

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

Jiajing Sheng^{1,2}, Xuan She¹, Xiaoyu Liu², Jia Wang³, Zhongli Hu^{Corresp. 1}

¹ Wuhan University, Wuhan, China, 430072

² Nantong University, School of Life Sciences, Jiangsu Key Laboratory of Neuroregeneration, Co-innovation Center of Neuroregeneration, Nantong, China, 226001

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 chloroplast genomes of the seven represent species were preference to A/T bases and A/T-ending codons. In addition, detection of 21 common high-frequency codons and 4-11 optimal codons were detected in the seven chloroplast genomes. Combining the results of ENC-plot, PR2-plot and neutrality analysis revealed the codon usage pattern of the seven chloroplast genomes is influenced by multiple factors, which the main influencing factor was nature selection. Comparative analysis of the codon usage frequencies between the seven represent species and four model organisms suggested that *Arabidopsis thaliana*, *Populus trichocarpa* and *Saccharomyces cerevisiae* could be preferential considered as appropriate exogenous expression receptors. These results might not only provide important reference information for evolutionary analysis, but also deepened insight in improving the expression efficiency of exogenous gene in transgenic research by codon optimization.

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

Jiajing Sheng^{1,2}, Xuan She², Xiaoyu Liu¹, Jia Wang³, Zhongli Hu^{2*}

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 chloroplast genomes of the seven represent species were preference to A/T bases and A/T-ending codons. In addition, detection of 21 common high-frequency codons and 4-11 optimal codons were detected in the seven chloroplast genomes. Combining the results of ENC-plot, PR2-plot and neutrality analysis revealed the codon usage pattern of the seven chloroplast genomes is influenced by multiple factors, which the main influencing factor was nature selection. Comparative analysis of the codon usage frequencies between the seven represent species and four model organisms suggested that *Arabidopsis thaliana*, *Populus trichocarpa* and *Saccharomyces cerevisiae* could be preferential considered as appropriate exogenous expression receptors. These results might not only provide important reference information for evolutionary analysis, but also deepened insight in improving the expression efficiency of exogenous gene in transgenic research by codon optimization.

Introduction

Miscanthus species are C4 photosynthetic plants, which have been widely investigated as potential second-generation bio-energy crops (Barling 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* (*M. lutarioriparius*), *Miscanthus sinensis* (*M. sinensis*), *Miscanthus sacchariflorus* (*M. sacchariflorus*) and *Miscanthus floridulus* (*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 2016). Typically, cp genome possess the small sizes, conserved gene content and large copy numbers, which have been extensive used as valuable source for evolution analysis and plastid engineering (Amiryousefi et al. 2018; Ravi et al. 2007; Yan et al. 2019). The lack genomic resources of the *Miscanthus* species 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 function studies of *Miscanthus* and their related species. Benefited by the rapid development of chloroplast engineering, plasmid DNA have been transferred into the chloroplasts 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 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. 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 2014; Quax 2015; Tuller 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 synonymous codon bias in the chloroplast genomes have been proven in many higher plants, including *Paaceae* (Zhang Y 2012), *Cinnamomum camphorn* (Chen et al. 2017), *Strawberry* (Cheng et al. 2017) and *Solanum* (Zhang et al. 2018). However, the codon usage pattern of chloroplast genomes in *Miscanthus* and related species has not been fully elucidated.

Currently, the codon usage patterns in chloroplast genomes of seven *Miscanthus* and related species were systematically analyzed based on the previous published genome-wide data. In addition, the codon usage bias of these seven species was compared with the other four model species including *Populus trichocarpa*, *Escherichia coli*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae*. These results will provide insight into the codon usage patterns of the *Miscanthus* and related species, which indicating an important theoretical basis 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 chloroplast genomes of *Miscanthus floridulus* (NC_035750.1), *Miscanthus sacchariflorus* (NC_028720.1), *Miscanthus sinensis* (NC_028721.1), *Miscanthus x giganteus* (NC_035753.1), *Miscanthus 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 script 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. After filtration, the amount of CDS, the contents of 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 is the ratio of its actual frequency of utilization to the expected usage frequency without bias. The RFSC value refers to the proportion of the actually observational number of a codon in the number of all synonymous codons. The RSCU and RFSC were calculated by CodonW version 1.4 as described previously (Wang et al. 2020). If the RSCU value of a codon is equal to 1, there is no bias for the use of the codon. However, the RSCU value is >1 , which reflect significant codon usage bias and vice versa (Sharp 1989).

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 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 reflecting 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) and the ENc value 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 (<http://bioweb.pasteur.fr/seqana/interfaces/codonw.html>). Optimal codons were determined by Δ RSCU method. Specifically, the average RSCU values of the two amino acid groups were computed and then minus (Δ RSCU). 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 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*; *Saccharomyces cerevisiae*; *Populus trichocarpa* and *Arabidopsis thaliana* were downloaded from the Codon Usage Database (<http://www.kazusa.or.jp/codon/cgi-bin/showcodon>). 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 2013).

Analysis of ENc-plot

ENc value was applied to evaluate the codon bias of an single amino acid (Wright 1990). 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). 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. In addition, GC content at the third positions of codon was counted eliminating the Codon Met (ATG) and Trp (TGG). Meanwhile, 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 chloroplast genomes were counted by Python script. The gradient of the curve regression is 0, indicating that there is no impact of mutation pressure, while 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 variation of codon usage in the seven chloroplast genomes were investigated based on the correspondence analyses (COA) using CodonW (Anue 2019). The usage pattern of 59 codons (excluding Met, Trp and three termination codons) were compared and all genes can be embedded into a 59-dimension hyperspace, in which each dimension responding to the synonymous codon usage of the gene (Xiang 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 scipy package (<https://docs.scipy.org/doc/scipy/reference/stats.html>). 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 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 (Table 1). The GC contents of three positions of codons were calculated respectively, which are strongly positively correlated with the patterns of codon usage (Shackelton 2006). It was found that the contents of GC at all three sites (GC1, GC2 and GC3) and the average GC content were all less than 0.500, which indicated the seven chloroplast genomes prone to use A/T bases and A/T-ending codons (Table 1). Furthermore, the mean GC content of three sites in *M. floridulus*, *M. giganteus* and *M. transmorrisonensis* is the same (0.375) and in *M. sacchariflorus* and *M. sinensis* is the same (0.393) but in *Saccharum spontaneum* (0.391) and *Sorghum bicolor* (0.39) are slightly different (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 usage of GC content in these seven chloroplast genomes were similar and were biased towards A/T bases.

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).

RSCU and RFSC

The chloroplast 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 closely in the seven chloroplast 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 and indicated the vitally positive bias. Furthermore, the pattern of codon usage were summarized in the seven *Miscanthus* and related chloroplasts (Figure 1). Specially, 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 cp 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 Δ RSCU value in the two expression library were calculated and listed in Supplementary Table S2. According to the values of Δ RSCU, 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 chloroplast 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 chloroplast 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 chloroplast genomes is mainly influenced by natural selection and other factors, while mutation pressure play slightly roles.

Figure 2. ENc-plot of chloroplast genomes of seven *Misacanthus* and related species.

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. It is generally accepted the similar proportion of four bases of degenerate codon in gene or genome indicated the codon usage preference is fully influenced by mutation pressure (Sueoka 1999). 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 spontaneum* 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 chloroplast genomes was lopsided, indicating that the base composition of the seven chloroplast genomes is not only influenced by mutation pressure, but also by natural selection.

Figure 3. PR2-plot of chloroplast genomes of seven *Misacanthus* and related species.

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 ($r_1=0.157$, $r_2=0.168$, $r_3=0.128$, $r_4=0.127$, $r_5=0.161$, $r_6=0.155$, $r_7=0.140$), GC1 with GC3 ($r_8=0.092$, $r_9=0.079$, $r_{10}=0.055$, $r_{11}=0.049$, $r_{12}=0.1$, $r_{13}=0.242$, $r_{14}=0.202$) and GC2 with GC3 ($r_{14}=0.054$, $r_{15}=0.053$, $r_{16}=-0.014$, $r_{17}=0.063$, $r_{18}=-0.032$, $r_{19}=-0.032$), 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 chloroplast gene in the seven species are mainly affected by natural selection.

Figure 4. Neutrality plot of chloroplast genomes of seven *Misacanthus* and related species.

Correspondence analysis (COA)

COA is used to explore the variations of codon usage in the chloroplast 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 *Misacanthus* and related chloroplast 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 $45\% \leq GC\% < 60\%$. 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 2, axis 1 for *M. floridulus*, *M. sacchariflorus*, *M. sinensis*, *M. transmorrisonensis*, *Saccharum spontaneum* and *Sorghum bicolor* possessed a remarkable correlation with GC3s ($p \leq 0.01$), 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 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)

Figure 5. Correspondence analysis of chloroplast genomes of seven *Misacanthus* and related species.

Discussion

This study compared the codon usage patterns of the five *Misacanthus* species and two related species, which will help further improvement our understanding in 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 diversity 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 to optimize the codon of heterologous gene and select appropriate

heterologous gene expression receptor system, which will be of great significance to the study of genetic engineering and genetic evolution.

Analysis of base composition of codon revealed that the CDS of the seven *Miscanthus* and related chloroplast genomes tended to use A/T codon, which was consistent with the results of Zhang et al. (2012) on the 23 Poaceae chloroplast genomes (Zhang 2012). According to previous study, the GC3S value of dicotyledonous plants is often less than 50% (codon use preference A/T), which is different from the monocotyledonous plants with high GC3S value (GC3S value >50%, showing that codon use preference G/C) (Murray 1989). The results of RSCU value analysis showed that there were mostly A/T codon usage bias in the chloroplast genomes of the seven represent plants, which was consistent with the patterns in mostly higher plants (Shang 2011). According to the conclusion of Novembre (Novembre 2002), the mutation pressure will affect the composition of the third base of the synonymous codon excluding the influence of natural selection. The results of the preferred codon ending with A/T in this study indicated that the mutation pressure of G/C > the mutation pressure of A/T in the chloroplast genomes of the seven species. 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 1993). The neutrality plot in this study revealed that there is weak correlation between GC12 and GC3 and the composition of the first two bases was different from the third base of the codon, which demonstrated that the codon usage patterns of the seven chloroplast genomes are mainly influenced by natural selection. In addition, combining the results of ENC-plot and PR2-plot suggested that the codon usage bias of the seven chloroplast genomes were affected by multiple factors, which the main influencing factor was nature selection.

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

This study conducted a comprehensive comparative analysis on codon usage pattern at the chloroplast genome-wide level of seven *Miscanthus* and related species. These results will promote our understanding in evolution analysis, the selection of appropriate heterologous gene expression receptor system and the optimization of codon components suitable for gene

expression, finally providing a theoretical basis for building a stable and efficient gene expression system in *Miscanthus* or other crops.

Conclusions

The codon usage patterns of chloroplast 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 represent 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 elected in the seven chloroplast genomes. Furthermore, the analysis of codon usage frequencies between the seven represent 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, combining the results of ENC-plot, PR2-plot and neutrality analysis revealed the codon usage pattern of the seven chloroplast genomes is influenced by multiple factors, in which the dominant influencing factor was nature selection. These results in the study might not only provide important reference information for evolutionary analysis, but also deepened insight 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.
- Anue MR, Xueli, S., Chunzhen, C., & Zhongxiong, L. 2019. Analysis of codon usage pattern of banana basic secretory protease gene. *Plant Diseases and Pests* 10:1-9.
- Barling A, Swaminathan, K., Mitros, T., James, B. T., Morris, J., Ngamboma, O., ... & Moose, S. P. 2013. A detailed gene expression study of the *Miscanthus* genus reveals changes in the transcriptome associated with the rejuvenation of spring rhizomes. *BMC genomics* 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 *Miscanthus*. 6:646-660.
- Chen C, Zheng Y, Liu S, Zhong Y, Wu Y, Li J, Xu LA, and Xu M. 2017. The complete chloroplast genome of *Cinnamomum camphora* and its comparison with related Lauraceae species. *PeerJ* 5:e3820. 10.7717/peerj.3820
- Cheng H, Li J, Zhang H, Cai B, Gao Z, Qiao Y, and Mi L. 2017. The complete chloroplast genome sequence of strawberry (*Fragaria x ananassa* Duch.) and comparison with related species of Rosaceae. *PeerJ* 5:e3919. 10.7717/peerj.3919
- Clark LV, Stewart JR, Nishiwaki A, Toma Y, Kjeldsen JB, Jorgensen U, Zhao H, Peng J, Yoo JH, Heo K, Yu CY, Yamada T, and Sacks EJ. 2015. Genetic structure of *Miscanthus sinensis* and *Miscanthus sacchariflorus* in Japan indicates a gradient of bidirectional but asymmetric introgression. *J Exp Bot* 66:4213-4225. 10.1093/jxb/eru511

399 Ge C, Liu X, Liu S, Xu J, Li H, Cui T, Yao Y, Chen M, Yu W, and Chen C. 2017. *Miscanthus*
400 sp.: Genetic Diversity and Phylogeny in China. *Plant Molecular Biology Reporter* 35:600-610.
401 10.1007/s11105-017-1048-9

402 Havaux M, Lutz C, and Grimm B. 2003. Chloroplast membrane photostability in chlP transgenic
403 tobacco plants deficient in tocopherols. *Plant Physiol* 132:300-310. 10.1104/pp.102.017178

404 Jarvis P, and López-Juez E. 2013. Biogenesis and homeostasis of chloroplasts and other plastids.
405 *Nature Reviews Molecular Cell Biology* 14:787-802. 10.1038/nrm3702

406 Khodakovskaya M, McAvoy R, Peters J, Wu H, and Li Y. 2006. Enhanced cold tolerance in
407 transgenic tobacco expressing a chloroplast omega-3 fatty acid desaturase gene under the control
408 of a cold-inducible promoter. *Planta* 223:1090-1100. 10.1007/s00425-005-0161-4

409 Kwak SY, Lew, T. T. S., Sweeney, C. J., Koman, V. B., Wong, M. H., Bohmert-Tatarev, K., ...
410 & Strano, M. S. 2019. Chloroplast-selective gene delivery and expression in planta using
411 chitosan-complexed single-walled carbon nanotube carriers. *Nature nanotechnology* 14(5):447-
412 455.

413 Murray EE, Lotzer, J., & Eberle, M. 1989. Codon usage in plant genes. *Nucleic Acids Research*
414 17: 477-498.

415 Nielsen AZ, Mellor, S. B., Vavitsas, K., Wlodarczyk, A. J., Gnanasekaran, T., Perestrello Ramos
416 H de Jesus, M., ... & Jensen, P. E. 2016. Extending the biosynthetic repertoires of cyanobacteria
417 and chloroplasts. *The Plant Journal* 87:87-102.

418 Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon
419 usage bias. *Molecular biology and evolution* 19 1390-1139.

420 Pan LL, Wang, Y., Hu, J. H., Ding, Z. T., & Li, C. 2013. Analysis of codon use features of
421 stearoyl-acyl carrier protein desaturase gene in *Camellia sinensis*. *Journal of theoretical biology*
422 334:80-86.

423 Pop C, Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S., & Koller, D. 2014.
424 Causal signals between codon bias, mRNA structure, and the efficiency of translation and
425 elongation. *Molecular systems biology* 10:770.

426 Quax TE, Claassens, N. J., Söll, D., & van der Oost, J. 2015. Codon bias as a means to fine-tune
427 gene expression. *Molecular cell* 59:149-161.

428 Ravi V, Khurana JP, Tyagi AK, and Khurana P. 2007. An update on chloroplast genomes. *Plant*
429 *Systematics and Evolution* 271:101-122. 10.1007/s00606-007-0608-0

430 Romero H, Zavala, A., & Musto, H. . 2000. Codon usage in *Chlamydia trachomatis* is the result
431 of strand-specific mutational biases and a complex pattern of selective forces. . *Nucleic Acids*
432 *Research* 28:2084-2090.

433 Ruf S, Hermann, M., Berger, I. J., Carrer, H., & Bock, R. 2001. Stable genetic transformation of
434 tomato plastids and expression of a foreign protein in fruit *Nature biotechnology* 19: 870-875.

435 Shackelton LA, Parrish, C. R., & Holmes, E. C. 2006. Evolutionary basis of codon usage and
436 nucleotide composition bias in vertebrate DNA viruses. *Journal of molecular evolution* 62: 551-
437 563.

438 Shang M, Liu, F., Hua, J., & Wang, K. . 2011. Analysis on codon usage of chloroplast genome
439 of *Gossypium hirsutum*. *Scientia Agricultura Sinica* 44:245-253.

440 Sharp PM, & Devine, K. M. . 1989. Codon usage and gene expression level in *Dictyostelium*
441 *discoideum*: highly expressed genes do [prefer [optimal codons. . *Nucleic Acids Research* 17:
442 5029-5040.

443 Sharp PM, & Li, W. H. . 1986. An evolutionary perspective on synonymous codon usage in
444 unicellular organisms. *Journal of molecular evolution* 24:28-38.

445 Sharp PM, Stenico, M., Peden, J. F., & Lloyd, A. T. . 1993. Codon usage: mutational bias,
446 translational selection, or both? *Biochemical Society Transactions* 21:835-841.

447 Sheng J, Hu X, Zeng X, Li Y, Zhou F, Hu Z, Jin S, and Diao Y. 2016. Nuclear DNA content in
448 *Miscanthus* sp. and the geographical variation pattern in *Miscanthus lutarioriparius*. *Sci Rep*
449 6:34342. 10.1038/srep34342

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

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

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

457 Tuller T, Waldman, Y. Y., Kupiec, M., & Ruppin, E. . 2010. Translation efficiency is determined
458 by both codon bias and folding energy. *Proceedings of the National Academy of Sciences*
459 107:3645-3650.

460 Vermerris W. 2008. *Miscanthus*: genetic resources and breeding potential to enhance bioenergy
461 production. In *Genetic improvement of bioenergy crops* (pp. 295-308). Springer, New York, NY.
462 Springer New York:295-308.

463 Wang Z, Xu B, Li B, Zhou Q, Wang G, Jiang X, Wang C, and Xu Z. 2020. Comparative analysis
464 of codon usage patterns in chloroplast genomes of six Euphorbiaceae species. *PeerJ* 8:e8251.
465 10.7717/peerj.8251

466 Wright F. 1990. The ‘effective number of codons’ used in a gene. *Gene* 87: 23-29.

467 Xiang H, Zhang, R., Butler III, R. R., Liu, T., Zhang, L., Pombert, J. F., & Zhou, Z. . 2015.
468 Comparative analysis of codon usage bias patterns in microsporidian genomes. *PloS one*
469 10:e0129223.

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

472 Zhang J, Yan J, Zhang Y, Ma X, Bai S, Wu Y, Dao Z, Li D, Zhang C, Zhang Y, You M, Yang F,
473 and Zhang J. 2013. Molecular insights of genetic variation in *Erianthus arundinaceus* populations
474 native to China. *PLoS One* 8:e80388. 10.1371/journal.pone.0080388

475 Zhang R, Zhang L, Wang W, Zhang Z, Du H, Qu Z, Li XQ, and Xiang H. 2018. Differences in
476 Codon Usage Bias between Photosynthesis-Related Genes and Genetic System-Related Genes of

477 Chloroplast Genomes in Cultivated and Wild Solanum Species. *Int J Mol Sci* 19.
 478 10.3390/ijms19103142

479 Zhang Y, Nie, X., Jia, X., Zhao, C., Biradar, S. S., Wang, L., ... & Weining, S. 2012. Analysis of
 480 codon usage patterns of the chloroplast genomes in the Poaceae family *Australian Journal of*
 481 *Botany* 60: 461-470.

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

484 Zhou M, Tong, C., & Shi, J. 2007. Analysis of codon usage between different poplar species.
 485 *Journal of Genetics and Genomics* 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	<i>Miscanthus floridulus</i>	<i>Miscanthus giganteus</i>	<i>Miscanthus sacchariflorus</i>	<i>Miscanthus sinensis</i>	<i>Miscanthus transmorrissonensis</i>	<i>Saccharum spontaneum</i>	<i>Sorghum bicolor</i>
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

3

Table 2(on next page)

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

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

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

2

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
<i>Miscanthus floridulus</i>	-0.65**	0.507**	-0.133	0.598**	-0.152	0.073	0.098	0.28**	0.639**	0.159	-0.315**
<i>Miscanthus x giganteus</i>	0.016	0.053	-0.104	-0.127	-0.038	0.044	0.078	0.09	-0.018	0.223**	0.057
<i>Miscanthus sacchariflorus</i>	0.66**	-0.463**	0.171*	-0.708**	0.176*	0.044	0.086	-0.43**	-0.7**	-0.154	0.118
<i>Miscanthus sinensis</i>	0.663**	-0.469**	0.18*	-0.711**	0.17*	0.043	0.083	-0.423**	-0.708**	-0.154	0.123
<i>Miscanthus transmorrisonensis</i>	0.647**	-0.509**	0.129	-0.604**	0.147	-0.076	-0.099	-0.285**	-0.641**	-0.152	0.318**
<i>Saccharum spontaneum</i>	0.693**	-0.479**	0.201*	-0.691**	0.193	0.02	-0.012	-0.217*	-0.667**	-0.19	0.197*
<i>Sorghum bicolor</i>	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*

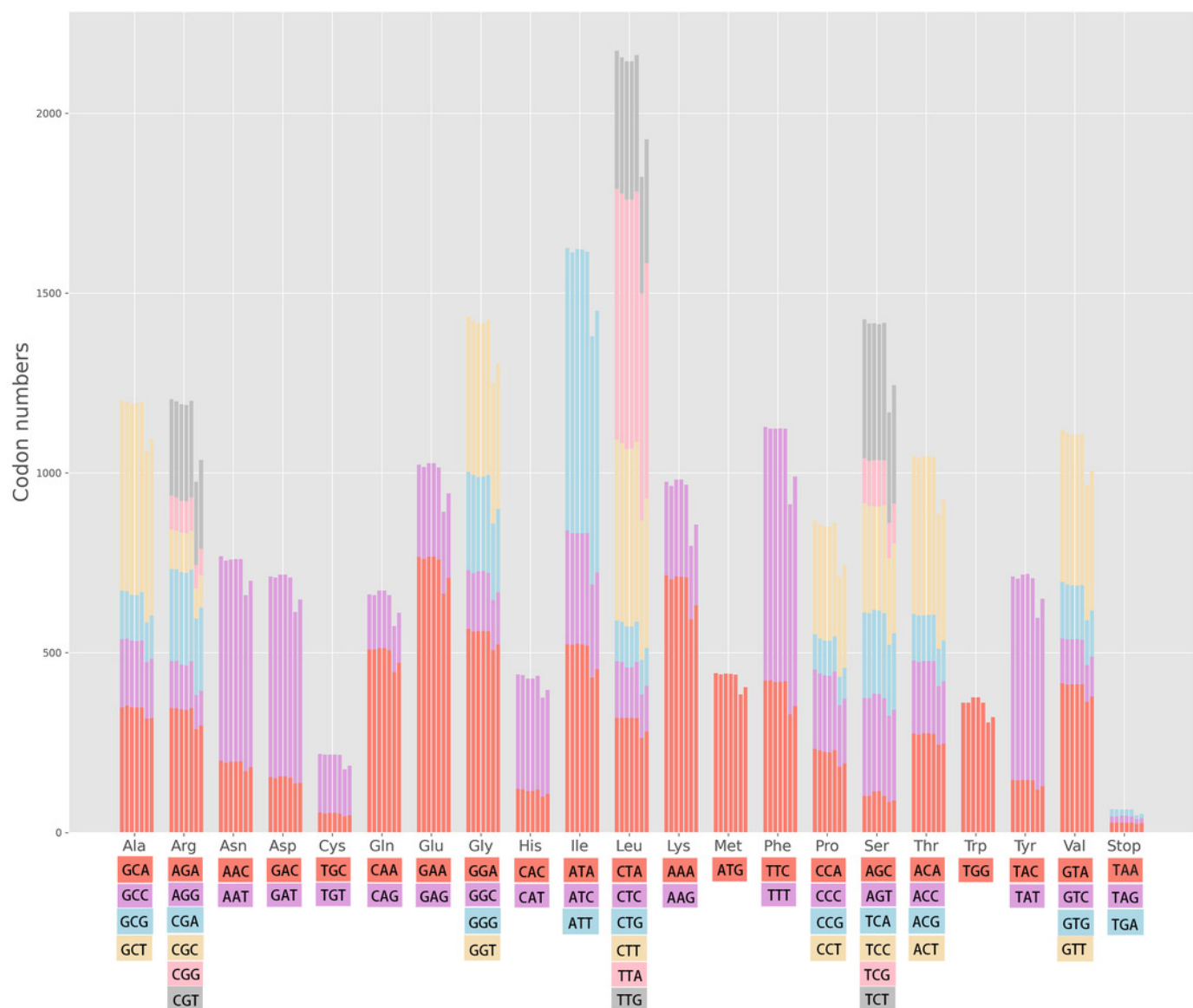


Figure 2

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

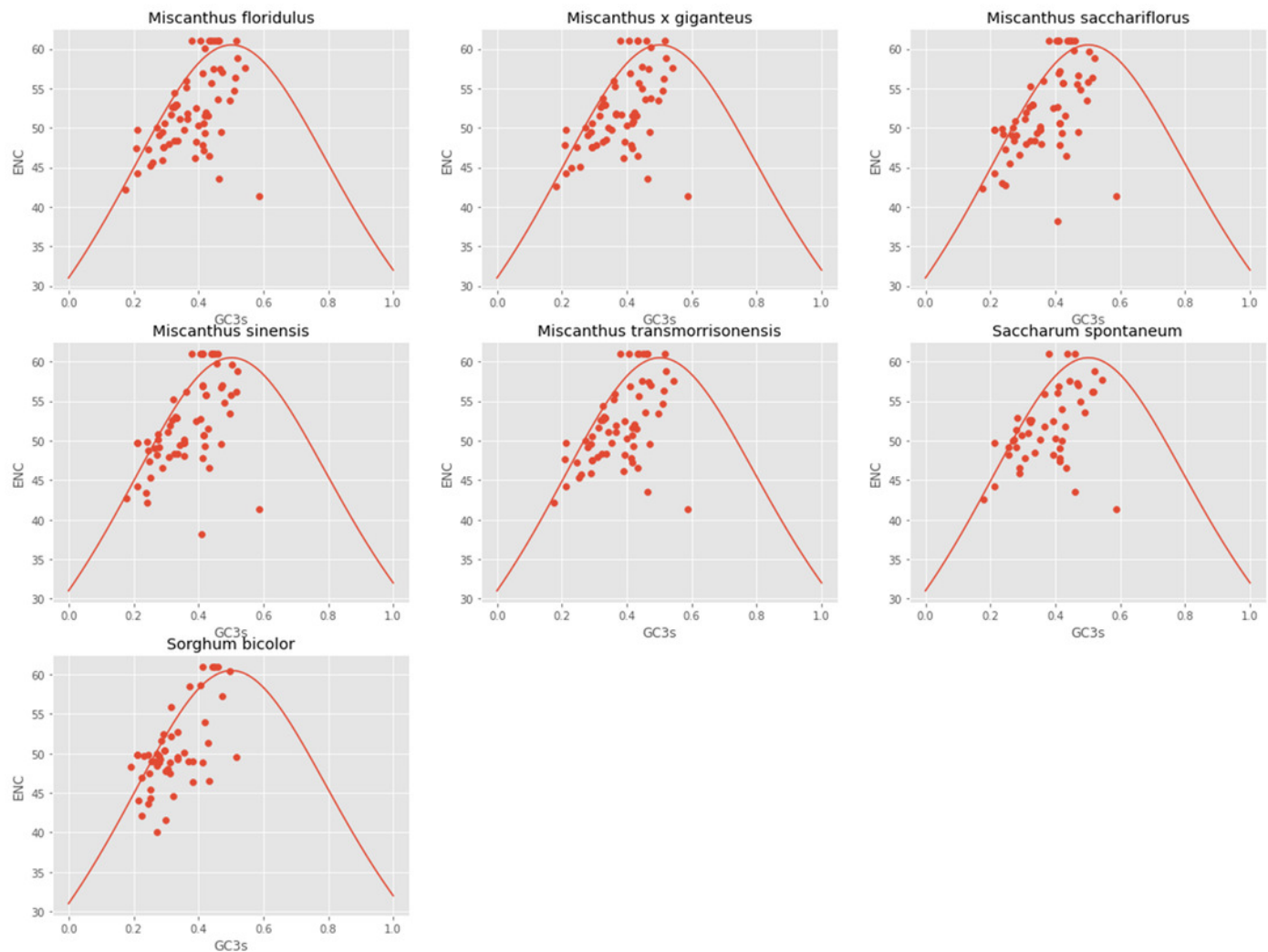


Figure 3

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

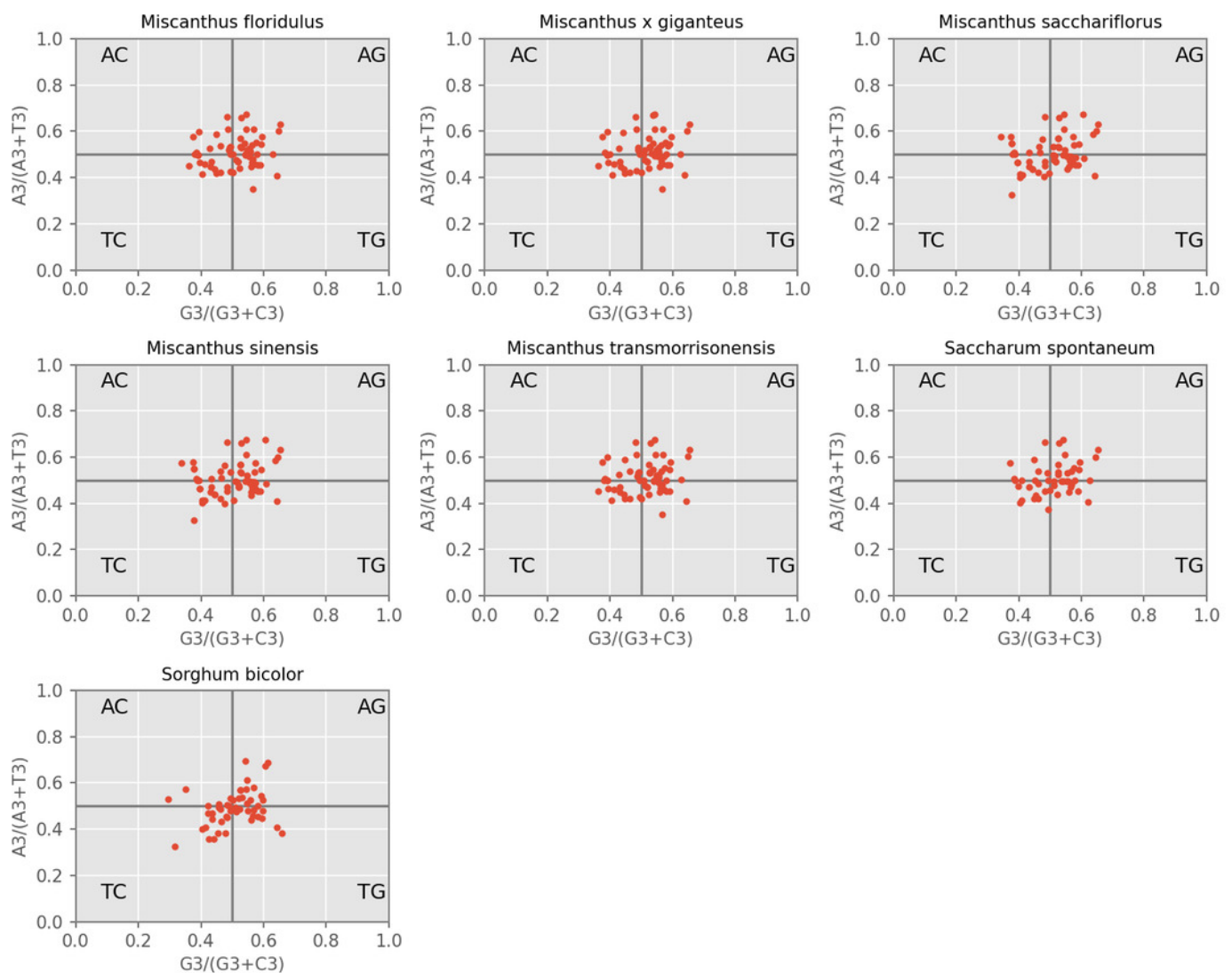


Figure 4

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

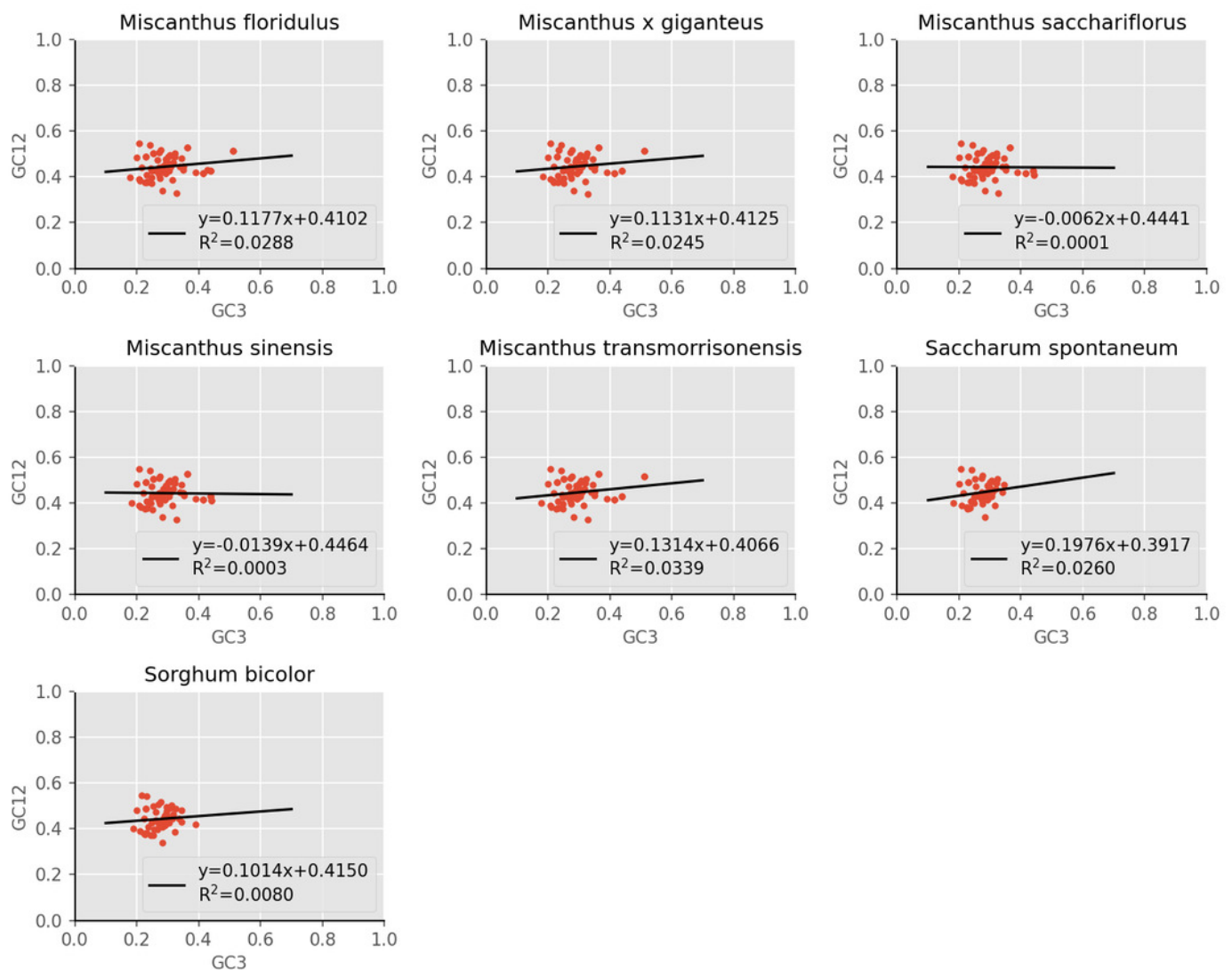


Figure 5

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

