Indian J. Genet., 79 (2) 485-493 (2019)
DOI: 10.31742/IJGPB.79.2.13

De novo transcriptome assembly and analysis of the codon usage bias of the MADS-box gene family in *Cymbidium kanran*

Huolin Luo, Yuan Tao, Wenjing Yu, Liping Luo and Boyun Yang*

Jiangxi Key Laboratory of Plant Resources, School of Life Science, Nanchang University, Nanchang, Jiangxi 330031, People's Republic of China

(Received: February 2019; Revised: April 2019; Accepted: May 2019)

Abstract

***Cymbidium kanran* is an important commercially grown member of the Chinese orchid family. However, little information regarding the molecular biology of this species is available. In this study, the *C. kanran* root, shoot, stem, leaf, and flower transcriptomes were sequenced with the Illumina HiSeq 4000 system, which resulted in 8.9 Gb of clean reads that were assembled into 74,620 unigenes, with an average length and N50 of 983 bp and 1,640 bp, respectively. The screening of seven databases (NR, NT, GO, KOG, KEGG, Swiss-Prot, and InterPro) for similar sequences resulted in the functional annotation of 49,813 unigenes. Additionally, 173 MADS-box genes, which help to control major aspects of plant development, were identified and their codon usage bias was analyzed. Only 26 genes had a low ENC ($\leq$35), suggesting the codon usage bias was weak. Base mutations were the major determinants of codon usage, although natural selection pressure also influenced codon usage bias. Moreover, 22 optimal codons were identified based on ΔRSCU, and 20 codons ended with A/U. The results of this study provide the foundation for the molecular breeding of new varieties.**

Key words: Codon usage bias; *Cymbidium kanran*; *de novo* transcriptome assembly; functional annotation

Introduction

Cymbidium kanran (Orchidaceae) is an important member of the Chinese orchid family because of its delicate fragrance, abundant colors, and long flowering period. However, the breeding of new *C. kanran* varieties has lagged behind the breeding in other ornamental plant species because of its long juvenile period. Molecular breeding techniques have become an important option because they can shorten the breeding period and enhance selection efficiency. However, they require many gene sequences and the identification of functional genes (Luo et al. 2014; Sullenberger et al. 2018). Unfortunately, there is only limited sequence information available for *C. kanran*. For example, there are currently only 240 nucleotide sequences or expressed sequence tags available for *C. kanran* in the NCBI database.

With the development of next-generation sequencing technologies, including Roche's 454 sequencing platforms and Illumina's Genome Analyzer systems, the transcriptome profiling of non-model organisms has become affordable and feasible (Escalona et al. 2016; Lee et al. 2017). Transcriptome analyses may generate abundant information regarding gene sequences, gene family differentiation, gene structure, and molecular markers (Jia et al. 2016; Zhang et al. 2015a). For example, the *de novo* characterization of the *Lilium* sp. 'Sorbonne' transcriptome resulted in the assembly of 39,636 unigenes, of which 30,986 were annotated, including 156 unigenes encoding key enzymes in the flavonoid biosynthesis pathway (e.g., chalcone synthase, chalcone isomerase, flavanone 3 hydroxylase, flavone synthase, and dihydroflavonol reductase). Moreover, 2,762 simple sequence repeats were identified with the MISA program (Zhang et al. 2015b).

Gene expression involves two stages: transcription and translation. During translation, gene information is transmitted by triplet codons, with each amino acid corresponding to 1–6 codons. Synonymous codons are codons that encode the same amino acid. Their presence in various genes of diverse species is not random, and codon usage bias has been confirmed

*Corresponding author's e-mail: 572460991@163.com

Published by the Indian Society of Genetics & Plant Breeding, A-Block, F2, First Floor, NASC Complex, IARI P.O., Pusa Campus, New Delhi 110 012; Online management by www.isgpb.org; indianjournals.com

(Presnyak et al. 2015). For example, there are six codons encoding arginine in the yeast genome, of which AGA is the most common (48%), with the other five synonymous codons (CGT, CGC, CGA, CGG, and AGG) occurring at a frequency of approximately 10% each (Trotta 2013).

Synonymous codon preference exists in all kinds of organisms. It reflects the evolution of genes and organisms and is associated with many biological phenomena (Kjaer et al. 2018). Therefore, there is considerable value to investigating codon usage bias. First, by studying codon bias, the optimal codons of specific gene types may be determined, after which the codons of a target gene can be modified to contain more of the optimal codons, thereby improving the expression of the gene in the host (Frumkin et al. 2018). Second, the function of unknown genes may be predicted by analyzing the correlation between codon usage bias and gene function (Liu et al. 2017). Third, the cluster analysis of codon preference parameters, such as relative synonymous codon usage (RSCU) and codon adaptation index (CAI), may be useful for characterizing horizontal gene transfer and gene family differentiation (Matsuo 2000).

In this study, a mixed library comprising RNA from *C. kanran* roots, shoots, stems, leaves, and flowers was sequenced with the Illumina platform, after which unigenes sequences were obtained by *de novo* assembly. A detailed functional annotation of the *C. kanran* transcriptome was also completed by screening seven databases (NR, NT, GO, KOG, KEGG, Swiss-Prot, and InterPro). Additionally, *C. kanran* MADS-box genes were identified, and their codon usage bias was thoroughly analyzed.

Materials and methods

Plant materials and RNA extraction

The seedlings of *Cymbidium kanran* 'Xuezhonghong' were cultured in a growth chamber with a 16-h light (25°C)/8-h dark (18°C) photoperiod and relative humidity of 75-85%. Leaf, root, and stem samples were collected at the 4 to 6 leaves stage (vegetative stage), whereas flowers were sampled at the bloom stage. All sampled plant tissues were immediately frozen in liquid nitrogen and kept frozen until processed. All analyses were completed with two biological replicates. Total RNA was extracted from the plant samples with the RNA plant reagent (Tiangen, Beijing, China).

Library construction and deep sequencing

Equal amounts of the total RNA from the collected root, shoot, stem, leaf, flower, and fruit samples were mixed, after which oligo (dT) was used to isolate mRNA to synthesize cDNA. Short fragments were purified and resolved with EB buffer for an end-repair step and the addition of adenine. Adapters were then added to the short fragments. Suitable fragments were selected for PCR amplification. The Agilent 2100 Bioanalyzer and ABI StepOnePlus Real-Time PCR System were used to quantify and assess the quality of the sample library. The library was subsequently sequenced with the Illumina HiSeq 4000 system.

De novo assembly and functional annotation of unigenes

Raw reads with adapters, reads with a high proportion of unknown bases (N > 5%), and other low-quality reads were removed with the SOAPdenovo program to obtain clean reads (Luo et al. 2015). SOAPdenovo was also used to integrate short reads into contigs with a specific overlap length (k-mer = 29). The read mate pairs were used to link the resulting contigs to scaffolds. To generate unigenes, the paired-end reads were reused to fill scaffold gaps (Liang et al. 2013).

The generated unigenes were compared with non-redundant sequences from the following databases with a BLASTX algorithm (E-value cutoff of 10^{-6}) (Li et al. 2015): NT (ftp://ftp.ncbi.nlm.nih.gov/blast/db), NR (ftp://ftp.ncbi.nlm.nih.gov/blast/db), KOG (http://www.ncbi.nlm.nih.gov/KOG), KEGG (http://www.genome.jp/kegg), and Swiss-Prot (http://ftp.ebi.ac.uk/pub/databases/swissprot). A published Perl script was used to annotate the unigenes based on the best hits (i.e., highest sequence similarity) (Sloan et al. 2012). The Bowtie program was used to map the high-quality reads to the annotated unigenes to determine unigene abundance (Toledo-Silva et al. 2013). The unigenes were functionally categorized with the GOslim tool of Blast2GO (Götz et al. 2008) according to the cellular component, biological process, and molecular function GO terms. Additionally, the InterProScan tool of Blast2GO along with the following databases were used to identify protein domains and gene families: Pfam, Gene3D, PRINTS, PANTHER, ProSITE, PIR, TIGERFAM, ProDom, SMART, and SUPERFAMILY (Cui 2011).

Identification of MADS-box sequences

The *C. kanran* MADS-box gene sequences comprising complete coding sequences (CDSs) were identified among the annotated unigenes. A Perl-based program was used to extract all CDSs, which were then examined to ensure they lacked a premature stop codon and contained the appropriate start and termination codons (Zhang et al. 2007; Swart et al. 2016). Additionally, we confirmed that the length of all CDSs was a multiple of 3.

Analysis of codon usage bias

The three stop codons (UGA, UAA, and UAG) and the single codons UGG (tryptophan) and AUG (methionine) were not included in the calculation of codon usage bias. All genes were evaluated based on the effective number of codons (ENC), which is similar to the factor used to calculate effective population size in population genetics, as well as the RSCU, which refers to the number of times a codon occurs in a gene relative to the number of expected occurrences under equal codon usage (Castells et al. 2017). Thus, the RSCU has often been used to standardize amino acid composition datasets and evaluate codon usage patterns (Velazquez-Salinas et al. 2016). A Perl-based program was used to calculate the GC content at the first, second, and third codon positions (GC1, GC2, and GC3). The frequency of GC in all codons in the sequence dataset was also calculated. As the average of GC1 and GC2, GC12 was applied for the neutrality plot analysis. The GC3 refers to the GC frequency at the third synonymously variable coding position (excluding termination codons and codons for tryptophan and methionine). The relationship between the GC3 and ENC for every gene was plotted to assess the influence of the GC content on codon usage (Gritsenko et al. 2013).

Determination of optimal codons

The unigenes were ranked according to their ENC, and the top and bottom 10% of unigenes were used to define the unigenes with high and low RSCUs, respectively. Optimal codons were those with △RSCU (i.e., difference between high and low RSCUs) > 0.08 (Zhang et al. 2007).

Results

De novo assembly of the C. kanran transcriptome

Approximately 8.9 Gb of sequence data was generated by the transcriptome sequencing completed with the Illumina HiSeq system. Specifically, 59.3 million high-quality reads (clean reads) were obtained following the removal of adapters, low-quality reads, short reads (< 45 bp), and primer sequences. The Trinity program was used to assemble the reads into 127,846 transcripts and 74,620 unigenes. All of the clean reads were deposited in the NCBI Short Read Archive database (accession number SRX8925700). The average unigene length and N50 were 983 bp and 1,640 bp, respectively (Table 1).

Table 1. Summary of the *Cymbidium kanran* transcriptome sequencing data obtained with the Illumina HiSeq system

Sequence	Number	Mean size	N50 size	Total nucleotides
Clean read	59,310,000	150	-	8,900,000,000
Transcripts	127, 846	633	1, 204	81, 028, 493
Unigen	74,620	983	1, 640	60, 894, 532

Transcriptome annotation

Of 74,620 unigenes, 49,813 (66.76%) were functionally annotated. The proportions of unigene with significant similarities to sequences in the NR, NT, GO, KOG, KEGG, Swiss-Prot, and InterPro databases were 63.36%, 54.52%, 45.41%, 47.67%, 47.77%, 44.79%, and 23.90%, respectively (Table 2). An examination of the species distribution of the top hits for the seven databases revealed that 60.13% of the annotated unigenes were most similar to sequences from *Dendrobium catenatum*, followed by *Phalaenopsis equestris* (23.95%) (Fig. 1a).

Table 2. Functional annotation of the assembled unigenes derived from the *Cymbidium kanran* transcriptome sequencing data

Category	Number	Percentage
Nr-Annotated	47,277	63.36%
Nt-Annotated	40,682	54.52%
GO-Annotated	33,882	45.41%
KOG-Annotated	35,574	47.67%
KEGG-Annotated	35,648	47.77%
Swissprot-Annotated	33,422	44.79%

On the basis of GO classifications, 33,882 matched unigenes were classified into the three main functional categories (biological process, cellular component, and molecular function). Regarding the unigenes in the biological process category, the most

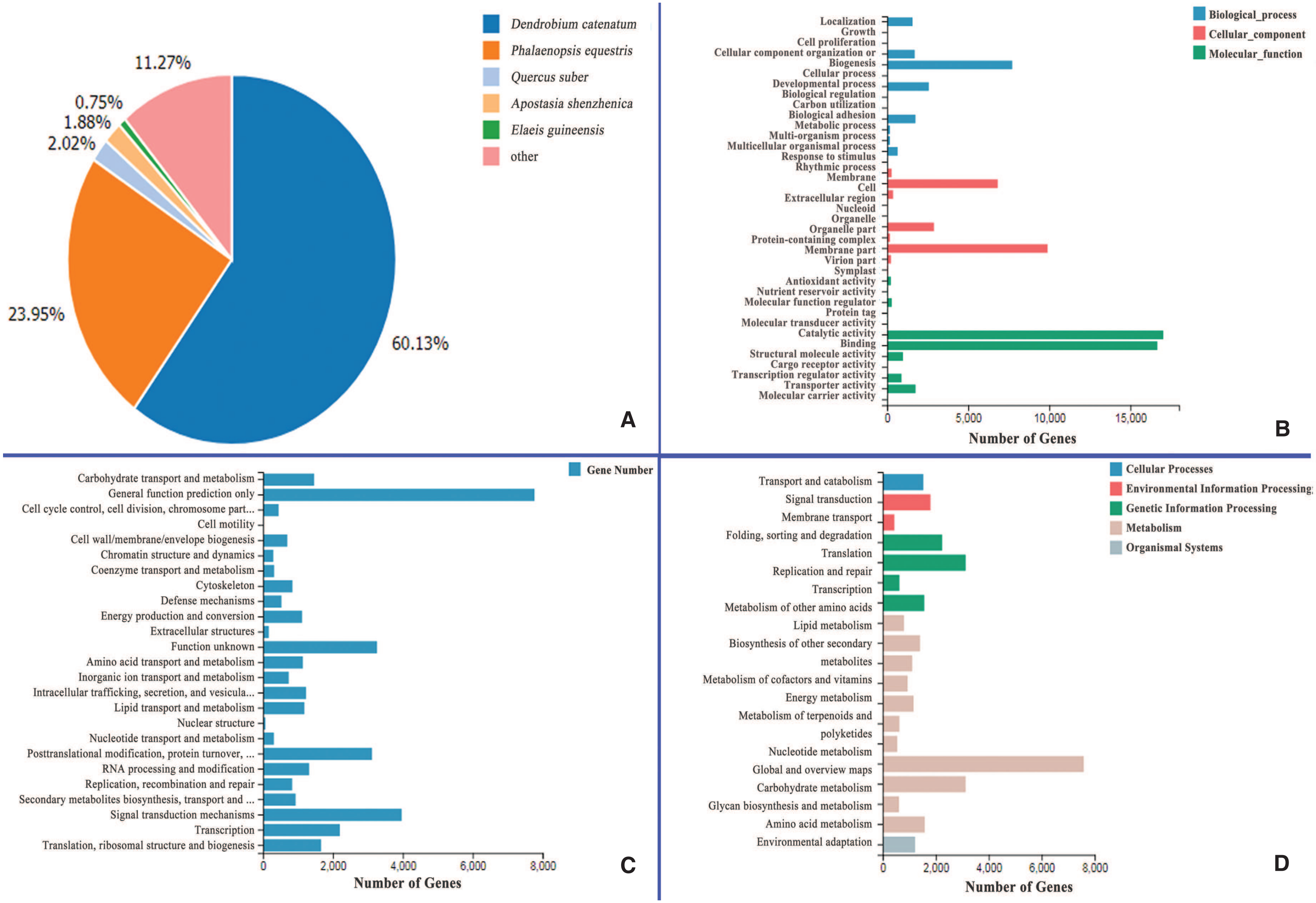

Fig. 1. Functional annotation of unigenes. (A) Distribution of annotated species. (B) Distribution of GO annotations. (C) Distribution of KOG annotations. (D) Distribution of KEGG annotations

commonly assigned GO terms were cellular process (7,714; 22.77%), biological regulation (2,568; 7.58%), and metabolic process (1,748; 5.16%). In the cellular component category, membrane part (9,893; 29.20%) and cell (6,816; 20.11%) were prominently represented. In terms of the molecular function category, catalytic activity (17,026; 50.25%) and binding (16,660; 49.17%) were the most common GO terms assigned to the unigenes (Fig. 1b).

All unigenes were used as queries to search the KOG database to predict functions and classifications. A total of 35,574 matched unigenes were classified into 25 categories. The unigenes in the same KOG categories were assumed to have been derived from the same ancestral gene or were paralogs or orthologs. The category with the most unigene was of general function prediction (7,766, 21.83%), followed by transcription (3,964; 11.14%) and signal transduction mechanisms (3,260, 9.14%). Nuclear structure (66; 1.86%) and cell motility (11; 0.31%) were the categories with the fewest unigenes (Fig. 1c).

The annotated unigenes were also used to screen the KEGG database to identify enriched biological pathways (Kanehisa et al. 2008). The 35,648 unigenes with significant matches in the KEGG database were mapped to 19 pathways representing the following five biochemical pathways (Fig. 1d): metabolism (19,507, 54.72%), genetic information processing (7,564, 21.21%), environmental information processing (2,237, 6.27%), cellular processes (1,532, 4.30%), and organismal systems (1,220, 3.42%). Regarding the unigenes in the metabolism category, most were associated with global and overview maps (7,587, 21.28%), carbohydrate metabolism (3,128, 8.77%), amino acid metabolism (1,583, 4.44%), lipid metabolism (1,407, 2.95%), energy metabolism (1,162, 3.26%), and biosynthesis of other secondary metabolites (1,113, 3.12%). The genetic information processing category included unigenes involved in translation (3,128, 8.77%), folding, sorting, and degradation (2,242, 6.29%), transcription (1,567, 4.40%), and replication and repair (627, 1.78%).

Analysis of the nucleotide composition of MADS-box genes

To investigate the nucleotide composition of *C. kanran* MADS-box genes, 173 MADS-box gene CDSs were identified with a Perl-based program. The GC1, GC2, and GC3 values were 0.397, 0.402, and 0.442, respectively, which revealed significant differences among the three synonymous codon positions. The GC content (GC_{all}) for the 173 MADS-box genes was 22.42%-63.02%, with an average of 41.4% (standard deviation of 0.076). The GC content for the unigenes in the 50th and 70th percentiles was 38.43%-46.16% and 33.83%-47.48%, respectively (Fig. 2a).

To study the relationships among the three codon positions, we calculated GC1, GC2, GC3, and GC12 for each unigene, and constructed neutrality plots (GC12 *vs* GC3) (Fig. 2b) (Sueoka 1988). The results revealed that GC3 varied considerably among the *C. kanran* MADS-box genes, ranging from 0.138 to 0.833. Additionally, the difference in GC3 usually reflected a neutral mutation bias, which altered the codon usage. Moreover, there was a significant positive correlation between GC_{all} and GC12, implying that the codon usage bias of *C. kanran* MADS-box genes was mainly affected by GC mutations.

The relationship between pyrimidines and purines at the third codon position was studied according to parity rule 2 analyses (Sueoka 1999; Kawabe and Miyashita 2003). If codon usage bias is determined by base mutations, the frequencies of G/C and A/T should be equal. In this study, the distribution of G/C and A/U at the third position was unequal, with a higher A content than G content in most MADS-box genes (Fig. 2c). These results indicated that the codon usage bias was also due to natural selection pressure.

Correlation between the ENC and GC3

To clarify the relationship between the ENC and GC3 in the absence of selection pressure, we drew a correlation curve with the calculated ENC and GC3 for each unigene (Fig. 2d). The data revealed that these values varied widely among the unigenes, with the ENC ranging from 23.2 to 61.0 and the GC3 ranging from 0.108 to 0.821. Additionally, there were 26 unigenes with a low ENC ($\leq$35, strongly biased) and 70 unigenes with a high ENC (50-61, weakly biased). The weakly and strongly biased unigenes represented 45.8% and 17.0% of the unigenes, respectively. An analysis of the correlation between the ENC and GC3 uncovered a negative correlation for all unigenes ($P < 0.05$, $r = 0.463$). The results suggested that codon usage may be influenced by a nucleotide composition mutational bias, and that codon usage is strongly biased for unigenes with a low ENC and a high GC3.

The correlation between the GC3 and ENC for

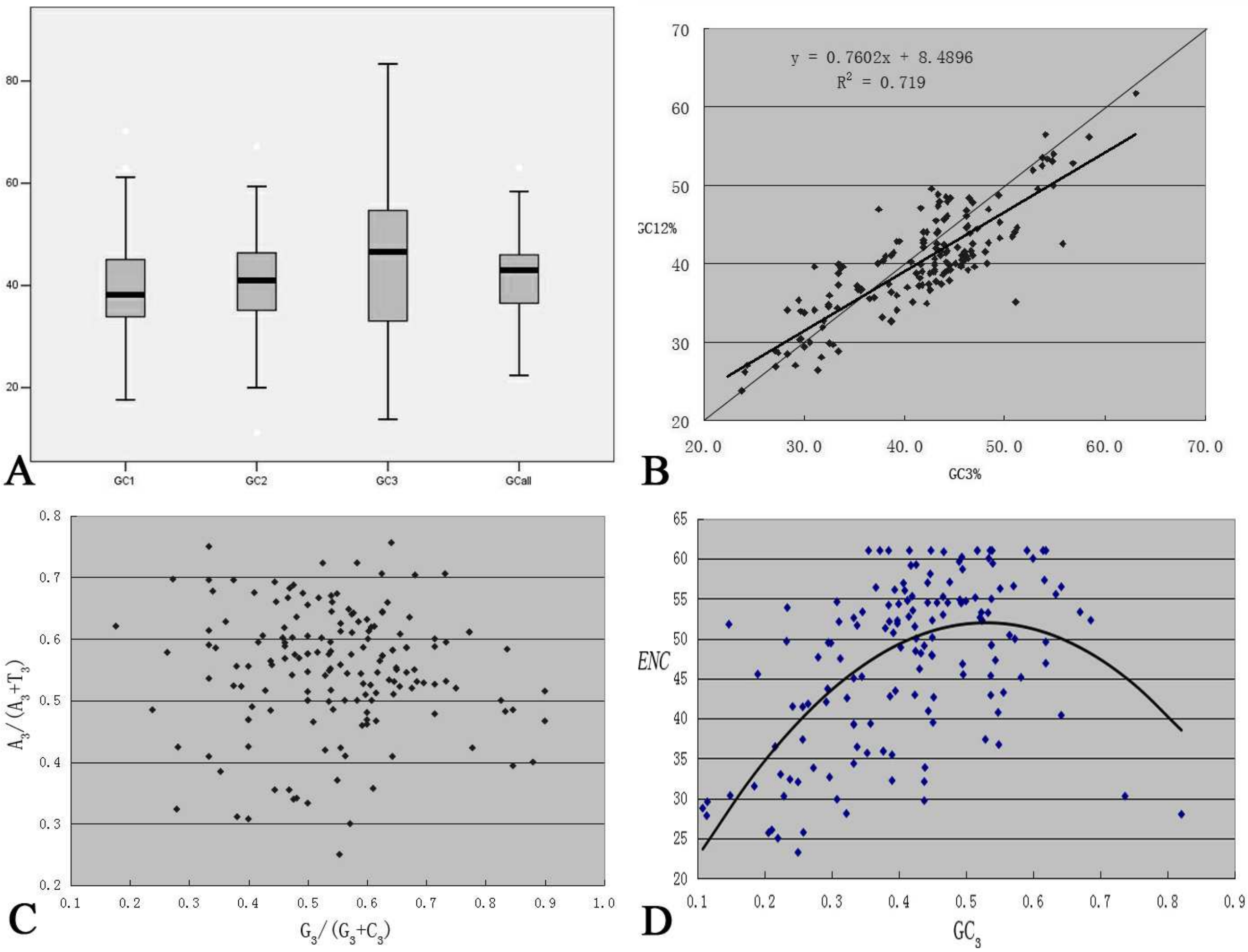

Fig. 2. Analysis of the codon usage bias of the MADS-box gene family. (A) GC content of the three codon positions. (B) Neutrality plot analysis. (C) Analysis of parity rule 2 bias. (D) Analysis of the relationship between ENC and GC3

each analyzed unigene is presented in Fig. 2d which indicates the codon usage patterns among unigenes that vary regarding their GC content. The curve refers to the expected unigene position where GC3 determines codon usage. Most unigenes were located on or near the reference line, but there were a few exceptions. Thus, GC3 appeared to be a major determinant of codon usage for *C. kanran* MADS-box genes, with natural selection pressure also having an effect.

Identification of optimal codons

On the basis of the ENC, the top and bottom 10% of unigenes represented the weakly biased and strongly biased unigene datasets, respectively. Additionally, 18 optimal codons were identified based on the ÄRSCU (Table 3). There was only one optimal codon for most of the amino acids, with the exceptions being isoleucine and leucine. Regarding the optimal codons, 16 ended with A/U, whereas only two ended with G/C. Therefore, the *C. kanran* MADS-box genes are biased toward codons ending with A/U.

Discussion

A *de novo* sequence assembly revealed 74,620 unigenes, of which 49,813 were functionally annotated. A total of 36,555 CDSs were obtained based on the functional annotation results and ESTscan predictions. Additionally, 1,917 SSRs distributed on 8,327 unigenes were detected with the MISA program (Zhang et al. 2017). Thus, the results of this study may provide relevant genetic information for future functional gene mining studies, and facilitate research into genetic

Table 3. Preferred codons in the *Cymbidium kanran* MADS-box genes

Amino acid	Codon	Amino acid	Codon	Amino acid	Codon
Ala	GCU	Ile	AUU	Phe	UUU
Arg	AGA	Ile	AUA	Ser	UCA
Asn	AAU	Leu	CUA	Ser	UCU
Gln	CAA	Leu	UUA	Thr	ACU
Glu	GAG	Leu	UUG	Tyr	UAU
Gly	GGU	Lys	AAA	Val	GUU

structures as well as molecular marker-assisted breeding of *C. kanran*.

The MADS-box genes encode transcription factors with fundamental roles related to developmental control or signal transduction processes, especially regarding flower development (Wang et al. 2017). A *de novo* transcriptome sequence assembly uncovered 173 MADS-box gene CDSs, which were then used to investigate codon usage bias in the MADS-box gene family. The results of association analyses (GC3 *vs* GC12 and GC3 *vs* ENC) and an examination of the proportions of pyrimidines and purines indicated that MADS-box genes exhibit a weak codon usage bias that is mainly due to base mutations but also influenced by natural selection pressure. The GC content in the unigenes consistent with those for rice, maize, and *Brachypodium distachyon* (Carels and Bernardi 2000; Wang and Hickey 2007; Liu H et al. 2010), but contradicted the data in previous reports regarding monocots (Carels and Bernardi 2000; Wang and Hickey 2007; Liu H et al. 2010). Most of the unigenes were distributed along the diagonal, indicating that base mutations were the major factor responsible for the codon usage bias rather than natural selection (Zhou et al. 2009). These results suggested that neutral mutation biases influenced codon selection.

Previous studies confirmed there are many factors affecting codon usage bias, including base mutations, natural selection, coding region length, and tRNA abundance (Taylor et al. 2017; Qin et al. 2017; Luo et al. 2016). Mutations are generally neutral and do not affect normal physiological mechanisms. However, the codon changes caused by natural selection may improve the efficiency of gene expression and the dynamics of polypeptide chain folding for specific genes or specific gene regions (Gingold et al. 2011; Sauna et al. 2011). Base mutations and natural selection pressure are the two major factors affecting the codon usage bias in plant nuclear genes, with natural selection having a more prominent role in some plant species, such as *Nicotiana tabacum*, *Picea asperata*, and *Silene pendula* (De La Torre et al. 2015; Qiu et al. 2010). In contrast, base mutations are the main factor affecting codon usage bias in some plant species. However, base mutations and natural selection pressure often simultaneously influence codon usage in many plant species, including *Arabidopsis thaliana* and *Oryza sativa* (Qiu et al. 2011; Mukhopadhyay et al. 2008). Therefore, the reasons for the codon usage bias in various species or genes are diverse, and the underlying evolutionary mechanism in plants remains to be characterized.

The codon usage bias reportedly differs between monocotyledonous and dicotyledonous species, with the former exhibiting a bias toward codons ending with G/C and the latter having a bias toward codons ending with A/U. However, the codon usage of Orchidaceae plant species appears to be biased toward codons ending with A/U. For example, the CDS of the *Dendrobium officinale* gene encoding UDP-glucose pyrophosphorylase has a higher AT content than a GC content, and 19 of the 27 biased codons end with A/U, similar to the frequently used codons (RSCU > 2) (Jing et al. 2014).

On the basis of the *C. kanran* codon usage bias described herein, the codons of MADS-box genes from other species may be modified to improve the efficiency of transcription and translation of these genes in transgenic *C. kanran* plants. Therefore, the results of this study may provide the foundation for the molecular breeding of new orchid varieties with enhanced traits (Li et al. 2017), especially for the breeding of transgenic varieties with modified MADS-box genes.

Authors' contribution

Conceptualization of research (HL, BY); Designing of the experiments (HL); Contribution of experimental materials (BY); Execution of field/lab experiments and data collection (HL, YT); Analysis of data and interpretation (HL, WY); Preparation of manuscript (HL, YT).

Declaration

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 31360491) and the Jiangxi Natural Science Foundation (Grant No. 20171ACB21014). We thank William Yajima, a PhD in Plant Science of University of Alberta, for editing the English text of a draft of this manuscript.

References

Castells M., Victoria M., Colina R., Musto H. and Cristina J. 2017. Genome-wide analysis of codon usage bias in Bovine Coronavirus, Virol. J., **14**(1): 115-123.

Cui Y. 2011. Bioinformatics Tools for Gene Function Prediction, Gene Discovery for Disease Models, 93.

De La Torre A. R., Lin Y. C., Van de Peer Y. and Ingvarsson P. K. 2015. Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in *picea* gene families, Genome Biol. Evol., **7**(4): 1002-1015.

Escalona M., Rocha S. and Posada D. 2016. A comparison of tools for the simulation of genomic next-generation sequencing data, Nat. Rev. Genet., **17**(8): 459-469.

Frumkin I., Lajoie M. J., Gregg C. J., Hornung G., Church G. M. and Pilpel Y. 2018. Codon usage of highly expressed genes affects proteome-wide translation efficiency, Proc. Natl. Acad. Sci. USA, **115**(21): 4940-4949.

Götz S., García-Gómez J. M., Terol J., Williams T. D., Nagaraj S. H., Nueda M. J., Robles M., Talón M., Dopazo J. and Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite, Nucleic Acids Res., **36**(10): 506-513.

Gingold H. and Pilpel Y. 2011. Determinants of translation efficiency and accuracy, Mol. Sys. Biol., **7**(1): 481-493.

Gritsenko A. A., Reinders M. J. and de Ridder D. 2013. Using predictive models to engineer biology: A case study in codon optimization. In: IAPR International Conference on Pattern Recognition in Bioinformatics, 159-171.

Jia X., Deng Y., Sun X., Liang L. and Su J. 2016. De novo assembly of the transcriptome of Neottopteris nidus using Illumina paired-end sequencing and development of EST-SSR markers, Mol. Breed., **36**(7): 94-105.

Jing S., Tao H. E., Wan R., Peng X. and Ze C. 2014. The Codon usage bias of UDP-glucose pyrophosphorylase gene (UGP) in Dendrobium officinale, Chinese J. Appl. Environ. Biol., **20**(5): 759-766.

Kjaer J. and Belsham G. J. 2018. Selection of functional 2A sequences within foot-and-mouth disease virus; requirements for the NPGP motif with a distinct codon bias, RNA, **24**(1): 12-17.

Lee S. J., Ban S. H., Kim G. H., Kwon S. I., Kim J. H. and Choi C. 2017. Identification of potential gene-associated major traits using GBS-GWAS for Korean apple germplasm collections, Pl. Breed., **136**(6): 977-986.

Li G., Zhao Y., Liu Z., Gao C., Yan F., Liu B. and Feng J. 2015. De novo assembly and characterization of the spleen transcriptome of common carp (*Cyprinus carpio*) using Illumina paired-end sequencing, Fish and Shellfish Immunol., **44**(2): 420-429.

Li J., Li H., Zhi J., Shen C., Yang X. and Xu J. 2017. Codon Usage of Expansin Genes in Populus trichocarpa, Current Bioinformatics, **12**(5): 452-461.

Liang C., Liu X., Yiu S.-M. and Lim B. L. 2013. De novo assembly and characterization of Camelina sativa transcriptome by paired-end sequencing, BMC Genomics, **14**(1): 146-156.

Liu H., Rahman S. U., Mao Y., Xu X. and Tao S. 2017. Codon usage bias in 5' terminal coding sequences reveals distinct enrichment of gene functions, Genomics, **109**(2017): 506-513.

Luo H., Luo K., Luo L., Xiang Li E., Guan B., Xiong D., Sun B., Peng K. and Yang B. 2014. Evaluation of candidate reference genes for gene expression studies in Cymbidium kanran, Scientia horticulturae, **167**(2014): 43-48.

Luo H., Zhang Y., Luo L., Xiong D. and Yang B. 2016. The Codon Usage Bias of FLOWERING LOCUS T Gene (FT) in Orchidaceae, Molecular Pl. Breed., **14**(1): 51-58.

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q. and Liu Y. 2015. Erratum: SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler, Giga Sci., **4**(1): 30.

Matsuo Y. 2000. Evolutionary change of codon usage for the histone gene family in Drosophila melanogaster and Drosophila hydei, Molecular Phylogenetics and Evolution, **15**(2): 283-291.

Mukhopadhyay P., Basak S. and Ghosh T. C. 2008. Differential Selective Constraints Shaping Codon Usage Pattern of Housekeeping and Tissue-specific Homologous Genes of Rice and Arabidopsis, DNA Res., **15**(6): 347-356.

Presnyak V., Alhusaini N., Chen Y.-H., Martin S., Morris N., Kline N., Olson S., Weinberg D., Baker K. E., Graveley B. R. and Coller J. 2015. Codon Optimality Is a Major Determinant of mRNA Stability, Cell, **160**(6): 1111-1124.

Qin W. Y., Gan L. N., Xia R. W., Sun S. Y., Zhu G. Q., Wu S. L. and Bao W. B. 2017. New insights into the codon usage patterns of the bactericidal/permeability-increasing (BPI) gene across nine species, Gene, **616**(2017): 45-51.

Qiu S., Bergero R., Zeng K. and Charlesworth D. 2010. Patterns of codon usage bias in Silene latifolia, Molecular Biology and Evolution, **28**(1): 771-780.

Qiu S., Zeng K., Slotte T., Wright S. and Charlesworth D. 2011. Reduced efficacy of natural selection on codon usage bias in selfing Arabidopsis and Capsella species, Genome Biol. Evol., **3**: 868-880.

Sauna Z. E. and Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease, Nature Rev. Genet., **12**(10): 683-691.

Sloan D. B., Keller S. R., Berardi A. E., Sanderson B. J., Karpovich J. F. and Taylor D. R. 2012. De novo transcriptome assembly and polymorphism detection in the flowering plant *Silene vulgaris* (Caryophy llaceae), Molecular Ecology Res., **12**(2): 333-343.

Sullenberger M. T., Jia M., Gao S. and Foolad M. R. 2018. Genetic analysis of late blight resistance in Solanum pimpinellifolium accession PI 270441: Heritability and response to selection, Plant Breed., **137**(1): 89-96.

Swart E. C., Serra V., Petroni G. and Nowacki M. 2016. Genetic codes with no dedicated stop codon: Context-dependent translation termination, Cell, **166**(3): 691-702.

Taylor T. L., Dimitrov K. M. and Afonso C. L. 2017. Genome-wide analysis reveals class and gene specific codon usage adaptation in avian paramyxoviruses 1, Infection, Genetics and Evolution, **50**: 28-37.

Toledo-Silva G., Cardoso-Silva C. B., Jank L. and Souza A. P. 2013. De novo transcriptome assembly for the tropical grass Panicum maximum Jacq, PLoS One, **8**(7): e70781.

Trotta E. 2013. Selection on codon bias in yeast: A transcriptional hypothesis, Nucleic Acids Res., **41**(20): 9382-9395.

Velazquez-Salinas L., Zarate S., Eschbaumer M., Lobo F. P., Gladue D. P., Arzt J., Novella I. S. and Rodriguez L. L. 2016. Selective factors associated with the evolution of codon usage in natural populations of arboviruses, PLoS One, **11**(7): e0159943.

Wang N., Xing Y., Lou Q., Feng P., Liu S., Zhu M., Yin W., Fang S., Lin Y. and Zhang T. 2017. Dwarf and short grain 1, encoding a putative U-box protein regulates cell division and elongation in rice, J. Pl. Physiol., **209**: 84-94.

Zhang L., Wan X., Xu J., Lin L. and Qi J. 2015a. De novo assembly of kenaf (*Hibiscus cannabinus*) transcriptome using Illumina sequencing for gene discovery and marker identification, Mol. Breed., **35**(10): 192-202.

Zhang M. F., Jiang L. M., Zhang D. M. and Jia G. X. 2015b. De novo transcriptome characterization of Lilium 'Sorbonne' and key enzymes related to the flavonoid biosynthesis, Mol. Genet. Genomics, **290**(1): 399-412.

Zhang W. J., Zhou J., Li Z. F., Wang L., Gu X. and Zhong Y. 2007. Comparative analysis of codon usage patterns among mitochondrion, chloroplast and nuclear genes in *Triticum aestivum* L., J. Integrative Pl. Biol., **49**(2): 246-254.

Zhang Y., Yang B., Xiong D. and Luo H. 2017. Analysis on SSR information based on transcriptome and development of molecular markers in Cymbidium kanran, J. Nanchang Univ., **41**(3): 249-254.

Zhou M. and Li X. 2009. Analysis of synonymous codon usage patterns in different plant mitochondrial genomes, Mol. Biol. Reports, **36**(8): 2039-2046.