A Strong Effect of AT Mutational Bias on Amino Acid Usage in *Buchnera* is Mitigated at High-Expression Genes

Carmen Palacios and Jennifer J. Wernegreen

Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory,
Woods Hole, Massachusetts

The advent of full genome sequences provides exceptionally rich data sets to explore molecular and evolutionary mechanisms that shape divergence among and within genomes. In this study, we use multivariate analysis to determine the processes driving genome-wide patterns of amino usage in the obligate endosymbiont *Buchnera* and its close free-living relative *Escherichia coli*. In the AT-rich *Buchnera* genome, the primary source of variation in amino acid usage differentiates high- and low-expression genes. Amino acids of high-expression *Buchnera* genes are generally less aromatic and use relatively GC-rich codons, suggesting that selection against aromatic amino acids and against amino acids with AT-rich codons is stronger in high-expression genes. Selection to maintain hydrophobic amino acids in integral membrane proteins is a primary factor driving protein evolution in *E. coli* but is a secondary factor in *Buchnera*. In *E. coli*, gene expression is a secondary force driving amino acid usage, and a correlation with tRNA abundance suggests that translational selection contributes to this effect. Although this and previous studies demonstrate that AT mutational bias and genetic drift influence amino acid usage in *Buchnera*, this genome-wide analysis argues that selection is sufficient to affect the amino acid content of proteins with different expression and hydropathy levels.

Introduction

The significance of symbiosis is now recognized for its abundance, wide distribution, and fundamental importance in many ecological processes (Douglas 1995). The advent of the molecular techniques has circumvented some of the initial difficulty in studying obligately intracellular, unculturable symbionts. *Buchnera* sp., the obligate endosymbiont located within specialized cells (bacteriocytes) in the body cavity of aphids (McLean and Houk 1973), is a Gram-negative γ-3 Proteobacteria that is closely related to *Escherichia coli* and other Enterobacteriaceae (Unterman, Baumann, and McLean 1989; Munson et al. 1991). Consistent with its strict maternal transmission (Buchner 1965, pp. 297–332 and pp. 640–659), phylogenetic data support cospeciation between *Buchnera* and its aphid hosts dating back to 150–250 MYA, when this association is thought to have originated (Moran et al. 1993). Recently, the genome sequence of *Buchnera* strain APS (Shigenobu et al. 2000) demonstrated that long-term intracellular transmission has dramatically affected the content of this small endosymbiont genome (641 kb, compared with 4.6 Mb of *E. coli* K-12; Blattner et al. 1997).

Vertically transmitted, obligate endosymbionts may have relatively small effective population size (N_e) caused by recurrent bottlenecks upon transmission between host generations (Moran 1996) and limited genetic recombination between endosymbionts of different hosts (Moran and Baumann 1994; Funk et al. 2000; Wernegreen and Moran 2001). Therefore, the efficacy of selection may be reduced in intracellular replicating genomes (Muller 1964; Ohta 1973; Moran 1996) compared with free-living, recombining organisms such as *E. coli*, which are thought to have large long-term N_e (Selander, Caugant, and Whittam 1987).

Analyses of synonymous codon usage and amino acid composition are useful tools to explore shifts in the mutation-selection balance across bacterial species with different lifestyles. For example, in contrast to adaptive codon usage in *E. coli* and other free-living bacteria (Ikemura 1981; Bennetzen and Hall 1982), synonymous codon usage of intracellular pathogens such as *Mycoplasma genitalium* and *Rickettsia prowazekii* corresponds with local base compositional biases, and selection seems to have little effect (Andersson and Sharp 1996; McInerney 1997). Likewise, *Buchnera* shows an extreme AT bias at synonymous codon positions and spacer regions (Shigenobu et al. 2000) and lacks the adaptive codon bias shown by *E. coli* (Wernegreen and Moran 1999). Analysis of several protein-coding genes in this endosymbiont shows that mutational bias and drift drive not only codon usage but also amino acid changes (Ohtaka and Ishikawa 1993; Moran 1996; Brynnel et al. 1998; Clark, Baumann, and Baumann 1998; Clark, Moran, and Baumann 1999) and may contribute to gene loss (Mira, Ochman, and Moran 2001; Silva, Latorre, and Moya 2001).

Recently, full genome sequence data have strengthened multivariate analyses to explore factors that drive variation in amino acid and codon usage among genes within a given genome (e.g., de Miranda et al. 2000; Romero, Zavala, and Musto 2000). In this article, we build upon previous genome-level studies by employing multivariate analysis to identify major factors shaping amino acid usage in the full genomes of *Buchnera* APS and *E. coli* K-12. Our results argue that mutation and selection have strong but distinct effects on protein evolution in these two bacterial species.

Abbreviations: N_e, effective population size; COA, correspondence analysis; RAAU, relative amino acid usage; D{H,L}, difference in RAAU of an amino acid at high- and low-expression genes; r_s, Spearman's Rho coefficient; RSCU, relative synonymous codon usage.

Key words: amino acid usage, selection, endosymbiosis, mutational bias, multivariate analysis, gene expression, *Escherichia coli*.

Address for correspondence and reprints: Jennifer J. Wernegreen, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, 7 MBL Street, Woods Hole, Massachusetts 02543. E-mail: jwernegreen@mbl.edu.

Mol. Biol. Evol. 19(9):1575–1584. 2002

Table 1
***Buchnera* Relative Amino Acid Usage**

		GC-		AT-					D{H,L}	D{H,L}
AA	Codons[a]	rich[b]	Aromatic[b]	rich[b]	RAAU T[c]	RAAU H[c]	RAAU L[c]	D{H,L}[d]	Lagging[e]	Leading[e]
Ala	GCN	+			0.0455	0.0594	0.0304	0.0290***	0.0298	0.0248
Arg	CGN; AG[AG]	+			0.0389	0.0732	0.0400	0.0332***	0.0189	0.0348
Asn	GA[TC]			+	0.0711	0.0521	0.0763	−0.0242**	−0.0194	−0.0203
Asp	AA[TC]				0.0440	0.0371	0.0347	0.0024	**−0.0058**	**0.0050**
Cys	TG[TC]				0.0118	0.0065	0.0104	−0.0039	**−0.0079**	**0.0012**
Gln	CA[AG]				0.0323	0.0308	0.0347	−0.0040	−0.0067	−0.0052
Glu	GA[AG]				0.0564	0.0572	0.0451	0.0121	0.0114	0.0128
Gly	GGN	+			0.0553	0.0721	0.0429	0.0291**	0.0291	0.0242
His	CA[TC]				0.0212	0.0218	0.0218	0.0000	**−0.0020**	**0.0019**
Ile	AT[TCA]			+	0.1143	0.0955	0.1183	−0.0227**	−0.0038	−0.0299
Leu	CTN; TT[AG]			+	0.0985	0.0795	0.1202	−0.0407***	−0.0474	−0.0413
Lys	AA[AG]			+	0.0988	0.1143	0.0994	0.0149	0.0304	0.0244
Met	ATG				0.0212	0.0257	0.0243	0.0014	**0.0093**	**−0.0030**
Phe	TT[TC]		+	+	0.0485	0.0311	0.0588	−0.0277***	−0.0237	−0.0269
Pro	CCN	+			0.0303	0.0285	0.0329	−0.0044	−0.0092	−0.0021
Ser	TCN; AG[TC]				0.0722	0.0666	0.0688	−0.0022	**0.0019**	**−0.0074**
Thr	ACN				0.0459	0.0515	0.0455	0.0060	**0.0193**	**−0.0065**
Trp	TGG	+	+		0.0090	0.0052	0.0139	−0.0087***	−0.0155	−0.0018
Tyr	TA[TC]		+	+	0.0358	0.0220	0.0380	−0.0160**	−0.0256	−0.0043
Val	GTN				0.0491	0.0697	0.0433	0.0264**	0.0169	0.0196

[a] Codons used by each amino acid; degenerate nucleotides are given in brackets.

[b] GC-rich amino acid codons, aromatic amino acids, and AT-rich amino acid codons are indicated by "+".

[c] RAAU of total data set (T), the fifty-four high-expression genes (H) and nineteen low-expression genes (L) considered in this article for each species.

[d] Subtraction of RAAU of high- and low-expression genes and its significance by means of a randomization test: * $0.05 \geq P > 0.01$, ** $0.01 \geq P > 0.001$, and *** $P \leq 0.001$.

[e] Subtraction of RAAU values when considering separately those genes situated in lagging (D{H,L} lagging) and leading (D{H,L} leading) strands of replication. Bold values indicate a change in sign when the statistic is calculated separately for leading and lagging strands.

Table 2
***Escherichia coli* Relative Amino Acid Usage**

		GC-		AT-					D{H,L}	D{H,L}
AA	Codons[a]	rich[b]	Aromatic[b]	rich[b]	RAAU T[c]	RAAU H[c]	RAAU L[c]	D{H,L}[d]	Lagging[e]	Leading[e]
Ala	GCN	+			0.0968	0.1095	0.0973	0.0122	0.0630	0.0064
Arg. . . .	CGN; AG[AG]	+			0.0552	0.0825	0.0669	0.0156*	0.0451	0.0110
Asn . . .	GA[TC]			+	0.0391	0.0356	0.0341	0.0015	0.0035	0.0058
Asp . . .	AA[TC]				0.0512	0.0466	0.0443	0.0023	**−0.0125**	**0.0028**
Cys. . . .	TG[TC]				0.0113	0.0047	0.0099	−0.0052**	−0.0070	−0.0044
Gln. . . .	CA[AG]				0.0440	0.0319	0.0520	−0.0202***	−0.0105	−0.0218
Glu. . . .	GA[AG]				0.0576	0.0676	0.0602	0.0075	**0.0212**	**−0.0003**
Gly. . . .	GGN	+			0.0754	0.0830	0.0679	0.0151	**−0.0294**	**0.0116**
His	CA[TC]				0.0224	0.0194	0.0252	−0.0058	**0.0063**	**−0.0090**
Ile.	AT[TCA]			+	0.0601	0.0562	0.0564	−0.0002	0.0002	0.0020
Leu. . . .	CTN; TT[AG]			+	0.1060	0.0708	0.1294	−0.0587***	−0.0899	−0.0488
Lys. . . .	AA[AG]			+	0.0437	0.0931	0.0334	0.0598***	0.0989	0.0604
Met . . .	ATG				0.0281	0.0256	0.0284	−0.0028	−0.0042	−0.0001
Phe. . . .	TT[TC]		+	+	0.0385	0.0284	0.0345	−0.0062	−0.0026	−0.0035
Pro	CCN	+			0.0442	0.0315	0.0496	−0.0182***	−0.0198	−0.0167
Ser	TCN; AG[TC]				0.0576	0.0438	0.0512	−0.0074	−0.0175	−0.0085
Thr. . . .	ACN				0.0542	0.0516	0.0498	0.0018	**−0.0243**	**0.0043**
Trp	TGG	+	+		0.0147	0.0052	0.0183	−0.0131***	−0.0164	−0.0143
Tyr	TA[TC]		+	+	0.0280	0.0176	0.0204	−0.0029	**0.0010**	**−0.0013**
Val	GTN				0.0720	0.0956	0.0707	0.0249**	**−0.0050**	**0.0243**

[a] Codons used by each amino acid; degenerate nucleotides are given in brackets.

[b] GC-rich amino acid codons, aromatic amino acids, and AT-rich amino acid codons are indicated by "+".

[c] RAAU of total data set (T), the fifty-four high-expression genes (H) and nineteen low-expression genes (L) considered in this article for each species.

[d] Subtraction of RAAU of high- and low-expression genes and its significance by means of a randomization test: * $0.05 \geq P > 0.01$, ** $0.01 \geq P > 0.001$, and *** $P \leq 0.001$.

[e] Subtraction of RAAU values when considering separately those genes situated in lagging (D{H,L} lagging) and leading (D{H,L} leading) strands of replication. Bold values indicate a change in sign when the statistic is calculated separately for leading and lagging strands.

Table 3
Low-Expression *Buchnera* Genes Considered in this Study

Gene	CAI[a]	Putative Gene Product
gloB	0.193	Hydroxyacylglutathione hydrolase
flgA	0.197	Flagella basal body P-ring formation protein FlgA precursor
rnpA	0.231	Ribonuclease P protein component
fliQ	0.235	Flagellar biosynthetic protein FliQ
bolA	0.244	BolA protein
miaA	0.244	tRNA delta(2)-isopentenylpyrophosphate transferase
ilvH	0.245	Acetolactate synthase small subunit
cls	0.255	Cardiolipin synthetase
mesJ	0.255	Cell cycle protein MesJ
ftsL	0.257	Cell division protein FtsL
holA	0.258	DNA polymerase III delta subunit
fliP	0.259	Flagellar biosynthetic protein FliP
mltE	0.259	Membrane-bound lytic murein transglycosylase E
fpr	0.262	Ferredoxin-NADP reductase
phrB	0.262	Deoxyribodipyrimidine photolyase
trmD	0.263	tRNA (guanine-n1)-methyltransferase
cysC	0.264	Adenylylsulfate kinase
flhA	0.264	Flagellar biosynthetic protein FlhA
lipB	0.265	Lipoate-protein ligase B

[a] CAI value of *Escherichia coli* homologous protein.

Material and Methods

Genome Sequence Data

Coding sequences were extracted from the complete genome sequences of *Buchnera* sp. APS (chromosome and two plasmids; Shigenobu et al. 2000) and *E. coli* K-12 (Blattner et al. 1997) available at GeneBank (June 2001). Hypothetical proteins and annotated genes with less than 50 codons were excluded from the analysis to reduce stochastic variation (as recommended by CodonW tutorial; see below), resulting in a final sample of 479 loci from *Buchnera* and 2,919 loci from *E. coli*.

Multivariate Analysis of Amino Acid Composition

We used correspondence analysis (COA, Greenacre 1984) to identify the major factors that shape variation in amino acid usage among proteins of *Buchnera* and *E. coli*, as implemented by CodonW v. 1.4.2 for UNIX (available with John Peden at http://www.molbiol.ox.ac.uk/cu/). Because COA vectors may be affected by unusual amino acid usage of *Buchnera* plasmid-encoded proteins, plasmid genes were added after COA, and their positions were calculated based on vectors obtained from nuclear genes only. Major axes did not distinguish plasmid from nuclear genes of *Buchnera*.

Identifying Sources of Trends in Amino Acid Usage

We used nonparametric tests of association to test the significance of associations between the position of loci (or amino acids) on principal axes of COA and 39 parameters relating to properties of loci or amino acids (JMP v. 4; SAS Institute). Given that multiple tests were performed, we adjusted values of type I error (α) by means of the Bonferroni correction (Sokal and Rohlf 1995, pp. 236–240). Parameters of loci included several measures of nucleotide composition (e.g., AT skew defined by $\{A - T/A + T\}$, base composition at each codon position, etc.), gene length, relative frequency of aromatic amino acids, and the overall hydropathicity score of a protein (GRAVY; Kyte and Doolittle 1982). Properties of amino acids included molecular weight, hydropathy level (i.e., degree of hydrophilicity or hydrophobicity), and AT-richness of codons scored as in Clark, Moran, and Baumann (1999). In this study, we selected the four amino acids at each extreme of that scale to define ‘‘GC-rich’’ and ‘‘AT-rich’’ categories (tables 1 and 2). We also included Leu in the AT-rich category, as this amino acid is encoded by TT[AG] and CTN.

Comparing High- Versus Low-Expression Genes

In *E. coli*, gene expression levels correlate closely with the codon adaptation index (CAI) (Sharp and Li 1987). But gene expression data in *Buchnera* are scarce. For high-expression genes, we selected the fifty-two ribosomal proteins, which are highly expressed across diverse taxa (Srivastava and Schlessinger 1990), *mopA* (*groEL*), which is highly expressed in *Buchnera* (Ishikawa 1984) and several other intracellular bacteria (e.g., the tsetse fly endosymbiont *Wigglesworthia*; Aksoy 1995) and its cotranscribed product *mopB* (*groES;* Sato and Ishikawa 1997). For low-expression genes, we selected 20 *Buchnera* loci that are homologous and named identically to the 639 *E. coli* genes with CAI values lower than 0.269 (table 3). From this pool of low-expression genes we omitted *ilvH* (acetolactate synthase involved in isoleucine and valine biosynthesis), a biosynthetic gene that may be highly expressed in this nutritional symbiont.

Comparisons of amino acid usage at high- and low-expression genes in *Buchnera* and *E. coli* were limited to the same set of homologous loci in both genomes. Homologs were identified conservatively as proteins with the highest degree of similarity and named identically in the two genomes. We can assume that pairs of homologs in *Buchnera* and *E. coli* are also orthologous because

Buchnera lacks any species-specific duplicated proteins (with the exception of *grpE*, which was not included in comparisons of high- and low-expression genes) (Shigenobu et al. 2001). We quantified differences in relative amino acid usage (RAAU) between high- and low-expression genes by developing a new statistic, D{H,L} (the difference in RAAU of an amino acid at high- and low-expression genes). This statistic is defined as the RAAU at high-expression genes in a given genome minus the RAAU at low-expression genes in that genome and was calculated separately for each amino acid. We tested the significance of D{H,L} values using a sampled permutation (randomization) test (Sokal and Rohlf 1995, pp. 803–819) and performed 1,000 permutations using a perl program kindly provided by E. T. Harley (http://www.cs.hmc.edu/∼eharley/research/tools/).

Locating Genes Situated on Leading Versus Lagging Strands of Replication

Asymmetrical mutational bias between the two complementary DNA strands may contribute to variation in both codon and amino acid usage (Karlin, Campbell, and Mrazek 1998; McInerney 1998; Lafay et al. 1999). To consider effects of leading versus lagging strands of replication, we determined strand orientation of genes on the basis of their position relative to the origin and terminus of replication. The presence of the DnaA-box of the *Buchnera* genome is thought to mark the origin of replication of the *Buchnera* chromosome (Shigenobu et al. 2000). But a shift of the GC skew in noncoding and synonymous third codon positions 13 kb upstream of this DnaA-box (Shigenobu et al. 2000) may correlate with the origin, as shown for other bacterial genomes (Lobry 1996; Blattner et al. 1997). To account for ambiguity in the location of the *Buchnera* origin, we considered this 13-kb region (from 627,681 to position 1 of the sequenced genome) as an ‘‘origin window.’’ Likewise, we defined the ‘‘terminus window’’ as the 13-kb region immediately opposite (180 degrees) the origin region (307,340 to 320,340). We excluded genes in these windows from comparisons of leading versus lagging strands. In contrast, the origin and terminus of *E. coli* are well defined experimentally (e.g., Yoshikawa and Ogasawara 1991) so that all genes can be assigned to the leading or lagging strand.

Results

Multivariate Analysis of Amino Acid Usage in *Buchnera*

Four of the nineteen axes generated by COA of *Buchnera* account for more than 50% of the total variance in amino acid composition of proteins.

Axis 1

The first axis accounts for 23.0% of the total variation of the data. This axis correlates positively with GC content at first and second codon positions (r_s (Spearman’s Rho coefficient) $= 0.89$, $P < 0.0001$; 0.78, $P < 0.0001$; respectively) but notably, not third codon positions. Axis 1 also correlates negatively with aromaticity levels of each protein ($r_s = -0.67$, $P < 0.0001$) and differentiates putative high- and low-expression genes in *Buchnera* (fig. 1*a*).

Axis 2

The second axis of COA accounts for 13.3% of the variance. This axis correlates with the global levels of hydropathy of each *Buchnera* protein ($r_s = 0.70$, $P < 0.0001$) and separates a group of presumed membrane proteins (with high GRAVY scores) from all other loci (data not shown). Axis 2 correlates negatively with AT skew ($r_s = -0.81$, $P < 0.0001$) because of the nucleotide composition of codons for amino acids situated at the extreme of this axis (fig. 1*b*, e.g., the hydrophobic Phe TT[T or C] vs. the hydrophilic Lys AA[A or G]).

Other Axes

The third and fourth axes of COA in *Buchnera* account for 8.1% and 6.3% of the variation in the data, respectively. Axis 3 does not correlate significantly with any parameter considered. The fourth axis separates proteins that are rich in Cys, the most rare amino acid in *Buchnera* (table 1). The low dispersion observed in these and subsequent axes did not warrant further consideration.

Multivariate Analysis of Amino Acid Usage in *E. coli*

The first four axes of COA of the complete genome sequence of *E. coli* K-12 explain 49.2% of the total variation of the data (distributed along the 19 total axes). Axis 1 (19.1% of the total variation) correlates positively with the GRAVY score of proteins ($r_s = 0.83$, $P < 0.0001$) and negatively with AT skew ($r_s = -0.67$, $P < 0.0001$). This result agrees with a previous multivariate analysis of 999 *E. coli* genes, in which integral membrane proteins (IMP’s) form a distinct group along Axis 1 (Lobry and Gautier 1994). Axis 2 (12.4% of the total variation) correlates with the gene expression (approximated using CAI) significantly ($r_s = 0.36$, $P < 0.0001$) but not as strongly as the correlation previously reported ($r_s = 0.55$, $P < 0.0001$; Lobry and Gautier 1994). A high correlation with C and A content at first codon positions ($r_s = -0.79$, $P < 0.0001$; and 0.68, $P < 0.0001$, respectively) does not coincide with that expected from the correlation with gene expression, i.e., excess of guanine at first codon position (Gutierrez, Marquez, and Marin 1996). Axis 3 (9.4% of the variation) correlates positively with aromaticity and correlates with CAI almost as well as does Axis 2 ($r_s = -0.31$, $P < 0.0001$). Both Axes 2 and 3 differentiate high- and low-expression genes considered in this study for *E. coli* (data not shown) as predicted by their correlations with CAI. As in *Buchnera*, the distinction of proteins rich in Cys (also the most rare amino acid in *E. coli*; table 2) along Axis 4 (8.3% of the variation) suggests that the frequency of Cys is highly variable among loci.

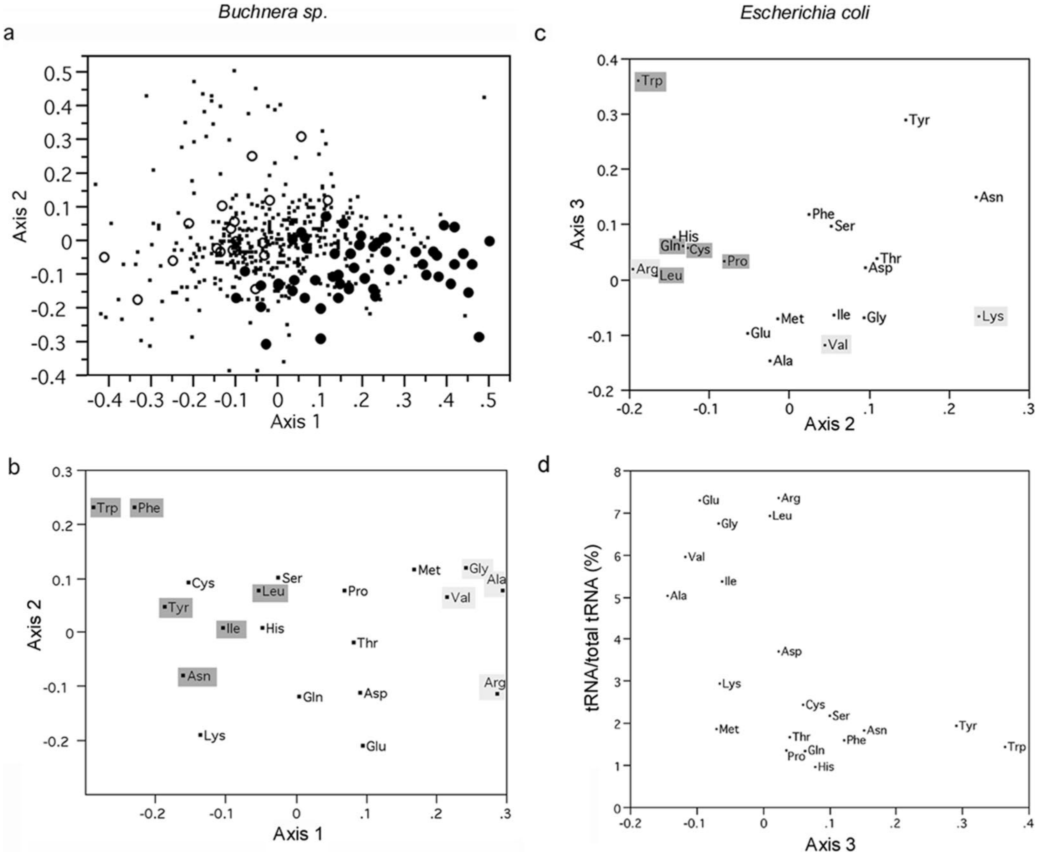

FIG. 1.—*a*, Positions of putative high- and low-expression genes (closed and open circles, respectively) considered in this study according to their location in the two main axes of COA of amino acid usage in *Buchnera*. *b*, Position of amino acids along the two main axes of COA of amino acid usage in *Buchnera*. Shaded amino acids have significantly different RAAU at high- and low-expression genes (refer to table 1 for significance values). Amino acids that are significantly overrepresented in high-expression genes (marked with light shadow) use relatively GC-rich codons with the exception of Val. Amino acids that are significantly underrepresented in high-expression *Buchnera* genes (dark shadow) tend to be aromatic and use relatively AT-rich codons. *c*, Position of amino acids along Axes 2 and 3 of COA of amino acid usage in *E. coli*. Amino acids that are significantly overrepresented in high-expression genes (see table 2) are marked with light shadows. Those significantly underrepresented in high-expression genes are marked with dark shadows. *d*, Position of amino acids along Axis 3 of COA of *E. coli*, plotted against frequencies of major tRNA molecules per cell according to Dong, Nilsson and Kurland (1996).

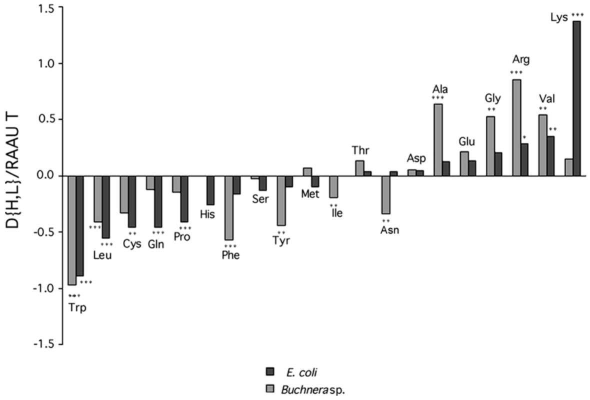

FIG. 2.—Difference in RAAU at high- and low-expression genes (D{H,L}) for *Buchnera* and *E. coli*. For comparative purposes, differences were divided by RAAU of all genes considered for each genome, respectively (“RAAU T,” see values in tables 1 and 2). Amino acids are ordered by *E. coli* values for the normalized statistic. Asterisks indicate significant values of D{H,L} according to a permutation test (tables 1 and 2).

Comparative Analysis of Major Trends Shaping Amino Acid Usage in *Buchnera* and *E. coli* *Gene Expression*

To test for differences in amino acid usage of high- and low-expression genes, we calculated the D{H,L} for each amino acid and determined the significance of observed values on the basis of the simulated null distribution of this statistic (tables 1 and 2). Amino acids that are significantly overrepresented in putative high-expression *Buchnera* genes are generally encoded by GC-rich codons (Arg, Ala, and Gly, with the exception of Val) and are not aromatic. Comparisons of *E. coli* and *Buchnera* indicate that most amino acids show similar trends in both species (i.e., are either over- or underrepresented in high-expression genes of both genomes) but to different degrees (fig. 2). No amino acid is significantly overrepresented in highly expressed genes of one species but significantly underrepresented in the other. Only two amino acids, Met and Asn, show opposite (but not significant) trends in *Buchnera* and *E. coli*. Notably, the aromatic amino acid Trp is severely reduced in high-expression genes of both species. The two other aromatic amino acids, Tyr and Phe (both of which are encoded by AT-rich codons), are significantly underrepresented in putative high-expression genes of *Buchnera* but not significantly in *E. coli*. Interestingly, Asn and Ile are relatively rare in the high-expression genes of *Buchnera* but show no strong bias in *E. coli*. Both Asn and Ile are encoded by AT-rich codons and are not aromatic. This pattern suggests selection against AT-rich amino acids at high-expression genes of *Buchnera* that is independent of selection against aromatic amino acids.

Because Axis 1 of COA in *Buchnera* distinguishes putative high- and low-expression genes, it is not surprising that D{H,L} values coincide with the positions of amino acids along Axis 1 (fig. 1*b*). That is, those amino acids that are overrepresented in putative high-expression *Buchnera* genes (i.e., D{H,L} > 0; Val, Arg, Gly, and Ala; fig. 2) are positioned at high values along Axis 1 (fig. 1*b*) (as are the high-expression genes; fig. 1*a*). Likewise, amino acids overrepresented in low-expression *Buchnera* genes (i.e., with D{H,L} < 0; Trp, Leu, Phe, Tyr, Ile, and Asn) are positioned at low values of Axis 1 (as are the low-expression genes).

In *E. coli*, D{H,L} values only partially account for variation at Axes 2 and 3, both of which correlate with gene expression. Amino acids that are underrepresented in high-expression genes (Trp, Leu, Cys, Gln, and Pro; fig. 2) appear at the extreme of Axis 2, but only Trp is extreme in Axis 3 (fig. 1*c*). Amino acids that are overrepresented in high-expression *E. coli* genes (Lys, Val, and Arg) are situated at the extreme of Axis 3, but only Lys is extreme in Axis 2.

Strand of Replication

Our COA of amino acid usage across the full genomes of *Buchnera* and *E. coli* did not clearly distinguish genes on the leading and lagging strands of replication in either species. But strand orientation and gene expression levels are not independent. As noted previously for *E. coli* and several other bacterial species (Francino and Ochman 1999), we found that putative high-expression genes tend to occur on the leading strand (78% of *Buchnera* and 96% of *E. coli* genes sampled here), perhaps because of selection to avoid collision between DNA and RNA polymerases (Brewer 1988).

Thus, we tested whether prevalence of high-expression genes on the leading strand, coupled with strand-specific mutational asymmetries, could account for the distinct amino acid profiles we observed at high- and low-expression genes. To distinguish the effects of strand orientation and gene expression level, we compared D{H,L} values calculated *separately* for genes on leading and lagging strands with D{H,L} values obtained when strand orientation was not considered (tables 1 and 2). In general, strand position did not affect the sign of significant D{H,L} values (i.e., had no effect on whether an amino acid is over- or underrepresented in high-expression genes). In *Buchnera*, switches in the sign of D{H,L} occurred only when this value was not significant; therefore, these sign changes may be attributed to random variation. The same result was found in *E. coli*, with the exception of Val (GTN) which is more abundant on the leading strand (see *Discussion*).

Hydrophobicity

We further explored the effect of hydrophobicity on amino acid usage in *Buchnera* and *E. coli* by comparing the inferred functions of genes at extreme positions of the axes that correlate with the hydropathy of proteins (i.e., loci positioned at >0.20 on Axis 2 of *Buchnera* and loci at >0.20 on Axis 1 for *E. coli*). In both genomes, these hydrophobic proteins tend to function as IMPs, with functions such as transport and anchoring of dehydrogenases. All but two *Buchnera* genes positioned >0.20 on Axis 2 were homologous to *E. coli* genes at extreme of Axis 1. These two exceptions, *znuB* and *secY*, also encode IMPs involved in transport. Moreover, *secY* is listed among *E. coli* genes at extreme of Axis 1 in a previous COA of this species (Lobry and Gautier 1994) but is absent from *E. coli* K-12.

Discussion

Shift in the Mutation-Selection Balance Contributes to Distinct Amino Acid Usage in Buchnera Versus *E. coli*

Previous studies demonstrate a strong effect of AT mutational bias on amino acid changes along *Buchnera* lineages, especially early in the symbiosis with aphids (Clark, Moran, and Baumann 1999). Virtually all *Buchnera* proteins are strongly influenced by directional mutational bias, as demonstrated for several other bacterial genomes (Singer and Hickey 2000). But we have addressed a distinct question: what processes drive variation in amino acid usage *among* loci within the *Buchnera* genome? The results of this study argue that selection relating to gene expression contributes to intragenomic variation in amino acid usage in this endosymbiont.

The first axis of the COA clearly distinguishes putative high- and low-expression *Buchnera* genes. To confirm this strong effect of gene expression on amino acid usage, we compared the RAAU of each amino acid at high- and low-expression genes using the new statistic D{H,L}. Several amino acids show significant differences in their abundance at putative high- and low-expression *Buchnera* genes. Amino acids significantly overrepresented at high-expression *Buchnera* genes include Ala, Gly, Arg, and Val, whereas those significantly underrepresented include Trp, Leu, Phe, Tyr, Ile, and Asn (table 1 and fig. 2).

This distinct amino acid profile of high-expression *Buchnera* loci may be shaped by selection against aromatic amino acids (Trp, Phe, and Tyr), which are expensive to biosynthesize (Craig and Weber 1998; Akashi and Gojobori 2002). In addition, amino acids that are more abundant at high-expression *Buchnera* genes tend to use codons that are relatively GC-rich at first and second positions (but not at third positions). This relative GC-richness suggests that selection counteracting a genome-wide AT mutational pressure in *Buchnera* is stronger at high-expression genes compared with low-expression genes.

Because many aromatic amino acids are encoded by AT-rich codons, selection against aromatic residues and selection against AT-rich codons are complementary and overlapping. But results of this study show their independent effects. For example, the amino acids Asn and Ile are encoded by AT-rich codons but are not aromatic. Thus, the low frequencies of Asn and Ile in high-expression *Buchnera* genes argues for selection against AT-rich codons at high-expression genes that cannot be explained by selection against aromaticity. Likewise, the low frequency of the very aromatic Trp (encoded by TGG) in high-expression genes suggests selection against aromatic amino acids that cannot be explained by selection against AT-rich codons.

The specific function of *Buchnera* as a nutritional endosymbiont may influence certain patterns of amino acid usage in this genome. Interestingly, the essential amino acids Trp and Leu are generally overproduced by *Buchnera* to supplement its host diet (Douglas and Prosser 1992; Bracho et al. 1995; Baumann et al. 1998) and might be relatively abundant amino acids in the endosymbiont cell. Therefore, the paucity of Trp and Leu in high-expression *Buchnera* genes may be shaped by host-level selection for energetic efficiency (see Rispe and Moran 2000 for models of host and symbiont-level selection). In addition, host-level selection may also influence the usage of several nonessential amino acids that *Buchnera* cannot synthesize but must acquire from the aphid host (Shigenobu et al. 2000).

Although strong AT bias in *Buchnera* may drive distinct amino acid usage at high- and low-expression genes, this mutational bias may actually narrow that difference in some cases. Notably, Lys, encoded by the AT-rich codons AA[AG], is significantly overrepresented in ribosomal proteins of *E. coli* (RAAU of 0.0948 in ribosomal proteins vs. 0.0437 genome-wide). But Lys is the most common amino acid across the *Buchnera* genome (RAAU of 0.0988 genome-wide), consistent with the strong AT mutational bias. The slightly higher frequency of Lys in high-expression *Buchnera* genes is not significant given the genome-wide abundance of this amino acid.

Despite a strong effect of gene expression level on amino acid usage in *Buchnera*, we found no effect of gene expression on patterns of relative synonymous codon usage (RSCU) in this species. That is, in a multivariate analysis of RSCU of the 479 *Buchnera* loci included in this study, no major axis distinguished putative high- and low-expression genes (data not shown). This genome-wide analysis supports previous evidence that mutational bias and drift shape codon usage in *Buchnera* (e.g., Brynnel et al. 1998; Wernegreen and Moran 1999) and adds to the accumulating evidence that translational selection is not sufficiently strong or effective (or both) to counter the effects of genetic drift and mutational bias on synonymous codon usage. Therefore, gene expression levels influence amino acid usage, where selection may act more strongly, but not synonymous codon usage, where weak selection is apparently ineffective in small *Buchnera* populations. Another plausible explanation for an absence of translational selection on codon usage in *Buchnera* could be the equal abundance of tRNAs in this genome, which contains only one or a few copies of each tRNA. The reduced tRNA populations in the AT-rich genome of the parasite *R. prowazekii* was also postulated as a major factor in the absence of codon usage biases (Andersson and Sharp 1996). Interestingly, a more detailed analysis of codon bias in *Buchnera* suggests that different mutational biases on leading and lagging strands affects synonymous codon usage (Claude Rispe, personal communication).

In *E. coli*, hydrophobicity is the primary factor shaping variation in amino acid usage among proteins, but the effects of gene expression (although secondary) are nonetheless apparent. Comparisons of putative high- and low-expression genes show many similarities between *Buchnera* and *E. coli* because no amino acid shows significantly different trends in the two genomes (fig. 2). The significant underrepresentation of Trp in high-expression genes of both genomes suggests selection against the use of this aromatic amino acid. On the basis of similar results for *E. coli*, Lobry and Gautier (1994) suggested that the energetic costs of aromatic amino acids may account for their low abundance in high-expression genes. Recently, a more detailed study of metabolic efficiency in bacteria analyzed the cost of each amino acid in terms of high-energy phosphate bonds (the most expensive amino acids being the aromatic Trp, Phe, and Tyr) (Akashi and Gojobori 2002). The authors found decreased abundance of costly amino acids in high-expression *E. coli* loci regardless of gene function. Interestingly, our comparison of high- and low-expression *E. coli* genes is based on a limited gene sample but yielded results entirely consistent with this previous genome-wide analysis. Only Phe, Pro, Ser, and Arg differ in whether they show significant changes with gene expression, but this could be attributed to our necessarily smaller gene sample. Consistent with our results, Akashi and Gojobori (2002) also found no relationship between gene expression and the abundance of amino acids encoded by AT-rich or GC-rich codons in *E. coli*. This pattern in *E. coli* contrasts with the striking correlation in *Buchnera* between GC-richness of amino acid codons and gene expression (as reflected in significant correlations between Axis 1 and GC content at first and second codon positions and the biased amino acid profiles at high- vs. low-expression genes [fig. 1*b*, fig. 2]).

Previous work shows that tRNA pools match overall amino acid usage of proteins in several genomes (Yamao et al. 1991). In *E. coli*, the amino acid composition of high-expression genes correlates more strongly with tRNA abundances than do low-expression genes (Lobry and Gautier 1994). This trend has been interpreted as coadaptation between amino acid composition of proteins and tRNA-pools to enhance translational efficiency (Lobry and Gautier 1994; Akashi and Eyre-Walker 1998). In this study, we plot the two axes that correlated with gene expression in the *E. coli* COA against major tRNA abundances (Dong, Nilsson and Kurland 1996). Axis 2 is not correlated with tRNA abundance of the corresponding amino acid, whereas Axis 3 correlates significantly with tRNA abundance (but only before applying the Bonferroni correction; $r_s = -0.66$, $P < 0.001$; fig. 1*d*). This data supports previous evidence that translational selection may shape amino acid usage in *E. coli* (Lobry and Gautier 1994). This pattern contrasts with *Buchnera*, in which equal abundances of tRNA molecules or reduced efficacy of selection may limit effects of translational selection on both codon usage (see above) and amino acid usage.

Other Processes that may Drive Intragenomic Variation in Amino Acid Composition of *Buchnera* and *E. coli*

We calculated D{H,L} on the leading and lagging strands separately to distinguish the effects of strand orientation and gene expression level. In *Buchnera*, strand orientation does not account for the observed differences between high- and low-expression genes (table 1). But in *E. coli*, switches in the sign of D{H,L} depending on strand orientation suggest that strand-specific mutational biases may affect amino acid usage (table 2). Namely, high-expression genes of *E. coli* may experience distinct mutational pressures by virtue of their prevalence on the leading strand of replication. For example, the strong bias of Val (encoded by GTN) on the leading strand in *E. coli* and other species, independent of gene expression level, is consistent with a G > C and T > A skew on the leading strand resulting from strand mutational asymmetries (Mackiewicz et al. 1999; Rocha, Danchin, and Viari 1999). Likewise, strand position apparently contributes to the overrepresentation of Val in high-expression *E. coli* genes in our study.

Strand-specific biases may also drive asymmetries between the coding and noncoding DNA strands, as a result of transcription-associated mutation or DNA repair (or both) (e.g., Francino and Ochman 1999). If this bias increases with transcription levels, then transcription-associated asymmetries may contribute to differences between high- and low-expression genes. For example, it is possible that C→T mutational bias on the coding strand (Beletskii and Bhagwat 1996) may contribute to the observed high frequencies of certain GT-rich amino acids (e.g., Gly [GGN] and Val [GTN]) at high-expression genes in *Buchnera*, and low frequencies of CA-rich amino acid codons ([Pro (CCN)], Gln [CA(AG)]) at these genes in *E. coli*. But transcription-associated biases alone cannot account for distinct amino acid profiles in high- and low-expression genes. For example, several amino acids that use GT-rich codons are significantly *under*represented at high-expression genes of one or both species (Trp [TGG], Phe [TT(TC)] Cys [TG(TC)]). Nor can transcription-associated biases account for the observed reduction in the AT content of amino acid codons at high-expression genes because a C→T mutational bias is expected to increase T-richness of high-expression genes.

In analyses of single genomes, apparent differences in amino acid usage at high- and low-expression genes may partially reflect distinct structural or functional requirements of the proteins selected. But gene-specific structural or functional constraints have a minimal effect on the conclusions of our study, which is based on a comparison of an identical set of genes in *Buchnera* and *E. coli*. Moreover, the consistency of our results with a previous genome-wide study of *E. coli* (Akashi and Gojobori 2002) suggests that, for the purposes of this study, our limited gene sample is largely characteristic of high- and low-expression genes. Paralogous genes also pose complications for analyses of individual genomes because amino acid profiles of paralogs may be similar because of common ancestry. But in this study the observed differences between *E. coli* and *Buchnera* cannot be explained by gene duplication in *Buchnera*, which basically represents a subset of the *E. coli* genome (Shigenobu et al. 2000).

Conclusions

In summary, this comparative genome analysis of *Buchnera* and *E. coli* K-12 highlights important differences in the effects of mutation and selection on amino acid usage in these species. Mutational bias and genetic drift likely explain trends toward genome reduction and AT richness in *Buchnera*, as well as other bacterial endosymbionts, intracellular parasites, and organelles (Andersson and Kurland 1998; Mira, Ochman, and Moran 2001; Selosse, Albert, and Godelle 2001). This analysis of the *Buchnera* genome suggests a shift in the mutation-selection balance across loci that depends on gene expression level. Strong AT mutational bias impacts all *Buchnera* loci, but its effect on amino acid usage is apparently greater at low-expression genes than at high-expression genes. Tighter selective constraints at high-expression *Buchnera* genes may limit changes toward AT-rich or aromatic amino acids (or both), and may act either at the level of the bacterium or the aphid host (Rispe and Moran 2000). In addition to selection associated with gene expression levels, selection to maintain hydrophobic amino acids in IMPs also shapes global amino acid composition of *Buchnera* as a secondary force. This and previous analyses of *E. coli* show that hydrophobicity of IMPs is the primary source of variation in amino acid usage among loci, whereas factors relating to gene expression (e.g., selection for amino acids with high tRNA abundances or biased strand orientation of high-expression genes) are secondary. These distinct patterns of protein evolution in *Buchnera* and *E. coli* may result from different magnitudes and effects of mutational pressure versus selection because of the obligate host association of the endosymbiont. Further genome-level studies of other endosymbionts will provide a comparative framework to determine whether bacterial species with similar lifestyles show parallel modes of evolution and whether processes shaping amino acid usage in *Buchnera* extend to other endosymbiont species.

Acknowledgments

We are grateful to Claude Rispe for helpful discussion about *Buchnera* genome evolution and Maile C.

Neel for statistical advice. We also thank Hilary G. Morrison, Laila Nahum, Michael P. Cummings, Adam Eyre-Walker, and two anonymous reviewers for helpful comments on the manuscript. This work is supported by an award from the NASA Astrobiology Institute (NCC2-1054) to C.P. and awards from the NIH (R01 GM62626-01), NSF (DEB 0089455), and the Josephine Bay Paul and C. Michael Paul Foundation to J.J.W.

LITERATURE CITED

AKASHI, H., and A. EYRE-WALKER. 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. **8**: 688–693.

AKASHI, H., and T. GOJOBORI. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc. Natl. Acad. Sci. USA **99**: 3695–3700.

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. E., and C. G. KURLAND. 1998. Reductive evolution of resident genomes. Trends Microbiol. **6**:263–268.

ANDERSSON, S. G. E., and P. M. SHARP. 1996. Codon usage and base composition in *Rickettsia prowazekii*. J. Mol. Evol. **42**:525–536.

BAUMANN, P., L. BAUMANN, M. A. CLARK, and M. L. THAO. 1998. Genetic properties and adaptations of *Buchnera aphidicola* to an endosymbiotic association with aphids. ASM News **64**:203–208.

BELETSKII, A., and A. S. BHAGWAT. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **93**:13919–13924.

BENNETZEN, J. L., and B. D. HALL. 1982. Codon selection in yeast. J. Biol. Chem. **257**:3026–3031.

BLATTNER, F. R., G. PLUNKETT, III, C. A. BLOCH et al. (17 coauthors). 1997. The complete genome sequence of *Escherichia coli* K-12. Science. **277**:1453–1474.

BRACHO, A. M., D. MARTINEZ-TORRES, A. MOYA, and A. LATORRE. 1995. Discovery and molecular characterization of a plasmid localized in *Buchnera* sp. bacterial endosymbiont of the aphid *Rhopalosiphum padi*. J. Mol. Evol. **41**:67–73.

BREWER, B. 1988. When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. Cell **53**:679–686.

BRYNNEL, E. U., C. G. KURLAND, N. A. MORAN, and S. G. E. ANDERSSON. 1998. Evolutionary rates for *tuf* genes in endosymbionts of aphids. Mol. Biol. Evol. **15**:574–582.

BUCHNER, P. 1965. Endosymbiosis of animals with plant microorganisms. Interscience-Wiley, New York.

CLARK, M. A., L. BAUMANN, and P. BAUMANN. 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., N. A. MORAN, and P. BAUMANN. 1999. Sequence evolution in bacterial endosymbionts having extreme base compositions. Mol. Biol. Evol. **16**:1586–1598.

CRAIG, C. L., and R. S. WEBER. 1998. Selection costs of amino acid substitutions in ColE1 and ColIa gene clusters harbored by *Escherichia coli*. Mol. Biol. Evol. **15**:774–776.

DE MIRANDA, A. B., F. ALVAREZ-VALIN, K. JABBARI, W. M. DEGRAVE, and G. BERNARDI. 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.

DONG, H., L. NILSSON, and C. G. KURLAND. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J. Mol. Biol. **260**:649–663.

DOUGLAS, A. E. 1995. The ecology of symbiotic micro-organisms. Adv. Ecol. Res. **26**:69–103.

DOUGLAS, A. E., and W. A. PROSSER. 1992. Synthesis of the essential amino acid tryptophan in the pea aphid (*Acyrthosiphon pisum*) symbiosis. J. Insect Physiol. **38**:565–568.

FRANCINO, M. P., and H. OCHMAN. 1999. A comparative genomics approach to DNA asymmetry. Ann. N. Y. Acad. Sci. **870**:428–431.

FUNK, D. J., L. HELBLING, J. J. WERNEGREEN, and N. A. MORAN. 2000. Intraspecific phylogenetic congruence among multiple symbiont genomes. Proc. R. Soc. Lond. B **267**: 2517–2521.

GREENACRE, M. J. 1984. Theory and applications of correspondence analysis. Academic Press, London.

GUTIERREZ, G., L. MARQUEZ, and A. MARIN. 1996. Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency. Nucleic Acids Res. **24**:2525–2527.

IKEMURA, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151**:389–409.

ISHIKAWA, H. 1984. Characterization of the protein species synthesized in vivo and in vitro by an aphid endosymbiont. Insect Biochem. **14**:417–425.

KARLIN, S., A. M. CAMPBELL, and J. MRAZEK. 1998. Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. **32**:185–225.

KYTE, J., and R. F. DOOLITTLE. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.

LAFAY, B., A. T. LLOYD, M. J. MCLEAN, K. M. DEVINE, P. M. SHARP, and K. H. WOLFE. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res. **27**: 1642–1649.

LOBRY, J. R. 1996. Origin of replication of *Mycoplasma genitalium*. Science. **272**:745–746.

LOBRY, J. R., and C. GAUTIER. 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.

MACKIEWICZ, P., A. GIERLIK, M. KOWALCZUK, M. DUDEK, and S. CEBRAT. 1999. How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res. **9**:409–416.

MCINERNEY, J. O. 1997. Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microb. Comparat. Genomics **2**:89–97.

———. 1998. Replication and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. USA **95**:10698–10703.

MCLEAN, D. L., and E. J. HOUK. 1973. Phase contrast and electron microscopy of the mycetocytes and symbiontes of the pea aphid *Acyrtosiphon pisum*. J. Insect Physiol. **19**: 625–633.

MIRA, A., H. OCHMAN, and N. A. MORAN. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. **17**:589–596.

Moran, N. 1996. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA **93**: 2873–2878.

Moran, N., and P. Baumann. 1994. Phylogenetics of cytoplasmically inherited microorganisms of arthropods. Trends Ecol. Evol. **9**:15–20.

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

Muller, J. 1964. The relation of recombination to mutational advance. Mutat. Res. **1**:2–9.

Munson, M. A., P. Baumann, M. A. Clark, L. Baumann, N. A. Moran, D. J. Voegtlin, and B. C. Campbell. 1991. Evidence for the establishment of aphid-eubacterium endosymbiosis in an ancestor of four aphid families. J. Bacteriol. **173**:6321–6324.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. Nature. **246**:96–98.

Ohtaka, C., and H. Ishikawa. 1993. Accumulation of adenine and thymine in a *gro*E-homologous operon of an intracellular symbiont. J. Mol. Evol. **36**:121–126.

Rispe, C., and N. A. Moran. 2000. Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. Am. Nat. **156**:425–441.

Rocha, E., A. Danchin, and A. Viari. 1999. Universal replication biases in bacteria. Mol. Microbiol. **32**:11.

Romero, H., A. Zavala, and H. Musto. 2000. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucleic. Acids. Res. **28**:2084–2090.

Sato, S., and H. Ishikawa. 1997. Expression and control of an operon from an intracellular symbiont which is homologous to the *gro*E operon. J. Bacteriol. **179**:2300–2304.

Selander, R. K., D. A. Caugant, and T. S. Whittam. 1987. Genetic structure and variation in natural populations of *Escherichia coli*. Pp. 1625–1648 *in* F. C. Neidhardt, J. L. Ingraham, K. B. Low, G. Magasanik, M. Schaechter, and H. E. Umbarger, eds. *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular Biology. American Society of Microbiology, Washington, D.C.

Selosse, M., B. Albert, and B. Godelle. 2001. Reducing the genome size of organelles favours gene transfer to the nucleus. Trends Ecol. Evol. **16**:135–141.

Sharp, P. M., and W. H. Li. 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., H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* APS. Nature. **407**: 81–86.

Shigenobu, S., H. Watanabe, Y. Sakaki, and H. Ishikawa. 2001. Accumulation of species-specific amino acid replacements that cause loss of particular functions in *Buchnera*, an endocellular bacterial symbiont. J. Mol. Evol. **53**:377–386.

Silva, F. J., A. Latorre, and A. Moya. 2001. Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. Trends Genet. **17**:615–618.

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

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. W.H. Freeman and Co., New York.

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

Unterman, B. M., P. Baumann, and D. L. McLean. 1989. Pea aphid symbiont relationships established by analysis of 16S rRNAs. J. Bacteriol. **171**:2970–2974.

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

———. 2001. Vertical transmission of biosynthetic plasmids in aphid endosymbionts (*Buchnera*). J. Bacteriol. **183**(2): 785–790.

Yamao, F., Y. Andachi, A. Muto, T. Ikemura, and S. Osawa. 1991. Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. Nucleic Acids Res. **19**:6119–6122.

Yoshikawa, H., and N. Ogasawara. 1991. Structure and function of DnaA and the DnaA-box in eubacteria: evolutionary relationships of bacterial replication origins. Mol. Microbiol. **5**:2589–2597.

Adam Eyre-Walker, reviewing editor

Accepted May 9, 2002