Mutational and Selective Pressures on Codon and Amino Acid Usage in *Buchnera*, Endosymbiotic Bacteria of Aphids

Claude Rispe, François Delmotte, Roeland C.H.J. van Ham, et al.

Genome Res. 2004 14: 44-53
Access the most recent version at doi:10.1101/gr.1358104

Letter

Mutational and Selective Pressures on Codon and Amino Acid Usage in *Buchnera*, Endosymbiotic Bacteria of Aphids

Claude Rispe,[1,6] François Delmotte,[2,3] Roeland C.H.J. van Ham,[4,5] and Andres Moya[2]

[1]*UMR BIO3P, Institut National de la Recherche Agronomique, BP35327, 35653 Le Rheu cedex, France;* [2]*Institut Cavanilles de Biodiversitat i Biologia Evolutiva and Departament de Genetica, Universitat de Valencia, 46071 Valencia, Spain*

We have explored compositional variation at synonymous (codon usage) and nonsynonymous (amino acid usage) positions in three complete genomes of *Buchnera*, endosymbiotic bacteria of aphids, and also in their orthologs in *Escherichia coli*, a close free-living relative. We sought to discriminate genes of variable expression levels in order to weigh the relative contributions of mutational bias and selection in the genomic changes following symbiosis. We identified clear strand asymmetries, distribution biases (putative high-expression genes were found more often on the leading strand), and a residual slight codon bias *within* each strand. Amino acid usage was strongly biased in putative high-expression genes, characterized by avoidance of aromatic amino acids, but above all by greater conservation and resistance to AT enrichment. Despite the almost complete loss of codon bias and heavy mutational pressure, selective forces are still strong at nonsynonymous sites of a fraction of the genome. However, *Buchnera* from *Baizongia pistaciae* appears to have suffered a stronger symbiotic syndrome than the two other species.

Bacteria have repeatedly quitted their normal free life to invade a new habitat, the eukaryotic cell. The complete transition, known as endosymbiosis, involved a revolution in the lifestyle of the prokaryote: A first major change concerned the availability of nutrients, because the new environment was more predictable, richer, and less competitive than the outside of cells. Another change was that the population dynamics of the symbiont became constrained by that of the host. Because of the several orders of magnitude difference in population size between free-living bacteria and eukaryotes, the transition implied a sharp decrease in population size for the symbiont. Additionally, with uniparental transmission of the symbionts, the reproductive mode became de facto asexual, bacteria losing the possibility of exchanging genetic material with other unrelated bacterial lineages, whereas within-host variability is kept low by regular intergenerational bottlenecks. Not surprisingly, endosymbiosis has systematically been characterized by spectacular genomic changes (Wernegreen 2002) defining a syndrome, which is comprised of (1) major reduction of the genome size due to loss of many genes, (2) AT enrichment, and (3) acceleration of the rate of evolution, notably at nonsynonymous sites.

The driving force behind each of these changes is not easy to identify; several conflicting interpretations have been proposed, which can be classified as negative, neutral, or even positive. For example regarding genome size, the reduction could be positively selected in the context of within-host competition, the shorter genomes replicating faster (Albert et al. 1996), or may be a result of gradual decay of genes, which are progressively eliminated from the genomes (Mira et al. 2001; Ochman and Moran 2001; Silva et al. 2001). In this view, gene loss could be the result of neutral evolution (by the loss of genes become useless in the host cell) or could be slightly deleterious.

However, the present article focuses on the second symptom associated with endosymbiosis, the shift in base composition. In *Buchnera*, the endosymbiont of aphids, most authors until recently attributed AT enrichment to a deleterious process (Moran 1996; Wernegreen and Moran 1999). Mutational bias probably results from greater basal mutabilities of GC versus AT bases (Graur and Li 2000), a general phenomenon which would be not countered by selection in this symbiont. In the context of reduced population sizes (resulting in increased drift) and vertical sequestration of bacterial lineages (resulting in effective asexuality), this would cause the progressive accumulation of slightly deleterious mutations. Genomic evolution in *Buchnera* has therefore been interpreted as an example of Muller's ratchet (Moran 1996; Rispe and Moran 2000) and in the framework of the near neutral theory (Ohta 1992), as it seems to be driven by the interaction of drift and weak selection. Several lines of evidence support this argument, that is, the general acceleration of evolutionary rates even in usually conserved RNAs (Moran 1996; Clark et al. 1999), the increased proportion of nonsynonymous mutations (Moran 1996), low levels of intraspecific polymorphisms (Funk et al. 2001; Abbot and Moran 2002), and the estimated decreased stability of rRNAs (Lambert and Moran 1998) and proteins (van Ham et al. 2002).

However, an alternative explanation has recently challenged this dominant view; it presents a purely neutral interpretation of *Buchnera* evolution (Itoh et al. 2002), assuming that the simple increase of mutation rates could explain the general acceleration of evolutionary rates in *Buchnera*. Itoh et al. criticized the "deleterious interpretation," arguing that long-term deleterious evolution should have led to the loss of genes, and that the proportion of nonsynonymous substitutions had been overestimated in previous studies.

We do not propose here to definitively settle the argument between these conflicting interpretations. However, we took advantage of the great wealth of information presented by the availability of three complete genome sequences of *Buchnera* (Shigenobu et al. 2000; Tamas et al. 2002; van Ham et al. 2002) to

Present addresses: [3]UMR Santé Végétale, Institut National de la Recherche Agronomique, BP81, 33883 Villenave d'Ornon, France; [4]Centro de Astrobiología, Instituto Nacional de Técnica Aeroespacial, Carr. de Ajalvir, 28850 Torrejón de Ardoz, Spain; [5]Plant Research International, B.u. Genomics, 6700 AA, Wageningen, The Netherlands.
[6]Corresponding author.
E-MAIL rispe@rennes.inra.fr; FAX 33-2-2348-5150.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1358104. Article published online before print in December 2003.

14:44–53 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

evaluate the effect of mutation and selection on the symbiont sequences. The original infection in the aphid/*Buchnera* association occurred 200–250 million years ago and has been followed by strict cospeciation (Moran and Baumann 1994, Clark et al. 2000). The three genomes analyzed here correspond to two moderately distant aphid species (*Acyrthosiphon pisum* and *Schizaphis graminum*) belonging to Aphididae and a more distant one (*Baizongia pistaciae*) belonging to Pemphigidae; they will be referred to as *Buchnera*-Aps, *Buchnera*-Sgr, and *Buchnera*-Bpi respectively. The symbionts of *A. pisum* and *S. graminum* show an amazingly high level of "genomic stasis" reflected by the very high proportion of genes shared (about 90%) and perfect synteny indicating the absence of rearrangements following divergence. In contrast, the more divergent *Buchnera*-Bpi showed also a nearly perfect gene-order conservation but a significantly more pronounced genome reduction than the two other closer strains of *Buchnera*. For the three genomes, we examined intragenomic variations of codon and amino acid usage. For comparison, we performed the same analysis in a set of orthologs from *Escherichia coli*, a close free-living relative of *Buchnera*. This was done to allow a more exact comparison than with complete sets of *E. coli* genes (comprising many instances of recent gene acquisition by horizontal transfer). Codon bias is high in highly expressed genes of *E. coli* where it matches abundances in tRNAs (Ikemura 1981). This has been interpreted as the result of translational selection, favoring the more rapid translation of the most needed proteins. Preliminary analyses detected no significant codon bias in *Buchnera*, studying either a moderate number of genes of a few *Buchnera* species (Wernegreen and Moran 1999) or one gene in many species (Moya et al. 2002). However, to our knowledge, no complete analysis of codon usage using whole sequences of *Buchnera* has been published. This is perhaps because these initial results and the apparent genome-wide uniformity in AT richness (about 85% AT at third positions) discouraged further investigation. However, codon bias might have persisted at least in some genes, which could only be revealed by analyzing the full array of data. Also, codon usage might respond to other subtle forces, as for example strand bias (see Ermolaeva 2001 for a short review of factors influencing codon usage). We have analyzed codon usage variations using both multivariate analysis and systematic χ^2 tests (details given in the Methods).

Our main objective was to search correlations between the level of expression and codon bias, in order to disentangle mutational and selective pressures on codon usage. For that purpose, we needed a marker of the level of expression. In their recent study, Palacios and Wernegreen (2002) defined a set of putatively high-expression genes including ribosomal proteins, GroEL and GroES, two chaperones which constitute a high fraction of the proteome (Ishikawa 1984) and are thought to play a special role in the symbiont, possibly compensating the effect of multiple slightly deleterious mutations (Moran 1996; Fares et al. 2002a). Another indicator of gene expression is the level of codon bias which can be quantified by the Codon Adaptive Index or CAI (Sharp and Li 1987), using a set of genes of recognized high expression (through experimental data) as a reference. However, calculating the CAI in this way for *Buchnera* can be misleading, because the extreme AT bias of the whole genome makes it difficult to assign "optimal" codons. We therefore rather used the CAI of *E. coli* orthologs, assuming that it remained a good indicator of gene expression and selective importance in the symbionts. Rather than comparing defined subsets of genes with putative high- and low-expression genes, we used the CAI as a continuous potential marker of the level of expression, for all genes. This had the advantage of using the full available information, which reduced the risk of relatively arbitrary choice within each class of genes, and increased the statistical power of the comparisons.

The second level of variation explored in this study concerns amino acid usage. Amino acid composition has been shown to be related to the level of expression through an apparent process of energetic optimization (Akashi and Gojobori 2002), because highly expressed genes avoid using energetically costly residues (in particular aromatic amino acids). A recent study investigated amino acid variation in *Buchnera*, using factorial correspondence analysis (FCA) on the complete sequence of *Buchnera*-Aps, and contrasting the composition of two groups of genes defined as highly and lowly expressed (Palacios and Wernegreen 2002). Its conclusions supported ongoing selection in *Buchnera* for energetic optimization at the cell level. We reanalyzed amino acid usage variation in the three complete *Buchnera* sequences, using multivariate analysis, and relating the patterns of variation with the values of the CAI of *E. coli* orthologs for every gene and not just the two groups defined above. Relative amino acid usage variation was first studied separately in the three symbionts and *E. coli*; then, to allow a direct comparison, the four genomes were pooled and the similarity of amino acid profiles was compared within (for different categories of genes) and between the genomes.

RESULTS

Global Base Composition

All three *Buchnera* showed a marked AT enrichment compared to the set of *E. coli* orthologs (Table 1), *Buchnera*-Bpi being only slightly less AT-rich than the two others; this enrichment was much stronger and more uniform at third positions ($\%AT_3$) than at the first two positions ($\%AT_{12}$) as revealed by the respective standard errors of these parameters. Also, $\%AT_3$ was practically equal to %AT of intergenic sequences, as expected if third positions were dominated by mutational bias.

Strand-specific skews, known to influence codon usage in other bacteria, were also reported in Table 1, for the third codon positions (these skews are generally stronger than at the more constrained first and second positions). As is generally seen in

Table 1. Percentage of AT at the Different Codon Positions for All CDs, %AT of Intergenic Spacers ≥50 bp in the Three Symbionts and in Their Unambigous Orthologs in *E. coli* K-12 MG1255, and Average AT and GC Skews at the Third Codon Position, for Leading and Lagging Strands

		$\%AT_{12}$		$\%AT_3$		$\%AT_{intergenic}$		AT_3 skew		GC_3 skew	
	N	**Mean**	**SD**	**Mean**	**SD**	**Man**	**SD**	**Leader**	**Lagger**	**Leader**	**Lagger**
Buchnera-Aps	*564*	**66.2**	5.4	**85.5**	2.6	**84.8**	4.6	−0.06	0.04	0.18	0.02
Buchnera-Sgr	*545*	**67.0**	5.6	**87.2**	2.4	**85.8**	4.6	−0.07	0.06	0.24	0.05
Buchnera-Bpi	*504*	**64.7**	4.6	**84.9**	2.6	**84.3**	3.7	−0.10	0.11	0.33	−0.47
E. coli K12 MG1655	*597*	**49.2**	3.2	**43.7**	6.2	**57.0**	7.1	−0.20	−0.17	0.01	−0.02

Table 2. Results of Factorial Correspondence Analyses on Relative Codon Usage in *Buchnera* and in *E. coli* Orthologs

		Inertia	CAI_{Eco}	GC_{12}	CG_3	AT_{s3}	GC_{s3}
B-aps	z1	6.1	**−0.11***	**−0.24****	0.06	**0.63****	**−0.56****
	z2	4.3	**0.14****	**0.16****	**0.29****	−0.07	0.01
B-sgr	z1	7.9	**−0.11***	**−0.21****	0.01	**0.71****	**−0.54****
	z2	3.8	**0.14****	**0.13***	0.09	**0.11***	**−0.31****
B-bpi	z1	22.9	−0.09	**−0.17****	**0.20****	**0.82****	**−0.91****
	z2	3.3	−0.05	−0.11	0.05	−0.08	−0.08
E. coli	z1	24.0	**−0.96****	0.10	0.01	**0.19****	**0.57****
	z2	5.8	−0.05	0.10	**−0.82****	0.09	−0.08

For each of the first two axes of the FCA (z1 and z2), % of total inertia, and nonparametric (Spearman's) correlation coefficients between z-values and CAI-Eco (codon adaptive index of the *E. coli* ortholog), GC content, AT skew (ATs) and GC skew (GCs) indices indicating codon positions. Bold values are significant at the 0.05 level (*) or at the 0.01 level (**) after Bonferroni's correction for multiple tests (k = 5); Sokal and Rohlf (2003).

bacteria, the leading strand shows an excess of T (vs. A) and G (vs. C) in *Buchnera*. However, a considerable variation was observed in the different genomes. In particular, the net GC-skewing attributable to replication direction, defined as half the difference between GC-skews on the leading and lagging strand respectively, was as low as 1.5% in *E. coli*, but reached 8% in *Buchnera*-Aps, 9.5% in *Buchnera*-Sgr, and the exceptional value of 39.6 % in *Buchnera*-Bpi.

Factorial Correspondence Analysis on Relative Frequencies of Synonymous Codons

We comment only the first two axes, because subsequent axes yielded little information and no significant correlations. With 6.1% only for the first axis in *Buchnera*-Aps, the level of explanation was very low (Fig.1; Table 2). This confirmed previous studies that showed very low levels of codon usage variability in *Buchnera* (Wernegreen and Moran 1999; Moya et al. 2002). However, our analysis revealed several new trends: first, the first axis was strongly correlated with AT and GC skews, particularly at the third position. At the right of the first axis, genes were characterized by richness in A (vs. T) and C (vs. G), whereas it was the opposite at the left. Base skews are related to the sense of replication in *Buchnera* (Shigenobu et al. 2000), as in other bacteria. Using the origin of replication proposed by those authors, and determining the terminus at the inflexion point of base shifts (situated at 325 kb), we identified genes on the leading and lagging strands. This showed that axis 1 discriminates both strands, although there is a rather large overlap between the two clouds of genes in *Buchnera*-Aps (Fig. 1). In *Buchnera*-Sgr (7.9%), and more markedly, in *Buchnera*-Bpi (22.9%), the level of description of the first axis was higher. This reflected the increased strand bias in these species, as shown by the better discrimination between the different orientations along axis 1 of the FCA (Fig. 1). Discriminating analysis allowed us to quantify this separation, showing that the percentage of correct classification among genes on the leading or lagging strand increased from *Buchnera*-Aps (84.9%) to *Buchnera*-Sgr (90.1%) to *Buchnera*-Bpi (97.9%).

However, the weight of subsequent axes remained very low in all *Buchnera*. Interestingly, there was a negative correlation between the first axis and the CAI of *E. coli* orthologs (rs = −0.09 to −0.11). This seems to be caused by a biased distribution of high- and low-expression genes on leading versus lagging strands. Indeed, the proportion of genes located on leading strands increases with CAI, from about 50% for low-CAI genes (<0.35) to 66% for high-CAI genes (>0.65) in the three *Buchnera* (Table 3). For ribosomal proteins and GroEL/S, the proportion of genes on the leading strand reaches 78% (44/56) .We analyzed the distribution of *E. coli* orthologs, which showed the same trend; however, the proportions of genes on the leading strand were higher than in *Buchnera* (by an 8%–14% margin) for every class of CAI (Table 3).

A second trend that could be identified was a significant correlation between CAI and the second axis in *Buchnera*-Aps and *Buchnera*-Sgr. In *Buchnera*-Aps and to a lesser extent in *Buchnera*-Sgr, variation on the second axis appeared to be correlated with GC richness at all codon positions. As this axis is orthogonal to the first one, it translates variations that are independent of gene orientation. This suggests that the level of expression had a slight residual effect on codon usage.

For comparison, the analysis of *E. coli* orthologs showed a high level of explanation for the first axis (24%), which was strongly correlated with CAI (rs = −0.96, $P < 0.0001$). The cloud had the typical V shape already identified in earlier FCA of unselected sets of *E. coli* genes, showing that the different classes of genes identified in these studies (Médigue et al. 1991; Kanaya et al. 1999) are all well represented in the orthologs of symbiont genes. *Buchnera* has therefore *E. coli* orthologs even in the class that Médigue et al. interpreted as probable recent lateral gene

Table 3. Percentages of Genes in *Buchnera* and *E. coli* Orthologs Located on the Leading (vs. Lagging) Strand for Three Classes of Increasing Codon Adaptive Index in *E. coli,* and in Ribosomal Proteins Plus GroEL/S

	Buchnera-Aps		*Buchnera*-Sgr		*Buchnera*-Bpl		*E. coli* K12	
	n	% leading strand	n	% leading strand	n	% leading strand	n	% leading strand
CAI-Eco <0.35	170	50.0%	165	49.1%	139	51.1%	173	59.0%
0.35 < CAI-Eco <0.65	328	56.4%	310	56.5%	302	57.0%	339	67.5%
CAI-Eco >0.65	57	66.7%	56	67.9%	55	67.3%	58	81.0%
groEL/S + ribosomal	56	78.6%	56	78.6%	56	78.6%	56	92.9%

The average for all *E. coli* genes is 55% (Blattner et al. 1997).

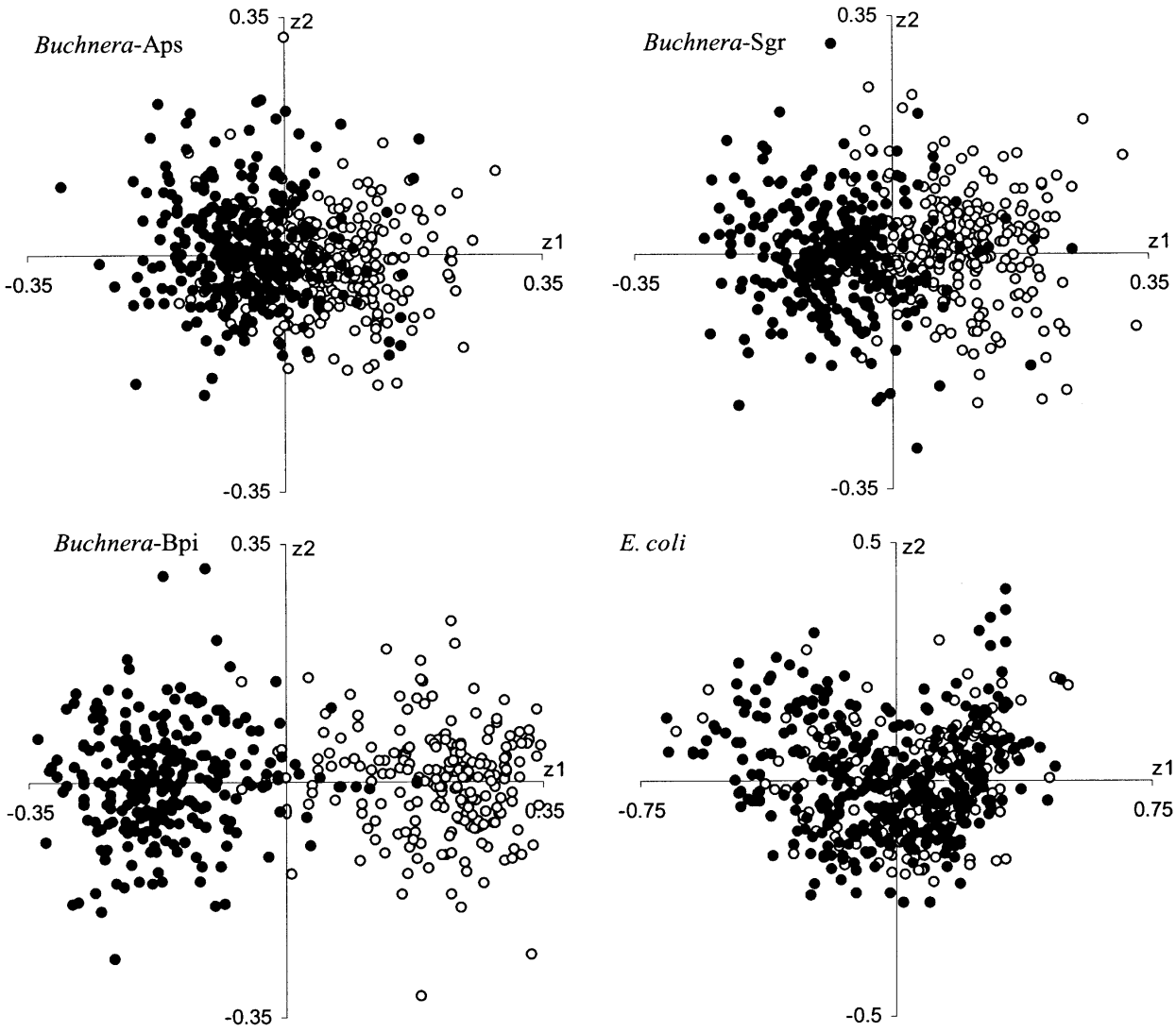

Figure 1 Dispersion along the first two axes (z1, z2) of the FCA of codon usage in three whole *Buchnera* genomes and the *E. coli* orthologs. (●) Genes on the leading strand; (○) genes on the lagging strand.

transfer (LGTs). Given that the separation between genomes occurred over 250 million years ago, this suggests that LGT was overestimated in Médigue et al.'s classification. The first axis was rather strongly correlated with GC_3 skew, but the overlap between leading and lagging strands was very high; finally, the second axis was primarily linked with GC_3 content (rs = −0.82).

χ^2 Tests

Pooling all the tests, we only found 6.2%, 6.6%, and 6.7% of biases significant at the 0.05 level, for *Buchnera*-Bpi, *Buchnera*-Aps, and *Buchnera*-Sgr, respectively. In the set of *E. coli* orthologs, this proportion was about fourfold, reaching 26.4%. However, we correlated the mean χ^2 values (standardized by gene size) with the CAI of *E. coli* orthologs separately for leading and lagging strands, in each of the three symbionts, and found positive relationships in all cases (Table 4). Yet, the slopes were significant (or borderline significant) only for *Buchnera*-Bpi and *Buchnera*-Sgr, and the fit values were very low, contrasting with the high significance and high fit for *E. coli*.

Factorial Correspondence Analysis of Relative Amino Acid Usage Within Each *Buchnera* Genome

In all three *Buchnera*, only three axes described about 45% of the total variability, which is slightly higher than in *E. coli* orthologs (43%). The first axis (z1) was strongly correlated with CAI-Eco (rs = −0.43 on average) and aromaticity (rs = 0.67 on average) but even more with $\%AT_{12}$ (rs = 0.78 on average). This axis therefore separates low-CAI, aromatic, GC_{12}-poor genes on the right and genes with opposed patterns on the left (Fig. 2). Palacios and Wernegreen (2002) found the same major determinants of amino acid usage in *Buchnera*-Aps, that is, aromaticity and AT content, but detected only minor effects of strand asymmetry. Focusing on two groups of genes (putatively highly and lowly expressed), they showed that the former tended to avoid aromatic residues and were less AT-rich than the latter. Because two of the aromatic amino acids are AT-rich (Tyr, Phe), the two trends were partly overlapping. However, Palacios and Wernegreen demonstrated that amino acid usage in *Buchnera* was influenced by both factors, that is, selection against AT-richness and selection against aromaticity. The specific amino acid usage of highly

Table 4. Regression Lines Between the Level of Codon Bias in *Buchnera* or *E. coli* (y) and the CAI of *E. coli* Orthologs (x)

	Regression line	r^2	*P*
B-Aps leading	y = 0.0016 x +0.0047	0.001	0.60
B-Aps lagging	y = 0.0022 x +0.0038	0.003	0.42
B-Sgr leading	y = 0.0045 x +0.0030	0.01	<0.05
B-Sgr lagging	y = 0.0062 x +0.0025	0.02	<0.05
B-Bpi leading	y = 0.0063 x +0.0021	0.05	<0.001
B-Bpi lagging	y = 0.0038 x +0.0033	0.02	0.06
E. coli K12	y = 0.0864 x −0.0153	0.45	<0.0001

The bias was measured by averaging χ^2 values (standardized by gene size) across fourfold and threefold (Ile) degenerated boxes. For *Buchnera*, χ^2 values were calculated separately for leading (●) and lagging strands (○), the reference being a set of 35 genes with the lowest CAI in *E. coli* for each orientation. For *E. coli*, the reference was the codon usage in the group of lowest CAI (<0.30) genes.

expressed *Buchnera* genes could therefore be guided by energetic optimization, leading to the avoidance of expensive aromatic amino acids (Akashi and Gojobori 2002). To precisely determine the level of energetic optimization of amino acid usage in *Buchnera*, relative to *E. coli*, we also examined the direction of changes in amino acid frequencies between orthologs in *Buchnera* and *E. coli*. We calculated the ratio between the frequency of each amino acid in *Buchnera* and in *E coli*, using the logarithm of this ratio (to normalize its distribution). To avoid zero values due to the absence of certain residues in small proteins, we added a unity for each residue. For each gene, the so-defined rate of change was therefore

$$r = \ln\left(\frac{(n_{buch}+1)/(nT_{buch}+1)}{(n_{eco}+1)/(nT_{eco}+1)}\right)$$

where n_{buch} and n_{eco} are the frequencies of a given residue, and nT_{buch} and nT_{eco} are the total numbers of codons. Values close to zero indicate no frequency change between *E. coli* and *Buchnera*, whereas positive values indicate an increase and negative values a decrease in the frequency of a given residue in *Buchnera*. We show only the results for *Buchnera*-Sgr (Fig. 3), but they are very close in the two other symbionts: All residues determined by AT-rich codons (among which are two aromatic amino acids, Phe and Tyr, and also Lys) increased, whereas all residues determined by GC-rich codons (including the aromatic Trp) decreased in frequency in the symbiont. Systematically, the intensity of the change was more moderate for high-CAI genes. Paradoxically, the CAI of *E. coli* orthologs, which marks codon biases in this bacteria, no longer does so in *Buchnera*; however, it had a very good predictive value of changes at the amino acid level.

The second axis was strongly linked with hydrophobicity (rs = 0.74 on average), highly hydrophobic proteins being situated on the upper part. This axis was also strongly correlated with AT skews at second positions (Table 5), which resulted in a partial discrimination of leading and lagging strands. This was most obvious in *Buchnera*-Bpi (Fig. 2), especially for positive values of z1 (for AT_{12}-rich genes). This interaction between hydrophobicity and gene orientation can be understood by a closer look at the distribution of "gravy" scores on a genetic code table (Kyte and Doolittle 1982). NTN and NAN codons have all positive and all negative scores, respectively. Then, for AT-rich genes, where NAN and NTN codons are abundant, genes on the leading strand (T > A) and lagging strand (A < T) tend to have positive and negative scores, respectively. This effect is much weaker in GC-rich genes, which results in a much weaker discrimination along the second axis between leading and lagging strands in the GC-rich region (for negative z1 values). This explains the peculiar triangular distribution of genes, with low z2 variation on the left and high z2 variation on the right (Fig. 2).

Global Analysis of Relative Amino Acid Usage in the Four Genomes

The amino acid profiles of all of the sequences from the four genomes were gathered and submitted to factorial correspondence analysis. A similar analysis was performed for codon usage but is not shown here, the differences between the three *Buchnera* appearing negligible at this scale, whereas those between *Buchnera* and *E. coli* were extreme. This was expected to reveal global effects separating the different symbionts, but also individual variations at the gene level. The first two axes of the FCA captured a large part of the variation (30.6% and 11.1%, respectively); the first axis was strongly correlated with GC content at the first two positions (rs = 0.98), GC content at the third position (rs = 0.88) and aromaticity (rs = −0.56), whereas the second axis was strongly correlated with hydrophobicity (rs = 0.85), aromaticity (rs = 0.50), AT_3 skew (rs = −0.39), and CAI_{Eco} (rs = −0.39). The distribution of the four genomes along the first two axes is shown in Figure 4: It shows large overlap between the *Buchnera* genomes, whereas overlap between *Buchnera* and *E. coli* is marginal. Interestingly, this overlap only occurs for negative z2 values (i.e., for less hydrophobic proteins). In contrast, the discrimination between *E. coli* and *Buchnera* genes is extreme for positive z2 value (i.e., for more hydrophobic proteins), with the exception of one outlier *Buchnera* sequence (atpE) which has both high positive z1 and z2 coordinates in the three *Buchnera*, and is among the most conserved proteins in the comparison between *Buchnera*-Aps and *Buchnera*-Sgr (Tamas et al. 2002).

We examined the individual information for orthologous genes, to evaluate the relative similarity of the amino acid usage profiles at different scales. We calculated the distances (*d*) between orthologous genes in the two-dimension plan of the first two axes of the FCA. To avoid biases due to differences in gene repertoires, we first compared distances to *E. coli* only for genes with an ortholog present in each of the three *Buchnera*. This comparison showed that *Buchnera*-Bpi have more divergent amino acid profiles from *E. coli*, whereas *Buchnera*-Aps is closest

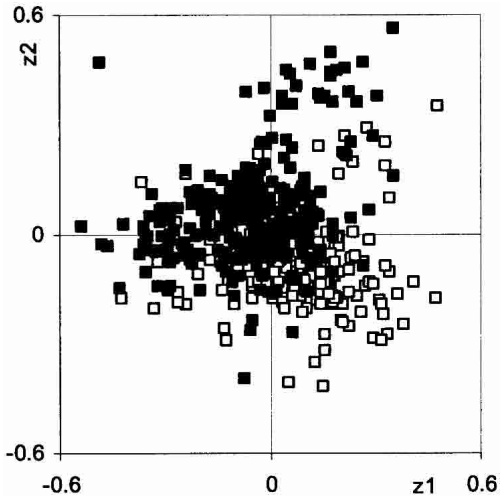

Figure 2 As in Figure 1, for *Buchnera*-Bpi only, but for amino acid usage.

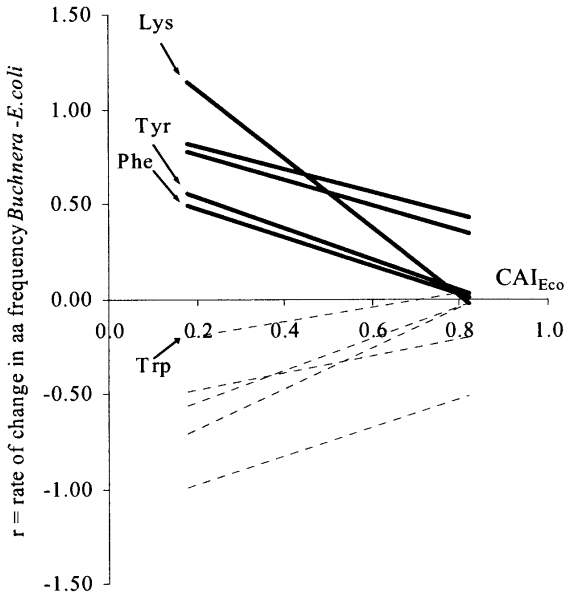

Figure 3 Relative change in amino acid frequency (*r*) between *Buchnera*-Sgr genes and their *E. coli* orthologs, in relation to the CAI in *E coli*. Solid lines, residues with AT-rich codons (at least 2 A or Ts); dotted lines, residues with GC-rich codons (at least 2 G or Cs; formula of *r* in equation 1).

to *E. coli* ($d_{Bpi\text{-}Eco} > d_{Sgr\text{-}Eco} > d_{Aps\text{-}Eco}$, all comparisons being significant, $P < 0.001$, two-tailed Student tests for paired observations, Table 6). For comparison, *Buchnera* orthologs were comparatively much closer, with on average $d_{Bpi\text{-}Aps} = d_{Bpi\text{-}Sgr} = 0.082 > d_{Aps\text{-}Sgr} = 0.056$ ($P < 0.001$).

Within each *Buchnera*, we also computed the mean distances to *E. coli* orthologs for genes which do not have an ortholog in one or two of the other complete *Buchnera* sequences, or whose at least one *Buchnera* ortholog is a pseudogene. For each of the three symbiotic genomes, the average distance between *Buchnera* genes and *E. coli* was much higher for this category of genes, compared to those present and functional in the three genomes. This increase was significant for the three *Buchnera* ($P < 0.001$, two-tailed Student tests, Table 6).

Finally, we analyzed Ka and Ks calculations provided by Tamas et al. (2002) for the comparison between *Buchnera*-Aps and *Buchnera*-Sgr: We separated genes shared by the three *Buchnera* genomes (i.e., the two former and *Buchera*-Bpi) *and* functional on one side, and genes present *and* functional in *Buchnera*-Aps and *Buchnera*-Sgr but lost or become pseudogene in *Buchnera*-Bpi on the other side. Both average nonsynonymous rates (Ka = 15.5 and Ka = 19.6, respectively, $P < 10^{-4}$) and synonymous rates (Ks = 89.0 and Ks = 96.4, respectively, $P = 0.04$) were significantly higher in the latter category.

DISCUSSION

Extending the conclusions of previous studies on smaller sets of genes, we confirmed that codon bias is extremely tenuous in *Buchnera*, a result of massive and almost uniform AT enrichment concentrated at third positions. By multivariate analysis, we could however determine trends affecting codon usage: In each of the three genomes the main determinant was gene orientation (with a respectively weak, moderate, and pronounced effect in *Buchnera*-Aps, *Buchnera*-Sgr, and *Buchnera*-Bpi). This was due to a strand asymmetry (such that G > C and T > A on the leading strand), a phenomenon generally explained by strand-specific mutational biases (Frank and Lobry 1999). The different intensity of strand bias is puzzling, particularly in *Buchnera*-Bpi, where it reaches an unprecedented level in bacteria, to our knowledge (surpassing the record high bias from *Borrelia burgdorferi*, McLean et al. 1998). This genome could have known, therefore, an increased asymmetry in substitutions on the different strands (e.g., C to T deaminations could be responsible for the strand-specific skews; Frank and Lobry 1999), due to increased mutation rates or decreased purifying selection. This symbiotic lineage has lost many genes compared to the two others, in particular some genes associated with the replication machinery (e.g., *topA*, *priA*, *dnaT*), which has been held as indicative of decreased efficiency of replication in *Buchnera*-Bpi (Van Ham et al. 2002). It may be advanced that these losses played a role in an increased differential replication fidelity between leading and lagging strands in this genome, which might have resulted in an increased strand-specific compositional bias (Frank and Lobry 1999).

Strand bias also dominates codon usage in other symbiotic or parasitic bacteria, such as *Rickettsia prowazekii* (Andersson et al. 1998) and *Borrelia burgdorferi* (McInerney 1998), in contrast to several free-living bacteria such as *E. coli*, where the prime cause of variation in synonymous codon usage is the level of expression of genes (Ikemura 1981). Translational selection, because of weak selective coefficients, is only efficient in large populations (Bulmer 1991), and seems therefore to be ineffective in parasitic/symbiotic bacteria due to reduced population sizes. As is com-

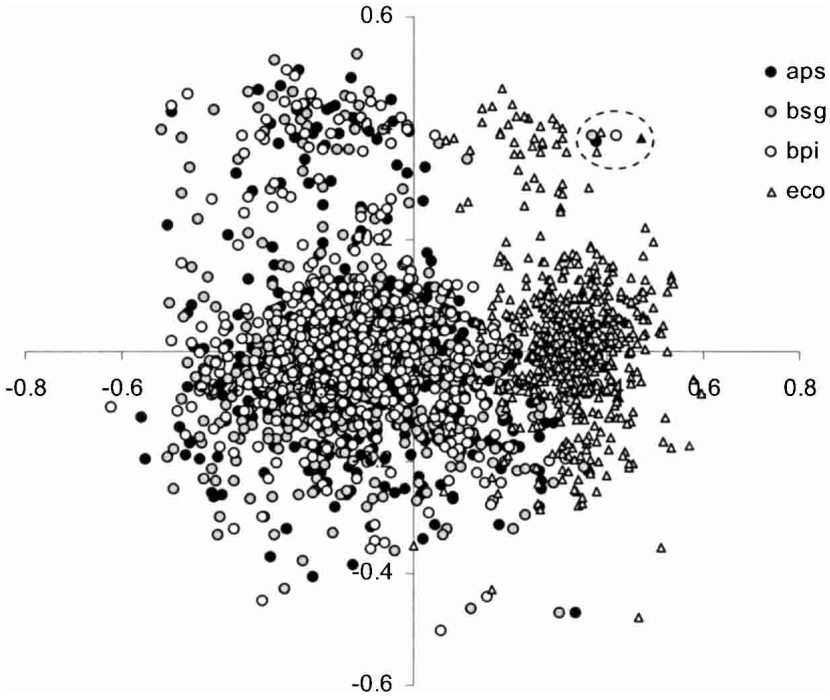

Figure 4 Dispersion along the first two axes (z1, z2) of the FCA of relative amino acid usage in the pooled *Buchnera* genomes and their *E. coli* orthologs. Symbols denote the genome as shown in the legend except for the *atpE* sequence from *E. coli* (▲); the dashed-line oval delineates the four sequences for this gene.

Table 5. Results of Factorial Correspondence Analyses on Amino Acid Usage in *Buchnera* and in *E. coli* Orthologs

		Inertia	CAI_{Eco}	GC_{12}	GC_3	AT_{s2}	Gravy	Aromo
B-aps	z1	22.7	**−0.43****	**−0.92****	−0.01	**0.13***	0.08	**0.70****
	z2	14.6	**−0.19****	**0.26****	**0.15****	**−0.76****	**0.74****	**0.29****
B-sgr	z1	24.6	**−0.39****	**−0.94****	**−0.18****	0.08	0.11	**0.71****
	z2	15.1	**−0.13***	**0.27****	**0.18****	**−0.78****	**0.76****	**0.29****
B-bpi	z1	22.0	**−0.46****	**−0.90****	0.07	0.01	**0.16****	**0.61****
	z2	15.5	**−0.16****	**0.13***	−0.12	**−0.78****	**0.72****	**0.45****
E. coli	z1	21.8	**−0.56****	**0.17****	**0.19****	**−0.19****	0.06	**0.38****
	z2	13.4	**−0.42****	−0.05	**0.30****	**−0.64****	**0.73****	**0.38****

Additional parameters; gravy score (level of hydrophobicity) and aromo (level of aromaticity = % Phe + Tyr + Trp). Bold values are significant at the 0.05 level (*) or at the 0.01 level (**) after Bonferroni's correction for multiple tests (k = 6); Sokal and Rohlf (2003).

mon in bacteria, we found a distribution bias of genes (particularly for those with a high CAI) on the leading stand. This is usually interpreted as the result of "replicational selection," by which presence on the leading strand would permit the avoidance of collision between polymerases when replication and transcription occur at the same time. However, this selective force has been considerably eroded if we compare *Buchnera* genes and *E. coli*, because the percentage of genes on the leading strand decreased for all categories of genes classified by CAI (Table 3). Because gene order and orientation are conserved in the three *Buchnera* studied, this evolution must have occurred before they diverged.

Still, we have identified a weak residual bias in presumably high-expression genes independent from strand asymmetries. This was observed through multivariate analysis (as shown by the correlation between CAI and the second axis) in *Buchnera*-Aps and *Buchnera*-Sgr but less so in *Buchnera*-Bpi, where the residual bias was weaker, and by $\chi 2$ values which very slightly increased with the CAI. The two methods were therefore concordant, although sometimes quantitatively different (e.g., a comparatively low effect of CAI on χ^2 values was observed for *Buchnera*-Aps, whereas multivariate analysis did detect a relatively strong effect of CAI), which is due to the fact that χ^2 tests were restricted to A- and T-ending codons (whereas GC-richness at the third position influenced codon usage in *Buchnera*-Aps).

It seems unlikely that the subsisting codon bias would result from ongoing translational selection. Actually, the tRNA battery is so depauperate in *Buchnera* that it simply corresponds to the minimal set allowing complete coverage of codons following rules of isoacceptance, and all tRNA genes are single-copy. Often, the present tRNA does not match the most used codons, although some of the lost tRNAs were those matching optimal codons in *E. coli* (for Gln, Leu, Ser). Thus, the residual bias in putative high-expression genes rather results from the greater conservation of these genes: Actually, using nonsynonymous and synonymous distances previously calculated for *Buchnera* (Tamas et al. 2002), we found that high-CAI genes have both lower Ka ($\rho_{KA\text{-}CAI} = -0.37$) and slightly lower Ks ($\rho_{KS\text{-}CAI} = -0.12$) than average. This supports the existence of a globally greater evolutionary constraint on the most essential genes, at both synonymous and nonsynonymous sites (King Jordan et al. 2002). Those authors offered two explanations for this pattern, a possible mechanistic bias in mutation, or the fact that synonymous sites are also subject to some degree of selection (King Jordan et al. 2002). The latter scenario could mean either selection on codon usage (a rather unlikely situation, as we argued above), or that synonymous substitutions might be not always silent (Graur and Li 2000). Interestingly, a similar interaction between the level of expression, the level of codon bias, and gene conservation (both synonymous and nonsynonymous) was demonstrated in *Mycobacterium* sp. (de Miranda et al. 2000), supporting the existence of common functional constraints between synonymous and nonsynonymous sites.

Amino acid usage showed a much greater range of variation than codon usage, suggesting a much stronger selective pressure on nonsynonymous positions. As first found for *Buchnera*-Aps only (Palacios and Wernegreen 2002), we found that amino acid usage was primarily driven by the percentage of A and T at the first two positions and aromaticity, and secondarily by hydrophobicity of proteins, in the three genomes. Palacios and Wernegreen explained the reduced usage of aromatic residues in high-expression genes by energetic optimization and avoidance of costly residues. Our results therefore allow a generalization of the results of Palacios and Wernegreen (2002) to the three *Buchnera* genomes and give further support to their argument that amino acid usage in this symbiont is shaped by both selection against AT-richness and selection against aromaticity. However, a closer look at the parallel evolution of amino acid usage between *Buchnera* genes and their *E. coli* orthologs allowed us to refine this conclusion, and hierarchize these selective forces. This was done through the comparison of the levels of energetic optimization in the symbiont and its free-living relative. This analysis revealed

Table 6. Degree of Similarity Between Amino Acid Usage Profiles of *E. coli* and *Buchnera* Genes, as Estimated by the Mean Euclidian Distances (*d*) Between *Buchnera* Genes and Their *E. coli* Ortholog in the 2-D Plan Defined by the First Two Axes of a Factorial Correspondence Analysis of Relative Amino Acid Usage, Made on Three Complete *Buchnera* Sequences and Their *E. coli* Orthologs

	$d_{Aps\text{-}Eco}$	n	$d_{Sgr\text{-}Eco}$	n	$d_{Bpi\text{-}Eco}$	n
Genes shared by the three *Buchnera*	**0.422**	459	**0.444**	459	**0.456**	459
Genes lost in one or two *Buchnera*	**0.502**	105	**0.526**	90	**0.556**	48
Shared/lost comparison	***		***		***	

These distances are computed separately for genes shared by the three *Buchnera* and functional, and for genes lost or pseudogene in at least one of the two other *Buchnera* genomes, compared to the reference. Bottom line: significance of Student tests for comparisons between the 'shared' and 'lost' categories; *** ($P < 0.001$).

that several trends that discriminated high-expression and low-expression genes in *E. coli* were substantially altered in *Buchnera*; although the *E. coli* sequences are not ancestral, it can be assumed that most of the directional changes in frequency occurred in *Buchnera*. For example, in *Buchnera*, putative low-expression genes sharply increased their usage of Lys, whereas usage of Lys remained stable between *E. coli* and *Buchnera* putative high-expression genes. Interestingly, Lys was among the residues used preferentially in high-expression genes of *E. coli*, which suggests that the trends shaping amino acid preferences in the free-living relative have been altered in the symbiont. In contrast, putative low-expression *Buchnera* genes significantly decreased their usage of Trp, which remained stable (or even slightly increased) in putative high-expression *Buchnera* genes. Because Trp is avoided by high-expression *E. coli* genes, this trend has been again altered in the symbiont, although it still subsists and still allows discrimination of putative high- and low-expression genes in *Buchnera* (Palacios and Wernegreen 2002).

Both changes (relative decrease of Trp and increase of Lys in putative low expression genes) can be explained by weaker purifying selection against AT enrichment in these genes. But this supports the idea that selection against AT-richness is the dominating factor shaping amino acid usage in the symbiont, whereas selection against aromaticity only comes next. Energetic optimization appears therefore as yet another selective force that has been eroded in this symbiont, and the discrimination between low- and high-expression genes appears firstly as the result of greater conservation and resistance to AT enrichment in the former, the two effects being tightly linked ($\rho_{Ka\ -\ \%AT12} = 0.77$).

Drift has obviously played a major role in these symbionts and eroded all selective forces shaping the genome in free-living bacteria, that is, translational selection, replicational selection, and energetic optimization of amino acid usage. It is not obvious however to qualify these evolutionary patterns as strictly deleterious, because we lack a point of comparison with symbionts that would not have suffered these changes. Also, relaxed selection in a less competitive environment may have explained a part of the genomic changes observed; for example, the speed of replication is considerably lower in *Buchnera*, probably because it is under the control of the host, which adjusts the development of its bacterial population to its own growth (Baumann and Baumann 1994). This may further explain why replicational selection has become less effective in the symbiont. *Buchnera* is also characterized by the loss of many regulatory proteins and functions (Shigenobu et al. 2000), suggesting that the range in level of expression is considerably altered compared to *E. coli*. This could result in a relaxed translational selection and the loss of codon bias. It is therefore risky to predict that aphid/symbiont pairs would be doomed because of gradual decay of the symbiotic genome.

Indeed, bacterial organelles of eukaryotes continue to be functional despite long-term asexuality, transgenerational bottlenecks, and reduced population sizes. In a severely reduced genome, the selective coefficient of individual amino acid changes might increase, reinforcing the power of purifying selection. Therefore, we might expect a slow-down of evolution in at least the most essential genes which keep the symbiosis functional, as suggested by the reduced AT enrichment and the reduced evolutionary rates in the group of high-CAI genes. Finally, other compensatory processes might be employed by these bacteria (Moran 1996), with a possible special role for GroEL which may allow the bacterial genome to cope with an increased number of mutations. This biological model would therefore be an example of the theoretical system discussed by Hartl and Taubes (1996), where many (therefore not so) deleterious mutations might be compensated by positive selection—which here would be concentrated on essentially one gene, as supported by recent molecular analyses (Fares et al. 2002b)—resulting in approximate stasis of the whole.

Although the three *Buchnera* compared in this study have shown rather similar patterns, it is interesting to note that the differences observed can bring interesting insights regarding the tempo and mode of genome evolution following symbiosis. Indeed, the more complete loss of codon bias and the much stronger strand asymmetry (influencing even amino acid usage) in *Buchnera*-Bpi, along with its more reduced genome, seem to indicate that this bacterium has gone further than the other two lineages in the genomic revolution accompanying the transition to symbiosis. This is further supported by the increased differences of amino acid profiles between *Buchnera*-Bpi genes and their *E. coli* orthologs compared to the two other *Buchnera*, as revealed by the distances calculated in the FCA gathering the four genomes. In agreement with this result, a phylogeny based on 61 concatenated conserved proteins, including the three *Buchnera* studied here and several other symbiotic or free-living bacteria, revealed a significantly longer branch length for *Buchnera*-Bpi (Gil et al. 2003) than for the two other *Buchnera* species considered. Finally, several genes showed increased proportions of nonsynonymous substitutions in *Buchnera* from aphids of the family Pemphigidae versus Aphididae (Clark et al. 1999).

Another interesting result from the latter analysis was that genes lost in one *Buchnera* lineage have considerably more differentiated amino acid profiles from *E. coli* in the *Buchnera* genomes where they are maintained, compared to the core group of genes shared and functional in the three symbiotic genomes. Also, the comparison between *Buchnera*-Aps and *Buchnera*-Sgr revealed increased Ka and Ks values for the genes present in these species but lost in *Buchnera*-Bpi. Together these findings suggest that genes lost in any *Buchnera* genome (but present in another one) correspond to more labile, less essential functions, and that they are subject to less intense purifying selection in the genomes where they are maintained. We can therefore predict that genes lost in any *Buchnera* have ultimately a strong chance to meet the same fate in the other symbiotic lineages, once many slightly deleterious mutations have accumulated in their sequence. In contrast, the core group of shared functional genes is more constrained (as revealed by their stronger proximity to the *E. coli* profile) and might allow the definition of a minimal set of genes essential to the viability of the symbiosis. It is possible that the stronger syndrome observed in *Buchnera*-Bpi could be related to relatively smaller population sizes of aphids of the family Pemphigidae (to which the host of *Buchnera*-Bpi belongs). It will likely be instructive when more sequences of *Buchnera* become available to relate the changes among genomes with differences in the ecology of their aphid hosts, and also with the genomic repertoires of symbionts, in order to better weigh the different forces, mutational or selective, intrinsic or extrinsic, that shaped their genomes.

METHODS

Buchnera sp. and *E. coli* Sequences

Complete genomes of *Buchnera*-Aps and *Buchnera*-Sgr were extracted from GenBank, and the genome annotation of *Buchnera*-Bpi (van Ham et al. 2002) was provided by the research group directed by A. Moya (Univ. Valencia, Spain). We compared the gene content of the three *Buchnera* sp. by reciprocal top scores of gapped BLAST in order to perform an alignment of the three genomes. Excluding plasmid and RNAs genes, we reconstructed an ancestor of *Buchnera* sp. with 609 nonredundant coding genes and compared it to the free-living relative *E. coli*-K12 gene rep-

ertoire. We finally identified 597 unquestionable orthologous genes in *E. coli* by using a similarity criterion measured by BLAST with a cutoff expect value of 10^{-5}.

Factorial Correspondence Analysis on Relative Frequencies of Synonymous Codons

We performed a factorial correspondence analysis (FCA) of codon usage for the three complete sequences (discarding plasmid sequences, which might respond to specific selective/mutational pressures). We followed methods described elsewhere (Chiapello et al. 1998, Kanaya et al. 1999), computing *relative* frequencies of informative codons (that sum up to 1 for each amino acid), to make the analysis independent from gene size and amino acid composition. The three stop codons and codons unique to an amino acid (Trp and Met) were not counted. For each amino acid in a given gene, codon frequencies were calculated; in order to reduce random noise, they were replaced by mean frequencies if the number of residues was less than five. The analysis was performed using the CORRESP procedure of the SAS software (SAS Institute Inc. 1988). Correlations between the coordinates on the first axes of the FCA and relevant genomic parameters were analyzed using nonparametric tests (Spearman's correlation coefficient, rs), a choice motivated by the nonnormality of the data, in particular for *Buchnera*-Bpi. The significance of the correlations at the .05 level was assessed, using Bonferroni's correction (Sokal and Rohlf 2003) in order to reduce the probability of a type I error in multiple tests (we assumed independence of the correlation tests between genomes and between axes within a genome). Finally, the discrimination among genes oriented on the leading and lagging strands, respectively, was quantified using the DISCRIM procedure of the SAS software, for each of three *Buchnera* complete sequences (SAS Institute Inc. 1988).

Estimation of Codon Bias by Systematic χ^2 Tests

We also tested nonrandom use of codons at eight fourfold degenerate sites (plus the threefold Ile) by χ^2 tests, a method usually more sensitive to distinguish biases from random usage than multivariate analysis (Ermolaeva 2001). A second motivation was to provide a comparison with a previous analysis (Wernegreen and Moran 1999) that was made on a more limited sample of genes. For *Buchnera*, G- and C-ending codons were excluded because of the small sample size of these codons, due to the low GC content at third positions. In their analysis, Wernegreen and Moran used an expected frequency determined by the "local base composition," a method which might be less sensitive because it integrates potential gene-specific biases in the reference value. Because the factorial correspondence analysis had shown that gene orientation was the major local determinant of codon usage, we distinguished the genes by their presence on the leading or lagging strand. For each orientation, we determined the expectation based on a pooled set of genes expected to be the most dominated by mutational bias (choosing 35 genes with the lowest CAI in *E. coli*), assuming that adaptive codon bias should increase steadily with CAI; χ^2 values were averaged across all amino acids, and divided by gene size to allow a normalized comparison of the level of bias.

Factorial Correspondence Analysis on Relative Frequencies of Amino Acids

We similarly performed an FCA of amino acid usage. The variables were the frequencies of the 20 amino acids (summing up to 1 for each gene). This analysis was done using the CODONW software (written by John Peden, Univ. of Nottingham, U.K.), for each of three *Buchnera* complete sequences. This software calculates hydrophobicity and aromaticity levels for each gene, which were correlated with positions on the main discriminating axis.

Finally, in an attempt to directly compare amino acid profiles between the three symbionts and their free-living relative, a correspondence analysis was conducted pooling the data of the four genomes. Orthologs were "aligned" in order to identify which genes might have changed most (or least) compared to *E. coli*, a comparison made between and within *Buchnera* genomes.

ACKNOWLEDGMENTS

We thank Carmen Palacios and Jenn Wernegreen, and an anonymous reviewer, for very helpful comments and suggestions which helped us improve the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

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

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

Albert, B., Godelle, B., Atlan, A., De Paepe, R., and Gouyon, P.-H. 1996. Dynamics of plant mitochondrial genome: Model of a three-level selection process. *Genetics* **144:** 369–382.

Andersson, S.G.E., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontén, T., Alsmark, U.C., Podowski, R.M., Näslund, A.K., Eriksson, A.-S., Winkler, H.H., and Kurland, C.G. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396:** 133–143.

Baumann, L. and Baumann, P. 1994. Growth kinetics of the endosymbiont *Buchnera aphidicola* in the aphid *Schizaphis graminum*. *Appl. Environ. Microbiol.* **60:** 3440–3443.

Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C.K., Mayhew, G.F., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1462.

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

Chiapello, H., Lisacek, F., Caboche, M., and Henaut, A. 1998. Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene* **209:** GC1–GC38.

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

Clark M.A., Moran, N.A., Baumann, P., and Wernegreen, J.J. 2000. Cospeciation between bacterial endosymbionts (*Buchnera*) and a recent radiation of aphids (*Uroleucon*) and pitfalls of testing for phylogenetic congruence. *Evolution* **54:** 517–525.

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

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

Fares, M.A., Ruiz-González, M.X., Moya, A., Elena, S.F., and Barrio, E. 2002a. GroEL buffers against deleterious mutations. *Nature* **417:** 398.

Fares, M.A., Barrio, E., Sabater-Muñoz, B., and Moya, A. 2002b. The evolution of the heat-shock protein GroEL from *Buchnera*, the primary endosymbiont of aphids, is governed by positive selection. *Mol. Biol. Evol.* **19:** 1162–1170.

Frank, A.C. and Lobry, J.R., 1999. Asymmetric substitution patterns: A review of possible underlying mutational or selective mechanisms. *Gene* **238:** 65–77.

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

Gil, R., Silva, F.J., Zientz, E., Delmotte, F., González-Candelas, F., Latorre, A., Rausell, C., Kamerbeek, J., Gadau, J., Hölldobler, B., et al. 2003. The genome sequence of *Blochmannia floridanus*: Comparative analysis of reduced genomes *Proc. Natl. Acad. Sci.* **100:** 9388–9393.

Graur, D. and Li, W.-H., 2000. *Fundamentals of molecular evolution*, 2nd ed., chapter 4. Sinauer, Sunderland.

Hartl, D.L. and Taubes, C.H. 1996. Compensatory nearly neutral mutations: selection without adaptation. *J. Theor. Biol.* **182:** 303–309.

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

Ishikawa, H. 1984. Characterization of the protein species synthesised in

vivo and in vitro by an aphid endosymbiont. *Insect Biochem.* **14:** 417–425.
Itoh, T., Martin, W., and Nei, M. 2002. Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proc. Natl. Acad. Sci.* **99:** 12944–12948.
Kanaya, S., Yamada, Y., Kudo, Y., and Ikemura, T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: Gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238:** 143–155.
King Jordan, I., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12:** 962–968.
Kyte, J. and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157:** 105–132.
Lambert, J.D. and Moran, N.A. 1998. Deleterious mutations destabilize ribosomal RNA in endosymbioytic bacteria. *Proc. Natl. Acad. Sci.* **95:** 4458–4462.
McInerney, J.O. 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci.* **95:** 10698–10703.
McLean, M.J., Wolfe, K.H., and Devine, K.M., 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol Evol.* **47:** 691–696.
Médigue, C., Rouxel, T., Vigier, P., Hénaut, A., and Danchin, A. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Evol.* **222:** 851–856.
Mira, A., Ochman, H., and Moran, N.A. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **10:** 589–596.
Moran, N.A. 1996. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci.* **93:** 2873–2878.
Moran, N.A. and Baumann, P. 1994. Phylogenetics of cytoplasmically inherited microorganisms of arthropods. *Trends Ecol. Evol.* **9:** 15–20.
Moya, A., Latorre, A., Sabater-Muñoz, B., and Silva, F.J. 2002. Comparative molecular evolution of primary (*Buchnera*) and secondary symbionts of aphids based on two protein-coding genes. *J. Mol. Evol.* **55:** 125–137.
Ochman, H. and Moran, N.A. 2001. Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science* **292:** 1096–1098.
Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23:** 263–286.
Palacios, C. and 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.
Rispe, C. and Moran, N.A. 2000. Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. *Am. Nat.* **156:** 425–441.
SAS Institute Inc. 1988. SAS/STAT User's guide, Release 6.03 Edition. Sas Institute Inc., Cary, NC.
Sharp, P.M. and Li, W.H. 1987. The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15:** 1281–1295.
Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. 2000. Genome sequence of the endocellular bacterial symbionts of aphids *Buchnera* sp. APS. *Nature* **407:** 81–86.
Silva, F.J., Latorre, A., and Moya, A. 2001. Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. *Trends Genet.* **17:** 615–618.
Sokal, R.R. and Rohlf, F.J. 2003. *Biometry: The principles and practice of statistics in biological research*, 3rd ed. W.H. Freeman and Co., New York.
Tamas, I., Klasson, L., Canbäck, B., Näslund, A.K., Eriksson, A.-S., Wernegreen, J.J., Sandström, J., Moran, N.A., and Andersson, S.G.E. 2002. Fifty million years of genomic stasis in endosymbiotic bacteria. *Science* **296:** 2376–2379.
Van Ham, R.C.H.J., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernández, J.M., Jiménez, L., Postigo, M., Silva, F.J., et al. 2002. Reductive genome evolution in *Buchnera aphidicola*. *Proc. Natl. Acad. Sci.* **100:** 581–586.
Wernegreen, J.J. 2002. Genome evolution in bacterial endosymbionts of insects. *Nat. Rev. Genet.* **3:** 850–861.
Wernegreen, J.J. and Moran, N.A. 1999. Evidence for genetic drift in endosymbionts (*Buchnera*): Analyses of protein-coding genes. *Mol. Biol. Evol.* **16:** 83–97.

Received March 21, 2003; accepted in revised form October 8, 2003.