanthus sacchariflorus (Maxim.) Nakai* (*M. sacchariflorus*) and *Miscanthus floridulus (Lab.) Warb. ex Schum. et Laut* (*M. floridulus*) are the four most widely distributed species. In addition to being bioenergy plants, *Miscanthus* species possess extensive breeding values due to their extremely advantageous agricultural characteristics, such as high photosynthetic efficiency, cold tolerance and extensive environmental adaptation (Vermerris 2008). Currently, the research focus of *Miscanthus* species is to utilize it as promising genetic resource (Clark et al. 2015; Zhang et al. 2013). The more diverse genetic resources we have, the better we can comprehend the adaptation, evolution and utilization of these significant economic crops.

Chloroplasts (cp) are key plastids involved in multifunctional processes of plant cell (Jarvis & López-Juez 2013; Nielsen et al. 2016). Typically, cp genome possesses small size, conserved gene content and large copy numbers, which have been extensively used as valuable source for evolution analysis and plastid engineering (Amiryousefi et al. 2018; Ravi et al. 2008; Yan et al. 2019). The lacking of genomic resources of the *Miscanthus* species has hindered the adequate understanding of their diversity traits (Chae et al. 2014; Sheng et al. 2016) The low expression efficiency of the exogenous gene may limit the research progress on the functional studies of *Miscanthus* and their related species (Wu et al. 2021). Benefited from the rapid development of cp engineering, plasmid DNA have been transferred into the cp of a variety of plants, such as *Nicotiana tabacum*, *Manihot esculenta Crantz* and *Eruca sativa Mill* (Havaux et al. 2003; Khodakovskaya et al. 2006; Kwak et al. 2019). Recently, the cp genomes of some *Miscanthus* species have been available in National Center for Biotechnology Information (NCBI) database (Sheng et al. 2021). These complete cp sequences in *Miscanthus* species can be used for studying population genetics, evolution analysis and plastid engineering (Amiryousefi et al. 2018; Yan et al. 2019).

Codon usage bias referring the usage frequencies of alternative synonymous is variable between genomes (Plotkin & Kudla 2011). The pattern of codon usage could be caused by multi-factors during the process of genome and gene evolution, including natural selection, compositional mutation mode, translational selection and so on (Pop et al. 2014; Quax et al. 2015; Tuller et al. 2010). The studies of codon preference can not only reveal the evolutionary rules between gene in a species or related species, but also improve the expression efficiency of exogenous in transgenic research by codon optimization. Recently, the applicability of codon optimization in the cp genomes have been proven in many vascular plants, including *Poaceae* （*Gramineae*）(Zhang et al. 2012), *Cinnamomum camphorn (L.) presl* (Chen et al. 2017), *Fragaria × ananassa Duch* (*strawberry*) (Cheng et al. 2017) and *Solanum tuberosum L.* (Zhang et al. 2018). However, the codon usage pattern of cp genomes in *Miscanthus* and related species has not been fully elucidated.

Currently, the codon usage patterns in cp genomes of seven *Miscanthus* and related species, including *M. sinensis*, *Miscanthus transmorrisonensis Hayata* (*M. transmorrisonensis*), *M. floridulus*, *M. sacchariflorus*, *Miscanthus x giganteus* (*M. x giganteus*), *Sorghum bicolor* and *Saccharum spontaneum* were systematically analyzed based on the previous published genome-wide data. Among them, *M. sinensis*, *M. transmorrisonensis* and *M. floridulus* belong to the Miscanthus clades under the *Miscanthus* species, *M. sacchariflorus* belong to the Triarrhena clades under the *Miscanthus* species, and *M. x giganteus* is a natural triploid hybrid. According to our previous studies, *Miscanthus* are most closely related to *Sorghum bicolor* (L.) *Moench* and *Saccharum L* (Sheng et al. 2021; Sheng et al. 2017). In addition, the codon usage bias of these seven species was compared with the other four model species including *Populus trichocarpa Torr & Gray*, *Escherichia coli*, *Arabidopsis thaliana (L.) Heynh* and *Saccharomyces cerevisiae*. *Miscanthus* are not only potential bio-energy crops, but also excellent breeding resources. So, it is of interest to understand the codon usage of *Miscanthus* for better utilization of *Miscanthus* and related resources as germplasm resources. Here, we revealed the codon usage patterns of the *Miscanthus* and related species and determined optimal codons for cp genetic engineering. The results in the current study will not only provide insight into genetic evolution studies, but also provide a reference for selecting appropriate heterologous gene expression receptor system to improve the gene expression of *Miscanthus* plants by optimizing codon.

Materials & Methods

Genomes and sequences selection

The complete cp genomes of *M. floridulus* (NC_035750.1), *M. sacchariflorus* (NC_028720.1), *M. sinensis* (NC_028721.1), *M. x giganteus* (NC_035753.1), *M. transmorrisonensis* (NC_035752.1), *Sorghum bicolor* (NC_008602), *Saccharum spontaneum* (NC_034802.1) with gene annotation were downloaded from the NCBI GeneBank database. The number of raw CDS of above seven species was 106, 122, 122, 106, 106, 84 and 76 respectively (Table 1). To avoid sampling bias, the CDS sequences were screened from genome-wide data by python scripts (https://github.com/shexuan/codon_analysis) according to the following principles: (1) CDS contains initiation codon (ATG), termination codons (TAA, TAG or TGA) and without intermediate stop codon in the sequences; (2) the number of bases in each CDS must be the fold of three (3) the length of sequence of CDS should be⩾300 bp, because the effective number of codon (ENC) could not be calculated correctly with shorter coding sequences (Wright 1990; Zhang 2007). After filtration, the amount of CDS, the base composition at the first/second/third site of codons (GC1/GC2/GC3) and average GC, as well as the total amino acid amounts were calculated.

Analysis of relative synonymous codon usage (RSCU) and relative synonymous codon usage frequency (RFSC)

RSCU value of a codon (the number of codon occurrences in a gene divided by the number of codon appearances expected under the same codon usage) is the ratio of its actual frequency of utilization to the expected usage frequency without bias. The RSCU was calculated as Eq. (1):

$$\text{RSCU} = \frac{x_{ij}}{\sum_{j}^{N_i} x_{ij}} n_i \qquad (1)$$

Where x_{ij} represents the frequency of codon j encoding the i th amino acid, and n_i represents the number of synonymous codon encoding the i th amino acid (Sharp P M 1986). If the RSCU value of a codon is equal to 1, the codon is used without bias. An RSCU value greater than 1 reflects a significant codon usage bias and vice versa (Sharp P M 1987).

The RFSC value refers to the proportion of the actually observational number of a codon in the number of all synonymous codons. The RFSC were calculated using Eq. (2):

$$\text{RFSC} = \frac{x_{ij}}{\sum_{j}^{N_i} x_{ij}} \qquad (2)$$

Where x_{ij} represents the frequency of codon j encoding for the I th amino acid. The high-frequency codon was screened based on the results of RFSC in all codon. The screening principles were as follows: the RFSC>60% of one codon; or the RFSC of a codon exceeds the average frequency of synonymous codon by 0.5 times (Zhou et al. 2007).

Determination of optimal codons

The effective number of codons (ENC) can be applied to describe the extent of deviation of codon usage from the random selection, which reflects the degree of unbalanced use of synonymous codon in genes. The ENC value range from 20 (each amino acid uses only one synonymous codon) to 61 (Each synonymous codon is equally used), which is inversely proportional to the codon bias (Wright 1990). The ENC value in each species was calculated by CodonW software and then 10% of the CDS with remarkable high and low expression levels were filtered out according to the ENC value. The RSCU of each codon was obtained from the sequence files of the high and low groups according to the cusp function of emboss (https://www.bioinformatics.nl/emboss-explorer/). Optimal codons were determined by ΔRSCU method. Specifically, the average RSCU values of the two amino acid groups were computed and then minus (ΔRSCU). The codon will be identified as the optimal codon through comparing the high and low group of the same codon Δ RSCU (> 0.08) and RSCU value (high group > 1, low group < 1) (Romero et al. 2000).

Comparative analysis of codon usage frequency

The ratio of codon usage frequency is one indicator of codon usage bias among species. To further explore the codon usage patterns in the seven species of *Miscanthus* and their relatives, codon usage bias data of four model species including *Escherichia coli* (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=199310); *Saccharomyces cerevisiae* (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=4932); *Populus trichocarpa* (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=3694) and *Arabidopsis thaliana* (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=3702), which have been used as the most common expression vectors were downloaded from the Codon Usage Database. Subsequently, the codon usage frequencies of the seven species in this study were compared with above four model organism. When the ratio is ⩾2 or ⩽0.5, it suggests that the codon bias difference between the two organisms is significant, otherwise, it means that is small (Pan et al. 2013).

Analysis of ENC-plot

GC3s value represents the proportion of G and C content at the third position of a codon to the total number of gene bases. ENC-plot is plotted with ENC values as ordinate and GC3 value as abscissa, which can be used to analyze the codon usage characteristics of each gene and to explore the relevance between gene base component and codon preference (Wright 1990). ENC values are located on or near the expected curve, when mutation pressure makes a key role in the formation of codon usage patterns. Conversely, when the use of codon is constrained by natural selection, the ENC value will be well below the prospective curve (Wright 1990).

PR2-plot analysis

PR2-plot is a graphical analysis takes G3/(G3+C3) as the abscisic and A3/(A3+T3) as the ordinate, which is performed to explore the composition of the four bases at the third site of amino acids (Sueoka 1999; Sueoka 1995). The pattern of splashes around the central spot (A=T, C=G) indicate the extent and orientation of the base offset.

Analysis of Neutrality plot

Neutrality analysis is used to exploring the degree of impact between natural selection and mutation pressure on the mode of codon usage (Sueoka 1988). GC12 indicates the mean GC content at the first and second sites of the codon, while GC3 represents the GC content of the third site. GC content at the third positions of codon was counted eliminating the Codon Met (ATG) and Trp (TGG). GC3 was counted eliminating the three stop codons (TAA, TAG and TGA) and three codons (ATT, ATC and ATA) of Ile (Sueoka 1988). Both GC12 and GC3 of the seven cp genomes were counted by Python scripts (https://github.com/shexuan/codon_analysis). If the gradient of the curve regression is 0, indicating that there is no impact of mutation pressure. Gradient 1 represents complete neutrality, which describes that codon usage preference is completely influenced by mutation pressure (Sueoka 1988).

Correspondence analysis of codon usage

The variations of codon usage in the seven cp genomes were investigated based on the correspondence analyses (COA) using CodonW (Anue et al. 2019). The usage patterns of 59 codons (exincluding Met, Trp and three termination codons) were compared and all genes can be embeded into a 59-dimension hyperspace, in which each dimensions responding to the synonymous codon usage of the gene (Xiang et al. 2015). Therefore, the major trends (Axis 1) of these axes in the 59-dimensional hyperspace can be used to determine the maximum fraction of genetic variation, indicating the major sources of codon usage variation. In addition, according to the results of COA, the correlation index between Axis1 and codon usage exponent, including the GC content of codons, GC3s, codon adaptation index (CAI) and the total numbers of amino acids (L_aa) were computed by python scripts package (https://github.com/shexuan/codon_analysis). CAI value is widely applied to assess gene expression levels, ranging from 0 to 1. Specifically, the larger CAI value is, the stronger codon usage preference is and vice versa (Sharp & Li 1986).

Results

Characteristics of codon usage bias

Analysis of base composition of codon

The screened CDSs processed by Python scripts contained 65, 64, 64, 64, 64, 48 and 52 for *M. floridulus*, *M. giganteus*, *M. sacchariflorus*, *M. sinensis*, *M. transmorrisonensis*, *Saccharum spontaneum* and *Sorghum bicolor* respectively. In addition, the GC contents of three positions of codons (GC1, GC2, GC3) were calculated respectively (Table 1). It was found that the contents of GC at all three sites and the average GC content (GC123) were all less than 0.5, which indicated the seven cp genomes prone to use A/T bases and A/T-ending codons (Table 1). Specifically, the mean GC content of three sites in *M. floridulus*, *M. giganteus* and *M. transmorrisonensis* is the same (0.375), the same in *M. sacchariflorus* and *M. sinensis* (0.393), but slightly different in *Saccharum spontaneum* (0.391) and *Sorghum bicolor* (0.39) (Table 1). Furthermore, the distribution trend of GC content was GC1 > GC2 >GC3, indicating that GC was not evenly distributed in the three positions of the codon. In summary, the codon usages of GC content in these seven cp genomes were similar and were biased towards A/T bases.

Table 1 Genomic features of cp genomes of the seven Miscanthus and related species (the total number of amino acids: L_aa ; the GC content at the first, second and third codon positions: GC1, GC2and GC3; average GC at three locations: GC123).

RSCU and RFSC

The cp genomes of the seven *Miscanthus* and related species have 30 common codons (RSCU > 1) with 28 codons ending with A/T (93.3%) (Table S1). Therefore, the codons of the seven plants (RSCU > 1) are likely to end with A/T. The variation ranges in the RSCU values were close in the seven cp genomes, i.e., 0.31–1.93 in *M. floridulus*, *M. giganteus* and *M. transmorrisonensis*, 0.32-1.94 in *M. sacchariflorus* and *M. sinensis*, 0.32- 2.01 in *Saccharum spontaneum* and 0.33-2.04 in Sorghum bicolor, respectively (Table S1). In addition, the maximum and the minimum RSCU values belonged to TTA and CTG which encode Leu, indicating the vitally positive bias. Furthermore, the pattern of codon usage were summarized in the seven *Miscanthus* and related species (Figure 1). Specifically, the high-frequency codons of seven *Miscanthus* and related species possess strong common base and share a total of 21 high-frequency codons (Table S1). Besides, *Saccharum spontaneum* and *Sorghum bicolor* possess two more high-frequency codon than other five *Miscanthus* species.

Figure 1. Codon content in all protein-coding genes of the seven *Miscanthus* and related chloroplast genomes. The histogram of each amino acid indicated codon usage within the seven species. (From left to right: *M. floridulus*, *M. giganteus*, *M. sacchariflorus*, *M. sinensis*, *M. transmorrisonensis*, *Saccharum spontaneum* and *Sorghum bicolor*).

Determination of optimal codons

The ENC values of each CDS were ranked and 10% of genes from both ends were selected to establish high and low expression gene banks respectively. The RSCU values and ΔRSCU values in the two expression library were calculated and listed in Supplementary Table S2. According to the values of ΔRSCU, the optimal codons in the seven represent species were determined as follows (Table 2).

Table 2 Optimal codons in chloroplast genomes of the seven *Miscanthus* and related species.

Codon usage frequency

The codon usage frequencies of the seven cp genomes were compared with four model species including *Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Populus trichocarpa* (Table S3). Results indicated that there are litter divergence in the codon usage frequencies among the seven represent plants with *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Populus trichocarpa*, possess 9–11 (accounting for 14.06%–17.19% of total codons), 13–14 (20.31%–21.88%), 12–13 (18.75%–20.31%) different codons, respectively (Table S3). However, the codon usage frequencies of the seven species with *Escherichia coli* were comparatively higher (27 different codons). The results indicated that the codon frequency difference between *Miscanthus* species and *Arabidopsis*, *Poppoplar* and *cerevisiae* was the least, while was the largest with *Escherichia coli*. Based on above results, it was preferred to select *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Populus trichocarpa* as heterologous gene expression receptor for *Miscanthus* and related species. In addition, the results shown that TAG is a different termination codon in comparison of the seven plants with the four model species.

Source analysis of variation in codon usage

ENC-plot

The ENC and GC3s of the seven *Miscanthus* and related plant cp genomic were analyzed and plotted. It can be seen from Figure 2 that the ENC values of most genes were lower than expected values and lie below the standard curve. The results of ENC-plot analysis suggested that codon usage preference of the seven cp genomes is mainly influenced by natural selection and other factors, while mutation pressure play slightly roles.

Figure 2. ENC-plot (ENC values vs GC3s) for chloroplast genomes of seven *Misacanthus* and related species. The continuous curve represents the expected curve between GC3s and ENC under random codon usage.

PR2-plot

PR2-plot is an efficient method to indicate the influence of mutation pressure by investigating the composition of A, T, C and G at the third position. Our results revealed that the AT-bias is 0.464, 0.463, 0.463, 0.463, 0.463, 0.465 and 0.463 for *M. floridulus*, *M. giganteus*, *M. sacchariflorus*, *M. sinensis*, *M. transmorrisonensis*, *Saccharum spontaneumand* and *Sorghum bicolor*, while the GC-bias is 0.512, 0.513, 0.516, 0.515, 0.512, 0.515 and 0.518, respectively (Fig 3). Therefore, T/G-bias was observed in all seven *Miscanthus* and related species. All in all, codon usage bias of A/T and G/C in the seven cp genomes was lopsided, indicating that the base composition of the seven cp genomes is not only influenced by mutation pressure, but also by natural selection.

Figure 3. PR2-plot (A3/(A3+T3) vs G3/(G3+C3)) for chloroplast genomes of seven *Misacanthus* and related species. The pattern of splashes around the central spot (A=T, C=G) indicate the extent and orientation of the base offset.

Neutrality plot

The distribution range of GC12 and GC3 is relatively concentrated, in which the range of GC12 is 0.3272 ~ 0.5469, and the range of GC3 is 0.1794 ~ 0.512 (Fig 4). No significant correlation was found for GC1 with GC2 (r1=0.157, r2=0.168, r3=0.128, r4=0.127, r5=0.161, r6=0.155, r7=0.140), GC1 with GC3 (r8=0.092, r9=0.079, r10=0.055, r11=0.049, r12=0.1, r13=0.242, r14=0.202) and GC2 with GC3 (r14=0.054, r14=0.053, r15=-0.014, r16=-0.02, r17=0.063, r18=-0.032, r19=-0.032), which suggested mutation pressure make a minor role in the codon usage preference. In addition, the regression coefficient (slope of neutrality plot) was 0.0062-0.1976, indicating that the correlation between GC12 and GC3 is not significant, and the composition of the first two bases may be different from the third base of the codon. These results demonstrated that the codon usage patterns of cp gene in the seven species are mainly affected by natural selection.

Figure 4. Neutrality plot (GC12 vs GC3) for chloroplast genomes of seven *Misacanthus* and related species. The solid line represents the regression line.

Correspondence analysis (COA)

COA is used to explore the variations of codon usage in the cp genomes. In the current study, RSCU-based COA was used to compare the usage patterns of 59 codons, which produced a series of orthogonal axes, reflecting the trend of change of codon usage in the seven *Miscanhus* and related cp genomes. The first four axes accounted for 36.22%, 38.18%, 38.47%, 38.31%, 36.8%, 40.72% and 40.94% of the overall changes, while the first axis proportion to 14.33%, 15.91%, 15.84%, 15.78%, 14.59%, 15.59% and 9.70% of the total variation in seven species respectively (Table 3). Axis 1, responsible for~10% of total variation, was the main source of variation, indicating that the codon usage should be influenced by multiple factor. In addition, the relationship between axis 1 and axis 2 was visualized to explore the effects of GC content on codon usage bias (Fig 5). Genes with different GC content are plotted as different colors, red with GC%<45% and blue with 45%⩽GC%<60%. In order to determine the factors leading to gene dispersion along axis 1 and axis 2, the correlation index were computed on axis 1 with CAI, GC3, L_aa and so on (Table 3). As can be seen from the results in Table 3, axis 1 for *M. floridulus*, *M. sacchariflorus*, *M. sinensis*, *M. transmorrisonensis*, *Saccharum spontaneum* and *Sorghum bicolor* possessed a remarkable correlation with GC3s (p⩽0.01), which indicated the base composition in mutation pressure was the main factor impacting codon usage preference.

Table 3 Correlation analysis of axis 1 and codon usage index of cp genomes of seven *Misacanthus* and related species (the T/C/A/G content at the third codon position of synonymous codons; codon adaptation index: CAI; codon bias index: CBI; frequency of optimal codons: Fop; the GC content at the third codon position of synonymous codons: GC3s; the GC content at the three position of synonymous codons: GC; total number of amino acids: L_aa)

Figure 5. Visualization of the first two axes from the correspondence analysis based on RSCU values for chloroplast genomes of seven *Misacanthus* and related species. Genes with different GC content are plotted as different colors, red with GC%<45% and blue with 45%⩽GC%<60%.

Discussion

The study compared the codon usage patterns of the five Miscanthus species and two related species. It will help further improve our understanding of evolution analysis and the optimization of codon components suitable for gene expression. During the evolutionary processes, specific codon usage patterns were obtained to adapt to the diverse factors including origin, evolution, natural selection and mutation pressure. In addition, analyzing the source of variation in genomic codon usage, the pattern of codon bias and the high frequency codon could provide insights into optimization of the codons of heterologous genes and selection of appropriate heterologous gene expression receptor system. Therefore, the research will be of great significance to the study of genetic engineering and genetic evolution.

Our analysis of base composition of codons revealed that the CDS of the seven *Miscanthus* and related cp genomes tended to use A/T codon, which was consistent with the results of Zhang et al. on the 23 Poaceae cp genomes (Zhang et al. 2012). According to previous study, the GC3S value of dicotyledonous plants is often less than 50% (codon use prefers A/T), which is different from the monocotyledonous plants with high GC3S value (GC3S value >50%, showing that codon use prefers G/C) (Murray et al. 1989). The results of RSCU value analysis showed an A/T codon usage bias in the cp genomes of the seven representative plants, which was consistent with the patterns in most higher plants (Shang et al. 2011). According to neutral evolution theory, the effects of mutation pressure and natural selection on the variation of the third base of codon are neutral or nearly neutral (Sharp et al. 1993). The study of Kawabe et al.

showed that when codon use is affected by natural selection, GC3 values tend to be distributed in a small range and there is no significant correlation between GC12 and GC3 (Kawabe et al. 2003). The neutrality plot in this study revealed a weak correlation between GC12 and GC3 and the composition of the first two bases were different from the third base of the codon, which demonstrated that the codon usage patterns of the seven cp genomes are mainly influenced by natural selection. This result is consistent with the codon usage of cp genomes of many species, such as *Oryza sativa* (Liu et al. 2003), *Zea mays*, *Triticum aestivum* (Liu & Xue 2005) and *Euphorbiaceae* (Wang et al. 2020). In addition, combining the results of ENC-plot, PR2-plot and COA suggested that the codon usage bias of the seven cp genomes were affected by multiple factors, such as mutation pressure, base composition and gene length, among which the main influencing factor was natural selection. Similar results have also been reported in cp genomes of *Poaceae* (Zhang et al. 2012), *Populus alba* (Zhou et al. 2008) and *Euphorbiaceae* (Wang et al. 2020), where multiple factors influences on codon usage bias.

The cp genomes of the seven *Miscanthus* and related plants possess strong common base and share totally 21 high-frequency codons. In addition, 4-11 codons were determined to be the optimal codons in each species, while no common optimal codon was defined in the seven representative species. These results of high frequency codons and optimal codons are not only beneficial to the codon optimization, but also promote further understanding of the relationship between gene expression and codon usage preference. In higher plants, the main obstacle to applying cp transformation to more species and especially, to other important crops is the limitation of available tissue culture systems and regeneration protocols (Ruf et al. 2001). Considering the variations in codon usage bias among the seven represent cp genomes and the receptors for heterologous genes expression, codon usage frequencies were analyzed in this study as well. Based on the results, it was suggested to select *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Populus trichocarpa* as heterologous gene expression receptor for *Miscanthus* and related crops, which possesse a little difference in codon usage frequency with the seven plants.

This study conducted a comprehensive comparative analysis on codon usage patterna at the cp genome-wide level of seven *Miscanthus* and related species. These results will improve our understanding on evolution analysis, the selection of appropriate heterologous gene expression receptor system and the optimization of codon components suitable for gene expression, finally provide a theoretical basis for building a stable and efficient gene expression system in *Miscanthus* or other crops.

Conclusions

The codon usage patterns of cp genomes of the five *Miscanthus* and two related species were compared and systematically analyzed for the first time. The results of codon usage bias and RSCU analysis indicated that the seven representative species prefers to A/T bases and A/T-ending codons. In addition, 21 common high-frequency codons and 4-11 optimal codons were elected in the seven cp genomes. Furthermore, the analysis of codon usage frequencies between the seven representative species and four model organisms suggested that *Arabidopsis thaliana*, *Populus trichocarpa* and *Saccharomyces cerevisiae* considered as appropriate exogenous expression receptors for *Miscanthus* and related species. Finally, combination of the results of ENC-plot, PR2-plot and neutrality analysis revealed the codon usage pattern of the seven cp genomes is influenced by multiple factors, among which the dominant influencing factor was natural selection. These results in the study might provide not only important reference information for evolutionary analysis, but also deepened insights in improving the expression efficiency of exogenous gene in transgenic research by codon optimization.

References

Amiryousefi A, Hyvönen J, and Poczai PJPo. 2018. The chloroplast genome sequence of bittersweet (Solanum dulcamara): Plastid genome structure evolution in Solanaceae. 13:e0196069.

Anue MR, Xueli S, Chunzhen C, Zhongxiong LJPD, and Pests. 2019. Analysis of codon usage pattern of banana basic secretory protease gene. 10:1-9.

Barling A, Swaminathan K, Mitros T, James BT, Morris J, Ngamboma O, Hall MC, Kirkpatrick J, Alabady M, and Spence AKJBg. 2013. A detailed gene expression study of the Miscanthus genus reveals changes in the transcriptome associated with the rejuvenation of spring rhizomes. 14:1-16.

Chae WB, Hong SJ, Gifford JM, Rayburn AL, Sacks EJ, and Juvik JAJGB. 2014. Plant morphology, genome size, and SSR markers differentiate five distinct taxonomic groups among accessions in the genus M iscanthus. 6:646-660.

Chen C, Zheng Y, Liu S, Zhong Y, Wu Y, Li J, Xu L-A, and Xu MJP. 2017. The complete chloroplast genome of Cinnamomum camphora and its comparison with related Lauraceae species. 5:e3820.

Cheng H, Li J, Zhang H, Cai B, Gao Z, Qiao Y, and Mi LJP. 2017. The complete chloroplast genome sequence of strawberry (Fragaria× ananassa Duch.) and comparison with related species of Rosaceae. 5:e3919.

Clark LV, Stewart JR, Nishiwaki A, Toma Y, Kjeldsen JB, Jørgensen U, Zhao H, Peng J, Yoo JH, and Heo KJJoEB. 2015. Genetic structure of Miscanthus sinensis and Miscanthus sacchariflorus in Japan indicates a gradient of bidirectional but asymmetric introgression. 66:4213-4225.

Ge C, Liu X, Liu S, Xu J, Li H, Cui T, Yao Y, Chen M, Yu W, and Chen CJPMBR. 2017. Miscanthus sp.: genetic diversity and phylogeny in China. 35:600-610.

Havaux M, Lütz C, and Grimm BJPP. 2003. Chloroplast membrane photostability in chlPtransgenic tobacco plants deficient in tocopherols. 132:300-310.

Jarvis P, and López-Juez EJNRMCB. 2013. Biogenesis and homeostasis of chloroplasts and other plastids. 14:787-802.

Kawabe A, Miyashita NTJG, and systems g. 2003. Patterns of codon usage bias in three dicot and four monocot plant species. 78:343-352.

Khodakovskaya M, McAvoy R, Peters J, Wu H, and Li YJP. 2006. Enhanced cold tolerance in transgenic tobacco expressing a chloroplast ω-3 fatty acid desaturase gene under the control of a cold-inducible promoter. 223:1090-1100.

Kwak S-Y, Lew TTS, Sweeney CJ, Koman VB, Wong MH, Bohmert-Tatarev K, Snell KD, Seo JS, Chua N-H, and Strano MSJNn. 2019. Chloroplast-selective gene delivery and expression in planta using chitosan-complexed single-walled carbon nanotube carriers. 14:447-455.

Liu Q-P, Tan J, and Xue Q-ZJYcxbAgS. 2003. Synonymous codon usage bias in the rice cultivar 93-11 (Oryza sativa L. ssp. indica). 30:335-340.

Liu Q, and Xue QJJog. 2005. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. 84:55-62.

Murray EE, Lotzer J, and Eberle MJNAR. 1989. Codon usage in plant genes. 17:477-498.

Nielsen AZ, Mellor SB, Vavitsas K, Wlodarczyk AJ, Gnanasekaran T, Perestrello Ramos H de Jesus M, King BC, Bakowski K, and Jensen PEJTPJ. 2016. Extending the biosynthetic repertoires of cyanobacteria and chloroplasts. 87:87-102.

Pan L-L, Wang Y, Hu J-H, Ding Z-T, and Li CJJotb. 2013. Analysis of codon use features of stearoyl-acyl carrier protein desaturase gene in Camellia sinensis. 334:80-86.

Plotkin JB, and Kudla GJNRG. 2011. Synonymous but not the same: the causes and consequences of codon bias. 12:32-42.

Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS, and Koller DJMsb. 2014. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. 10:770.

Quax TE, Claassens NJ, Söll D, and van der Oost JJMc. 2015. Codon bias as a means to fine-tune gene expression. 59:149-161.

Ravi V, Khurana J, Tyagi A, Khurana PJPS, and Evolution. 2008. An update on chloroplast genomes. 271:101-122.

Romero H, Zavala A, and Musto HJNar. 2000. Codon usage in Chlamydia trachomatis is the result of strand-specific mutational biases and a complex pattern of selective forces. 28:2084-2090.

Ruf S, Hermann M, Berger IJ, Carrer H, and Bock RJNb. 2001. Stable genetic transformation of tomato plastids and expression of a foreign protein in fruit. 19:870-875.

Shang M, Liu F, Hua J, and Wang KJSAS. 2011. Analysis on codon usage of chloroplast genome of Gossypium hirsutum. 44:245-253.

Sharp P M LWH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of molecular evolution* 24:28-38.

Sharp P M LWH. 1987. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* 15:1281-1295.

Sharp PM, and Li W-HJJome. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. 24:28-38.

Sharp PM, Stenico M, Peden JF, and Lloyd ATJBST. 1993. Codon usage: mutational bias, translational selection, or both? 21:835-841.

Sheng J, Hu X, Zeng X, Li Y, Zhou F, Hu Z, Jin S, and Diao Y. 2016. Nuclear DNA content in Miscanthus sp. and the geographical variation pattern in Miscanthus lutarioriparius. *Sci Rep* 6:34342. 10.1038/srep34342

Sheng J, Yan M, Wang J, Zhao L, Zhou F, Hu Z, Jin S, and Diao Y. 2021. The complete chloroplast genome sequences of five Miscanthus species, and comparative analyses with other grass plastomes. *Industrial Crops and Products* 162. 10.1016/j.indcrop.2021.113248

Sheng J, Zheng X, Wang J, Zeng X, Zhou F, Jin S, Hu Z, and Diao Y. 2017. Transcriptomics and proteomics reveal genetic and biological basis of superior biomass crop Miscanthus. *Scientific Reports* 7. 10.1038/s41598-017-14151-z

Sueoka NJG. 1999. Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+ C content of third codon position. 238:53-58.

Sueoka NJJome. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. 40:318-325.

Sueoka NJPotNAoS. 1988. Directional mutation pressure and neutral molecular evolution. 85:2653-2657.

Tuller T, Waldman YY, Kupiec M, and Ruppin EJPotNAoS. 2010. Translation efficiency is determined by both codon bias and folding energy. 107:3645-3650.

Vermerris W. 2008. Miscanthus: genetic resources and breeding potential to enhance bioenergy production. *Genetic improvement of bioenergy crops*: Springer, 295-308.

Wang Z, Xu B, Li B, Zhou Q, Wang G, Jiang X, Wang C, and Xu Z. 2020. Comparative analysis of codon usage patterns in chloroplast genomes of six Euphorbiaceae species. *PeerJ* 8:e8251. 10.7717/peerj.8251

Wright FJG. 1990. The ‘effective number of codons’ used in a gene. 87:23-29.

Wu Y, Zhou N, Ni X, Okoye CO, Wang Y, Li X, Gao L, Zhou G, Jiang JJIC, and Products. 2021. Developing a long-term and powerful in vitro culture and Agrobacterium-mediated transformation system for Miscanthus sinensis (Poaceae). 161:113190.

Xiang H, Zhang R, Butler III RR, Liu T, Zhang L, Pombert J-F, and Zhou ZJPo. 2015. Comparative analysis of codon usage bias patterns in microsporidian genomes. 10:e0129223.

Yan M, Zhao X, Zhou J, Huo Y, Ding Y, and Yuan ZJIjoms. 2019. The complete chloroplast genomes of Punica granatum and a comparison with other species in Lythraceae. 20:2886.

Zhang J, Yan J, Zhang Y, Ma X, Bai S, Wu Y, Dao Z, Li D, Zhang C, and Zhang YJPo. 2013. Molecular insights of genetic variation in Erianthus arundinaceus populations native to China. 8:e80388.

Zhang R, Zhang L, Wang W, Zhang Z, Du H, Qu Z, Li X-Q, and Xiang HJIjoms. 2018. Differences in codon usage bias between photosynthesis-related genes and genetic system-related genes of chloroplast genomes in cultivated and wild solanum species. 19:3142.

Zhang WJ, Zhou, J., Li, Z. F., Wang, L., Gu, X., & Zhong, Y. 2007. Comparative analysis of codon usage patterns among mitochondrion, chloroplast and nuclear genes in Triticum aestivum L. . *Journal of Integrative Plant Biology* 49:246-254.

Zhang Y, Nie X, Jia X, Zhao C, Biradar SS, Wang L, Du X, and Weining SJAJoB. 2012. Analysis of codon usage patterns of the chloroplast genomes in the Poaceae family. 60:461-470.

Zhou M, Long W, and Li XJJoFR. 2008. Analysis of synonymous codon usage in chloroplast genome of Populus alba. 19:293-297.

Zhou M, Tong C, Shi JJJoG, and Genomics. 2007. Analysis of codon usage between different poplar species. 34:555-561.

Table 1(on next page)

Genomic features of chloroplast genomes of the seven *Miscanthus* and related species

the total number of amino acids: L_aa ; the GC content at the first, second and third codon positions: GC1, GC2and GC3; average GC at three locations: GC123

Table 1 Genomic features of chloroplast genomes of the seven *Miscanthus* and related species (the total number of amino acids: L_aa ; the GC content at the first, second and third codon positions: GC1, GC2and GC3; average GC at three locations: GC123).

Parameters	**Miscanthus floridulus**	**Miscanthus giganteus**	**Miscanthus sacchariflorus**	**Miscanthus sinensis**	**Miscanthus transmorrisonensis**	**Saccharum spontaneum**	**Sorghum bicolor**
L_aa	19611	19469	19508	19506	19486	16553	17490
CDSs number (before filting)	106	106	122	122	106	76	84
CDSs number (after filting)	65	64	64	64	64	48	52
GC1	0.473	0.474	0.472	0.472	0.473	0.477	0.476
GC2	0.397	0.397	0.396	0.396	0.397	0.395	0.393
GC3	0.312	0.311	0.311	0.311	0.311	0.302	0.303
GC123	0.394	0.394	0.393	0.393	0.394	0.391	0.39

Table 2(on next page)

Optimal codons in chloroplast genomes of the seven *Miscanthus* and related species

Table 2 Optimal codons in chloroplast genomes of the seven *Miscanthus* and related species.

Species	**Optimal codon numbers**	**Optimal codon**
Miscanthus floridulus	4	'CAC', 'CCA', 'TCA', 'TAG'
Miscanthus giganteus	6	'TTC', 'CTT', 'CCA', 'AGG', 'TCA', 'TGA'
Miscanthus sacchariflorus	4	'GCC', 'CTT', 'AGG', 'ACG'
Miscanthus sinensis	11	'GCC', 'TTC', 'GGC', 'ATA', 'CTA', 'AGA', 'AGG', 'TCT', 'ACC', 'ACG', 'GTC'
Miscanthus transmorrisonensis	4	'CTT', 'CCA', 'TCA', 'TGA'
Saccharum spontaneum	4	'CTT', 'CCA', 'TCA', 'TAG'
Sorghum bicolor	8	'GCC', 'GGA', 'CAT', 'ATA', 'TCA', 'ACA', 'TAG', 'TGA'

Table 3(on next page)

Correlation analysis of axis 1 and codon usage index of chloroplast genomes of seven *Misacanthus* and related species

the T/C/A/G content at the third codon position of synonymous codons; codon adaptation index: CAI; codon bias index: CBI; frequency of optimal codons: Fop; the GC content at the third codon position of synonymous codons: GC3s; the GC content at the three position of synonymous codons: GC; total number of amino acids: L_aa

Table 3 Correlation analysis of axis 1 and codon usage index of chloroplast genomes of seven *Misacanthus* and related species (the T/C/A/G content at the third codon position of synonymous codons; codon adaptation index: CAI; codon bias index: CBI; frequency of optimal codons: Fop; the GC content at the third codon position of synonymous codons: GC3s; the GC content at the three position of synonymous codons: GC; total number of amino acids: L_aa)

Species	**T3s**	**C3s**	**A3s**	**G3s**	**CAI**	**CBI**	**Fop**	**Nc**	**GC3s**	**GC**	**L_aa**
Miscanthus floridulus	-0.65**	0.507**	-0.133	0.598**	-0.152	0.073	0.098	0.28**	0.639**	0.159	-0.315**
Miscanthus x giganteus	0.016	0.053	-0.104	-0.127	-0.038	0.044	0.078	0.09	-0.018	0.223**	0.057
Miscanthus sacchariflorus	0.66**	-0.463**	0.171*	-0.708**	0.176*	0.044	0.086	-0.43**	-0.7**	-0.154	0.118
Miscanthus sinensis	0.663**	-0.469**	0.18*	-0.711**	0.17*	0.043	0.083	-0.423**	-0.708**	-0.154	0.123
Miscanthus transmorrisonensis	0.647**	-0.509**	0.129	-0.604**	0.147	-0.076	-0.099	-0.285**	-0.641**	-0.152	0.318**
Saccharum spontaneum	0.693**	-0.479**	0.201*	-0.691**	0.193	0.02	-0.012	-0.217*	-0.667**	-0.19	0.197*
Sorghum bicolor	0.412**	-0.464**	0.25**	-0.726**	0.33**	0.253**	0.252**	-0.48**	-0.685**	-0.035	0.034

Notes: *P<0.05; **P<0.01

Figure 1

Codon content in all protein-coding genes of the seven *Miscanthus* and related cp genomes

From left to right: *M. floridulus*, *M. giganteus*, *M. sacchariflorus*, *M. sinensis*, *M. transmorrisonensis*, *Saccharum spontaneum* and *Sorghum bicolor*

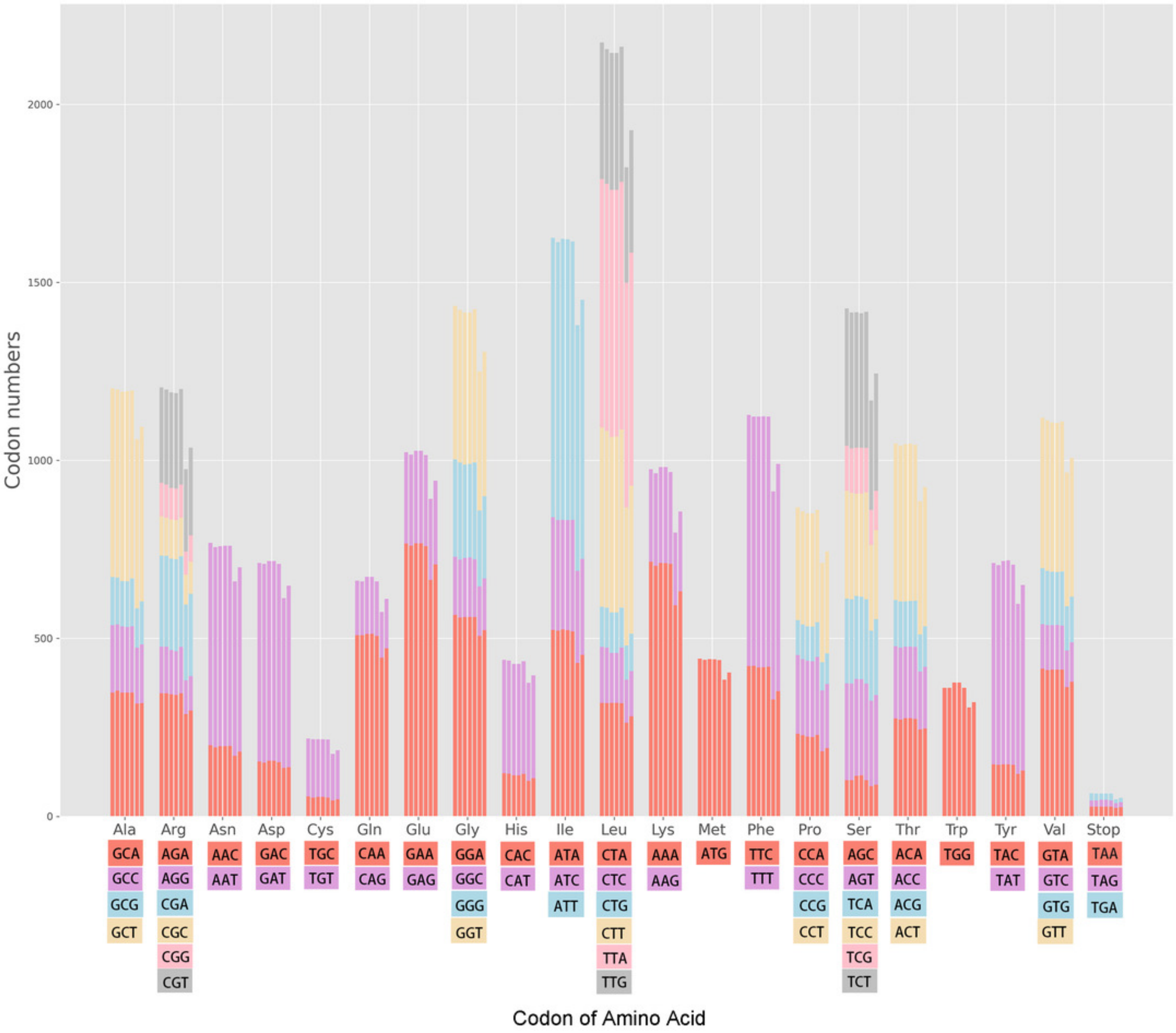
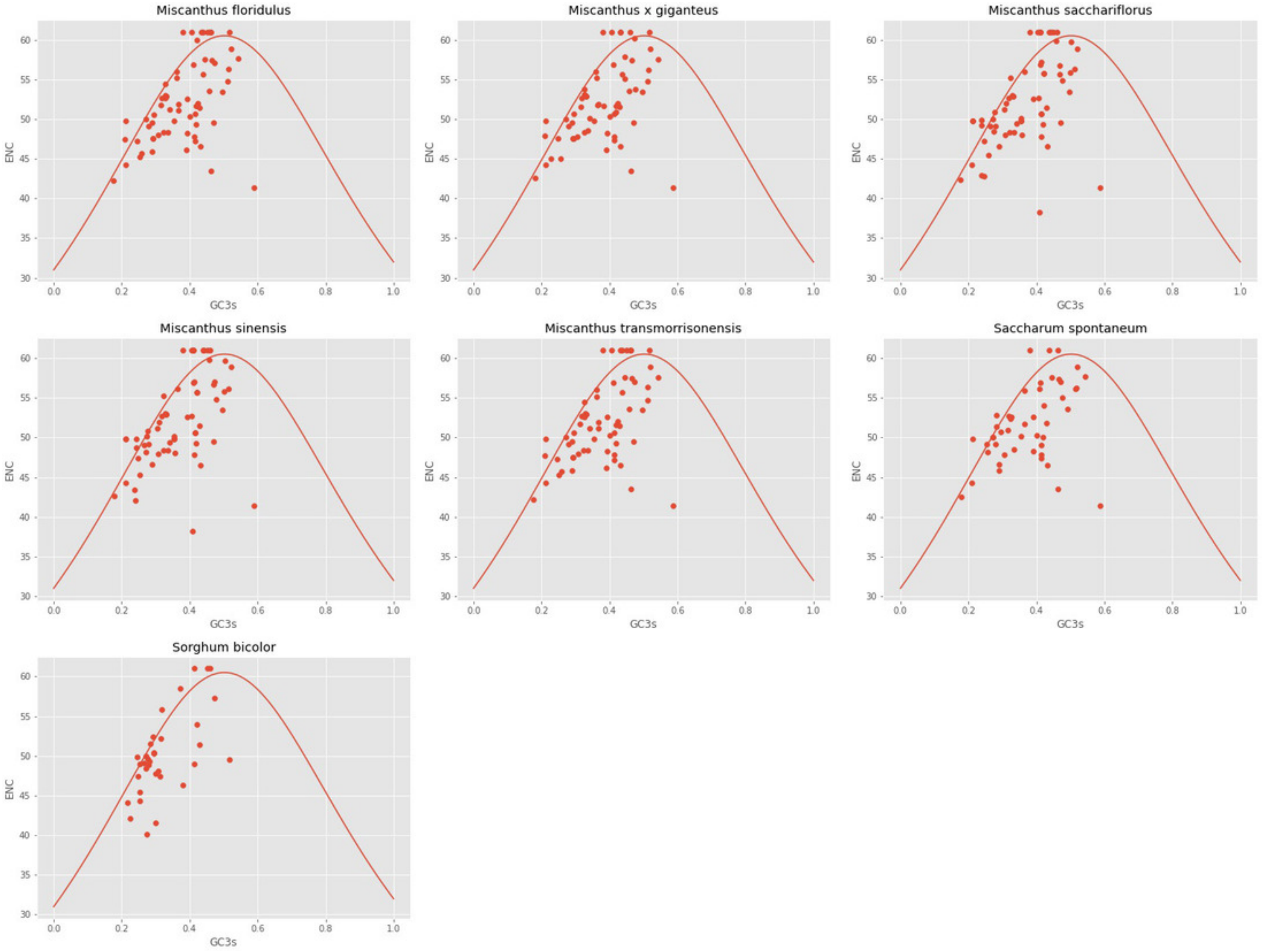


Figure 2

ENc-plot of chloroplast genomes of seven *Misacanthus* and related species



Miscanthus floridulus

Miscanthus x giganteus

Miscanthus sacchariflorus

Miscanthus sinensis

Miscanthus transmorrisonensis

Saccharum spontaneum

Sorghum bicolor

Figure 3

PR2-plot of chloroplast genomes of seven *Misacanthus* and related species

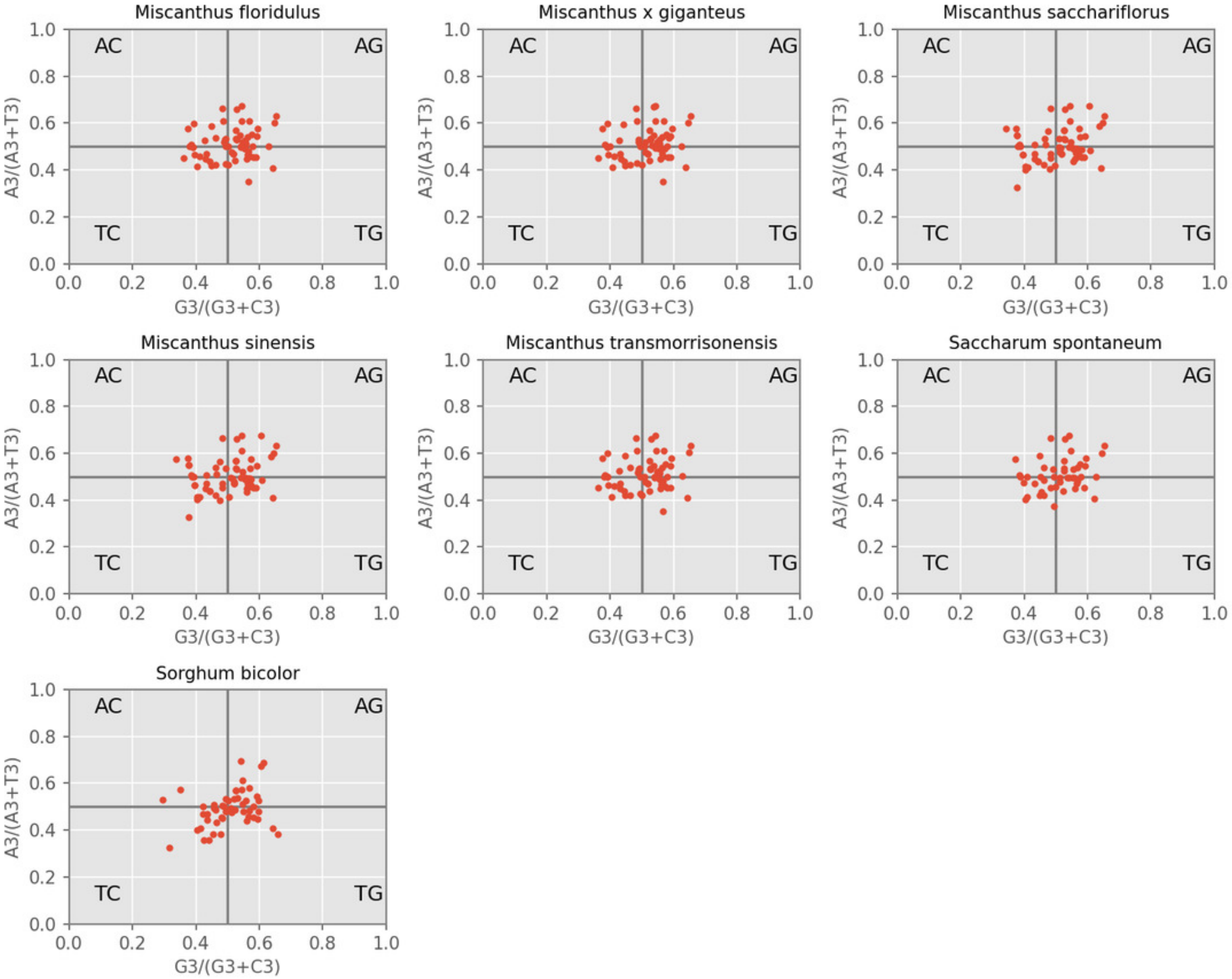


Figure 4

Neutrality plot of chloroplast genomes of seven *Misacanthus* and related species

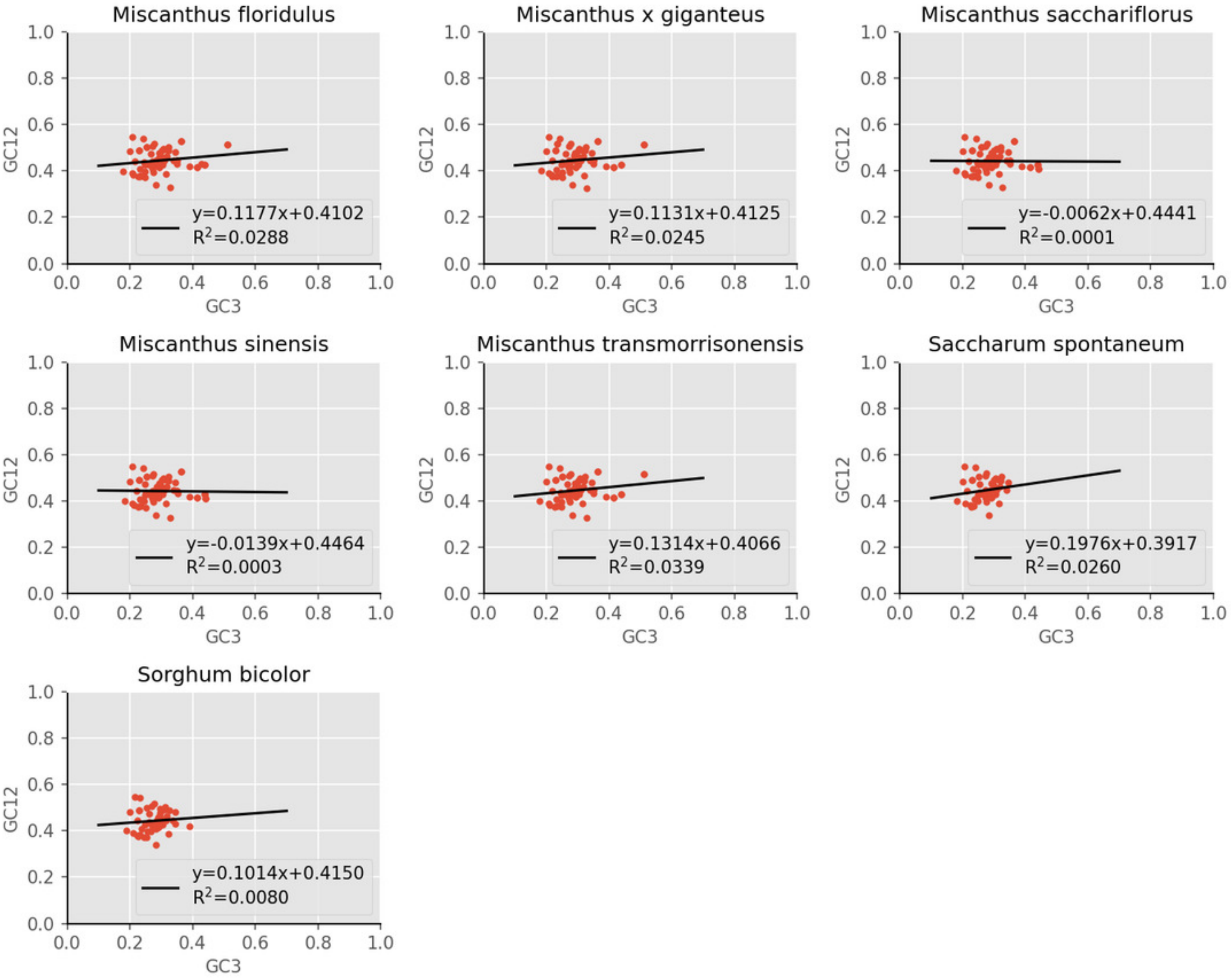


Figure 5

Correspondence analysis of chloroplast genomes of seven *Misacanthus* and related species