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

Jiajing Sheng1, 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 a changing environment. In the present study, the codon usage patterns of five different *Miscanthus* species and that of two other related species were systematically analyzed. The results indicated that the chloroplast genomes of the seven representative species showed preference for A/T bases and A/T-ending codons. In addition, 21 common high-frequency codons and 4 to 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 natural 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 heterologous expression hosts. 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) is a C4 photosynthetic plant, which have been widely investigated as a potential second-generation bio-energy crop (Barling et al. 2013). The genus Miscanthus includes approximately 20 species, which could be classified into Miscanthus clades

40    and Triarrhena clades (Ge et al. 2017). China is the biological diversity center of *Miscanthus*
41    species, of which *Miscanthus lutarioriparius L.Liou* (*M. lutarioriparius*), *Miscanthus sinensis*
42    *Andersson* (*M. sinensis*), *Miscanthus sacchariflorus (Maxim.) Nakai* (*M. sacchariflorus*) and
43    *Miscanthus floridulus (Lab.) Warb. ex Schum. et Laut* (*M. floridulus*) are the four most widely
44    distributed species. In addition to being bioenergy plants, *Miscanthus* species possess extensive
45    breeding values due to their extremely advantageous agricultural characteristics, such as high
46    photosynthetic efficiency, cold tolerance and an extensive ability to adapt to environmental
47    change (Vermerris 2008). Currently, the research focus of *Miscanthus* species is to utilize it as a
48    promising genetic resource (Clark et al. 2015; Zhang et al. 2013). The more diverse genetic
49    resources available, the more likely it is that scientific research will be able to comprehend the
50    adaptation, evolution and utilization of these significant economic crops.

51        Chloroplasts (cp) are key plastids involved in multifunctional processes of the plant cell,
52    including photosynthesis, carbon fixation, starch storage, nitrogen metabolism, fatty acid and
53    nucleic acid synthesis (Jarvis & López-Juez 2013; Nielsen et al. 2016). Typically, cp genomes
54    (cpDNAs) possess a small size, conserved gene content and large copy numbers, which have
55    been extensively used as valuable source for evolution analysis and plastid engineering
56    (Amiryousefi et al. 2018; Ravi et al. 2008; Yan et al. 2019). The lack of genomic resources of the
57    *Miscanthus* species has hindered the adequate understanding of their diversity traits (Chae et al.
58    2014; Sheng et al. 2016) The low expression efficiency of an exogenous gene may limit the
59    research progress on the functional studies of *Miscanthus* and their related species (Wu et al.
60    2021). Stemming from the rapid development of cp engineering, plasmid DNA have been
61    transferred into the cpDNA of a variety of plants, such as *Nicotiana tabacum*, *Manihot esculenta*
62    *Crantz* and *Eruca sativa Mill* (Havaux et al. 2003; Khodakovskaya et al. 2006; Kwak et al.
63    2019). Recently, the cpDNA of some *Miscanthus* species have been made available in the
64    National Center for Biotechnology Information (NCBI) database (Sheng et al. 2021). These
65    complete cpDNA sequences of *Miscanthus* species can be used for studying population genetics,
66    evolution analysis and plastid engineering (Amiryousefi et al. 2018; Yan et al. 2019).

67        Codon usage bias refers to the variations on the usage frequencies of synonymous codons
68    (Plotkin & Kudla 2011). The pattern of codon usage could be caused by multi-factors during the
69    process of genome and gene evolution, including natural selection, compositional mutation
70    mode, translational selection, gene length, tRNA abundance and mRNA secondary structure (Liu
71    et al. 2004; Pop et al. 2014; Quax et al. 2015; Tuller et al. 2010). The studies of codon preference
72    can not only reveal the evolutionary rules between genes in a species or related species, but also
73    improve the expression efficiency of exogenous sequences in transgenic research by codon
74    optimization. Recently, the applicability of codon optimization in the cpDNA have been
75    determined for many vascular plants, including *Poaceae*(Gramineae)(Zhang et al. 2012),
76    *Cinnamomum camphorn (L.) presl* (Camphor tree ) (Chen et al. 2017), *Fragaria × ananassa*
77    *Duch* (strawberry) (Cheng et al. 2017) and *Solanum tuberosum L.* (potato) (Zhang et al. 2018).
78    However, the codon usage pattern of cpDNA in *Miscanthus* and related species has not been
79    fully elucidated.

80        In this study, the codon usage patterns of the cpDNA of seven *Miscanthus* and related
81  species, including *M. sinensis, Miscanthus transmorrisonensis Hayata (M. transmorrisonensis),*
82  *M. floridulus, M. sacchariflorus, Miscanthus x giganteus (M. x giganteus), Sorghum bicolor* and
83  *Saccharum spontaneum* were systematically analyzed based on the previously published
84  genome-wide data. Among them, *M. sinensis, M. transmorrisonensis* and *M. floridulus* belong to
85  the Miscanthus clades under the *Miscanthus* species, *M. sacchariflorus* belongs to the Triarrhena
86  clades under the *Miscanthus* species, and *M. x giganteus* is a natural triploid hybrid. According
87  to our previous studies, *Miscanthus* are most closely related to *Sorghum bicolor* (L.) *Moench* and
88  *Saccharum L* (Sheng et al. 2021; Sheng et al. 2017). In addition, the codon usage bias of these
89  seven species was compared with the other four model species including *Populus trichocarpa*
90  *Torr & Gray, Escherichia coli, Arabidopsis thaliana (L.) Heynh* and *Saccharomyces cerevisiae.*
91  *Miscanthus* are not only potential bio-energy crops, but also forms an excellent breeding
92  resources. So, it is of interest to understand the codon usage of *Miscanthus* for better utilization
93  of *Miscanthus* and related resources as germplasm resources. Here, we revealed the codon usage
94  patterns of the *Miscanthus* and related species and determined optimal codons for cpDNA
95  genetic engineering. The results in the current study will not only provide insight into genetic
96  evolution studies, but also provide a reference for selecting appropriate heterologous expression
97  hosts to improve the gene expression of *Miscanthus* plants by optimizing codon.
98

## Materials & Methods

### Genomes and sequences selection

101  The complete cpDNAs of *M. floridulus* (NC_035750.1), *M. sacchariflorus* (NC_028720.1), *M.*
102  *sinensis* (NC_028721.1), *M. x giganteus* (NC_035753.1), *M. transmorrisonensis*
103  (NC_035752.1), *Sorghum bicolor* (NC_008602), *Saccharum spontaneum* (NC_034802.1) with
104  gene annotation were downloaded from the NCBI GeneBank database. The number of raw
105  sequence coding for amino acids in protein (CDS) of above seven species was 106, 122, 122,
106  106, 106, 84 and 76 respectively (Table 1). To avoid sampling bias, the CDS sequences were
107  screened from genome-wide data by python scripts (https://github.com/shexuan/codon_analysis)
108  according to the following principles: (1) CDS contains initiation codon (ATG), termination
109  codons (TAA, TAG or TGA) and without intermediate stop codons in the sequences; (2) the
110  number of bases in each CDS must be the fold of three (3) the length of the sequence of CDS
111  should be≥300 bp (Wright 1990; Zhang 2007). After filtration, the CDS number, the base
112  composition at the first/second/third site of codons (GC1/GC2/GC3) and average GC, as well as
113  the total amino acids encoded by each CDS were calculated.
114

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

117  Relative synonymous codon usage value of a codon (the number of codon occurrences in a gene
118  divided by the number of codon appearances expected under the same codon usage) is the ratio

119   of its actual frequency of utilization to the expected usage frequency without bias. The RSCU
120   was calculated as Eq. (1):

121

122   $$\text{RSCU} = \frac{x_{ij}}{\sum_{j}^{N_i} x_{ij}} n_i \tag{1}$$

123   Where $x_{ij}$ represents the frequency of codon j encoding the i th amino acid, and $n_i$ represents the
124   number of synonymous codon encoding the i th amino acid (Sharp, 1986). If the RSCU value of
125   a codon is equal to 1, the codon is used without bias, whereas a RSCU value greater than 1
126   reflects a significant codon usage bias (Sharp, 1987).
127         The RFSC value refers to the proportion of the actually observed number of a codon in
128   the number of all synonymous codons. The RFSC were calculated using Eq. (2):

129   $$\text{RFSC} = \frac{x_{ij}}{\sum_{j}^{N_i} x_{ij}} \tag{2}$$

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

134

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

149

150   **Comparative analysis of codon usage frequency**
151   The ratio of codon usage frequency is one indicator of codon usage bias among species. To
152   further explore the codon usage patterns in the seven species of *Miscanthus* and their relatives,
153   codon usage bias data of four model species including *Escherichia coli* ([http://ww](http://ww)
154   w.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=199310);    *Saccharomyces    cerevisiae*
155   (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=4932);    *Populus    trichocarpa*
156   (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=3694) and *Arabidopsis thaliana*

157 (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=3702), which have been used as
158 the most common heterologous expression hosts were downloaded from the Codon Usage
159 Database. Subsequently, the codon usage frequencies of the seven species in this study were
160 compared with the above four model organisms. When the ratio is $\geq 2$ or $\leq 0.5$, it suggests that
161 the codon bias difference between the two organisms is significant, whereas other values outside
162 of this range represent a lack of significance (Pan et al. 2013).
163
164 **Analysis of ENC-plot**
165 The proportion of G and C content at the third position of a codon to the total number of gene
166 bases are defined as GC3s. ENC-plot is plotted with ENC values as ordinate and GC3 value as
167 abscissa, which can be used to analyze the codon usage characteristics of each gene and to
168 explore the relevance between gene base component and codon preference (Wright 1990). ENC
169 values are located on or near the expected curve, when mutation pressure plays a key role in the
170 formation of codon usage patterns. Conversely, when the use of a codon is constrained by natural
171 selection, the ENC value will be well below the prospective curve (Wright 1990).
172
173 **PR2-plot analysis**
174 In G3/(G3+C3) as the abscisic and A3/(A3+T3) as the ordinate graphic mapping (PR2-plot),
175 which is performed to explore the composition of the four bases at the third nucleotide position
176 of each codon (Sueoka 1999; Sueoka 1995). The pattern of splashes around the central spot
177 (A=T, C=G) indicate the extent and orientation of the base offset.
178
179 **Analysis of Neutrality plot**
180 Neutrality analysis is used to exploring the degree of impact between natural selection and
181 mutation pressure on the mode of codon usage (Sueoka 1988). GC12 indicates the mean GC
182 content at the first and second nucleotide positions of the codon, while GC3 represents the GC
183 content of the third site. GC content at the third nucleotide position of a codon was counted post
184 eliminating the Codon Met (ATG) and Trp (TGG). GC3 was counted post the eliminating of the
185 three stop codons (TAA, TAG and TGA) and three codons (ATT, ATC and ATA) of Ile (Sueoka
186 1988). Both the GC12 and GC3 values of the seven cpDNAs were counted by Python scripts
187 (https://github.com/shexuan/codon_analysis). If the gradient of the curve regression is 0,
188 indicating that there is no impact of mutation pressure. Gradient 1 represents complete neutrality,
189 which describes that codon usage preference is completely influenced by mutation pressure
190 (Sueoka 1988).
191
192 **Correspondence analysis of codon usage**
193 The variations of codon usage in the seven analyzed cpDNAs were investigated based on the
194 correspondence analyses (COA) using CodonW (Anue et al. 2019). The usage patterns of 59
195 codons (exincluding Met, Trp and three termination codons) were compared and all genes can be
196 embedded into a 59-dimension hyperspace, in which each dimension corresponds to the

197 synonymous codon usage of the gene (Xiang et al. 2015). Therefore, the major trends (Axis 1) of
198 these axes in the 59-dimensional hyperspace can be used to determine the maximum fraction of
199 genetic variation, indicating the major sources of codon usage variation. In addition, according to
200 the results of COA, the correlation index between Axis1 and codon usage exponent, including
201 the GC content of codons, GC3s, codon adaptation index (CAI) and the total numbers of amino
202 acids in the encoded polypeptide (L_aa) were computed by python scripts package
203 (https://github.com/shexuan/ codon_analysis). CAI value is widely applied to assess gene
204 expression levels, ranging from 0 to 1. Specifically, the larger the CAI value is, the stronger the
205 codon usage preference is, and vice versa (Sharp & Li 1986).
206

## Results

207
208

### Characteristics of codon usage bias

209
210

**Analysis of codon base composition**
211
212 The screened CDSs numbers processed by Python scripts are 65, 64, 64, 64, 64, 48 and 52 for *M.*
213 *floridulus*, *M. giganteus*, *M. sacchariflorus*, *M. sinensis*, *M. transmorrisonensis*, *Saccharum*
214 *spontaneum* and *Sorghum bicolor* respectively. In addition, the GC contents of three positions of
215 codons (GC1, GC2, GC3) were calculated respectively (Table 1). It was found that the contents
216 of GC at all three sites and the average GC content (GC123) were all less than 0.5, which
217 indicated the seven analyzed cpDNAs were prone to use A/T bases and A/T-ending codons
218 (Table 1). Specifically, the mean GC content of three sites in *M. floridulus*, *M. giganteus* and *M.*
219 *transmorrisonensis* was determined to equal the same value (0.375), as well as a matching value
220 determined for *M. sacchariflorus* and *M. sinensis* (0.393), but slightly different in  *Saccharum*
221 *spontaneum* (0.391) and *Sorghum bicolor* (0.39) (Table 1). Furthermore, the distribution trend of
222 GC content was GC1 > GC2 >GC3, indicating that GC was not evenly distributed in the three
223 positions of a codon of the seven assessed species. In summary, the codon usages of GC content
224 in these seven cpDNAs were similar and were biased towards A/T bases.
225
226 **Table 1** Genomic features of cp genomes of the seven Miscanthus and related species (the total
227 number of amino acids: L_aa ; the GC content at the first, second and third codon positions: GC1,
228 GC2and GC3; average GC at three locations: GC123).
229

**RSCU and RFSC**
230
231 The cp genomes of the seven *Miscanthus* and related species have 30 common codons (RSCU >
232 1) with 28 codons ending with the nucleotides A/T (93.3%) (Table S1). Therefore, almost the
233 majority of codons of the seven plants species analyzed (RSCU > 1) are likely to end with A/T.
234 The variation ranges in the RSCU values were close in the seven cpDNAs, i.e., 0.31–1.93 in *M.*
235 *floridulus*, *M. giganteus* and *M. transmorrisonensis*, 0.32-1.94 in *M. sacchariflorus* and *M.*
236 *sinensis*, 0.32- 2.01 in *Saccharum spontaneum* and 0.33-2.04 in *Sorghum bicolor*, respectively

237  (Table S1). In addition, the maximum and the minimum RSCU values belonged to TTA and
238  CTG which encode Leu, indicating the vitally positive bias. Furthermore, the pattern of codon
239  usage were summarized in the seven *Miscanthus* and related species (Figure 1). Specifically, the
240  high-frequency codons of seven *Miscanthus* and related species possess strong common base and
241  share a total of 21 high-frequency codons (Table S1). Furthemore, *Saccharum spontaneum* and
242  *Sorghum bicolor* were determined to possess two more high-frequency codons, specifically CGT
243  and TAA, than the other five *Miscanthus* species.

244

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

249

250  **Determination of optimal codons**
251  The ENC values of each CDS were ranked and 10% of genes from both ends were selected to
252  establish high and low expression gene banks respectively. The RSCU values and ΔRSCU
253  values in the two expression libraries were calculated and are listed in Supplementary Table S2.
254  According to the values of ΔRSCU, the optimal codons in the seven assessed species were
255  determined as follows (Table 2).

256

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

258

259  **Codon usage frequency**
260  The codon usage frequencies of the seven cpDNAs were compared with four model species
261  including *Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Populus*
262  *trichocarpa* (Table S3). The Results of our analyses indicated that there is little divergence in the
263  codon usage frequencies among the seven assessed plant species with *Saccharomyces cerevisiae*,
264  *Arabidopsis thaliana* and *Populus trichocarpa*, have 9 to11 (accounting for 14.1%–17.2% of
265  total codons), 13 to 14 (20.3%–21.9%), 12 to 13 (18.8%–20.3%) different codons, respectively
266  (Table S3). However, the codon usage frequencies of the seven species with *Escherichia coli*
267  were comparatively higher (27 different codons). The results indicated that the codon frequency
268  difference between *Miscanthus* species and *Arabidopsis thaliana*, *Populus trichocarpa* and
269  *Saccharomyces cerevisiae* was the lowest, while was the largest with *Escherichia coli*. Based on
270  above results, it was optimal to select *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and
271  *Populus trichocarpa* as heterologous expression hosts for *Miscanthus* and related species.
272  Furthermore, the results indicated that TAG is a different termination codon in usage frequency
273  when comparing the seven assessed plants with the four model species (Table S3).

274

275  **Source analysis of variation in codon usage**
276  **ENC-plot**

277   The ENC and GC3s of the seven analyzed cpDNAs were plotted. It can be seen from Figure 2
278   that the ENC values of most genes were lower than expected values and lie below the standard
279   curve. The results of ENC-plot analysis suggested that codon usage preference of the seven
280   cpDNAs is mainly influenced by natural selection and other factors, while mutation pressure was
281   determined to play only a minor role.
282
283   **Figure 2.**  ENC-plot (ENC values vs GC3s) for chloroplast genomes of seven *Misacanthus* and related
284   species. The continuous curve represents the expected curve between GC3s and ENC under random
285   codon usage.
286
287   **PR2-plot**
288   PR2-plot is an efficient method to indicate the influence of mutation pressure by investigating
289   the composition of A, T, C and G at the third nucleotide position of a codon. Our results revealed
290   that the AT-bias is 0.464, 0.463, 0.463, 0.463, 0.463, 0.465 and 0.463 for *M. floridulus*, *M.*
291   *giganteus*, *M. sacchariflorus*, *M. sinensis*, *M. transmorrisonensis*, *Saccharum spontaneumand*
292   and *Sorghum bicolor*, while the GC-bias is 0.512, 0.513, 0.516, 0.515, 0.512, 0.515 and 0.518,
293   respectively (Fig 3). Therefore, T/G-bias was observed in all seven assessed species. When
294   considered together, codon usage bias of A/T and G/C in the seven cp genomes was unbalanced,
295   indicating that the base composition of the seven analyzed cpDNAs is not only influenced by
296   mutation pressure, but also by natural selection.
297
298   **Figure 3.**  PR2-plot (A3/(A3+T3) vs G3/(G3+C3)) for chloroplast genomes of seven *Misacanthus* and
299   related species. The pattern of splashes around the central spot (A=T, C=G) indicate the extent and
300   orientation of the base offset.
301
302   **Neutrality plot**
303   The distribution range of GC12 (the mean GC content at the first and second nucleotide positions
304   of a codon) and GC3 is relatively concentrated, in which the range of GC12 is 0.3272 to 0.5469,
305   and the range of GC3 is 0.179 to 0.512 (Fig 4). No significant correlation was found for GC1
306   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
307   GC3 (r8=0.092, r9=0.079, r10=0.055, r11=0.049, r12=0.100, r13=0.242, r14=0.202) and GC2
308   with GC3 (r14=0.054, r14=0.053, r15=-0.014, r16=-0.020, r17=0.063, r18=-0.032, r19=-0.032),
309   which suggested mutation pressure only contributes a minor role in the codon usage preference.
310   In addition, the regression coefficient (slope of neutrality plot) was 0.006 to 0.198, indicating
311   that the correlation between GC12 and GC3 is not significant, and the composition of the first
312   two bases may be different from the third base of the codon. These results demonstrated that the
313   codon usage patterns of cp coding sequences in the seven species are mainly affected by natural
314   selection.
315
316   **Figure 4.**  Neutrality plot (GC12 vs GC3) for chloroplast genomes of seven *Misacanthus* and related
317   species. The solid line represents the regression line.

318
319 **Correspondence analysis (COA)**
320 Correspondence analysis is used to explore the variations of codon usage among the analyzed
321 cpDNAs. In the current study, RSCU-based COA was used to compare the usage patterns of 59
322 codons, which produced a series of orthogonal axes, reflecting the trend of change of codon
323 usage in the seven *Miscanhus* and related plant species cpDNAs. The first four axes accounted
324 for 36.2%, 38.2%, 38.5%, 38.3%, 36.8%, 40.7% and 40.9% of the overall changes, while the
325 first axis proportion to 14.3%, 15.9%, 15.8%, 15.8%, 14.6%, 15.6% and 9.7% of the total
326 variation in seven species respectively (Table 3). Axis 1, responsible for∼10% of total variation,
327 was the main source of variation, indicating that the codon usage was influenced by multiple
328 factors. In addition, the relationship between axis 1 and axis 2 was visualized to explore the
329 effects of GC content on codon usage bias (Fig 5). Genes with different GC content are plotted
330 as different colors, red with GC%<45% and blue with 45%≤GC%<60% (Fig 5). In order to
331 determine the factors leading to gene dispersion along axis 1 and axis 2, the correlation index
332 were computed on axis 1 with CAI, CBI, Fop, GC3, GC and L_aa (Table 3). As can be seen
333 from the results in Table 3, axis 1 for *M. floridulus*, *M. sacchariflorus*, *M. sinensis*, *M.
334 transmorrisonensis*, *Saccharum spontaneum* and *Sorghum bicolor* possessed a remarkable
335 correlation with GC3s (p≤0.01), which indicated the base composition stemming from mutation
336 pressure was the main factor impacting codon usage preference.
337
338 **Table 3** Correlation analysis of axis 1 and codon usage index of cp genomes of seven *Misacanthus* and
339 related species (the T/C/A/G content at the third codon position of synonymous codons; codon adaptation
340 index: CAI; codon bias index: CBI; frequency of optimal codons: Fop; the GC content at the third codon
341 position of synonymous codons: GC3s; the GC content at the three position of synonymous codons: GC;
342 total number of amino acids: L_aa)
343
344 **Figure 5.** Visualization of the first two axes from the correspondence analysis based on RSCU values
345      for chloroplast genomes of seven *Misacanthus* and related species. Genes with different GC content
346      are plotted as different colors, red with GC%<45% and blue with 45%≤GC%<60%.
347

# Discussion
349 The study compared the codon usage patterns of the five *Miscanthus* species and two related
350 species. The findings reported here will help to improve our understanding of evolution analysis
351 and the optimization of codon components suitable for gene expression. During the evolutionary
352 processes, specific codon usage patterns were obtained to adapt to the diverse factors including
353 origin, evolution, natural selection and mutation pressure. In addition, analyzing the source of
354 variation in genomic codon usage, the pattern of codon bias and the codon frequency could
355 provide insights into optimization of the codons of heterologous genes and selection of
356 appropriate heterologous expression hosts. Therefore, the research will be of great significance to
357 the study of genetic engineering and genetic evolution.

358    Our analysis of base composition of codons revealed that the CDSs of the seven
359  *Miscanthus* and related cpDNAs analyzed tended to use an A/T codon, which was consistent
360  with the results of Zhang et al. (2012) on the 23 Poaceae cpDNAs that this previous study
361  analyzed. According to a previous study, the GC3 values of dicotyledonous plants is often less
362  than 50% (codon use prefers A/T), which is different from monocotyledonous plants with high
363  GC3 values (GC3 values >50%, showing that codon use prefers G/C) (Murray et al. 1989). The
364  results of RSCU value analysis showed an A/T codon usage bias in the cpDNAs of the seven
365  analyzed species, which was consistent with the patterns in most higher plants (Shang et al.
366  2011). According to neutral evolution theory, the effects of mutation pressure and natural
367  selection on the variation of the third base of codon are neutral or nearly neutral (Sharp et al.
368  1993). The study of Kawabe et al. (2003) showed that when codon use is affected by natural
369  selection, GC3 values tend to be distributed in a small range and there is no significant
370  correlation between GC12 and GC3 (Kawabe et al. 2003). The neutrality plot in this study
371  revealed a weak correlation between GC12 and GC3 and the composition of the first two bases
372  were different from the third base of the codon, which demonstrated that the codon usage
373  patterns of the seven cpDNAs analyzed are mainly influenced by natural selection. This result is
374  consistent with the codon usage of cpDNAs of many species, such as *Oryza sativa* (Liu et al.
375  2003), *Zea mays* (maize) (Liu et al. 2009), *Triticum aestivum* (Liu & Xue 2005) and
376  *Euphorbiaceae* (Wang et al. 2020). In addition, combining the results of ENC-plot, PR2-plot and
377  COA analysis suggested that the codon usage bias of the seven cpDNAs were affected by
378  multiple factors, including mutation pressure, base composition and gene length, which the
379  dominant influencing factor was natural selection. This results was consistent with the analysis in
380  *Poaceae* (Zhang et al. 2012), *Populus alba* (Zhou et al. 2008) and *Euphorbiaceae* (Wang et al.
381  2020) .

382    The cpDNAs of the seven *Miscanthus* and related plants assessed here are highly
383  conserved and share a total of 21 high-frequency codons. In addition, 4 to 11 codons were
384  determined to be the optimal codons in each species, while no common optimal codon was
385  defined in the seven representative species. These results of high frequency codons and optimal
386  codons are not only beneficial for codon optimization, but also promote further understanding of
387  the relationship between gene expression and codon usage preference. In higher plants, the main
388  obstacle to applying cp transformation to more species and especially, to other important crops is
389  the limitation of available tissue culture systems and regeneration protocols (Ruf et al. 2001).
390  Considering the differences of codon usage bias between the seven cpDNAs analyzed in this
391  study and heterologous expression hosts, codon usage frequencies were analyzed to select the
392  suitable host system. Based on the results, it was suggested to select *Saccharomyces cerevisiae*,
393  *Arabidopsis thaliana* and *Populus trichocarpa* as heterologous expression hosts for *Miscanthus*
394  and related crops, which possesse a little difference in codon usage frequency with the seven
395  plants.

396    This study conducted a comprehensive comparative analysis on codon usage patterns at
397  the cpDNA-wide level of seven *Miscanthus* and related species. These results will improve our

398  understanding on evolution analysis, the selection of appropriate heterologous expression hosts
399  and the optimization of codon components suitable for gene expression, finally provide a
400  theoretical basis for building a stable and efficient gene expression system in *Miscanthus* or other
401  crops.
402

## Conclusions

404  The codon usage patterns of cp genomes of the five *Miscanthus* and two related species were
405  compared and systematically analyzed for the first time. The results of codon usage bias and
406  RSCU analysis indicated that the seven representative species prefer the use of A/T bases and
407  A/T-ending codons. In addition, 21 common high-frequency codons and 4 to11 optimal codons
408  were elevated in their frequency in the seven cpDNAs. Furthermore, the analysis of codon usage
409  frequencies between the seven representative species and four model organisms suggested that
410  *Arabidopsis thaliana*, *Populus trichocarpa* and *Saccharomyces cerevisiae* could be considered
411  as appropriate heterologous expression hosts for *Miscanthus* and related species. Finally, when
412  considered together, the results of ENC-plot, PR2-plot and neutrality analysis revealed that the
413  codon usage pattern of the seven cpDNAs analyzed here were influenced by multiple factors,
414  which the dominant influencing factor was natural selection. These results presented in this study
415  potentially provide not only important reference information for evolutionary analysis, but also
416  potentially provide additional important insights for the improvement in the expression
417  efficiency of exogenously introduced genic sequences in transgenic research by codon
418  optimization.
419

## References

421  Amiryousefi A, Hyvönen J, and Poczai PJPo. 2018. The chloroplast genome sequence of
422      bittersweet (Solanum dulcamara): Plastid genome structure evolution in Solanaceae.
423      13:e0196069.
424  Anue MR, Xueli S, Chunzhen C, Zhongxiong LJPD, and Pests. 2019. Analysis of codon usage
425      pattern of banana basic secretory protease gene.  10:1-9.
426  Barling A, Swaminathan K, Mitros T, James BT, Morris J, Ngamboma O, Hall MC, Kirkpatrick J,
427      Alabady M, and Spence AKJBg. 2013. A detailed gene expression study of the
428      Miscanthus genus reveals changes in the transcriptome associated with the rejuvenation
429      of spring rhizomes.  14:1-16.
430  Chae WB, Hong SJ, Gifford JM, Rayburn AL, Sacks EJ, and Juvik JAJGB. 2014. Plant
431      morphology, genome size, and SSR markers differentiate five distinct taxonomic groups
432      among accessions in the genus M iscanthus.  6:646-660.
433  Chen C, Zheng Y, Liu S, Zhong Y, Wu Y, Li J, Xu L-A, and Xu MJP. 2017. The complete
434      chloroplast genome of Cinnamomum camphora and its comparison with related
435      Lauraceae species.  5:e3820.
436  Cheng H, Li J, Zhang H, Cai B, Gao Z, Qiao Y, and Mi LJP. 2017. The complete chloroplast
437      genome sequence of strawberry (Fragaria× ananassa Duch.) and comparison with
438      related species of Rosaceae.  5:e3919.
439  Clark LV, Stewart JR, Nishiwaki A, Toma Y, Kjeldsen JB, Jørgensen U, Zhao H, Peng J, Yoo
440      JH, and Heo KJJoEB. 2015. Genetic structure of Miscanthus sinensis and Miscanthus

441     sacchariflorus in Japan indicates a gradient of bidirectional but asymmetric introgression.
442     66:4213-4225.
443  Ge C, Liu X, Liu S, Xu J, Li H, Cui T, Yao Y, Chen M, Yu W, and Chen CJPMBR. 2017.
444     Miscanthus sp.: genetic diversity and phylogeny in China.  35:600-610.
445  Havaux M, Lütz C, and Grimm BJPP. 2003. Chloroplast membrane photostability in
446     chlPtransgenic tobacco plants deficient in tocopherols.  132:300-310.
447  Jarvis P, and López-Juez EJNRMCB. 2013. Biogenesis and homeostasis of chloroplasts and
448     other plastids.  14:787-802.
449  Kawabe A, Miyashita NTJG, and systems g. 2003. Patterns of codon usage bias in three dicot
450     and four monocot plant species.  78:343-352.
451  Khodakovskaya M, McAvoy R, Peters J, Wu H, and Li YJP. 2006. Enhanced cold tolerance in
452     transgenic tobacco expressing a chloroplast ω-3 fatty acid desaturase gene under the
453     control of a cold-inducible promoter.  223:1090-1100.
454  Kwak S-Y, Lew TTS, Sweeney CJ, Koman VB, Wong MH, Bohmert-Tatarev K, Snell KD, Seo
455     JS, Chua N-H, and Strano MSJNn. 2019. Chloroplast-selective gene delivery and
456     expression in planta using chitosan-complexed single-walled carbon nanotube carriers.
457     14:447-455.
458  Liu H, He R, Zhang H, Huang Y, Tian M, and Zhang J. 2009. Analysis of synonymous codon
459     usage in Zea mays. *Molecular Biology Reports* 37:677-684. 10.1007/s11033-009-9521-7
460  Liu Q-P, Tan J, and Xue Q-ZJYcxbAgS. 2003. Synonymous codon usage bias in the rice
461     cultivar 93-11 (Oryza sativa L. ssp. indica).  30:335-340.
462  Liu Q, Feng Y, Dong H, and Xue QJPS. 2004. Synonymous codon usage bias in Oryza sativa.
463     167:101-105.
464  Liu Q, and Xue QJJog. 2005. Comparative studies on codon usage pattern of chloroplasts and
465     their host nuclear genes in four plant species.  84:55-62.
466  Murray EE, Lotzer J, and Eberle MJNAR. 1989. Codon usage in plant genes.  17:477-498.
467  Nielsen AZ, Mellor SB, Vavitsas K, Wlodarczyk AJ, Gnanasekaran T, Perestrello Ramos H de
468     Jesus M, King BC, Bakowski K, and Jensen PEJTPJ. 2016. Extending the biosynthetic
469     repertoires of cyanobacteria and chloroplasts.  87:87-102.
470  Pan L-L, Wang Y, Hu J-H, Ding Z-T, and Li CJJotb. 2013. Analysis of codon use features of
471     stearoyl-acyl carrier protein desaturase gene in Camellia sinensis.  334:80-86.
472  Plotkin JB, and Kudla GJNRG. 2011. Synonymous but not the same: the causes and
473     consequences of codon bias.  12:32-42.
474  Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS, and Koller DJMsb. 2014.
475     Causal signals between codon bias, mRNA structure, and the efficiency of translation
476     and elongation.  10:770.
477  Quax TE, Claassens NJ, Söll D, and van der Oost JJMc. 2015. Codon bias as a means to fine-
478     tune gene expression.  59:149-161.
479  Ravi V, Khurana J, Tyagi A, Khurana PJPS, and Evolution. 2008. An update on chloroplast
480     genomes.  271:101-122.
481  Romero H, Zavala A, and Musto HJNar. 2000. Codon usage in Chlamydia trachomatis is the
482     result of strand-specific mutational biases and a complex pattern of selective forces.
483     28:2084-2090.
484  Ruf S, Hermann M, Berger IJ, Carrer H, and Bock RJNb. 2001. Stable genetic transformation of
485     tomato plastids and expression of a foreign protein in fruit.  19:870-875.
486  Shang M, Liu F, Hua J, and Wang KJSAS. 2011. Analysis on codon usage of chloroplast
487     genome of Gossypium hirsutum.  44:245-253.
488  Sharp PM, and Li W-HJJome. 1986. An evolutionary perspective on synonymous codon usage
489     in unicellular organisms.  24:28-38.
490  Sharp PM, Stenico M, Peden JF, and Lloyd ATJBST. 1993. Codon usage: mutational bias,
491     translational selection, or both?  21:835-841.

492    Sheng J, Hu X, Zeng X, Li Y, Zhou F, Hu Z, Jin S, and Diao Y. 2016. Nuclear DNA content in
493        Miscanthus sp. and the geographical variation pattern in Miscanthus lutarioriparius. *Sci
494        Rep* 6:34342. 10.1038/srep34342
495    Sheng J, Yan M, Wang J, Zhao L, Zhou F, Hu Z, Jin S, and Diao Y. 2021. The complete
496        chloroplast genome sequences of five Miscanthus species, and comparative analyses
497        with other grass plastomes. *Industrial Crops and Products* 162.
498        10.1016/j.indcrop.2021.113248
499    Sheng J, Zheng X, Wang J, Zeng X, Zhou F, Jin S, Hu Z, and Diao Y. 2017. Transcriptomics
500        and proteomics reveal genetic and biological basis of superior biomass crop Miscanthus.
501        *Scientific Reports* 7. 10.1038/s41598-017-14151-z
502    Sueoka NJG. 1999. Translation-coupled violation of Parity Rule 2 in human genes is not the
503        cause of heterogeneity of the DNA G+ C content of third codon position. 238:53-58.
504    Sueoka NJJome. 1995. Intrastrand parity rules of DNA base composition and usage biases of
505        synonymous codons. 40:318-325.
506    Sueoka NJPotNAoS. 1988. Directional mutation pressure and neutral molecular evolution.
507        85:2653-2657.
508    Tuller T, Waldman YY, Kupiec M, and Ruppin EJPotNAoS. 2010. Translation efficiency is
509        determined by both codon bias and folding energy. 107:3645-3650.
510    Vermerris W. 2008. Miscanthus: genetic resources and breeding potential to enhance bioenergy
511        production. *Genetic improvement of bioenergy crops*: Springer, 295-308.
512    Wang Z, Xu B, Li B, Zhou Q, Wang G, Jiang X, Wang C, and Xu Z. 2020. Comparative analysis
513        of codon usage patterns in chloroplast genomes of six Euphorbiaceae species. *PeerJ*
514        8:e8251. 10.7717/peerj.8251
515    Wright FJG. 1990. The 'effective number of codons' used in a gene. 87:23-29.
516    Wu Y, Zhou N, Ni X, Okoye CO, Wang Y, Li X, Gao L, Zhou G, Jiang JJIC, and Products. 2021.
517        Developing a long-term and powerful in vitro culture and Agrobacterium-mediated
518        transformation system for Miscanthus sinensis (Poaceae). 161:113190.
519    Xiang H, Zhang R, Butler III RR, Liu T, Zhang L, Pombert J-F, and Zhou ZJPo. 2015.
520        Comparative analysis of codon usage bias patterns in microsporidian genomes.
521        10:e0129223.
522    Yan M, Zhao X, Zhou J, Huo Y, Ding Y, and Yuan ZJIjoms. 2019. The complete chloroplast
523        genomes of Punica granatum and a comparison with other species in Lythraceae.
524        20:2886.
525    Zhang J, Yan J, Zhang Y, Ma X, Bai S, Wu Y, Dao Z, Li D, Zhang C, and Zhang YJPo. 2013.
526        Molecular insights of genetic variation in Erianthus arundinaceus populations native to
527        China. 8:e80388.
528    Zhang R, Zhang L, Wang W, Zhang Z, Du H, Qu Z, Li X-Q, and Xiang HJIjoms. 2018.
529        Differences in codon usage bias between photosynthesis-related genes and genetic
530        system-related genes of chloroplast genomes in cultivated and wild solanum species.
531        19:3142.
532    Zhang WJ, Zhou, J., Li, Z. F., Wang, L., Gu, X., & Zhong, Y. 2007. Comparative analysis of
533        codon usage patterns among mitochondrion, chloroplast and nuclear genes in Triticum
534        aestivum L. . *Journal of Integrative Plant Biology* 49:246-254.
535    Zhang Y, Nie X, Jia X, Zhao C, Biradar SS, Wang L, Du X, and Weining SJAJoB. 2012.
536        Analysis of codon usage patterns of the chloroplast genomes in the Poaceae family.
537        60:461-470.
538    Zhou M, Long W, and Li XJJoFR. 2008. Analysis of synonymous codon usage in chloroplast
539        genome of Populus alba. 19:293-297.
540    Zhou M, Tong C, Shi JJJoG, and Genomics. 2007. Analysis of codon usage between different
541        poplar species. 34:555-561.
542

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

1   **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
2   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 |

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 |
|---|---|---|
| *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' |

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

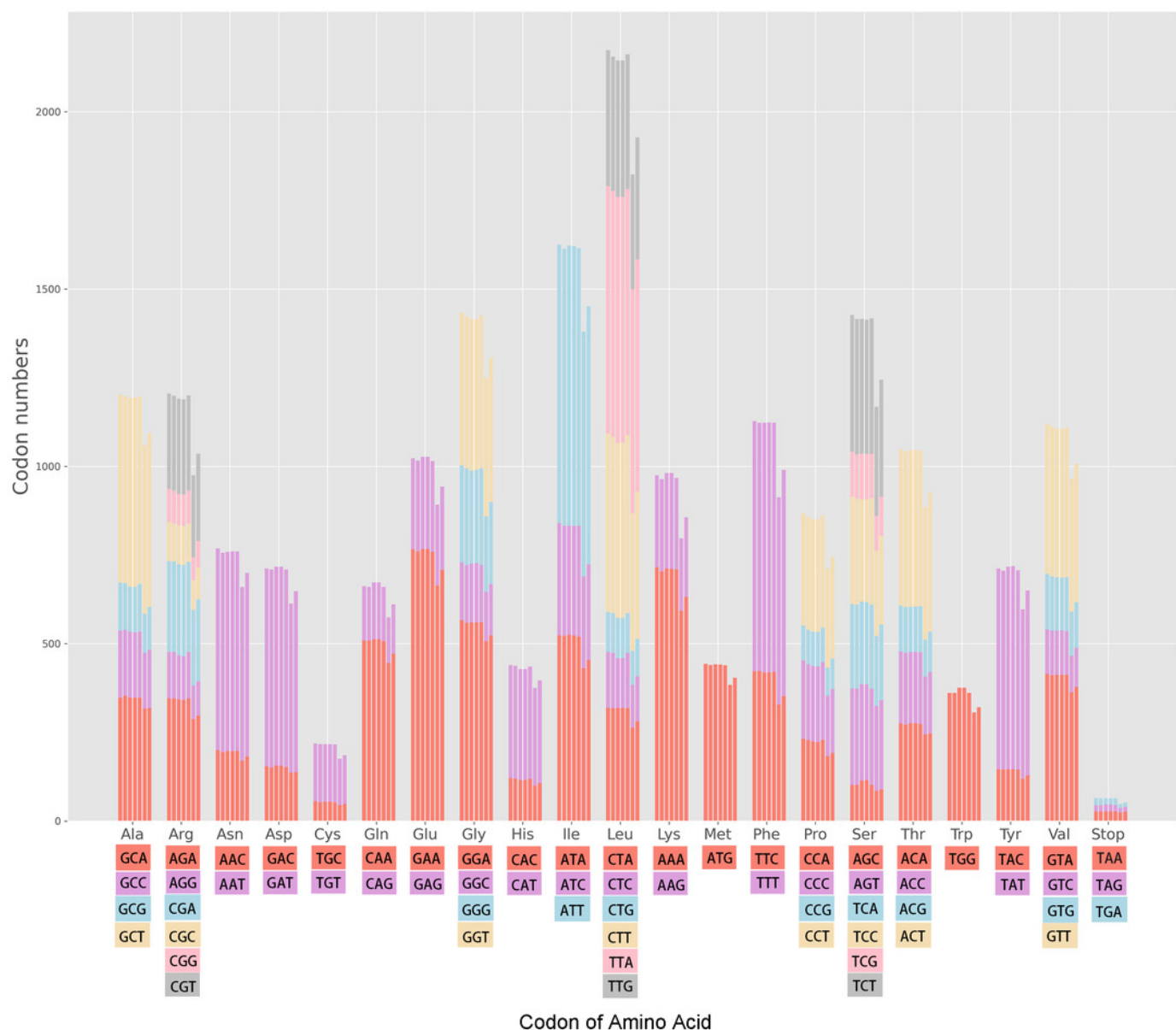
1  **Table 3** Correlation analysis of axis 1 and codon usage index of chloroplast genomes of seven *Misacanthus* and related species (the T/C/A/G content at
2  the third codon position of synonymous codons; codon adaptation index: CAI; codon bias index: CBI; frequency of optimal codons: Fop; the GC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

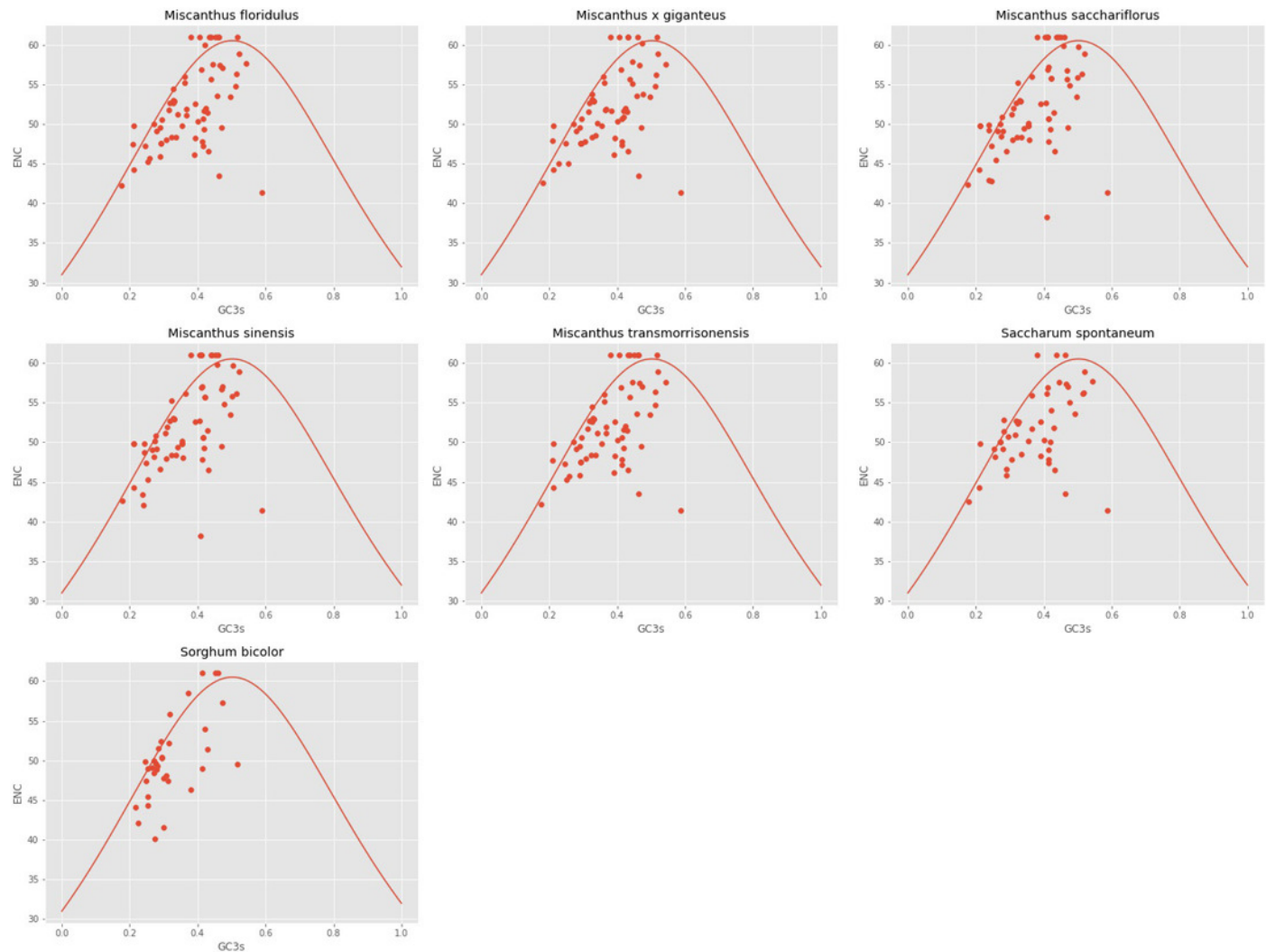
| Species | T3s | C3s | A3s | G3s | CAI | CBI | Fop | Nc | GC3s | GC | L_aa |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Miscanthus floridulus* | -0.65** | 0.507** | -0.133 | 0.598** | -0.152 | 0.073 | 0.098 | 0.28** | 0.639** | 0.159 | -0.315** |
| *Miscanthus x giganteus* | 0.016 | 0.053 | -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 |

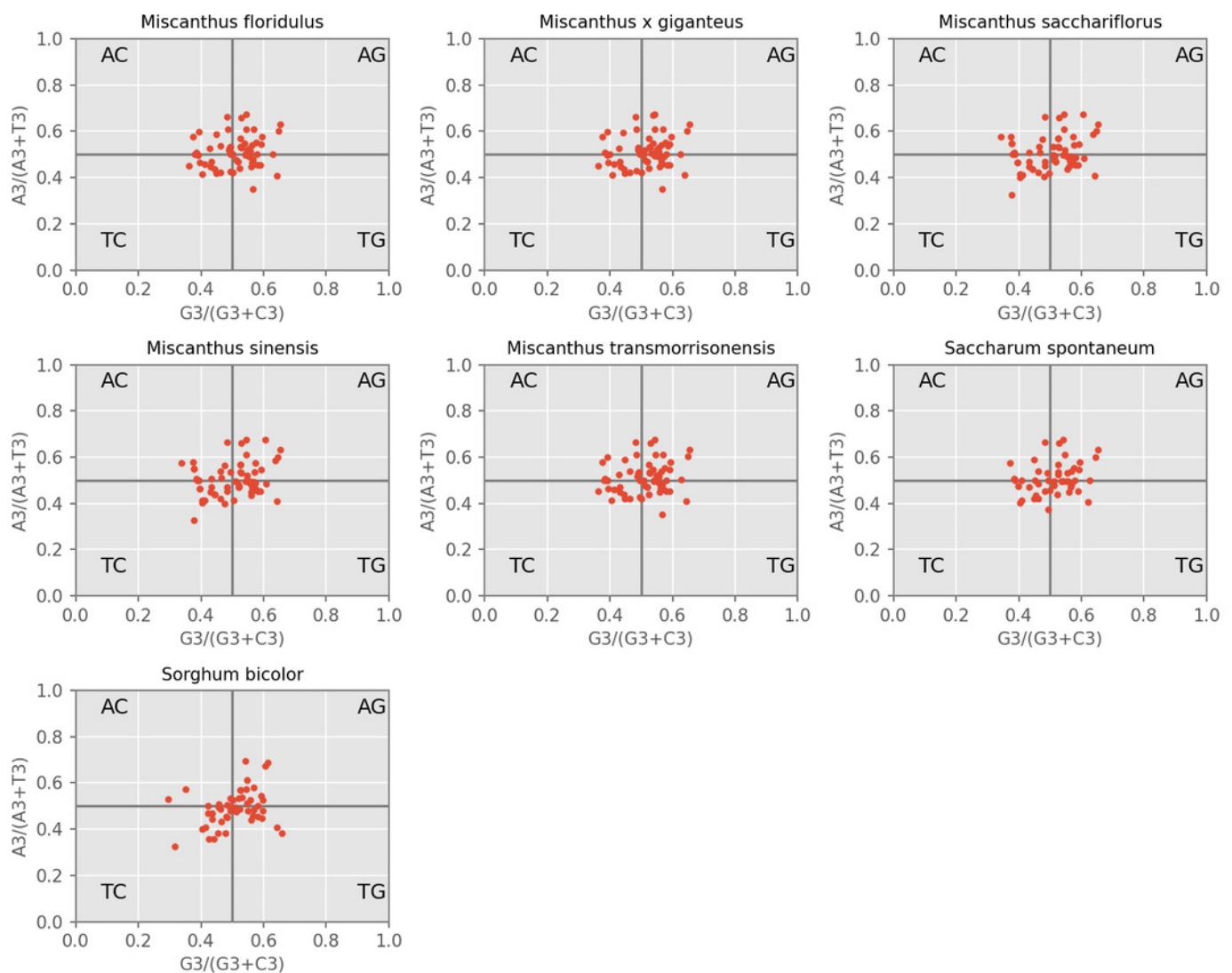
5  Notes: *P<0.05; **P<0.01

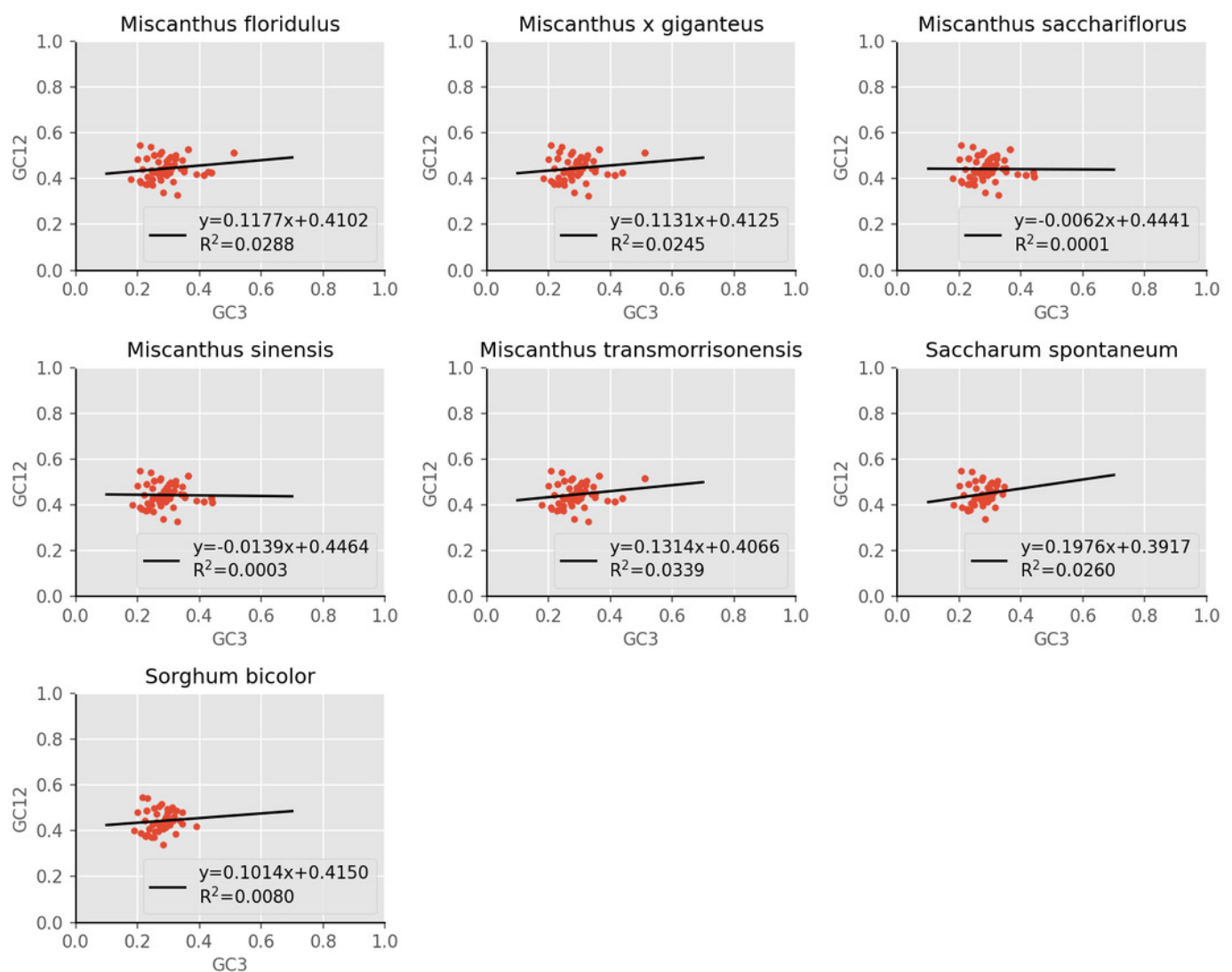
# Figure 1

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 *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