A Phylogenomic Study of Endosymbiotic Bacteria

Björn Canbeck,1 Ivica Tamas,2 and Siv G. E. Andersson

Department of Molecular Evolution, Evolutionary Biology Center, University of Uppsala, Uppsala, Sweden

Endosymbiotic bacteria of aphids, Buchnera aphidicola, and tsetse flies, Wigglesworthia glossinidia, are descendents of free-living γ-Proteobacteria. The acceleration of sequence evolution in the endosymbiotic genomes is here estimated from a phylogenomic analysis of the γ-Proteobacteria. The tree topologies associated with the most highly conserved genes suggest that the endosymbionts form a sister group with Escherichia coli, Salmonella sp., and Yersinia pestis. Our results indicate that deviant tree topologies result from high substitution rates and biased nucleotide patterns, rather than from lateral gene transfer, as previously suggested. A reinvestigation of the relative rate increase in the endosymbiont genomes reveals variability among genes that correlate with host-associated metabolic dependencies. The conclusion is that host-level selection has retarded both the loss of genes and the acceleration of sequence evolution in endocellular symbionts.

Introduction

Endosymbiotic bacteria of aphids (Buchnera aphidicola) and tsetse flies (Wigglesworthia glossinidia) have undergone massive genome reductions, with little or no influx of external DNA since their divergence from a free-living γ-proteobacterial ancestor (Shigenebou et al. 2000; Tamas et al. 2002; Akman et al. 2002; van Ham et al. 2003). No rearrangements, duplications, or horizontal gene transfer events were detected in the genomes of two aphid endosymbiont species that diverged more than 50 MYA (Tamas et al. 2002). In contrast, free-living γ-Proteobacteria, such as Escherichia coli, Salmonella typhimurium, Vibrio cholerae, and Haemophilus influenzae are highly dynamic in genome structures with more than 2000-fold higher ratios of insertion/deletion and rearrangement mutations relative to nonsynonymous substitutions (Tamas et al. 2002). The difference is attributed to two forces acting on the evolution of endosymbiotic genomes: reduced rearrangement frequencies and enhanced rates of nucleotide sequence evolution (Tamas et al. 2002).

An important but, yet, unresolved question concerns the evolutionary forces that drive the acceleration of nucleotide substitutions in endosymbiotic populations. Factors such as enhanced mutation rates, reduced levels of purifying selection, and increased fixation rates for slightly deleterious mutations combined with the effects of Mullers ratchet on small, asexual populations have been postulated to account for the rate increase (Moran 1996; Andersson and Kurland 1998; Brynnel et al. 1998; Clark, Moran, and Baumann 1999). A recent study revealed a uniform rate increase for B. aphidicola genes relative to those of E. coli, suggesting that the elevated substitution rate is solely the result of a mutation rate increase (Itoh, Martin, and Nei 2002). This conclusion was drawn from an analysis that only included one-third of all B. aphidicola genes. These were selected on the basis that they produced phylogenetic trees in which B. aphidicola formed a sister group with E. coli to the exclusion of H. influenzae (Itoh, Martin, and Nei 2002). The majority of genes, two-thirds, were associated with deviant tree topologies, and these were excluded from the analyses based on the assumption that they could represent horizontal gene transfer events (Itoh, Martin, and Nei 2002).

A large number of conflicting tree topologies is a characteristic feature of whole-genome–based phylogenetic analyses (Sicheritz-Ponten and Andersson 2001; Eisen and Fraser 2003), and these incongruent topological features have been used as evidence for horizontal gene transfer events (Gogarten, Doolittle, and Lawrence 2002; Doolittle et al. 2003). Taken to the extreme, it is argued that the widespread occurrence of such anomalies challenges the concept of a universal tree of life (Doolittle 1999; Doolittle et al. 2003). Others, however, argue that a core set of genes with a common history can be identified that reflects the true species tree, despite the many conflicting tree topologies (Daubin, Gouy, and Perriere 2002; Daubin, Moran, and Ochman 2003; Rokas et al. 2003). Alternative explanations for unexpected placements of taxa in single-gene trees include long-branch artifacts, gene paralogy, and lack of resolution because of saturation effects or low levels of sequence divergence (Eisen and Fraser 2003; Daubin, Moran, and Ochman 2003).

Because the correct species tree is normally not known, there are few realistic scenarios under which the sources and implications of deviant tree topologies can be examined. The aphid endosymbiont lineages provide an interesting exception; because all genes in these genomes are thought to share the same evolutionary history, they represent an ideal reference system for phylogenomic analyses. Because it is of general interest to examine whether unexpected placements of taxa in single-gene trees represent horizontal gene transfer events (Doolittle et al. 2003; Gogarten, Doolittle, and Lawrence 2002; Daubin, Gouy, and Perriere 2002; Daubin, Moran, and Ochman 2003; Eisen and Fraser 2003; Kurland, Canbäck, and Berg 2003) and of specific interest to examine the causes for lineage-specific rate increases in endosymbiotic bacteria (Itoh, Martin, and Nei 2002), we have here used a phylogenomic approach to re-examine the congruence of gene trees from B. aphidicola. Our results suggest that distorted tree topologies for endosymbiont genes result from biased mutation pressures and enhanced substitution rates, rather than from horizontal gene transfer events. The analysis provides evidence for relative rate variations.
among functional classes of genes in endosymbiotic bacteria of aphids and tsetse flies, indicating that the acceleration of evolutionary rates is affected by purifying selection acting at the host level.

Materials and Methods
Strategy for Screening the $B$. aphidicola Phylome

To study the phylome of $B$. aphidicola, that is, the total set of trees generated from the complete set of genes in the $B$. aphidicola genome, we performed three series of whole-genome–based phylogenetic studies that differ with respect to the aim of the study, the reconstruction methods used, and the set of species included in the analysis. For each global analysis, smaller subsets of genes were compiled and analyzed to illustrate the results obtained in the whole-genome–based investigation.

For the first global phylogenetic survey, we reconstructed phylogenetic trees from protein alignments that contained all of the 545 inferred $B$. aphidicola (Sg) gene products and the homologous protein sequences from other proteobacterial species. Based on the results, we conducted a second study to test the hypothesis that different tree topologies were obtained as a result of rate heterogeneity and nucleotide composition bias in the data set. In this case, phylogenetic trees were built from a more restricted data set that consisted of 180 $B$. aphidicola genes aligned with their orthologs from free-living proteobacterial species for which complete genome data were available. Finally, to examine rate variation among endosymbiont genes, we determined the relative rate of sequence evolution for orthologous genes of $B$. aphidicola and $W$. glossinidia from the branch lengths of these lineages in 258 phylogenetic trees that also included four newly sequenced proteobacterial species (including $W$. glossinidia).

Survey of the $B$. aphidicola Phylome: Selection of Species, Genes, and Methods

To perform a first survey of the $B$. aphidicola phylome, two databases were created with all 545 putative protein sequences of $B$. aphidicola (Sg), along with all proteobacterial sequences found in the SPTR protein database (accessible as SWALL on the EBI SRS server, http://srs6.ebi.ac.uk/). At the time of the analysis (April 2002), the total number of sequences in the SPTR database amounted to 127,974. The 545 putative proteins encoded by the genome of the Buchnera (Sg) were searched for homologous proteins in the constructed proteobacterial database by using the Blast version 2.0 algorithm (Altschul et al. 1997) with a threshold E-value of $1 \times 10^{-10}$. The amino acid sequences for the 30 best hits were aligned together with the query sequence using ClustalW version 1.8 (Thompson, Higgins, and Gibson 1994).

As many as 518 alignments contained four or more taxa, from which minimum-evolution (ME) trees were reconstructed with raw amino acid fraction differences as distances and 500 bootstrap replicates by using PAUP* version 4.0b10 for Unix (Swofford 1998). In each reconstruction, out-groups were defined as the species with the lowest scoring Blast hit among the $\alpha$-Proteobacteria, $\beta$-Proteobacteria, $\delta$-Proteobacteria, or $\varepsilon$-Proteobacteria (i.e., among taxa that are not members of the $\gamma$-Proteobacteria). In 41 trees, only $\gamma$-proteobacterial homologs were identified. Consequently, no out-group could be defined, and these genes were excluded from further analyses, as were also 121 trees that did not contain genes from either $E$. coli, $H$. influenzae, or both. Another 30 trees were unresolved at the nodes of interest, and these were also excluded from the analysis.

For the 326 remaining trees, the relationships between $B$. aphidicola, $E$. coli, and $H$. influenzae were analyzed. If two species shared a node exclusive of the third species, a relationship between these two species was assumed for that particular protein. The resulting tree topologies differed mainly with respect to the placement of $B$. aphidicola. To illustrate the difference among the most frequently observed tree topologies, nine or 10 proteins were selected for a more detailed investigation of topology structures. The selected genes supported one of the alternative topologies with high bootstrap support values. The sequences of the selected proteins were concatenated, aligned, and subjected to phylogenetic reconstructions using both neighbor-joining (Saitou and Nei 1987) and maximum-parsimony methods. Bootstrap trees were calculated with 500 replicates in both cases.

Hypothesis Testing: Selection of Species, Genes, and Methods

To examine the influence of nucleotide sequence heterogeneity on tree topologies, a database was created that contained all nucleotide and protein sequences available from completely sequenced proteobacterial genomes. Thus, a restricted set of species was included in the analysis (table 1) so as to produce phylogenetic trees with high consistency and, thereby, reduce species sampling effects and, most importantly, to allow simple comparisons among trees. In the few cases where two or more strains had been sequenced from the same species, only one strain was selected. All the 545 putative proteins encoded by the $B$. aphidicola (Sg) genome were searched for homologous proteins in the proteobacterial genome database using a threshold E-value of $1 \times 10^{-5}$. Proteins without homologs in all of the defined $\gamma$-proteobacterial species (table 1) and the selected out-group taxa were discarded from further analysis. As out-groups we selected the $\alpha$-Proteobacteria Rickettsia conorii and Rickettsia prowazekii and the $\beta$-Proteobacteria Neisseria meningitidis and Ralstonia solanacearum, R. prowazekii and R. conorii were selected as out-groups because of their A+T-richness so as to also test for possible A+T artifacts with the A+T-rich $B$. aphidicola. After this screening process, 180 genes remained for an in-depth phylogenetic analysis.

Protein sequences were aligned using ClustalW 1.8 (Thompson, Higgins, and Gibson 1994). Protein trees were constructed with PAUP* (Swofford 1998) with minimum evolution using the raw amino acid fraction as distances and 500 bootstrap replicates. Because biased base composition may affect amino acid usage, we also wanted...
to explore substitution models that account for asymmetric substitution rates and biased nucleotide composition. Nucleotide sequence alignments were created from the protein alignments by replacing each amino acid with the corresponding codon. Gene trees were constructed with two methods based on the first and second codon positions in the gene alignments.

First, the neighbor-joining method (Saitou and Nei 1987) was used with the Galtier and Gouy substitution model (Galtier and Gouy 1995) that accommodates asymmetric substitution frequencies (e.g., the substitution rate from C to G may not equal the rate from G to C), unequal base compositions, and different transversion/transition rates and is implemented in Phylo_win (Galtier, Gouy, and Gautier 1996). Second, maximum-likelihood trees were built with the aid of PAUP*. Parameters used in the likelihood analyses were estimated with Modeltest (Posada and Crandall 1998), which helps to determine whether to use a gamma distribution, the size of the associated shape parameter, and the proportion of invariant sites and selects an appropriate substitution model. Likelihood estimations are demanding from a computational perspective. Therefore, we restricted the search in PAUP* by using only one repetition in the heuristic search. For the bootstrapped trees, 10 replicates were generated using the FastStep search method with HKY85 as substitution model and a gamma distribution with a shape parameter estimated by Modeltest.

To illustrate the different topologies obtained in this analysis, we selected four sets of genes that supported each of four different topologies with high bootstrap support values in the minimum-evolution trees. Nucleotide alignments were created by the concatenation of the individual bootstrap support values. Phylogenetic relationships were inferred from the first and second codon positions with the methods described above. However, with the limited number of taxa, the likelihood method could now be used with heuristic searches with 10 repetitions. For the bootstrapped trees, 100 replicates were generated using FastStep search. In both cases, all estimated parameters from Modeltest were used.

Studies of Rate Variation Among Genes: Selection of Species, Genes, and Methods

For the comparative analysis of the relative rate increase in endosymbiont genomes, we compiled a third data set. In addition to the selected set of species described above, this data set also included the gene and protein sequences of (by the time) newly sequenced Proteobacteria W. glossinidia, Shigella flexneri, Shewanella oneidensis, and Brucella suis (table 1). In this analysis, we selected as out-groups the α-Proteobacteria Caulobacter crescentus and Mesorhizobium loti, together with the β-Proteobacteria N. meningitidis and R. solanacearum. The genomes of C. crescentus and M. loti have a relatively high gene content as compared with R. prowazekii and R. conorii, which allowed more genes to be kept for the analysis.

The resulting 258 sequences were aligned with ClustalW 1.8 (Thompson, Higgins, and Gibson 1994), with default settings. From the protein alignments, minimum-evolution (ME) trees were constructed with the aid of PAUP* 4.0b10 for Unix (Swofford 1998). Five-hundred bootstrap replicates were produced, with the raw amino acid fraction differences as distances. Phylogenetic trees were also reconstructed based on first and second codon positions using the neighbor-joining method and the Galtier and Gouy (1995) substitution model. Branch-length ratios between taxa (table 2) were calculated from trees derived from the Galtier and Gouy (1995) method by dividing the branch length from the common node to the external node (taxon) for one of the taxons with the corresponding length for the other taxon.

Table 1
Species and Strains Included in the Phylogenomic Analyses

<table>
<thead>
<tr>
<th>Species/strain</th>
<th>Subdivision</th>
<th>Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agrobacterium tumefaciens C58</td>
<td>α</td>
<td>NC_003062–5</td>
</tr>
<tr>
<td>Brucella melitensis</td>
<td>α</td>
<td>NC_003317–8</td>
</tr>
<tr>
<td>Brucella suis 130b</td>
<td>α</td>
<td>NC_004310–1</td>
</tr>
<tr>
<td>Caulobacter crescentus</td>
<td>α</td>
<td>NC_002696</td>
</tr>
<tr>
<td>Mesorhizobium loti</td>
<td>α</td>
<td>NC_002678–9, NC_002682</td>
</tr>
<tr>
<td>Rickettsia conorii</td>
<td>α</td>
<td>NC_003103</td>
</tr>
<tr>
<td>Rickettsia prowazekii</td>
<td>α</td>
<td>NC_000963</td>
</tr>
<tr>
<td>Sinorhizobium meliloti</td>
<td>α</td>
<td>NC_003037, NC_003047, NC_003078</td>
</tr>
<tr>
<td>Neisseria meningitidis</td>
<td>MC58</td>
<td>β</td>
</tr>
<tr>
<td>Ralstonia solanacearum</td>
<td>β</td>
<td>NC_003295–6</td>
</tr>
<tr>
<td>Buchnera aphidicola (Ap)</td>
<td>γ</td>
<td>NC_002252–3, NC_002528</td>
</tr>
<tr>
<td>Buchnera aphidicola (Sg)</td>
<td>γ</td>
<td>NC_004061</td>
</tr>
<tr>
<td>Escherichia coli K12</td>
<td>γ</td>
<td>NC_000913</td>
</tr>
<tr>
<td>Haemophilus influenzae</td>
<td>γ</td>
<td>NC_000907</td>
</tr>
<tr>
<td>Pasteurella multocida</td>
<td>γ</td>
<td>NC_002663</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>γ</td>
<td>NC_002516</td>
</tr>
<tr>
<td>Salmonella typhi</td>
<td>γ</td>
<td>NC_003198, NC_003384, NC_003385</td>
</tr>
<tr>
<td>Salmonella typhimurium LT2</td>
<td>γ</td>
<td>NC_003197, NC_003277</td>
</tr>
<tr>
<td>Shewanella oneidensis</td>
<td>γ</td>
<td>NC_004347, NC_004349</td>
</tr>
<tr>
<td>Shigella flexneri 2a</td>
<td>γ</td>
<td>NC_004337</td>
</tr>
<tr>
<td>Vibrio cholerae</td>
<td>γ</td>
<td>NC_002505, NC_002506</td>
</tr>
<tr>
<td>Wigglesworthia brevipalpis</td>
<td>γ</td>
<td>NC_004344</td>
</tr>
<tr>
<td>Xanthomonas axonopodis</td>
<td>γ</td>
<td>NC_003919</td>
</tr>
<tr>
<td>Xanthomonas campestris</td>
<td>γ</td>
<td>NC_003902</td>
</tr>
<tr>
<td>Xylella fastidiosa</td>
<td>γ</td>
<td>NC_002488–90</td>
</tr>
<tr>
<td>Yersinia pestis</td>
<td>γ</td>
<td>NC_003131–2, NC_003134, NC_003143</td>
</tr>
<tr>
<td>Campylobacter jejuni</td>
<td>ε</td>
<td>NC_002163</td>
</tr>
<tr>
<td>Helicobacter pylori</td>
<td>ε</td>
<td>NC_000915</td>
</tr>
</tbody>
</table>

* Species that were only included in the analysis of rate variation among genes in endosymbiont genomes.
Genomic Positions

Tree Topologies, Functional Roles, and Results

B. aphidicola of species were estimated with CODEML from the PAML were then replaced by the corresponding codons. From the neighbor-joining method with the Galtier and Gouy substitution model. The functional categories were taken from the annotation of the B. aphidicola (Sg) genome (Tamas et al. 2002).

were then replaced by the corresponding codons. From the resulting alignments nonsynonymous substitution frequencies were estimated with CODEML from the PAML package (Yang 1997). For each gene in B. aphidicola (Sg) the G+C content value was calculated.

Results

Tree Topologies, Functional Roles, and Genomic Positions

In a first survey of the B. aphidicola phylome, a total of 518 protein trees were produced from the complete set of B. aphidicola genes using the minimum-evolution algorithm on protein alignments (ME_pro). A subset of 326 trees contained homologs in B. aphidicola, E. coli, and H. influenzae and the out-group species from the α-Proteobacteria, β-Proteobacteria, δ-Proteobacteria, and/or ε-Proteobacteria. In 78 cases, B. aphidicola clustered together with E. coli; these trees will here be referred to as the A-type of trees. The remaining 248 trees suggested that E. coli and H. influenzae are sister taxa, with B. aphidicola diverging before the separation of these two species, although the exact position of B. aphidicola in relation to the other γ-proteobacterial species was found to be gene dependent. The latter category of trees will here be referred to as the B-type of trees.

To illustrate the different topologies obtained, we concatenated three sets of approximately 10 genes in each set that were selected on the basis that they supported each of the conflicting tree topologies with high bootstrap support values (fig. 1). Gene order is conserved between E. coli and B. aphidicola in a shared set of 91 fragments that contain a total of 503 genes in B. aphidicola (Ap) (Moran and Mira 2001). Genes producing alternative tree topologies mapped onto these fragments in a random manner (data not shown), with no indications of clusters of genes having similar histories, as might be expected if longer DNA segments had been acquired by horizontal gene transfer. We were also unable to find any support for a correlation between tree topologies and functional categories in B. aphidicola, as would be the case if genes with different functions had different ancestries.

Tree Topologies Are Influenced by Rates and Patterns of Sequence Evolution

A high fraction of deviant tree topologies was also noted in previous studies of the phylogenetic placement of B. aphidicola in γ-proteobacterial gene trees (Itoh, Martin, and Nei 2002; Snell, Bork, and Huynen 2002). One interpretation is that the atypical topologies represent horizontal gene transfer events (Itoh, Martin, and Nei 2002); another is that they are the result of long-branch attraction problems (Snell, Bork, and Huynen 2002). It is well known that procedures commonly used for evolutionary analysis can be misleading in sequences with biased base composition (Eyre-Walker 1998; Galtier and Gouy 1998). Because B. aphidicola is a rapidly evolving lineage with a strong AT-bias that has influenced both its codon and amino acid usage (Palacios and Wernigreen 2002; Herbeck, Wall, and Wernigreen 2003; Rispe et al. 2004) methodological artifacts might be expected.

To examine the effect of biased rates and patterns of sequence evolution on the placement of B. aphidicola in individual gene trees, we compared topological features with substitution frequencies and genetic G+C content values. Nonsynonymous substitution frequencies (dN) were estimated for a set of orthologous genes in B. aphidicola (Sg) and (Ap). To begin, we noted a strong correlation between substitution frequencies and genetic G+C content values (fig. 2), indicative of a recent mutation bias towards A+T. Both of these parameters were found to correlate with tree topology structures (fig. 2). On the average, genes in the A-type category were found to be more slowly evolving and less biased towards A+T (dN = 0.13; G+C = 28.2% [n = 78 genes]) than genes in the B-type category (dN = 0.17; G+C = 26.2% [n = 248 genes]). Likewise, the topology favoring a recent divergence of B. aphidicola near to E. coli and Salmonella with 100% bootstrap support (fig. 1A) was associated with a set of genes that on the average are highly conserved and have low G+C content values (fig. 2A). In contrast, topologies providing evidence for deeper divergences of B. aphidicola within the γ-Proteobacteria with strong bootstrap support (fig. 1B, C) were obtained from sets of genes with higher substitution rates and higher A+T content values (fig. 2A).

Thus, our results argue against rampant horizontal gene transfer (Itoh, Martin, and Nei 2002) as the explanation for the deviant tree topologies and support the hypothesis that rapid rates of sequence evolution (Snell, Bork, and Huynen 2002) and biased protein composition have distorted some of the individual protein trees in such a way that B. aphidicola artificially diverges closer to the root of the γ-proteobacterial tree with increasing nucleotide sequence bias (figs. 1 and 2).

Application of Substitution Models for Unequal Base Composition

To improve the phylogenetic signal and minimize problems caused by the A+T-richness of the B. aphidicola

Table 2

Relative Rate Enhancement for Genes Involved in Different Functional Categories

<table>
<thead>
<tr>
<th>Functional Category</th>
<th>$B_{uwh}/W_{uw}^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.70 ± 0.25</td>
</tr>
<tr>
<td>Amino acid biosynthesis</td>
<td>0.60 ± 0.09</td>
</tr>
<tr>
<td>Cofactor biosynthesis</td>
<td>0.93 ± 0.18</td>
</tr>
<tr>
<td>Cell envelope</td>
<td>0.88 ± 0.29</td>
</tr>
<tr>
<td>Cell processes</td>
<td>0.78 ± 0.20</td>
</tr>
<tr>
<td>Energy metabolism</td>
<td>0.73 ± 0.16</td>
</tr>
<tr>
<td>Fatty acid</td>
<td>1.16 ± 0.06</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>0.70 ± 0.28</td>
</tr>
<tr>
<td>Other</td>
<td>0.81 ± 0.22</td>
</tr>
<tr>
<td>Purines, pyrimidines</td>
<td>0.84 ± 0.21</td>
</tr>
<tr>
<td>Replication</td>
<td>0.75 ± 0.17</td>
</tr>
<tr>
<td>Transcription</td>
<td>0.46 ± 0.13</td>
</tr>
<tr>
<td>Translation</td>
<td>0.60 ± 0.22</td>
</tr>
</tbody>
</table>

$^a$ The analysis was based on 116 gene trees supporting a clustering of B. aphidicola and W. glossinidia with bootstrap support values above 70%, using the neighbor-joining method with the Galtier and Gouy substitution model.
genes, we performed phylogenetic reconstructions using nucleotide first and second codon positions with the Galtier and Gouy substitution model (Galtier and Gouy 1995) applied to the neighbor-joining algorithm (NJGG). The Galtier and Gouy model was developed for estimating evolutionary distances for data sets with heterogeneous base composition and have previously been shown to compensate for such biases (Galtier and Gouy 1995). To eliminate species sampling effects and to make the individual gene trees as comparable as possible, we selected for this analysis a set of 180 genes, all of which are present in the γ-proteobacterial genomes examined, as

FIG. 1.—Phylogenetic relationships of a selected set of γ-Proteobacteria based on the concatenated gene products of (A) dapA, fisH, prxA, deaD, rpoD, rho, hflC, ybeX, yciA, and yidC, (B) rplAFIKLSUV and rpmA, and (C) aAS, glyS, hisS, leuS, metG, pheS, pheT, proS, and cca. Branch lengths are proportional to those reconstructed with the neighbor-joining method. Values at nodes indicate the percentages of 500 bootstraps in the neighbor-joining and maximum-parsimony methods, in this order. Stars at nodes indicate that the order of divergence was unresolved.
well as in four out-group taxa (table 1). A comparison of figure 2A and B shows that this gene set spans the entire range of G+C content values and nonsynonymous substitution frequencies in *B. aphidicola* and is thus a good representation of the complete gene set.

As in the previous analysis, several different topologies were obtained when the minimum-evolution algorithm was applied to each of the 180 protein alignments (MEprot). To illustrate the topology characteristics, we concatenated alignments of sets of individual proteins that supported each of four different topologies (A1, A2, B1, and B2) with high support. In topology A1 (fig. 3A), *B. aphidicola* formed a monophyletic group with *Y. pestis*, *E. coli*, and *Salmonella* species (the yess cluster). Topology A2 (fig. 3B) is similar to topology A1 except that *V. cholerae* is positioned in between the yess cluster and *B. aphidicola*. In the B-type topologies, *B. aphidicola* was found to diverge before *H. influenzae*, *Pasteurella multocida*, *V. cholera*, and the yess cluster. In the B1-type of trees (fig. 3C), *P. aeruginosa* diverged immediately after *B. aphidicola*, whereas in the B2-type of trees (fig. 3D), *B. aphidicola* diverged before *P. aeruginosa*.
FIG. 3.—Influence of the Galtier and Gouy substitution method on the position of *B. aphidicola* in γ-Proteobacteria gene trees. The phylogenetic relationships were inferred from concatenations of (A and E) *dapA*, *ftsH*, *hflC*, *rpoC*, *yidC*, and *yfjB*; (B and F) *dnaB*, *dnaK*, *ftsA*, *ftsW*, *lon*, *nusA*, *nusG*, and *ppiD*; (C and G) *dnaG*, *hslU*, *rplEF*, *rpoA*, *rps*BDE, *thr*S, *tsf*, and *yggJ*; and (D and H) *atpB*, *queA*, *rpsS*, *topA*, *ycfH*, and *ydhD*. Tree topologies are from (A–D) the minimum-evolution method applied to the concatenated protein sequence alignments (MEprot) and (E–H) the neighbor-joining method with the Galtier and Gouy substitution model (NJGG) (Galtier and Gouy 1995) and the maximum-likelihood (ML) method applied to the concatenated nucleotide sequence alignments. Third codon positions were removed. Branch lengths were taken from the NJGG reconstruction analyses. Values at nodes indicate the percentages of bootstraps obtained in NJGG and ML analyses in this order.
FIG. 3 (Continued)
The application of the NJ_{GG} methods to the first and second nucleotide positions of the four concatenated nucleotide alignments discussed above yielded strong support for the A1 topology in which B. aphidicola forms a monophyletic group with the vess cluster (fig. 3E–H). This topology was also obtained in each of the four cases with the implementation of the maximum-likelihood (ML) method (fig. 3E–H). The fraction of trees yielding the four tree topologies was radically altered when the NJ_{GG} model was applied to the complete data set of 180 genes (fig. 4). When considering topologies with high bootstrap support values (>80%), the fraction of trees supporting the A1 topology increased from 21% to 46%. Simultaneously, the fraction of trees supporting topology B2 (bootstrap >80%) decreased from 23% to 9%. Overall, one third of the individual gene trees shifted topology from the B-type to the A-type (bootstrap support values >80%). Most of the conversions were from the B1 to the A1 topology, but several shifts from the B2 to the B1 topology and from the B1 to the A2 topology were also observed. No examples of shifts in the reverse direction (from the A-type to the B-type) were identified. Also in this analysis, we found a strong correlation between tree topologies and biased mutation pressures (fig. 2B), except that a much larger fraction of genes yielded the A1-type topology than observed in the ME_{pro} reconstruction analyses.

When the ML method was applied to the 180 single-gene alignments, the A1 topology was the preferred topology. However, when bootstraps were applied, only a small fraction of these topologies were well supported: only 8% (>80% bootstrap support) and 18% (>50% bootstrap support). Unlike the ML reconstruction analysis, the substitution model used in the NJ_{GG} analysis accounts for both asymmetric substitution patterns and variable G+C content levels, which may explain the higher resolution obtained with this method. The strong support given to the A1 topology when the NJ_{GG} methods were applied to both single and concatenated gene alignments suggests that the grouping of B. aphidicola with the enterics is our best approximation of the underlying species tree.

**Acceleration of Genomic Evolution Is Counterbalanced by Host-Level Selection**

To study the role of host selection on endosymbiont genome evolution, we compiled a data set of 258 genes, all of which are present in B. aphidicola and W. glossinidia, as well as in the free-living γ-proteobacterial species selected and the out-group taxa M. loti, C. crescentus, N. meningitidis, and R. solanacearum. Also for this data set, approximately 50% of the trees yielded the A-type topology with strong support when the NJ_{GG} method was applied to the analysis. From the relative branch lengths in these trees, the rate of sequence evolution at nonsynonymous sites were estimated to be approximately twofold higher on the average in B. aphidicola than in E. coli (B_{avg}/E_{coli} = 2.4), with a less than twofold variation among functional categories (data not shown). However, we were unable to infer the full range of the rate enhancement from the branch lengths of B. aphidicola and E. coli to a shared node, because the most rapidly evolving genes are associated with deviant tree topologies.

Another approach to search for variations in evolutionary rates among genes is to estimate the relative rate enhancement in aphid and tsetse fly endosymbionts. The individual gene trees supported a clustering of B. aphidicola and W. glossinidia with high bootstrap support (>70%), as illustrated in figure 5. On the average, we estimated genes from the endosymbionts of aphids to have evolved 1.4-fold slower than those of tsetse flies (B_{ach}/W_{igg} = 0.70) (table 2). The rate variation among functional categories is within a factor of 2; however, it is interesting to note that many of the outliers represent putative host-selected functions. At the upper end, we find genes coding for proteins involved in the biosynthesis of fatty acids and phospholipids (B_{ach}/W_{igg} = 1.16), cofactors (B_{ach}/W_{igg} = 0.93), and cell envelope structures (B_{ach}/W_{igg} = 0.88) (table 2 and fig. 5A). At the lower end, we find genes coding for amino acid biosynthetic functions (B_{ach}/W_{igg} = 0.60) (table 2 and fig. 5B).

The substitution rate variation among functional categories is inversely related to differences in gene contents in the endosymbiont genomes. For example, the tsetse fly endosymbionts contain a larger number of genes for fatty acid and cofactor biosynthesis, and the remaining orthologs present in both genomes evolve relatively slower in the tsetse fly endosymbiont genome. Vice versa, the aphid endosymbionts contain a higher fraction of the genome devoted to amino acid biosynthesis, and the remaining orthologs evolve relatively slower in the aphid endosymbiont genomes. Finally, it is worth noting that also genes coding for proteins involved in the transcription and translation processes evolve slower in the aphid than in the tsetse fly endosymbionts (B_{ach}/W_{igg} = 0.46 to 0.60) (table 2). The rate difference is particularly striking for ribosomal protein genes, indicative of relaxed selective constraints on ribosome functions in the tsetse fly endosymbiont.

Taken together, our analysis suggests that the rate and mode of sequence evolution has not been uniform among genes in the endosymbiont genomes. The observation is that the rate acceleration has been counterbalanced by selection for genes in functional categories that contain a larger number of genes in one endosymbiont genome compared with the other. The two effects, gene retention and reduced substitution rates, are best explained as the result of purifying selection acting at the host level.

**Discussion**

The phylogenomic analyses presented in this paper strongly suggest that B. aphidicola is closely related to E. coli and Salmonella, with H. influenzae being an out-group to these species, also noted previously in a concatenated protein tree inferred with the aid of the neighbor-joining method (Lerat, Daubin, and Moran 2003). We conclude that the atypical placement of B. aphidicola observed in many of the single-gene trees is caused by high rates and uneven patterns of nucleotide substitutions, rather than by horizontal gene transfer, as was assumed in a recent publication by Itoh, Martin, and Nei (2002). By excluding...
the most rapidly evolving genes that produce atypical tree topologies, their study was inadvertently biased towards a subset of highly conserved genes. In effect, much of the enhancement and variation of evolutionary rates among genes was not recorded, leading to the possibly incorrect conclusion that there is no relative rate difference among orthologs in *B. aphidicola* and *E. coli* (Itoh, Martin, and Nei 2002).

The finding that many of the individual gene trees supported topologies that deviated from the presumed species tree (figs. 1 and 3) is consistent with results from previous phylogenetic analyses of the γ-Proteobacteria (Itoh, Martin, and Nei 2002; Snel, Bork, and Huynen 2002). In contrast, a lack of phylogenetic conflict with the assumed species tree for 203 of 205 genes was suggested by the Shimodaira-Hasegawa test (Lerat, Daubin, and Moran 2003). However, it is difficult to evaluate the outcome of this test, because it is dependent on the confidence level (Lerat, Daubin, and Moran 2003). The fraction of single-gene trees supporting the assumed species trees increased upon the use of substitution models that take asymmetric base composition and variable G+C content levels into account (fig. 4). The ML method failed to resolve the divergence node for *B. aphidicola* in most of the individual gene trees, although a single, fully resolved species tree was obtained with high support in analyses of concatenated gene sets (fig. 3). Likewise, incongruence among single-gene phylogenies for eight yeast species were found to be resolved using larger concatenated data sets (Rokas et al. 2003).

This confirms the suspicion that phylogenetic trees inferred from single-gene sequences may be seriously flawed because of too few sites and/or because the data set violates the assumptions of the models used for the phylogenetic reconstructions, as was also noted in previous studies of sequences with biased G+C content values (Galtier and Gouy 1995, 1998; Eyre-Walker 1998; Foster and Hickey 1999). In many cases, the assumed incorrect internal nodes are highly supported by bootstrap analysis (>95%), as also observed for dubious internal branches in other analyses where no correction was made for compositional biases (Galtier and Gouy 1995). Thus, the assumption that atypical gene trees correspond to horizontal gene transfers (Gogarten, Doolittle, and Lawrence 2002; Doolittle et al. 2003) should be taken with great caution, in particular when there is a compositional bias in the data set.

The elevated substitution rates in endosymbiont genomes relative to those of free-living species may be explained by the loss of genes involved in repair and recombination pathways, leading to higher mutation rates (Tamas et al. 2002; Wernegreen 2002; Dale et al. 2003). If the rate difference is solely the result of a mutation rate difference (Itoh, Martin, and Nei 2002), all proteins will be affected to similar extent. Also if the rate acceleration is influenced by slightly deleterious mutations or by a general relaxation of selection (Moran 1996; Andersson and Kurland 1998; Brynnel et al. 1998; Wernegreen and Moran 1999; Woolfit and Bromham 2003), the rate increase will be similar for all proteins. However, under this scenario, higher *ds/ds* ratios are expected for the endosymbiont populations. Because of selective constraints on synonymous substitutions in *E. coli*, these ratios are difficult to compare, and previous attempts to compare *ds/ds* ratios for endosymbionts and enterics have given discrepant results (Clark, Moran, and Baumann 1999; Ochman, Elwyn, and Moran 1999; Itoh, Martin, and Nei 2002).

If functional constraints are relaxed only for some genes in the endosymbiont populations, the rate acceleration is expected to differ among functional classes of genes. Genes coding for proteins involved in flagellar biosynthesis and protein export seem to evolve under relaxed functional constraints in the endosymbiont genomes, as indicated by atypically high levels of amino acid divergence for these genes (Tamas et al. 2002). A more direct approach to infer amino acid rate variation among genes is to compare branch lengths for endosymbionts and free-living lineages in individual gene trees (Itoh, Martin, and Nei 2002). However, until the correct species topology can be reconstructed for all genes in *B. aphidicola*, including those that are most rapidly evolving, no conclusive results can be obtained from these studies (Itoh, Martin, and Nei 2002).

Less focus has been placed on the role of the host, although simulation studies point at the population size of the host as the most important factor determining the fixation rate of mutations in endosymbiont genomes (Rispe and Moran 2000). According to the results of this simulation, fixation rates for mutations in endosymbiont genomes should decrease with increasing host population sizes (Rispe and Moran 2000). The interpretation is that selection on the host may counterbalance the accumulation of deleterious mutations through Mullers ratchet, thereby delaying fitness decline and extinction of endosymbiont populations (Rispe and Moran 2000). The relatively small increase of substitution rates for proteins involved in
FIG. 5.—Relative rates of sequence evolution in *B. aphidicola* and *W. glossinidia* are affected by host selection. Phylogenetic trees were inferred from concatenated alignments of genes involved in (A) cofactor biosynthesis, *dxs*, *ispA*, *ribF*, *ribH*, and *trxB* and (B) amino acid metabolism, *aroK*, *glyA*, and *metK*. The neighbor-joining trees were calculated with the Galtier and Gouy substitution model (NJGG) (Galtier and Gouy 1995). Third codon positions were removed in the analyses. The clustering of *B. aphidicola* and *W. glossinidia* was supported by 100% in both NJGG and ML analyses. Values at nodes indicate the percentages of 500 bootstraps.
amino acid biosynthesis in *B. aphidicola* (Itoh, Martin, and Nei 2002) provides some support for this view. If the evolution of endosymbiont genes is indeed influenced by selective constraints imposed by the host, the relative rate enhancement is expected to be variable among endosymbionts with different host-interaction profiles. On the other hand, if the acceleration of genomic evolution is solely influenced by enhanced mutation rates, no such correlation is to be expected.

To examine the influence of host-selection on endosymbiont gene evolution, we compared the relative rate variation among genes for endosymbionts with different hosts. Because the test is based on a comparative analysis of branch lengths for the endosymbionts of aphids (*B. aphidicola*) with those of tsetse flies (*W. glossinidia*), it is important that the endosymbionts are correctly placed in the γ-proteobacterial tree. It was previously thought that the symbiosis event for the tsetse fly endosymbionts dates back approximately 40 Myr and was independent from that of the aphid endosymbionts (Moran and Wernegreen 2000), estimated to 150 to 250 MYA. However, a more recent analysis suggests that *B. aphidicola* and *W. glossinidia* are sister clades and have originated from a single symbiosis event (Lerat, Daubin, and Moran 2003), which is consistent with our phylogenetic analysis of both single and concatenated gene alignments (fig. 5).

The results of our analysis suggest that the relative rate increase differs among endosymbiont genes in a manner that is best explained by host-level selection. For example, the higher genomic fraction and the lower rate of sequence evolution for genes involved in amino acid biosynthesis in *B. aphidicola* is consistent with host selection for amino acid biosynthetic genes in aphid endosymbionts (Shigenebou et al. 2000; Tamas et al. 2002). Likewise, endosymbiont supply of B-vitamins to the tsetse flies is consistent with the retention of a greater proportion of genes involved in cofactor biosynthesis and a relatively lower rate of sequence evolution for these genes in the tsetse fly endosymbionts. Another difference is that the *B. aphidicola* genome only contains 50% of the genes coding cell-structure components present in the *W. glossinidia* genome, possibly because of a higher requirement for a functional cell wall during transmission to larvae in the tsetse fly endosymbionts (Akman et al. 2002). Also, for genes encoding cell wall components, the relative rate acceleration in the tsetse fly endosymbionts is lower than for other genes.

The inverse correlation between gene retention and substitution rates suggests that the rate of sequence evolution in endosymbiont genomes is influenced by selective constraints acting at the level of the host. Thus, we have shown here that host-beneficial genes accumulate mutations at reduced levels in relation to the metabolic dependencies of hosts and endosymbionts. To our knowledge, this is the first demonstration of species-specific, host-selected retardation of sequence evolution in endosymbiont genomes.

**Acknowledgments**

We thank Lisa Klasson, Alex Mira, and Sean Hooper for assistance, helpful discussions, and valuable comments on the manuscript. This work was supported by the Swedish Research Council (VR), the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS), the Swedish Foundation for Strategic Research (SSF), the Knut and Alice Wallenberg Foundation (KAW), and the European Union (EU).

**Literature Cited**


**Endosymbiont Phylogeny 1121**

**Downloaded from https://academic.oup.com/mbe/article-abstract/21/6/1110/1050787 by guest on 07 March 2019**
usage, in the tsetse fly endosymbiont *Wigglesworthia*. Microbiology **149**:2585–2596.


Jonathan Eisen, Associate Editor

Accepted February 17, 2004