ISSN (Print) 0023-4001
ISSN (Online) 1738-0006

Korean J Parasitol Vol. 53, No. 6: 689-697, December 2015
http://dx.doi.org/10.3347/kjp.2015.53.6.689

▣ ORIGINAL ARTICLE

Codon Usage Bias and Determining Forces in *Taenia solium* Genome

Xing Yang[1,2,†], Xusheng Ma[2,†], Xuenong Luo[1], Houjun Ling[1], Xichen Zhang[1,*], Xuepeng Cai[1,2*]

[1]College of Veterinary Medicine, Jilin University, Changchun, 130000, P. R. China; [2]State Key Laboratory of Veterinary Etiological Biology, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou 730046, P. R. China

Abstract: The tapeworm *Taenia solium* is an important human zoonotic parasite that causes great economic loss and also endangers public health. At present, an effective vaccine that will prevent infection and chemotherapy without any side effect remains to be developed. In this study, codon usage patterns in the *T. solium* genome were examined through 8,484 protein-coding genes. Neutrality analysis showed that *T. solium* had a narrow GC distribution, and a significant correlation was observed between GC12 and GC3. Examination of an NC (ENC vs GC3s)-plot showed a few genes on or close to the expected curve, but the majority of points with low-ENC (the effective number of codons) values were detected below the expected curve, suggesting that mutational bias plays a major role in shaping codon usage. The Parity Rule 2 plot (PR2) analysis showed that GC and AT were not used proportionally. We also identified 26 optimal codons in the *T. solium* genome, all of which ended with either a G or C residue. These optimal codons in the *T. solium* genome are likely consistent with tRNAs that are highly expressed in the cell, suggesting that mutational and translational selection forces are probably driving factors of codon usage bias in the *T. solium* genome.

Key words: *Taenia solium*, codon usage bias, translation selection, mutation bias, intron number

INTRODUCTION

It is well known that 64 different codons (61 codons encoding for amino acids plus 3 stop codons) encode 20 standard amino acids. Many amino acids are coded by more than 1 codon, and the different codons that code the same amino acid are called 'synonymous codons'. Numerous studies have shown that the synonymous codons are used with unequal frequency, and some codons are used preferentially, a feature known as codon usage bias (CUB) [1,2]. Various hypotheses have been proposed to explain the origin of codon usage bias. Among these are the neutral theory [3] and the selection-mutation-drift balance model [4]. According to the neutral theory, mutations at degenerate coding positions should be selectively neutral, resulting in random synonymous codon choice and loss of natural selection power. In the selection-mutation-drift model, codon bias is thought to be a process mediated by a balance between mutation pressure, genetic drift, and weak selection. In other words, if a gene experiences high selective pressure, such as elevated expression, it may be inclined to result in stronger codon usage bias.

However, in recent years, with the completion of genome projects of many organisms, the 2 hypotheses are no longer sufficient to explain observed codon usage biases. Many important factors have been reported to influence this phenomenon, including gene length [5], GC-content [6], recombination rate [7], gene expression level [5], intron length [8], the hydrophobicity and the aromaticity of the encoded proteins [9], and so on.

Investigations of codon usage patterns have contributed to an understanding of the basic features of molecular organization of a genome, heterologous gene expression [10], and the prediction of gene expression levels [11], gene function [12], and gene position on chromosomes [13], and have also revealed information about the molecular evolution of individual genes. However, most of the studies on codon usage patterns focus on some model organisms and pathogenic agents, such as *Caenorhabditis, Drosophila, Arabidopsis* [5], yeast [14], *Giardia lamblia* [15], and *Borrelia burgdorferi* [16].

Taenia solium is one of the most important zoonotic parasites transmitted by consumption of the pork. *T. solium* can cause significant health problems and even death of their in-

• Received 16 May 2015, revised 10 August 2015, accepted 6 October 2015.
* Corresponding author (caixp@vip.163.com; zhangxichen2008@aliyun.com)
† Xing Yang and Xusheng Ma contributed equally to this work.

termediate host (= pigs), causing considerable losses to the world economy [17]. Until now, a systematic examination of the codon usage for *T. solium* genome has not been performed. Here, we carried out a whole genome analysis of codon usage bias of the *T. solium* genome by using the multivariate analysis technique [18]. This information will contribute to a better understanding of the parasite biology and provide a new resource to help with the development of urgently needed anti-parasite drugs and vaccines.

MATERIALS AND METHODS

Sequence data

The protein-coding sequences (CDS) from the *T. solium* genome were downloaded from http://www.genedb.org/Homepage. To minimize sampling error, we selected sequences that were longer than 300 bp and had the initiation and termination codons annotated. After removing redundant and incomplete DNA sequences, a total of 8,484 genes were selected for analysis.

Indices of codon usage

Codon usage in the selected genes was assessed using the program CodonW (version 1.4.2, http://codonw.sourceforge.net). Relative synonymous codon usage (RSCU) is the observed frequency of a codon divided by the expected frequency if all synonyms for that amino acid are used equally. Thus, RSCU values close to 1.0 indicate lack of bias, whereas the values more than 1 indicate that the codons are used more frequently than expected; conversely when the RSCU values are less than 1, the codon is used less frequently than expected. The codon adaptation index (CAI) is a simple and effective way to measure the extent of bias toward codons that were known to be preferred in highly expressed genes. A CAI value ranges from 0 to 1.0, and a higher value means a likely stronger codon usage bias and a potential higher expression level [19]. A codon usage reference table (data not shown) was constructed with a reference set of highly expressed genes for the calculation of CAI (of the 196 genes used in the reference set, 86 encode ribosomal proteins, 6 transcription elongation factor genes, 5 pyruvate kinase genes, 1 phosphoglycerate kinase gene, 3 glyceraldehyde-3-phosphate dehydrogenase genes, 4 enolase genes, 46 actin genes, and 45 tubulin genes). GC content of the entire gene, the first, second, and third codon positions (GCall, GC1, GC2, and GC3, respectively), and effective number of codons (ENC) were calculated. GC12 values (the average of GC1 and GC2) were calculated and used for neutrality plot analyses.

Correspondence analysis

Correspondence analysis (CA) has been widely used to explore the variation in synonymous codon usage among genes [20]. CA is a sophisticated multivariate statistical technique in which codon usage data (59 codons excluding Met, Trp, and stop codons) was plotted in a multidimensional space of 59 axes. The plot was then used to identify the axes that represent the most prominent factors contributing to variation among genes.

tRNA abundance and intron number

tRNA genes in the *T. solium* genome were searched using the tRNAscan-SE program with the eukaryote-specific parameters [21]. tRNAscan-SE was used to predict 161 tRNA genes and 22 pseudogenes in the genome sequence (data not shown). In the present study, the pseudogenes have been removed. We used tRNA gene copy numbers as an estimate of cellular tRNA abundance. The intron/exon number of the *T. solium* genes was obtained from the CDS annotation (ftp://bioinformatica.biomedicas.unam.mx/TsM1_13.12.11/).

Determination of optimal codons

We selected 5% of the total genes with extremely high and low CAI values which were regarded as the high and low expression gene datasets, respectively. Codon usage was compared using chi-squared contingency test of the 2 groups, and codons whose frequency of usage were significantly higher ($P < 0.01$) in highly expressed genes than in genes with low level of expression would be defined as the optimal codons [22].

Statistical analysis

Correlation analysis was carried out using the Spearman's rank correlation analysis method in SPSS version 19.0.

RESULTS

Nucleotide content of *T. solium* genes

The nucleotide content of *T. solium* coding sequences (expressed as % GC) is summarized in Fig. 1. This figure shows that there is a distinctly unimodal distribution of GC content

among the 8,484 *T. solium* genes, which is similar to *T. pisiformis* [23]. The GC contents of *T. solium* genes vary from 20.8% to 72.6% with a SD of 3.88. To understand the nucleotide distribution in 3 positions of codons, we investigated G+C content in all codons. The results showed that the G+C contents at 3 positions of codon were obviously different. GC3 was higher than the first and second positions, and GC1 was the lowest in all 3 positions of codon. The average GC content of all codons was 50.4%.

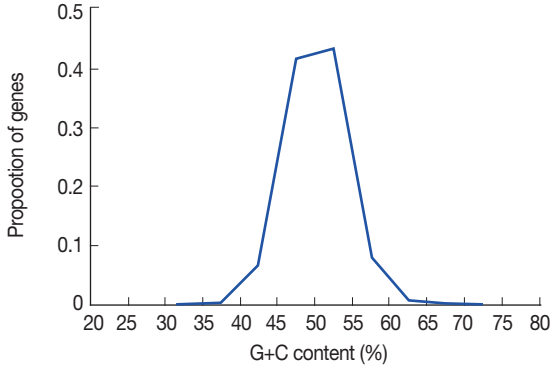

Fig. 1. The distribution of GC contents in *T. solium* genes. The GC content of the 8,484 *T. solium* genes (shown in blue) is unimodal.

Codon usage in *T. solium*

The overall codon usage of 8,484 genes from *T. solium* was calculated, with each codon, with the exception of stop codons, represented at least 29,081 times (Table 1). The genome of *T. solium* had a GC content of 43.7%. Although the genome appeared to be (at least slightly) A+T-rich, overall codon usage was biased toward C- and G-ending codons. Thirty-one of the 59 codons were found to be the preferred codons, and 54.8% (17/31) of the preferred codons were either G-ending or C-ending.

Table 1. Codon usage table in *T. solium*

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	73863	0.96	Ser	**UCU**	66945	1.09
	UUC	80528	1.04		**UCC**	74835	1.22
Leu	UUA	29081	0.45		UCA	62935	1.03
	UUG	71331	1.10		UCG	54145	0.88
	CUU	79963	1.23	Pro	CCU	64730	1.10
	CUC	91326	1.41		CCC	61128	1.04
	CUA	40023	0.62		CCA	68271	1.16
	CUG	77936	1.20		CCG	42041	0.71
Ile	**AUU**	81264	1.29	Thr	ACU	67286	1.12
	AUC	73474	1.16		ACC	69492	1.16
	AUA	34730	0.55		ACA	59130	0.99
Met	AUG	85450	1.00		ACG	43999	0.73
Val	**GUU**	70073	1.09	Ala	**GCU**	89508	1.19
	GUC	65525	1.02		**GCC**	84269	1.12
	GUA	34933	0.54		GCA	71732	0.96
	GUG	86210	1.34		GCG	54139	0.72
Tyr	UAU	40934	0.80	Cys	UGU	42340	0.98
	UAC	61575	1.20		**UGC**	44305	1.02
TER	UAA	2562	0.91	TER	UGA	3493	1.24
	UAG	2429	0.86	Trp	UGG	44634	1.00
His	CAU	47676	0.94	Arg	**CGU**	55022	1.30
	CAC	53846	1.06		**CGC**	50562	1.19
Gln	CAA	77299	0.96		**CGA**	51124	1.21
	CAG	83180	1.04		CGG	30983	0.73
Asn	**AAU**	83583	1.06	Ser	AGU	58100	0.95
	AAC	74133	0.94		AGC	50685	0.83
Lys	AAA	84273	0.91	Arg	AGA	34529	0.82
	AAG	99955	1.09		AGG	31888	0.75
Asp	**GAU**	108334	1.09	Gly	**GGU**	78723	1.35
	GAC	90233	0.91		**GGC**	63945	1.10
Glu	GAA	113616	0.91		GGA	58560	1.00
	GAG	137061	1.09		GGG	31858	0.55

Relative synonymous codon usage (RSCU) values were calculated by summing over all the genes. A total of 8,484 genes comprising 4,001,735 codons were analyzed. N is the number of codons, AA is the amino acid. Preferentially used codons are displayed in bold.

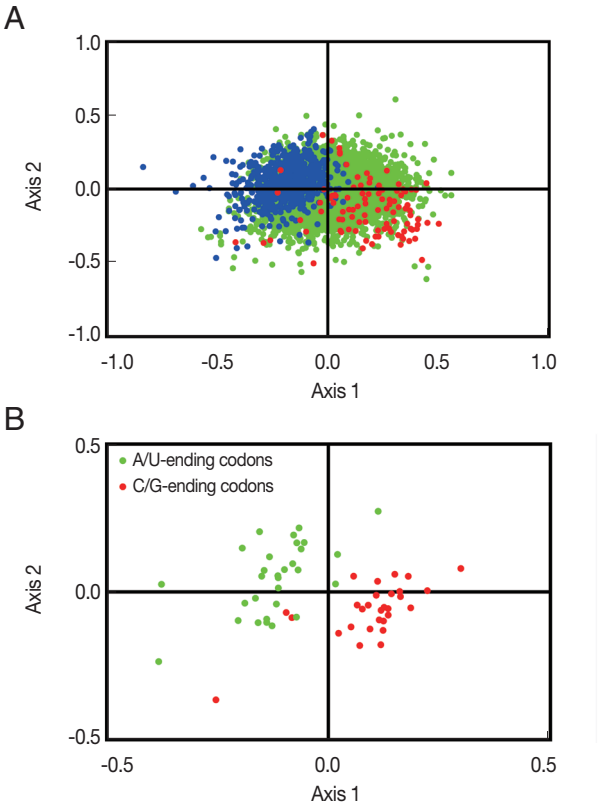

Fig. 2. Correspondence analysis of relative synonymous codon usage (RSCU) for all 8,484 *T. solium* genes. (A) This panel shows the distribution of genes on the primary and secondary axes (accounting for 16.7% and 14.1% of the total variation, respectively). The 2 classes of genes (High GC and Low GC) are color coded; the high GC genes are shown in red and the low GC genes are shown in blue. (B) This panel shows the underlying distribution of codons on the same 2 axes. Codons ending with G or C are shown in red, and codons ending with A or U are shown in green.

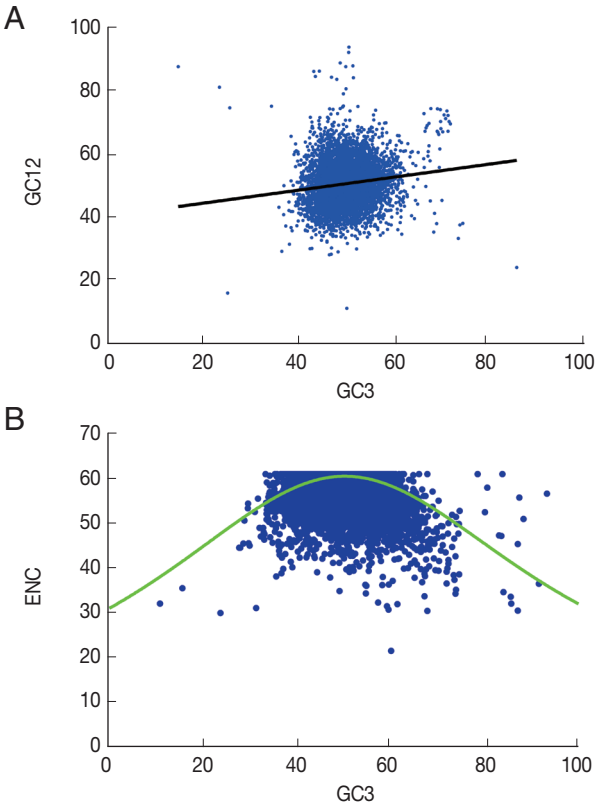

Fig. 3. GC12 or ENC vs GC3s plot for 8,484 genes of *T. solium*. (A) GC12 vs GC3 plot (Neutrality plot analyses). The regression line is y=0.5513x+0.2196, R2=0.7309, OP=0.4894. The range of the GC3 values was 12.0%-93.4%. The cross point of the regression line and the diagonal line is defined as the optimum point (OP). (B) ENC versus GC3s plot(NC plot), the solid red line indicates the expected ENC. The ENC values of different genes ranged 21.5 to 61.0; values of GC3s ranged 10.8 to 93.3.

Correspondence analysis

To investigate the synonymous codon usage variation among *T. solium* genes, correspondence analysis was employed to explore it in RSCU. The result revealed a single major trend in codon usage, namely, that the first axis accounted for 16.7% of the total variation, while the next 3 axes accounted for 14.1%, 8.5%, and 7.2%, respectively, confirming that the primary axis is the main factor explaining codon usage in these genes. The plot of the first and second axis of each gene is shown in Fig. 2A. The distance between genes on the plot is a reflection of their diversity in RSCU, with respect to the first 2 axes. To investigate the effect of GC content of genes on codon usage bias, different GC contents of genes are color-coded. The genes, GC content of which was more than or equal to 60%, plotted in red, and less than 45% plotted in blue. Green dots indicate genes in which the GC content is between 45% and 60%. In Fig. 2A, the high and low GC content of genes separated along the primary axis is shown.

The corresponding distribution of synonymous codons (see Fig. 2B) showed the separation of C/G-ending codons and A/U-ending codons along this same axis, indicating that the variations in synonymous codon usage among the *T. solium* genes were based on the nucleotide content of the genes. The separation of genes on the second axis appeared to be largely due to frequency differences in C-ending and G-ending codons among the GC rich genes (see right side of Fig. 2B).

Although the color-coding in Fig. 2A suggests a general relationship between the nucleotide content of genes and their position on the first axis of the correspondence analysis, it does not give us any statistical measure of this relationship. To

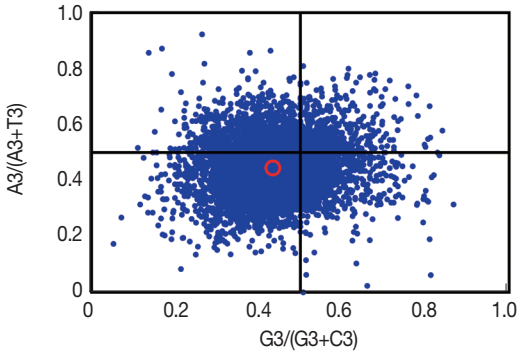

Fig. 4. Parity rule (PR2)-bias plot. The red open circle indicates the average position x = 0.4340 ± 0.0887 and y = 0.4458 ± 0.0927.

do this, we calculated the correlation between the GC content of individual *T. solium* genes and their location on the primary axis of the Correspondence Analysis. The results were highly significant ($r = 0.6563$, $P < 0.001$), indicating that the variations in codon usage are strongly correlated with the nucleotide content (i.e., GC content) of the genes.

Neutrality plot and NC plot analyses

Mutation bias and translation selection are considered to be the main factors that contribute to codon usage bias in different organisms [24,25]. To identify the main factors that shape codon usage bias in *T. solium*, neutrality plots (GC12 vs GC3) were used to analyze the influences of mutation bias and translation selection on codon usage [26]. When the correlation between GC12 and GC3 is statistically significant and the slope of the regression line is close to 1, mutation bias is assumed to be the main force shaping codon usage. Conversely, if selection is the dominant factor, then the slope of the regression line is close to 0. The results revealed significant correlations between GC12 and GC3 (Fig. 3A). The slope of the regression line in *T. solium* was 0.55. This significantly positive correlation in the neutrality plots indicated that mutation pressure and selection both contribute to the codon bias in *T. solium*.

A plot of ENC versus GC3s (NC plot) has been used to explore the codon usage variation among genes in different organisms [27]. It is argued that if the codon choice of a gene is constrained only by a G/C mutation bias, the gene would lie either on or just below the expected curve. As shown in Fig. 3B, it is clear that although a few of the genes lie on the expected curve, the majority with a low ENC fall below the expected curve. This suggests that not only mutation but also other factors, such as translational selection, are likely to be involved in determining the variation of codon usage.

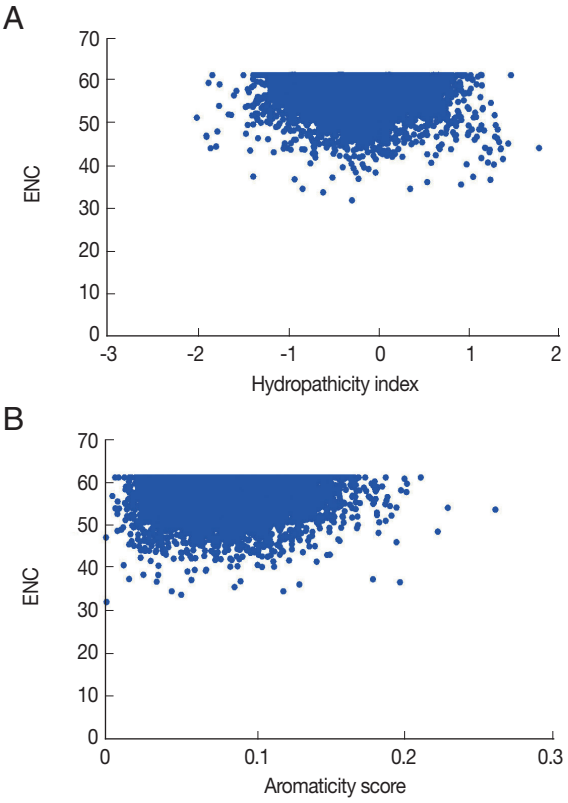

Fig. 5. Plot of ENC versus hydropathicity index and aromaticity score for *T. solium* genes.

PR2-bias plot

If only mutation pressure determined codon usage bias in a gene, G and C (A and T) should be used proportionally among the 4-fold degenerate codon families. Natural selection, however, would not necessarily cause proportional use of G and C (A and T) [28]. Here, we analyzed the associations between the purine (A and G) and the pyrimidine (C and T) content in the 4-fold degenerate codon families using the type 2 parity rule (PR2) bias plot (Fig. 4). Our results showed that C and T are used more frequently than G and A in these degenerate codons, suggesting that together with mutation pressure, other factors (such as selection) may influence the codon usage bias.

Relationship between CUB and the gene expression level

To explore the correlation between CUB and expression levels, we calculated the correlation coefficient between ENC and CAI to measure of the expression level of selected genes. The

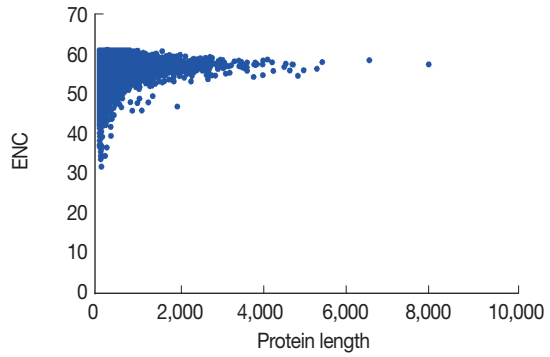

Fig. 6. Plot of ENC versus protein length for *T. solium*.

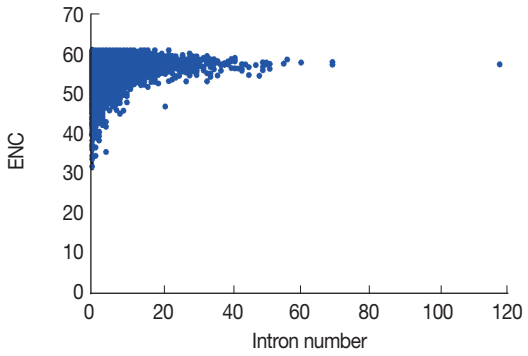

Fig. 7. Plot of ENC versus intron number for *T. solium*.

results show CUB is weakly and negatively correlated with the gene expression levels (r=-0.1804, P<0.001), suggesting that the genes with higher expression levels tend to have a higher codon usage bias (such as those encoding actin genes and ribosomal proteins tend to have lower ENC values) (Fig. 5).

Relationship between codon bias and hydropathicity index and aromaticity score

Numerous studies have shown that the hydropathicity and aromaticity of encoded protein play important roles in shaping codon usage of some species [29]. In order to investigate if the same thing happens to *T. solium*, we conducted a correlation analysis between codon usage bias and hydropathicity index (Gravy) and aromaticity score (Aromo) values. The correlation coefficients for the Gravy and Aromo scores (r=-0.0468, P<0.001; r=0.055, P<0.001, respectively) indicated that the hydropathicity and aromaticity of the encoded proteins were associated with the codon usage bias.

Relationship between CUB and protein length

In general, codon usage bias might be affected by the gene length. Here, we conducted a correlation analysis between codon bias and protein length. Our data demonstrated that ENC values are significantly and positively correlated with protein length, as expected (r=0.1138, P<0.001) (Fig. 6). The results indicated that protein length shaped codon usage in *T. solium* and the longer genes had a lower degree of codon bias.

Relationship between CUB and intron number

Evidence has been assembled to suggest that CUB has a close relationship with intron number [8,30]. However, the relationship between CUB and intron number is at present unclear. Here, we performed a correlation analysis to evaluate whether ENC values were related to intron number. From this analysis, results showed that ENC values are significantly and positively correlated with the intron number (r=0.1324, P<0.001) in *T. solium* (Fig. 7). The analysis results suggested that the intron number was associated with codon usage variation.

Optimal codons and tRNA abundance in *T. solium*

The average RSCU values of high/low expressed gene sample group are listed in Table 2. Twenty-six codons were determined to be the optimal codons, which were significantly more frequent among the highly expressed genes (P<0.01) according to the chi-square test. Almost all of optimal codons (except GGU and CGU) ended with G or C.

Previous studies suggested that the optimal codons tend to correspond to highly expressed tRNAs and tRNA gene copy numbers [31,32]. Here, we conducted an analysis to test whether this trend also existed in *T. solium* genome. We used tRNA gene copy numbers as a substitute for tRNA abundance in the cell. We found that there is good correspondence between tRNA abundance and optimal codons, 13 of the 26 optimal codons corresponded to the most abundant tRNAs (Table 2).

DISCUSSION

Base composition is an important feature of a genome and is the main force that affects codon usage. GC-rich organisms, such as bacteria, archea, fungi, *Triticum aestivum*, *Hordium vulgare*, and *Oryza sativa* [33,34], tend to use G or C at the third position. However, AT-rich organisms show a preference for A or T at third position, such as *Onchocerca volvulus*, *Mycoplasma capricolum*, and *Plasmodium falciparum* [35]. The genome of *T. solium* has a G+C content of 43.7%. Although the genome would thus appear to be slightly A+T rich, overall codon usage

Table 2. Optimal codons and tRNA abundance in *T. solium*

AA	Codon	tRNA gene	High	Low	AA	Codon	tRNA gene	High	Low
Phe	UUU	AAA (NA)	0.69 (2047)	1.21 (2707)	Ser	UCU	AGA (4)	0.88 (1710)	1.21 (2609)
	UUC*→	GAA (2)	1.31 (3863)	0.79 (1751)		UCC*	GGA (NA)	1.59 (3088)	0.87 (1863)
Leu	UUA	TAA (4)	0.21 (492)	0.94 (1719)		**UCA**	TGA (2)	0.72 (1400)	1.40 (3003)
	UUG	CAA (NA)	0.87 (2090)	1.32 (2422)		UCG*→	CGA (7)	1.00 (1937)	0.75 (1613)
	CUU	GAA (NA)	0.99 (2362)	1.32 (2405)	Pro	**CCU**	AGG (2)	0.97 (1976)	1.20 (2262)
	CUC*	GAG (NA)	2.01 (4816)	0.76 (1397)		**CCC***	GGG (NA)	1.39 (2820)	0.71 (1342)
	CUA	TAG (3)	0.45 (1068)	0.74 (1355)		**CCA**	TGG (2)	0.84 (1710)	1.49 (2806)
	CUG*	CAG (1)	1.47 (3525)	0.92 (1675)		CCG*→	CGG (5)	0.79 (1607)	0.60 (1142)
Ile	AUU	AAT (3)	1.04 (2444)	1.39 (2828)	Thr	**ACU**	AGT (1)	0.98 (2130)	1.17 (2287)
	AUC*	GAT (NA)	1.61 (3779)	0.79 (1613)		**ACC***	GGT (NA)	1.53 (3327)	0.84 (1644)
	AUA	TAT (4)	0.35 (813)	0.82 (1668)		ACA	TGT (3)	0.73 (1593)	1.31 (2576)
Met	AUG	CAT (11)	1.00 (3403)	1.00 (2759)		ACG	CGT (1)	0.76 (1661)	0.68 (1341)
Val	GUU	AAC (1)	0.75 (1794)	1.32 (2544)	Ala	**GCU**	AGC (NA)	1.05 (3038)	1.32 (2669)
	GUC*	GAC (NA)	1.20 (2893)	0.75 (1437)		**GCC***	GGC (NA)	1.44 (4176)	0.79 (1590)
	GUA	UAC (2)	0.38 (903)	0.81 (1557)		GCA	TGC (1)	0.70 (2031)	1.30 (2630)
	GUG*→	CAC (6)	1.68 (4040)	1.12 (2156)		GCG*→	CGC (1)	0.82 (2383)	0.59 (1193)
Tyr	UAU	ATA (NA)	0.52 (1043)	1.10 (1609)	Cys	UGU	ACA (NA)	0.82 (1326)	1.11 (1650)
	UAC*→	GTA (3)	1.48 (2947)	0.90 (1306)		UGC*	GCA (1)	1.18 (1899)	0.89 (1314)
His	CAU	ATG (NA)	0.69 (1259)	1.19 (1850)	Trp	UGG	CCA (11)	1.00 (1862)	1.00 (1138)
	CAC*→	GTG (2)	1.31 (2371)	0.81 (1260)	Arg	CGU*	ACG (4)	1.47 (2200)	0.88 (1097)
Gln	CAA	TTG (2)	0.73 (2009)	1.14 (3043)		CGC*	GCG (NA)	1.92 (2866)	0.60 (746)
	CAG*→	CTG (8)	1.27 (3511)	0.86 (2285)		CGA	TCG (8)	1.05 (1569)	1.05 (1306)
Asn	AAU	ATT (NA)	0.82 (2145)	1.25 (3689)		CGG	CCG (4)	0.76 (1139)	0.60 (742)
	AAC*→	GTT (1)	1.18 (3106)	0.75 (2193)	Ser	AGU	ACT (NA)	0.83 (1603)	1.04 (2233)
Lys	AAA	TTT (4)	0.67 (2071)	1.11 (4356)		AGC*→	GCT (5)	0.99 (1915)	0.74 (1581)
	AAG*	CTT (1)	1.33 (4138)	0.89 (3489)	Arg	AGA	TCT (10)	0.34 (501)	1.70 (2115)
Asp	GAU	ATC (NA)	0.86 (2935)	1.26 (4243)		AGG	CCT (10)	0.46 (680)	1.17 (1462)
	GAC*→	GTC (1)	1.14 (3888)	0.74 (2502)	Gly	GGU*	ACC (NA)	1.38 (3213)	1.19 (2098)
Glu	GAA	TTC (2)	0.62 (2561)	1.12 (5859)		GGC*	GCC (3)	1.47 (3423)	0.77 (1357)
	GAG*→	CTC (3)	1.38 (5755)	0.88 (4598)		GGA	TCC (6)	0.67 (1548)	1.39 (2450)
						GGG	CCC (3)	0.48 (1127)	0.65 (1140)

Codon usage was compared using chi-squared contingency test to identify optimal codons. That occur significantly more often ($P<0.01$) are indicated with asterisk denote codons that occurred significantly more often ($P<0.01$). The number of codons in the high expressed genes was 143953; the number of codons in the low expressed genes was 129698.

*denotes the codons that occurred significantly more often in the high expressed genes ($P<0.01$); they are the codons that were designated as 'optimal' codons. Thirteen (indicated by arrows) of the 26 optimal codons corresponded to the most abundant tRNAs in the *T. solium* genome.

was biased toward C- and G-ending codons (Table 2), this phenomenon is also found in *G. lamblia* [17].

Neutrality analysis found that there was a significant correlation between GC12 and GC3, which suggests that mutations may play a more important role in codon usage bias in *T. solium*. Meanwhile, ENC-plot analysis revealed that a majority of the points with low-ENC values lay below the expected curve, with only a few genes observed to lie on the expected curve, an indication that besides mutation bias, selection was also involved in determining the codon bias of some genes. However, because NC plot analysis cannot distinguish between mutation bias and selection within a species, Wright [27] suggested 2 ways to distinguish between selection and mutation bias. If mutation bias is the cause of codon bias, GC or AT should be used proportionally among the degenerate codon groups in a gene. In contrast, natural selection for codon choice would not necessarily cause proportional use of G and C (A and T). However, the PR2 plot showed that the CDS in the *T. solium* genome did not use GC and AT equally. The unequally used GC and AT in the degenerate codon positions in our current analysis further reflects the fact that selection pressure has played an important role in driving CUB of *T. solium*. From these findings, we can conclude that both mutation bias and selection have contributed to the codon bias in the *T. solium* genome.

Generally, it was thought that codon usage bias was affected by gene length. In this paper, protein length appeared to play a significant role in shaping codon bias in *T. solium*. Thus, we found that codon bias was negatively correlated with protein

length. Similar results have been found in many organisms, such as yeast [36], *Caenorhabditis elegans* [37], *Drosophila melanogaster* [6], and *Arabidopsis thaliana* [5]. An explanation as been proposed by Moriyama and Powell [38] for this phenomenon; namely, if shorter proteins could perform similar functions to those of longer ones, longer proteins become energy-expensive and disadvantageous, thus the selection constraint acts to reduce the size of highly expressed genes, dominantly determines the relationship between codon bias and gene length.

Until now, the role of introns in the codon bias usage of eukaryotic genes remains enigmatic. As mentioned above, recent studies have shown that intron length is closely related to codon usage, suggesting that introns may play a role in gene regulation [30]. However, the relationship between codon bias and intron number is at present unclear. To explore this relationship, the sequences of a set of genes containing between 0 and 77 introns was extracted from the published genome sequences of *T. solium*. Our results suggested that CUB was negatively related to intron number; in other words, genes with the higher codon bias were found to have fewer introns. The loss of intron is a major feature of eukaryotic evolution [39]. It has been shown that introns in highly expressed genes are substantially shorter than those in genes that are expressed at low levels [40], and that rapidly regulated genes are intron poor [41]. These reports suggested that introns might play a role in the negative regulation and expression of these genes. On the other hand, numerous studies have shown that codon bias is generally positively correlated with gene expression level whereas highly expressed genes (such as ribosomal proteins) usually exhibit higher levels of codon bias [42]. Based on these studies, we conclude that the negative relationship between codon bias and intron number may have a role in gene expression in *T. solium*.

In this study, we identified 26 codons as the optimal codons in the *T. solium* genome. Most importantly, optimal codons in the *T. solium* genome were found to end either with G or C. This is very similar to the pattern observed in other eukaryotic genomes, such as *D. melanogaster* [43], *C. elegans* [44], *G. lamblia* [17], and *Schizosaccharomyces pombe* [45]. The identification of optimal codons in this parasite will impact the design of degenerate primers, introduction of point mutations, and investigation of mechanism(s) of evolution of the species at the molecular level.

ACKNOWLEDGMENTS

This work was supported by the 863 program (no. 2006AA 10A207), the Gansu Natural Science Foundation (no. 1010RJZ A002), and the National Key Project of Scientific and Technical Supporting program (no. 2007BAD40B03), China.

CONFLICT OF INTEREST

We have no conflict of interest related to this work.

REFERENCES

1. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 1985; 2: 13-34.
2. Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 1981; 151: 389-409.
3. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from the international DNA sequence databases. Nucleic Acids Res 1997; 25: 244-245.
4. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 1986; 24: 28-38.
5. Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis, Drosophila,* and *Arabidopsis*. Proc Natl Acad Sci USA 1999; 96: 4482-4487.
6. Marais G, Mouchiroud D, Duret L. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci USA 2001; 98: 5688-5692.
7. Kliman RM, Hey J. Hill-Robertson interference in *Drosophila melanogaster*: reply to Marais, Mouchiroud and Duret. Genet Res 2003; 81: 89-90.
8. Vinogradov AE. Intron length and codon usage. J Mol Evol 2001; 52: 2-5.
9. Romero H, Zavala A, Musto H. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucleic Acids Res 2000; 28: 2084-2090.
10. Kane JF. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. Curr Opin Biotechnol 1995; 6: 494-500.
11. Naya H, Romero H, Carels N, Zavala A, Musto H. Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii*. FEBS Lett 2001; 501: 127-130.
12. Lin K, Kuang Y, Joseph JS, Kolatkar PR. Conserved codon composition of ribosomal protein coding genes in *Escherichia coli,*

Mycobacterium tuberculosis and *Saccharomyces cerevisiae*: lessons from supervised machine learning in functional genomics. Nucleic Acids Res 2002; 30: 2599-2607.
13. Salamov AA, Solovyev VV. Ab initio gene finding in *Drosophila genomic* DNA. Genome Res 2000; 10: 516-522.
14. Kliman RM, Irving N, Santiago M. Selection conflicts, gene expression, and codon usage trends in yeast. J Mol Evol 2003; 57: 98-109.
15. Lafay B, Sharp PM. Synonymous codon usage variation among *Giardia lamblia* genes and isolates. Mol Biol Evol 1999; 16: 1484-1495.
16. McInerney JO. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc Natl Acad Sci USA 1998; 95: 10698-10703.
17. Flisser A, Sarti E, Lightowlers M, Schantz P. Neurocysticercosis: regional status, epidemiology, impact and control measures in the Americas. Acta Trop 2003; 87: 43-51.
18. Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol 2001; 53: 290-298.
19. Sharp PM, Li WH. The codon Adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 1987; 15: 1281-1295.
20. Greenacre MJ. Theory and applications of correspondence analysis. Información General. Academic Press. 1984.
21. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997; 25: 955-964.
22. Liu Q. Analysis of codon usage pattern in the radioresistant bacterium *Deinococcus radiodurans*. Biosystems 2006; 85: 99-106.
23. Chen L, Liu T, Yang D, Nong X, Xie Y, Xie Y, Fu Y, Wu X, Huang X, Gu X, Wang S, Peng X, Yang G. Analysis of codon usage patterns in *Taenia pisiformis* through annotated transcriptome data. Biochem Biophys Res Commun 2013; 430: 1344-1348.
24. Zhong J, Li Y, Zhao S, Liu S, Zhang Z. Mutation pressure shapes codon usage in the GC-rich genome of foot-and-mouth disease virus. Virus Genes 2007; 35: 767-776.
25. Sau K, Sau S, Mandal SC, Ghosh TC. Factors influencing the synonymous codon and amino acid usage bias in AT-rich *Pseudomonas aeruginosa* phage PhiKZ. Acta Biochim Biophys Sin (Shanghai) 2005; 37: 625-633.
26. Sueoka N. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 1988; 85: 2653-2657.
27. Wright F. The 'effective number of codons' used in a gene. Gene 1990; 87: 23-29.
28. Sueoka N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 1995; 40: 318-325.
29. Chen L, Yang DY, Liu TF, Nong X, Huang X, Xie Y, Fu Y, Zheng WP, Zhang RH, Wu XH, Gu XB, Wang SX, Peng XR, Yang GY. Synonymous codon usage patterns in different parasitic platyhelminth mitochondrial genomes. Genet Mol Res 2013; 12: 587-596.
30. Rao Y, Wu G, Wang Z, Chai X, Nie Q, Zhang X. Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. DNA Res 2011; 18: 499-512.
31. Goetz RM, Fuglsang A. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. Biochem Biophys Res Commun 2005; 327: 4-7.
32. Rocha EP. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res 2004; 14: 2279-2286.
33. Hershberg R, Petrov DA. General rules for optimal codon choice. PLoS Genet 2009; 5: e1000556.
34. Kawabe A, Miyashita NT. Patterns of codon usage bias in three dicot and four monocot plant species. Genes Genet Syst 2003; 78: 343-352.
35. Waterkeyn JG, Gauci C, Cowman AF, Lightowlers MW. Codon usage in *Taenia* species. Exp Parasitol 1998; 88: 76-78.
36. Ingvarsson PK. Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*. Mol Biol Evol 2007; 24: 836-844.
37. Comeron JM, Kreitman M, Aguadé M. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics 1999; 151: 239-249.
38. Moriyama EN, Powell JR. Codon usage bias and tRNA abundance in *Drosophila*. J Mol Evol 1997; 45: 514-523.
39. Coulombe-Huntington J, Majewski J. Intron loss and gain in *Drosophila*. Mol Biol Evol 2007; 24: 2842-2850.
40. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. Selection for short introns in highly expressed genes. Nat Genet 2002; 31: 415-418.
41. Jeffares DC, Penkett CJ, Bähler J. Rapidly regulated genes are intron poor. Trends Genet 2008; 24: 375-378.
42. Andersson S, Kurland C. Codon preferences in free-living microorganisms. Microbiol Rev 1990; 54: 198-210.
43. Shields DC, Sharp PM, Higgins DG, Wright F. " Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 1988; 5: 704-716.
44. Stenico M, Lloyd AT, Sharp PM. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res 1994; 22: 2437-2446.
45. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res 1988; 16: 8207-8211.