RESEARCH ARTICLE

Nucleotide composition bias and codon usage trends of gene populations in *Mycoplasma capricolum* subsp. *capricolum* and *M. agalactiae*

XIAO-XIA MA, YU-PING FENG, JIA-LING BAI, DE-RONG ZHANG, XIN-SHI LIN and ZHONG-REN MA*

College of Life Science and Engineering, Northwest University for Nationalities; Engineering and Technology Research Center for Animal Cell; Key Laboratory of Bioengineering and Biotechnology of State Ethnic Affairs Commission; Lanzhou, Gansu 730030, People's Republic of China

Abstract

Because of the low GC content of the gene population, amino acids of the two mycoplasmas tend to be encoded by synonymous codons with an A or T end. Compared with the codon usage of ovine, *Mycoplasma capricolum* and *M. agalactiae* tend to select optimal codons, which are rare codons in ovine. Due to codon usage pattern caused by genes with key biological functions, the overall codon usage trends represent a certain evolutionary direction in the life cycle of the two mycoplasmas. The overall codon usage trends of a gene population of *M. capricolum* subsp. *capricolum* can be obviously separated from other mycoplasmas, and the overall codon usage trends of *M. agalactiae* are highly similar to those of *M. bovis*. These results partly indicate the independent evolution of the two mycoplasmas without the limits of the host cell's environment. The GC and AT skews estimate nucleotide composition bias at different positions of nucleotide triplets and the protein consideration caused by the nucleotide composition bias at codon positions 1 and 2 largely take part in synonymous codon usage patterns of the two mycoplasmas. The correlation between the codon adaptation index and codon usage variation indicates that the effect of codon usage on gene expression in *M. capricolum* subsp. *capricolum* is opposite to that of *M. agalactiae*, further suggesting independence of the evolutionary process influencing the overall codon usage trends of gene populations of mycoplasmas.

[Ma X.-X., Feng Y.-P., Bai J.-L., Zhang D.-R., Lin X.-S. and Ma Z.-R. 2015 Nucleotide composition bias and codon usage trends of gene populations in *Mycoplasma capricolum* subsp. *capricolum* and *M. agalactiae*. *J. Genet*. **94**, 251–260]

Introduction

Mycoplasma is a genus of bacteria whose members lack a cell wall. These spherical microorganisms belong to the class Mollicutes and are distinguished from other bacteria by their small size, which resulted from regressive evolution from gram-positive ancestors with a massive genome reduction (Fraser *et al*. 1995; Ciccarelli *et al*. 2006; Nouvel *et al*. 2010). The family of mycoplasmas can cause some of the most serious diseases that largely affect the husbandry industry (Rosengarten *et al.* 2001). *M. agalactiae* and *M. capricolum* can cause contagious agalactia, which is a major animal health problem of small ruminants because of its economic significance (Ariza-Miguel *et al.* 2012). *M. agalactiae* is a causative pathogen that results in contagious agalactia, which occurs in sheep and goats, and is marked by pneumonia, mastitis, arthritis and conjunctivitis (Bergonier *et al.* 1997). *M. agalactiae* strains show very little genetic diversity and are highly homogeneous with the type strain PG2 (Solsona *et al.* 1996; Marenda *et al.* 2005; McAuliffe *et al.* 2008). Analysing genetic features of *M. agalactiae* can assist investigations of its evolutionary direction and interactions with its susceptible host.

M. capricolum is a causative agent of acute arthritis, mastitis and relative respiratory diseases, and its DNA structure shows that *M. capricolum* is a derivative of gram-positive bacteria (Tarshis *et al.* 1993). Based on the multiple sequence alignment of the 16S rDNA sequence, significant genetic diversity was identified between *M. capricolum* and *M. agalactiae* (Weisburg *et al.* 1989). With the development of comparative genomics, a single gene has a limited ability to indicate the genetic diversity of a gene population of specific microorganisms. Although, biological functions, such as virulence and physiological properties of a target microorganism depend on a specific subset of genes that is responsible for the species-specific lifestyle and may not be equally distributed within the species (Dobrindt and Hacker 2001;

*For correspondence. E-mail: mzr@xbmu.edu.cn; maxiaoxia956@163.com. Xiao-Xia Ma and Yu-Ping Feng contributed equally to this work.

Keywords. nucleotide composition bias; codon usage trends; evolutionary process; *Mycoplasma agalactiae*; *M. capricolum* subsp. *capricolum*.

Zhou *et al.* 2014), comparative genomics represent a novel pathway to investigate what drives the target microorganism in a certain evolutionary direction (Medini *et al.* 2005). Recently, *M. agalactiae* PG2 (NC_009497) and *M. capricolum* subsp. *capricolum* (NC_007633) genome sequencing projects have been completed. Information about the gene populations of *M. agalactiae* and *M. capricolum* may provide some significant clues to understand the evolutionary dynamic sustaining the unique genetic features of the mycoplasmas. With the increasing number of genes, with gene annotation for the two mycoplasmas, new resources have become available to investigate the factors that influence the codon usage trends of gene populations and to analyse the evolutionary processes that are reflected by the overall codon usage trends.

In nature, 20 canonical amino acids can be encoded by 61 nucleotide triplets (codons), and there are three codons that terminate gene expression. Different codons that translate the same amino acid are known as synonymous codons. In a wide variety of organisms, synonymous codons are selected with different frequencies rather than equal and random frequencies; this phenomenon is termed synonymous codon usage bias (Hershberg and Petrov 2008). Many genetic studies indicated that the third position (synonymous site) of a synonymous codon is subject to weak selection and that, synonymous codon usage bias is sustained by an equilibration between mutation pressure, translation selection and genetic drift (Bulmer 1991). Synonymous codon usage varies between both organisms and among genes within a genome and it arises from differences in GC content, replication strand skew, or gene expression (Suzuki *et al.* 2008). The interaction of these factors may vary among different species depending on their evolutionary process. Here, based on the complete genome information of *M. agalactiae* PG2 and *M. capricolum* subsp. *capricolum*, we carried out comprehensive analyses of synonymous codon usage and nucleotide content to recognize the roles of different factors in the evolutionary process. In addition, studies of synonymous codon usage patterns could better reveal the two species' dynamics relevant to disease control measures and exploring reasons for adaptation to their host and environment.

Materials and methods

Basic information about the M. agalactiae and M. capricolum genomes

To investigate synonymous codon usage patterns of *M. agalactiae* and *M. capricolum*, the genomes of *M. agalactiae* PG2 and *M. capricolum* subsp. *capricolum* were downloaded from the NCBI Entrez Genomes Division website (http://www.ncbi.nlm.nih.gov) (table 1). Using the target mycoplasma genome sequences, the location information of each gene on the leading or lagging strand was processed; this information provided a resource for estimating the role of the strand-specific mutational bias in the formation of synonymous codon usage for the two mycoplasmas. In addition, to estimate the similarity or deviation between the two mycoplasmas and ovine animals (the susceptible host), the codon usage frequencies of ovine animals were obtained from the codon usage database (Nakamura *et al.* 2000), and the synonymous codon usage was calculated.

Nucleotide composition statistics in two mycoplasmas

In many prokaryotic genomes and some archaeal genomes, asymmetry exists between the nucleotide compositions of the leading and lagging strands (Lobry 1996; Kunst *et al.* 1997). Specifically, guanine (G) and thymine (T) tend to be selected in the leading strand, whereas cytosine (C) and adenine (A) tend to be selected in the lagging strand. This phenomenon is termed GC skew or AT skew (Blattner *et al.* 1997). Here, the GC skew, namely, the ratio of $(G - C)/(G + C)$ was calculated across the target genome as the sum of a series of sliding windows of a specified length (1000 nucleotides) to predict origins and termini of replication. In addition, a series of GC skew and AT skew data was calculated for codon positions 1, 2 and 3, to further investigate the effects of nucleotide composition bias at different nucleotide positions on the overall codon usage trends of a gene population.

Calculating the relative synonymous codon usage values

The relative synonymous codon usage (RSCU) values for each gene in the gene population of the targeted mycoplasmas were calculated by a previously reported formula (Sharp and Li 1986). To effectively identify the usage bias of the 59 synonymous codons, we applied a standard from the previous reports (Wong *et al.* 2010; Zhou *et al.* 2013a). Correspondence analysis (COA) is a multivariate statistical method. Here, the COA method was introduced to categorize the overall codon usage trend derived from synonymous codon usage patterns of a gene population from the target organisms and analyse the similarity or deviation of synonymous codon usage pattern of genes in the leading and lagging strands. This was carried out by the program CodonW1.3 (devised by John Peden), which is available at http://sourceforge.net/projects/codonw/. In addition, principal component analysis (PCA) was applied to compress the multidimensional information into a two-dimensional or three-dimensional map; this supplies a more convenient method to visualize genetic features of the target gene family with respect to synonymous codon usage (Gustafsson *et al.* 2004). This method was introduced in this study to visually analyse the overall codon usage trends of a gene population in the target mycoplasmas and investigate codon usage trends of the specific genes (such as 30S, 50S and lipoprotein genes).

Other indices concerning codon usage

To estimate the effect of GC content at the third nucleotide position of a codon (GC3s%) on synonymous codon usage, the GC3s% of each gene was calculated. GC3s% is regarded as a useful link to other codon usage indices (e.g. the effective number of codons (ENC) and codon adaptation index) to represent the synonymous codon usage from different

Table 1. The information of genomes of the two target mycoplasmas.

	Accession no.	30S protein	50S protein	Lipoproteins
M. agalactiae PG2	NC_009497	20	31	41
M. capricolum subsp. *capricolum*	NC_007633	19	31	61
M. bovis	NC_014760	–	–	–
M. bovis	NC_018077	–	–	–
M. bovis	NC_015725	–	–	–

perspectives (Ermolaeva 2001). The ENC values range from 20 (when only one synonymous codon is chosen by the corresponding amino acid) to 61 (when all synonymous codons are used equally), and lower the ENC value for one gene is stronger the overall codon usage bias for this gene is (Wright 1990). The codon adaptation index (CAI) is also a useful tool for estimating codon usage bias. In microorganisms, poorly expressed coding sequences in the genome have a relatively low codon usage bias; in contrast, highly expressed coding sequences have the highest codon usage bias (Sharp and Li 1987). In addition, to better understand the role of nucleotide composition bias of a gene population in overall codon usage trends, GC and AT skews were calculated for different nucleotide positions of codons following the formula from a previous report (Lobry 1996).

Codon usage bias in the translation elongation region of a gene population

Based on the definition of the translation elongation region of gene (Tuller *et al.* 2010), we analysed the translation elongation region of genes of *M. agalactiae* and *M. capricolum* as the coding sequence, which is composed of 30 codon positions, ranging from the translation start codon to the 30th codon position. The first 30 codons of each gene were obtained manually and were named cod_1, cod_2, cod_3 ..., cod_28, cod_29 and cod_30. Codons with the same name were mixed and recorded in FASTA format; then, the new rearranged sequences were structured and calculated using the ENC formula to represent the codon distribution at different positions.

Overall codon usage pattern of a gene population of mycoplasma species

Since *M. bovis* infects ruminants, induces similar symptoms (contagious agalactia), and is closely related to *M. agalactiae* (Pfutzner and Sachse 1996), three complete genomes of *M. bovis* were also downloaded (table 1) and were used to investigate the genetic diversity of a gene population in different mycoplasma species and estimate the similarity of overall codon usage trends of a gene population in *M. agalactiae* and *M. bovis*.

Statistical tests

One-way analysis of variance was used to estimate the differentiation of nucleotide content of genes on the leading and lagging strands of the genome of the given mycoplasma. Correlation analysis in this study employed Spearman's rank correlation (with the following levels of significance: $P < 0.05$, $P < 0.01$ and $P < 0.001$) and was performed by SPSS ver. 11.5 for Windows.

Results

Nucleotide compositions of a gene population of the two mycoplasmas

The AT content of a gene population of the two targeted mycoplasmas is much higher than the GC content (figure 1; table 1 in electronic supplementary material at http://www.ias.ernet.in/jgenet/). This feature was common in members of the Mycoplasmataceae family (Razin 1985; Muto *et al.* 1987; Citti and Blanchard 2013). A comparison of nucleotide compositions of the two mycoplasmas revealed that the average G and C contents of a gene population of *M. capricolum* subsp. *capricolum* are lower than those of *M. agalactiae* (table 1 in electronic supplementary material). Several previous reports have addressed the issue of base composition bias in the leading and lagging strands in bacterial genomes

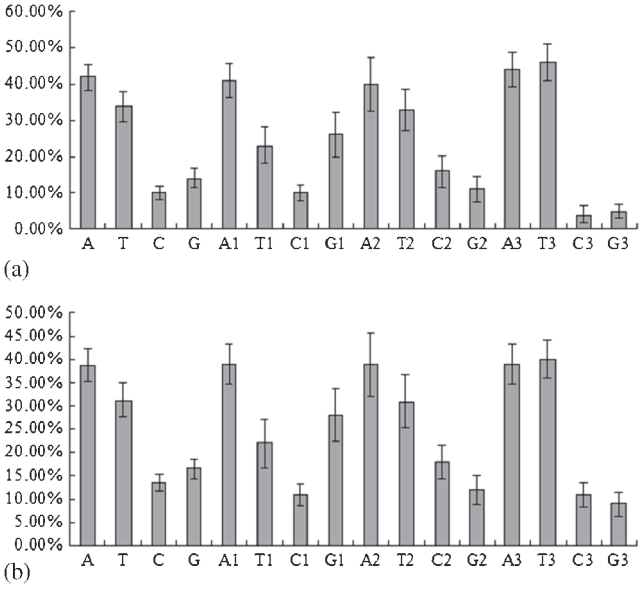

Figure 1. Variations of the nucleotide composition of genes of the two mycoplasma species. (a) *M. capricolum* subsp. *capricolum*; (b) *M. agalactiae*.

and indicated that base composition bias exists in the DNA strands of mycoplasmas (Kunst *et al.* 1997; McLean *et al.* 1998; Mrazek and Karlin 1998; Nayak 2013). Based on the suggestive information, the nucleotide composition of a gene population was analysed using the specific nucleotide composition at different nucleotide positions of codons in the leading and lagging strands of mycoplasmas. In *M. capricolum* subsp. *capricolum*, the base compositions of the whole coding sequence and those at the first, second and third codon positions have little strand-specific mutation bias. In contrast, in *M. agalactiae*, the G, T and G compositions at the first codon position, and the nucleotide composition at the second codon position have a strand-specific bias ($P < 0.01$). It is interesting that the nucleotide composition at the third codon position do not show the strand-specific bias of the two mycoplasmas. These results indicate that the mutation pressure of the nucleotide composition at codon position 3 takes part in the formation of synonymous codon usage, and is free from the effect of the strand-specific bias of codon usage, even though the variation of the base composition of the gene population follows the fluctuation of the base composition of the genome.

To investigate the effect of replication related to mutation pressure caused by the base composition, the normal and cumulative skew of two selectable nucleotides (G and C) for a given genome were calculated by the GC skew formula. The normal GC skew of the two mycoplasmas indicated that even though the GC content is poor in a given genome, the distribution of the G/C content in different parts of a given genome is generally balanced. Based on the V-/inverted V-shape of bidirectional replication between a singular origin and terminus (Nayak 2013), the cumulative GC skew suggested that there is a singular origin and terminus of replication for a given mycoplasma (figures 1 and 2 in electronic supplementary material).

Comparison of synonymous codon usage between the two mycoplasmas and their host

Because ovine animals are susceptible host for the two mycoplasmas, a comparison of the synonymous codon usage between the pathogen and host can be used to investigate the reaction of the targeted pathogen to its host's environment. Interestingly, synonymous codons, which are strongly selected in the ovine genomes, are generally weakly selected by the two mycoplasmas, and this phenomenon is true for all amino acid families (figure 2; table 2 in electronic supplementary material). According to the first major variation (f_1') in 59 synonymous codon usage between the leading and lagging strands, differences in the synonymous codon usage patterns of DNA strands were not found in the two mycoplasmas (figures 3 and 4 in electronic supplementary material). Comparing the synonymous codon usage patterns in the same amino acid family revealed that the two mycoplasmas tend to select synonymous codons with an A/T end rather than those with a G/C end (table 2 in electronic supplementary material), suggesting that nucleotide composition plays a key role in the selection of synonymous codons. Estimating the underrepresented and overrepresented codons revealed that *M. capricolum* subsp. *capricolum* uses TTT (Phe); TTA (Leu); ATT (Ile); GTT (Val); TCA and AGT (Ser); CCA (Pro); ACT (Thr); GCT (Ala); TAT (Tyr); CAT (His); CAA (Gln); AAT (Asn); AAA (Lys); GAT (Glu); GAA (Asp); AGA (Arg); and GGT and GGA (Gly) as the overrepresented codons and TTC (Phe); TTG, CTT, CTC and CTG (Leu); ATC (Ile); GTC and GTG (Val); TCC, TCG and AGC (Ser); CCC and CCG (Pro); ACC and ACG (Thr); GCC and GCG (Ala); TAC (Tyr); CAC (His); CAG (Gln); AAC (Asn); AAG (Lys); GAC (Asp); GAG (Glu); TGC (Cys); CGC, CGA, CGG and AGG (Arg); and GGC and GGG (Gly) as the underrepresented codons (table 2 in electronic supplementary material). *M. agalactiae* selects TTT

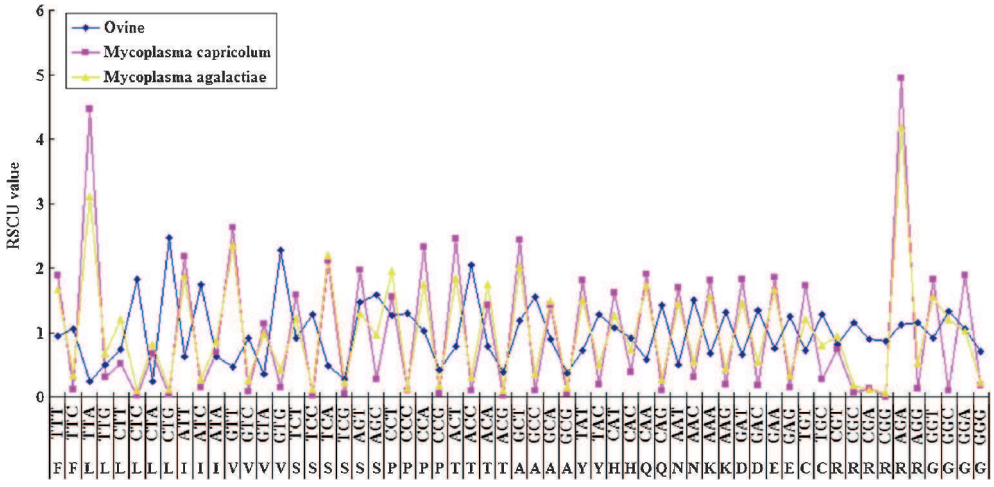

Figure 2. Comparison of RSCU data of 59 synonymous codon among *M. capricolum* subsp. *capricolum*, *M. agalactiae* and ovine. Blue colour represented ovine; pink colour represented *M. capricolum* subsp. *capricolum*; yellow colour represented *M. agalactiae*.

(Phe); TTA (Leu); ATT (Ile); GTT (Val); TCA (Ser); CCT and CCA (Pro); ACT and ACA (Thr); GCT (Ala); CAA (Qln); GAA (Glu); and AGA (Arg) as the overrepresented codons and TTC (Phe); CTC and CTG (Leu); ATC, GTC and GTG (Val); TCC and TCG (Ser); CCC and CCG (Pro); ACC and ACG (Thr); GCC and GCG (Ala); TAC (Tyr); CAG (Gln); AAC (Asn); AAG (Lys); GAC (Asp); GAG (Glu); CGC, CGA, CGG and AGG (Arg); and GGG (Gly) as the underrepresented codons (table 2 in electronic supplementary material). With respect to synonymous codon usage patterns, most synonymous codon usage has a general usage bias in both species, and synonymous codons containing the dinucleotide CpG tend not to be selected. The underrepresentation of CpG or TpA dinucleotides has been reported in many organisms (Karlin and Burge 1995; Cooper and Youssoufian 1988; DeAmicis and Marchetti 2000); however, only CpG was underrepresented in the two mycoplasmas.

Overall codon usage trends of a gene population in the two mycoplasmas

To better understand the overall codon usage trends of a gene population in the two mycoplasmas, corresponding analysis (COA) of the gene population of the given genomes was performed. The plots for the overall codon usage trends of the gene population were drawn using the first major variation (f_1'= 12.43%) and the second major variation (f_2'= 6.77%) of COA on the plane. This showed that the overall codon usage trends of the gene population of *M. agalactiae* are generally separated from those of *M. capricolum* subsp. *capricolum* (figure 3). This phenomenon suggests that even though the synonymous codon usage patterns of *M. capricolum* subsp. *capricolum* and *M. agalactiae* are highly similar, the overall codon usage trends of a gene population of a given species can represent its specific genetic diversity during its evolution. The effects of the strand-specific mutational bias on the overall codon usage trends of the gene population were estimated by COA of the leading and lagging strands, and the large-scale overlay plots of the leading and lagging strands did not reflect a strand-specific mutational bias that influenced the codon usage trends of the gene population of the two mycoplasmas (figure 4).

To better understand the GC content and GC3s% that influenced the overall codon usage trends of a gene population, the overall codon usage trends of a gene population of *M. capricolum* subsp. *capricolum* and *M. agalactiae* were analysed by PCA. The overall codon usage trends of *M. capricolum* subsp. *capricolum* represent the first major variation (f_1'= 6.75%) and the second major variation (f_2'= 5.82%); the overall codon usage trends for *M. agalactiae* show the first major variation (f_1'= 9.98%) and the second major variation (f_2'= 4.99%). Depending on the bivariate correlation analysis, the effects of the GC content and GC3s% on the overall codon usage trends of the gene population were estimated. The GC content and the two major variations of the overall codon usage trends in the two mycoplasmas were correlated.

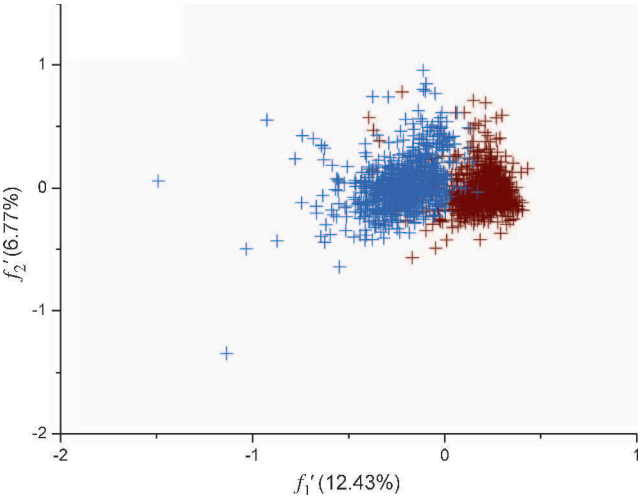

Figure 3. Plot of the first and second major axes generated by PCA. The red points represent the overall codon usage trends of a gene population of *M. capricolum* subsp. *capricolum*. The blue points represent the overall usage trends of a gene population of *M. agalactiae*.

However, unlike the correlation between the GC3s% and the two major variations of the overall codon usage trends of *M. agalactiae*, the GC3s% and the second major variation of the overall codon usage trends of *M. capricolum* subsp. *capricolum* were not correlated (table 3 in electronic supplementary material). The correlation between the GC content and the two major variations differ for *M. capricolum* subsp. *capricolum* and *M. agalactiae*, while the correlations between the GC3s% and the first major variation are both

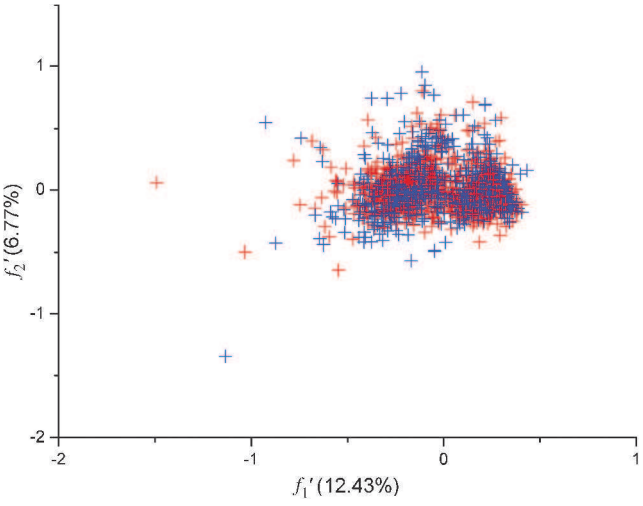

Figure 4. Overall codon usage trends of genes of the leading and lagging strands of two mycoplasmas. The red points represent the overall codon usage trends of genes located on the leading strands of the two mycoplasmas, and the blue points represent the overall codon usage trends of genes located on the lagging strands of the two species. Note: the left cloud of plots represents a gene population of *M. capricolum* subsp. *capricolum*; the right cloud of plots represents a gene population of *M. agalactiae*.

positive for these mycoplasmas (table 3 in electronic supplementary material). These results suggest that the mutation pressure of the GC3s% plays a more important role in the overall codon usage trends of a gene population in the two mycoplasmas than that of protein considerations caused by the GC content.

The effective number of codons (ENC) and codon adaptation index (CAI) were applied to estimate the codon usage bias from different perspectives. Interestingly, significant positive correlation coefficients were found between the ENC and the two major variations (f_1' and f_2') of the two mycoplasmas. However, the correlation coefficients between the CAI and the first major variation for codon usage are in opposite directions (table 3 in electronic supplementary material). In addition, a significant negative correlation (r value is -0.375, $P< 0.001$) was found between the CAI and ENC of the gene population of *M. agalactiae*, while no correlation was found in *M. capricolum* subsp. *capricolum*. These results suggest that even though the overall codon usage trends of the gene population of the two mycoplasma species are strongly influenced by nucleotide composition, the effect of codon usage bias on synonymous codon usage is stronger for *M. agalactiae* than for *M. capricolum* subsp. *capricolum*.

The codon usage trends for genes that play important roles in metabolic processes were also analysed by PCA. The codon usage trends of genes encoding ribosomal 30S and 50S subunits have a similar evolutionary direction in the two mycoplasmas (figures 5 and 6). The codon usage trends of lipoprotein genes for *M. capricolum* subsp. *capricolum* represent a more obvious cloud than those of *M. agalactiae* (figures 7 and 8). These results imply that even though

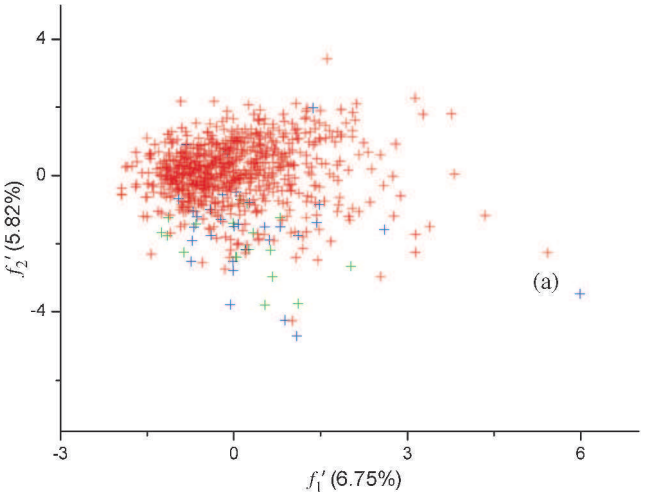

Figure 5. Overall codon usage trends of the 30S/50S ribosome subunits and other genes of *M. capricolum* subsp. *capricolum*. The blue points represent the overall codon usage trends of genes encoding the 30S ribosome subunit, the green points represent the overall codon usage trends of genes encoding the 50S ribosome subunit, and the red points represent the overall codon usage trends of other genes.

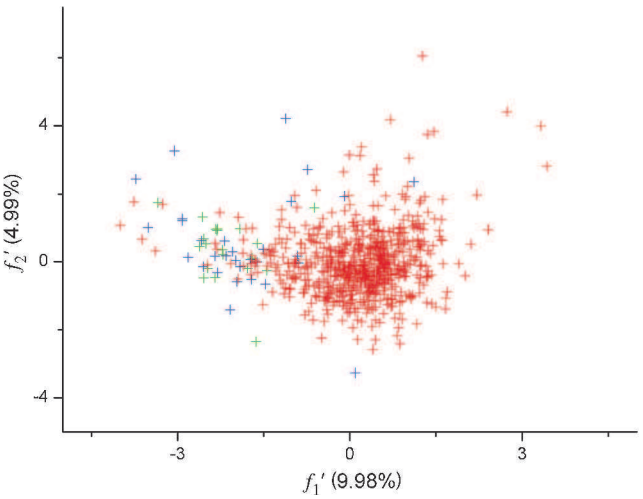

Figure 6. Overall codon usage trends of the 30S/50S ribosome subunits and other genes of *M. agalactiae*. The blue points represent the overall codon usage trends of genes encoding the 30S ribosome subunit, the green points represent the overall codon usage trends of genes encoding the 50S ribosome subunit and the red points represent the overall codon usage trends of other genes.

the codon usage trends of a gene population seem to be disorderly and unsystematic in the evolutionary process, codon usage trends of specific genes representing similar functions can be shaped in a certain evolutionary direction, to some degree.

Base composition bias for synonymous codon usage

To better understand the roles of base composition bias at different codon positions in the formation of synonymous codon usage, a series of correlations between composition bias and the major variation of codon usage was performed. The

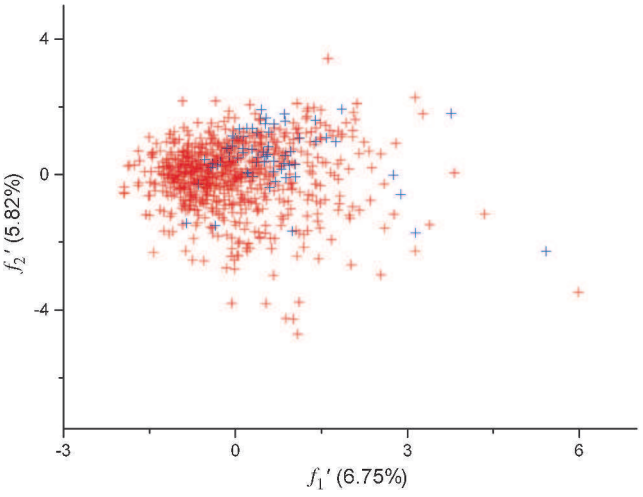

Figure 7. Comparison of overall codon usage trends between genes encoding lipoproteins and other genes of *M. capricolum* subsp. *capricolum*. The blue points represent the overall codon usage trends of genes encoding lipoproteins and the red points represent the overall codon usage trends of other genes.

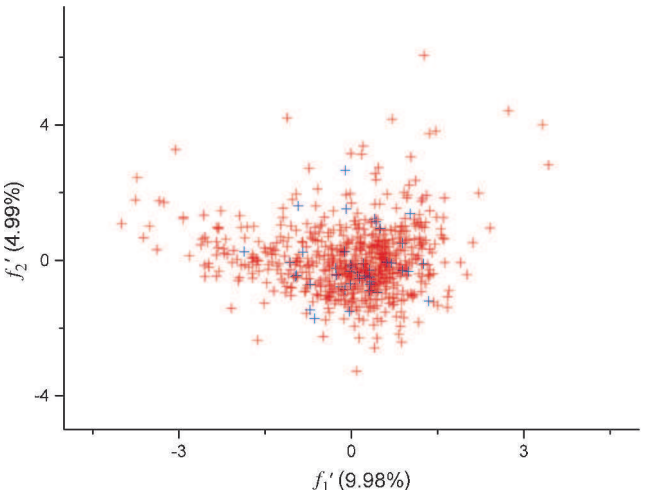

Figure 8. Comparison of overall codon usage trends between genes encoding lipoproteins and other genes of *M. agalactiae*. The blue points represent the overall codon usage trends of genes encoding lipoproteins and the red points represent the overall codon usage trends of other genes.

correlations between GC and AT skews and the two major variations obviously indicate that the AT composition bias plays a more important role in the synonymous codon usage than the GC composition bias (table 4 in electronic supplementary material). Based on correlations between GC and AT skews at the first and second codon positions and the two major variations, the effect of protein consideration with respect to the GC composition is stronger than that of the AT composition (table 4 in electronic supplementary material). AT composition bias at the third codon position mainly takes part in the formation of synonymous codon usage for *M. capricolum* subsp. *capricolum*, while GC composition bias at the third codon position mainly takes part in the formation of synonymous codon usage for *M. agalactiae* (table 4 in electronic supplementary material). These results imply that (i) the formation of synonymous codon usage for the two mycoplasmas is linked to multiple factors including mutation pressure from nucleotide composition bias at the third codon position and protein consideration caused by nucleotide composition bias at the first and second codon positions. Also, (ii) although the two genomes are generally AT-rich and GC-poor, the role of GC composition bias should not be neglected in the formation of synonymous codon usage.

Comparative preference of the translation elongation region of a gene population in the two mycoplasmas

The gradient of codon usage preference of the translation elongation region (the first 30 codons of an open reading frame (ORF)) of the gene population of *M. capricolum* is generally stronger than that of *M. agalactiae* (figure 9). This suggests that the effect of translation selection plays a more important role in the formation of synonymous codon usage to sustain the normal gene expression of *M. capricolum* subsp. *capricolum* than *M. agalactiae*.

Comparative codon usage among M. agalactiae, M. capricolum and M. bovis

The ENC and CAI were applied again to estimate the codon usage bias of the two mycoplasmas and three strains of *M. bovis*. The relationship between the GC3s% and ENC value was drawn and applied to represent the overall codon usage pattern in the genome of a certain target organism. According to the standard of estimating the multiple factors that affect the overall codon usage, the phenomenon of data points that are only scattered on the expected curve indicates that nucleotide compositional constraints are the single factor that affects the codon usage pattern with no other selection factors. In contrast, the phenomenon of data points that are scattered around the expected curve suggests that some other selection factors rather than compositional constraints affect the codon usage pattern. In the plot of the ENC value versus GC3s% of the five strains of mycoplasmas, although most of the data points are below the expected curve, these plots have an obvious tendency to lay towards the GC3s%-poor region (figure 5 in electronic supplementary material), suggesting that nucleotide composition, such as the TA content of the third codon position, plays a major role in the codon usage for the five mycoplasma strains. Further, the correlation coefficient between the CAI and GC3s% of *M. capricolum* subsp. *capricolum* is positive, while that of *M. agalactiae* and three *M. bovis* strains is negative (figure 6 in electronic supplementary material). To further identify the difference in the overall codon usage trends of the gene population of the five strains of mycoplasmas, the COA analysis was performed again. Interestingly, the overall codon usage trends of the gene population of *M. agalactiae* are highly similar to those of the three *M. bovis* strains, while the overall codon

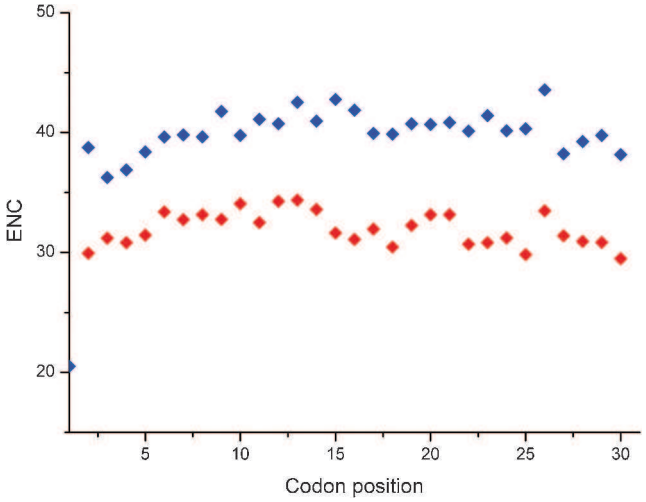

Figure 9. Overall codon usage bias for each codon position in the translation elongation region of genes. It is noted that the first position is the initiation codon for the two species. The red points represent the overall codon usage bias of the translation elongation region of *M. capricolum* subsp. *capricolum*. The blue points represent the overall codon usage bias of the translation elongation region of *M. agalactiae*.

usage trends of the gene population of *M. capricolum* subsp. *capricolum* are obviously different from those of the other mycoplasma strains (figure 7 in electronic supplementary material). These results suggest that the genetic feature of synonymous codon usage patterns of *M. agalactiae* is closely linked to that of *M. bovis*.

Discussion

The overall codon usage trends caused by synonymous codon usage patterns can reflect a distinctive feature of organisms (Ermolaeva 2001; Cutter *et al*. 2006; Lobo *et al*. 2009), particularly bacteria and viruses. Because of the limited cell membrane, genome and metabolic pathways of mycoplasmas, the selfreplication of this microorganism relies on the host cell's machinery and environment. Thus, we analysed the synonymous codon usage patterns of mycoplasmas to better understand the strategies used by mycoplasmas during the course of host adaptation. In the current study, the role of nucleotide composition in shaping synonymous codon usage patterns was estimated through a series of tests. In many previous reports concerning synonymous codon usage of viruses, which only contain the single-nucleotide sequence with a positive or negative sense, the GC3s% which reflects the mutation pressure can be used to estimate the synonymous codon usage. The GC3s% may contribute to the synonymous codon usage, particularly in bacterial genomes with low GC content. Other factors that are related to nucleotide composition need to be considered because nucleotide composition at the first and second codon positions influence the protein consideration and is a part of the genetic code. In theory, the synonymous codon usage of a gene is related to nucleotide changes at all positions of nucleotide triplets. There have been many previous reports about codon usage concerning the role of nucleotide composition at each codon position in single-stranded sequences of viruses (Belalov and Lukashev 2013; Zhang *et al.* 2013; Zhou *et al.* 2013b). However, bacteria have double-stranded DNA, which shows nucleotide composition bias in bacterial genomes (Francino and Ochman 1997; Mrazek and Karlin 1998). GC and AT skews have been accepted for analysing the genetic features between the leading and lagging strands in bacterial genome and therefore for analysing the nucleotide composition bias of a gene population (Lobry 1996; Kerr *et al*. 1997; McLean *et al*. 1998). In this study, GC and AT skews at different positions of nucleotide triplets give further evidence for the factors of protein consideration and mutation pressure influencing the synonymous codon usage in *M. agalactiae* and *M. capricolum* subsp. *capricolum*.

PCA and COA based on relative synonymous codon usage (RSCU) values are effective tools to estimate overall codon usage trends for a gene population and the specific gene family. RSCU values for a gene population of a given mycoplasma species can be mapped to the first and second major variations derived from PCA or COA to represent the relationship between the gene population and existing populations as shown by the validation tests performed in this study. COA is conceptually similar to PCA, but it is applied to categorical data rather than continuous data. Similar to PCA, COA can provide a method of displaying a set of data in two-dimensional or three-dimensional graphical form. Although previous studies have analysed the genetic diversity of different mycoplasma species, these studies were all carried out using some specific sequences that can indicate the genetic diversity of mycoplasmas (Marenda *et al.* 2005; Breton *et al.* 2012). The biological and genetic features of organisms are shaped under the control of a gene population rather than several types of genes. A comparison of gene populations can represent the relationship between different microorganisms during the evolutionary process. PCA of synonymous codon usage was carried out to reveal evolutionary relationships and overall codon usage trends of a gene population in mycoplasmas. Through this analysis, GC3s% was found to play an important role in the differentiation of the overall codon usage trends between *M. agalactiae* and*M. capricolum*. Because of the similar GC3s% between *M. agalactiae* and *M. bovis*, the overall codon usage trends of the analysed gene population between the two species of mycoplasma have a generally similar evolutionary direction. Mycoplasma infections are often associated with the presence of asymptotic carriers that result in the disease to different degree, but mycoplasma species which have similar genetic features may have a similarity in pathogenicity (Baseman and Tully 1997; Brown *et al*. 2001, 2004, 2011). Although mycoplasma species are regarded as having a strict host field, some studies showed the occurrence of mycoplasmas producing diseases outside of their natural hosts (Rosengarten *et al*. 2001; Pitcher and Nicholas 2005; Yuan *et al.* 2009).

In conclusion, a series of comprehensive analyses of synonymous codon usage patterns has supported a basic understanding of mechanisms for codon usage bias in *M. agalactiae* and *M. capricolum*. Nucleotide composition bias and synonymous codon usage patterns provide an additional pathway to investigate the evolutionary direction of the two mycoplasma species, and the information could be helpful in further investigations of evolutionary mechanisms of mycoplasma, cloning and heterologous expression of functionally important proteins.

Acknowledgements

This study was supported by programme for Changjiang Scholars and Innovative Research Team in University (IRT13091), programme for Leading Talent of SEAC and programme for Innovative Research Team of SEAC ([2013]231).

References

Ariza-Miguel J., Rodriguez-Lazaro D. and Hernandez M. 2012 A survey of *Mycoplasma agalactiae* in dairy sheep farms in Spain. *BMC Vet. Res.* **8**, 171.

Baseman J. B. and Tully J. G. 1997 Mycoplasmas: sophisticated, reemerging, and burdened by their notoriety. *Emerg. Infect. Dis.* **3**, 21–32.

Belalov I. S. and Lukashev A. N. 2013 Causes and implications of codon usage bias in RNA viruses. *PLoS One* **8**, e56642.

Bergonier D., Berthelot X. and Poumarat F. 1997 Contagious agalactia of small ruminants: current knowledge concerning epidemiology, diagnosis and control. *Rev. Sci. Tech.* **16**, 848–873.

Blattner F. R., Plunkett G. 3rd, Bloch C. A., Perna N. T., Burland V., Riley M. *et al.* 1997 The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462.

Breton M., Tardy F., Dordet-Frisoni E., Sagne E., Mick V., Renaudin J. *et al.* 2012 Distribution and diversity of mycoplasma plasmids: lessons from cryptic genetic elements. *BMC Microbiol.* **12**, 257.

Brown D. R., Farley J. M., Zacher L. A., Carlton J. M., Clippinger T. L., Tully J. G and Brown M. B. 2001 *Mycoplasma alligatoris* sp. nov., from American alligators. *Int. J. Syst. Evol. Microbiol.* **51**, 419–424.

Brown D. R., Farmerie W. G., May M., Benders G. A., Durkin A. S., Hlavinka K. *et al.* 2011 Genome sequences of *Mycoplasma alligatoris* A21JP2T and *Mycoplasma crocodyli* MP145T. *J. Bacteriol.* **193**, 2892–2893.

Brown D. R., Zacher L. A. and Farmerie W. G. 2004 Spreading factors of *Mycoplasma alligatoris*, a flesh-eating mycoplasma. *J. Bacteriol.* **186**, 3922–3927.

Bulmer M. 1991 The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897–907.

Ciccarelli F. D., Doerks T., von Mering C., Creevey C. J., Snel B. and Bork P. 2006 Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287.

Citti C. and Blanchard A. 2013 Mycoplasmas and their host: emerging and re-emerging minimal pathogens. *Trends Microbiol.* **21**, 196–203.

Cooper D. N. and Youssoufian H. 1988 The CpG dinucleotide and human genetic disease. *Hum. Genet.* **78**, 151–155.

Cutter A. D., Wasmuth J. D and Blaxter M. L. 2006 The evolution of biased codon and amino acid usage in nematode genomes. *Mol. Biol. Evol.* **23**, 2303–2315.

De Amicis F. and Marchetti S. 2000 Intercodon dinucleotides affect codon choice in plant genes. *Nucleic Acids Res.* **28**, 3339–3345.

Dobrindt U. and Hacker J. 2001 Whole genome plasticity in pathogenic bacteria. *Curr. Opin. Microbiol.* **4**, 550–557.

Ermolaeva M. D. 2001 Synonymous codon usage in bacteria. *Curr. Issues Mol. Biol.* **3**, 91–97.

Francino M. P. and Ochman H. 1997 Strand asymmetries in DNA evolution. *Trends Genet.* **13**, 240–245.

Fraser C. M., Gocayne J. D., White O., Adams M. D., Clayton R. A., Fleischmann R. D. *et al.* 1995 The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.

Gustafsson C., Govindarajan S. and Minshull J. 2004 Codon bias and heterologous protein expression. *Trends Biotechnol.* **22**, 346–353.

Hershberg R. and Petrov D. A. 2008 Selection on codon bias. *Annu. Rev. Genet.* **42**, 287–299.

Karlin S. and Burge C. 1995 Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**, 283–290.

Kerr A. R., Peden J. F and Sharp P. M. 1997 Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae*. *Mol. Microbiol.* **25**, 1177–1179.

Kunst F., Ogasawara N., Moszer I., Albertini A. M., Alloni G., Azevedo V. *et al.* 1997 The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.

Lobo F. P., Mota B. E., Pena S. D., Azevedo V., Macedo A. M., Tauch A. *et al.* 2009 Virus–host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. *PLoS One* **4**, e6282.

Lobry J. R. 1996 Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**, 660–665.

Marenda M. S., Sagne E., Poumarat F. and Citti C. 2005 Suppression subtractive hybridization as a basis to assess *Mycoplasma agalactiae* and *Mycoplasma bovis* genomic diversity and species-specific sequences. *Microbiology* **151**, 475–489.

McAuliffe L., Churchward C. P., Lawes J. R., Loria G., Ayling R. D. and Nicholas R. A. 2008 VNTR analysis reveals unexpected genetic diversity within *Mycoplasma agalactiae*, the main causative agent of contagious agalactia. *BMC Microbiol.* **8**, 193.

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.

Medini D., Donati C., Tettelin H., Masignani V. and Rappuoli R. 2005 The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594.

Mrazek J. and Karlin S. 1998 Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA* **95**, 3720–3725.

Muto A., Yamao F. and Osawa S. 1987 The genome of *Mycoplasma capricolum*. *Prog. Nucleic Acid Res. Mol. Biol.* **34**, 29–58.

Nakamura Y., Gojobori T. and Ikemura T. 2000 Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* **28**, 292.

Nayak K. C. 2013 Comparative genome sequence analysis of Sulfolobus acidocaldarius and 9 other isolates of its genus for factors influencing codon and amino acid usage. *Gene* **513**, 163–173.

Nouvel L. X., Sirand-Pugnet P., Marenda M. S., Sagne E., Barbe V., Mangenot S. *et al.* 2010 Comparative genomic and proteomic analyses of two *Mycoplasma agalactiae* strains: clues to the macro- and micro-events that are shaping mycoplasma diversity. *BMC Genomics* **11**, 86.

Pfutzner H. and Sachse K. 1996 *Mycoplasma bovis* as an agent of mastitis, pneumonia, arthritis and genital disorders in cattle. *Rev. Sci. Tech.* **15**, 1477–1494.

Pitcher D. G. and Nicholas R. A. 2005 Mycoplasma host specificity: fact or fiction. *Vet. J.* **170**, 300–306.

Razin S. 1985 Molecular biology and genetics of mycoplasmas (Mollicutes). *Microbiol. Rev.* **49**, 419–455.

Rosengarten R., Citti C., Much P., Spergser J., Droesse M. and Hewicker-Trautwein M. 2001 The changing image of mycoplasmas: from innocent bystanders to emerging and reemerging pathogens in human and animal diseases. *Contrib. Microbiol.* **8**, 166–185.

Sharp P. M. and Li W. H. 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**, 28–38.

Sharp P. M. and Li W. H. 1987 The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295.

Solsona M., Lambert M. and Poumarat F. 1996 Genomic, protein homogeneity and antigenic variability of *Mycoplasma agalactiae*. *Vet. Microbiol.* **50**, 45–58.

Suzuki H., Brown C. J., Forney L. J. and Top E. M. 2008 Comparison of correspondence analysis methods for synonymous codon usage in bacteria. *DNA Res.* **15**, 357–365.

Tarshis M., Salman M. and Rottem S. 1993 Cholesterol is required for the fusion of single unilamellar vesicles with *Mycoplasma capricolum*. *Biophys. J.* **64**, 709–715.

Tuller T., Carmi A., Vestsigian K., Navon S., Dorfan Y., Zaborske J. *et al.* 2010 An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354.

Weisburg W. G., Tully J. G., Rose D. L., Petzel J. P., Oyaizu H., Yang D. *et al.* 1989 A phylogenetic analysis of the mycoplasmas: basis for their classification. *J. Bacteriol.* **171**, 6455–6467.

Wong E. H., Smith D. K., Rabadan R., Peiris M. and Poon L. L. 2010 Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. *BMC Evol. Biol.* **10**, 253.

Wright F. 1990 The effective number of codons used in a gene. *Gene* **87**, 23–29.

Yuan C. L., Liang A. B., Yao C. B., Yang Z. B., Zhu J. G., Cui L. *et al.* 2009 Prevalence of *Mycoplasma suis* (Eperythrozoon suis) infection in swine and swine-farm workers in Shanghai, China. *Am. J. Vet. Res.* **70**, 890–894.

Zhang Z., Dai W. and Dai D. 2013 Synonymous codon usage in TTSuV2: analysis and comparison with TTSuV1. *PLoS One* **8**, e81469.

Zhou J. H., Ding Y. Z., He Y., Chu Y. F., Zhao P., Ma L. Y. *et al.* 2014 The effect of multiple evolutionary selections on synonymous codon usage of genes in the *Mycoplasma bovis* genome. *PLoS One* **9**, e108949.

Zhou J. H., Gao Z. L., Zhang J., Ding Y. Z., Stipkovits L., Szathmary S. *et al.* 2013a The analysis of codon bias of foot-and-mouth disease virus and the adaptation of this virus to the hosts. *Infect. Genet. Evol.* **14**, 105–110.

Zhou J. H., Zhang J., Sun D. J., Ma Q., Chen H. T., Ma L. N. and *et al.* 2013b The distribution of synonymous codon choice in the translation initiation region of dengue virus. *PLoS One* **8**, e77239.

Received 10 October 2014, in revised form 10 December 2014; accepted 12 December 2014
Unedited version published online: 30 December 2014
Final version published online: 10 June 2015