ISSN 1672-9145 **Acta Biochimica et Biophysica Sinica** 2005, **37**(1): 1–10 **CN** 31-1940/Q

Analysis of Synonymous Codon Usage Bias in *Chlamydia*

Hui LÜ, Wei-Ming ZHAO*, Yan ZHENG, Hong WANG, Mei QI, and Xiu-Ping YU

Department of Medical Microbiology, School of Medicine, Shandong University, Jinan 250012, China

Abstract Chlamydiae are obligate intracellular bacterial pathogens that cause ocular and sexually transmitted diseases, and are associated with cardiovascular diseases. The analysis of codon usage may improve our understanding of the evolution and pathogenesis of *Chlamydia* and allow reengineering of target genes to improve their expression for gene therapy. Here, we analyzed the codon usage of *C. muridarum*, *C. trachomatis* (here indicating biovar trachoma and LGV), *C. pneumoniae*, and *C. psittaci* using the codon usage database and the CUSP (Create a codon usage table) program of EMBOSS (The European Molecular Biology Open Software Suite). The results show that the four genomes have similar codon usage patterns, with a strong bias towards the codons with A and T at the third codon position. Compared with *Homo sapiens*, the four chlamydial species show discordant seven or eight preferred codons. The ENC (effective number of codons used in a gene)-plot reveals that the genetic heterogeneity in *Chlamydia* is constrained by the G+C content, while translational selection and gene length exert relatively weaker influences. Moreover, mutational pressure appears to be the major determinant of the codon usage variation among the chlamydial genes. In addition, we compared the codon preferences of *C. trachomatis* with those of *E. coli*, yeast, adenovirus and *Homo sapiens*. There are 23 codons showing distinct usage differences between *C. trachomatis* and *E. coli*, 24 between *C. trachomatis* and adenovirus, 21 between *C. trachomatis* and *Homo sapiens*, but only six codons between *C. trachomatis* and yeast. Therefore, the yeast system may be more suitable for the expression of chlamydial genes. Finally, we compared the codon preferences of *C. trachomatis* with those of six eukaryotes, eight prokaryotes and 23 viruses. There is a strong positive correlation between the differences in coding GC content and the variations in codon bias (r=0.905, P<0.001). We conclude that the variation of codon bias between *C. trachomatis* and other organisms is much less influenced by phylogenetic lineage and primarily determined by the extent of disparities in GC content.

Key words *Chlamydia*; codon usage bias; GC content; gene expression

Genetic codes are sets of three nucleotides (codons) in an mRNA molecule that are translated into amino acids in the course of protein synthesis. There are a total of 64 codons, with 61 of them coding 20 different amino acids and other three serving as stop codons. The genetic code is degenerate, meaning that each amino acid may be coded by two or more codons (synonymous codons). Synonymous codons are not used at equal frequencies both within and between organisms [1–3]; the patterns of codon usage vary considerably among organisms, and also among genes from the same genome [4]. Analysis of codon usage patterns can provide a basis for understanding the relevant mechanism for biased usage of synonymous codons and for selecting appropriate host expression systems to improve the expression of target genes *in vivo* and *in vitro*.

Chlamydiae are Gram-negative obligate intracellular bacterial pathogens, which are classified into four species: *Chlamydia trachomatis*, *C. psittaci*, *C. pneumoniae* and *C. pecorum*. *C. trachomatis* and *C. pneumoniae* are mainly human pathogens, while *C. psittaci* and *C. pecorum* are usually pathogens of birds and mammals. *C. trachomatis* (Ct) can be divided into three biological variants: biovar

Received: September 29, 2004 Accepted: December 10, 2004

This work was supported by the grants from the National Natural Science Foundation of China (No. 30271193), International Cooperation and Exchange Fund from the NSFC (No. 30170045) and the Natural Science Foundation of Shandong Province (No. Y2002004)

*Corresponding author: Tel, 86-531-8382579; E-mail, zhaowm@sdu.edu.cn

trachoma, biovar lymphogranuloma venereum (LGV) and biovar mouse pneumonitis (*C. muridarum*, MoPn). *C. trachomatis* is the causative agent of trachoma, which leads to preventable blindness worldwide, and it also causes several sexually transmitted diseases, such as urethritis, cervicitis and salpingitis [5,6]. *C. pneumoniae*, which causes pneumonia, sinusitis and bronchitis, has attracted great scientific attention because it has also been found to be associated with atherosclerosis, acute myocardial infarction and chronic neurological diseases [7].

Romero *et al*. [8] have reported that four factors (strand-specific mutational biases, replicational-transcriptional selection, the hydropathy of each encoded protein and the degree of amino acid conservation) are involved in the codon usage bias in *C. trachomatis*. However, it is not clear whether genetic and environmental factors affect codon usage in *Chlamydia*.

In this study, we have analyzed the codon usage data of MoPn, Ct (here indicating biovar trachoma and LGV), *C. pneumoniae* and *C. psittaci* and examined how other factors may affect codon usage variation in *Chlamydia*. We have also compared the codon preferences of Ct with those of *Escherichia coli*, *Saccharomyces cerevisiae*, adenovirus and *Homo sapiens*. Knowledge of the codon usage pattern in *Chlamydia* and a comparison of codon preference between *Chlamydia* and other species may therefore assist in the development of nucleic acid vaccines and improve the understanding of factors shaping codon usage patterns.

Materials and Methods

Overall codon usage patterns in *Chlamydia*

Complete genomic sequences of MoPn, Ct, *C. pneumoniae* and *C. psittaci* were obtained from GenBank (Bethesda, Maryland, USA; http://www.ncbi.nlm.nih.gov/). Codon usage data were analyzed using the codon usage database (Chiba, Japan; http://www.kazusa.or.jp/codon/) and the CUSP program of EMBOSS (The European Molecular Biology Open Software Suite, Cambridge, UK; http://bioinfo.pbi.nrc.ca:8090/EMBOSS/).

Codon usage variation among *Chlamydia* genes

The protein-coding sequences (*.ffn files) of complete genomes were downloaded from the GenBank FTP site (ftp://ftp.ncbi.nih.gov/). Because there is a negative correlation between codon usage bias and gene length—that is, codon usage is restricted in short coding sequences—genes having sequences of less than 100 codons were excluded from the analysis.

GC3 indicates the G+C content at the third position of synonymously degenerate codons. The effective number of codons of a gene (ENC) was used to quantify how far the codon usage of a gene departs from equal usage of synonymous codons [9]. ENC can take values from 20 (when only one codon is used per amino acid) to 61 (when all synonyms are used in equal frequencies). ENC appears to be a good measure of the extent of codon preference in a gene. Therefore, the GC3, ENC and gene length of each gene were calculated using the program INteractive Codon Analysis 1.0 (Division of Biology, Faculty of Science, Zagreb, Croatia; http://www.bioinfo-hr.org/inca). The codon usage pattern across genes was examined by the ENC-plot, which is a plot of ENC versus GC3.

Comparison of codon preferences between Ct and other organisms

We compared the codon preferences among Ct, six eukaryotes, eight prokaryotes and 23 viruses. The genomic sequences were retrieved from GenBank (ftp://ftp.ncbi.nih.gov/), including those of *Staphylococcus aureus*, *Candida albicans*, *Lactobacillus acidophilus*, *Haemophilus influenzae*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Vibrio cholerae*, *Salmonella typhi*, *Escherichia coli*, *Homo sapiens*, *Eremothecium gossypii*, *Neurospora crassa*, *Bifidobacterium adolescentis*, *Mycobacterium tuberculosis*, eliothis armigera entomopoxvirus, human rotavirus, human respiratory syncytial virus, vaccinia virus, human papillomavirus, human coronavirus, SARS coronavirus, human metapneumovirus, mice minute virus, human immunodeficiency virus, human endogenous retrovirus, human echovirus, human enterovirus, hepatitis B virus, human adenovirus, human T-cell lymphotropic virus, avian sarcoma virus, groundnut rosette virus, hepatitis C virus, Abelson murine leukemia virus, frog virus, avian retrovirus and human herpesvirus. The codon usage analysis of these species was carried out using the codon usage database and the CUSP program of EMBOSS.

Statistical methods

The correlation between codon usage variation among *Chlamydia* genes and two parameters (G+C content at synonymous sites and gene length) was analyzed using the linear regression analysis model from microsoft excel with significance-of-difference levels of $P<0.05$ or $P<0.01$. A similar method was used to analyze the relationship between interspecific codon usage variation and coding GC content.

Results

Overall codon usage patterns and codon usage data in MoPn, Ct, *C. pneumoniae* and *C. psittaci* are listed in **Table 1**. *C. pecorum* was not analyzed because sufficient data were not available. In the four genomes that were analyzed, the amino acids Arg, Leu, Gly and Val have different codon usage biases because they have six-fold and four-fold coding degeneracy, while the preferred codons of amino acids that have two-fold or three-fold coding degeneracy are uniform.

The amino acids Arg, Leu and Ser have six-fold coding degeneracy. For MoPn and *C. pneumoniae*, Arg uses CGT most frequently, while AGA is most commonly used for Ct and *C. psittaci*. Although the most and the least commonly used codons of Arg are different, all four genomes prefer to use the codons with A and T ending, not with G and C ending. Codons with A and T ending are used 1.37 times more often than codons with G and C ending for MoPn, 1.7 times more often for Ct, 1.16 times more often for *C. pneumoniae* and 2 times more often for *C. psittaci*. AGA and CGA are the two codons with A ending among the six codons coding for Arg. The four chlamydial species prefer to use AGA rather than CGA. Leu is encoded by six codons with two having A at the third position: TTA and CTA. TTA is used 1.55 ± 0.65 times more often than CTA. Leu uses TTA the most and CTG/CTC the least. TTA is used 3.21 times more often than codon with the lowest frequency for MoPn, 2.68 times for Ct, 2.52 times for *C. pneumoniae* and 7.31 times for *C. psittaci*. As a result, Leu uses codons with A and T ending 0.88 times more often than codons with G and C ending for MoPn, 0.76 times more often for Ct, 0.79 times more often for *C. pneumoniae*, and 1.06 times more often for *C. psittaci*. For Ser, TCT is most commonly used, and TCG is used with the lowest frequency. Codons with A and T ending are used 1.07 times more than codons with G and C ending for MoPn, 0.93 times more often for Ct, 1.07 times more often for *C. pneumoniae*, and 0.98 times more often for *C. psittaci*. In a word, Arg, Leu and Ser prefer to use the codons with A and T ending and the usage bias of *C. psittaci* is more evident than others.

The amino acids Ala, Gly, Pro, Thr and Val have four-fold coding degeneracy (XYA, XYC, XYG and XYT). For Ala and Pro, XYT is used most frequently while XYG is used the least. For Gly, *C. psittaci* uses XYT the most and XYG the least, while the other species show a different bias; that is, they use XYA the most and XYC the least. For Thr, the usage of XYA is approximately the same as that of XYT, XYA/XYT is the most often used synonymous codon and XYG is the least commonly used synonymous codon. Val uses XYT most often and XYC the least, except for *C. pneumoniae*, for which XYG is used least often by Val. The usage of codons with A and T ending is 2.46 ± 1.05 times higher than codons with G and C ending for Ala, 0.90 ± 0.19 times higher for Gly, 2. 99 ± 1.49 times higher for Pro, 1.30 ± 0.34 times higher for Thr and 1.26 ± 0.68 times higher for Val.

The amino acids Asn, Asp, Cys, His, Phe and Tyr have two-fold codon degeneracy (XYC and XYT). They prefer to use XYT (10.1 to 36.2 per 1000 codons) rather than XYC (4.54 to 20.8 per 1000 codons). XYT is used 0.91 ± 0.32 times more often than XYC for Asn, 1.91 ± 0.41 times more often for Asp, 0.59 ± 0.29 times more often for Cys, 1.35 ± 0.37 times more often for His, 0.77 ± 0.27 times more often for Phe and 0.88 ± 0.55 times more often for Tyr.

For the amino acids Gln, Glu and Lys, whose two-fold degeneracy is of the form XYA or XYG, XYG (8.12 to 23.6 per 1000 codons) is used less often than XYA (26.2 to 51.5 per 1000 codons). XYA is used 1.44 ± 0.74 times more than XYG for Gln, 1.07 ± 0.35 times more for Glu and 1.69 ± 0.59 times more for Lys.

Ile is the only amino acid that has three-fold codon degeneracy (XYA, XYC and XYT). XYT is used most frequently (0.78 ± 0.31 times more than XYC and 1.66 ± 0.52 times more than XYA). Compared with all the other amino acids where codons with G or C at the third position are used the least, Ile uses XYA with the least frequency. However, for Ile, codons with A and T ending (XYA and XYT) are used 1.46 ± 0.35 times more often than codons with G and C ending (XYC).

As a whole, all chlamydial species or biovars analyzed show significant preference for one postulate codon for each amino acid. They show a high bias of codon usage toward the codons with T and/or A ending rather than C and/or G ending for all degenerate codons (1.16–5.96 times). At the same time, there are some differences in codon usage patterns among various chlamydial species or biovars. *C. psittaci* shows the greatest bias towards optimal codons. For example, the codon used for Arg with the highest frequency (18 per 1000 codons) was used 15 times more than the codon with the lowest frequency (1.12 per 1000 codons) in *C. psittaci*, but only 2.7 times more in MoPn, 6.5 times more in Ct and 3.3 times more in *C. pneumoniae*.

Generally, codons used more than twice as frequently as host consensus codons are regarded as preferred codons of heterologous genes. Compared with *Homo*

Table 1 Codon usage data in MoPn, Ct, *C. pneumoniae* and *C. psittaci*

Amino acid	MoPn			Ct			*C. pneumoniae*			*C. psittaci*			Human
	Codon	Fract	1/1000	Codon	Fract	1/1000	Codon	Fract	1/1000	Codon	Fract	1/1000	1/1000
Arg	CGC	0.15	6.91	CGC	0.17	7.79	CGC	0.15	6.83	CGC	0.15	6.29	10.68
	AGG	0.07	3.36	AGG	0.04	2.07	AGG	0.10	4.49	AGG	0.07	2.84	11.71
	AGA	0.24	11.30	**AGA**	0.30	13.80	AGA	0.24	10.90	**AGA**	0.44	18.00	11.72
	CGG	0.08	4.03	CGG	0.06	2.73	CGG	0.07	3.08	CGG	0.03	1.12	11.65
	CGA	0.21	9.78	CGA	0.19	8.65	CGA	0.15	6.72	CGA	0.08	3.25	6.24
	CGT	0.26	12.20	CGT	0.24	11.20	**CGT**	0.30	13.50	CGT	0.23	9.34	4.63
Leu	TTG	0.18	20.50	TTG	0.18	18.10	TTG	0.14	15.90	TTG	0.18	16.30	12.75
	TTA	0.30	34.00	**TTA**	0.29	29.30	**TTA**	0.27	31.00	**TTA**	0.40	35.40	7.43
	CTG	0.07	8.06	CTG	0.11	10.70	CTG	0.08	8.81	CTG	0.05	4.26	40.13
	CTA	0.12	13.40	CTA	0.13	13.00	CTA	0.14	15.80	CTA	0.12	10.30	7.04
	CTT	0.23	25.60	CTT	0.22	22.50	CTT	0.23	25.90	CTT	0.16	14.10	13.01
	CTC	0.09	10.20	CTC	0.08	7.96	CTC	0.14	15.90	CTC	0.09	8.42	19.67
Ser	**TCT**	0.42	34.80	**TCT**	0.43	34.30	**TCT**	0.38	30.60	**TCT**	0.32	24.90	14.89
	AGT	0.13	10.30	AGT	0.11	8.93	AGT	0.15	11.80	AGT	0.11	8.93	12.05
	AGC	0.10	8.21	AGC	0.13	10.10	AGC	0.11	8.76	AGC	0.14	10.60	19.45
	TCG	0.07	6.09	TCG	0.07	5.89	TCG	0.08	6.33	TCG	0.09	6.70	4.47
	TCA	0.12	10.20	TCA	0.12	9.72	TCA	0.14	11.60	TCA	0.23	18.10	12.00
	TCC	0.15	12.40	TCC	0.14	11.40	TCC	0.14	11.00	TCC	0.11	8.93	17.63
Ala	**GCT**	0.48	34.50	**GCT**	0.47	39.70	**GCT**	0.44	30.40	**GCT**	0.53	47.90	18.57
	GCG	0.12	8.27	GCG	0.11	9.66	GCG	0.12	8.21	GCG	0.04	3.96	7.55
	GCA	0.26	18.90	GCA	0.31	25.80	GCA	0.28	19.60	GCA	0.31	27.80	16.00
	GCC	0.14	10.00	GCC	0.11	9.45	GCC	0.16	11.20	GCC	0.12	11.40	28.28
Gly	GGG	0.25	15.70	GGG	0.19	12.20	GGG	0.20	12.30	GGG	0.13	8.73	16.48
	GGT	0.18	11.00	GGT	0.20	13.10	GGT	0.22	13.90	**GGT**	0.36	23.40	10.80
	GGC	0.13	7.91	GGC	0.15	9.58	GGC	0.15	9.47	GGC	0.19	12.60	22.56
	GGA	0.45	28.10	**GGA**	0.47	30.40	**GGA**	0.43	26.50	GGA	0.32	21.20	16.42
Pro	CCC	0.17	7.17	CCC	0.10	4.28	CCC	0.21	9.22	CCC	0.09	3.76	20.03
	CCT	0.52	22.20	**CCT**	0.55	22.70	**CCT**	0.52	23.30	**CCT**	0.50	20.30	17.42
	CCG	0.08	3.41	CCG	0.08	3.27	CCG	0.07	3.18	CCG	0.05	2.13	7.04
	CCA	0.23	10.00	CCA	0.26	10.70	CCA	0.20	8.74	CCA	0.36	14.80	16.84
Thr	ACC	0.17	8.83	ACC	0.14	8.48	ACC	0.19	10.30	ACC	0.16	11.20	19.09
	ACG	0.15	7.38	ACG	0.13	7.49	ACG	0.14	7.61	ACG	0.13	8.83	6.15
	ACA	0.33	16.80	**ACA**	0.38	22.50	**ACA**	0.33	17.30	**ACA**	0.36	24.20	14.91
	ACT	0.35	17.70	**ACT**	0.36	21.20	**ACT**	0.33	17.50	**ACT**	0.35	23.60	13.01
Val	**GTT**	0.42	27.10	**GTT**	0.41	29.20	**GTT**	0.33	20.40	**GTT**	0.41	24.40	10.98
	GTG	0.18	11.80	GTG	0.18	12.70	GTG	0.19	11.60	GTG	0.14	8.63	28.56
	GTA	0.25	16.00	GTA	0.30	21.20	GTA	0.27	16.40	GTA	0.35	21.00	7.06
	GTC	0.14	9.16	GTC	0.12	8.84	GTC	0.21	12.80	GTC	0.10	5.79	14.63
Asn	**AAT**	0.70	25.80	**AAT**	0.65	25.60	**AAT**	0.67	25.40	**AAT**	0.59	30.10	16.72
	AAC	0.30	11.00	AAC	0.35	13.60	AAC	0.33	12.80	AAC	0.41	20.80	19.17
Asp	**GAT**	0.78	35.20	**GAT**	0.74	36.20	**GAT**	0.74	33.10	**GAT**	0.71	34.30	21.98
	GAC	0.22	10.10	GAC	0.26	12.50	GAC	0.26	11.80	GAC	0.29	13.80	25.50
Cys	**TGT**	0.63	10.20	**TGT**	0.63	12.10	**TGT**	0.64	10.30	**TGT**	0.54	13.40	10.31
	TGC	0.37	5.89	TGC	0.37	7.18	TGC	0.36	5.72	TGC	0.46	11.60	12.55
His	CAC	0.28	6.23	CAC	0.27	4.54	CAC	0.32	7.56	CAC	0.34	5.18	15.03
	CAT	0.72	16.10	**CAT**	0.73	12.40	**CAT**	0.68	16.20	**CAT**	0.66	10.10	10.69
Phe	**TTT**	0.67	32.70	**TTT**	0.63	27.30	**TTT**	0.66	31.20	**TTT**	0.59	25.20	17.16
	TTC	0.33	15.90	TTC	0.37	16.30	TTC	0.34	16.30	TTC	0.41	17.50	20.39
Tyr	**TAT**	0.70	21.40	**TAT**	0.62	17.00	**TAT**	0.70	22.70	**TAT**	0.55	14.70	12.09
	TAC	0.30	9.14	TAC	0.38	10.40	TAC	0.30	9.80	TAC	0.45	12.10	15.41
Gln	CAG	0.33	13.80	CAG	0.30	11.70	CAG	0.35	13.90	CAG	0.22	8.12	34.39
	CAA	0.67	27.80	**CAA**	0.70	27.30	**CAA**	0.65	26.20	**CAA**	0.78	28.50	12.03
Glu	GAG	0.34	22.40	GAG	0.34	20.60	GAG	0.36	23.60	GAG	0.28	14.10	39.98
	GAA	0.66	42.80	**GAA**	0.66	40.60	**GAA**	0.64	42.30	**GAA**	0.72	36.40	28.92
Lys	AAG	0.29	17.20	AAG	0.24	15.20	AAG	0.34	21.00	AAG	0.24	16.20	32.22
	AAA	0.71	43.00	**AAA**	0.76	48.00	**AAA**	0.66	40.70	**AAA**	0.76	51.50	24.04
Ile	ATA	0.18	12.30	ATA	0.17	10.20	ATA	0.20	13.70	ATA	0.24	15.30	7.28
	ATT	0.57	38.10	**ATT**	0.51	31.20	**ATT**	0.50	34.70	**ATT**	0.47	30.20	15.79
	ATC	0.25	17.00	ATC	0.32	19.20	ATC	0.30	20.60	ATC	0.30	19.30	21.07

Fract refers to the proportion of all synonymous codons encoding the same amino acid. The frequency of each codon that appears in the coding sequence of the individual gene is 1/1000. Shaded codons are the preferred codons in *Chlamydia*. Triplets in bold face indicate a high frequency in coding the amino acid. Rimmed codons appear during low-frequency coding of the amino acid. Ct denotes biovar trachoma and LGV.

sapiens, seven preferred codons were found in MoPn, Ct or *C. pneumoniae* and eight preferred codons in *C. psittaci*. Four codons—CGT coding for Arg, TTA coding for Leu, GTA coding for Val and CAA coding for Gln—are mutually preferred in four genomes. Except *C. psittaci*, the other three species prefer to use TCT encoding Ser. Three genomes, except *C. pneumoniae*, use GTT encoding Val at least twice as often as *Homo sapiens*. In addition, ATT coding for Ile is the preferred codon in MoPn and *C. pneumoniae*, while GCT encoding Ala is preferred in Ct and *C. psittaci*. CTA coding for Leu is preferred only in *C. pneumoniae*, while ATA coding for Ile, GGT for Gly and AAA for Lys are exclusively preferred in *C. psittaci*. In conclusion, there are distinct differences in codon usage patterns between *Chlamydia* and *Homo sapiens*. Four chlamydial genomes do not have the same preferred codons as *Homo sapiens*. Therefore, although four genomes show similar codon usage patterns, the preference of specific codons among the chlamydial species or biovars must be specified when human consensus codons are used to modify targeted chlamydial genes for optimal expression.

Codon usage variation in chlamydial genes

The codon usage pattern also varies between genes in the same chlamydial genome. Plotting ENC values against GC3 is one effective way to explore this heterogeneity [9]. The ENC value of each chlamydial gene is plotted against its corresponding GC3 in **Fig. 1(A)**. Due to limited data, *C. psittaci* was not included in the analysis. The curve shows the expected position of genes whose codon usage

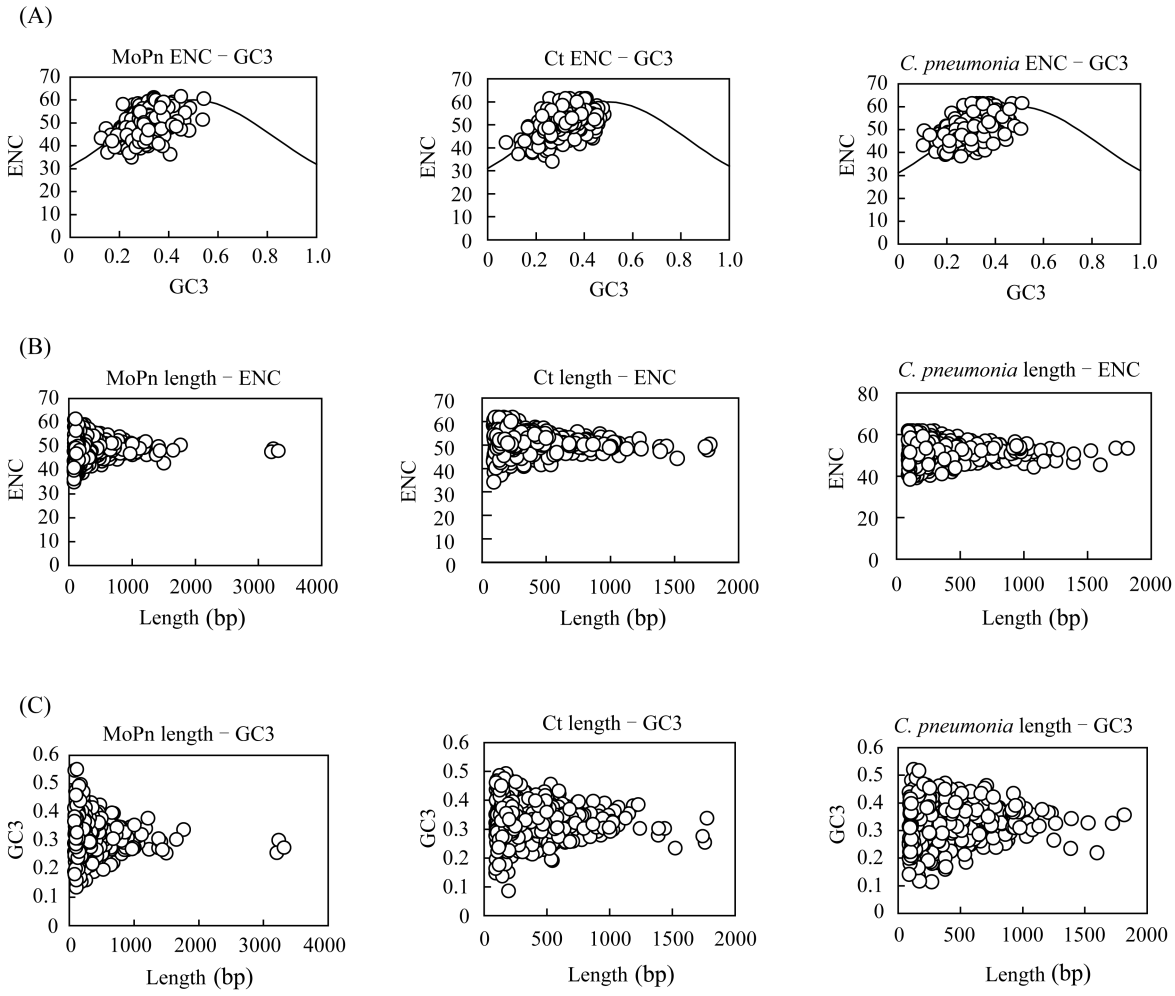

Fig. 1 Relationship between ENC, GC3 and gene length

(A) Plots of ENC versus GC3 for all chlamydial genes. ENC denotes the effective number of codons of each gene, and GC3 denotes the G+C content at the third synonymous codon position of each gene. The solid curve shows the expected position of genes whose codon usage is only determined by the variation in GC3. (B) Plots of ENC versus gene length. (C) Plots of GC3 versus gene length (bp).

is only determined by variation in GC3 content. If a particular gene is subject to G+C compositional constraints, it will lie on or just below the expected curve. If a gene is subject to selection for translationally optimal codons, it will lie considerably below the expected curve. Among most chlamydial genes, the GC3 values vary from 0.08 to 0.44, while the ENC values vary from 40 to 56. If the genes have low codon usage bias, the translational selection factor does not appear to be important for gene expression. Genes with lower GC3 values also have lower ENC values, indicating a stronger codon bias. A large number of points lie near the solid curve on the left side of this distribution, suggesting that these genes are subject to GC compositional constraints. Statistically, the relationship between ENC and GC3 is significantly positive (P<0.001), suggesting that mutational bias may be the major determinant of codon usage variation among chlamydial genes. The genetic heterogeneity correlates positively with the A+T content at the third position, which is consistent with the preference for codons with T and/or A ending as discussed above.

The relationship between gene length and synonymous codon usage bias has been reported for *Drosophila melanogaster*, *Escherichia coli*, *Saccharomyces cerevisiae*, *Pseudomonas aeruginosa* and *Yersinia pestis* [10–12]. Here, the plot of gene length against ENC or against GC3 [**Fig. 1 (B,C)**] appears to assume the shape of a normal distribution. Shorter genes have a much wider variance in ENC values, vice versa for longer genes. We have analyzed the relationship between ENC value and gene length, and the relationship between GC3 and gene length in chlamydial genes. None of the correlations were statistically significant. Evidently, gene length affects codon usage of *Chlamydia* only in a minor way. Similar results were also found in *S. pneumoniaee*, *P. aeruginosa* and SARS coronavirus [12–14].

Comparison of codon preferences between Ct and *E. coli*, yeast, adenovirus and *Homo sapiens*

As mentioned above, MoPn, Ct, *C. pneumoniae* and *C. psittaci* adopt similar codon usage patterns. Thus, the codon preferences of Ct, as a representative of Chlamydiae, were compared with those of *E. coli*, yeast, adenovirus and *Homo sapiens* to see which will be the suitable host for the optimal expression of Chlamydia genes.

From **Table 2**, there are 23 codons showing a Ct-to-*E. coli* ratio higher than 2 or lower than 0.50, 24 codons showing a Ct-to-adenovirus ratio higher than 2 or lower than 0.50 and 21 codons showing a Ct-to-human ratio higher than 2 or lower than 0.50, but only 6 codons showing a Ct-to-yeast ratio higher than 2 or lower than 0.50, suggesting that codon usage of Ct genes more closely resembles that of yeast genes than that of *E. coli*, adenovirus and human genes. Thus, to express chlamydial genes efficiently in *E. coli* or human cell systems, codon optimization of the chlamydial genes may be required. At the same time, we can speculate that the Chlamydia genes may be more efficiently expressed in the yeast system.

Comparison of codon preferences among different species

On the basis of the above observations, we compared the codon usage of several other eukaryotes and prokaryotes with that of *Chlamydia* (**Table 3**).

To examine whether different species comply with the same codon usage rule, we compared Ct not only with six eukaryotes and eight prokaryotes, but also with 23 viruses, taking into account that both *Chlamydia* and viruses belong to obligate intracellular microorganisms whose codon usage may be restricted by their hosts (**Table 3**).

From **Table 3**, it is clear that Ct presents similar codon preferences to *Vibrio cholerae*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* because less than 10 codons show Ct-to-species ratios either higher than 2 or lower than 0.5 irrespective of their phylogenetic lineages. However, there are 23 codons with a Ct-to-*E. coli* ratio either higher than 2 or lower than 0.5, 40 codons with a Ct-to-*Mycobacterium tuberculosis* ratio either higher than 2 or lower than 0.5, 47 codons with a Ct-to-*Bifidobacterium adolescentis* ratio either higher than 2 or lower than 0.5, 21 codons with a Ct-to-human ratio either higher than 2 or lower than 0.5, 28 codons with a Ct-to-*Eremothecium goss* ratio either higher than 2 or lower than 0.5 and 31 codons with a Ct-to-*Neurospora crassa* ratio either higher than 2 or lower than 0.5, indicating that the codon preferences are significantly different between them. If codon usage is a major determinant of gene expression, Ct genes may be expressed more efficiently in species such as *Vibrio cholerae*, *Saccharomyces cerevisia* and *Schizosaccharomyces pombe.*

From **Table 3**, it can also be seen that the codon preference of Ct is most similar to the vaccinia virus, followed by the human coronavirus and human immunodeficiency virus, but is least similar to the human adenovirus and human herpesvirus. Thus, it can be speculated that Ct genes can probably express well in the vaccinia virus system. This hypothesis remains to be tested.

To investigate whether the changes in coding GC content among various species are associated with the observed variations in codon bias, the absolute values of the

Table 2 Comparison of codon preferences between Ct and *E. coli*, yeast, adenovirus (ad), and *Homo sapiens* (human)

Amino acid	Codon	1/1000					Ct/			
		Ct	*E. coli*	yeast	ad	human	*E. coli*	yeast	ad	human
Ala	GCT	39.66	15.37	21.13	17.95	18.57	2.58	1.88	2.21	2.136
	GCG	9.66	33.57	6.17	12.85	7.55	0.29	1.57	0.75	1.279
	GCA	25.79	20.31	16.20	11.13	16.00	1.27	1.59	2.32	1.612
	GCC	9.45	25.49	12.57	31.68	28.28	0.37	0.75	0.30	0.334
Arg	CGC	7.79	21.96	2.59	20.99	10.68	0.35	3.01	0.37	0.729
	AGG	2.07	1.24	9.25	6.96	11.71	1.67	0.22	0.30	0.177
	AGA	13.82	2.08	21.28	9.76	11.72	6.64	0.65	1.42	1.179
	CGG	2.73	5.40	1.74	5.49	11.65	0.51	1.57	0.50	0.234
	CGA	8.65	3.56	3.01	3.19	6.24	2.43	2.87	2.71	1.386
	CGT	11.22	20.96	6.48	5.35	4.63	0.54	1.73	2.10	2.423
Asn	AAT	25.56	17.70	35.90	22.71	16.72	1.44	0.71	1.13	1.529
	AAC	13.64	21.71	24.89	40.02	19.17	0.63	0.55	0.34	0.712
Asp	GAT	36.22	32.24	37.74	21.28	21.98	1.12	0.96	1.70	1.648
	GAC	12.48	19.04	20.28	32.07	25.50	0.66	0.62	0.39	0.489
Cys	TGT	12.10	5.19	7.98	4.95	10.31	2.33	1.52	2.44	1.174
	TGC	7.18	6.44	4.72	7.70	12.55	1.11	1.52	0.93	0.572
Gln	CAG	11.66	28.81	12.17	22.81	34.39	0.40	0.96	0.51	0.339
	CAA	27.32	15.44	27.41	17.26	12.03	1.77	1.00	1.58	2.271
Glu	GAG	20.57	17.74	19.19	25.01	39.98	1.16	1.07	0.82	0.515
	GAA	40.63	39.50	45.84	27.66	28.92	1.03	0.89	1.47	1.405
Gly	GGG	12.23	11.01	6.01	9.86	16.48	1.11	2.03	1.24	0.742
	GGT	13.09	24.88	23.88	14.42	10.80	0.53	0.55	0.91	1.212
	GGC	9.58	29.41	9.76	24.13	22.56	0.33	0.98	0.40	0.425
	GGA	30.44	7.94	10.91	17.46	16.42	3.83	2.79	1.74	1.854
His	CAC	4.54	9.70	7.75	14.27	15.03	0.47	0.59	0.32	0.302
	CAT	12.43	12.89	13.72	4.86	10.69	0.96	0.91	2.56	1.163
Ile	ATA	10.23	4.33	17.81	10.64	7.28	2.36	0.57	0.96	1.405
	ATT	31.21	30.35	30.14	19.96	15.79	1.03	1.04	1.56	1.977
	ATC	19.22	24.98	17.05	11.03	21.07	0.77	1.13	1.74	0.912
Leu	TTG	18.11	13.73	27.06	14.96	12.75	1.32	0.67	1.21	1.420
	TTA	29.30	13.91	26.24	2.26	7.43	2.11	1.12	13.00	3.943
	CTG	10.73	52.65	10.45	23.39	40.13	0.20	1.03	0.46	0.267
	CTA	13.01	3.86	13.35	18.64	7.04	3.37	0.97	0.70	1.848
	CTT	22.49	11.04	12.22	21.53	13.01	2.04	1.84	1.04	1.729
	CTC	7.96	11.02	5.45	12.65	19.67	0.72	1.46	0.63	0.405
Lys	AAG	15.19	10.20	30.82	21.14	32.22	1.49	0.49	0.72	0.471
	AAA	48.02	33.64	42.11	20.79	24.04	1.43	1.14	2.31	1.998
Met	ATG	18.31	27.75	20.94	26.39	22.18	0.66	0.87	0.69	0.826
Phe	TTT	27.29	22.38	26.10	26.04	17.16	1.22	1.05	1.05	1.590
	TTC	16.29	16.59	18.30	14.57	20.39	0.98	0.89	1.12	0.799
Pro	CCC	4.28	5.54	6.78	23.84	20.03	0.77	0.63	0.18	0.214
	CCT	22.72	7.05	13.54	18.98	17.42	3.22	1.68	1.20	1.304
	CCG	3.27	23.21	5.26	10.94	7.04	0.14	0.62	0.30	0.464
	CCA	10.73	8.53	18.19	14.66	16.84	1.26	0.59	0.73	0.637
Ser	TCT	34.33	8.51	23.46	14.08	14.89	4.03	1.46	2.44	2.306
	AGT	8.93	8.76	14.22	8.83	12.05	1.02	0.63	1.01	0.741
	AGC	10.12	16.07	9.69	14.96	19.45	0.63	1.04	0.68	0.520
	TCG	5.89	8.94	8.55	5.89	4.47	0.66	0.69	1.00	1.318
	TCA	9.72	7.16	18.70	12.11	12.00	1.36	0.52	0.80	0.810
	TCC	11.37	8.61	14.21	15.69	17.63	1.32	0.80	0.72	0.645
Thr	ACC	8.48	23.39	12.60	29.72	19.09	0.36	0.67	0.29	0.444
	ACG	7.49	14.38	7.94	6.82	6.15	0.52	0.94	1.10	1.218
	ACA	22.48	7.07	17.77	17.51	14.91	3.18	1.27	1.28	1.508
	ACT	21.21	8.95	20.29	19.91	13.01	2.37	1.05	1.07	1.630
Trp	TGG	8.65	15.31	10.34	11.82	13.05	0.56	0.84	0.73	0.663
Tyr	TAT	17.01	16.33	18.77	11.38	12.09	1.04	0.91	1.49	1.407
	TAC	10.38	12.27	14.72	26.92	15.41	0.85	0.71	0.39	0.674
Val	GTT	29.17	18.39	21.99	14.08	10.98	1.59	1.33	2.07	2.657
	GTG	12.67	26.24	10.72	24.47	28.56	0.48	1.18	0.52	0.444
	GTA	21.21	10.89	11.80	10.25	7.06	1.95	1.80	2.07	3.004
	GTC	8.84	15.22	11.64	9.07	14.63	0.58	0.76	0.97	0.604

1/1000 represents the frequency of each codon that appears in the whole coding gene. Ct/*E. coli*, Ct/yeast, Ct/ad and Ct/human indicate the ratio of codon usage frequency in Ct to that in *E. coli*, *Saccharomyces cerevisiae*, adenovirus and *Homo sapiens* respectively. A ratio higher than 2 or lower than 0.5 indicates that the codon preference differs greatly, and vice versa.

Table 3 Comparison of codon preference among different species

Species	GCcod (%)	GCd (%)	Number
Ct (prokaryote)	41.35		
Staphylococcus aureus (prokaryote)	33.10	8.25	17
Candida albicans (eukaryote)	36.93	4.42	16
Lactobacillus acidophilus (prokaryote)	37.55	3.80	15
Haemophilus influenzae (prokaryote)	38.76	2.59	11
Saccharomyces cerevisiae (eukaryote)	39.73	1.62	6
Schizosaccharomyces pombe (eukaryote)	39.80	1.55	3
Vibrio cholerae (prokaryote)	44.33	2.98	6
Salmonella typhi (prokaryote)	48.16	6.81	14
Escherichia coli (prokaryote)	51.80	10.45	23
Homo sapiens (eukaryote)	52.54	11.19	21
Eremothecium gossypii (eukaryote)	52.70	11.35	28
Neurospora crassa (eukaryote)	56.13	14.78	31
Bifidobacterium adolescentis (prokaryote)	61.87	20.52	47
Mycobacterium tuberculosis (prokaryote)	65.77	24.42	40
Eliothis armigera entomopoxvirus (dsDNA)	25.81	15.54	26
Human rotavirus (dsRNA)	32.74	8.61	21
Human respiratory syncytial virus (ssRNA)	33.44	7.91	17
Vaccinia virus (dsDNA)	34.20	7.15	8
Human papillomavirus (dsDNA)	37.89	3.46	19
Human coronavirus (ssRNA)	38.42	2.93	11
SARS coronavirus (ssRNA)	40.98	0.37	11
Human metapneumovirus (ssRNA)	41.01	0.34	19
Mice minute virus (ssDNA)	43.01	1.66	17
Human immunodeficiency virus (ssRNA)	43.13	1.78	16
Human endogenous retrovirus (ssRNA)	46.69	5.34	13
Human echovirus (ssRNA)	47.62	6.27	15
Human enterovirus (ssRNA)	47.94	6.59	14
Hepatitis B virus (dsDNA)	49.65	8.30	18
Human adenovirus (dsDNA)	51.02	9.67	24
Human T-cell lymphotropic virus (ssRNA)	54.80	13.45	32
Avian sarcoma virus (ssRNA)	56.78	15.43	33
Groundnut rosette virus (ssRNA)	57.04	15.69	24
Hepatitis C virus (ssRNA)	57.88	16.53	32
Abelson murine leukemia virus (ssRNA)	58.46	17.11	36
Frog virus (dsDNA)	58.74	17.39	38
Avian retrovirus (ssRNA)	63.19	21.84	37
Human herpesvirus	66.08	24.73	40

GCcod indicates the G+C content of protein genes. GCd indicates the absolute value of the difference in coding GC content between Ct and other species. Number indicates the number of codons with great diversities in codon bias (ratios higher than 2 or lower than 0.5) between Ct and other organisms.

difference in coding GC content between Ct and other species were calculated. A linear regression analysis was then performed to examine the relationship between the differences in coding GC content and the corresponding numbers of codons with obviously different usage frequency. As a result, a strong positive correlation between differences in coding GC content and variations in codon bias (r=0.905, P<0.001, **Table 3**) was observed, indicating that the codon usage variation between Ct and other species is strongly constrained by coding GC content. In summary, the variation of codon usage bias between Ct and other organisms appears to be determined by the extent of disparities in coding GC content, and is less influenced by phylogenetic lineage.

Discussion

Previous analyses of codon usage have suggested that both a huge interspecific variation and a clear intragenomic variability exist. Codon usage bias is found to be related to different biological factors, such as tRNA abundance, strand-specific mutational bias, gene expression level, gene length, amino acid composition, protein structure, mRNA structure and GC composition [15–18]. However, directional mutation pressure on DNA sequences and natural selection affecting gene translation are the two major factors that have been widely accepted to account for both interspecific codon usage variation and intragenomic codon usage variability. With regard to the codon usage variation among genes within the same organism, this phenomenon has been observed in a wide range of species. In some unicellular organisms, such as *E. coli* and *Saccharomyces cerevisiae*, highly expressed genes have a strong selective preference for the codons complementary to the most abundant tRNA species, whereas lowly expressed genes display more uniform codon usage patterns largely compatible with the mutational bias in the absence of translational selection [19,20]. In mammals and birds, the diverse patterns of codon usage may arise from compositional constraints of the genomes [21–23].

Romero *et al*. have conducted the correspondence analysis for *C. trachomatis* genes and found that the most important source of variations among the genes comes from whether the sequence is located on the leading or lagging strand of replication, resulting in an over-representation of G or C, respectively [8]. In the present study, we used the ENC-plot to analyze the factors affecting codon usage variation among genes and extended the analysis to other chlamydial species. Our analysis has reinforced the above-mentioned findings. Here, genetic heterogeneity in the *Chlamydia* species is observed to be constrained by GC content, while the gene length has only a minor impact on the codon choice. In various species of *Chlamydia*, genetic heterogeneity seems to be the result of similar factors. In *C. trachomatis*, the major trend in codon choices is not as strong as in other species [8], so it can be postulated that there may be several major factors which shape chlamydial gene codon usage. In addition to strand-specific mutational biases, GC3 may be another important factor. Certainly, illustrating all the factors which shape chlamydial codon usage variation is a complex issue. However, mutational pressure, not the selective forces acting at the translational level, may play an important role in determining the codon usage variation among chlamydial genes. All these findings that hold true for *C. trachomatis* as well as for other tested chlamydial species support the "mutational bias–translational selection" hypothesis.

Our studies indicate that the codon usage variation between Ct and other species (including eukaryotes, prokaryotes and viruses) is much less influenced by phylogenetic lineage and primarily determined by the extent of disparities in coding GC content. It can be inferred that codon usage variability among different species may not depend on their phylogenetic relationships, but on their coding GC content. Consequently, the coding GC content may be more useful in predicting the amino acid or nucleotide sequence rather than the phylogenetic reconstruction. Our conclusion is consistent with previous studies that were mostly focused on genome GC content and limited to the three domains of life (Bacteria, Archaea, and Eukarya). For example, Chen *et al*. have proposed that only two parameters, genome GC content and context-dependent nucleotide bias, effectively differentiate the genome-wide codon bias of 100 eubacterial and archaeal organisms [24]. Moreover, they found that genome GC content variation is the most important parameter differentiating codon bias between different organisms [24]. It has been reported that seven GC-rich microbial genomes belonging to different domains of life adopt similar codon usage patterns regardless of their phylogenetic lineages [25]. In the present study, we have examined the relationship between the coding GC content and cross-species disparities in codon usage. Since the whole genome contains coding regions and non-coding regions, coding GC content may presumably be a more accurate reflection of the codon usage bias than genome GC content. Most importantly, our analysis revealed that coding GC content is correlated with cross-species disparities in codon usage not only in all three domains of life but in non-cellular microbes.

In the present study, a comprehensive analysis of codon usage and genome base composition in chlamydial species has revealed that: (1) MoPn, Ct, *C. pneumoniae* and *C. psittaci* adopt similar codon usage patterns, although *C. psittaci* shows the greatest bias towards optimal codons; (2) the chlamydial species prefer to use the codons with A and T at the third codon position; and (3) the gene codon usage pattern is significantly different between *Chlamydia* (MoPn, Ct, *C. pneumoniae* and *C psittaci*) and human genomes. Compared with human genomes, the four chlamydial genomes do not have the same preferred codons. Furthermore, the biased trend towards A and T coincides with high A+T content at silent sites in *Chlamydia* (the mean value is 70%). Since *Chlamydia* are AT-rich

organisms, it is reasonable that A and/or T ending codons are predominant in their genomes. These findings suggest that it is still necessary to differentiate various chlamydial species—even biovars—when designing specific strategies for optimizing chlamydial codons even though there are significant similarities in the codon usage pattern among all tested *Chlamydia* species.

An assumption made in the present study that chlamydial genes may express more efficiently in *Saccharomyces cerevisiae* and vaccinia virus systems is potentially important. This may serve as a guide for manipulating the expression of the targeted genes. Chlamydial genes optimizing with host-preferred codons are likely to improve the expression levels of the chlamydial genes in a given host. Our preliminary experiments have proved that the chlamydial major outer membrane protein (MOMP) gene optimized with human-preferred codon usage shows a higher level of expression in mammalian cells than the wild-type MOMP gene (data not shown). Thus, yeast and vaccinia virus expression systems may be better applied to the production of chlamydial proteins. We plan to use yeast and vaccinia virus expression systems to test our hypothesis. In summary, our work has provided a basic understanding of the evolution and pathogenesis of *Chlamydia*, with some new insights into the mechanisms for codon usage bias and vaccine development to prevent chlamydial diseases.

References

1 Grantham R, Gautier C, Gouy M. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res, 1980, 8(9): 1893–1912

2 Martin CE, Scheinbach S. Expression of proteins encoded by foreign genes in *Saccharomyces cerevisiae*. Biotechnol Adv, 1989, 7(2): 155–185

3 Lloyd AT, Sharp PM. Evolution of codon usage patterns: The extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. Nucleic Acids Res, 1992, 20(20): 5289–5295

4 Grocock RJ, Sharp PM. Synonymous codon usage in *Cryptosporidium parvum*: Identification of two distinct trends among genes. Int J Parasitol, 2001, 31(4): 402–412

5 Schatcher J. Chlamydial infections. N Engl J Med, 1978, 298(10): 540–548 West SK, Rapoza P, Muñoz B, Katala S, Taylor HR. Epidemiology of ocular

6 Chlamydial infection in a trachoma-hyperendemic area. J Infect Dis, 1991, 163(4): 752–756

7 Ayaslioglu E, Duzgun N, Erkek E, Inal A. Evidence of chronic *Chlamydia pneumoniae* infection in patients with Behcet's disease. Scand J Infect Dis, 2004, 36(6-7): 428–430

8 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(10): 2084–2090

9 Wright F. The "effective number of codons" used in a gene. Gene, 1990, 87 (1): 23–29

10 Moriyama EN, Powell JR. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res, 1998, 26(13): 3188–3193

11 Hou ZC, Yang N. Factors affecting codon usage in *Yersinia pestis*. Acta Biochim Biophys Sin, 2003, 35(6): 580–586

12 Gupta SK, Ghosh TC. Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. Gene, 2001, 273(1): 63–70

13 Hou ZC, Yang N. Analysis of factors shaping *S. pneumoniae* codon usage. Yi Chuan Xue Bao, 2002, 29(8): 747–752

14 Gu W, Zhou T, Ma J, Sun X, Lu Z. Analysis of synonymous codon usage in SARS coronavirus and other viruses in the Nidovirales. Virus Res, 2004, 101 (2): 155–161

15 Wan XF, Xu D, Kleinhofs A, Zhou J. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evol Biol, 2004, 4(1): 19

16 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

17 Sueoka N. Directional mutation pressure, selective constraints, and genetic equilibria. J Mol Evol, 1999, 34(2): 95–114

18 Blake WJ, KAErn M, Cantor CR, Collins JJ. Noise in eukaryotic gene expression. Nature, 2003, 422(6932): 633–637

19 Lesnik T, Solomovici J, Deana A, Ehrlich R, Reiss C. Ribosome traffic in *E. coli* and regulation of gene expression. J Theor Biol, 2000, 202(2): 175–185

20 Sharp PM, Tuohy TM, Mosurski KR. Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res, 1986, 14(13): 5125–5143

21 Romero H, Zavala A, Musto H, Bernardi G. The influence of translational selection on codon usage in fishes from the family Cyprinidae. Gene, 2003, 317(1-2): 141–147

22 Ghosh TC, Gupta SK, Majumdar S. Studies on codon usage in *Entamoeba histolytica*. Int J Parasitol, 2000, 30(6): 715–722

23 Karlin S, Mrazek J. What drives codon choices in human genes? J Mol Biol, 1996, 262(4): 459–472

24 Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci USA, 2004, 101(10): 3480–3485

25 Chen LL, Zhang CT. Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. Biochem Biophys Res Commun, 2003, 306(1): 310–317

Edited by
You-Xin JIN