Comparative analysis of the codon usage patterns in two closely related *Marsupenaeus* species based on comparative transcriptomics

Panpan Wang

Jiangsu ocean university

Yong Mao (✉ maoyong@xmu.edu.cn)

Xiamen University <https://orcid.org/0000-0002-0522-0614>

Yongquan Su

Xiamen university

Jun Wang

Xiamen university

Research article

Keywords: Codon usage pattern, *Marsupenaeus japonicus*, *Marsupenaeus pulchricaudatus*, Orthologous genes, Phylogenetics

Posted Date: October 9th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-88141/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: *Marsupenaeus japonicus*, a major commercial shrimp species in the world, has two cryptic or sibling species, *Marsupenaeus japonicus* and *Marsupenaeus pulchricaudatus*. Due to the lack of genomic information, little is known about the correlations among codon usage bias, gene expression, and evolutionary trends in *Marsupenaeus* orthologs.

Results: Using the CodonW 1.4.2 software, we performed the codon bias analysis of two *Marsupenaeus* species transcriptomes. The average contents of GC and ENc were 51.61% and 52.1 for VI (*M. japonicus*), 51.54% and 52.22 for VII (*M. pulchricaudatus*), respectively. Parity Rule 2 (PR2) plot analysis showed that purines (A and G) were used more frequently than pyrimidines (C and T) in two *Marsupenaeus* species. The average ENc value was 52.1 and 52.22 for *M. japonicus* and *M. pulchricaudatus*, respectively. Overall, orthologous genes that underwent positive selection ($\omega > 1$) had a higher correlation coefficient than that experienced purifying selection ($\omega < 1$). In *M. japonicus*, the relationships were highly significant positive about Axis 1 and A3, T3 and ENc ($p < 0.01$). However, all relationships in *M. pulchricaudatus* were the opposite. We determined 12 and 14 optimal codons for *M. japonicus* and *M. pulchricaudatus*, respectively. Two *Marsupenaeus* species had 31 different codon pairs. The results of multi-species clustering based on codon preference were consistent with traditional classification.

Conclusions: We characterized the codon usage patterns of the two *Marsupenaeus* species and the evolutionary trends in *Marsupenaeus* orthologs, which provides new insights into the genetic divergence and the phylogenetic relationships of two *Marsupenaeus* species.

Background

The codon is the basic genetic code of all biological messenger RNA and 62 codons encode 20 different amino acids [1–3]. For different genes or genomes, the selection of synonymous codons is non-random, which is called codon bias (CUB) [4, 5]. Codon preference is specific to the organism and may be influenced by GC content, gene expression level, and gene length [6–8]. Besides, codon usage patterns may affect the biological functions of mRNA biosynthesis, translation elongation rate, protein folding, and other downstream expressions [7, 9–12]. It is now thought that the CUB is mainly affected by natural selection and mutational pressure [13–16].

Most of the research in the past mainly studied the codon preference of species with genome-wide information [17–19]. Recent rapid development in next-generation sequencing has provided large amounts of genomic and transcriptome data. Machado et al detected and quantified strong selection on synonymous sites of *Drosophila melanogaster* by using deep genomic population sequencing [20]. Utilizing the Robo-seq and RNA-seq data, Chu et al studied how codon usage bias could impact the translation patterns of *Arabidopsis thaliana* [21]. Guan et al analyzed codon usage of *Hirudinaria manillensis* RNA-seq data and found that the genetic evolution was driven by mutation pressure and

natural selection [22]. Additional studies of codon usage bias based on transcriptome data include *Bombyx mori* [23], *Taenia multiceps* [24], and *Megalobrama amblycephala* [25].

The kuruma shrimp (*Marsupenaeus japonicus*) includes two cryptic species, which are distributed mostly allopatrically but co-occur in the northern South China Sea [26]. Our previous study showed that there were obvious genetic differentiation and reproductive isolation between the two cryptic species [27]. We sequenced and compared the transcriptomes of two *Marsupenaeus* species, and obtained a large number of putative orthologs [28]. Yi et al compared and analyzed the codon bias of three *Misgurnus* species and revealed the potential importance of expression-mediated selection in shaping the genome evolution [29]. Song et al found that the complex correlations among gene expression, codon usage bias, and substitution rate in *Arachis duranensis* and *Arachis ipaënsis* orthologs [30].

In this study, we performed codon usage bias analysis based on two *Marsupenaeus* species transcriptomes using the CodonW software. We systematically compared the codon usage patterns of the two *Marsupenaeus* species and evaluated the comprehensive effects of various factors. Moreover, the results provided new insights into the genetic divergence and the phylogenetic relationships of two *Marsupenaeus* species.

Methods

Data collection and filtering

Raw Illumina sequences are accessible from NCBI Sequence Read Archive (SRA) (<https://trace.ncbi.nlm.nih.gov/Traces/sra/>) under accession SRR7786082 (*Marsupenaeus pulchricaudatus*) and SRR7786083 (*Marsupenaeus japonicus*). Transcript assembly and functional annotation were described in detail in our previous work [28]. A total of 14126 and 13695 unigenes with full-length CDS (Coding sequence) regions were identified from *M. japonicus* and *M. pulchricaudatus* libraries, respectively. Orthologous groups were screened using OrthoMCL with default settings [31]. Gene expression levels as fragments per kilobase million [32] were estimated by RSEM software [33]. Coding sequences of the other 9 species were downloaded from NCBI. All CDSs were filtered using the Omics hare online platform (<ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>), and those sequences with length less than 400 bp or unknown bases were eliminated.

Codon usage indices analysis

The GC1, GC2, GC3 contents were calculated using Perl scripts, and GC12 was the average value of GC1 and GC2. Using the CodonW 1.4.2 software (<http://codonw.sourceforge.net>), we performed codon bias analysis. The calculation indices included GC content, the nucleotide composition at the 3rd codon position (A3s, T3s, G3s, and C3s), the effective number of codons (ENc), the codon adaptation index (CAI), codon bias index (CBI), frequency of optional codon (Fop), relative synonymous codon usage (RSCU), and so on. The Parity Rule 2 (PR2) plot analysis was based on the third codon position, using

A3/(A3 + T3) as the ordinate and G3/(G3 + C3) as the abscissa. The PR2 plot can be used to estimate the impact of selection and mutation pressure on codon usage bias [34].

ENc-plot and GO annotation

The effective number of codons (ENc), with a value between 20 and 61, is a key parameter to interpret codon bias. The value 20 indicates that only one synonymous codon is chosen and 61 represent no usage bias and all synonymous codons have the same probability. The lower the value for a coding sequence, the stronger the codon usage bias [35, 36]. In general, a gene possesses strong codon usage bias when the ENc value lower than 35 [37, 38]. The ENc-plot was drawn by Origin 2020 (OriginLab Corporation, USA), which uses ENc value as the ordinate and GC3s as the abscissa. The expected ENc values were calculated based on the equation: $ENc(exp) = 2 + GC3s + 29/[GC3s^2 + (1 - GC3s)^2]$ [39]. CAI was an important index for estimating synonymous codon usage bias and gene expression level, and the higher CAI value signifies the stronger codon usage bias [40–42]. Gene ontology annotation was performed using Blast2GO v2.5 ($E\text{-value} < 1e^{-6}$) [43]. GO classifications were compared among different groups of GC3s (High, Mid, and Low) using the Omicshare online platform (<https://www.omicshare.com/tools/>).

Correlation analysis

The codon usage patterns were often shaped by many factors, such as GC content, expression level, tRNA abundance, protein structure, and hydrophilicity [44, 45]. We performed a correlation analysis between codon bias parameters and expression level (FPKM). Using the PAML toolkit [46], we calculate the non-synonymous substitution ratio (Ka) and synonymous substitution ratio (Ks). In genetics, the Ka/Ks (ω) can be used to determine whether there is selective pressure on protein-encoding genes [47, 48]. The $\omega > 1$ suggests that the gene evolved under positive selection and the ω closing to zero indicates that the gene is under heavy selection pressure [47, 49].

Correspondence analysis (COA)

To further investigate the factors related to the codon usage pattern, the correspondence analysis was conducted by CodonW based on the RSCU values. The COA was used to compare the usage patterns of 59 codons (except Met, Trp, Taa, Tag, and Tga) and reflect the variation trend in codon usage. COA creates a series of orthogonal axes, which were used to estimate the main source of variation. Using SPSS v22, the relative coefficient between ten codon bias parameters and axis1 and axis2 was calculated, respectively.

Relative synonymous codon usage and optimal codons

According to Sharp et al [50], the relative synonymous codon usage (RSCU) was an index to measure the codon usage preference. The higher the RSCU value, the stronger the preference. Based on the calculated ENc values, 10% of the genes with extremely high and low ENc values were regarded as the high and low RSCU datasets [51]. The optimal codons were confirmed based on the Δ RSCU value and chi-square test [38, 52, 53].

Clustering and principal component analysis

The protein-coding sequences of nine species were downloaded from the ensemble database (<http://asia.ensembl.org/index.html>) and NCBI (<https://www.ncbi.nlm.nih.gov/>), and codon usage preference was analyzed using CodonW. The heatmap was performed based on RSCU values using the Omicshare online platform (<https://www.omicshare.com/tools/>). Based on the RSCU of 59 codons, the Principal component analysis (PCA) of 11 species were performed by Origin 2020 (OriginLab Corporation, USA).

Results

Nucleotide composition and PR2-plot analysis

A total of 9414 and 9420 unigenes with length larger than 400 were screened from *M. japonicus* and *M. pulchricaudatus* libraries, respectively (Fig. 1A). The length distribution of the two groups was similar. In the *M. japonicus*, the mean contents of A and T nucleotide were 31.89% (SD = 10.45%) and 30.8% (SD = 9.04%), and the mean contents of C and G nucleotide were 33.63% (SD = 10.71%) and 28.03% (SD = 9.21%). In the *M. pulchricaudatus*, the average contents of A and T nucleotide were 31.86% (SD = 10.55%) and 30.8% (SD = 9.18%), and the average contents of C and G nucleotide were 33.56% (SD = 10.73%) and 28.27% (SD = 9.29%) (Fig. 1B). The average contents of GC were 51.61% and 51.54% for *M. japonicus* and *M. pulchricaudatus*, respectively. The mean contents of GC3s were 49.1% and 49.17% for *M. japonicus* and *M. pulchricaudatus*, respectively, which were significantly higher than that of GC12. The median of GC-biases [$G3/(G3 + C3)$] were 0.4563 and 0.4582 for *M. japonicus* and *M. pulchricaudatus*, and the median of AT-biases [$A3/(A3 + T3)$] were 0.5047 and 0.5051, respectively (Fig. 1C and D). Parity Rule 2 (PR2) plot analysis showed that purines (A and G) were used more frequently than pyrimidines (C and T) in two *Marsupenaeus* species. The unbalanced use of the third base suggested that mutation pressure and natural selection contribute to codon usage bias.

Correlation analysis of codon usage parameters

All parameters have similar correlations between *M. japonicus* and *M. pulchricaudatus* (Fig. 2). The results indicated that FPKM was negatively correlated with T3s and A3s ($p < 0.05$), and positively correlated with other parameters ($p < 0.05$) in *M. japonicus* and *M. pulchricaudatus*. There was a significant ($p < 0.05$) positive correlation among T3s, A3s, and ENc values. These three values were negatively ($p < 0.05$) correlated with other parameters. Correlation analysis indicated that the third base content of synonymous codons significantly affects gene expression and codon usage bias. The significant correlation ($p < 0.05$) between GC3 and GC content indicated that the nucleotide contents play an important role in codon usage bias. The first and second base content was often determined by natural selection and the third base content was affected by mutation pressure [54, 55].

The average ENc values were 52.1 and 52.22 for *M. japonicus* and *M. pulchricaudatus*, respectively. There were 268 (2.85%) and 249 (2.64%) genes with ENc value equal to 61 for *M. japonicus* and *M. pulchricaudatus*, which represent all synonymous codons have the same probability. There were 187 (1.99%) and 133 (1.41%) genes with ENc value less than 35 for *M. japonicus* and *M. pulchricaudatus*, and the minimum values were 23.6 and 27.46, respectively. The 35 value was the standard of codon bias [35, 37]. To explore the effect of GC3s on the codon usage bias, we performed the ENc-plot analysis. In figure S1A and S1B, most genes were aggregated close to the expected curve, which showed that codon usage bias was mainly affected by mutation pressure. We found lower ENc values in *M. japonicus* than *M. pulchricaudatus*. Meanwhile, we estimated the difference between the expected and the observed ENc values and calculated the $(\text{ENc}_{\text{exp}} - \text{ENc}_{\text{obs}})/\text{ENc}_{\text{exp}}$ (Fig. S1C and S1D). The frequencies distribution of unigenes with values within 0 ~ 0.1 were highest, which showed that most ENc values from GC3s were larger than observed ENc values. Moreover, there was a significant positive correlation between GC3s and CAI values (Fig. S1E and S1F).

Go ontology (GO) annotation based on GC3s

To further understand the influence of GC3s on the gene function, the CDSs with low, mid, and high GC3, including 1000, 1001, and 1005 genes in *M. japonicus* and 1005, 1001, and 1002 genes in *M. pulchricaudatus*, were performed GO annotation. Gene ontology terms provided three categories, including cellular component, molecular function, and biological process. Overall, these three groups' genes in two shrimp species presented similar functional categories (Fig. S2). In the biological process categories, with 13 subtypes, most corresponding genes were involved in cellular process, metabolic process, single-organism process, and biological regulation. Thirteen subtypes were annotated with the cellular component, and the highest gene number was observed in "cell part" and "cell" categories. In the molecular function, the "binding" was the highest category.

Correlation analysis between codon usage parameters and the substitution rate

A total of 5036 pairs of single-copy orthologous genes were identified between *M. japonicus* and *M. pulchricaudatus* libraries [28]. Among these orthologs, the Ka/Ks values of 2491 pairs were calculated with the mean values were 0.002, 0.019, and 0.175 for Ka , Ks , and Ka/Ks (ω), respectively. There were 49 pairs of orthologous genes with a ω value greater than 1 (positive selection), and 2225 pairs with a ω value less than 1 (purifying selection).

Overall, orthologous genes that underwent positive selection ($\omega > 1$) had a higher correlation coefficient than that experienced purifying selection ($\omega < 1$), which could be more genes with $\omega < 1$ lead to large differences. Almost all parameters had different significance levels with Ka , Ks , or Ka/Ks (Fig. 3). In the *M. japonicus*, the Ka/Ks of genes with ω less than 1 was positively correlated with ENc, A3s, and T3s ($p < 0.01$), but negatively correlated with other parameters ($p < 0.01$). There was no significant correlation between all parameters with the Ka/Ks of genes with ω greater than 1. However, GC content and CBI value were positively correlated with Ks , and G3s was negatively correlated with Ks . Besides, Fop and CBI

values were positively correlated with K_a . In *M. pulchricaudatus*, the K_a/K_s of genes with ω less than 1 were positively correlated with A3s and T3s but negatively correlated with other parameters. Like *M. japonicus*, there was no significant correlation between all parameters with the K_a/K_s of genes with ω greater than 1. However, CBI and T3s values were positively correlated with K_a and K_s .

Correspondence analysis (COA)

Using the RSCU values, the correspondence analysis was used to investigate the factors related to codon usage patterns and reflect the variation trend in codon usage. The result indicated that the first five axes accounted for 43.8% and 44.3% of the amino-acid variation for *M. japonicus* and *M. pulchricaudatus* (Fig. 4A). In *M. japonicus*, Axis 1 and Axis 2 explain 25.16% and 6.54% of the variance, respectively. In *M. pulchricaudatus*, Axis 1 and Axis 2 explain 26.38% and 6.29% of the variance, respectively. In *M. japonicus*, the relationships were highly significant positive between Axis 1 and A3, T3, and ENc ($p < 0.01$), and others were significantly negative correlated ($p < 0.01$). However, all relationships in *M. pulchricaudatus* were opposite (Fig. 4B).

To identify the effect of GC content on codon bias, GC contents of genes were color-coded on the plot, which uses Axis 1 as the abscissa and Axis 2 as the ordinate (Fig. 4C for *M. japonicus* and 4D for *M. pulchricaudatus*). Overall, the distribution of GC content was the opposite along Axis 1. In *M. japonicus*, the larger value of Axis 1, the smaller the GC content. The negative correlation (-0.562 with p -value < 0.01) between Axis 1 and GC content was presented in the Fig. 6B. Instead, the larger value of Axis 1, the larger the GC content in *M. pulchricaudatus*, and the positive correlation was 0.7 ($p < 0.01$).

Determination of optimal codons

There were 32 codons with the RSCU value > 1 in *M. japonicas* and *M. pulchricaudatus*, respectively, which indicated that these codons were preferences for two *Marsupenaeus* species (Table S1). Except for Trp and Met, the codons of Ala, Arg, Gly, Pro, Ser, and Thr had a higher bias. Besides, the codons with the RSCU value > 1 mainly end with C and A. Based on ENc values, we obtained the RSCU datasets of high and low expression genes and calculated the Δ RSCU value (Table S2). We determined 12 and 14 optimal codons for *M. japonicus* and *M. pulchricaudatus*, respectively (Table 1). In *M. japonicus*, 9 optimal codons were C-ending and 3 optimal codons were G-ending. In *M. pulchricaudatus* species, 9 optimal codons were C-ending and 5 optimal codons were G-ending. Most optimal codons were the same in two *Marsupenaeus* species. *M. japonicus* had a special ACC (Thr) codon, and *M. pulchricaudatus* had special CCG (Pro), GCG (Ala), and GGC (Gly) codons.

Table 1
The optimal codons based on high and low levels of expression. AA: amino acids.

		<i>M. japonicus</i>			<i>M. pulchricaudatus</i>		
AA	Codon	RSCU-H	RSCU-L	Δ RSCU	RSCU-H	RSCU-L	Δ RSCU
Val	GUC	1.557	0.995	0.563	1.454	0.984	0.470
Ser	UCG	1.005	0.682	0.324	1.202	0.717	0.486
Pro	CCC	1.552	0.906	0.646	1.764	0.920	0.844
	CCG	0.896	0.614	0.282	1.047	0.606	0.441
Thr	ACC	1.756	0.959	0.797	1.627	1.001	0.626
	ACG	1.170	0.698	0.472	1.356	0.720	0.636
Ala	GCG	0.836	0.564	0.271	1.002	0.537	0.464
Tyr	UAC	1.602	0.944	0.658	1.609	0.943	0.666
His	CAC	1.551	0.986	0.565	1.541	0.998	0.544
Asn	AAC	1.605	0.966	0.639	1.590	0.998	0.591
Asp	GAC	1.470	0.991	0.479	1.545	0.986	0.559
Glu	GAG	1.329	0.924	0.405	1.465	0.917	0.548
Cys	UGC	1.407	0.929	0.478	1.469	0.980	0.488
Arg	CGC	1.832	0.835	0.997	2.120	0.830	1.289
Gly	GGC	1.819	1.001	0.818	2.108	0.978	1.130

Codon pairs in two *Marsupenaeus* species

The synonymous codon that encodes two amino acids was called a duplex codon or codon pairs, which was more commonly used than the single codon. Two *marsupenaeus* species had different use frequency of codon pairs (Table 2), such as GlyAla (GGAGCU vs GGAGCA), GlnArg (CAGAGA vs CAAAGA) and GluAsn (GAGAAC vs GAAAAU). In *M. japonicus*, the high-frequency codon pair of ArgArg was AGAAGA, while the optimal codon of Arg was CGC (Fig. S3). The high-frequency codon pair of AspAsp was GAUGAU, while the optimal codon of Asp was GAC. The high-frequency codon pair of GluGlu was GAAGAA, while the optimal codon of Glu was GAG. There were other inconsistencies, including GlyGly (GGAGGA) and Gly (GGC), HisHis (CAUCAU) and His (CAC), ProPro (CCACCA) and Pro (CCC), SerSer (AGCAGC) and Ser (UCG), ThrThr (ACAACA) and Thr (ACC/ACG), and ValVal (GUGGUG) and Val (GUC) (Fig. S3). In *M. pulchricaudatus*, the high-frequency codon pair of HisHis was CACCAC, while differentiates it from that of *M. japonicus*. The high-frequency codon pair of ProPro was CCACCA, while the optimal codon of Pro was CCC and CCG. The high-frequency codon pair of AlaAla was GCAGCA,

while the optimal codon of Ala was GCG (Fig. S4). Codon pair utilization biases play an important role in protein synthesis by the interaction with tRNA isoacceptors [56]. Codon pair analysis enables us to get a clear picture of the codon usage bias during transcription and translation.

Table 2
The different duplex codons of two *Marsupenaeus* species

Codons	<i>M. japonicus</i>	<i>M. pulchricaudatus</i>	Codons	<i>M. japonicus</i>	<i>M. pulchricaudatus</i>
Arg_Pro	AGGCCA	AGACCA	Phe_Thr	TTCACA	TTCACC
Asn_Ile	AACATT	AACATC	Pro_His	CCTCAT	CCTCAC
Asn_Leu	AACCTC	AACCTG	Pro_Lys	CCCAAG	CCAAAG
Asp_Pro	GATCCA	GACCCA	Pro_Ser	CCTTCA	CCATCA
Asp_Val	GATGTG	GATGTT	Pro_Val	CCAGTG	CCTGTG
Cys_Met	TGTATG	TGCATG	Ser_Gln	TCACAG	AGCCAG
Gln_Arg	CAGAGA	CAAAGA	Ser_Met	TCCATG	TCAATG
Gln_Ile	CAGATC	CAGATT	Thr_Leu	ACCCTC	ACTTTG
Glu_Asn	GAGAAC	GAAAAT	Trp_His	TGGCAC	TGGCAT
Gly_Al	GGAGCT	GGAGCA	Trp_Tyr	TGGTAC	TGGTAT
Gly_Val	GGAGTG	GGTGTT	Tyr_Al	TATGCT	TATGCA
His_His	CATCAT	CACCAC	Tyr_Pro	TACCCA	TATCCA
Leu_Leu	CTGCTG	CTCCTC	Val_Cys	GTGTGC	GTGTGT
Leu_Met	CTGATG	TTGATG	Val_Gly	GTGGGC	GTTGGA
Lys_Glu	AAAGAA	AAGGAA	Val_Ser	GTGTCA	GTCAGC
Phe_Al	TTTGCA	TTTGCT			

Multi-species clustering analysis

Based on the RSCU values of 59 codons (except Met, Trp, Taa, Tag, and Tga), the heat map (Fig. 5) showed that two *Marsupenaeus* species were clustered with *Daphnia pulex* and *Litopenaeus vannamei*, and then *Crassostrea gigas*. The *Larimichthys crocea*, *Cyprinus carpio*, and *Danio rerio* were classified into the same cluster. *Homo sapiens* and *Mus musculus* were clustered into one group. It was interesting that *Drosophila melanogaster* and mammals were grouped at first, then Arthropoda and Granulifusus kiranus (*Crassostrea gigas*) joined in them. This may be mainly since *D. melanogaster* has a stronger codon preference than other arthropods. Similar to the clustering results, the PCA showed two *Marsupenaeus* species overlapped almost completely, and the relationship between *C. gigas* and

arthropods was not as strong as the result of heat map clustering (Fig. 6). These clustering results were consistent with traditional species classification.

Discussion

Given the significant biological effects of different codon patterns, identifying these patterns in a given gene or genome is important to understand the molecular mechanisms of gene expression and to uncover the effects of long-term evolution on the genome [15, 57, 58]. Moreover, identifying these patterns is helpful to the phylogenetic analysis of species and to improve the expression of the target gene by optimizing codons [59–62].

In this study, we analyzed the codon preference of transcripts of two *Marsupenaeus* species, which was consistent on the whole. There was no significant difference in the content of AT and GC of the third base. The correlation between codon preference parameters was similar. The gene expression level (FPKM) was significantly negatively correlated with A/T3s, and the effective codon number ENc was significantly positively correlated with A/T3s. This result indicated that the third codon base significantly affects codon preference and gene expression level. The first and second base content of a codon is usually affected by natural selection, while the third-base content is affected by mutation pressure [54, 55]. The mean effective codon number (ENc) of the two cryptic species was 52.1 and 52.2, respectively, indicating the weak codon preference of *M. japonicus* transcripts. The PR2 analysis and ENc-GC3s correction analysis showed that the codon usage patterns of *M. japonicus* were influenced by both mutation pressure and natural selection.

It was consistent on the whole that the correlation analysis between Ka/Ks value and codon preference parameters of orthologous genes in two cryptic species. In the group with $\omega < 1$, the Ka/Ks of variety I was significantly positively correlated with ENc. In the group with $\omega > 1$, Ka and Ks were significantly positively correlated with T3s, and the Fop and GC content of variety I was significantly positively correlated with Ka and Ks , respectively. Those orthologous genes under positive selective stress are usually associated with abiotic or biotic stress resistance [63, 64]. Axis 1 was the most important evaluation index, which showed a highly significant correlation with C3 and GC3. The results of the correspondence analysis showed that the codon preference parameters of the two cryptic species had an opposite correlation with axis 1. The gene expression level was significantly positively correlated with GC content. The results from Camiolo et al showed that gene sequences with higher GC content showed a higher expression level and better codon preference [65]. Natural selection can improve gene transcription and translation efficiency by giving priority to the use of optimal synonymous codons [53, 66].

In this study, RSCU and ENc values were combined to determine 12 optimal codons in the variety I, among which 9 ended in C and 3 ended in G, and 14 optimal codons in variety II, among which 9 ended in C and 5 ended in G. these results showed *M. japonicus* is genetically more likely to end in C/G, which was similar to the codon usage characteristic of carp (*Cyprinus carpio* L.), zebrafish (*Danio rerio* L.), *Acanthopagrus schlegelii* and *Pagrus major* [67, 68]. This may be since the evolution of *M. japonicus* is

mainly mutated from AT to CG. Al-Saif et al showed that reducing the proportion of UU or UA could enhance the resistance to mRNA attenuation, thus increasing protein expression [69]. The codon preference of different species is generally influenced by mutation and natural selection pressure [70, 71]. There were 31 different double codon pairs between the two cryptic species, and the optimization of the codon pair could improve the efficiency of protein translation than the single optimal codon [72–74]. The genetic distance of species is closely related to the codon preference difference, which can be used for species classification [25]. The results of multi-species heat map analysis and clustering based on RSCU values are consistent with traditional species classification, indicating that the size of inter-species codon preference difference can reflect the proximity of species, which is also verified in other species [74–76].

In conclusion, we systematically compared the codon usage patterns of two *Marsupenaeus* species and evaluated the comprehensive effects of various factors. This study provided a relatively comprehensive understanding of the correlations among codon usage bias, gene expression, and evolutionary trends in *Marsupenaeus* orthologs. Moreover, the results provided new insights into the genetic divergence and the phylogenetic relationships of two *Marsupenaeus* species.

Abbreviations

NCBI: National Center for Biotechnology Information; GO:Gene Ontology; CDS:the coding sequences; SRA:Sequence Read Archive; ENc:the effective number of codons; CAI:the codon adaptation index; CBI:codon bias index; Fop:frequency of optional codon; RSCU:relative synonymous codon usage; COA:correspondence analysis; PR2:parity rule 2; FPKM:Fragments Per Kilobase per Million; PCA:principal component analysis; Ka:non-synonymous substitution ratio; Ks:synonymous substitution ratio.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Project of China Agriculture Research System (Grant No. CARS-48), the Science and Technology Plan Project of Ningbo (2019B10011), and the Open Research Fund of Jiangsu Key Laboratory of Marine Biotechnology (HS2020001). The authors are grateful to Dongshan Swire Marine Station (Xiamen University) for providing the laboratory space.

Authors' contributions

PPW and YM conceived and designed the experiments, YM, YQS, and JW obtained funds for the study. PPW, YQS, and JW performed bioinformatics analysis. PPW and YM drafted the manuscript. YM, YQS, and JW participated in the manuscript revision. All authors read and approved the final manuscript.

References

1. Duret L. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev.* 2002;12(6):640–9.
2. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research.* 2009;19(9):1655–64.
3. Biro JC. Studies on the origin and evolution of codon bias. *arXiv preprint arXiv:08073901* 2008.
4. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology evolution.* 1985;2(1):13–34.
5. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet.* 2011;12(1):32–42.
6. Komar AA. The Yin and Yang of codon usage. *Human molecular genetics.* 2016;25(R2):R77–85.
7. Yu C-H, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, Liu Y. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Molecular cell.* 2015;59(5):744–54.
8. Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. *Molecular biology evolution.* 2018;35(5):1092–103.
9. Eyre-Walker AC. An analysis of codon usage in mammals: selection or mutation bias? *Journal of molecular evolution.* 1991;33(5):442–9.
10. Powell JR, Dion K. Effects of codon usage on gene expression: empirical studies on *Drosophila*. *Journal of molecular evolution.* 2015;80(3–4):219–26.
11. Zalucki YM, Power PM, Jennings MP. Selection for efficient translation initiation biases codon usage at second amino acid position in secretory proteins. *Nucleic acids research.* 2007;35(17):5748–54.
12. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y: Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proceedings of the National Academy of Sciences* 2018, 115(21):E4940-E4949.

13. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from the international DNA sequence databases. *Nucleic acids research*. 1997;25(1):244–5.
14. Sharp PM, Li W-H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of molecular evolution*. 1986;24(1–2):28–38.
15. Yannai A, Katz S, Hershberg R. The codon usage of lowly expressed genes is subject to natural selection. *Genome Biol Evol*. 2018;10(5):1237–46.
16. Archetti M. Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. *J Mol Evol*. 2004;59(2):258–66.
17. Hershberg R, Petrov DA. Selection on codon bias. *Annual review of genetics*. 2008;42:287–99.
18. Mukhopadhyay P, Basak S, Ghosh TC. Differential selective constraints shaping codon usage pattern of housekeeping and tissue-specific homologous genes of rice and Arabidopsis. *DNA research*. 2008;15(6):347–56.
19. Trotta E. Selection on codon bias in yeast: a transcriptional hypothesis. *Nucleic acids research*. 2013;41(20):9382–95.
20. Machado HE, Lawrie DS, Petrov DA. Pervasive Strong Selection at the Level of Codon Usage Bias in *Drosophila melanogaster*. *Genetics*. 2020;214(2):511–28.
21. Chu D, Wei L. Characterizing the heat response of *Arabidopsis thaliana* from the perspective of codon usage bias and translational regulation. *Journal of plant physiology*. 2019;240:153012.
22. Guan D-L, Ma L-B, Khan MS, Zhang X-X, Xu S-Q, Xie J-Y. Analysis of codon usage patterns in *Hirudinaria manillensis* reveals a preference for GC-ending codons caused by dominant selection constraints. *BMC Genomics*. 2018;19(1):542.
23. Jia X, Liu S, Zheng H, Li B, Qi Q, Wei L, Zhao T, He J, Sun J. Non-uniqueness of factors constraint on the codon usage in *Bombyx mori*. *BMC Genomics*. 2015;16(1):356.
24. Huang X, Xu J, Chen L, Wang Y, Gu X, Peng X, Yang G. Analysis of transcriptome data reveals multifactor constraint on codon usage in *Taenia multiceps*. *BMC Genomics*. 2017;18(1):308.
25. Duan X, Yi S, Guo X, Wang W. A comprehensive analysis of codon usage patterns in blunt snout bream (*Megalobrama amblycephala*) based on RNA-Seq data. *Int J Mol Sci*. 2015;16(6):11996–2013.
26. Tsoi KH, Ma KY, Wu TH, Fennessy ST, Chu KH, Chan TY. Verification of the cryptic species *Penaeus pulchricaudatus* in the commercially important kuruma shrimp *P. japonicus* (Decapoda: Penaeidae) using molecular taxonomy. *Invertebrate Systematics*. 2014;28(5):476–90.
27. Wang P, Chen B, Zheng J, Cheng W, Zhang H, Wang J, Su Y, Xu P, Mao Y: Fine-Scale Population Genetic Structure and Parapatric Cryptic Species of Kuruma Shrimp (*Marsupenaeus japonicus*), Along the Northwestern Pacific Coast of China. *Frontiers in Genetics* 2020, 11(118).
28. Wang P, Xing C, Wang J, Su Y, Mao Y. Evolutionary adaptation analysis of immune defense and hypoxia tolerance in two closely related *Marsupenaeus* species based on comparative transcriptomics. *Fish Shellfish Immunol*. 2019;92:861–70.

29. Yi S, Li Y, Wang W. Selection shapes the patterns of codon usage in three closely related species of genus *Misgurnus*. *Genomics*. 2018;110(2):134–42.
30. Song H, Gao H, Liu J, Tian P, Nan Z. Comprehensive analysis of correlations among codon usage bias, gene expression, and substitution rate in *Arachis duranensis* and *Arachis ipaënsis* orthologs. *Scientific reports*. 2017;7(1):1–12.
31. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*. 2003;13(9):2178–89.
32. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*. 2010;28(5):511.
33. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform*. 2011;12(1):323.
34. Sueoka N, Kawanishi Y. DNA G + C content of the third codon position and codon usage biases of human genes. *Gene*. 2000;261(1):53–62.
35. Wright F. The 'effective number of codons' used in a gene. *Gene*. 1990;87(1):23–9.
36. Fuglsang A. The 'effective number of codons' revisited. *Biochem Biophys Res Commun*. 2004;317(3):957–64.
37. Comeron JM, Aguadé M. An evaluation of measures of synonymous codon usage bias. *Journal of molecular evolution*. 1998;47(3):268–74.
38. Jiang Y, Deng F, Wang H, Hu Z. An extensive analysis on the global codon usage pattern of baculoviruses. *Archives of virology*. 2008;153(12):2273.
39. Hartl DL, Moriyama EN, Sawyer SA. Selection intensity for codon bias. *Genetics*. 1994;138(1):227–34.
40. Lee S, Weon S, Lee S, Kang C. Relative codon adaptation index, a sensitive measure of codon usage bias. *Evolutionary Bioinformatics*. 2010;6:EBO. S4608.
41. Tsai IJ, Zarowiecki M, Holroyd N, Garcarrubio A, Sanchez-Flores A, Brooks KL, Tracey A, Bobes RJ, Fragoso G, Sciotto E. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature*. 2013;496(7443):57.
42. Zheng H, Zhang W, Zhang L, Zhang Z, Li J, Lu G, Zhu Y, Wang Y, Huang Y, Liu J. The genome of the hydatid tapeworm *Echinococcus granulosus*. *Nat Genet*. 2013;45(10):1168.
43. Götz S, Garcíagómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa AJNAR. High-throughput functional annotation and data mining with the Blast2GO suite. 2008, 36(10):3420–3435.
44. Shah P, Gilchrist MA. Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. *PLoS Genet*. 2010;6(9):e1001128.
45. Xu C, Cai X, Chen Q, Zhou H, Cai Y, Ben A. Factors affecting synonymous codon usage bias in chloroplast genome of *oncidium gower ramsey*. *Evolutionary Bioinformatics*. 2011;7:EBO. S8092.

46. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology evolution*. 2007;24(8):1586–91.
47. Nekrutenko A, Makova KD, Li W-H. The KA/KS ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome research*. 2002;12(1):198–202.
48. Hurst LD. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet*. 2002;18(9):486–6.
49. Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genom Proteom Bioinform*. 2006;4(4):259–63.
50. Sharp PM, Li W-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*. 1987;15(3):1281–95.
51. Bellgard M, Schibeci D, Trifonov E, Gojobori T. Early detection of G + C differences in bacterial species inferred from the comparative analysis of the two completely sequenced *Helicobacter pylori* strains. *Journal of molecular evolution*. 2001;53(4–5):465–8.
52. Liu Q. Analysis of codon usage pattern in the radioresistant bacterium *Deinococcus radiodurans*. *Biosystems*. 2006;85(2):99–106.
53. Duret L, Mouchiroud D: Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proceedings of the National Academy of Sciences* 1999, 96(8):4482–4487.
54. Hu C, Chen J, Ye L, Chen R, Zhang L, Xue X. Codon usage bias in human cytomegalovirus and its biological implication. *Gene*. 2014;545(1):5–14.
55. RoyChoudhury S, Mukherjee D. A detailed comparative analysis on the overall codon usage pattern in herpesviruses. *Virus Res*. 2010;148(1–2):31–43.
56. Irwin B, Heck JD, Hatfield GW. Codon pair utilization biases influence translational elongation step times. *J Biol Chem*. 1995;270(39):22801–6.
57. Quax TE, Claassens NJ, Söll D, van der Oost J. Codon bias as a means to fine-tune gene expression. *Molecular cell*. 2015;59(2):149–61.
58. Paul P, Malakar AK, Chakraborty S. Codon usage and amino acid usage influence genes expression level. *Genetica*. 2018;146(1):53–63.
59. Acosta-Rivero N, Sánchez JC, Morales J. Improvement of human interferon HUIFN α 2 and HCV core protein expression levels in *Escherichia coli* but not of HUIFN α 8 by using the tRNAAGA/AGG. *Biochem Biophys Res Commun*. 2002;296(5):1303–9.
60. Sharp PM, Matassi G. Codon usage and genome evolution. *Curr Opin Genet Dev*. 1994;4(6):851–60.
61. Miller JB, Whiting MF, Kauwe JS, Ridge PG. How Codon Usage Bias Affects Our Ability to Recover the Tree of Life. 2019(2019100086 (doi: 10.20944/preprints201910.0086.v1)).
62. Mauro VP, Chappell SA: **Considerations in the use of codon optimization for recombinant protein expression**. In: *Recombinant Protein Expression in Mammalian Cells*. Springer; 2018: 275–288.

63. De La Torre AR, Lin Y-C, Van de Peer Y, Ingvarsson PK. Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in picea gene families. *Genome Biol Evol.* 2015;7(4):1002–15.
64. Buschiazzi E, Ritland C, Bohlmann J, Ritland K. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC evolutionary biology.* 2012;12(1):8.
65. Camiolo S, Melito S, Porceddu A. New insights into the interplay between codon bias determinants in plants. *DNA Res.* 2015;22(6):461–70.
66. Akashi H. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics.* 1995;139(2):1067–76.
67. Cao G, Zhang Z, Zhang Z, Chen S, Zhu F, Jia C, Chen Z, Zeng H, Tang X. The analysis of the microsatellite sequences and codon bias of the coding sequence in *acanthopagrus schlegelii*, *pagrus major* and their hybrid progenies. *Oceanology Limnology.* 2019;50(5):1108–15.
68. Zhang X, Shun X. The synonymous codon bias of carp (*Cyprinus carpio* L.) and Zebrafish (*Danio rerio* L.). *Chinese Journal Fisheries.* 2010;23(4):23–9.
69. Al-Saif M, Khabar KS. UU/UA dinucleotide frequency reduction in coding regions results in increased mRNA stability and protein expression. *Mol Ther.* 2012;20(5):954–9.
70. Liu H, Huang Y, Du X, Chen Z, Zeng X, Chen Y, Zhang H. Patterns of synonymous codon usage bias in the model grass *Brachypodium distachyon*. *Genet Mol Res.* 2012;11(4):4695–706.
71. Ingvarsson PK. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. *Mol Biol Evol.* 2009;27(3):650–60.
72. Shao Z-Q, Zhang Y-M, Feng X-Y, Wang B, Chen J-Q. Synonymous codon ordering: a subtle but prevalent strategy of bacteria to improve translational efficiency. *PLoS One.* 2012;7(3):e33547.
73. Qian W, Yang J-R, Pearson NM, Maclean C, Zhang J. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet.* 2012;8(3):e1002603.
74. Feng C, Xu C-j, Wang Y, Liu W-l, Yin X-r, Li X, Chen M, Chen K. -s: Codon usage patterns in Chinese bayberry (*Myrica rubra*) based on RNA-Seq data. *BMC Genomics.* 2013;14(1):732.
75. Lu Y, Peng C, Chen Z, Chang X, Qiu J, Lin Z, Pan J. Analysis of Codon Usage Bias in *Clausena lansium* Transcriptome. *Molecular Plant Breeding.* 2018;16(18):5904–13.
76. Chen Z, Hu F, Wang X, Fan H, Zhang Z. Analysis of codon usage bias of *Ananas comosus* with genome sequencing data. *Journal of Fruit Science.* 2017;34(08):946–55.

Figures

Figure 1

Length distribution of unigenes (A), nucleotide composition (B), and PR2-plot analysis for *M. japonicus* (C) and *M. pulchricaudatus* (D).

Figure 2

Correlation analysis of codon usage parameters. Significant difference at $p < 0.05$; ** significant difference at $p < 0.01$.

Figure 3

The correlation analysis between Ka and Ks value and codon preference. *Significant difference at $p < 0.05$; ** significant difference at $p < 0.01$.

Figure 4

Correspondence analysis (A), correlation analysis (B) and GC content effect codon preference of *M. japonicus* (C) and *M. pulchricaudatus* (D).

Figure 5

The clustering analysis based on RSCU values. Cg = *Crassostrea gigas*, Dp = *Daphnia pulex*, Lv = *Litopenaeus vannamei*, Mj = *Marsupenaeus japonicus*, Mp = *Marsupenaeus pulchricaudatus*, Dm = *Drosophila melanogaster*, Hs = *Homo sapiens*, Mu = *Mus musculus*, Dr = *Danio rerio*, Cc = *Cyprinus carpio*, Lc = *Larimichthys crocea*.

Figure 6

The principal component analysis (PCA) of RSCU values.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS1S2.xlsx](#)
- [FigS4.tif](#)
- [FigS3.tif](#)
- [FigS2.tif](#)
- [FigS1.tif](#)