Supplementary data:

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

J. Genet. **94**, 251–260

Table 1. Variations of the nucleotide composition of genes of the two mycoplasma species.

	Nucleotide	Range (%)	Average (%)	Standard deviation (%)
M. capricolum subsp. *capricolum*	A	29.1–54.9	41.9	± 3.66
	T	16.2–48.9	34.0	± 4.20
	C	5.8–20.3	10.1	± 1.95
	G	6.5–29.5	13.9	± 2.73
	A1	25.7–59.0	41.0	± 4.73
	T1	5–40	23.0	± 5.04
	C1	1.6–21.1	10.0	± 2.35
	G1	10.5–53.7	26.0	± 6.17
	A2	15.7–59.7	40.0	± 7.33
	T2	7.0–54.0	33.0	± 5.80
	C2	5.6–46.3	16.0	± 4.40
	G2	0–33.3	11.0	± 3.44
	A3	32.2–63.4	44.0	± 4.65
	T3	26.0–63.0	46.0	± 5.12
	C3	0–21.3	4.0	± 2.31
	G3	1.4–12.8	5.0	± 1.71
M. agalactiae	A	23.0–50.6	38.8	± 3.53
	T	18.7–47.2	31.2	± 3.60
	C	7.4–23.3	13.5	± 1.71
	G	10.2–23.7	16.5	± 2.18
	A1	15.6–54.0	39.0	± 4.27
	T1	8.0–47.0	22.0	± 5.05
	C1	3.4–22.4	11.0	± 2.42
	G1	7.3–45.2	28.0	± 5.70
	A2	17.1–56.5	39.0	± 6.85
	T2	13.0–53.0	31.0	± 5.71
	C2	6.9–34.9	18.0	± 3.69
	G2	2.4–31.2	12.0	± 3.13
	A3	18.6–53.9	39.0	± 4.26
	T3	23.0–54.0	40.0	± 4.03
	C3	3.7–35.7	11.0	± 2.59
	G3	2.1–23.9	9.0	± 2.67

Table 2. Comparative synonymous codon usage patterns of the two mycoplasmas and ovine genomes.

Codon (amino acid)	Ovine*	*M. capricolum*	*M. agalactiae*
TTT(F)	0.94	1.89	1.67
TTC(F)	1.06	0.11	0.33
TTA(L)	0.24	4.46	3.11
TTG(L)	0.49	0.30	0.67
CTT(L)	0.74	0.51	1.20
CTC(L)	1.83	0.02	0.09
CTA(L)	0.24	0.68	0.81
CTG(L)	2.46	0.03	0.11
ATT(I)	0.63	2.17	1.87
ATC(I)	1.74	0.14	0.27
ATA(I)	0.63	0.69	0.86
GTT(V)	0.46	2.63	2.35
GTC(V)	0.91	0.08	0.26
GTA(V)	0.36	1.14	0.98
GTG(V)	2.27	0.14	0.42
TCT(S)	0.91	1.58	1.22
TCC(S)	1.28	0.02	0.12
TCA(S)	0.48	2.11	2.20
TCG(S)	0.28	0.05	0.22
AGT(S)	1.48	1.97	1.29
AGC(S)	1.58	0.27	0.96
CCT(P)	1.26	1.55	1.96
CCC(P)	1.29	0.09	0.13
CCA(P)	1.03	2.32	1.75
CCG(P)	0.42	0.05	0.17
ACT(T)	0.78	2.45	1.84
ACC(T)	2.05	0.10	0.31
ACA(T)	0.78	1.42	1.75
ACG(T)	0.38	0.02	0.10
GCT(A)	1.18	2.43	2.02
GCC(A)	1.55	0.10	0.36
GCA(A)	0.90	1.42	1.49
GCG(A)	0.37	0.05	0.14
TAT(Y)	0.72	1.81	1.50
TAC(Y)	1.28	0.19	0.50
CAT(H)	1.08	1.62	1.27
CAC(H)	0.92	0.38	0.73
CAA(Q)	0.57	1.91	1.73
CAG(Q)	1.43	0.09	0.27
AAT(N)	0.49	1.70	1.45
AAC(N)	1.51	0.30	0.55
AAA(K)	0.68	1.81	1.56
AAG(K)	1.32	0.19	0.44
GAT(D)	0.66	1.83	1.45
GAC(D)	1.34	0.17	0.55
GAA(E)	0.75	1.85	1.67
GAG(E)	1.25	0.15	0.33
TGT(C)	0.72	1.73	1.20
TGC(C)	1.28	0.27	0.80
CGT(R)	0.82	0.73	0.94
CGC(R)	1.15	0.07	0.18
CGA(R)	0.89	0.13	0.12
CGG(R)	0.86	0.00	0.06
AGA(R)	1.12	4.95	4.18
AGG(R)	1.16	0.13	0.53
GGT(G)	0.92	1.83	1.55
GGC(G)	1.33	0.10	1.20
GGA(G)	1.05	1.89	1.03
GGG(G)	0.71	0.18	0.22

*Data are from the previous report (Zhou *et al*. 2013a).

Table 3. Bivariate correlation between the major variations (f_1' and f_2') and codon usage index of *M. capricolum* subsp. *capricolum* and *M. agalactiae*.

		f_1'	f_2'
M. capricolum	GC3s%	$R = 0.774, P < 0.001$	$R = 0.054, P > 0.05$
	GC%	$R = 0.260, P < 0.001$	$R = -0.490, P < 0.001$
	CAI	$R = 0.107, P < 0.01$	$R = -0.475, P < 0.001$
	ENC	$R = 0.617, P < 0.001$	$R = 0.183, P < 0.001$
M. agalactiae	GC3s%	$R = 0.364, P < 0.001$	$R = -0.211, P < 0.001$
	GC%	$R = -0.487, P < 0.001$	$R = 0.472, P < 0.001$
	CAI	$R = -0.671, P < 0.001$	$R = -0.133, P < 0.001$
	ENC	$R = 0.492, P < 0.001$	$R = 0.421, P < 0.001$

CAI, codon adaptation index; ENC, effective number of codons.

Table 4. Nucleotide composition statistics for a gene population in the two mycoplasma species.

	$f_1' = 6.75\%^{\#}$	$f_2' = 5.82\%^{\#}$	$f_1' = 9.98\%^{\$}$	$f_2' = 4.99\%^{\$}$
GC skew	$r = -0.350$***	$r = 0.055$	$r = 0.091$*	$r = -0.335$***
	$P = 8.01\times10^{-25}$	$P = 0.11$	$P = 0.013$	$P = 6.85\times10^{-21}$
AT skew	$r = 0.280$***	$r = -0.084$*	$r = -0.145$***	$r = -0.088$*
	$P = 1.86\times10^{-16}$	$P = 0.016$	$P = 7.22\times10^{-5}$	$P = 0.017$
GC1 skew	$r = -0.207$***	$r = -0.172$***	$r = -0.219$***	$r = -0.397$***
	$P = 2.82\times10^{-9}$	$P = 8.12\times10^{-7}$	$P = 1.59\times10^{-9}$	$P = 2.27\times10^{-29}$
AT1 skew	$r = 0.065$	$r = -0.470$	$r = -0.192$***	$r = -0.129$**
	$P = 0.63$	$P = 0.185$	$P = 1.38\times10^{-7}$	$P = 0.004$
GC2 skew	$r = -0.110$**	$r = 0.079$*	$r = 0.085$*	$r = 0.02$
	$P = 0.002$	$P = 0.024$	$P = 0.021$	$P = 0.58$
AT2 skew	$r = 0.110$**	$r = 0.068$	$r = -0.005$	$r = -0.237$***
	$P = 0.002$	$P = 0.052$	$P = 0.9$	$P = 6.79\times10^{-11}$
GC3 skew	$r = -0.139$***	$r = 0.552$***	$r = 0.767$***	$r = 0.052$
	$P = 7.54\times10^{-5}$	$P = 5.46\times10^{-66}$	$P = 1.65\times10^{-144}$	$P = 0.16$
AT3 skew	$r = 0.516$***	$r = -0.249$***	$r = -0.103$**	$r = 0.32$***
	$P = 2.14\times10^{-56}$	$P = 1.77\times10^{-13}$	$P = 0.005$	$P = 3.59\times10^{-19}$

[#]Represents two major variations for the codon usage of *M. capricolum* subsp. *capricolum*.
[$]Represents two major variations for the codon usage of *M. agalactiae*.
*Means $P < 0.05$; **means $P < 0.01$; ***means $P < 0.001$.

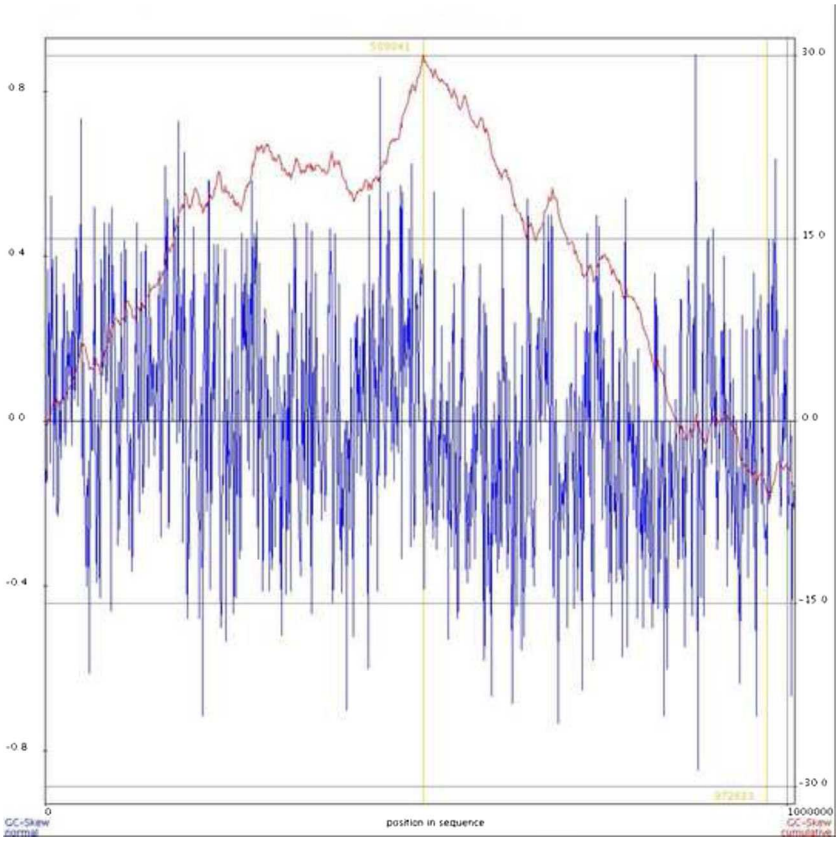

Figure 1. GC skew for the whole genome of *M. capricolum* subsp. *capricolum*. The blue line represents the normal GC skew data which were calculated from 1000-bp windows. The red line represents the cumulative GC skew data which is the sum of the values for all previous windows up to a certain position. The minimum cumulative GC skew gives a predictable origin of replication and the maximum cumulative GC skew gives a predictable terminus of replication.

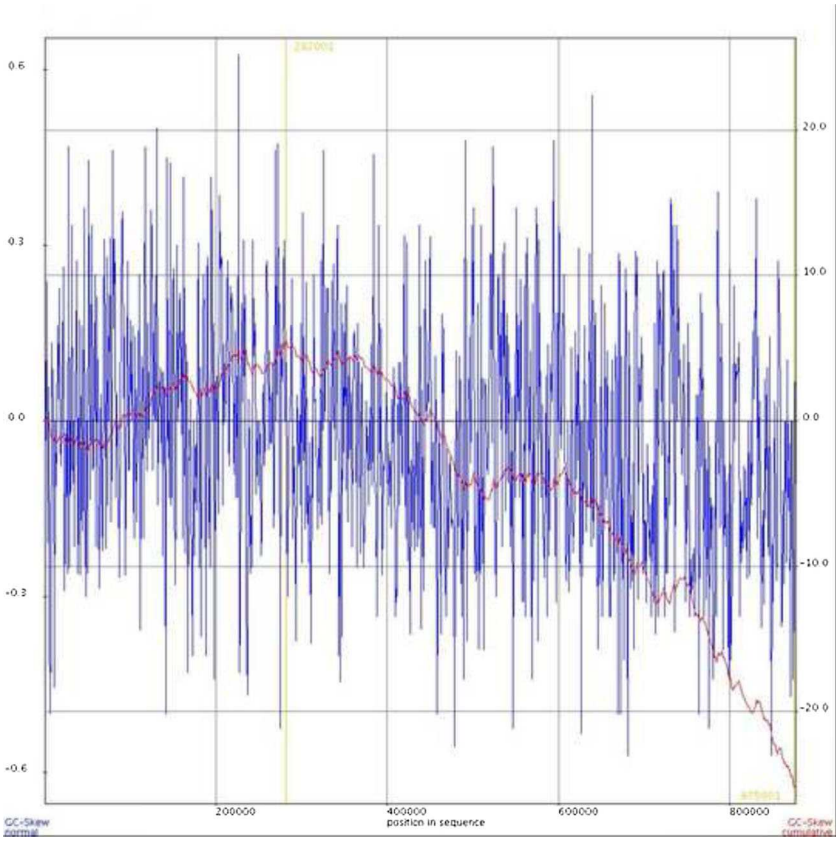

Figure 2. GC skew for the whole genome of *M. agalactiae*. The blue line represents the normal GC skew data which were calculated from 1000-bp windows. The red line represents the cumulative GC skew data which is the sum of the values for all previous windows up to a certain position. The minimum cumulative GC skew gives a predictable origin of replication and the maximum cumulative GC skew gives a predictable terminus of replication.

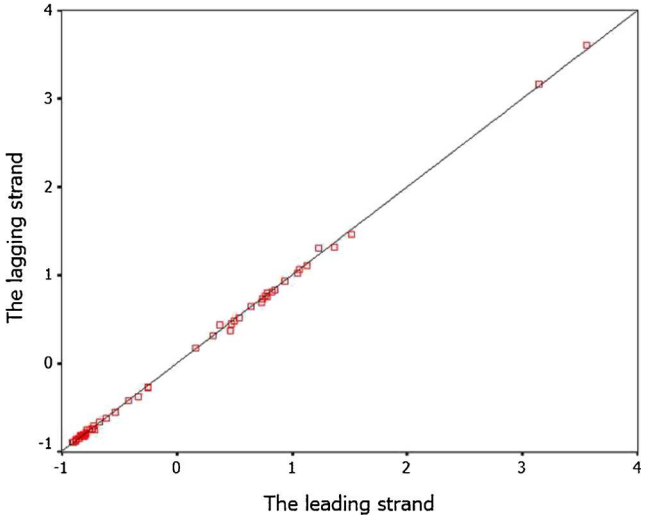

Figure 3. Comparison of 59 synonymous codon usage patterns in between the leading strand and the lagging strand of *M. capricolum* subsp. *capricolum*. The x-axis represents the first major variations for each gene of the leading strand, and the y-axis represents the first major variations for each gene of the lagging strand. The red plots represent the 59 synonymous codon usage patterns. If a red plot is located in this black linear, this phenomenon reflects the same codon usage of a certain synonymous codon of genes located in the leading strand and the lagging strand of *M. capricolum* subsp. *capricolum*.

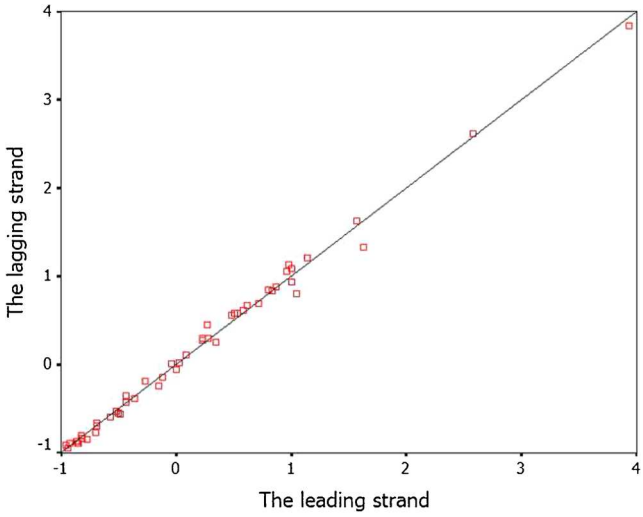

Figure 4. Comparison of 59 synonymous codon usage patterns of the leading and lagging strands of *M. agalactiae*. The *x*-axis represents the first major variation for each gene of the leading strand, and the *y*-axis represents the first major variation for each gene of the lagging strand. The red points represent the 59 synonymous codon usage patterns. If a red point is located on the black line the same codon of a certain synonymous codon is used in genes located on the leading and lagging strands of *M. agalactiae*.

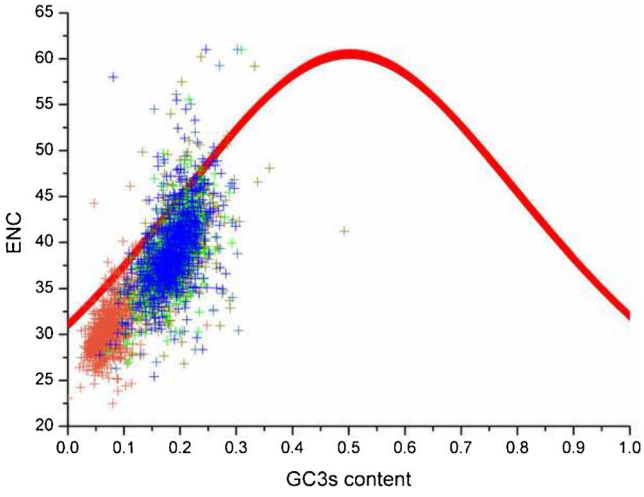

Figure 5. Effective number of codon (ENC) values for genes of five strains of mycoplasmas. The rust red points represent the ENC value against the GC content at the third nucleotide position of a codon (GC3s%) of a gene population of *M. capricolum* subsp. *capricolum*. The green points represent *M. agalactiae*. The blue points indicate the three strains of *M. bovis*. The continuous bright red curve indicates the expected curve between the ENC and GC3s% with random codon usage.

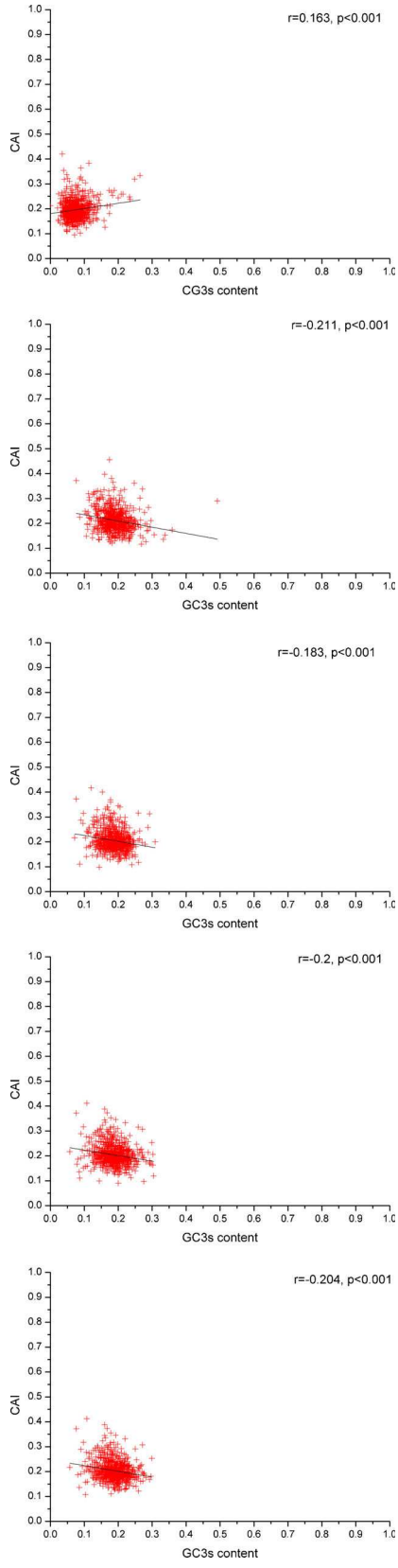

Figure 6. The codon adaptation index (CAI) values for a gene population of five strains of mycoplasmas. The correlation between CAI values and the GC content at the third nucleotide position of a codon (GC3s%) was estimated by Spearman's rank. The red point represents the CAI value against the GC3s% for each gene in the given gene population, and the black line was generated by correlation analysis. (a) *M. capricolum* subsp. *capricolum*; (b) *M. agalactiae*; (c), (d), and (e) three strains of *M. bovis*.

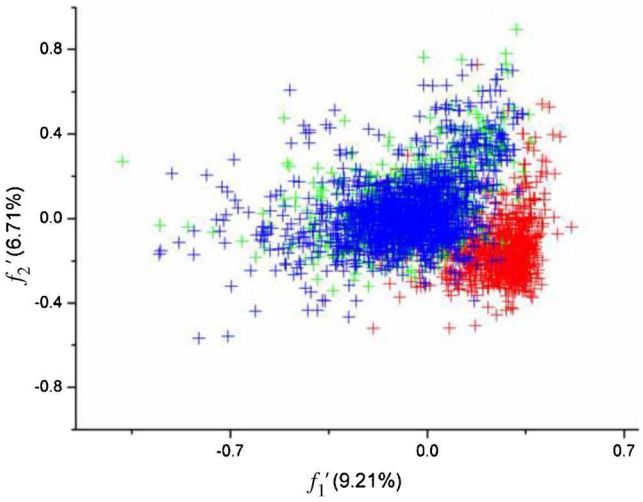

Figure 7. Comparative overall codon usage trends of five strains of mycoplasmas. The red points indicate the gene population of *M. capricolum* subsp. *capricolum*; the green plots indicate the gene population of *M. agalactiae*; and blue plots indicate the gene population of three strains of *M. bovis*.