DNA Research **16**, 91–104, (2009) doi:10.1093/dnares/dsp001

Codon Usages of Genes on Chromosome, and Surprisingly, Genes in Plasmid are Primarily Affected by Strand-specific Mutational Biases in *Lawsonia intracellularis*

Feng-Biao Guo*, and Jian-Bo Yuan

School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China

(Received 13 November 2008; accepted 8 January 2009; published online 15 February 2009)

Abstract

In this study, the factors driving genome-wide patterns of codon usages in *Lawsonia intracellularis* genome are determined. For genes on the chromosome of the bacterium, it is found that the most important source of variation results from strand-specific mutational biases. A lesser trend of variation is attributable to genes that are presumed as horizontally transferred. These putative alien genes are unusually GC richer than the other genes, whereas horizontally transferred genes have been observed to be AT rich in bacteria with medium and relatively low G + C contents. Hydropathy of encoded protein and expression level are also found to influence codon usage. Therefore, codon usage in *L. intracellularis* chromosome is the result of a complex balance among the different mutational and selectional factors. When analyzing genes in the largest plasmid, for the first time it is found that the strand-specific mutational biases are responsible for the primary variation of codon usages in plasmid. Genes, particularly highly expressed genes of this plasmid, are mainly located on the leading strands and this supposed to be the effects exerted by replicational–transcriptional selection. These facts suggest that this plasmid adopts the similar mechanism of replication as the chromosome in *L. intracellularis*. Common characters among the 10 bacteria in whose genomes the strand-specific mutational biases are the primary source of variation of codon usage are also investigated. For example, it is found that genes *dnaT* and *fis* that are involved in DNA replication initiation and re-initiation pathways are absent in all of the 10 bacteria.

Key words: *Lawsonia intracellularis*; codon usage; strand-specific mutational bias; plasmid; replication mechanism

1. Introduction

When sequences of hundreds of microbial protein-coding genes became available in 1980, Grantham et al.[1,2] analyzed frequencies of 61 codons of all these genes. Consequently, they found that a surprising consistency of choices exists among genes of the same or similar genome. The 'genome hypothesis' was hereby proposed.[1,2] Soon after that, it was shown that evident intra-genomic variability existed in many microorganisms.[3] This variation was interpreted as the effect of nature selection acting at the level of translation, which resulted in the preferential usage of optimal codons.[4] The interpretation was reinforced by the finding that the preferred codons in highly expressed genes were recognized by the most abundant tRNA in *Escherichia coli*[5] and as well as in *Saccharomyces cerevisiae*.[6] The selective advantage of optimal codons seems to lie in maximizing

Edited by Minoru Ko

[*] To whom correspondence should be addressed. Fax. +86 28-83201018. E-mail: fbguo@uestc.edu.cn

the efficiency of translation, particularly during periods of competitive exponential growth.[7] In bacteria with slow growth rate, selected codon usage bias may be relatively weak.[7] There may be no such bias in those bacteria for which the competitive growth is unimportant.[8] On the other hand, the preferred codons vary among species based on the changes in the complement of tRNAs in that bacterium.[9]

Besides translational selection, replicational and transcriptional selection may also have influence on the codon usage of a gene.[10,11] Replicational selection is responsible for the higher number of genes on the leading strands, and transcriptional selection appears to be responsible for the enrichment of highly expressed genes on these strands.

The effects of mutation may be superimposed on biases generated by natural selection.[12] In most bacteria, there are short chromosome segments of unusual base composition due to the relatively recent import of the region through horizontal transfer.[13,14] Genes located in these regions possess distinct codon usage or nucleotide composition from other genes, for example, in *E. coli*[15] and in *Bacillus subtilis*.[16] In a single known example, *Mycoplasma genitalium*, codon usage variation is continuous and associated very strongly with position on the chromosome, perhaps reflecting change in the spectrum of mutations around the genome.[17] In addition, many bacteria exhibit skewed base composition between the leading and lagging strands of replication, although the magnitude of this skew varies considerably among species.[18]

Based on the above narration, the variation of codon usage of genes within a species is due to the combined effect of mutation and selection.[12] Among these factors, the bias from asymmetric replication mechanism received special attention of researchers in the past 10 years.[19] Many researches have been performed to analyze the effect on the codon usage (and/or amino acid composition) exerted by the asymmetric mutation and to investigate the underlying mechanism of the different mutation.[20–27] The skewed base composition between two replicating strand was first observed in *E. coli*, *M. genitalium*, *Haemophilus influenzae* and *B. subtilis*.[28,29] Then similar observations were obtained in most of the other bacteria. In 1998, for the first time it was found that the asymmetric replication is the major source of codon usage variation.[10] This observation was obtained in *Borrelia burgdorferi* genome.[10] The effect of asymmetry was so strong that the codon usages of genes on the two replicating strands were separated, distinct. After that, the separated codon usages between two replicating strands were also observed in *Treponema pallidum*,[30] *Chlamydia trachomatis*,[11] *Buchnera aphidicola*,[31] *Blochmannia floridanus*,[32] *Bartonella henselae*,[33] *Bartonella quintana*,[33] *Tropheryma whipplei*[34] and *Chlamydia muridarum*.[35]

Lawsonia intracellularis is an obligate intracellular Gram-negative bacterial pathogen.[36] Though primarily recognized in pigs, *L. intracellularis* is spreading to a wide range of mammals such as horses, and hamsters in North America and elsewhere. The bacterial pathogen invades the intestinal epithelial cells, which causes hyperplasia of the infected cells and leads to the process of disease pathogenesis. The disease has two clinical manifestations: an acute hemorrhagic form often referred as porcine hemorrhagic enteropathy, and a more chronic proliferative form often called porcine intestinal adenomatosis. Genome of *L. intracellularis* PHE/MN1-00 was determined in 2006, which provided a wonderful opportunity to extract a wealth of information on biochemistry, genetics, evolutionary history and pathogenicity of this organism.

Traditionally, codon usage data have been used in a wide variety of areas.[10] It is often desirable to use codon usage information to reduce the redundancy of primers for the PCR. Optimizing the codon usage of a gene could increase its expression level. Codon usage tables have been used to identify those ORFs that may encode proteins. Codon usage patterns also have been used to identify ORFs that probably do not code for functional proteins. Because of the importance of the intracellular pathogen and the potential usage of codon usage patterns, the intragenomic variation in codon usage in *L. intracellularis* PHE/MN1-00 has been investigated through multivariate analysis method in this study.

2. Materials and methods

2.1. The database

The complete genome sequence of *L. intracellularis* PHE/MN1-00 was downloaded from GenBank ftp site. One chromosome and three plasmids are contained in the complete genome. In this work, Plasmid 1 and Plasmid 2 are not taken into account because they contain too little genes to be analyzed statistically. Plasmid 3 is the largest plasmid and also analyzed. Chromosome has 1 457 619 bp and Plasmid 3 has 194 553 bp. A total of 1180 and 104 protein-coding genes are listed in the annotations of the chromosome and the largest plasmid, respectively. No attempt was made to alter the sequences or to remove those genes of unknown function. The FASTA formatted files, which are used as input files for codonW software, are proved as Supplementary data. In Supplementary File 1, the DNA sequences of 1180 genes located on the chromosome of

L. intracellularis are contained. The first 607 genes correspond to those located on the leading strands and the last 573 ones correspond to those on the lagging strands. In Supplementary File 2, the DNA sequences of 104 genes in the largest plasmid (Plasmid 3) of *L. intracellularis* are contained. The first 68 genes correspond to those located on the leading strands and the last 36 ones correspond to those on the lagging strands.

2.2. *Statistical analysis*

Most analyses were carried out by using codonW,[37] which can be freely downloaded from the website (http://sourceforge.net/projects/codonw/). $GC3_S$ denotes the frequencies of G and C at the third synonymously variable coding position (excluding Met, Trp and termination codons). N_C means the 'effective number of codons' used in a gene.[38] When all sense codons are used randomly, N_C takes a value of 61. Lower values of N_C indicate stronger bias, with an extreme value of 20 when only one synonymous codon is used for each amino acid. After calculating N_C and $GC3_S$ for each gene, N_C–$GC3_S$ plot can be made, which shows whether there are genes whose codon usage is affected by genome composition pressure and natural selection or mutation. An expected curve is plotted through the formula: $N = 2 + s + \{29/[s^2 + (1 - s)^2]\}$. For each gene, codon adaptive index (CAI) and hydropathy values (gravy) are also calculated by codonW.

Correspondence analysis (COA), as implemented in codonW, was used to determine the major source of variation of codon usage among the genes on the chromosome and the genes in Plasmid 3. As suggested by Perrière and Thioulouse,[39] parallel COA on codon counts and on relative codon frequencies were performed and then the results were compared. In addition, COA was carried out for genes on the chromosome, for genes in the largest plasmid and for the genes located on the leading strands of the chromosome, respectively. Relative synonymous codon usage (RSCU) is defined as the observed frequency of a codon divided by that expected when all codons for that amino acid are used equally. Therefore, RSCU values close to 1.0 indicate a lack of bias for that codon. Compared with simple measurements of codon abundance, RSCU values are normalized and are much more independent of amino acid usage. Only those codons for which there is a synonymous alternative were used in the analysis. Hence, the three termination codons and the codons that encode methionine and tryptophan are excluded. Consequently, each gene is described by a vector of 59 variables (codons). COA maps all the genes analyzed into the 59-dimensional space and attempts to identify a series of new orthogonal axes accounting for the greatest variation among genes. The first principal axis is chosen to maximize the standard deviation of the derived variable and the second principal axis is the direction to maximize the standard deviation among directions un-correlated with the first, and so forth. For details about this method, refer to Dillon and Goldstein.[40]

GC skew $[(G - C)/(G + C)]$ was used to determine the origin and termination of replication for the chromosome and the plasmid.[28] A non-overlapping sliding window of 1000 bp was employed for the GC skew. For the chromosome, the origin site is assumed to lie between genes LI0775 and LI0776, whereas the termination between genes LI0227 and LI0228. For the largest plasmid, the origin lies before gene LIC020, whereas the terminus between genes LIC062 and LIC063.

In order to check whether there have significant differences of codon usage between genes on the leading strands and those on the lagging strands, a χ^2 test was employed. Significance was examined at the 5% level (χ^2 value of 3.841). Significance was evaluated for the 59 sense codons for which there are a synonymous alternative.

3. Results

3.1. *Global codon usage of genes on the chromosome*

It is widely accepted that global codon usage in unicellular species that displays extremely biased genomic composition is predominantly shaped by the compositional pressure.[41] This viewpoint is confirmed again in *L. intracellularis* genome. As can be seen from Table 1, the global codon usage in 1180 genes on the chromosome in *L. intracellularis* shows the expected bias toward AT-rich codons. This enrichment is much stronger at the third codon position than at the first two positions. For all of the 18 amino acids (excepting Met and Trp), the frequencies of A- or T-ending codons are much more than those of G- or C-ending synonyms.

As suggested by Wright,[38] a plot of N_C against $GC3_S$ can gives a useful visual display of the main features of codon usage patterns for a number of genes. If a gene is only subject to G + C-biased mutational pressure, it will lie on the $GC3_S$ curve. It will lie just below the $GC3_S$ curve if a gene is under selection (either negative or positive) for codons in C and/or G. In other words, the gene will lie over the $GC3_S$ curve if it is subject to other kinds of selection and/or any kinds of mutation pressure. Such a plot for genes on the chromosome is shown in Fig. 1. According to the figure, there exists significant heterogeneity of codon usage among genes on the

Table 1. Global codon usage of 1180 genes on the chromosome of *L. intracellularis*

AA	Codon	*N*	RSCU	AA	Codon	*N*	RSCU
Phe	UUU	15 112	(1.72)	Ser	UCU	10 541	(2.23)
	UUC	2411	(0.28)		UCC	1463	(0.31)
Leu	UUA	15 959	(2.19)		UCA	7238	(1.53)
	UUG	2922	(0.40)		UCG	525	(0.11)
Tyr	UAU	10 979	(1.67)	Cys	UGU	4395	(1.68)
	UAC	2160	(0.33)		UGC	841	(0.32)
Ter	UAA	777	(0.00)	Ter	UGA	152	(0.00)
	UAG	246	(0.00)	Trp	UGG	4624	(1.00)
Leu	CUU	16 136	(2.21)	Pro	CCU	9194	(2.07)
	CUC	2328	(0.32)		CCC	1075	(0.24)
	CUA	5306	(0.73)		CCA	7189	(1.62)
	CUG	1076	(0.15)		CCG	334	(0.08)
His	CAU	7724	(1.68)	Arg	CGU	6148	(2.21)
	CAC	1494	(0.32)		CGC	1033	(0.37)
Gln	CAA	13 060	(1.58)		CGA	2413	(0.87)
	CAG	3462	(0.42)		CGG	637	(0.23)
Ile	AUU	19 619	(1.65)	Thr	ACU	8245	(1.36)
	AUC	4019	(0.34)		ACC	1853	(0.31)
	AUA	12 077	(1.01)		ACA	13 220	(2.18)
Met	AUG	8968	(1.00)		ACG	904	(0.15)
Asn	AAU	14 715	(1.57)	Ser	AGU	6659	(1.41)
	AAC	4058	(0.43)		AGC	1911	(0.40)
Lys	AAA	20 477	(1.62)	Arg	AGA	5064	(1.82)
	AAG	4835	(0.38)		AGG	1384	(0.50)
Val	GUU	11 365	(1.88)	Ala	GCU	12 191	(1.85)
	GUC	2237	(0.37)		GCC	2009	(0.30)
	GUA	8764	(1.45)		GCA	11 581	(1.75)
	GUG	1793	(0.30)		GCG	639	(0.10)
Asp	GAU	15 082	(1.67)	Gly	GGU	10 262	(1.55)
	GAC	3017	(0.33)		GGC	2397	(0.36)
Glu	GAA	19 734	(1.57)		GGA	10 629	(1.60)
	GAG	5331	(0.43)		GGG	3217	(0.49)

chromosome: $GC3_S$ values range from 0.07 to 0.61, whereas N_C values range from 31.1 to 60.0. It can also be seen that one-third of genes lie above the $GC3_S$ curve, and the retaining two-thirds lie below the $GC3_S$ curve. For genes that lie above the $GC3_S$ curve, there may exist mutational or selectional pressure that leads to A- and/or T-ending codons. On inspection, almost all of these genes are located on the lagging strands and most of these genes are those likely to be lowly expressed. It should also be noted that dozens of genes are located far from the majority. These genes have high $GC3_S$ values and high N_C values, which are marked by open triangles in the plot. Origin of these genes will be discussed in the later section.

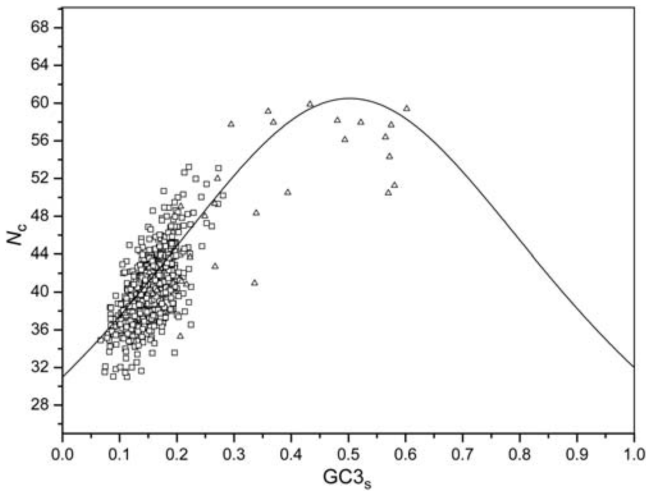

Figure 1. The effective number of codons (N_C) plotted against the G + C content at the synonymously variable third position ($GC3_S$), for the 1180 genes on the chromosome. Open triangles indicate genes with unusual G + C content, speculated as foreign genes. Open squares denote the other genes.

3.2. *Strand-specific composition bias at three codon positions of genes on the chromosome*

GC-skew analysis shows a clear polarity switch at two points, around 284 and 975 kb on the chromosome in *L. intracellularis*, suggesting that the putative replication terminus and origin sites might be located in these regions. According to the record for this bacterium in the Doric database,[42] GC disparity that is a component of Z curve also shows a clear minimum and maximum at these points for this genome. Comparisons with the consensus sequence for the non-perfect DnaA box motif (ttttcaaca) reveal that a non-translatable region between 983 356 and 984 329 bp possesses cluster of three putative DnaA boxes, thereby confirming the possible locations of the functional chromosomal origins between two genes, LI0775 (trkH) and LI0776 (psd). The existence of a DnaA gene around 980 kb furthermore confirms this location as replication origin.

Table 2 shows the frequencies of nucleotide A, C, G, T and the mean G + T contents at three codon positions of the genes located on the leading and lagging strands of the *L. intracellularis* chromosome. Strand-specific skews, known to influence codon usage in other bacteria,[18] are also found to have the same influences on *L. intracellularis*. The leading strand genes show an excess of G over C and T over A, whereas the case for the lagging strands is opposite. The *t*-test shows that the mean G + T content of each codon position of the genes located on the leading strands is significantly different from that on the lagging strands of replication. This indicates that the strand-specific compositional bias has significant

Table 2. Base compositions and medium G + T content at three codon positions for genes located on the leading and lagging strands of the chromosome

		A Mean (SD)	C Mean (SD)	G Mean (SD)	T Mean (SD)	G + T Mean (SD)
First codon position	Leading	0.304 (0.036)	0.177 (0.031)	0.309 (0.047)	0.209 (0.046)	0.519 (0.039)
	Lagging	0.331 (0.040)	0.204 (0.032)	0.276 (0.054)	0.189 (0.040)	0.464 (0.043)
Second codon position	Leading	0.308 (0.058)	0.206 (0.035)	0.162 (0.032)	0.325 (0.054)	0.486 (0.055)
	Lagging	0.319 (0.060)	0.220 (0.031)	0.141 (0.030)	0.320 (0.049)	0.461 (0.057)
Third codon position	Leading	0.364 (0.045)	0.065 (0.038)	0.127 (0.031)	0.444 (0.047)	0.571 (0.041)
	Lagging	0.391 (0.040)	0.102 (0.024)	0.078 (0.022)	0.429 (0.037)	0.507 (0.037)
All codon positions	Leading	0.325 (0.032)	0.149 (0.022)	0.199 (0.033)	0.326 (0.036)	0.525 (0.033)
	Lagging	0.347 (0.035)	0.175 (0.017)	0.165 (0.024)	0.313 (0.029)	0.478 (0.034)

influence on nucleotide selection not only at the third codon position but also at the first and second positions. The inter-strand variation in G + T content is highest in the third and lowest in the second codon position, whereas the intra-strand variation is highest in the second codon position, which is reflected by the highest standard deviations.

3.3. *The first trend is associated with strand-specific mutational biases*

COA of codon usage was used to study extensively and quantitatively the variation of codon usage among the 1180 genes on the chromosome. Because the use of relative codon frequency sometimes introduces other biases and often diminishes the quantity of information to analyze, occasionally resulting in interpretation errors, we compute in parallel COA on codon counts and on RSCU and then compare the results in this study. Fig. 2(a) and (b) shows the positions of the genes along the first and second major axes produced by COA on codon counts and RSCU values, respectively. The closeness of any two genes on each plot reflects the similarities of their codon usages. In the following sections, the factors that drive variation of synonymous codon usage of genes on the chromosome are discussed.

Both in Fig. 2(a) and (b), the first axis separate the genes into two clusters with little overlap between them. The following two results could indicate that the two clusters correspond to genes on the leading and lagging strands of replication. (i) The first axis is found to strongly correlate with GC and AT skews, particularly at the third position. At the left of the first axis, genes are characterized by richness in nucleotides G and T, whereas it is opposite at the right. On the other hand, it has been found that there is an excess of nucleotides G relative to C in the leading strands and of C to G in the lagging strands in most bacterial genomes, which is frequently accompanied by an abundance of T over A in the leading strand.[18] (ii) The coordinates of individual genes along the first axis are plotted against the chromosomal locations of the corresponding genes in Fig. 3. Genes on the Watson strand and those on the Crick strand are denoted by black and gray squares, respectively. It is found that

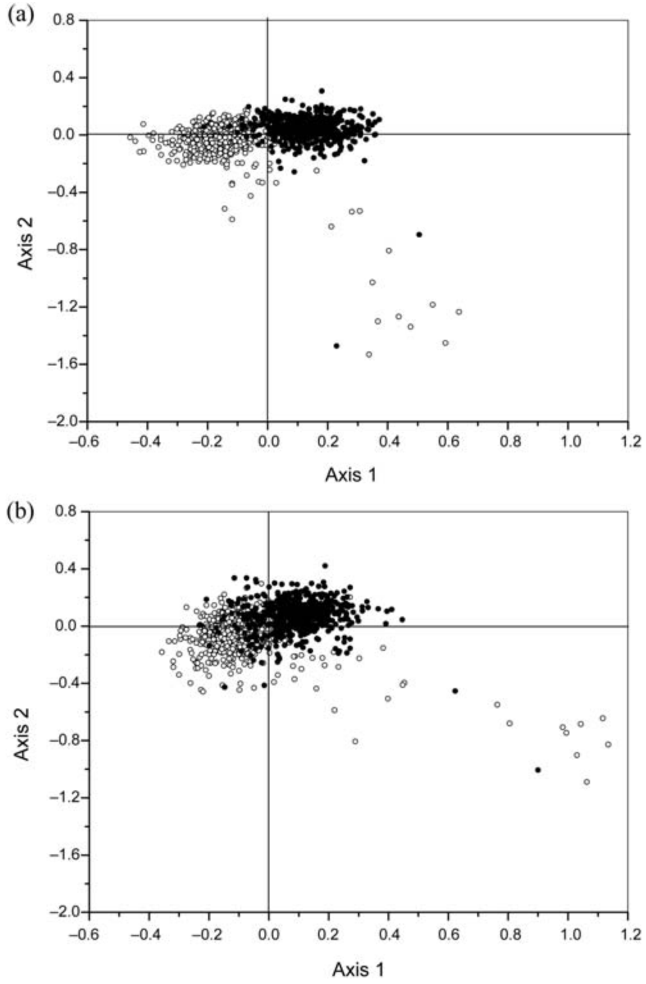

Figure 2. Plot of the two most important axes after COA for the 1180 genes on the chromosome. The open circles indicate the genes that are transcribed on the leading strands of replication. The filled circles indicate the genes that are transcribed on the lagging strands of the replication. (**a**) Use codon count to compute COA (COA/counts), the first and second axes account for 10.5% and 7.5% of the total inertia of the 59-dimensional space. (**b**) Use RSCU to compute COA (COA/RSCU), the first and second axes account for 9.2% and 7.7% of the total inertia of the 59-dimensional space.

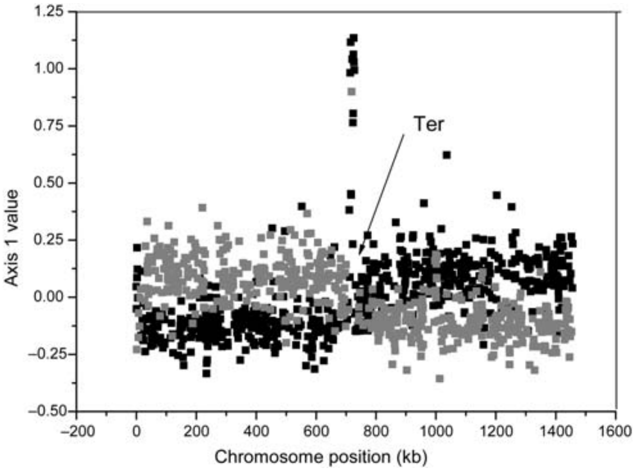

Figure 3. Plot of Axis 1 values of chromosomal genes against their corresponding chromosomal locations in the *L. intracellularis* genome. Black squares represent genes located on the Watson strand and gray squares represent genes on the Crick strand. To allow convenient observation, the beginning point of the horizontal axis is shifted to the origin of replication.

genes on the left side of Watson strand and those on the right side of Crick strand have low values of coordinates along the Axis 1, whereas, for the other genes, the case is opposite. In fact, genes on the left side of Watson strand and those on the right side of Crick strand just correspond to genes on the leading strands, the other ones correspond to lagging strands. Therefore, it is reasonable to say that two clusters in Fig. 2 correspond to genes on the leading strands and lagging strands, respectively. After marking genes located on the leading and lagging strands by different symbols in Fig. 2, the speculation is confirmed.

A χ^2 test was performed for RSCU of genes located on the leading versus lagging strands and the results are listed in Table 3. As can be seen, 49 among 59 codons are found to be significantly different between genes on the two strands of replication. The 23 codons used more frequently in the leading strands are G-ending or T-ending, except TTA, ACA, AGA and GCA. Among the 26 codons used more frequently in the lagging strands, 16 are C-ending, eight are A-ending and the exceptions are CTT and ACT. Results of the test confirm that there is a bias toward G, T in the leading strands and toward C, A in the lagging strands of replication. Therefore, it can be concluded that in *L. intracellularis*, the leading and lagging strands of replication display an asymmetry in the mutational biases and or the differential correction/repair rates, and as shown in several other bacteria,[10,11,30–35] this difference is the most important source of codon usage variation.

Furthermore, there are more annotated genes on the leading strands than on the lagging strands. The numbers are 607 and 573, respectively. However, the two numbers in *L. intracellularis* are less different than that in *B. burgdorferi*.[10] Therefore, the effect of replication selection is weaker than that in *B. burgdorferi*. Three sets of genes, including ribosomal proteins, translation/transcription processing factors and the major chaperones and degradation genes,[43] are chosen as representative of highly expressed genes. For these putative highly expressed genes, the distribution on the two replicating strands is much more skewed. More than 59% of the 61 putative highly expressed genes are transcribed on the leading strands. So, the differences between the leading and lagging strands indicate the combined effects of mutation and selection induced by the replication–transcription. Replicational selection, although weak, may be responsible for the higher number of genes located on the leading strands,[10,11] and transcriptional selection appears to be responsible for the enrichment of highly expressed genes on these strands.[10,11] Replicational–transcriptional selection coupled with asymmetric mutational bias is, therefore, the most important cause of intra-chromosome variations of synonymous codon usages in *L. intracellularis*.

As mentioned above, the separated codon usages between the two replicating strands have been found in nine bacteria. Among these species, *B. burgdorferi* shows an extremely strong bias of codon usage.[10] Lobry and Sueoka[22] once described a method of graphical display of the influence of replication bias on leading versus lagging strands. It would be interesting to compare the graphics, namely PR2-plots, obtained for *L. intracellularis* chromosome with those of *B. burgdorferi*. As can be seen from Fig. 4(a) and (b), the strand biases of G/C, which could be reflected by the values of the horizontal axis of the plot, are slightly weaker in *L. intracellularis* than those in *B. burgdorferi*, whereas the strand biases of T/A (reflected by the values of the vertical axis) are much weaker in the former than those in the latter. In both figures, the strand-specific biases are strong enough to separate the genes on the two replicating strands.

3.4. *The second trend may be associated with horizontal gene transfer*

When analyzing the second trend, it is found that there exists a strong negative correlation between the Axis 2 (COA/RSCU) and $GC3_S$ ($r = -0.6559$). If analysis is restricted to the 87 genes whose coordinates along Axis 2 are less than −0.2, the correlation is more significant. Marking these genes by open triangles in the N_C–$GC3_S$ plot, it is found that these genes are located far from the other ones and have higher N_C and $GC3_S$. And this fact means the different codon usage of these genes with that of the majority.

Table 3. Results of χ^2 test for RSCU of genes on the leading and lagging strands of the chromosome

AA	Codon	RSCU leading	Significant[a]	RSCU lagging	AA	Codon	RSCU Leading	Significant[a]	RSCU lagging
Phe	UUU	1.81	≫	1.64	Ser	UCU	2.21	–	2.25
	UUC	0.19	≪	0.36		UCC	0.23	≪	0.39
Leu	UUA	2.53	≫	1.86		UCA	1.47	≪	1.60
	UUG	0.60	≫	0.20		UCG	0.14	≫	0.08
Tyr	UAU	1.74	≫	1.60	Cys	UGU	1.77	≫	1.59
	UAC	0.26	≪	0.40		UGC	0.23	≪	0.41
Ter	UAA	0.00	–	0.00	Ter	UGA	0.00	–	0.00
Ter	UAG	0.00	–	0.00	Trp	UGG	1.00	–	1.00
Leu	CUU	1.98	≪	2.44	Pro	CCU	2.08	–	2.06
	CUC	0.17	≪	0.46		CCC	0.18	≪	0.30
	CUA	0.56	≪	0.89		CCA	1.64	–	1.59
	CUG	0.15	–	0.14		CCG	0.10	–	0.05
His	CAU	1.75	≫	1.61	Arg	CGU	2.24	–	2.18
	CAC	0.25	≪	0.39		CGC	0.26	≪	0.51
Gln	CAA	1.41	≪	1.74		CGA	0.74	≪	1.02
	CAG	0.59	≫	0.26		CGG	0.25	–	0.20
Ile	AUU	1.72	≫	1.59	Thr	ACU	1.33	≪	1.39
	AUC	0.26	≪	0.41		ACC	0.24	≪	0.36
	AUA	1.02	–	1.01		ACA	2.24	≫	2.14
Met	AUG	1.00	–	1.00		ACG	0.19	≫	0.12
Asn	AAU	1.66	≫	1.49	Ser	AGU	1.58	≫	1.24
	AAC	0.34	≪	0.51		AGC	0.36	≪	0.45
Lys	AAA	1.47	≪	1.76	Arg	AGA	1.88	≫	1.75
	AAG	0.53	≫	0.24		AGG	0.63	≫	0.34
Val	GUU	1.93	≫	1.82	Ala	GCU	1.85	–	1.84
	GUC	0.29	≪	0.47		GCC	0.25	≪	0.36
	GUA	1.40	≪	1.51		GCA	1.78	≫	1.73
	GUG	0.37	≫	0.20		GCG	0.12	–	0.08
Asp	GAU	1.73	≫	1.60	Gly	GGU	1.67	≫	1.41
	GAC	0.27	≪	0.40		GGC	0.32	≪	0.41
Glu	GAA	1.43	≪	1.73		GGA	1.46	≪	1.76
	GAG	0.57	≫	0.27		GGG	0.55	≫	0.42

[a] ≫ indicates that the leading strand genes use the codon more frequently than the lagging strand genes; ≪ indicates the lagging strand genes use the codon more frequently than the leading strand genes; – indicates that there is no significant difference in the usage of the codon on either strand. Significance is examined at the level of 5%.

As well known, genes that were recently imported through horizontal transfer would have unusual codon usages and base compositions. In *Pseudomonas aeruginosa*, putative alien genes showed higher N_C values than the majority of genes.[12] From the widely used public database HGT-DB,[14] we downloaded the information for all the predicted horizontally transferred genes in *L. intracellularis*. Among the 10 genes that have the lowest values of Axis 2, nine are predicted as horizontally transferred genes according to the results in HGT-DB.[14] The exceptional gene is as short as 104 codons. Based on the above analysis, it is reasonable to come to the conclusion that most of the 87 genes may be transferred horizontally. It should be noted that 20 of these genes are located around the replication terminus, which has been shown to be a hot spot of mutation and chromosome recombination.

3.5. *Hydropathy of encoded proteins influences the codon usage variation*

Fig. 5 shows the positions of the 1180 genes along the third and forth major axes produced by COA on

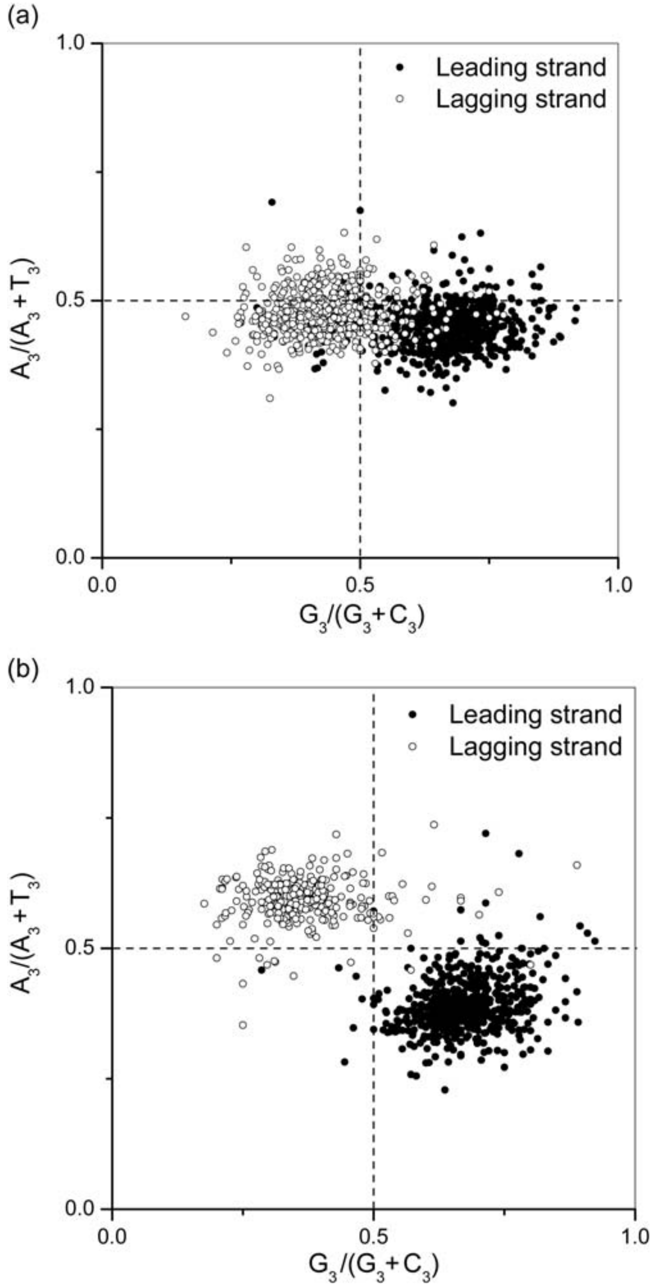

Figure 4. PR2-plots for *L. intracellularis* chromosome and *B. burgdorferi* genome. G/(G + C) and A/(A + T) at the third codon positions of protein-coding genes are calculated and plotted as the horizontal and vertical axes, respectively. For details of the PR2-plot, please refer to Lobry and Sueoka.[22]

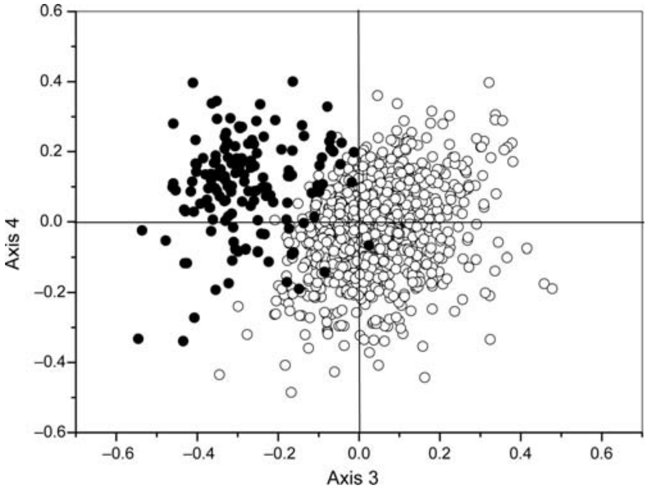

Figure 5. Plot of the third and fourth axes after COA of codon counts for the 1180 genes on the chromosome. Proteins having a gravy score >0.3 are represented as filled circles and the other genes are represented as open circles.

codon counts. As can be seen from the figure, the third axis separates another group of genes from the main group. After computing the gravy score (a measure of hydrophobicity) of all the proteins encoded by the genes, it is found that the proteins separated from the others on the third axis have very high scores (gravy > 0.3) and so are hydrophobic. These results are nothing but the superimposition of amino acid bias on codon usage bias. If we use RSCU to compute the COA, the third major axis could not discriminate the genes according to the hydrophobicity values. It should be noted that this is the same case as that in the genome of *Lactococcus lactis*.[44]

3.6. *Influence of the expression level on the codon usage*

In order to investigate whether codon usage patterns are further shaped by factors at the level of translation, we conduct COA of codon counts and RSCU values on the genes located on the leading strands of replication, because most of the 61 putative expressed genes are located in that strand. Fig. 6(a) and (b) shows the positions of the genes along the first and second major axes produced by such COA on codon counts and RSCU, respectively. As can be seen from the two figures, almost all of the putative highly expressed genes have positive scores along the second axis, whereas genes presumably expressed at low level are scattered everywhere of the whole distribution area. In addition, we calculate the CAI value for each gene located on the lagging strands. The correlation between the values of Axis 2 and CAI is statistically significant ($r = 0.494$, $P < 0.0001$). The above analyses suggest that the second axis is associated with expression level.

3.7. *COA of RSCU of genes in the largest plasmid*

COA of RSCU of genes in the largest plasmid in *L. intracellularis* was performed to determine the most important factor that drives synonymous codon usage patterns. The other two plasmids are not analyzed because there are too little genes to perform multivariate statistical analysis. One plot of the two most important axes after the COA is shown in Fig. 7. The first and second axes account for 12.7% and 8.2% of the total inertia of the

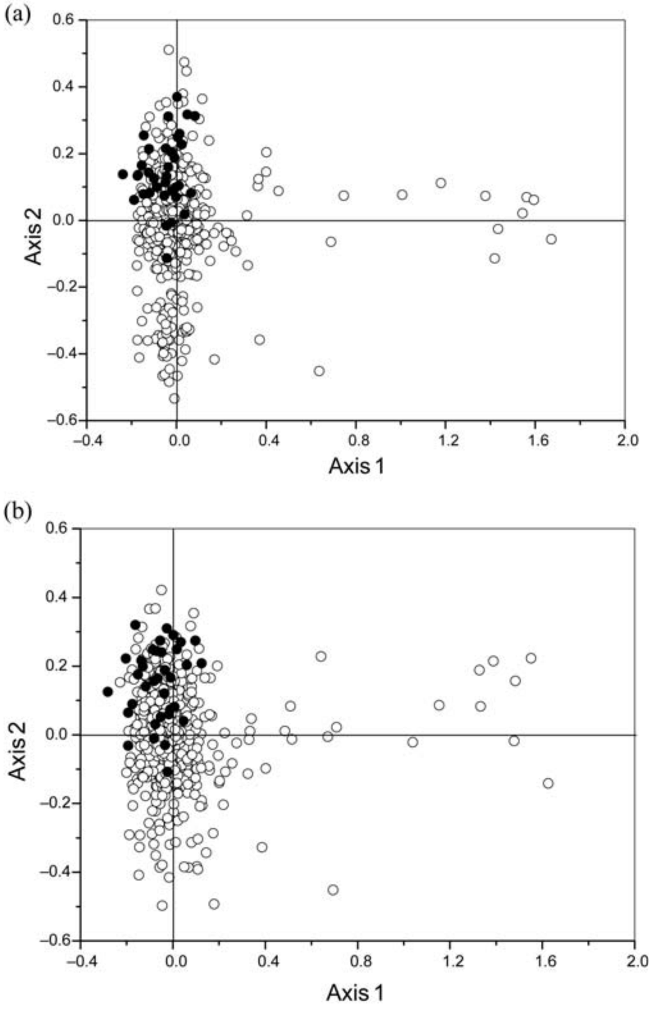

Figure 6. Plot of the two most important axes after COA for 607 chromosome genes that are located on the leading strands of replication. The filled circles indicate 36 putative highly expressed genes. The open circles denote other genes. (**a**) Use codon count to compute COA (COA/counts), the first and second axes account for 12.7% and 7.2% of the total inertia of the 59-dimensional space. (**b**) Use RSCU to compute COA (COA/RSCU), the first and second axes account for 13.5% and 6.3% of the total inertia of the 59-dimensional space.

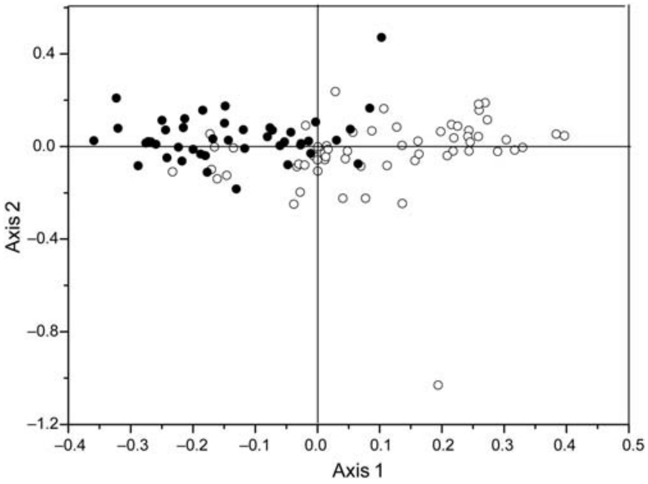

Figure 7. Plot of the two most important axes after COA of RSCU values for the 104 genes in the largest plasmid. The open circles indicate the genes that are transcribed on the leading strands of replication. The filled circles indicate the genes that are transcribed on the lagging strands of the replication.

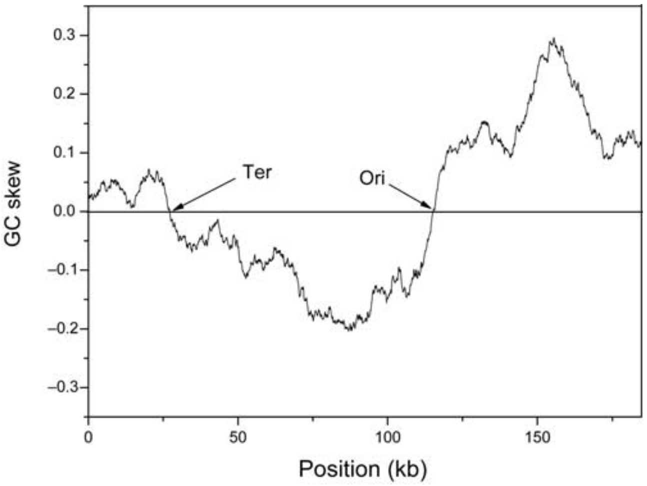

Figure 8. GC-skew along the DNA sequence of the largest plasmid. It is calculated using a non-overlapping sliding window of 1000 bp. As can be seen, there are two clear polarity switches, suggested as putative origin and terminus of replication.

59-dimensional space, respectively. After determining replication origin and terminus sites by GC-skew analysis shown in Fig. 8, 68 genes are found to lie on the leading strands and 36 ones on the lagging strands. There is more significant difference between the numbers of genes on the two replicating strands in the plasmid than those on the chromosome, suggesting that replicational selection in the former is stronger than that in the latter. Marking the two groups of genes by open and filled circles, and then it is found that genes on the leading strands lie on the right side of the Axis 1, whereas lagging strand genes lie on the left side. χ^2 test was performed for RSCU of genes located on the leading versus lagging strands in this plasmid. Consequently, 23 among 59 codons are found to be significantly different between genes on the two strands of replication. Among the 13 codons used more frequently in the leading strands, nine are G-ending or T-ending, except TTA, ATA, CCA and GCA. Among the 10 codons used more frequently in the lagging strands, eight are C-ending or A-ending and the exceptions are CTT and TCT. Therefore, the strand-specific mutational biases are responsible for the major variation of synonymous codon usages of genes in the plasmid. No known factors are found to correlate with the second axis of this COA.

To test whether the transcriptional selection exerts influence on the genes, CAI values are calculated for 104 genes in the plasmid by using ribosomal protein-coding genes as reference set. Consequently, it is found that all of the seven genes that have the highest CAI values are located on the leading strands. This suggests that transcriptional selection do have influence on the genes of the plasmid. Replicational selection and transcriptional selection are two different kinds of selective pressures. Although the two selective pressures yield similar consequences, they are very distinct.[10] The asymmetric mechanism of replication in the plasmid may also be shown by the following result. CMR database at TIGR lists the function categories of known genes in this plasmid. After comparing the numbers of the known genes on the two replicating strands, it is found that 15 of the 22 cell envelope-related genes are located on the lagging strands. On the other hand, 10 of the 11 genes with mobile and extrachromosomal element functions are located on the leading strands.

4. Discussion

4.1. *Underlying mechanisms of replicational–transcriptional mutation and selection*

It is widely accepted that codon usage variation in a species is the combined effect of mutation bias and nature selection.[12] In *L. intracellularis* chromosome, genes located on the two replicating strands are shown to have distinct codon usages. On the other hand, genes, particularly those are highly expressed, are mainly located on the leading strands. It is important to investigate the underlying mechanisms of the two effects. According to McInerney[10] and Romero et al.,[11] the former is caused by the strand mutational bias and the latter results from replication–transcription selection. Also, some researchers believe that the strand-specific compositional biases are not only the result of strand mutation biases but also the superimposition of differential mutation rate and differential correction/repair rates.[45]

Among the theories aimed at explaining strand mutation biases, it seems that the cytosine deamination theory enjoys the most attention.[19] The deamination of cytosine results in the formation of uracil. In normal circumstance *in vivo*, cytosine is effectively protected against deamination because of the Watson–Crick base paring. But the rate of cytosine deamination increases 140 times when the DNA is single-stranded.[46] If the resulting uracil is not replaced with cytosine, C to T mutation occurs. During the replicating process, the leading strand is much more exposed in the single-stranded state. Therefore, the C to T mutation occurs more frequently in the leading strands than in the lagging strands and then the excesses of G relative to C and T relative to A are formed in the leading strands. According to Furusawa and Doi,[47] such fidelity difference between the leading and lagging strands may make it possible to accelerate the evolution of unicellular and multicellular organisms and avoid the extinction of the population.

As far as gene orientation biases are concerned, there exist similar explanations. For genes on the leading strands, RNA polymerase, when transcribing them, moves in the same direction as a replication fork would move during replication, whereas opposite for genes on the lagging strands. Transcription–replication should be more effective if one organism maintains most of its genes, particularly highly expressed genes on the leading strands. The high efficiency results from the three factors: (i) the same direction reduces the probability of head-on collisions between the polymerases involved in the replicational and transcriptional processes; (ii) transcription may not be aborted by the replication complex; and (iii) the inverse orientation is very disadvantageous because of the possible lack of solution mechanism for head-on collisions.[48] In highly expressed genes, the selective advantage of transposition to the leading strands is more significant than that of lowly expressed genes. Therefore, highly expressed genes are much more likely to overcome random genetic drift, and these genotypes become fixed more easily in the population. Lowly expressed genes do not interfere with replication to such an extent as highly expressed genes, and, also, the interruption of lowly expressed gene transcription is not nearly as deleterious. So, the selective advantage is not so great in lowly expressed genes. Transposition of a lowly expressed gene from a lagging strand to a leading strand may not offer a sufficient selective advantage and therefore may not become fixed so easily in the population.[10]

4.2. *Codon usages of genes in the plasmid and asymmetric mechanism of replication*

Usually, bacterial plasmids replicate using a different mechanism than that of the chromosome of their host cell. As an exception, cumulative skew diagrams showed that plasmid and chromosome of *B. burgdorferi* adopted a similar bi-directional replication.[49] Such common replication mechanism was consistent with previous suggestions that *Borrelia* plasmids were actually mini-chromosomes.[50] In this work, GC-skew analysis in Fig. 8 shows clear polarity switch at two points, around 27 and 115 kb in the largest plasmid of *L. intracellularis*. This suggests that this plasmid replicates bi-directionally from an internal origin as the chromosome does. Leading strands and lagging strands are hence determined based on the putative origin and terminus. COA shows that genes on the leading and lagging strands have distinct codon usages. The same results are observed in genes on the chromosome, which is supposed to be caused by the strand-specific mutational bias. Similarly, the difference between codon usages of genes on the two replicating strands in the plasmid is very likely to result from the different mutation and/or repair rates. If this speculation is the real case, then common asymmetric replication would be involved in the chromosome and the largest plasmid of *L. intracellularis*. Not only both replicate bi-directionally from internal origin, but also have biased mutation/repair rate between the two replicating strands.

Replicational–transcriptional selection is also found to exert pressure in the largest plasmid of *L. intracellularis*. The fact that most $(68/104 = 65.4\%)$ genes are located on the leading strands suggests the existence of replicational selection. In addition, all of the seven genes that have the highest CAI values are located on the leading strands. This suggests that transcriptional selection has influence on the genes of the plasmid. All the above facts suggest that this plasmid adopts the similar mechanism of replication as the chromosome in *L. intracellularis*. Perhaps, the largest plasmid of *L. intracellularis* is one mini-chromosome, as that in *B. burgdorferi*.[50]

4.3. *The 'GC-richness' of the putative alien genes in L. intracellularis*

Over 10 years ago, the tendency of horizontally transferred genes to be A + T-rich had been noted in species having intermediate G + C contents.[51] After that, the same phenomenon was observed for *Helicobacter pylori* and *Streptococcus pneumoniae*,[52] which have low G + C contents. This striking pattern raises questions about the nature and the source of these horizontally transferred genes. Lawrence and Ochman[53] hypothesized that the recently transferred genes were adapted to the genomic context of other distant species. The results of Daubin et al.[52] suggest that either the donor genomes are always more A + T rich than the acceptor genomes or there is a bias toward the internalization of A + T-rich exogenous DNA in the genome.

However, the putative alien genes in *L. intracellularis* are found to have higher GC than the other genes. And even, some of the putative alien genes are GC-rich. Perhaps, someone will think this is an exception. But we do not think so. In contrast, we suppose that this may be a usual pattern of bacterial genomes with very low GC contents. Genomic GC contents of *H. pylori* and *S. pneumoniae*, in which AT-rich alien genes are found, are low but not very low. Their GC contents are still 6% higher than those of *L. intracellularis*, which is 33%. For bacteria with GC content low as 33%, it is difficult to obtain a donor species that have higher AT contents. Therefore, alien genes will be inclined to GC-richer than the other genes in the acceptor species.

4.4. *Common genomic characters of bacteria in which strand-specific mutational biases are strong*

Up to now, the strand-specific mutational bias has been found to be the most important factor that affects codon usage in genomes of 10 bacteria. Names of the 10 bacteria are *B. aphidicola*, *B. burgdorferi*, *B. floridanus*, *B. henselae*, *B. quintana*, *C. muridarum*, *C. trachomatis*, *T. pallidum*, *T. whipplei* and *L. intracellularis*, which is found in this work. Investigation on the common genomic characters of these bacteria may be useful. Several characters are analyzed in the following sections.

First, chromosomes of the 10 bacteria are all shorter than 2000 kb. According to statistics on fully sequenced genomes, bacteria vary from 160 to >10 000 kb in their chromosomal lengths. However, these species are all small bacteria based on their chromosomal length, although some of these bacteria are not endosymbiont. Hence, we hypothesize that the short length of chromosome is a necessary condition to generate strong enough strand-specific mutational bias. Perhaps, in bacteria with larger chromosome, the mutation pressure is hard to prevail translational selection. Alternatively, among genomes that have suffered reductive evolution, the repair mechanism of replication may be inefficient.

Secondly, all of the 10 bacteria have medium or low genomic G + C content. Among these species, *B. aphidicola* has the lowest G + C content as 26%, whereas *T. pallidum* has the highest ones as 52%. Perhaps, the environment of high G + C contents is adverse to the generation of strong strand mutation biases. Future experimental works are needed to clarify the relationship between replication mechanism and genomic GC content or genome size.

Thirdly, the strong mutation bias may be associated with presence or absence of certain genes involved in chromosome replication. As suggested by Klasson and Andersson,[25] the strong strand-specific mutational bias in endosymbiont genomes coincides with the absence of genes for replication restart pathways. They performed a comparative analysis of 20 γ-proteobacterial genomes and found that endosymbiont bacteria lacking *recA* and other genes involved in replication restart processes, such as *priA*, displayed the strongest strand bias.[25] Driven by this viewpoint, we investigate the presence and absence of replication restart-related genes in the 10 genomes, in which the strand mutation biases are strong enough to generate distinct codon usages. The analysis is involved with genes *mutH*, *priA*, *topA*, *dnaT*, *fis* and *recA*, which are all initiation and re-initiation associated genes.[25] Consequently, all of these genes are found to be absent in *B. floridanus* and five absent in *B. aphidicola*, with *mutH* as an exception. For the other eight bacteria, *priA*, *topA* and *recA* are present and the other three genes are absent. In a word, *dnaT* and *fis* are absent in all of the 10 genomes, whereas both the genes exist in *E. coli* and other γ-proteobacteria, which have not strong mutation biases. Klasson and Andersson hypothesized that cytosine deaminations accumulate during single-strand exposure at stalled replication forks and the extent of strand bias may depend on the time spent repairing such lesions. Inefficient restart mechanisms result in the long time of the replication forks being arrested and hereby lead to high DNA strand asymmetry.[25] As a common character of the genomes mentioned here, we believe that the absence of replication restart involved genes is very likely to appear in the other genomes, found in the future, with strong strand mutational bias.

Finally, the strong mutation bias may reflect as the strong cumulative excess of G over C plus T over A along the chromosome. In our previously published work,[35] the Z curve was used to compare the chromosome sequences between genomes with or without strand-specific codon usage. The y-component of the Z curve represents the plus of cumulative excess of G over C and T over A. An index is defined as the changing rate of y-component per unit base and denoted by symbol k. In fact, k equals to $(\mathrm{G}-\mathrm{C}+\mathrm{T}-\mathrm{A})/(\mathrm{G}+\mathrm{C}+\mathrm{T}+\mathrm{A})$, where A, C, G and T denote the total number of the corresponding base appearing in the half of chromosome (from replication origin to terminus). After calculating k values for the 10 bacteria mentioned above, it is found that this values for these species are all larger than 0.035 and some even larger than 0.1. However, the value of k for *E. coli* K-12 is less than 0.02. Therefore, the k value that represents the changing rate of cumulative excess of keto $(\mathrm{G}+\mathrm{T})$ over amino $(\mathrm{A}+\mathrm{C})$ per unit base could be a good measurement of magnitude of strand composition biases and even strand mutation bias.

4.5. Conclusion

Complex factors are found to be responsible for variation of codon usage in *L. intracellularis* chromosome. All of these factors could be interpreted by the paradigm 'mutational bias-translational selection'. When analyzing genes in the largest plasmid of this bacterium, for the first time it is found that the strand-specific mutational biases are responsible for the primary variation of synonymous codon usages in plasmid. Genes, particularly highly expressed genes of this plasmid, are mainly located on the leading strands and this is supposed to be the effects exerted by replicational–transcriptional selection. These facts suggest that this plasmid may adopt the similar replication mechanism as the chromosome in *L. intracellularis*. Finally, common genomic characters are found among *L. intracellularis* and other bacteria in whose genomes the strand-specific mutational biases are the most important source of variation of codon usage.

Supplementary data: Supplementary data are available online at www.dnaresearch.oxfordjournals.org.

Funding

The present study was supported by Doctoral Fund of Ministry of Education of China (20070614011) and National Natural Science Foundation of China (60801058).

References

1. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pavé, A. 1980, Codon catalog usage and the genome hypothesis, *Nucleic Acids Res.*, **8**, r49–r62.
2. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. 1981, Codon catalog usage is a genome strategy modulated for gene expressivity, *Nucleic Acids Res.*, **9**, r43–r74.
3. Gouy, M. and Gautier, C. 1982, Codon usage in bacteria: correlation with gene expressivity, *Nucleic Acids Res.*, **10**, 7055–7074.
4. Ikemura, T. 1981, 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.*, **151**, 389–409.
5. Dong, H. J., Nilsson, L. and Kurland, C. G. 1996, Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates, *J. Mol. Biol.*, **260**, 649–663.
6. Bennetzen, J. L. and Hall, B. D. 1982, Codon selection in yeast, *J. Biol. Chem.*, **257**, 3026–3031.
7. Andersson, S. G. E. and Sharp, P. M. 1996, Codon usage in the *Mycobacterium tuberculosis* complex, *Microbiology*, **142**, 915–925.
8. Lafay, B., Atherton, J. C. and Sharp, P. M. 2000, Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*, *Microbiology*, **146**, 851–860.
9. Kanaya, S., Yamada, Y., Kudo, Y. and Ikemura, T. 1999, Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific

diversity of codon usage based on multivariate analysis, *Gene*, **238**, 143–155.

10. McInerney, J. O. 1998, Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*, *Proc. Natl Acad. Sci. USA*, **95**, 10698–10703.
11. Romero, H., Zavala, A. and Musto, H. 2000, Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces, *Nucleic Acids Res.*, **28**, 2084–2090.
12. Grocock, R. J. and Sharp, P. M. 2002, Synonymous codon usage in *Pseudomonas aeruginosa* PA01, *Gene*, **289**, 131–139.
13. Garcia-Vallve, S., Romeu, A. and Palau, J. 2000, Horizontal gene transfer in bacterial and archaeal complete genomes, *Genome Res.*, **10**, 1719–1725.
14. Garcia-Vallve, S., Guzman, E., Montero, M. A. and Romeu, A. 2003, HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes, *Nucleic Acids Res.*, **31**, 187–189.
15. Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A. 1991, Evidence for horizontal gene transfer in *Escherichia coli* speciation, *J. Mol. Biol.*, **222**, 851–856.
16. Moszer, I., Rocha, E. P. and Danchin, A. 1999, Codon usage and lateral gene transfer in *Bacillus subtilis*, *Curr. Opin. Microbiol.*, **2**, 524–528.
17. McInerney, J. O. 1997, Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns, *Microb. Comp. Genomics*, **2**, 1–10.
18. McLean, M. J., Wolfe, K. H. and Devine, K. M. 1998, Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes, *J. Mol. Evol.*, **47**, 691–696.
19. Frank, A. C. and Lobry, J. R. 1999, Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms, *Gene*, **238**, 65–77.
20. Mrázek, J. and Karlin, S. 1998, Strand compositional asymmetry in bacterial and large viral genomes, *Proc. Natl Acad. Sci. USA*, **95**, 3720–3725.
21. Mackiewicz, P., Gierlik, A., Kowalczuk, M., Dudek, M. R. and Cebrat, S. 1999, How does replication-associated mutational pressure influence amino acid composition of proteins?, *Genome Res.*, **9**, 409–416.
22. Lobry, J. R. and Sueoka, N. 2002, Asymmetric directional mutation pressures in bacteria, *Genome Biol.*, **3**, RESEARCH0058.
23. Rocha, E. P. and Danchin, A. 2003, Essentiality, not expressiveness, drives gene-strand bias in bacteria, *Nat Genet.*, **34**, 377–378.
24. Price, M. N., Alm, E. J. and Arkin, A. P. 2005, Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication, *Nucleic Acids Res.*, **33**, 3224–3234.
25. Klasson, L. and Andersson, S. G. 2006, Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways, *Mol. Biol. Evol.*, **23**, 1031–1039.
26. Rocha, E. P., Touchon, M. and Feil, E. J. 2006, Similar compositional biases are caused by very different mutational effects, *Genome Res.*, **16**, 1537–1547.
27. Necsulea, A. and Lobry, J. R. 2007, A new method for assessing the effect of replication on DNA base composition asymmetry, *Mol. Biol. Evol.*, **24**, 2169–2179.
28. Lobry, J. R. 1996, Origin of replication of *Mycoplasma genitalium*, *Science*, **272**, 745–746.
29. Lobry, J. R. 1996, Asymmetric substitution patterns in the two DNA strands of bacteria, *Mol. Biol. Evol.*, **13**, 660–665.
30. Lafay, B., Lloyd, A. T., McLean, M. J., Devine, K. M., Sharp, P. M. and Wolfe, K. H. 1999, Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases, *Nucleic Acids Res.*, **27**, 1642–1649.
31. Rispe, C., Delmotte, F., van Ham, R. C. and Moya, A. 2004, Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids, *Genome Res.*, **14**, 44–53.
32. Banerjee, T., Basak, S., Gupta, S. K. and Ghosh, T. C. 2004, Evolutionary forces in shaping the codon and amino acid usages in *Blochmannia floridanus*, *J. Biomol. Struct. Dyn.*, **22**, 13–23.
33. Das, S., Paul, S., Chatterjee, S. and Dutta, C. 2005, Codon and amino acid usage in two major human pathogens of genus *Bartonella*–optimization between replicational–transcriptional selection, translational control and cost minimization, *DNA Res.*, **12**, 91–102.
34. Das, S., Paul, S. and Dutta, C. 2006, Evolutionary constraints on codon and amino acid usage in two strains of human pathogenic actinobacteria *Tropheryma whipplei*, *J. Mol. Evol.*, **62**, 645–658.
35. Guo, F. B. and Yu, X. J. 2007, Separate base usages of genes located on the leading and lagging strands in *Chlamydia muridarum* revealed by the Z curve method, *BMC Genomics*, **8**, 366.
36. Mølbak, L., Johnsen, K., Boye, M., Jensen, T. K., Johansen, M., Møller, K. and Leser, T. D. 2008, The microbiota of pigs influenced by diet texture and severity of *Lawsonia intracellularis* infection, *Vet Microbiol.*, **128**, 96–107.
37. Peden, J. F. 1999, Analysis of codon usage, Ph.D. Thesis, University of Nottingham.
38. Wright, F. 1990, The ‘effective number of codons’ used in a gene, *Gene*, **87**, 23–29.
39. Perrière, G. and Thioulouse, J. 2002, Use and misuse of correspondence analysis in codon usage studies, *Nucleic Acids Res.*, **30**, 4548–4555.
40. Dillon, W. R. and Goldstein, M. 1984, *Multivariate Analysis, Method and Application*, Willey Press: New York, USA.
41. Romero, H., Zavala, A. and Musto, H. 2000, Compositional pressure and translational selection determine codon usage in the extremely GC-poor unicellular eukaryote *Entamoeba histolytica*, *Gene*, **242**, 307–311.
42. Gao, F. and Zhang, C. T. 2007, DoriC: a database of oriC regions in bacterial genomes, *Bioinformatics*, **23**, 1866–1867.
43. Fu, Q. S., Li, F. and Chen, L. L. 2005, Gene expression analysis of six GC-rich Gram-negative phytopathogens, *Biochem. Biophys. Res. Commun.*, **332**, 380–387.
44. Gupta, S. K., Bhattacharyya, T. K. and Ghosh, T. C. 2004, Synonymous codon usage in *Lactococcus lactis*:

mutational bias versus translational selection, *J. Biomol. Struct. Dyn.*, **21**, 527–536.

45. Francino, M. P., Chao, L., Riley, M. A. and Ochman, H. 1996, Asymmetries generated by transcription-coupled repair in enterobacterial genes, *Science*, **272**, 107–109.
46. Beletskii, A. and Bhagwat, A. S. 1996, Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*, *Proc. Natl Acad. Sci. USA*, **93**, 13919–13924.
47. Furusawa, M. and Doi, H. 1998, Asymmetrical DNA replication promotes evolution: disparity theory of evolution, *Genetica*, **102–103**, 333–347.
48. French, S. 1992, Consequences of replication fork movement through transcription units in vivo, *Science*, **258**, 1362–1365.
49. Picardeau, M., Lobry, J. R. and Hinnebusch, B. J. 2000, Analyzing DNA strand compositional asymmetry to identify candidate replication origins of *Borrelia burgdorferi* linear and circular plasmids, *Genome Res.*, **10**, 1594–1604.
50. Barbour, A. G. 1993, Linear DNA of *Borrelia* species and antigenic variation, *Trends Microbiol*, **1**, 236–239.
51. Syvanen, M. 1994, Horizontal gene transfer: evidence and possible consequences, *Annu. Rev. Genet.*, **28**, 237–261.
52. Daubin, V., Lerat, E. and Perrière, G. 2003, The source of laterally transferred genes in bacterial genomes, *Genome Biol.*, **4**, R57.
53. Lawrence, J. G. and Ochman, H. 1997, Amelioration of bacterial genomes: rates of change and exchange, *J. Mol. Evol.*, **44**, 383–397.