Microbiology (2003), **149**, 2585–2596 DOI 10.1099/mic.0.26381-0

Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont *Wigglesworthia*

Joshua T. Herbeck,[1] Dennis P. Wall[2] and Jennifer J. Wernegreen[1]

[1]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA

[2]Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA

Correspondence
Jennifer J. Wernegreen
jwernegreen@mbl.edu

Received 31 March 2003
Revised 5 June 2003
Accepted 18 June 2003

Wigglesworthia glossinidia brevipalpis, the obligate bacterial endosymbiont of the tsetse fly *Glossina brevipalpis*, is characterized by extreme genome reduction and AT nucleotide composition bias. Here, multivariate statistical analyses are used to test the hypothesis that mutational bias and genetic drift shape synonymous codon usage and amino acid usage of *Wigglesworthia*. The results show that synonymous codon usage patterns vary little across the genome and do not distinguish genes of putative high and low expression levels, thus indicating a lack of translational selection. Extreme AT composition bias across the genome also drives relative amino acid usage, but predicted high-expression genes (ribosomal proteins and chaperonins) use GC-rich amino acids more frequently than do low-expression genes. The levels and configuration of amino acid differences between *Wigglesworthia* and *Escherichia coli* were compared to test the hypothesis that the relatively GC-rich amino acid profiles of high-expression genes reflect greater amino acid conservation at these loci. This hypothesis is supported by reduced levels of protein divergence at predicted high-expression *Wigglesworthia* genes and similar configurations of amino acid changes across expression categories. Combined, the results suggest that codon and amino acid usage in the *Wigglesworthia* genome reflect a strong AT mutational bias and elevated levels of genetic drift, consistent with expected effects of an endosymbiotic lifestyle and repeated population bottlenecks. However, these impacts of mutation and drift are apparently attenuated by selection on amino acid composition at high-expression genes.

INTRODUCTION

A balance of mutational biases, selection and genetic drift determines patterns of synonymous codon and amino acid usage (Akashi, 1997; Bulmer, 1991). Variation in codon and amino acid usage among genes and among species can elucidate evolutionary processes that alter this mutation-selection balance (Kreitman & Antezana, 1999; Sharp & Li, 1987; Singer & Hickey, 2000). Underlying processes that affect codon and amino acid usage include translational selection related to gene expression levels (Akashi, 1994; Sharp & Li, 1987), directional mutational biases (prevalence of mutations from GC to AT or vice versa) (Bernardi, 1985; D'Onofrio *et al.*, 1991), distinct mutational spectra on the leading or lagging strands of replication (McInerney, 1998a) and genetic drift as influenced by effective population size (N_e) (Bulmer, 1991; Powell & Moriyama, 1997). With the increasing number of fully sequenced genomes, it is now possible to study the distinct genome patterns that reflect these varying evolutionary pressures.

Prokaryotic genomes are ideal for comparative analyses of the mutation-selection balance because diverse bacterial lifestyles influence mutation, selection and genetic drift, and can be linked to distinct patterns of genome evolution. Synonymous codon usage in the free-living bacteria *Escherichia coli* and *Salmonella typhimurium* is dictated by selection for translational efficiency, as evidenced by bias toward the use of 'optimal codons' at highly expressed genes (Sharp, 1991; Sharp & Li, 1987). In contrast, synonymous codon usage is shaped by mutational bias and genetic drift in many obligately host-associated bacteria, such as the aphid mutualist *Buchnera aphidicola* (Wernegreen & Moran, 1999) and the parasites *Rickettsia prowazekii* (Andersson & Sharp, 1996) and *Micrococcus luteus* (Ohama *et al.*, 1990), among other species. Strong effects of mutation and drift on codon usage may reflect a reduced N_e caused by recurrent bottlenecks related to an intracellular lifestyle (Mira & Moran, 2002; Moran & Wernegreen, 2000). For example,

Abbreviations: CAI, codon adaptation index; COA, correspondence analysis; d_N, non-synonymous divergences; GC_3, % GC at third codon position; GC_{12}, % GC at first and second codon positions; N_c, effective number of codons; N_e, effective population size; RAAU, relative amino acid usage; RSCU, relative synonymous codon usage; r_s, Spearman's rank correlation coefficient.

independent population genetic studies of two divergent aphid species estimated the N_e of *Buchnera* to be $\sim 10^7$ (Abbot & Moran, 2002; Funk *et al.*, 2001) whereas the N_e for the closely related bacterium *E. coli* has been estimated at $2{\cdot}5 \times 10^9$ (Ochman & Wilson, 1987). Reduced N_e causes stronger genetic drift and decreases the efficiency of selection on preferred codons or amino acids. Among intracellular bacterial genomes, the irrevocable loss of DNA repair genes may contribute to elevated underlying mutation rates and biases. Such mutational bias can affect not only synonymous codon usage, but also amino acid usage (Sueoka, 1961) as seen in *Buchnera* (Palacios & Wernegreen, 2002), *Rickettsia* (Andersson & Sharp, 1996) and a broad array of other prokaryotic genomes (Singer & Hickey, 2000).

This study examines codon and amino acid usage in the full genome sequence of *Wigglesworthia glossinidia brevipalpis*, the γ-3 proteobacterial endosymbiont of the blood-feeding tsetse fly *Glossina brevipalpis*, to gain a more comprehensive understanding of the mutation-selection balance in this bacterial genome. The extremely reduced, 697 kb chromosome of *Wigglesworthia* includes just 621 coding regions (Akman *et al.*, 2002), compared to the 4·6 Mb and nearly 4000 genes in the closely related *E. coli* genome (Blattner *et al.*, 1997). Like many other intracellular bacteria, *Wigglesworthia* shows a heavily biased nucleotide composition (22 % GC; Akman *et al.*, 2002). The tsetse–*Wigglesworthia* symbiosis is obligate and mutual, and it is believed that *Wigglesworthia* plays a role in vitamin synthesis for the host (Nogge, 1981). Tsetse flies cured of *Wigglesworthia* infection experience lower fecundity, although a supplementary diet of B-complex vitamins can restore the host's reproductive ability (Nogge, 1981). The retention of vitamin biosynthetic capabilities in the small *Wigglesworthia* genome may reflect host-level selection to supply these nutrients which are lacking in the tsetse diet of vertebrate blood [analogous to aphid host-level selection on amino acid production by *Buchnera* (Shigenobu *et al.*, 2000)].

Wigglesworthia provides an independent lineage in which to compare the effects of endosymbiosis on the mutation-selection balance in bacteria. Although *Wigglesworthia* is closely related to the relatively well characterized *Buchnera*, the two lineages apparently represent independent transitions to an endosymbiotic lifestyle within the γ-3 *Proteobacteria* (Wernegreen *et al.*, 2003). Shared lifestyle characteristics of *Wigglesworthia* and *Buchnera* include an obligate association with insects, maternal transmission to host offspring and strict cospeciation with their hosts, and may account for the overall similarities between their AT-rich, small genomes. The *Wigglesworthia*–tsetse fly endosymbiosis is estimated to be $\sim$40 million years old (Moran & Wernegreen, 2000) compared to the much older 150–250 million year old *Buchnera*–aphid endosymbiosis (Moran *et al.*, 1993). Here, we test whether these phylogenetically independent endosymbiotic bacteria of different ages experience similar evolutionary processes that affect genome-wide variation in predictable ways and with the same severity.

METHODS

Sequence analysis. The annotated genome sequence of *Wigglesworthia* was downloaded from GenBank (accession nos AB063521–AB063522; November 2002). We excluded hypothetical proteins and genes shorter than 50 codons to minimize random variation in codon or amino acid usage. The final dataset included 564 genes. Because clear distinctions of the mutation-selection balance should be most apparent by comparing high- and low-expression genes (Sharp & Li, 1987), we focused on two gene categories: highly expressed genes in *E. coli* whose orthologues are likely to be highly expressed in *Wigglesworthia*, and lowly expressed *E. coli* genes whose orthologues may be lowly expressed in *Wigglesworthia*. This identification of highly and lowly expressed genes is not without complications. Due to their distinct lifestyles and gene contents, *E. coli* and *Wigglesworthia* may show distinct patterns of gene expression, such that a highly expressed gene in one genome is not necessarily highly expressed in the other. In addition, because the *Wigglesworthia* genome lacks many regulatory functions, the extent of transcriptional differences among genes in the genomes of *Wigglesworthia* and *E. coli* is not necessarily equal. Although gene expression data for *Wigglesworthia* is currently limited, ribosomal proteins are highly expressed in all bacterial species studied thus far (Srivastava & Schlessinger, 1990). Therefore, we included 54 ribosomal proteins in the high-expression gene category for this study. This use of ribosomal proteins as putative high-expression genes is a common practice in genome analyses where experimental expression data are limited or unavailable (de Miranda *et al.*, 2000; Lafay *et al.*, 2000; Palacios & Wernegreen, 2002). The high-expression category also includes the bacterial chaperonin genes *groEL* and *groES* (*mopA* and *mopB*) which have been shown experimentally to be overexpressed in *Wigglesworthia* and other intracellular bacteria (Aksoy, 1995; Baumann *et al.*, 1996). We selected putative low-expression genes in *Wigglesworthia* by first identifying *E. coli* genes with codon adaptation index (CAI) values $<0{\cdot}270$. Since CAI values correlate well with gene expression in *E. coli* (Sharp & Li, 1987), genes with such low CAI values are likely to be lowly expressed in *E. coli*. We then identified orthologues of these *E. coli* genes in the *Wigglesworthia* genome by using a BLAST search (cutoff of $E=10^{-2}$) and the restriction that the gene names must match. This approach identified 17 putative low-expression genes in *Wigglesworthia* (Table 1). These putative low-expression genes included three flagellar genes, *fliO*, *fliP* and *fliQ*, whose presence, along with 28 other flagellar genes, in the *Wigglesworthia* genome suggests the retention of a functional flagellum, although no flagellum or motility has been observed to date (Akman *et al.*, 2002).

Multivariate analyses. We identified major factors shaping variation in relative synonymous codon usage (RSCU) and relative amino acid usage (RAAU) among *Wigglesworthia* genes using correspondence analyses (COA) (Greenacre, 1984) as implemented by CodonW (version 1.3 for UNIX; J. Peden, http://www.molbiol.ox.ac.uk/cu/). We tested for significance of association between the position of loci or amino acids along major axes of the COA with biological variables using non-parametric tests of association (JMP v. 5; SAS Institute). The following properties of loci were plotted against axes 1–4 of COA of RSCU and/or RAAU: nucleotide content at codon positions, hydrophobicity, aromaticity, gene length and codon bias level. Levels of codon bias were estimated in CodonW using the effective number of codons, N_c, an index that ranges from 20 (extreme bias with one codon used per amino acid) to 61 (uniform codon usage) (Wright, 1990). In addition, we tested for a distinction of high- and low-expression gene categories along major axes. We adjusted values of type I error (α) using a Bonferroni correction for multiple tests (Sokal & Rohlf, 1995).

Differences in amino acid usage between high- and low-expression genes. We calculated differences in the RAAU of

Table 1. *Wigglesworthia* genes identified as putatively low-expression in this study based on CAI of *E. coli* orthologues

The gene function corresponds to the annotated genome sequence (Akman *et al.*, 2002).

Gene name	CAI (in *E. coli* orthologue)	Gene function
emrE	0·167	Cellular processes
rnpA	0·231	Protein component; processes tRNA, 4·5S RNA
fliQ	0·235	Flagellar biosynthesis
yaeS	0·235	Biosynthesis of cofactors, prosthetic groups, and carrier
miaA	0·244	Δ(2)-Isopentenylpyrophosphate tRNA-adenosine transferase
fliO	0·246	Flagellar biosynthesis
emrD	0·249	Cellular processes
cdsA	0·252	CDP-diglyceride synthetase
mesJ	0·255	Cell cycle protein
kup	0·256	Transport and binding proteins
ftsL	0·257	Cell division protein
yaeL	0·258	Protein fate
fliP	0·259	Flagellar biosynthesis
phrB	0·262	Deoxyribodipyrimidine photolyase
ubiA	0·262	Biosynthesis of cofactors, prosthetic groups and carrier
birA	0·263	Regulatory functions
dadX	0·267	Cell envelope

putative high- and low-expression genes for each amino acid using the index D{H,L} described previously (Palacios & Wernegreen, 2002). This statistic quantifies the difference in RAAU for each amino acid at high- and low-expression genes. For this calculation, we determined RAAU using General Codon Usage Analysis (McInerney, 1998b). We estimated the significance of D{H,L} values by a randomization test (Sokal & Rohlf, 1995) including 1000 permutations of gene-specific amino acid counts. We designated amino acids as AT-rich, GC-rich or unbiased (neither AT- nor GC-rich) based on the base composition of their associated codons, following accepted conventions (Foster *et al.*, 1997). For brevity, in this paper we indicate amino acids encoded by relatively GC-rich codons as 'GC-rich amino acids', refer to those amino acids encoded by relative AT-rich codons as 'AT-rich amino acids' and call amino acids that are neither GC- nor AT-rich 'unbiased amino acids'.

Estimation of amino acid conservation levels. We explored amino acid differences between *Wigglesworthia* and *E. coli* to help distinguish between two hypotheses to explain an observed abundance of GC-rich amino acids at highly expressed *Wigglesworthia* genes: selection against AT-rich amino acids at high-expression genes or selection for overall maintenance of amino acids since divergence from an ancestor with moderate base composition (Fig. 1). Endosymbionts within the γ-3 *Proteobacteria*, including *Wigglesworthia*, have shifted to extreme AT base compositional bias since their divergence from an ancestor with a relatively moderate base composition (Charles *et al.*, 2001; Heddi *et al.*, 1998). Given the close phylogenetic position of *E. coli*, its moderate base composition of 50·8 % GC (Blattner *et al.*, 1997) and its slow rate of amino acid substitution, all relative to *Wigglesworthia*, many amino acid differences between *E. coli* and *Wigglesworthia* proteins most likely reflect changes in the lineage leading to *Wigglesworthia*. Therefore, if we consider *E. coli* protein sequences a proxy for ancestral states, we can develop a heuristic approach to study the relatively rapid evolution of extremely biased amino acid composition in *Wigglesworthia*.

For sets of 10 high- and 10 low-expression genes, we determined the frequencies of amino acid differences between *E. coli* and *Wigglesworthia* across various base composition categories. For example, we compared all GC-rich amino acids in *E. coli* with their homologous positions in *Wigglesworthia*, which were categorized as one of four states: conserved (the same amino acid); a GC-rich amino acid different from that of *E. coli*; an AT-rich amino acid; or an unbiased amino acid (neither GC- nor AT-rich). Amino acid differences between *E. coli* and *Wigglesworthia* were totalled using MacClade 4.04 (Maddison & Maddison, 2002). Among those amino acids that differed in *E. coli* and *Wigglesworthia*, we compared the configuration of differences in high- and low-expression genes using a χ^2 test. We did not correct for multiple substitutions or the available amino acid types (i.e. there are 5 GC-rich, 6 AT-rich and 9 other amino acids, with related probabilities of substitution to each class), because we were primarily interested in the differences between high- and low-expression genes rather than differences among categories of amino acid changes. We also calculated uncorrected pairwise amino acid divergence using PAUP* 4.0b10 (Swofford, 2002) for the same pairwise alignments with sequence partitions defined by the type of amino acid (GC-rich, AT-rich, unbiased) in *E. coli*.

In addition, we explored the relationship between protein conservation and gene expression by estimating pairwise non-synonymous divergences (d_N) between all *Wigglesworthia* and *E. coli* orthologues using the CODEML program in the PAML version 3.13 software package (Yang, 2002). We tested for a significant difference between the mean d_N values of the 56 putative high-expression genes and the mean d_N values of all other orthologue pairs. We also compared pairwise d_N estimates to CAI values, using the respective CAI value of each *E. coli* gene in all orthologue pairs.

RESULTS

Synonymous codon usage

The GC content of the 564 *Wigglesworthia* genes included in this study ranged from 13 to 37·7 % (mean value of

Possible differences between *E. coli* and *Wigglesworthia* at GC-rich amino acid sites in *E. coli*			Predicted comparisons of relative frequencies between high- and low-expression genes	
			(a)	(b)
E. coli		*Wigglesworthia*	With selection against AT-rich amino acids at high-expression genes	With selection for overall amino acid conservation at high-expression genes
GC-rich amino acids		AT-rich amino acids	**high < low**	**high = low**
		Other GC-rich amino acids	high ≥ low	high = low
		Unbiased amino acids	high ≥ low	high = low

Fig. 1. Expected patterns of amino acid differences between *E. coli* and *Wigglesworthia* under two hypotheses about the nature of selection at high-expression loci. Both hypotheses could explain the relative GC-richness of high-expression genes in *Wigglesworthia*, but yield distinct predictions about the differences between *E. coli* and *Wigglesworthia* proteins. On the left, all possible differences between *E. coli* and *Wigglesworthia*, at homologous amino acid sites that have GC-rich amino acids in *E. coli*. On the right, the predicted configurations of relative frequencies of all possible amino acid differences under the two hypotheses, compared between high- and low-expression genes. (a) Hypothesis 1: selection against changes to AT-rich amino acids at high-expression *Wigglesworthia* genes. (b) Hypothesis 2: overall conservation of amino acids at high-expression *Wigglesworthia* genes, with no selective distinction between AT-rich and GC-rich amino acids. These predictions are used to evaluate the observed relative frequencies shown in Table 4(b).

$23{\cdot}7 \pm 4{\cdot}1$ %) and GC_3 (% GC at third codon position) ranged from 3·8 to 20 % (mean of $9{\cdot}4 \pm 2{\cdot}1$ %), consistent with the extreme AT bias of this genome (Akman *et al.*, 2002). This strong AT bias is reflected in the predominance of synonymous codons ending in A or T, and a low N_c that ranges from 24·7 to 51·4 across the genes sampled (mean of 32·7), rather than the 61 expected under uniform codon usage.

If translational selection influences codon usage in *Wigglesworthia*, then codon bias is expected to differ between high- and low-expression genes, as observed in *E. coli* (Sharp, 1991) and other genomes (Andersson & Kurland, 1990). We tested this prediction by performing COA on the RSCU of 564 *Wigglesworthia* genes, and by testing for distinct codon usage patterns between loci with predicted high- versus low-expression. Axes 1 and 2 account for just 10 and 6 % of the observed variation in RSCU, respectively, distributed across a total of 19 axes. This result indicates little variation among genes in codon usage patterns (Table 2). Genes plotted against the first two axes were distinguished neither by predicted expression levels (Fig. 2a) nor by protein gravy score (hydrophobicity) (Fig. 2b). Rather, a strong correlation of axis one and GC_3 (r_s, Spearman's rank

Table 2. Non-parametric tests of association between the first two axes of COA and multiple genetic and amino acid parameters

COA was performed on both the RSCU and RAAU of 564 *Wigglesworthia* genes. Notable significant relationships are indicated in bold type.

	Variation explained (%)	Correlations (r_s)						
		A_3	T_3	GC_{12}	GC_3	GC	Gravy	Aromo
COA of RSCU								
Axis 1	0·10	0·29*	0·21*	−0·24*	**−0·60***	−0·32*	0·10	0·17*
Axis 2	0·06	0·11	−0·02	−0·15	−0·03	−0·15	0·06	0·06
COA of RAAU								
Axis 1	0·28	−0·22*	−0·58*	**0·95***	0·29*	**0·94***	−0·16*	−0·72*
Axis 2	0·13	0·57*	−0·05	−0·24*	0·00	−0·22*	**−0·75***	−0·32*

*$P < 0{\cdot}0001$, with Bonferroni correction.

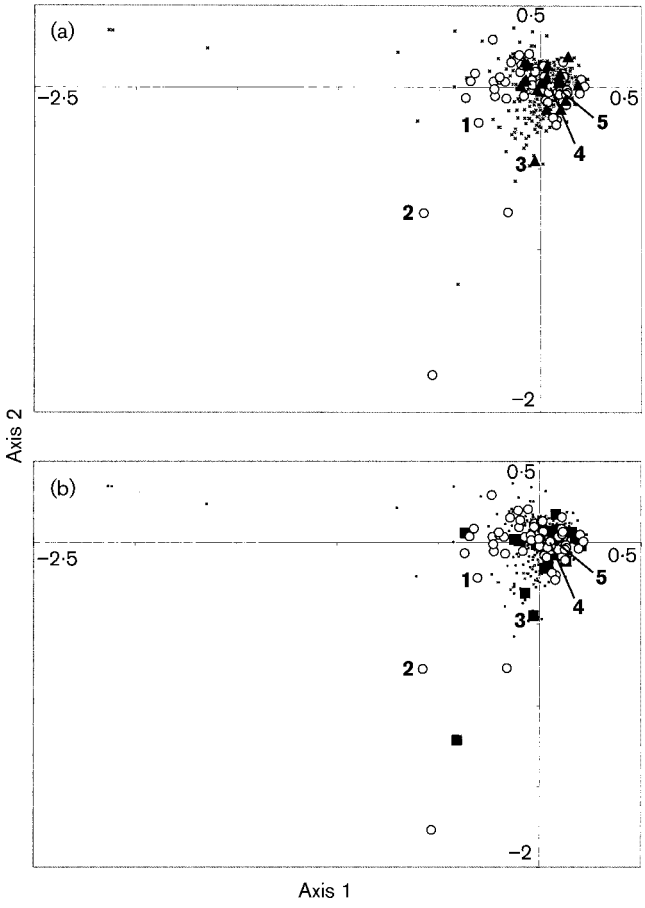

Fig. 2. Position of *Wigglesworthia* genes along the two main axes of variation in the COA of RSCU. (a) The lack of distinction between high-expression genes (open circles) and putative low-expression genes (closed triangles) argues against the hypothesis of adaptive codon bias. (b) Proteins with relatively high levels of hydrophobicity (closed squares) are not distinct from other loci. Due to their potential functional significance, the following genes are labelled by number: 1, *groEL*; 2, *groES*; 3, *fliQ*; 4, *fliP*; 5, *fliO*.

correlation coefficient, $=-0{\cdot}60$, $P<0{\cdot}0001$) suggests that local variation in synonymous base composition drives variation in codon usage patterns. The location of genes along axes 2 and 3 reveals no distinction based on expression level, hydrophobicity, aromaticity or GC content (not shown).

We further explored the effect of mutational bias on codon usage by comparing the relationship of N_c and GC_3 to the pattern expected if variation in local base composition determines variation in codon bias (Wright, 1990). If variation in codon bias is due primarily to base compositional variation across the genome, then N_c values should lie on or just below the expected curve, a pattern that is seen in *Wigglesworthia* (Fig. 3). This pattern contrasts with that expected under translational selection, which predicts N_c

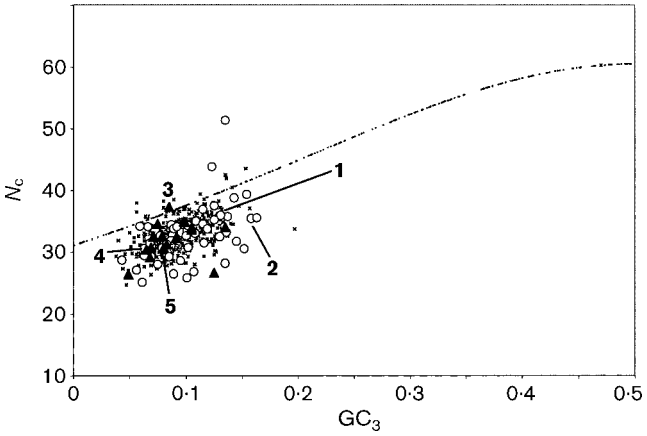

Fig. 3. Relationship between GC_3 and N_c in *Wigglesworthia*, compared to the expected curve if local variation in base composition accounts for variation in codon bias. Putative high-expression genes are marked with open circles and low-expression genes are marked with closed triangles. The plot illustrates the extremely low GC_3 of all *Wigglesworthia* genes, the close approximation of the data to the expected curve and the lack of distinction between putative high- and low-expression genes. Due to their potential functional significance, the following genes are labelled by number: 1, *groEL*; 2, *groES*; 3, *fliQ*; 4, *fliP*; 5, *fliO*.

values considerably lower than those found in the expected curve, especially at high-expression genes.

Amino acid usage

The first two axes of COA of RAAU account for 28 and 13 % of the genome-wide variation in protein amino acid content, respectively, distributed across a total of 20 axes (Table 2). Putative high- and low-expression genes form distinct groups when plotted against these two major axes (Fig. 4a). Axis 1 shows a strong correlation with gene total GC content ($r_s=0{\cdot}94$, $P<0{\cdot}0001$) and with gene GC_{12} (% GC at first and second codon positions) ($r_s=0{\cdot}95$, $P<0{\cdot}0001$) (Fig. 4b), but only a weak correlation with GC_3 ($r_s=0{\cdot}29$, $P<0{\cdot}0001$) (Table 2). This trend reflects the higher GC_{12} content of high-expression genes due to their use of relatively GC-rich codons. Axis 2 is correlated with gene amino acid hydrophobicity as estimated by the gravy score ($r_s=-0{\cdot}75$, $P<0{\cdot}0001$) and distinguishes a group of hydrophobic *Wigglesworthia* proteins from all others (not shown). Gene location on the leading or lagging strands of replication does not show an association with RAAU, and the high- and low-expression genes selected for this study show uniform distributions on the two strands. Across analyses, the three flagellar genes *fliO*, *fliP* and *fliQ* generally group with other putative low-expression genes, and the positions of the chaperonin genes *groEL* and *groES* group with other high-expression (ribosomal) genes selected, thus supporting our expression classifications of these genes in *Wigglesworthia*. The *Wigglesworthia* genes involved in

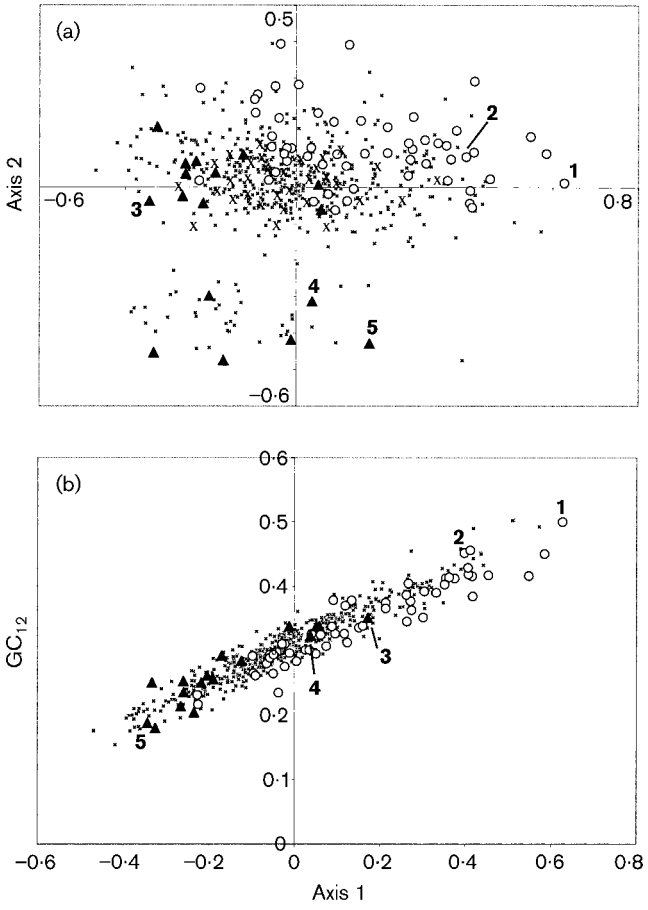

Fig. 4. Position of *Wigglesworthia* genes on the two main axes of COA of RAAU. (a) The distinction of high-expression genes (open circles) from putative low-expression genes (closed triangles) along the first two axes of COA indicates distinct RAAU patterns at high- versus low-expression genes. Genes involved in cofactor biosynthetic pathways, including the production of B-complex vitamins (Akman *et al.*, 2002), are marked with an X. (b) A strong correlation between axis 1 and GC_{12} ($r_s = 0{\cdot}95$, $P < 0{\cdot}0001$) reflects the relative abundance of amino acids encoded by GC-rich codons in high-expression genes. Due to their potential functional significance, the following genes are labelled by number: 1, *groEL*; 2, *groES*; 3, *fliQ*; 4, *fliP*; 5, *fliO*.

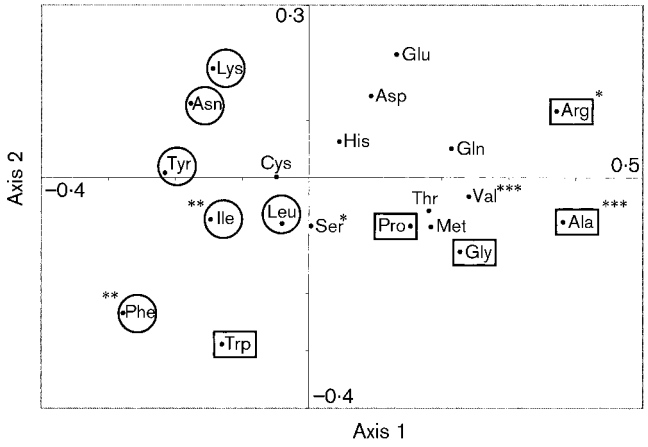

Fig. 5. Position of amino acids on axes 1 and 2 of COA on RAAU in *Wigglesworthia*. Amino acids in rectangles represent GC-rich amino acids and amino acids in circles represent AT-rich amino acids. Asterisks indicate significantly over- or under-represented in putative high-expression genes according to the D{H,L} analysis (*, $0{\cdot}05 \geqslant P > 0{\cdot}01$; **, $0{\cdot}01 \geqslant P > 0{\cdot}001$; ***, $P \leqslant 0{\cdot}001$; randomization test) (see Table 3).

vitamin biosynthesis and considered to have functional roles in symbiosis (Akman *et al.*, 2002) are not differentiated from other genes in the COA. The positions of amino acids along axes 2 and 3 of COA of RAAU highlight those residues that contribute to the observed variation among genes (Fig. 5). The extreme right-hand positions (positive values of axis 1) of GC-rich amino acids Arg and Ala suggests their relative abundance in high-expression genes (also positioned at the extreme positive values of axis 1). Likewise, the far left-hand position (negative value of axis 1) of AT-rich, aromatic Phe reflects its relative abundance in low-expression genes.

The COA results above indicate distinct RAAU at putative high- and low-expression genes. Therefore, we quantified differences in the RAAU of each amino acid separately, using the index D{H,L}. Three relatively GC-rich amino acids, Ala, Arg and Gly, are over-represented in high-expression genes, Ala and Arg significantly so (Table 3, Fig. 5). Four of the six AT-rich amino acids are under-represented in high-expression genes, with Ile and Phe significantly under-represented (Table 3, Fig. 5). These trends resemble those seen in the *Buchnera* genome, in which GC-rich codons are more common in high-expression genes (Palacios & Wernegreen, 2002). D{H,L} analyses performed here for *Wigglesworthia* and reported previously for *Buchnera* show no notable differences, although Trp and Leu are significantly under-represented in *Buchnera* high-expression genes and are not significantly under-represented in *Wigglesworthia* high-expression genes. Only Gln shows an opposite trend in the two genomes (under-represented in high-expression genes of *Wigglesworthia* and over-represented in *Buchnera*) but the D{H,L} value for this amino acid is not significant in either genome.

We compared amino acid differences between *Wigglesworthia* and *E. coli* at high- versus low-expression genes to evaluate alternative hypotheses to account for the relative GC-richness of amino acids at high-expression loci (Fig. 1). The hypothesis of selection against AT-rich amino acids per se predicts that high-expression genes, compared to low-expression genes, will show fewer amino acids that are GC-rich in *E. coli* but AT-rich in *Wigglesworthia* (Fig. 1a). Alternatively, if the relatively GC-rich amino acid profiles of high-expression genes reflects overall conservation of amino acids, then we would expect higher conservation at high-expression *Wigglesworthia* genes, but

Table 3. RAAU in *Wigglesworthia*

D{H,L} values in bold indicate significant over-representation in high-expression genes; D{H,L} values in italics indicate significant under-representation in high-expression genes.

Amino acid	Codon	GC-rich	AT-rich	Aromatic	$RAAU^T$	$RAAU^H$	$RAAU^L$	D{H,L}
Ala	GCN	+			0·0371	0·0482	0·0266	**0·0216***
Arg	CGN; AG(AG)	+			0·0327	0·0607	0·0231	**0·0376**†
Asp	GA(TC)				0·0834	0·0634	0·0789	−0·0155
Asn	AA(TC)		+		0·04	0·0417	0·0296	0·0121
Cys	TG(TC)				0·0129	0·0082	0·0139	−0·0057
Gln	CA(AG)				0·0229	0·0256	0·0157	0·0099
Glu	GA(AG)				0·0508	0·0628	0·0376	0·0252
Gly	GGN	+			0·0528	0·0661	0·0452	0·0209
His	CA(TC)				0·0164	0·0175	0·0131	−0·0044
Ile	AT(TCA)		+		0·1295	0·1053	0·1594	*−0·0541*‡
Leu	CTN; TT(AG)		+		0·0941	0·0818	0·1042	*−0·0224*
Lys	AA(AG)		+		0·1151	0·1229	0·1095	0·0134
Met	ATG				0·0199	0·0251	0·0233	0·0018
Phe	TT(TC)		+	+	0·0541	0·0349	0·0830	*−0·0481*‡
Pro	CCN	+			0·0285	0·0298	0·0268	−0·0030
Ser	TCN; AG(TC)				0·0768	0·0683	0·0779	−0·0096†
Thr	ACN				0·0374	0·0456	0·0392	0·0064
Trp	TGG	+		+	0·0085	0·0049	0·0153	−0·0104
Tyr	TA(TC)		+	+	0·0402	0·0247	0·0417	−0·0170
Val	GTN				0·0429	0·0575	0·0327	**0·0248***

*$P \leqslant 0{\cdot}001$, randomization test.
†$0{\cdot}05 \geqslant P > 0{\cdot}01$.
‡$0{\cdot}01 \geqslant P > 0{\cdot}001$.

similar configurations of amino acid differences at high- and low-expression loci (Fig. 1b).

Our results show lower uncorrected amino acid divergence at high-expression (0·35) than low-expression genes (0·54) (Table 4a), indicating fewer amino acid changes at high-expression loci. Likewise, uncorrected pairwise divergences for all amino acid partition types in *E. coli* (whether an amino acid site in *E. coli* has a GC-rich, AT-rich or unbiased amino acid) are significantly lower in high-expression genes than in low-expression genes [$P < 0{\cdot}05$ for all comparisons, (Student's t distribution)] (Table 4a). The apparent conservation of amino acids that are AT-rich in *E. coli* may be partially explained by the lack of correction for multiple substitutions and the extreme AT bias of the *Wigglesworthia* genome. Consistent with the hypothesis of overall amino acid conservation (Fig. 1b), the GC-rich amino acids of *E. coli* that differ in *Wigglesworthia* show the same relative frequencies of GC-rich, AT-rich and unbiased amino acids in *Wigglesworthia* high- and low-expression genes ($\chi^2 = 0{\cdot}13$, $P = 0{\cdot}94$; Table 4b). This similarity of configurations at high- and low-expression genes is predicted under the model of selection for overall amino acid conservation at high-expression loci, but not under the hypothesis of selection against AT-rich amino acids (Fig. 1a).

DISCUSSION

Analyses of bacterial genomes can provide insights into the balance of mutation, selection and drift in species with varied lifestyles and distinct patterns of genome variation. The extreme AT richness of the *Wigglesworthia* genome suggests this balance has shifted toward stronger mutational bias and/or genetic drift as a result of an endosymbiotic lifestyle. The results of this study show a dominant effect of AT mutational bias on codon usage and amino acid usage across the genome. However, distinct RAAU patterns at high- and low-expression genes indicate that selection is sufficiently strong at these loci to attenuate the effects of a strong mutational pressure.

Synonymous codon usage

Adaptive codon bias shaped by translational selection is characterized by distinct synonymous codon usage at high- and low-expression genes. This type of selection shapes codon usage in many free-living microbial species, where

Table 4. Amino acid differences between orthologous genes of *Wigglesworthia* and *E. coli*

(a) Uncorrected pairwise amino acid divergences between *E. coli* and *Wigglesworthia* categorized by the base composition of codons for amino acids in *E. coli*. All calculations were based on 10 putative high-expression genes (1713 total amino acid sites) and 10 putative low-expression genes (2926 total amino acids sites). $P<0{\cdot}05$ for all comparisons (Student's t distribution). (b) Uncorrected pairwise amino acid divergences between *E. coli* and *Wigglesworthia* categorized by putative gene expression level (high versus low) and the base composition of codons for the amino acid in each species (GC-rich, AT-rich or unbiased, as indicated in Table 3). There are 608 total amino acid differences between *E. coli* and *Wigglesworthia*. $\chi^2=0{\cdot}13$, $P=0{\cdot}94$; among all types of differences at GC-rich amino acid sites.

(a) Pairwise divergence between *E. coli* and *Wigglesworthia* at homologous amino acid sites

***E. coli* amino acids**	**Pairwise amino acid divergence, uncorrected**	
	High-expression genes	**Low-expression genes**
All amino acids	0·35	0·54
AT-rich amino acid sites	0·22	0·37
GC-rich amino acid sites	0·36	0·57
Unbiased amino acid sites	0·43	0·57

(b) Frequencies of amino acid differences between *E. coli* and *Wigglesworthia*

Type of amino acid differences	**High-expression genes**	**Low-expression genes**
GC-rich *E. coli*, GC-rich *Wigglesworthia*	0·07 (14/193)	0·04 (23/538)
GC-rich *E. coli*, AT-rich *Wigglesworthia*	0·46 (89/193)	0·53 (285/538)
GC-rich *E. coli*, unbiased *Wigglesworthia*	0·47 (90/193)	0·42 (230/538)
Percentage due to differences at *E. coli* AT-rich sites	0·16 (97/608)	0·15 (240/1594)
Percentage due to differences at *E. coli* GC-rich sites	0·32 (193/608)	0·34 (538/1594)
Percentage due to differences at *E. coli* unbiased sites	0·52 (318/608)	0·51 (816/1594)

the efficiency of gene translation is related to maximum growth rates (Bulmer, 1991). However, our COA of RSCU in the *Wigglesworthia* genome shows no distinction of high- and low-expression genes, and thus suggests that translational selection has little, if any, effect on codon bias (Fig. 1). Our identification of putative high- and low-expression genes was based on known expression patterns in *E. coli* and, in the case of ribosomal proteins, across all studied bacterial species (Srivastava & Schlessinger, 1990). This criterion has important limitations, since the acquisition of an endosymbiotic lifestyle may alter relative gene expression levels. However, experimental analyses of *Wigglesworthia* proteins show very high expression levels of the chaperonin *groEL* (Aksoy, 1995). Therefore, the observation that *groEL* did not show distinct codon usage compared to other genes, including those of putative lower expression, strengthens evidence against adaptive codon usage in *Wigglesworthia*.

Mutational biases that may shape codon usage include strand-specific biases, in which the two DNA strands experience different configurations of mutations. For example, the relative abundance of Ts and Gs in the leading strand of replication may be explained by strand-specific mutational spectra during replication (Francino & Ochman, 1999; Rocha & Danchin, 2001). Such strand-specific biases can play a dominant role in shaping codon usage [e.g. *Borrelia burgdorferi* (McInerney, 1998a)]. However, we found no relationship between strand orientation and codon usage patterns in *Wigglesworthia*, consistent with the notable lack of strand-specific nucleotide bias, or skew, in this genome (Akman *et al.*, 2002). Rather, the predominant use of A- and T-ending codons across the genome indicates that strong directional AT mutational bias (from GC to AT pairs) drives codon usage in *Wigglesworthia*. In addition, the observed correlation of N_c and GC_3 implies that variation in codon bias among *Wigglesworthia* genes reflects slight differences in local base composition. Directional mutation also drives codon usage in many intracellular bacterial genomes such as the AT-rich genomes of *Buchnera* (Wernegreen & Moran, 1999) and *Rickettsia prowazekii* (Andersson & Sharp, 1996), and the GC-rich genome of *Micrococcus luteus* (Ohama *et al.*, 1990).

Changes in population structure that accompany an endosymbiotic lifestyle may explain the lack of translational selection and strong effects of mutational bias in the *Wigglesworthia* genome. Although the population dynamics of *Wigglesworthia* are currently unclear, these bacteria may experience bottlenecks during transmission to new host offspring through the tsetse milk gland (Ma & Denlinger, 1974), analogous to the bottlenecks experienced

by *Buchnera* when transmitted to aphid eggs or embryos (Mira & Moran, 2002). Such bottlenecks would reduce N_e, increase genetic drift and limit the ability of weak selection to maintain optimal codons in a gene (Bulmer, 1991). In this case, background mutational biases would tend to dominate over selection and eventually shape codon usage. In addition to the potential effects of genetic drift, the substantial loss of DNA repair functions, such as the nucleotide excision repair genes *uvrABC*, may also contribute to the strong mutational biases in *Wigglesworthia* and other endosymbiont genomes. Reduced tRNA gene number may also explain the lack of adaptive codon bias in certain prokaryote genomes, a hypothesis that has been considered in previous studies (Andersson & Sharp, 1996; Palacios & Wernegreen, 2002). Compared to 86 tRNA genes in *E. coli* (Blattner *et al.*, 1997), the reduced genome of *Wigglesworthia* contains just 34 (Akman *et al.*, 2002), resulting in only one or a few tRNA genes per amino acid or codon. However, even those amino acid families with only one corresponding tRNA gene (representing extremely biased tRNA pools) do not show distinct RSCU between high- and low-expression genes (data not shown).

Amino acid usage

Previous studies have demonstrated strong effects of mutational bias on amino acid composition of bacterial proteins (Singer & Hickey, 2000). Among bacterial endosymbionts, strong effects of mutational biases on synonymous codon and amino acid usage have been interpreted as shifts of the mutation-selection balance due to elevated levels of genetic drift (Clark *et al.*, 1998, 2001; Wernegreen & Moran, 1999). The extremely AT-rich *Buchnera* exemplifies this evolutionary phenomenon, as it lacks adaptive codon usage and exhibits a genome-wide increased frequency of AT-rich amino acids (Clark *et al.*, 1999). However, the strength of mutational bias on amino acid usage in *Buchnera* is attenuated in high-expression genes, where selection on amino acid composition is apparently strongest (Palacios & Wernegreen, 2002). Likewise, the full genome sequence of *Wigglesworthia* documented a predominance of amino acids with AT-rich codons (Akman *et al.*, 2002). In this study, we found that *Wigglesworthia* genes of putative high- and low-expression form distinct groups on the first two axes of COA of RAAU (Fig. 3). A major force driving this distinction is the tendency of high-expression genes to use amino acids with relatively GC-rich codons. Of the three amino acids that are over-represented in high-expression genes (i.e. significant D{H,L} values), Ala and Arg are GC-rich while Val is considered unbiased. Likewise, Gly tends to be over-represented at high-expression genes and is also GC-rich. These results suggest that stronger selection on amino acid usage in high-expression genes counters the effects of a genome-wide mutational bias toward AT-rich amino acids.

We considered three possible explanations for the relative GC-richness of high-expression genes. First, this pattern may reflect selection against aromatic and energetically costly amino acids, two of which (Phe and Tyr) are also AT-rich (Akashi & Gojobori, 2002). This hypothesis predicts that aromatic amino acids will be under-represented at high-expression genes. For example, the proteomes of *E. coli* and *Bacillus subtilis* reflect selection to enhance metabolic efficiency by reducing the abundance of aromatic and other energetically expensive amino acids in high-expression genes (Akashi & Gojobori, 2002). This prediction partially holds in *Wigglesworthia*, since the aromatic amino acid Phe is significantly under-represented at high-expression genes (i.e. significantly negative D{H,L}). However, while the aromatic amino acids Trp and Tyr are less frequent in high-expression genes, this result is not significant (Table 3), suggesting that selection against aromaticity does not completely explain the distinct profiles at high-expression genes.

Second, the distinct RAAU patterns may result from selection against the use of AT-rich amino acids at high-expression genes. Consistent with that hypothesis, four of the six AT-rich amino acids (all but Lys and Asn) are under-represented at high-expression genes, and two have significant D{H,L} values. We compared specific amino acid differences between *E. coli* and *Wigglesworthia* to distinguish this hypothesis from a third: that relative GC-richness of high-expression genes reflects overall conservation of amino acids since divergence from a relatively GC-rich ancestor. In this case, high-expression genes are expected to show lower amino acid divergence compared to other loci. We found that amino acids of high-expression genes are significantly more conserved (mean $d_N = 0{\cdot}66$ for high-expression genes, mean $d_N = 1{\cdot}25$ for all genes; difference in means is significantly greater than zero; $P = 5{\cdot}75 \times 10^{-15}$, Student's t distribution of differences). Consistent with this result, we found a significant negative correlation between amino acid divergence and gene expression level as estimated by the *E. coli* CAI (Fig. 6) ($r_s = -0{\cdot}42$, $P < 0{\cdot}0001$). This negative relationship has been shown before in *E. coli* and *S. typhimurium* (Sharp, 1991). Given the moderate 50·8 % GC nucleotide composition of *E. coli* (Blattner *et al.*, 1997), the more frequent use of GC-rich amino acids in *Wigglesworthia* high-expression genes is likely to reflect the maintenance of ancestral amino acid composition, rather than selection against AT-rich amino acids per se.

Indistinguishable configurations of amino acid changes at high- and low-expression genes further supported an overall conservation of amino acids at high-expression genes, rather than selection against AT-rich amino acids. Specifically, we determined if amino acid codons that are GC-rich in *E. coli* are GC-rich, AT-rich or unbiased amino acids in *Wigglesworthia*. If GC-rich amino acids in *Wigglesworthia* high-expression genes reflect selection against AT-rich amino acids, then we would expect to find different configurations at high- and low-expression genes. Namely, we would expect fewer amino acids that are GC-rich in *E. coli* but AT-rich in *Wigglesworthia* at high-expression genes,

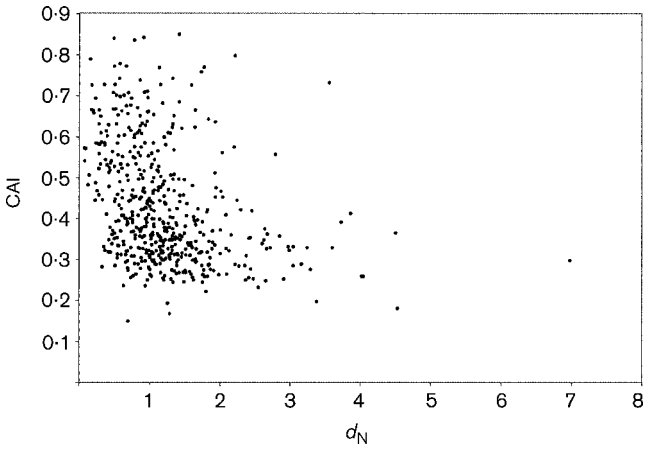

Fig. 6. Correlation between amino acid divergence and gene expression level, as estimated by pairwise d_N for *E. coli* and *Wigglesworthia* orthologues and CAI of the *E. coli* gene ($r_s = -0{\cdot}42$, $P < 0{\cdot}0001$). For this comparison we identified orthologues using the reciprocal smallest distance algorithm (Wall *et al.*, 2003), which estimates corrected evolutionary distances in PAML (Yang, 2002) for multiple, highly significant BLAST hits between genomes (here we used a cutoff of $E = 10^{-5}$). In contrast to BLAST-based homology searches based on local protein sequence alignment, this method uses global maximum-likelihood estimation of evolutionary distances between sequences. This method corroborated the previous *Wigglesworthia* genome annotation, identifying only five genes named differently in the two genomes; these mismatched genes were not among the putative high or low-expression genes used in our analyses.

reflecting selection against changes to AT-rich amino acids, compared to low-expression genes. In contrast, if the relatively GC-rich amino acid profile of high-expression genes of *Wigglesworthia* reflects conservation of amino acids, then we would expect (i) that GC-rich amino acids in *E. coli* are more conserved at high-expression *Wigglesworthia* genes and (ii) among those amino acids that differ in the two species, high- and low-expression genes show similar configurations of changes (Fig. 1). Our results support the latter hypothesis. *Wigglesworthia* high-expression genes show greater conservation of amino acids that are GC-rich in *E. coli* (35 % difference at high-expression genes, 54 % difference at low-expression genes), yet changes at *E. coli* GC-rich amino acid sites had similar configurations in high- and low-expression genes (Table 4). These results support the hypothesis that variation in amino acid usage in *Wigglesworthia* is influenced by stronger selection for overall amino acid conservation at high-expression genes and provide no evidence for stronger selection against the use of AT-rich amino acids at high-expression *Wigglesworthia* genes.

Although the specific mechanism of transmission of *Wigglesworthia* to host offspring is unclear, transmission occurs maternally via the tsetse milk gland (Ma & Denlinger, 1974). The similar genome-wide patterns in *Wigglesworthia* and *Buchnera* observed in this and previous studies suggest that *Wigglesworthia* also experiences genetic drift, perhaps due to repeated bottlenecks associated with transmission through host generations. This may be corroborated as more information on the *Wigglesworthia*–tsetse endosymbiosis becomes available. In addition to small population and genetic drift, other peculiarities of the bacterial intracellular lifestyle may drive extreme genome reduction and strong AT mutational bias. Effects of genetic drift may be enhanced by irreversible gene loss in endosymbionts that lack opportunities and mechanisms for recombination with genetically distinct strains (Moran & Wernegreen, 2000). Previous studies have suggested a universal AT mutational bias, because many types of spontaneous mutations (e.g. the deamination of cytosine) cause GC to AT changes (Birdsell, 2002). The effects of this mutational bias may be more pronounced in small genomes that are deficient in DNA repair and that experience genetic drift (e.g. mitochondria, *Buchnera* and many other small genome bacteria). In addition, a possible relaxation of selection in the intracellular environment compared to free-living existence may allow more rapid gene loss and stronger mutational biases.

Host-level selection might also play a role in *Wigglesworthia* amino acid usage, as *Wigglesworthia* has very limited capability for amino acid biosynthesis but possesses numerous transporters that mediate the acquisition of amino acids from the tsetse host (Akman *et al.*, 2002). Host-level selection on amino acid usage is predicted to occur when the host has certain amino acid deficiencies, whether due to limited biosynthetic capacity or restricted diet. Because the tsetse fly feeds on an amino-acid-rich diet of vertebrate blood, it is not clear which, if any, amino acids could limit host growth and drive such host-level selection. In fact, we found no evidence that host-level selection shapes amino acid usage in *Wigglesworthia*. That is, the distinct amino acid profile of high-expression *Wigglesworthia* genes reflects conservation of these proteins, not a decreased use of rare or energetically expensive amino acids. In striking contrast, *Buchnera* retains biosynthetic capacity for essential amino acids, which it supplies to the aphid host in exchange for non-essential amino acids (Shigenobu *et al.*, 2000). For example, the aphid depends on *Buchnera* to supply Trp and Leu, which occur at particularly low concentrations in the aphids' plant sap diet (Sandström & Moran, 1999). The location of Trp and Leu biosynthetic genes on multicopy plasmids in some *Buchnera* species has been interpreted as host-level selection (Baumann *et al.*, 1999; Lai *et al.*, 1994; Wernegreen & Moran, 2001). The patterns of amino acid usage in *Buchnera* and *Wigglesworthia* are strikingly similar, except for Leu and Trp, which are significantly under-represented in *Buchnera* high-expression genes (Palacios & Wernegreen, 2002), but not significantly under-represented in *Wigglesworthia* high-expression genes.

The relative youth of the *Wigglesworthia* endosymbiosis is corroborated by genetic signatures of parasitism in the *Wigglesworthia* genome, such as maintenance of genes for a flagellum and a more robust cell membrane structure that are lacking in *Buchnera* (Akman *et al.*, 2002). However, the effects of the intracellular lifestyle on genome size and nucleotide and amino acid composition of *Wigglesworthia* have been extreme, as shown previously (Akman *et al.*, 2002) and in this study. Comparisons of diverse *Buchnera* lineages show that severe genome reduction and AT-biased amino acid changes occurred very early in the symbiosis, before the divergence of major aphid subfamilies (Clark *et al.*, 1999; van Ham *et al.*, 2003; Wernegreen *et al.*, 2000). Such rapid genome changes apparently occur in the relatively young *Wigglesworthia* as well, as this endosymbiont also shows severe effects of AT mutational bias on both codon and amino acid usage.

ACKNOWLEDGEMENTS

We gratefully acknowledge Serap Aksoy for helpful discussions about the evolution of the *Wigglesworthia* genome. We thank Seth Bordenstein, Patrick Degnan, Adam Lazarus and two anonymous reviewers for comments on the manuscript. We also thank Eric Harley for the use of his Perl script for data randomization (freely available from the authors), and Carmen Palacios and Maile Neel for assistance with statistical approaches. This work was made possible by support to J. J. W. from the NIH (R01 GM62626-01) and NSF (DEB 0089455), and the Josephine Bay Paul and C. Michael Paul Foundation. Analysis was performed using shared computational facilities funded by the NASA Astrobiology Institute (grant NCC2-1054).

REFERENCES

Abbot, P. & Moran, N. A. (2002). Extremely low levels of genetic polymorphism in endosymbionts (*Buchnera*) of aphids (*Pemphigus*). *Mol Ecol* **11**, 2649–2660.

Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**, 927–935.

Akashi, H. (1997). Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene* **205**, 269–278.

Akashi, H. & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci U S A* **99**, 3695–3700.

Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M. & Aksoy, S. (2002). Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat Genet* **32**, 402–407.

Aksoy, S. (1995). Molecular analysis of the endosymbionts of tsetse flies: 16S rDNA locus and over-expression of a chaperonin. *Insect Mol Biol* **4**, 23–29.

Andersson, S. G. & Kurland, C. G. (1990). Codon preferences in free-living microorganisms. *Microbiol Rev* **54**, 198–210.

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

Baumann, L., Baumann, P. & Clark, M. A. (1996). Levels of *Buchnera aphidicola* chaperonin *groEL* during growth of the aphid *Schizaphis graminum*. *Curr Microbiol* **32**, 279–285.

Baumann, L., Baumann, P., Moran, N. A., Sandstrom, J. & Thao, M. L. (1999). Genetic characterization of plasmids containing genes encoding enzymes of leucine biosynthesis in endosymbionts (*Buchnera*) of aphids. *J Mol Evol* **48**, 77–85.

Bernardi, G. (1985). Codon usage and genome composition. *J Mol Evol* **22**, 363–365.

Birdsell, J. A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol* **19**, 1181–1197.

Blattner, F. R., Plunkett, G. I., Bloch, C. A. & 14 other authors (1997). The complete genome sequence of *Escherichia coli* K12. *Science* **277**, 1453–1474.

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

Charles, H., Heddi, A. & Rahbe, Y. (2001). A putative insect intracellular endosymbiont stem clade, within the *Enterobacteriaceae*, inferred from phylogenetic analysis based on a heterogeneous model of DNA evolution. *C R Acad Sci III* **324**, 489–494.

Clark, M. A., Baumann, L. & Baumann, P. (1998). Sequence analysis of a 34·7-kb DNA segment from the genome of *Buchnera aphidicola* (endosymbiont of aphids) containing *groEL*, *dnaA*, the *atp* operon, *gidA*, and *rho*. *Curr Microbiol* **36**, 158–163.

Clark, M. A., Moran, N. A. & Baumann, P. (1999). Sequence evolution in bacterial endosymbionts having extreme base compositions. *Mol Biol Evol* **16**, 1586–1598.

Clark, M. A., Baumann, L., Thao, M. L., Moran, N. A. & Baumann, P. (2001). Degenerative minimalism in the genome of a psyllid endosymbiont. *J Bacteriol* **183**, 1853–1861.

de Miranda, A. B., Alvarez-Valin, F., Jabbari, K., Degrave, W. M. & Bernardi, G. (2000). Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. *J Mol Evol* **50**, 45–55.

D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C. & Bernardi, G. (1991). Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J Mol Evol* **32**, 504–510.

Foster, P. G., Jermiin, L. S. & Hickey, D. A. (1997). Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol* **44**, 282–288.

Francino, M. P. & Ochman, H. (1999). A comparative genomics approach to DNA asymmetry. *Ann N Y Acad Sci* **870**, 428–431.

Funk, D. J., Wernegreen, J. J. & Moran, N. A. (2001). Intraspecific variation in symbiont genomes: bottlenecks and the aphid–*Buchnera* association. *Genetics* **157**, 477–489.

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

Heddi, A., Charles, H., Khatchadourian, C., Bonnot, G. & Nardon, P. (1998). Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G+C content of an endocytobiotic DNA. *J Mol Evol* **47**, 52–61.

Kreitman, M. & Antezana, M. (1999). *The Population and Evolutionary Genetics of Codon Bias*. Cambridge: Cambridge University Press.

Lafay, B., Atherton, J. C. & Sharp, P. M. (2000). Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology* **146**, 851–860.

Lai, C. Y., Baumann, L. & Baumann, P. (1994). Amplification of *trpEG*: adaptation of *Buchnera aphidicola* to an endosymbiotic association with aphids. *Proc Natl Acad Sci U S A* **91**, 3819–3823.

Ma, W.-C. & Denlinger, D. L. (1974). Secretory discharge and microflora of milk gland in tsetse flies. *Nature* **247**, 301–303.

Maddison, D. & Maddison, W. (2002). *MacClade: Analysis of Phylogeny and Character Evolution*. Sunderland, MA: Sinauer Associates.

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

McInerney, J. O. (1998b). GCUA (General Codon Usage Analysis). *Bionformatics* **14**, 372–373.

Mira, A. & Moran, N. A. (2002). Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. *Microb Ecol* **44**, 137–143.

Moran, N. A. & Wernegreen, J. J. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol Evol* **15**, 321–326.

Moran, N. A., Munson, M. A., Baumann, P. & Ishikawa, H. (1993). A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proc R Soc Lond B Biol Sci* **253**, 167–171.

Nogge, G. (1981). Significance of symbionts for the maintenance of an optional nutritional state for successful reproduction in hematophagous arthropods. *Parasitology* **82**, 101–104.

Ochman, H. & Wilson, A. C. (1987). Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J Mol Evol* **26**, 74–86.

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

Palacios, C. & Wernegreen, J. J. (2002). A strong effect of AT mutational bias on amino acid usage in *Buchnera* is mitigated at high expression genes. *Mol Biol Evol* **19**, 1575–1584.

Powell, J. R. & Moriyama, E. N. (1997). Evolution of codon usage bias in *Drosophila*. *Proc Natl Acad Sci U S A* **94**, 7784–7790.

Rocha, E. P. & Danchin, A. (2001). Ongoing evolution of strand composition in bacterial genomes. *Mol Biol Evol* **18**, 1789–1799.

Sandström, J. & Moran, N. (1999). How nutritionally imbalanced is phloem sap for aphids? *Entomol Exp Appl* **91**, 203–210.

Sharp, P. M. (1991). Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. *J Mol Evol* **33**, 23–33.

Sharp, P. M. & 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.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. & Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407**, 81–86.

Singer, G. A. & Hickey, D. A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* **17**, 1581–1588.

Sokal, R. R. & Rohlf, F. J. (1995). *Biometry*. New York: W. H. Freeman.

Srivastava, A. K. & Schlessinger, D. (1990). Mechanism and regulation of bacterial ribosomal RNA processing. *Annu Rev Microbiol* **44**, 105–129.

Sueoka, N. (1961). Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harb Symp Quant Biol* **26**, 35–43.

Swofford, D. L. (2002). *PAUP**. *Phylogenetic Analysis Using Parsimony* (**and Other Methods*). Sunderland, MA: Sinauer Associates.

van Ham, R. C., Kamerbeek, J., Palacios, C. & 13 other authors (2003). Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci U S A* **100**, 581–586.

Wernegreen, J. J. & Moran, N. A. (1999). Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. *Mol Biol Evol* **16**, 83–97.

Wernegreen, J. J. & Moran, N. A. (2001). Vertical transmission of biosynthetic plasmids in aphid endosymbionts (*Buchnera*). *J Bacteriol* **183**, 785–790.

Wernegreen, J. J., Ochman, H., Jones, I. B. & Moran, N. A. (2000). Decoupling of genome size and sequence divergence in a symbiotic bacterium. *J Bacteriol* **182**, 3867–3869.

Wernegreen, J., Degnan, P., Lazarus, A., Palacios, C. & Bordenstein, S. (2003). Genome evolution in an insect cell: distinct features of an ant–bacterial partnership. *Biol Bull* **204**, 221–231.

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

Yang, Z. (2002). *Phylogenetic Analysis by Maximum Likelihood* (*PAML*), version 3.12. London: University College London.