Absence of translationally selected synonymous codon usage bias in Helicobacter pylori.

Bénédicte Lafay, John C. Atherton, Paul M. Sharp

Microbiology (2000), **146**, 851–860 Printed in Great Britain

Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*

Bénédicte Lafay,[1]† John C. Atherton[2] and Paul M. Sharp[1]

Author for correspondence: Paul M. Sharp. Tel: +44 115 970 9263. Fax: +44 115 970 9906.
e-mail: paul@evol.nott.ac.uk

Institute of Genetics[1], and Division of Gastroenterology, Department of Medicine and Institute of Infections and Immunity[2], University of Nottingham, Queen's Medical Centre, Nottingham NG7 2UH, UK

Synonymous codon usage in the complete genome of *Helicobacter pylori* was investigated. The moderate A+T-richness of the genome (G+C = 39 mol%) is reflected in the overall synonymous codon usage but the frequencies of some codons cannot be explained by simple mutational biases. A low level of heterogeneity among genes was observed, but this does not appear to be due to varying mutational bias or translational selection. Some of the heterogeneity was due to amino acid composition variation among the encoded proteins, and some may be attributable to recent acquisition of genes from other species. Since *Hel. pylori* codon usage is not dominated by biased mutation patterns, the absence of evidence for translationally mediated selection among synonymous codons is striking. This has implications with regard to the life history of this species, and in particular suggests that *Hel. pylori* strains are not subject to periods of competitive exponential growth. Despite the lack of selected codon usage, base composition immediately after the translation initiation site is skewed, consistent with selection against secondary structure formation in this region.

Keywords: codon usage, *Helicobacter pylori*, translational selection, *cag* pathogenicity island, molecular evolution

INTRODUCTION

Alternative synonymous codons are generally not used in equal frequencies. From studies of various bacterial species, two major paradigms have emerged. The enteric bacterium *Escherichia coli* (a member of the γ-subclass of the proteobacteria) represents the first of these paradigms. Codon usage in this species has been extensively studied, using both experimental and statistical approaches; it was also the first species to be studied in any detail (Post & Nomura, 1980; Ikemura, 1981a, b; Gouy & Gautier, 1982). In *E. coli*, synonymous codons that are recognized more efficiently and/or accurately by the most abundant tRNA species are preferred, and the strength of preference is correlated with the level of gene expression. Highly expressed genes have very high frequencies of these optimal codons, whilst in genes expressed at low levels the preference is weaker and codon usage seems more to reflect mutation patterns (Sharp & Li, 1986a; Bulmer, 1990). Thus, there is considerable heterogeneity among genes, and the codon usage in any gene reflects a particular point of balance between the forces of natural selection, mutational bias and random genetic drift (Sharp & Li, 1986b; Bulmer, 1991). Other species in which analogous patterns of codon usage are apparent include another γ-proteobacterium, *Haemophilus influenzae* (McInerney, 1997), the low-G+C Gram-positive *Bacillus subtilis* (Shields & Sharp, 1987) and the high-G+C Gram-positive *Mycobacterium tuberculosis* (Andersson & Sharp, 1996b). These species differ from *E. coli* insofar as the particular codons favoured by translational selection or mutational bias vary. Also, the degree of bias in highly expressed genes varies, presumably reflecting different strengths of natural selection in different species.

The second paradigm is represented by species where patterns of codon usage appear to be dominated by

†**Present address:** Centre d'Océanologie de Marseille, CNRS-UMR 6540, Station Marine d'Endoume, rue Batterie des Lions, 13007 Marseille, France.

Abbreviations: $GC3_s$, G+C content at synonymously variable third positions of sense codons; N_c, effective number of codons; RSCU, relative synonymous codon usage.

strong mutational biases. The clearest examples are bacteria with extreme, either G + C-rich or A + T-rich, genomic base compositions. Thus, the high-G + C Gram-positives *Micrococcus luteus* (Ohama *et al*., 1990) and *Streptomyces* (various species; Wright & Bibb, 1992) have genomic G + C contents around 74 mol% and almost exclusively use G- or C-ending codons. At the other extreme, the α-proteobacterium *Rickettsia prowazekii* and the Gram-positive *Mycoplasma capricolum* have a genomic G + C content of 29 or 25 mol%, respectively, and A- and U-ending codons are heavily used for all amino acids (where there is a choice) in all genes (Andersson & Sharp, 1996a; Ohkubo *et al*., 1987). Generally, in these species there is little heterogeneity among genes; all feature similar extremely biased codon usage. However, two types of intragenomic variation in codon usage related to gene location have recently been described. First, in *Mycoplasma genitalium* the extent of bias towards A + T-rich codons varies substantially and systematically around the genome (Kerr *et al*., 1997; McInerney, 1997). Second, in a number of species, the leading strand of replication has been found to be more G + T-rich than the lagging strand (Perrière *et al*., 1996; Francino & Ochman, 1997; McLean *et al*., 1998) and this is reflected in codon usage. In particular, in the spirochaete *Borrelia burgdorferi*, whilst codon usage is strongly influenced by the A + T-richness of the genome (G + C = 29 mol%), nevertheless it varies between genes from the two strands, with all U-ending and most G-ending codons occurring more frequently in genes located on the leading strand (McInerney, 1998; Lafay *et al*., 1999). For both types of variation, the simplest explanation would appear to be intragenomic differences in mutation biases. Certainly in neither case is the codon usage variation correlated with gene expression level.

Here, we investigate synonymous codon usage patterns in *Helicobacter pylori*, a species for which the complete genome sequence has been determined (Tomb *et al*., 1997). *Hel. pylori* is a Gram-negative, spiral-shaped bacterium assigned to the ε-subclass of the Proteobacteria (Olsen *et al*., 1994). It chronically infects more than half of the human population worldwide (Taylor & Blaser, 1991), and is involved in the pathogenesis of gastritis and peptic ulcer disease (Blaser, 1992). The results of our analyses indicate that codon usage in *Hel. pylori* does not conform to either of the two major paradigms outlined above. The genome of *Hel. pylori* is A + T-rich but not extremely so (G + C = 39 mol%), and it is apparent that codon usage in this species is not merely dominated by A + T-richness; however, there is little heterogeneity among genes and no sign that natural selection has shaped codon usage. These results are interpreted with reference to the ecology of *Hel. pylori*.

METHODS

Sequences. The complete genome sequence of *Hel. pylori* strain 26695 (Tomb *et al*., 1997) appears in the GenBank/EMBL/DDBJ DNA sequence database under the accession number AE000511. It contains 1566 annotated potential protein-encoding sequences, which were extracted using WWW-Query at the PBIL (Pôle Bio-Informatique Lyonnais) World Wide Web site (http://pbil.univ-lyon1.fr/). A number of genes were excluded from our main analyses. Twenty-six genes occur in the 40 kb *cag* pathogenicity island, which has a different G + C content compared to the remainder of the chromosome and is most likely the result of horizontal transfer (Covacci *et al*., 1997); these genes were analysed separately. Thirteen genes occur in insertion element sequences and were excluded because these can have compositional properties quite different from chromosomal genes. A further 509 genes had been identified only on the basis of being ORFs with compositional properties similar to other reliably identified genes; because we cannot be absolutely certain that these are true genes, we excluded them. Finally, one gene containing an internal stop codon, and a second containing an internal frameshift, were also excluded. The remaining dataset contained 1016 genes, all with homology either to known genes (834) or to unidentified reading frames conserved in other species (182).

For comparative purposes, codon usage datasets were compiled from the complete genome sequences of *E. coli* (Blattner *et al*., 1997), *Hae. influenzae* (Fleishmann *et al*., 1995) and *Bacillus subtilis* (Kunst *et al*., 1997). In each case, certain genes were excluded using the same criteria as were applied to the *Hel. pylori* dataset.

Analyses. Numbers of each codon, as well as the relative synonymous codon usage (RSCU), were calculated for each gene. The RSCU is the observed frequency of a codon divided by the frequency expected if all synonyms for that amino acid were used equally; thus RSCU values close to 1·0 indicate a lack of bias. Two indices of overall codon usage bias were calculated for each gene. The first was the G + C content at synonymously variable third positions of sense codons ($GC3_s$), which can potentially vary from 0 to 1. The variation of $GC3_s$ among genes was characterized by its standard deviation. The expected value of this standard deviation was calculated from binomial theory, using the harmonic mean of the number of synonymously variable third positions of codons. The second was the 'effective number of codons' (N_c) used in a gene (Wright, 1990). This is a measure of general nonuniformity of codon usage within groups of synonyms, which can vary from 20 (in a gene with extreme bias, where only one codon is used for each amino acid) to 61 (random codon usage). Since one potential source of nonuniformity is mutational bias, characterized by $GC3_s$, the expected value of N_c can be calculated for any given value of $GC3_s$ (Wright, 1990; see also Andersson & Sharp, 1996b). In addition, the base composition at each of the three codon positions, and amino acid composition of the predicted product, were calculated for each gene. All of these calculations were performed using the program CODONS (Lloyd & Sharp, 1992).

The pattern of codon usage variation among genes was investigated using correspondence analysis (Greenacre, 1984). This method plots genes according to their codon usage in a 59-dimensional space, and then identifies the major trends in codon usage as those axes through this multidimensional hyperspace which account for the largest fractions of the variation among genes. This method has been applied extensively in the analysis of codon usage (e.g. Grantham *et al*., 1981; Shields & Sharp, 1987; Andersson & Sharp, 1996a, b; McInerney, 1997).

RESULTS

Overall codon usage

Overall codon usage, summed across the 1016 selected *Hel. pylori* coding sequences (see Methods), is shown in Table 1. Codon usage for datasets excluding the 182 unidentified (but conserved) genes, or including the 509 additional ORFs without known homologues in other species, did not differ significantly from the pattern of relative usage of synonyms in Table 1. In this moderately A + T-rich genome (G + C = 39 mol%), we might expect to see an excess of U- and A-ending codons. This is true overall, in that the usage of G + C at synonymously variable third codon positions is 41 mol%. Also, for eight of the nine amino acids encoded by just two codons, the U- or A-ending codon is the most frequently used. However, for the amino acids encoded with three-, four- or sixfold degeneracy, among the 19 codons with RSCU values greater than 1·0, only 8 are U- or A-ending. Thus, overall codon usage is biased, but these patterns cannot be explained by a simple mutational bias towards A + T-richness. This complexity in the array of codon usage values could be due to a more complex mutational bias, or to natural selection favouring certain codons.

Some codons may be preferred by natural selection because of the population of tRNA molecules available for translation: the various species of isoaccepting tRNAs may be present in differing abundances and

Table 1. Codon usage in *Helicobacter pylori*

Total, 1016 genes; high, 57 highly expressed genes; low, 47 lowly expressed genes; *N*, number of codons; RSCU, relative synonymous codon usage; ter, termination codon; codons in bold are complementary to unmodified tRNA anticodons.

Amino acid	Codon	Total		High		Low		Amino acid	Codon	Total		High		Low	
		N	RSCU	*N*	RSCU	*N*	RSCU			*N*	RSCU	*N*	RSCU	*N*	RSCU
Phe	UUU	15167	1·59	218	1·49	638	1·60	Ser	UCU	5612	1·36	126	1·56	278	1·37
	UUC	3964	0·41	75	0·51	160	0·40		**UCC**	2000	0·49	39	0·48	114	0·56
Leu	**UUA**	16365	2·38	222	1·99	819	2·37		**UCA**	2062	0·50	51	0·63	102	0·50
	UUG	11113	1·62	200	1·80	520	1·50		UCG	1399	0·34	25	0·31	66	0·32
Leu	CUU	5796	0·84	111	1·00	298	0·86	Pro	CCU	6045	1·97	139	2·01	283	1·95
	CUC	3551	0·52	60	0·54	192	0·56		**CCC**	3247	1·06	39	0·56	185	1·27
	CUA	2810	0·41	48	0·43	166	0·48		**CCA***	1678	0·55	63	0·91	65	0·45
	CUG	1580	0·23	27	0·24	80	0·23		CCG	1281	0·42	36	0·52	48	0·33
Ile	AUU	13528	1·51	338	1·61	699	1·50	Thr	ACU	4787	1·21	142	1·40	227	1·19
	AUC	10324	1·15	242	1·22	554	1·19		**ACC**	5052	1·28	134	1·32	264	1·38
	AUA	3014	0·34	33	0·17	149	0·32		**ACA**	2191	0·56	57	0·56	114	0·60
Met	**AUG**	8338	–	239	–	389	–		ACG	3755	0·95	72	0·71	158	0·83
Val	GUU	5646	1·04	173	0·93	253	1·05	Ala	GCU	10032	1·55	270	1·63	442	1·53
	GUC	2953	0·55	87	0·47	154	0·64		**GCC**	5308	0·82	107	0·65	260	0·90
	GUA*	2091	0·39	103	0·55	83	0·34		**GCA***	2567	0·40	95	0·57	113	0·39
	GUG	10968	2·03	382	2·05	473	1·96		GCG	7992	1·23	190	1·15	338	1·17
Tyr	UAU	8877	1·39	164	1·34	381	1·37	Cys	UGU	1267	0·63	22	0·65	54	0·67
	UAC	3909	0·61	81	0·66	175	0·63		**UGC**	2745	1·37	46	1·35	108	1·33
ter	UAA	579	1·71	38	2·00	25	1·60	ter	UGA	284	0·84	8	0·42	16	1·02
ter	UAG	153	0·45	11	0·58	6	0·38	Trp	**UGG**	2426	–	47	–	67	–
His	CAU	5506	1·37	141	1·31	279	1·41	Arg	CGU	1791	0·86	74	0·80	105	0·88
	CAC	2508	0·63	74	0·69	118	0·59		**CGC**	3293	1·58	136	1·48	186	1·56
Gln	**CAA**	10744	1·70	208	1·63	566	1·72		**CGA**	837	0·40	22	0·24	61	0·51
	CAG	1896	0·30	48	0·38	91	0·28		CGG	388	0·19	7	0·08	21	0·18
Asn	AAU	11357	1·13	218	1·18	537	1·08	Ser	AGU	3422	0·83	56	0·69	152	0·75
	AAC	8736	0·87	151	0·82	454	0·92		**AGC**	10230	2·48	187	2·32	508	2·50
Lys	**AAA**	24266	1·54	762	1·43	1237	1·59	Arg	**AGA***	3041	1·46	199	2·16	160	1·34
	AAG*	7196	0·46	303	0·57	321	0·41		**AGG**	3128	1·51	115	1·25	182	1·53
Asp	GAU	12403	1·46	262	1·38	613	1·42	Gly	GGU*	3386	0·61	158	1·01	137	0·62
	GAC	4628	0·54	118	0·62	249	0·58		**GGC**	8093	1·46	240	1·53	342	1·55
Glu	**GAA**	18428	1·47	502	1·50	1039	1·53		**GGA**	2173	0·39	56	0·36	79	0·36
	GAG	6569	0·53	169	0·50	318	0·47		GGG	8546	1·54	172	1·10	325	1·47

* Codons with significantly ($P < 0·01$) higher frequencies in highly expressed genes.

alternative codons translated by a single tRNA are bound with differing efficiency and/or accuracy (Ikemura, 1981a, b). It is not clear that either of these factors can explain the high frequency of some C- and G-ending codons in *Hel. pylori*. The abundance of various tRNA species in *Hel. pylori* is not known. However, in *E. coli*, major variations in the abundance of different isoaccepting tRNAs are correlated with the copy numbers of the genes for those tRNAs, whereas in *Hel. pylori* only one tRNA (that for Glu) is encoded by more than one gene. Similarly, the anticodons of *Hel. pylori* tRNAs are not known, but indirect inferences can be made from the sequences of the anticodon sites in the tRNA genes, i.e. prior to possible modifications (codons best matching the unmodified anticodons of tRNA genes are shown in bold characters in Table 1). Few of the C- or G-ending codons with RSCU values greater than 1·0 are directly complementary to these unmodified anticodon sequences. For example, the purine-ending codons for Val, Ala and Gly are all translated by tRNAs with T at the wobble site in the gene sequence, which should favour A-ending codons; in each case the G-ending codons are 3–5 times more numerous. Furthermore, for the five amino acids with twofold degeneracy of pyrimidine-ending codons, the fact that the wobble site in the tRNA gene is G has not led to an excess of C-ending codons (Table 1). Finally, as discussed in detail below, there is no sign that codons which might be expected to be translationally optimal are more abundant in genes where translational selection is expected to be stronger, i.e. highly expressed genes. These factors suggest that translational selection has not been an important factor in shaping overall codon usage in *Hel. pylori*.

The high frequency of some C- and G-ending codons could be due to a complexity in mutational biases, such as an influence of the neighbouring base. Again, however, this does not seem sufficient to explain the patterns of codon usage in *Hel. pylori*. For example, if nearest-neighbour-influenced mutational biases were important, the relative frequencies of the four codons might be similar across the four sets of NCN codons (where N is A, C, G or T), encoding Ser, Pro, Thr and Ala. This is not the case: for all four, the U-ending codon is heavily used and the A-ending codon is underused, but the usage of C- and G-ending codons relative to each other and to the other codons varies substantially among the four sets (Table 1).

Heterogeneity among genes

In most genomes, other than those with extremely biased genomic base composition, there is considerable heterogeneity in codon usage patterns among genes. In bacterial species, this heterogeneity is usually associated with gene expression level, such that highly expressed genes have much higher frequencies of those codons that are translationally optimal. We have taken two approaches to examine the codon usage heterogeneity among *Hel. pylori* genes. First, we have examined the variation among genes statistically, using two different methods. Second, we have identified and compared two subsets of genes expected to have extreme expression levels. One subset included 53 ribosomal protein genes plus four translation elongation factor genes, all of which are expected to be expressed at high levels in all bacterial species. The other subset comprised 47 genes expected to be expressed at low levels, including regulatory genes and genes involved in cell division, chemotaxis and DNA repair. It is not necessarily true that all of these genes would be expressed at low levels in all species. However, since the modal level of expression across the entire complement of genes from a genome is low, the pattern of codon usage in lowly expressed genes should not differ much from the overall pattern.

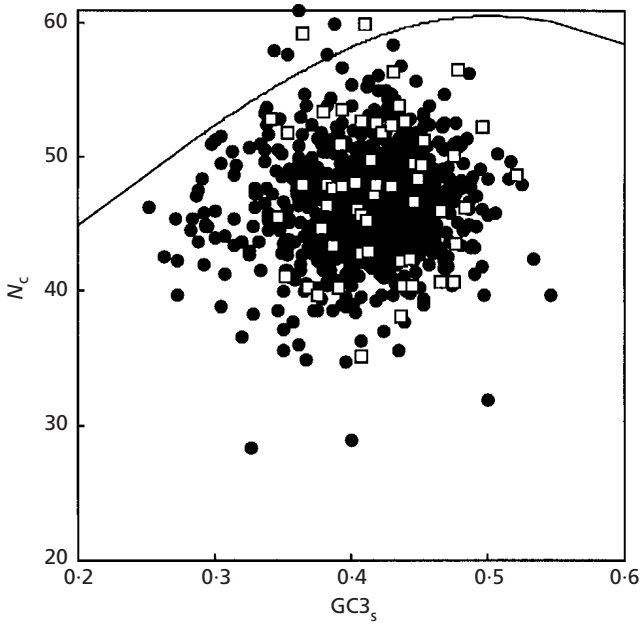

Fig. 1. Effective number of codons used (N_c) in each gene plotted against the G+C content at synonymously variable third positions of codons ($GC3_s$). The curve represents the expected value of N_c if bias is only due to G+C content. Genes expected to be highly expressed (genes encoding ribosomal proteins and translation elongation factors) are represented by open squares.

Wright (1990) suggested plotting the effective number of codons (N_c) against $GC3_s$ as a means of characterizing codon usage variation among genes. Genes subject to G+C compositional constraints will lie on or just below the $GC3_s$ curve, whereas genes subjected to selection for a subset of translationally optimal codons will depart from this 'neutral' position (e.g. see Wright, 1990; Andersson & Sharp, 1996a, b; McInerney, 1997). Such a plot for the 1016 *Hel. pylori* genes (Fig. 1) shows that most of the points are tightly clustered around the mean values of $GC3_s = 0{\cdot}41$, $N_c = 47$. The extent of dispersal is surprisingly low compared to other species. For example, the observed standard deviation of $GC3_s$ values is only a little larger than that expected with

Table 2. Comparison of genomic and codon usage statistics of *Helicobacter pylori*, *Escherichia coli*, *Haemophilus influenzae* and *Bacillus subtilis*

	Genome			ORFs†	$GC3_s$‡			C.A. axes§			
	Size (Mbp)	G + C content (mol %)	tRNAs*		Mean	SD	Ratio	1st	2nd	3rd	4th
Helicobacter pylori	1·7	39	36	1016	0·41	0·042	1·31	6·3	5·2	5·0	4·7
Escherichia coli	4·6	51	86	3984	0·53	0·084	2·40	15·2	7·7	4·0	3·6
Haemophilus influenzae	1·8	38	58	1473	0·26	0·053	1·71	11·1	7·2	5·4	4·5
Bacillus subtilis	4·2	44	88	2435	0·43	0·072	2·25	12·4	7·6	4·7	4·4

* Number of tRNA genes identified within the genome.

† Number of protein-encoding genes included in the analyses.

‡ Mean and standard deviation, across genes, and ratio of observed-to-expected standard deviation, of G + C content at synonymously variable third position of codons.

§ Percentage of variation explained by each of the first four axes of correspondence analyses.

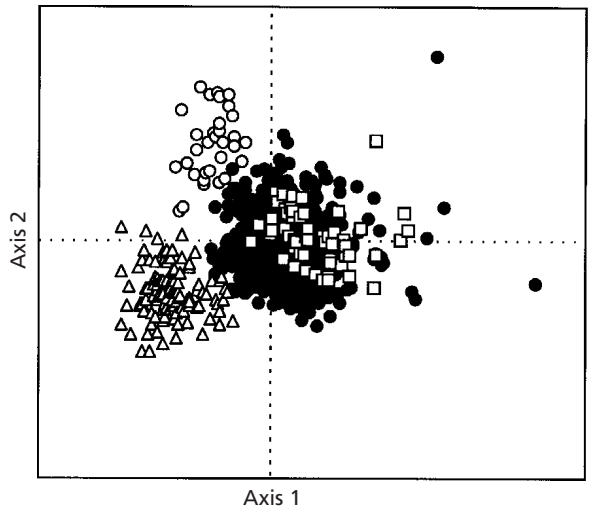

Fig. 2. Correspondence analysis of codon usage frequency variation among *Hel. pylori* genes. Genes are plotted at their coordinates on the first two axes produced by the analysis. Genes expected to be highly expressed (genes encoding ribosomal proteins and translation elongation factors) are represented by open squares. Genes rich in codons for hydrophobic acids, encoding integral membrane proteins, are represented by open triangles. Genes unusually rich in codons for Ser, Thr, Tyr, Gly, Asn and Gln, encoding outer-membrane proteins, flagellar proteins or secreted proteins, are represented by open circles.

random variation, and rather smaller than seen in similar analyses of *E. coli*, *Hae. influenzae* or *Bacillus subtilis* genes (Table 2). What little dispersal is evident is not at all related to gene expression level, in that genes from neither the highly expressed subset (Fig. 1) nor the lowly expressed subset (not shown) are distributed differently from other genes. The fact that the cluster of genes is centred somewhat below the curve for the expected values is consistent with the observation (see above) that the overall bias is not simply explicable in terms of bias in G + C content.

A second approach to explore the variation among genes is to use multivariate statistical analysis. In particular, correspondence analysis can be used to investigate whether there are any major trends in codon usage by identifying and assessing the importance of the major axes through a plot of genes in 59-dimensional hyperspace; typically, the first two or three axes reveal any interesting trends among genes. Correspondence analysis of codon usage frequencies was found to separate *Hel. pylori* genes on the basis of amino acid composition, rather than synonymous codon usage. On a plot of the positions of genes at their coordinates on the first two axes (Fig. 2), the great majority of points are clustered around the null coordinates, but two discrete subsets of genes are apparent at one extremity of axis 1. One group (lower left on Fig. 2) comprises 112 integral membrane protein genes, many of which encode transport proteins; these are differentiated from other chromosomal coding sequences by a particular amino acid usage which is reflected at the gene level by an unusual overall richness in codons for the hydrophobic amino acids Phe, Leu, Ile, Met and Val (Lobry & Gautier, 1994). The second group (upper left in Fig. 2) comprises 32 genes encoding outer membrane (*omp*), flagellar or secreted proteins. Overall, these genes were unusually rich in codons for Ser, Thr, Tyr, Gly, Asn and Gln. Both groups of genes had an excess of Trp codons. Neither the highly expressed gene subset, although comparatively slightly shifted to the right because of their function-related differential amino acid composition (Fig. 2), nor the lowly expressed gene subset (not shown) appeared distinct from the main cluster of genes.

In an attempt to examine variation of synonymous codon usage, rather than of amino acid composition, we performed correspondence analysis on RSCU values. This analysis revealed no obvious trend among genes. In particular, highly expressed genes are not distinguished on any of the first four axes produced by the analysis (the first two axes are shown in Fig. 3). No statistic of codon usage, base or amino acid composition was found

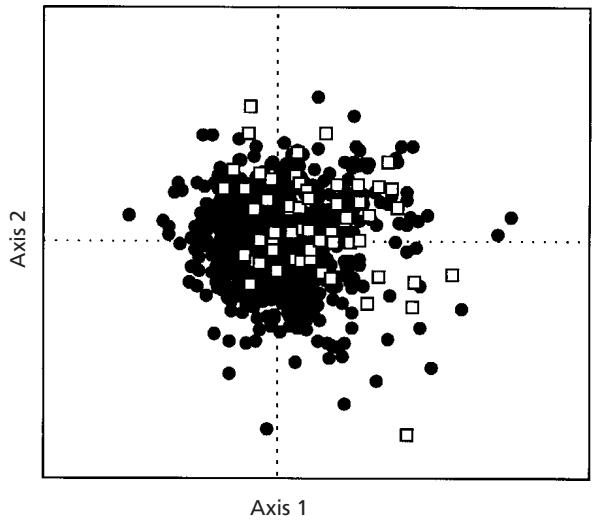

Fig. 3. Correspondence analysis of RSCU variation among *Hel. pylori* genes. Genes are plotted at their coordinates on the first two axes produced by the analysis. Genes expected to be highly expressed (genes encoding ribosomal proteins and translation elongation factors) are represented by open squares.

to be strongly correlated with position on any of the first four axes. The best correlation we could find was between position on axis 1 and $GC3_s$: the coefficient value (0·32) was highly significant ($P < 0{\cdot}001$), but indicates that $GC3_s$ can explain only 10% of the variation on axis 1. Among the genes on the left in Fig. 3, mostly with low G+C content at the third codon position, several encode restriction-modification systems. However, at either extreme of axes 1 and 2 are found genes corresponding to a variety of protein functional classes, expected levels of expression and locations on the chromosome; these include genes encoding conserved secreted or integral membrane proteins, restriction enzymes, ribosomal proteins, translation elongation factors, ATP-binding cassette (ABC) transporters and VirB4 homologues. Whilst some of these genes, such as those encoding restriction enzymes and the VirB4 homologues (Tomb *et al*., 1997), may well have a foreign origin, this explanation cannot simply account for all of the genes found at the extreme coordinates on axes 1 and 2.

Additionally, a moderately good correlation was observed between position on axis 3 and G+T content at the third codon position (GT3; $r = 0{\cdot}48$). This GT3 variation was further explored, taking into account the replication strand (leading or lagging). In contrast to the G+T content variation in the spirochaetes *Borrelia burgdorferi* and *Treponema pallidum* (McInerney, 1998; Lafay *et al*., 1999), genes on both leading and lagging strands exhibited similar GT3 range ($0{\cdot}52 \pm 0{\cdot}054$ and $0{\cdot}54 \pm 0{\cdot}052$, respectively) and distribution along axis 3.

Examination of the values for the amounts of variation (in the total dataset) explained by the first four axes of the correspondence analysis in comparison with results from similar analyses of other species (Table 2) revealed two aspects of the *Hel. pylori* data. First, the fraction of variation explained by the first axis was very low in *Hel. pylori*. Second, the amounts of variation explained by the first and second (and subsequent) axes differed little in *Hel. pylori*. These results suggest that there are no strong trends among the data. Consistent with this, when 101 genes (10% of the total dataset) from either extreme of the first axis produced by the correspondence analysis were compared, their patterns of codon usage were quite similar, with little differences between the two datasets mainly restricted to the usage of CCC (Pro), ACG (Thr), CGC (Arg) and GGG (Gly) codons.

We also compiled codon usage data for subsets of highly and lowly expressed genes and compared them with each other and with the total dataset (Table 1). There was little difference in codon usage between the highly expressed genes and either the lowly expressed genes or the total dataset. Using a chi square statistic to compare the relative frequencies (within each group of synonyms) of codons between the highly and lowly expressed genes, only six codons were found at a significantly ($P < 0{\cdot}01$) higher frequency in the highly expressed group (indicated by an asterisk in Table 1). In the highly expressed genes, three of these codons (AAG, GUA, GCA) are still comparatively rare (RSCU $< 0{\cdot}6$), and only one (AGA, for Arg) is the most highly used within a group of synonyms. There is no clear tendency for the highly expressed genes to differ from the others in having an increased frequency of the codons expected to be translationally optimal. For example, in *Hel. pylori*, as in other species, the only genes for tRNAs for Phe and Tyr have G at the wobble site in the anticodon; in other species, UUC and UAC have been found to be the optimal codons for Phe and Tyr, respectively, but in *Hel. pylori*, both of these codons are used only marginally more often in the highly expressed genes (Table 1).

Intragenic variation in codon usage

Base composition and codon usage have been found to vary within genes in other species, with, in particular, differences between the region close to the translation initiation site and the rest of the gene (Eyre-Walker & Bulmer, 1993). Therefore, we examined base composition and codon usage in the *Hel. pylori* genes, as a function of distance from the start codon. The first few codons after the initiation codon were found to have compositional biases different from the rest of the gene (Fig. 4). The first two codons are unusually A-rich, and T- and G-poor. Although less marked, the excess of A and lack of G extends over about 10 codons. The excess of A was found at all three positions of codons, and the frequency of the codon AAA (for Lys) was found to be elevated to about 15% (around two times its normal level) at the first two positions after the start codon, making it by far the most common codon at those sites. In contrast, the alternative Lys codon (AAG) was not increased at these positions.

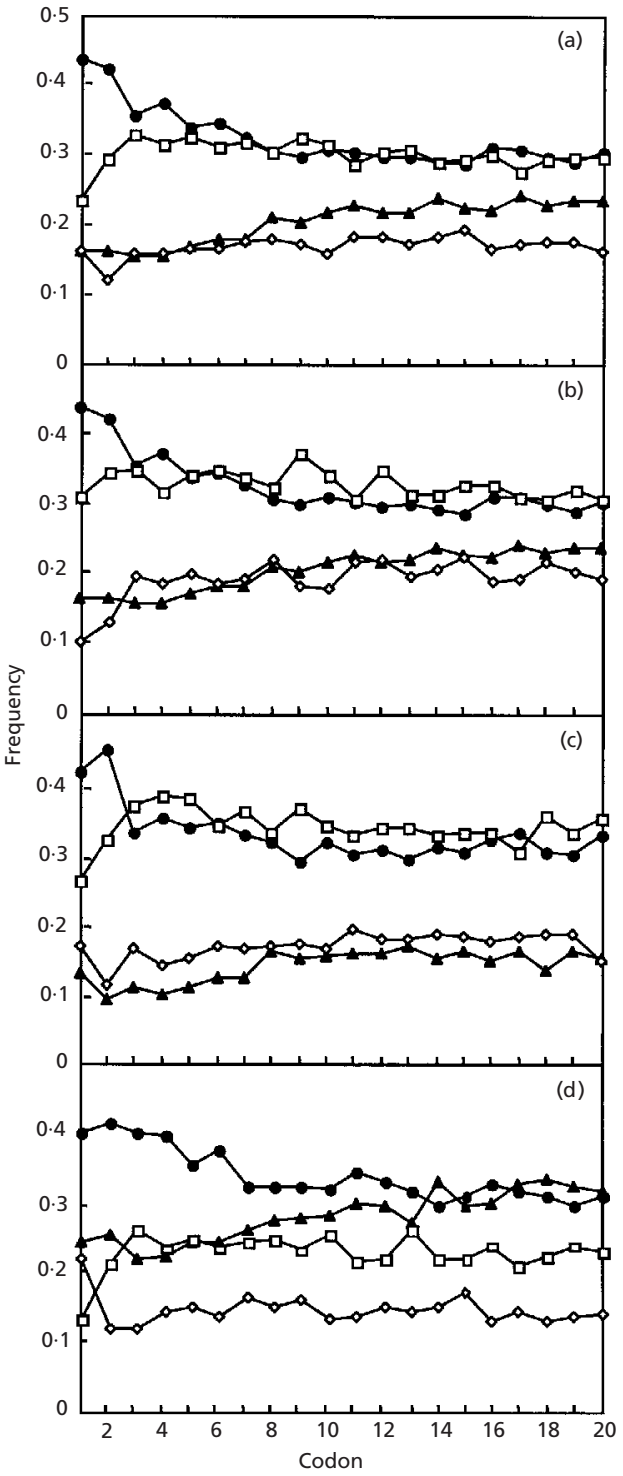

Fig. 4. Mean base composition at the beginning of genes. Composition of each nucleotide is plotted against position for each of the 20 first codons after the initiation codon. T is represented by open squares, C by open diamonds, A by filled circles and G by filled triangles. (a) Overall codon base composition; (b) base composition at the third position of codon; (c) base composition at the second position of codon; (d) base composition at the first position of codon.

cag pathogenicity island

Covacci *et al*. (1997) proposed that the *cag* pathogenicity island involved in *Hel. pylori* virulence arrived in the genome through horizontal transfer. Indeed, the G+C content of this region (35 mol%) is significantly lower than the rest of the genome (39 mol%), which justified the exclusion of the 26 genes corresponding to this region from our codon usage analysis. The difference appears more clearly if only the third codon position is considered; the mean value is 34 mol% for the *cag* pathogenicity island genes whereas the 1016 genes in our dataset have a mean value of 41 mol%. The heterologous origin of the pathogenicity island is further supported by more detailed examination of codon usage in the *cag* genes. Within these genes, there are 16 codons used significantly more often, and 13 codons used significantly less often ($P < 0{\cdot}01$), than in genes from the rest of the genome. The codons relatively underused included UUU and UUA, indicating that the difference in the *cag* genes is not merely one of base composition (G+C content), but of more complex codon usage patterns.

DISCUSSION

Synonymous codon usage patterns in *Hel. pylori* strain 26695 have been investigated. The complete genome sequence contains 1566 annotated putative genes (Tomb *et al*., 1997); we focused on 1016 genes exhibiting homology to genes in other species, but analyses using all 1566 genes produced very similar results. Overall codon usage was found to be moderately A+U-rich, with the mean frequency of $GC3_s$ being 41 mol%. This reflects the 39 mol% G+C content of the genome as a whole, but is not simply interpretable in terms of nucleotide composition since some of the more abundant codons end in C or G. The most striking observation was a relative lack of heterogeneity in codon usage among genes, and in particular an absence of differentiation between highly expressed genes and others. Such homogeneous codon usage has previously been found only in bacterial species with extreme base compositional bias. That it occurs in *Hel. pylori* suggests that natural selection has been largely ineffective in shaping codon usage in this species, and raises the question why?

In *E. coli*, codon selection is partly geared to variation in the abundance of different tRNAs for the same amino acid, and this in turn is correlated with tRNA gene copy number. The *Hel. pylori* genome contains only 36 tRNA genes and only one tRNA has more than one gene. Thus, it is possible that different isoaccepting tRNAs have similar abundances in *Hel. pylori*, and that this potential source of codon selection does not exist. However, where multiple (two, or possibly three) codons are translated by a single tRNA, one synonym is expected to be more efficiently and/or accurately recognized than the other(s) (Ikemura, 1981a, b). This source of potential selective differences among codons should exist in *Hel. pylori*, as in other species.

Since the selective differences among alternative synonyms are expected to be very small, natural selection can only have been effective if the long-term evolutionary effective population size (N_e) of the species has been large (Shields *et al*., 1988; Bulmer, 1991). Small values of N_e may explain why codon selection appears to have been ineffective in mammals (Sharp *et al*., 1993), but N_e values for bacterial species are expected to be quite large, and the very high level of allelic diversity at allozyme loci in *Hel. pylori* (Go *et al*., 1996) is consistent with a large effective population size. Furthermore, whilst the clonal population structure of some bacterial species would be expected to reduce the efficacy of weak natural selection on synonymous codons (due to selection at linked sites in the genome), *Hel. pylori* appears to have a highly recombinational population structure (Go *et al*., 1996; Göttke *et al*., 1996; Blaser, 1997; Cao & Cover, 1997; Hazell *et al*., 1997; Salaün *et al*., 1998; Suerbaum *et al*., 1998).

The implication of the above discussion would appear to be that there has been no opportunity for natural selection to influence synonymous codon usage in *Hel. pylori*. In *E. coli*, for example, the genes with the most extreme codon usage bias include those encoding ribosomal proteins and elongation factors, i.e. components of the translational machinery. These proteins are particularly abundant during periods of exponential growth (Kurland, 1993), suggesting that competition during such periods provides the major selective force for codon usage adaptation. The absence of selected codon usage bias in *Hel. pylori* suggests that this species does not go through periods of competitive exponential growth. This might be explained by the particularly hostile environment inhabited by *Hel. pylori*, i.e. the gastric mucosa, with low pH and an active immune response, and continual washing generated by peristalsis (Marshall *et al*., 1998). Multiple infections by different strains of *Hel. pylori* (Beji *et al*., 1989; Owen *et al*., 1993; Fujimoto *et al*., 1994; Akopyants *et al*., 1995; Taylor *et al*., 1995; Berg *et al*., 1997), as well as localized mixed occurrences (Kitamoto *et al*., 1998; J. C. Atherton, unpublished) have been reported, but there is no evidence that these strains compete via exponential growth. Nor is there evidence that strains of *Hel. pylori* compete with other species since other bacteria do not normally survive in this environment.

The correlation of codon usage frequencies and tRNA abundances in (for example) *E. coli* probably represents a coadaptive strategy to maximize growth rates (Bulmer, 1987; Berg & Kurland, 1997). The lack of duplicated tRNA genes in *Hel. pylori* may be a response to a lack of selection on codon usage, rather than the cause. In this context, it is interesting also to compare *Hel. pylori* with *Hae. influenzae* (Table 2). The two species have similar genome sizes and genomic G+C contents but (in contrast to *Hel. pylori*) *Hae. influenzae* genes show evidence of selected codon usage bias (McInerney, 1997). However, the strength of selected bias in *Hae. influenzae*, as revealed by comparisons of codon usage in genes expressed at high and low levels (unpublished observations) or by consideration of the amount of variation explained by the first axis of correspondence analysis (Table 2), is less than that seen in *E. coli*. Consistent with this, the number of tRNA genes encoded within the *Hae. influenzae* genome is intermediate between *Hel. pylori* and *E. coli* (Table 2).

Despite the lack of evidence that natural selection has shaped codon usage generally in *Hel. pylori*, base composition immediately after the translation initiation site is different from other sites within genes. These 5′ regions are comparatively A-rich and G-poor. This does not appear to be related to codon usage effects since it extends to all three codon positions, or to amino acid composition since the frequency of the codon AAA is greatly elevated, but that of the alternative Lys codon, AAG, is not. Rather, this appears to be related to the nucleotide sequence, and may reflect selection against intramolecular base pairing and formation of secondary structure of the mRNAs which could interfere with ribosome binding at the initiation site (Stormo *et al*., 1982; Eyre-Walker & Bulmer, 1993).

The strongest, albeit still weak, source of codon usage variation among genes was found to be correlated with $GC3_s$. Genes may have atypical base composition (and/or codon usage) if they are the product of relatively recent horizontal transfer (Lawrence & Ochman, 1997). As noted above, evidence has been found for extensive recombination among strains of *Hel. pylori*, and it is also possible that the *Hel. pylori* genome has acquired sequences from more distantly related species (Go *et al*., 1996). Indeed, the 40 kb *cag* pathogenicity island has been suggested to be the result of horizontal transfer because of its unusually low G+C content (Covacci *et al*., 1997) and the codon usage of genes from this island further supports this hypothesis. For this reason, we excluded genes from this region from our analyses. It may be that there are many more genes that have arrived in the *Hel. pylori* genome through horizontal transfer. Such 'foreign' genes are more likely to have unusual base composition. For example, restriction-modification system genes are liable to horizontal transfer, and in *Hel. pylori* some of these genes are among those with the lowest $GC3_s$ values ($GC3_s < 0{\cdot}30$). The degree to which the codon usage in a foreign gene deviates from that of other genes depends on both the extent of difference in codon usage pattern of the donor species and the time since gene acquisition. Thus a range of codon usage patterns are expected in foreign genes, and it is not possible from codon usage studies to estimate what fraction of genes have arrived through horizontal transfer.

Finally, we note that the low level of codon usage heterogeneity among *Hel. pylori* genes has (at least) two practical implications. First, Suerbaum *et al*. (1998) recently applied the homoplasy test (Maynard Smith & Smith, 1998) to investigate the extent of recombination among strains for three *Hel. pylori* genes. That method requires an estimate of the effective number of sites (S_e) free to vary. Suerbaum *et al*. (1998) examined syn-

onymous nucleotide variation and assumed that different expression levels of the *flaA*, *vacA* and *flaB* genes would entail different extents of selected codon usage bias, and thus yield different fractions of potentially variable synonymous sites in the three genes. Our results suggest that this is not appropriate, although it is not clear that this has affected the results reported by Suerbaum *et al*. (1998) in a substantial way. Second, the relative uniformity of codon usage may be helpful for computational methods of gene detection that rely on overall compositional biases. In the *Hel. pylori* genome sequencing project (Tomb *et al*., 1997), genes without database homologues were identified using GeneMark (Borodovsky & McIninch, 1993), which predicts coding regions based on compositional similarity to a training set of known genes. Salzberg *et al*. (1998) have recently reported that a more complex interpolated Markov-model-based approach seemed to find *Hel. pylori* genes extremely efficiently. Such methods may have less success in genomes where genes exhibit greater codon usage heterogeneity.

ACKNOWLEDGEMENTS

This work was supported by award G04905 from the BBSRC.

REFERENCES

Akopyants, N. S., Eaton, K. A. & Berg, D. E. (1995). Adaptive mutation and cocolonization during *Helicobacter pylori* infection of gnotobiotic piglets. *Infect Immun* **63**, 116–121.

Andersson, S. G. E. & Sharp, P. M. (1996a). Codon usage and base composition in *Rickettsia prowazeckii*. *J Mol Evol* **42**, 525–536.

Andersson, S. G. E. & Sharp, P. M. (1996b). Codon usage in the *Mycobacterium tuberculosis* complex. *Microbiology* **142**, 915–925.

Beji, A., Vincent, P., Dachis, I., Husson, M. O., Cortol, A. & Leclerc, H. (1989). Evidence of gastritis with several *Helicobacter pylori* strains. *Lancet* **2**, 1402–1405.

Berg, D. E., Gilman, R. H., Lelwala-Guruge, J. & 9 other authors (1997). *Helicobacter pylori* populations in Peruvian patients. *Clin Infect Dis* **25**, 996–1002.

Berg, O. G. & Kurland, C. G. (1997). Growth rate-optimised tRNA abundance and codon usage. *J Mol Biol* **270**, 544–550.

Blaser, M. J. (1992). *Helicobacter pylori* – its role in disease. *Clin Infect Dis* **15**, 386–393.

Blaser, M. J. (1997). Heterogeneity *of Helicobacter pylori*. *Eur J Gastroenterol Hepatol* **9**, S3–S6.

Blattner, F. R., Plunkett, G. I., Bloch, A. & 14 other authors (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462.

Borodovsky, M. & McIninch, J. D. (1993). Parallel gene recognition for both DNA strands. Comput Chem **17**, 123–133.

Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. *Nature* **325**, 728–730.

Bulmer, M. (1990). The effect of context on synonymous codon usage genes with low codon usage bias. *Nucleic Acids Res* **18**, 2869–2873.

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

Cao, P. & Cover, T. L. (1997). High-level genetic diversity in the *vapD* chromosomal region of *Helicobacter pylori*. *J Bacteriol* **179**, 2852–2856.

Covacci, A., Falkow, S., Berg, D. E. & Rappuoli, R. (1997). Did the inheritance of a pathogenicity island modify the virulence of *Helicobacter pylori*? *Trends Microbiol* **5**, 205–208.

Eyre-Walker, A. & Bulmer, M. (1993). Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res* **21**, 4599–4603.

Fleischmann, R. D., Adams, M. D., White, O. & 37 other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

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

Fujimoto, S., Marshall, B. & Blaser, M. J. (1994). PCR-based restriction fragment length polymorphism typing of *Helicobacter pylori*. *J Clin Microbiol* **32**, 331–334.

Go, M. F., Kapur, V., Graham, D. Y. & Musser, J. M. (1996). Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. *J Bacteriol* **178**, 3934–3938.

Göttke, M. U., Groody, J. M., Loo, V., Fallone, C. A., Barkum, A. N. & Beech, R. N. (1996). Panmyctic population structure due to frequent recombination in *Helicobacter pylori*. *Gut* **39**, A121.

Gouy, M. & Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* **10**, 7055–7074.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* **9**, r43–r74.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Hazell, S. L., Andrews, R. H., Mitchell, H. M. & Daskalopoulos, G. (1997). Genetic relationship among isolates of *Helicobacter pylori*: evidence for the existence of a *Helicobacter pylori* species-complex. *FEMS Microbiol Lett* **150**, 27–32.

Ikemura, T. (1981a). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol* **146**, 1–21.

Ikemura, T. (1981b). 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.

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

Kitamoto, N., Nakamoto, H., Katai, A., Takahara, N., Nakata, H., Tamaki, H. & Tanaka, T. (1998). Heterogeneity of protein profiles of *Helicobacter pylori* isolated from individual patients. *Helicobacter* **3**, 152–162.

Kunst, F., Ogasawara, N., Moszer, I. & 148 other authors (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.

Kurland, C. G. (1993). Major codon preference: theme and variations. *Biochem Soc Trans* **21**, 841–846.

Lafay, B., Lloyd, A. T., McLean, M. J., Devine, K. M., Sharp, P. M. & 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.

Lawrence, J. G. & Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**, 383–397.

Lobry, J. R. & Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res* **22**, 3174–3180.

Lloyd, A. T. & Sharp, P. M. (1992). CODONS: a microcomputer program for codon usage analysis. *J Hered* **83**, 239–240.

McInerney, J. O. (1997). Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. *Microb Comp Genomics* **2**, 1–10.

McInerney, J. O. (1998). Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc Natl Acad Sci USA* **95**, 10698–10703.

McLean, M. J., Wolfe, K. H. & Devine, K. M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol* **47**, 691–696.

Marshall, D. G., Dundon, W. G., Beesley, S. M. & Smyth, C. J. (1998). *Helicobacter pylori* – a conundrum of genetic diversity. *Microbiology* **144**, 2925–2939.

Maynard Smith, M. & Smith, N. H. (1998). Detecting recombination from gene trees. *Mol Biol Evol* **15**, 590–599.

Ohama, T., Muto, A. & Osawa, S. (1990). Role of GC-biased mutation pressure on synonymous codon choice in *Micrococcus luteus*, a bacterium with a high genomic GC-content. *Nucleic Acids Res* **18**, 1565–1569.

Ohkubo, S., Muto, A., Kawauchi, Y., Yamao, F. & Osawa, S. (1987). The ribosomal protein gene cluster of *Mycoplasma capricolum*. *Mol Gen Genet* **210**, 314–322.

Olsen, G. J., Woese, C. R. & Overbeek, R. (1994). The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol* **176**, 1–6.

Owen, R. J., Desai, M., Figura, N., Bayeli, P. F., Di Gregorio, L., Russi, M. & Musmanno, R. A. (1993). Comparisons between degree of histological gastritis and DNA fingerprints, cytotoxicity and adhesivity of *Helicobacter pylori* from different gastric sites. *Eur J Epidemiol* **9**, 315–321.

Perrière, G., Lobry, J. R. & Thioulouse, J. (1996). Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences. *Comput Appl Biosci* **12**, 519–524.

Post, L. E. & Nomura, M. (1980). DNA sequences from the *str* operon of *Escherichia coli*. *J Biol Chem* **255**, 4660–4666.

Salaün, L., Audibert, C., Le Gay, G., Burucoa, C., Fauchère, J. L. & Picard, B. (1998). Panmictic structure of *Helicobacter pylori* demonstrated by the comparative study of six genetic markers. *FEMS Microbiol Lett* **161**, 231–239.

Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **26**, 544–548.

Sharp, P. M. & Li, W.-H. (1986a). Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for ''rare'' codons. *Nucleic Acids Res* **14**, 7737–7749.

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

Sharp, P. M., Stenico, M., Peden, J. F. & Lloyd, A. T. (1993). Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans* **21**, 835–841.

Shields, D. C. & Sharp, P. M. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res* **15**, 8023–8040.

Shields, D. C., Sharp, P. M., Higgins, D. G. & Wright, F. (1988). ''Silent'' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol* **5**, 704–716.

Stormo, G. D., Schneider, T. D. & Gold, L. M. (1982). Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Res* **10**, 2971–2996.

Suerbaum, S., Maynard Smith, J., Bapumia, K., Morelli, G., Smith, N. H., Kunstmann, E., Dyrek, I. & Achtman, M. (1998). Free recombination within *Helicobacter pylori*. *Proc Natl Acad Sci USA* **95**, 12619–12624.

Taylor, D. N. & Blaser, M. J. (1991). The epidemiology of *Helicobacter pylori*. *Epidemiol Rev* **13**, 42–59.

Taylor, N. S., Fox, J. G., Akopyants, N. S. & 8 other authors (1995). Long-term colonization with single and multiple strains of *Helicobacter pylori* assessed by DNA-fingerprinting. *J Clin Microbiol* **33**, 918–923.

Tomb, J.-F., White, O., Kerlavage, A. R. & 39 other authors (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547.

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

Wright, F. & Bibb, M. J. (1992). Codon usage in the G+C-rich *Streptomyces* genome. *Gene* **113**, 55–65.

Received 31 December 1999; accepted 13 January 2000.