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

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

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

[2] Wuhan University, Wuhan, China

[3] Anhui University of Science and Technology, Huainan, China

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

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

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

4
5

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

7 1 School of Life Sciences, Jiangsu Key Laboratory of Neuroregeneration, Co-innovation Center of Neuroregeneration, Nantong University,

8 Nantong, 226019, PR China

9 2 State Key Laboratory of Hybrid Rice, Lotus Engineering Research Center of Hubei Province, College of Life Sciences, Wuhan University,

10 Wuhan, 430072, PR China

11 3 Key Laboratory of Industrial Dust Prevention and Control & Occupational Safety and Health of the Ministry of Education, Anhui University of

12 Science and Technology, Huainan, 232001 China

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

18

19 ## Abstract

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

35

36 ## Introduction

37 *Miscanthus Andersson* (Poaceae) are C4 photosynthetic plants, which have been widely
38 investigated as potential second-generation bio-energy crops (Barling et al. 2013). The genus
39 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 extensive environmental adaptation (Vermerris
47 2008). Currently, the research focus of *Miscanthus* species is to utilize it as promising genetic
48 resource (Clark et al. 2015; Zhang et al. 2013). The more diverse genetic resources we have, the
49 better we can comprehend the adaptation, evolution and utilization of these significant economic
50 crops.
51   Chloroplasts (cp) are key plastids involved in multifunctional processes of plant cell,
52 including photosynthesis, carbon fixation, starch storage, nitrogen metabolism, fatty acid and
53 nucleic acid synthesis in plant cell (Jarvis & López-Juez 2013; Nielsen et al. 2016). Typically, cp
54 genomes possess a small size, conserved gene content and large copy numbers, which have been
55 extensively used as valuable source for evolution analysis and plastid engineering (Amiryousefi
56 et al. 2018; Ravi et al. 2008; Yan et al. 2019). The lack of genomic resources of the *Miscanthus*
57 species has hindered the adequate understanding of their diversity traits (Chae et al. 2014; Sheng
58 et al. 2016) The low expression efficiency of an exogenous gene may limit the research progress
59 on the functional studies of *Miscanthus* and their related species (Wu et al. 2021). Benefited
60 from the rapid development of cp engineering, plasmid DNA have been transferred into the cp of
61 a variety of plants, such as *Nicotiana tabacum*, *Manihot esculenta Crantz* and *Eruca sativa Mill*
62 (Havaux et al. 2003; Khodakovskaya et al. 2006; Kwak et al. 2019). Recently, the cp genomes of
63 some *Miscanthus* species have been made available in the National Center for Biotechnology
64 Information (NCBI) database (Sheng et al. 2021). These complete cp sequences in *Miscanthus*
65 species can be used for studying population genetics, evolution analysis and plastid engineering
66 (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
71 (Pop et al. 2014; Quax et al. 2015; Tuller et al. 2010). The studies of codon preference can not
72 only reveal the evolutionary rules between genes in a species or related species, but also improve
73 the expression efficiency of exogenous sequences in transgenic research by codon optimization.
74 Recently, the applicability of codon optimization in the cp genomes have been proven in many
75 vascular plants, including *Poaceae*（Gramineae）(Zhang et al. 2012), *Cinnamomum camphorn*
76 *(L.) presl* (Camphor tree) (Chen et al. 2017), *Fragaria × ananassa Duch* (strawberry) (Cheng et
77 al. 2017) and *Solanum tuberosum L.* (potato) (Zhang et al. 2018). However, the codon usage
78 pattern of cp genomes in *Miscanthus* and related species has not been fully elucidated.

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

Materials & Methods

Genomes and sequences selection

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

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

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

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

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

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

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

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

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

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

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

196　synonymous codon usage of the gene (Xiang et al. 2015). Therefore, the major trends (Axis 1) of
197　these axes in the 59-dimensional hyperspace can be used to determine the maximum fraction of
198　genetic variation, indicating the major sources of codon usage variation. In addition, according to
199　the results of COA, the correlation index between Axis1 and codon usage exponent, including
200　the GC content of codons, GC3s, codon adaptation index (CAI) and the total numbers of amino
201　acids in the encoded polypeptide (L_aa) were computed by python scripts package
202　(https://github.com/shexuan/ codon_analysis). CAI value is widely applied to assess gene
203　expression levels, ranging from 0 to 1. Specifically, the larger the CAI value is, the stronger the
204　codon usage preference is, and vice versa (Sharp & Li 1986).
205
206　**Results**
207
208　**Characteristics of codon usage bias**
209
210　**Analysis of base composition of codon**
211　The screened CDSs numbers processed by Python scripts are 65, 64, 64, 64, 64, 48 and 52 for *M.*
212　*floridulus*, *M. giganteus*, *M. sacchariflorus*, *M. sinensis*, *M. transmorrisonensis*, *Saccharum*
213　*spontaneum* and *Sorghum bicolor* respectively. In addition, the GC contents of three positions of
214　codons (GC1, GC2, GC3) were calculated respectively (Table 1). It was found that the contents
215　of GC at all three sites and the average GC content (GC123) were all less than 0.5, which
216　indicated the seven cp genomes were prone to use A/T bases and A/T-ending codons (Table 1).
217　Specifically, the mean GC content of three sites in *M. floridulus*, *M. giganteus* and *M.*
218　*transmorrisonensis* is the same (0.375), and matches between *M. sacchariflorus* and *M. sinensis*
219　(0.393), but slightly different in *Saccharum spontaneum* (0.391) and *Sorghum bicolor* (0.39)
220　(Table 1). Furthermore, the distribution trend of GC content was GC1 > GC2 >GC3, indicating
221　that GC was not evenly distributed in the three positions of a codon. In summary, the codon
222　usages of GC content in these seven cp genomes were similar and were biased towards A/T
223　bases.
224
225　**Table 1** Genomic features of cp genomes of the seven Miscanthus and related species (the total
226　number of amino acids: L_aa ; the GC content at the first, second and third codon positions: GC1,
227　GC2and GC3; average GC at three locations: GC123).
228
229　**RSCU and RFSC**
230　The cp genomes of the seven *Miscanthus* and related species have 30 common codons (RSCU >
231　1) with 28 codons ending with A/T (93.3%) (Table S1). Therefore, the codons of the seven
232　plants species analyzed (RSCU > 1) are likely to end with A/T. The variation ranges in the
233　RSCU values were close in the seven cp genomes, i.e., 0.31–1.93 in *M. floridulus*, *M. giganteus*
234　and *M. transmorrisonensis*, 0.32-1.94 in *M. sacchariflorus* and *M. sinensis*, 0.32- 2.01 in
235　*Saccharum spontaneum* and 0.33-2.04 in *Sorghum bicolor*, respectively (Table S1). In addition,

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

242

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

247

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

254

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

256

257 **Codon usage frequency**
258 The codon usage frequencies of the seven cp genomes were compared with four model species
259 including *Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Populus*
260 *trichocarpa* (Table S3). Results indicated that there is little divergence in the codon usage
261 frequencies among the seven assessed plant species with *Saccharomyces cerevisiae*, *Arabidopsis*
262 *thaliana* and *Populus trichocarpa*, have 9–11 (accounting for 14.06%–17.19% of total codons),
263 13–14 (20.31%–21.88%), 12–13 (18.75%–20.31%) different codons, respectively (Table S3).
264 However, the codon usage frequencies of the seven species with *Escherichia coli* were
265 comparatively higher (27 different codons). The results indicated that the codon frequency
266 difference between *Miscanthus* species and *Arabidopsis thaliana*, *Populus trichocarpa* and
267 *Saccharomyces cerevisiae* was the lowest, while was the largest with *Escherichia coli*. Based on
268 above results, it was optimal to select *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and
269 *Populus trichocarpa* as heterologous gene expression receptor for *Miscanthus* and related
270 species. In addition, the results shown that TAG is a different termination codon in comparison
271 of the seven plants with the four model species (Table S3).

272

273 **Source analysis of variation in codon usage**
274 **ENC-plot**

275　The ENC and GC3s of the seven *Miscanthus* and related plant cp genomic were analyzed and
276　plotted. It can be seen from Figure 2 that the ENC values of most genes were lower than
277　expected values and lie below the standard curve. The results of ENC-plot analysis suggested
278　that codon usage preference of the seven cp genomes is mainly influenced by natural selection
279　and other factors, while mutation pressure plays only a minor roles.
280

281　**Figure 2.** ENC-plot (ENC values vs GC3s) for chloroplast genomes of seven *Misacanthus* and related
282　species. The continuous curve represents the expected curve between GC3s and ENC under random
283　codon usage.
284

285　**PR2-plot**
286　PR2-plot is an efficient method to indicate the influence of mutation pressure by investigating
287　the composition of A, T, C and G at the third position. Our results revealed that the AT-bias is
288　0.464, 0.463, 0.463, 0.463, 0.463, 0.465 and 0.463 for *M. floridulus*, *M. giganteus*, *M.*
289　*sacchariflorus*, *M. sinensis*, *M. transmorrisonensis*, *Saccharum spontaneumand* and *Sorghum*
290　*bicolor*, while the GC-bias is 0.512, 0.513, 0.516, 0.515, 0.512, 0.515 and 0.518, respectively
291　(Fig 3). Therefore, T/G-bias was observed in all seven *Miscanthus* and related species. All in all,
292　codon usage bias of A/T and G/C in the seven cp genomes was unbalanced, indicating that the
293　base composition of the seven cp genomes is not only influenced by mutation pressure, but also
294　by natural selection.
295

296　**Figure 3.** PR2-plot (A3/(A3+T3) vs G3/(G3+C3)) for chloroplast genomes of seven *Misacanthus* and
297　related species. The pattern of splashes around the central spot (A=T, C=G) indicate the extent and
298　orientation of the base offset.
299

300　**Neutrality plot**
301　The distribution range of GC12 (the mean GC content at the first and second nucleotide positions
302　of a codon) and GC3 is relatively concentrated, in which the range of GC12 is $0.3272 \sim 0.5469$,
303　and the range of GC3 is $0.1794 \sim 0.512$ (Fig 4). No significant correlation was found for GC1
304　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
305　GC3 (r8=0.092, r9=0.079, r10=0.055, r11=0.049, r12=0.1, r13=0.242, r14=0.202) and GC2 with
306　GC3 (r14=0.054, r14=0.053, r15=-0.014, r16=-0.02, r17=0.063, r18=-0.032, r19=-0.032), which
307　suggested mutation pressure only contributes a minor role in the codon usage preference. In
308　addition, the regression coefficient (slope of neutrality plot) was 0.0062 to 0.1976, indicating that
309　the correlation between GC12 and GC3 is not significant, and the composition of the first two
310　bases may be different from the third base of the codon. These results demonstrated that the
311　codon usage patterns of cp coding sequences in the seven species are mainly affected by natural
312　selection.
313

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

316

Correspondence analysis (COA)

318 Correspondence analysis is used to explore the variations of codon usage among the cp genomes.
319 In the current study, RSCU-based COA was used to compare the usage patterns of 59 codons,
320 which produced a series of orthogonal axes, reflecting the trend of change of codon usage in the
321 seven *Miscanhus* and related plant species cp genomes. The first four axes accounted for
322 36.22%, 38.18%, 38.47%, 38.31%, 36.8%, 40.72% and 40.94% of the overall changes, while the
323 first axis proportion to 14.33%, 15.91%, 15.84%, 15.78%, 14.59%, 15.59% and 9.70% of the
324 total variation in seven species respectively (Table 3). Axis 1, responsible for~10% of total
325 variation, was the main source of variation, indicating that the codon usage was influenced by
326 multiple factors. In addition, the relationship between axis 1 and axis 2 was visualized to explore
327 the effects of GC content on codon usage bias (Fig 5). Genes with different GC content are
328 plotted as different colors, red with GC%<45% and blue with 45%≤GC%<60% (Fig 5). In
329 order to determine the factors leading to gene dispersion along axis 1 and axis 2, the correlation
330 index were computed on axis 1 with CAI, CBI, Fop, GC3, GC and L_aa (Table 3). As can be
331 seen from the results in Table 3, axis 1 for *M. floridulus*, *M. sacchariflorus*, *M. sinensis*, *M.*
332 *transmorrisonensis*, *Saccharum spontaneum* and *Sorghum bicolor* possessed a remarkable
333 correlation with GC3s (p≤0.01), which indicated the base composition stemming from mutation
334 pressure was the main factor impacting codon usage preference.

335

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

341

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

345

Discussion

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

356   Our analysis of base composition of codons revealed that the CDS of the seven
357 *Miscanthus* and related cp genomes tended to use A/T codon, which was consistent with the
358 results of Zhang et al. (2012) on the 23 Poaceae cp genomes. According to a previous study, the
359 GC3S value of dicotyledonous plants is often less than 50% (codon use prefers A/T), which is
360 different from the monocotyledonous plants with high GC3S value (GC3S value >50%, showing
361 that codon use prefers G/C) (Murray et al. 1989). The results of RSCU value analysis showed an
362 A/T codon usage bias in the cp genomes of the seven representative plants, which was consistent
363 with the patterns in most higher plants (Shang et al. 2011). According to neutral evolution
364 theory, the effects of mutation pressure and natural selection on the variation of the third base of
365 codon are neutral or nearly neutral (Sharp et al. 1993). The study of Kawabe et al. (2003) showed
366 that when codon use is affected by natural selection, GC3 values tend to be distributed in a small
367 range and there is no significant correlation between GC12 and GC3 (Kawabe et al. 2003). The
368 neutrality plot in this study revealed a weak correlation between GC12 and GC3 and the
369 composition of the first two bases were different from the third base of the codon, which
370 demonstrated that the codon usage patterns of the seven cp genomes are mainly influenced by
371 natural selection. This result is consistent with the codon usage of cp genomes of many species,
372 such as *Oryza sativa* (Liu et al. 2003), *Zea mays* (maize), *Triticum aestivum* (Liu & Xue 2005)
373 and *Euphorbiaceae* (Wang et al. 2020). In addition, combining the results of ENC-plot, PR2-plot
374 and COA suggested that the codon usage bias of the seven cp genomes were affected by multiple
375 factors, such as mutation pressure, base composition and gene length, among which the main
376 influencing factor was natural selection. Similar results have also been reported in cp genomes of
377 *Poaceae* (Zhang et al. 2012), *Populus alba* (Zhou et al. 2008) and *Euphorbiaceae* (Wang et al.
378 2020), where multiple factors influences on codon usage bias.
379   The cp genomes of the seven *Miscanthus* and related plants assessed here possess strong
380 common base and share a total of 21 high-frequency codons. In addition, 4 to 11 codons were
381 determined to be the optimal codons in each species, while no common optimal codon was
382 defined in the seven representative species. These results of high frequency codons and optimal
383 codons are not only beneficial for codon optimization, but also promote further understanding of
384 the relationship between gene expression and codon usage preference. In higher plants, the main
385 obstacle to applying cp transformation to more species and especially, to other important crops is
386 the limitation of available tissue culture systems and regeneration protocols (Ruf et al. 2001).
387 Considering the variations in codon usage bias among the seven represent cp genomes and the
388 plasmid vector receptors for heterologous genes expression, codon usage frequencies were
389 analyzed in this study as well. Based on the results, it was suggested to select *Saccharomyces
390 cerevisiae*, *Arabidopsis thaliana* and *Populus trichocarpa* as foreign gene expression plasmid
391 vector receptors for *Miscanthus* and related crops, which possesse a little difference in codon
392 usage frequency with the seven plants.
393   This study conducted a comprehensive comparative analysis on codon usage patterns at
394 the cp genome-wide level of seven *Miscanthus* and related species. These results will improve
395 our understanding on evolution analysis, the selection of appropriate heterologous gene

396 expression plasmid vector receptors and the optimization of codon components suitable for gene
397 expression, finally provide a theoretical basis for building a stable and efficient gene expression
398 system in *Miscanthus* or other crops.

399

Conclusions

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

414

References

416 Amiryousefi A, Hyvönen J, and Poczai PJPo. 2018. The chloroplast genome sequence of
417   bittersweet (Solanum dulcamara): Plastid genome structure evolution in Solanaceae.
418   13:e0196069.
419 Anue MR, Xueli S, Chunzhen C, Zhongxiong LJPD, and Pests. 2019. Analysis of codon usage
420   pattern of banana basic secretory protease gene. 10:1-9.
421 Barling A, Swaminathan K, Mitros T, James BT, Morris J, Ngamboma O, Hall MC, Kirkpatrick J,
422   Alabady M, and Spence AKJBg. 2013. A detailed gene expression study of the
423   Miscanthus genus reveals changes in the transcriptome associated with the rejuvenation
424   of spring rhizomes. 14:1-16.
425 Chae WB, Hong SJ, Gifford JM, Rayburn AL, Sacks EJ, and Juvik JAJGB. 2014. Plant
426   morphology, genome size, and SSR markers differentiate five distinct taxonomic groups
427   among accessions in the genus M iscanthus. 6:646-660.
428 Chen C, Zheng Y, Liu S, Zhong Y, Wu Y, Li J, Xu L-A, and Xu MJP. 2017. The complete
429   chloroplast genome of Cinnamomum camphora and its comparison with related
430   Lauraceae species. 5:e3820.
431 Cheng H, Li J, Zhang H, Cai B, Gao Z, Qiao Y, and Mi LJP. 2017. The complete chloroplast
432   genome sequence of strawberry (Fragaria× ananassa Duch.) and comparison with
433   related species of Rosaceae. 5:e3919.
434 Clark LV, Stewart JR, Nishiwaki A, Toma Y, Kjeldsen JB, Jørgensen U, Zhao H, Peng J, Yoo
435   JH, and Heo KJJoEB. 2015. Genetic structure of Miscanthus sinensis and Miscanthus
436   sacchariflorus in Japan indicates a gradient of bidirectional but asymmetric introgression.
437   66:4213-4225.
438 Ge C, Liu X, Liu S, Xu J, Li H, Cui T, Yao Y, Chen M, Yu W, and Chen CJPMBR. 2017.
439   Miscanthus sp.: genetic diversity and phylogeny in China. 35:600-610.

440   Havaux M, Lütz C, and Grimm BJPP. 2003. Chloroplast membrane photostability in
441   chlPtransgenic tobacco plants deficient in tocopherols. 132:300-310.
442   Jarvis P, and López-Juez EJNRMCB. 2013. Biogenesis and homeostasis of chloroplasts and
443   other plastids. 14:787-802.
444   Kawabe A, Miyashita NTJG, and systems g. 2003. Patterns of codon usage bias in three dicot
445   and four monocot plant species. 78:343-352.
446   Khodakovskaya M, McAvoy R, Peters J, Wu H, and Li YJP. 2006. Enhanced cold tolerance in
447   transgenic tobacco expressing a chloroplast ω-3 fatty acid desaturase gene under the
448   control of a cold-inducible promoter. 223:1090-1100.
449   Kwak S-Y, Lew TTS, Sweeney CJ, Koman VB, Wong MH, Bohmert-Tatarev K, Snell KD, Seo
450   JS, Chua N-H, and Strano MSJNn. 2019. Chloroplast-selective gene delivery and
451   expression in planta using chitosan-complexed single-walled carbon nanotube carriers.
452   14:447-455.
453   Liu Q-P, Tan J, and Xue Q-ZJYcxbAgS. 2003. Synonymous codon usage bias in the rice
454   cultivar 93-11 (Oryza sativa L. ssp. indica). 30:335-340.
455   Liu Q, and Xue QJJog. 2005. Comparative studies on codon usage pattern of chloroplasts and
456   their host nuclear genes in four plant species. 84:55-62.
457   Murray EE, Lotzer J, and Eberle MJNAR. 1989. Codon usage in plant genes. 17:477-498.
458   Nielsen AZ, Mellor SB, Vavitsas K, Wlodarczyk AJ, Gnanasekaran T, Perestrello Ramos H de
459   Jesus M, King BC, Bakowski K, and Jensen PEJTPJ. 2016. Extending the biosynthetic
460   repertoires of cyanobacteria and chloroplasts. 87:87-102.
461   Pan L-L, Wang Y, Hu J-H, Ding Z-T, and Li CJJotb. 2013. Analysis of codon use features of
462   stearoyl-acyl carrier protein desaturase gene in Camellia sinensis. 334:80-86.
463   Plotkin JB, and Kudla GJNRG. 2011. Synonymous but not the same: the causes and
464   consequences of codon bias. 12:32-42.
465   Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS, and Koller DJMsb. 2014.
466   Causal signals between codon bias, mRNA structure, and the efficiency of translation
467   and elongation. 10:770.
468   Quax TE, Claassens NJ, Söll D, and van der Oost JJMc. 2015. Codon bias as a means to fine-
469   tune gene expression. 59:149-161.
470   Ravi V, Khurana J, Tyagi A, Khurana PJPS, and Evolution. 2008. An update on chloroplast
471   genomes. 271:101-122.
472   Romero H, Zavala A, and Musto HJNar. 2000. Codon usage in Chlamydia trachomatis is the
473   result of strand-specific mutational biases and a complex pattern of selective forces.
474   28:2084-2090.
475   Ruf S, Hermann M, Berger IJ, Carrer H, and Bock RJNb. 2001. Stable genetic transformation of
476   tomato plastids and expression of a foreign protein in fruit. 19:870-875.
477   Shang M, Liu F, Hua J, and Wang KJSAS. 2011. Analysis on codon usage of chloroplast
478   genome of Gossypium hirsutum. 44:245-253.
479   Sharp PM, and Li W-HJJome. 1986. An evolutionary perspective on synonymous codon usage
480   in unicellular organisms. 24:28-38.
481   Sharp PM, Stenico M, Peden JF, and Lloyd ATJBST. 1993. Codon usage: mutational bias,
482   translational selection, or both? 21:835-841.
483   Sheng J, Hu X, Zeng X, Li Y, Zhou F, Hu Z, Jin S, and Diao Y. 2016. Nuclear DNA content in
484   Miscanthus sp. and the geographical variation pattern in Miscanthus lutarioriparius. *Sci*
485   *Rep* 6:34342. 10.1038/srep34342
486   Sheng J, Yan M, Wang J, Zhao L, Zhou F, Hu Z, Jin S, and Diao Y. 2021. The complete
487   chloroplast genome sequences of five Miscanthus species, and comparative analyses
488   with other grass plastomes. *Industrial Crops and Products* 162.
489   10.1016/j.indcrop.2021.113248

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

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

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

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
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'

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

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: *$P<0.05$; **$P<0.01$

6

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*
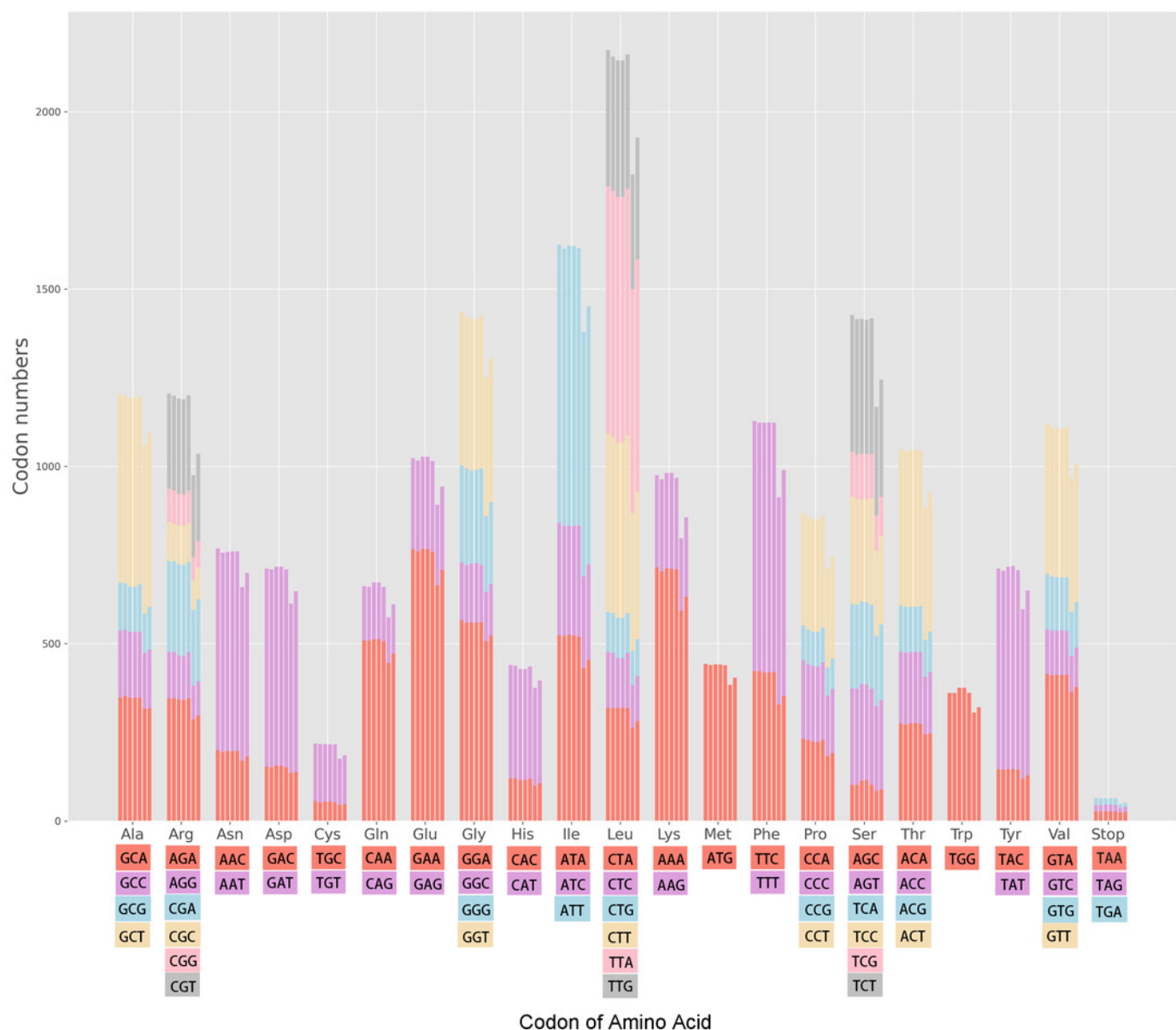
Figure 2

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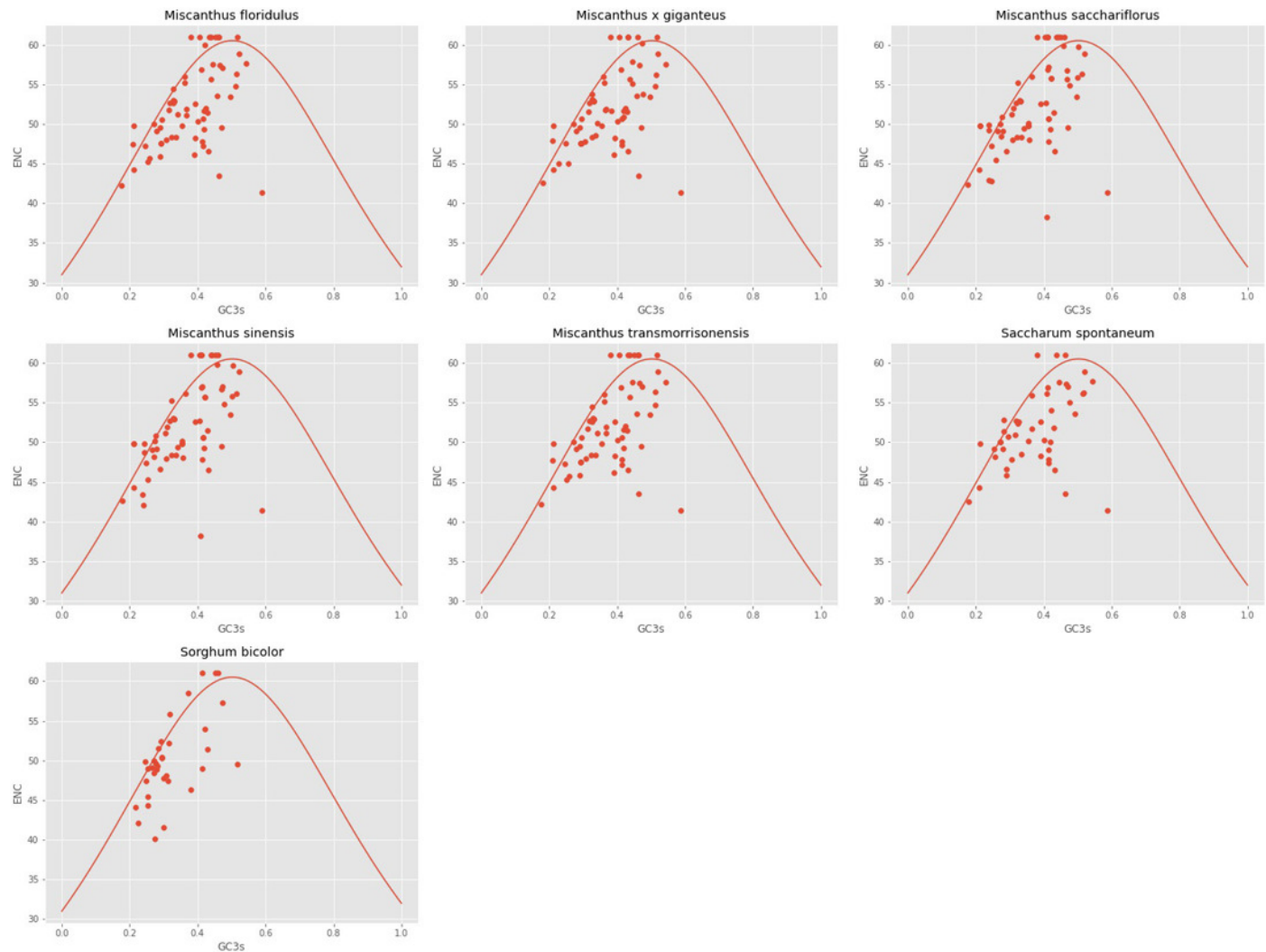
Figure 3

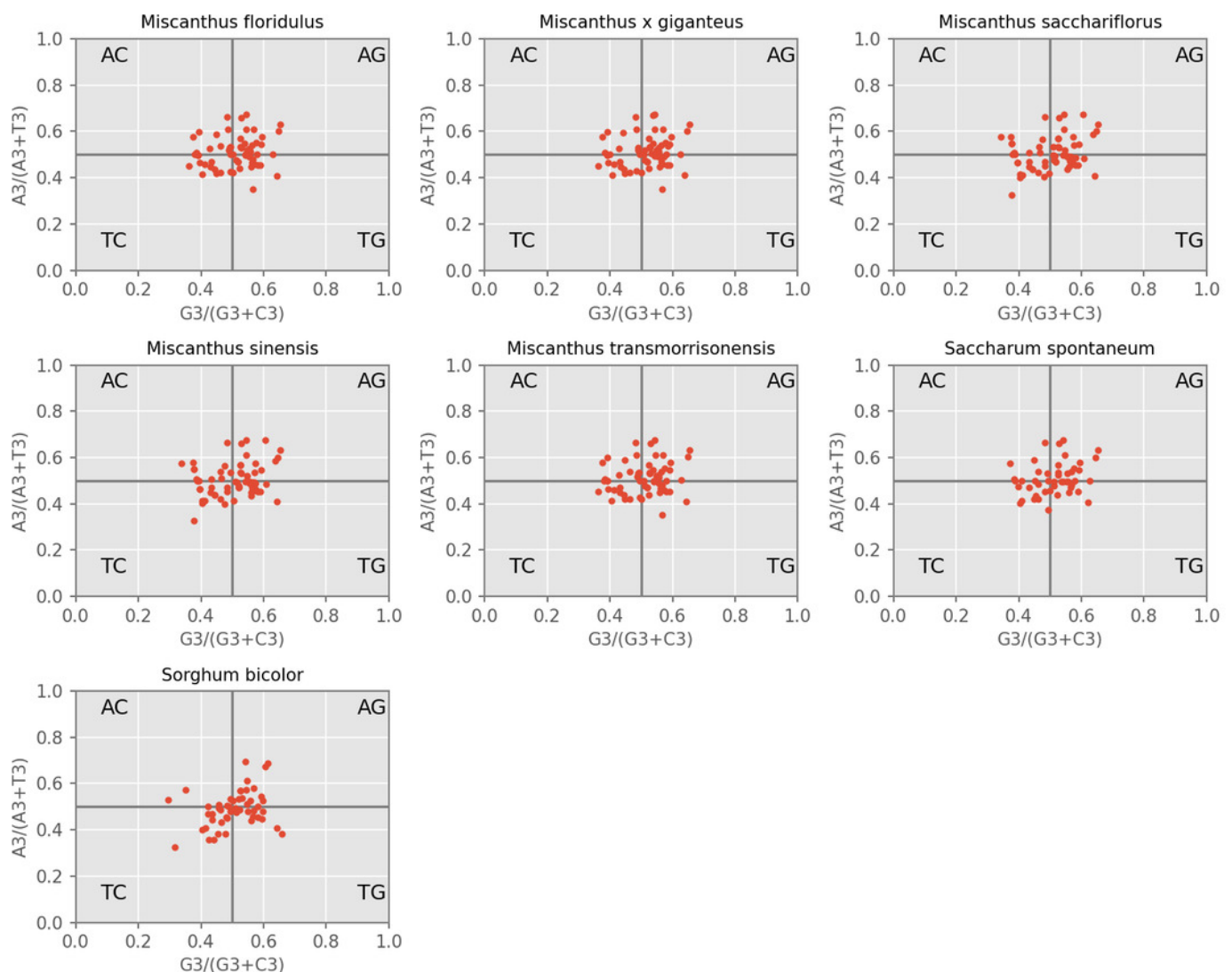PR2-plot of chloroplast genomes of seven *Misacanthus* and related species

Figure 4

Neutrality plot of chloroplast genomes of seven *Misacanthus* and related species
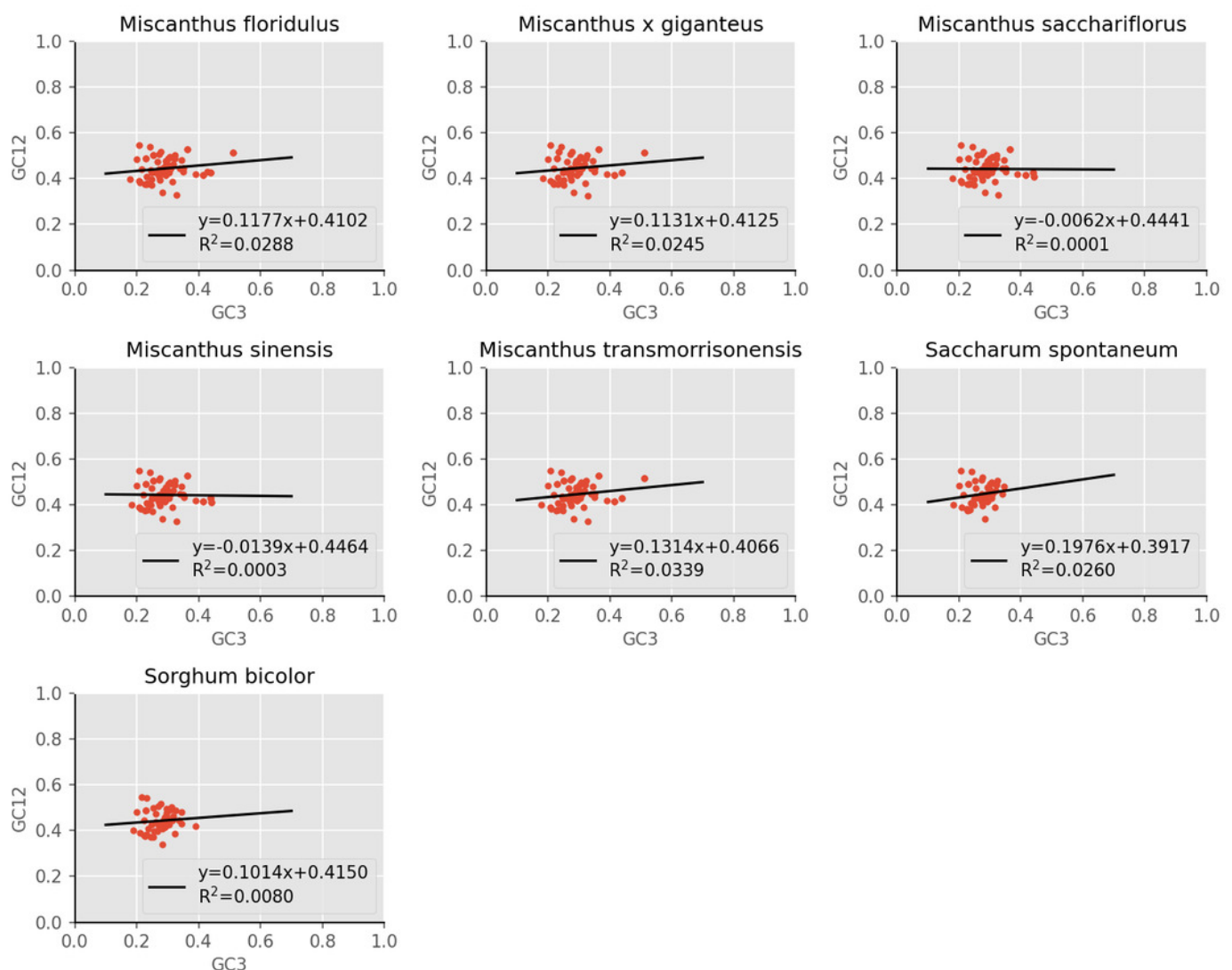
Figure 5

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