From Phylogenetics to Phylogenomics: The Evolutionary Relationships of Insect Endosymbiotic γ-Proteobacteria as a Test Case

Iñaki Comas, Andrés Moya, and Fernando González-Candelas

Institut Cavanilles de Biodiversitat i Biologia Evolutiva. Universitat de València, Valencia, Spain;
E-mail: inaki.comas@uv.es (I.C.); andres.moya@uv.es (A.M.); fernando.gonzalez@uv.es (F.G.-C.)

Abstract.—The increasing availability of complete genome sequences and the development of new, faster methods for phylogenetic reconstruction allow the exploration of the set of evolutionary trees for each gene in the genome of any species. This has led to the development of new phylogenomic methods. Here, we have compared different phylogenetic and phylogenomic methods in the analysis of the monophyletic origin of insect endosymbionts from the γ-Proteobacteria, a hotly debated issue with several recent, conflicting reports. We have obtained the phylogenetic tree for each of the 579 identified protein-coding genes in the genome of the primary endosymbiont of carpenter ants, Blochmannia floridanus, after determining their presumed orthologs in 20 additional Proteobacteria genomes. A reference phylogeny reflecting the monophyletic origin of insect endosymbionts was further confirmed with different approaches, which led us to consider it as the presumed species tree. Remarkably, only 43 individual genes produced exactly the same topology as this presumed species tree. Most discrepancies between this tree and those obtained from individual genes or by concatenation of different genes were due to the grouping of Xanthomonadales with β-Proteobacteria and not to uncertainties over the monophyly of insect endosymbionts. As previously noted, operational genes were more prone to reject the presumed species tree than those included in information-processing categories, but caution should be exerted when selecting genes for phylogenetic inference on the basis of their functional category assignment. We have obtained strong evidence in support of the monophyletic origin of γ-Proteobacteria insect endosymbionts by a combination of phylogenetic and phylogenomic methods. In our analysis, the use of concatenated genes has shown to be a valuable tool for analyzing primary phylogenetic signals coded in the genomes. Nevertheless, other phylogenomic methods such as supertree approaches were useful in revealing alternative phylogenetic signals and should be included in comprehensive phylogenomic studies. [Concatenate alignment; incongruence; microbial evolution; phylome; supermatrix alignment; supertree analysis.]

The evolution of Bacteria is strongly influenced by the fluid nature of their genomes. Processes such as duplication, gene gain and loss limit the inference of their evolutionary relationships. However, in the last decade the genomic revolution has represented not only a change in scale for the analysis of sequences but also for phylogenetic inference. The phylogenomic approach, initially proposed as the application of phylogenetic analysis to help annotating complete genome sequences (Eisen, 1998), has transformed into the (possibly) most appropriate way to derive the phylogenetic history of organisms through genome-scale phylogenetic inference (O'Brien and Stany, 1999; Sicheritz-Pontén and Andersson, 2001; Eisen and Fraser, 2003). This leap from phylogenetics to phylogenomics has allowed avoiding some of the inherent problems in the inference of species trees from single gene phylogenies. However, it has also brought new issues in the reconstruction of species trees and even questioned the existence of such a single tree for all the genes in the genome of a bacterial species (Doolittle, 1999; Baptiste et al., 2005).

The single-gene approach reached its peak with the use of ribosomal RNA sequences as phylogenetic markers for microorganisms (Woese, 1987). These genes have been, and still are, a powerful tool in bacterial phylogenetic analysis. Their properties as a good marker for phylogenetic inference, such as universal presence and evolutionary conservation, have enabled the proposal of a universal tree of life (Woese et al., 1990) and the classification and reconstruction of evolutionary relationships for the three domains in the absence of genomic data (Woese and Fox, 1977). However, even ribosomal markers have been shown to be laterally transferred in some cases (Asai et al., 1999; Yap et al., 1999).

The advantages of multiple gene approaches versus those based on single genes are a priori evident. In theory, we can evade the single gene evolutionary histories in favor of a common “true” phylogenetic signal; it is possible to avoid problems derived from insufficient sample size by addition of more sites from multiple genes or to compensate for biased base compositions. In practice, some of these problems will also affect phylogenies reconstructed from large data sets, but others will be substantially reduced and/or easily diagnosed. Several alternative methods to single-gene tree phylogenies have been proposed recently (reviewed in Lelsuc et al., 2005). These methods are based on different kinds of sequence characters and genome structure, such as gene content (Snel et al., 1999; Gu et al., 2004; Huson and Steel, 2004), gene order (Wol et al., 2001; Korbel et al., 2002; Belda et al., 2005), concatenated sequences (Brown et al., 2001; Rokas et al., 2003), or gene tree-based techniques (Bininda-Emonds et al., 2002; Bininda-Emonds, 2004b). In this article we focus on methods based on direct or indirect analyses of genomic sequence data.

The concatenation approach has proven useful in many cases and its use is increasingly common (Baptiste et al., 2002; Rokas et al., 2003). The method has been justified in cases where single-gene phylogenetic signals are insufficient (Herniou et al., 2001), when there is heterogeneity in the evolutionary rates (Baptiste et al., 2002; Gontcharov et al., 2004), or when a high influence of nonstandard evolutionary patterns in the shaping of gene trees, such as horizontal gene transfer, is to...
be expected (Brochier et al., 2002). This method increases phylogenetic signal by joining the sequences from multiple genes, thus creating a supermatrix of characters, and generally recovers accepted (and presumably correct) phylogenies with highly supported nodes, thus evading many of the above pitfalls (for instance, insufficient amounts of informative sites or particular gene histories) as far as the correct model of evolution is used, although it may not overcome problems associated to systematic biases in the data (Phillips et al., 2004).

Contrary to the concatenation approach, consensus and supertree approaches have an indirect association with the genome sequence. The consensus method is based on the integration of multiple source trees into a single topology. When the initial data set does not include the same taxa in all the gene trees, then a supertree is constructed combining the overlapping topologies (Bininda-Emonds, 2004a). All these methods have proven useful for phylogenetic inference (Daubin et al., 2002; Rokas et al., 2003), but they have different possible associated errors. For instance, the strength of the supermatrix approach decreases when the number of shared orthologous genes is low or when many of the genes concatenated are influenced by horizontal gene transfer events or hidden paralogies (Daubin et al., 2002). On the other hand, a supertree approach has also its own sources of problems. The use of wrong source trees, the indirect relationship with molecular data, taxon heterogeneity among the trees, or lack of a universal methodology for assessing the reliability of the nodes are among the most common problems encountered by consensus and supertree approaches (Gatesy et al., 2002; Creevey, 2004; Bininda-Emonds, 2004a).

Our primary concerns in this work are to carry out a comparative analysis of different phylogenetic methodologies and to derive, as far as possible, a presumed species tree for a group of Proteobacteria, although this last concept itself is controversial for bacteria. Opinions range from those who consider nonvertical evolutionary events, mainly horizontal gene transfer, as background noise that, after appropriate filtering, still allows to recover the true phylogenetic relationships among the species (Jain et al., 1999; Kurland et al., 2003) to those who sustain the impossibility of describing bacteria genome evolution through a single tree-like phylogeny (Doolittle, 1999; Creevey et al., 2004; Baptiste et al., 2005). We will not address directly this question in this work, although we will use the term “species tree” to refer to that phylogenetic hypothesis that encompasses most of the vertically transmitted phylogenetic signal. To study the evolution of bacteria we must identify the different vertical and nonvertical phylogenetic signals; the former, if sufficiently strong, will be reflected in the species tree.

In this work, we have approached a conflicting issue in bacterial phylogenetics through the use of different phylogenetic and phylogenomic methods. Apart from allowing the comparison of the problems and advantages of the different methodologies employed, the particular set of Proteobacteria used in this study also poses some interesting evolutionary questions. In particular, there is an ongoing debate on whether bacterial endosymbionts of insects from the gamma subdivision of Proteobacteria conform to a monophyletic group (Gil et al., 2003; Lerat et al., 2003; Canback et al., 2004) or they are paraphyletic and their grouping results from artefacts in the phylogeny reconstruction process (Charles et al., 2001; Herbeck et al., 2004; Belda et al., 2005). This study investigates the evolutionary relationships of five γ-Proteobacteria endosymbionts, including the three Buchnera strains sequenced so far. Several evolutionary features of these genomes, such as their high evolutionary rates and low G+C content, and genome disintegration led to different methodological problems (Moreira and Philippe, 2000; Sanderson and Shaffer, 2002).

Our work is based on the determination of the phylo- lome (Sicheritz-Pontén and Andersson, 2001)—the set of phylogenetic trees for each protein-coding gene in the genome—of Blochmannia floridanus (Gil et al., 2003), the primary endosymbiont of carpenter ants. We have used it as a starting point to explore the phylogenetic landscape of γ-Proteobacteria. In this work we have used an array of phylogenetic and phylogenomic techniques to infer the phylogenetic relationships among 21 Proteobacteria. In consequence, we have analyzed several phylogenetic hypotheses for these species through the examination of the phylogenies derived from the set of protein-coding genes in Blochmannia and their comparison with topologies obtained from the 16S rDNA and different phylogenomic methods. Our aim has been not only to test alternative species tree hypotheses but also to analyze the advantages and pitfalls associated to each class of approximation. Additionally, we have attempted an approximation to the genomic core concept (Nesbø et al., 2001; Gil et al., 2004; Strobel and Arnold, 2004) through the analysis of trees and sequences that has allowed us to question the relevance of the informational/operational gene distinction in phylogenetic inference and in the conformation of a potential bacterial core set of genes.

**Materials and Methods**

*Genomes and Homologous Genes Selection*

In this study we have used 21 complete genome sequences of Proteobacteria (Table 1) retrieved from GenBank (Benson et al., 2004). We included three β-Proteobacteria, *Neisseria meningitidis* MC58, *Neisseria meningitidis* Z2491, and *Ralstonia solanacearum*; one α-Proteobacteria, *Rickettsia prowazekii*; and 17 γ-Proteobacteria, including five insect endosymbionts: the three *Buchnera* species sequenced so far, *Wigglesworthia gossinidia brevipalpis*, and *Blochmannia floridanus*, which is our initial genome. A complete and general scheme of the analyses carried out in this study and the relationships among them is presented in Figure 1.

The procedure to obtain the putative orthologs for each gene in the *Blochmannia floridanus* genome started from an initial reference tree. Because it is not possible to be certain about the truly orthologous nature of a gene until the phylogenetic analysis is completed, we considered...
Table 1. Complete genome sequences of Proteobacteria used in this study.

<table>
<thead>
<tr>
<th>Species (strain)</th>
<th>Accession no.</th>
<th>Division</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rickettsia prowazekii</td>
<td>NC_000963</td>
<td>α-Proteobacteria</td>
<td>Rickettsiales</td>
</tr>
<tr>
<td>Neisseria meningitidis MC58</td>
<td>NC_003112</td>
<td>β-Proteobacteria</td>
<td>Neisseriales</td>
</tr>
<tr>
<td>Neisseria meningitidis Z2491</td>
<td>NC_003116</td>
<td>β-Proteobacteria</td>
<td>Neisseriales</td>
</tr>
<tr>
<td>Ralstonia solanacearum</td>
<td>NC_003295</td>
<td>β-Proteobacteria</td>
<td>Enterobacteriales</td>
</tr>
<tr>
<td>Blochmannia floridanus</td>
<td>NC_005061</td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriales</td>
</tr>
<tr>
<td>Buchnera aphidicola BPH</td>
<td>NC_004545</td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriales</td>
</tr>
<tr>
<td>Buchnera aphidicola SG1</td>
<td>NC_004061</td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriales</td>
</tr>
<tr>
<td>Buchnera sp. APS</td>
<td>NC_002528</td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriales</td>
</tr>
<tr>
<td>Escherichia coli K12</td>
<td>NC_000913</td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriales</td>
</tr>
<tr>
<td>Escherichia coli O157:H7 EDL933</td>
<td>NC_002655</td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriales</td>
</tr>
<tr>
<td>Haemophilus influenzae Rd</td>
<td>NC_000907</td>
<td>γ-Proteobacteria</td>
<td>Pasteurellales</td>
</tr>
<tr>
<td>Pasteurella multocida</td>
<td>NC_002663</td>
<td>γ-Proteobacteria</td>
<td>Pasteurellales</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>NC_002516</td>
<td>γ-Proteobacteria</td>
<td>Pseudomonadales</td>
</tr>
<tr>
<td>Salmonella enterica subsp. enterica serovar Typhi</td>
<td>NC_003198</td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriales</td>
</tr>
<tr>
<td>Salmonella typhimurium LT2</td>
<td>NC_003197</td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriales</td>
</tr>
<tr>
<td>Vibrio cholerae</td>
<td>NC_002505</td>
<td>γ-Proteobacteria</td>
<td>Vibrionales</td>
</tr>
<tr>
<td>Wigglesworthia glossinidia brevipalpis</td>
<td>NC_004344</td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriales</td>
</tr>
<tr>
<td>Xanthomonas axonopodis pv. citri strain 306</td>
<td>NC_003919</td>
<td>γ-Proteobacteria</td>
<td>Xanthomonadales</td>
</tr>
<tr>
<td>Xanthomonas campestris pv. campestris strain ATCC 33913</td>
<td>NC_003902</td>
<td>γ-Proteobacteria</td>
<td>Xanthomonadales</td>
</tr>
<tr>
<td>Xylella fastidiosa</td>
<td>NC_002488</td>
<td>γ-Proteobacteria</td>
<td>Xanthomonadales</td>
</tr>
<tr>
<td>Yersinia pestis KIM</td>
<td>NC_004088</td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriales</td>
</tr>
</tbody>
</table>

The homologous genes found in the procedure described below as putative orthologs, which may eventually become true orthologs, paralogs, or xenologs, depending on their inferred evolutionary history. The reference tree was obtained as described in Gil et al. (2003). Briefly, we obtained the orthologs for 60 informational genes present in all the genomes considered. Protein sequences were aligned with CLUSTALW 1.8 (Thompson et al., 1994) and processed with GBLOCKS (Castresana, 2000) with default parameters to eliminate areas of uncertain homology or low phylogenetic content before concatenation. The resulting concatenate was used to obtain a
maximum likelihood tree using the quartet method implemented in TREEPUZZLE 5.1 (Schmidt et al., 2002) with the following options: JTT model of substitution (Jones et al., 1992), proportion of invariant sites (I) estimated from the data, eight discrete categories to approximate a gamma distribution accounting for evolutionary rate heterogeneity across sites (G), empirical amino acids frequencies (F), and 4000 puzzling steps. We also obtained phylogenetic trees by Bayesian inference using MrBayes 3.0 (Ronquist and Huelsenbeck, 2003) (JTT+I+F+G and 100,000 generations). This tree is an expanded version of the tree reported in Gil et al. (2003) with additional sequences from the non-γ-Proteobacteria genomes. With this reference tree, we assigned each genome to one of nine different groups (see Fig. 2) in order to reduce the BLAST database and to speed up and refine the searches (see below).

The B. floridanus genome contains 579 annotated protein coding genes (Gil et al., 2003) and each of these was used as query for a BLASTP (Altschul et al., 1997) search against the remaining genomes. To retrieve the homologous genes from the other genomes, we first performed a BLAST search using the representative genome from each of the nine previously described groups as target (Fig. 2). Next, we used the best hit from each representative species of the nine groups in the first BLAST search as query for a second BLAST against the remaining genomes in the corresponding group. This strategy was slightly modified for those B. floridanus genes for which no reliable homolog (see below a more detailed description of the criteria used) was found in the representative species of some group. In this case, a new BLAST search was performed against all the genomes in the group and the best hit was used as query for a second BLAST to retrieve the homologs from the remaining species in the group as described. In all cases, we used an E-value < 1E−3 as threshold for considering a matching sequence as a putative ortholog in the BLAST searches.

Once homologs for each of the 579 genes in B. floridanus were retrieved we performed a new filtering to ensure

---

**Figure 2.** Guide species tree obtained with a trimmed alignment of 60 concatenated proteins. For clarity, only genus names are represented except for cases with more than one species or strain (see Table 1). Numbers in nodes indicate support values in the form of proportion of quartets and Bayesian a posteriori probabilities for the corresponding inner branch. The symbol next to each species indicates the group of adscription for BLAST searches. Species with underlined names were used as initial target species for BLAST searches of the corresponding groups. The taxonomic classification (order and Proteobacteria division) of each species is also indicated according to the NCBI taxonomy database.
that these genes corresponded to putative orthologs. We first proceeded to obtain a multiple alignment and the corresponding gene tree as described below. With this information we considered five different criteria for each case: annotation (coincidence in functional annotation), multiple alignment (similarity extended over the complete sequence and not a short fragment), gene tree (strong conflict with the reference tree), BLAST significance, and information from the Microbial Genome Database for Comparative Analysis (Uchiyama, 2003). Homologous genes were accepted as putative orthologs after simultaneous consideration of all these five criteria. Because these are independent, they allowed us to analyze orthology in the absence of a good annotation or when the position in the gene tree was unexpected. Should we have based strictly and only on annotation or position in the gene tree, then we would have biased the data set towards genes with the most congruent phylogenies or best annotations. In most cases the identification of putative orthologs was easy. In the most difficult cases, which corresponded to nonannotated proteins or to very divergent gene trees from the usually accepted species phylogenies, we considered as putative orthologs those genes that were supported by at least three of the above mentioned criteria. This resulted in a set of 200 genes in B. floridanus with reliable orthologs in all of the other 20 genomes considered. We will refer to this subset as the ‘200-gene’ data set. Whenever possible, and in order to test the effect of missing data, we performed the analyses described below with the remaining genes with reliable orthologs missing from at least one genome (‘379-gene’ data set) and also with the complete set of coding genes in the B. floridanus genome (‘579-gene’ data set).

Obtaining the Phylome of Blochmannia floridanus

After obtaining putative orthologs for each protein of the B. floridanus genome we proceeded to obtain its phylome, the set of phylogenetic trees for each protein coding gene. The alignment for each set of homologous proteins was obtained using CLUSTALW 1.8 (Thompson et al., 1994) and the result was trimmed with GBlocks (Castesana, 2000) using default parameters. The maximum likelihood tree for each multiple alignment was inferred with the program PHYML 2.1b (Guindon and Gascuel, 2003) as it implements one of the fastest and most accurate algorithms for heuristic searches. For all the alignments we used JTT + I + G + F as model and parameters of evolution. Some of the initial alignments and trees were modified based on the a posteriori selection of homologs explained above. We decided not to obtain support values by bootstrap resampling to save computation time and not to bias the analysis towards the best supported topologies.

We used four phylogenetic approaches to obtain a 16S rDNA topology for the 21 species: maximum parsimony (MP), maximum likelihood (ML), and two different distances with the neighbor-joining (NJ) algorithm (Saitou and Nei, 1987). MP analysis was performed using PAUP* 4b10 (Swofford, 2002) with heuristic search based in tree-bisection-reconnection (TBR) as branch-swapping algorithm and 1000-replicate to assess bootstrap support values. The best model for ML inference was selected using the ModelTest program (Posada and Crandall, 1998) following the Akaike information criterion (AIC) and implemented in PHYML 2.1b with 500 bootstrap replicates. For NJ we used two models that could handle heterogeneous base composition among lineages. Firstly we used MEGA 3 (Kumar et al., 2004) to implement the LogDet distance modified by Tamura and Kumar (Tamura and Kumar, 2002) to account for substitution pattern heterogeneity. Secondly we used PHYLO_WIN (Galtier et al., 1996), implementing the Galtier and Gouy distance (Galtier and Gouy, 1995) for reducing nucleotide bias. A 1000-replicate bootstrap was carried out for both NJ analyses.

Phylogenomic Analysis

We used three different approaches for phylogenomic analysis. Concatenated sequences, consensus trees, and supertrees were used to explore the phylogenetic landscape of γ-Proteobacteria.

Concatenates.—The analysis of concatenates was based on the subset of putative orthologous genes shared by the 21 genomes, the ‘200-gene’ data set. We concatenated the GBlocks-trimmed alignments of these proteins and analyzed the resulting alignment using several methods. PAUP* 4b10 (Swofford, 2002) was used for parsimony analysis with stepwise addition and tree bisection-reconnection for heuristic search and 500 bootstrap replicates. PHYML and TREEPUZZLE were used for maximum likelihood inference. Due to computational limitations, we had to reduce the number of parameters in the evolutionary models. In the analysis with TREEPUZZLE we used a JTT model with four gamma categories and 4000 puzzling steps. The JTT model of substitution with estimated proportion of invariant sites and 100 bootstrap replicates was used for inference with PHYML.

We also used this alignment of the ‘200-gene’ data set to explore the possible influence of heterogeneous base composition in the retrieval of the different identified clades, especially those involving endosymbiotic bacteria. We removed from the alignment those positions that included the amino acids most strongly affected by nucleotide bias (FYMINK) (Singer and Hickey, 2000). With this data set we carried out a Bayesian analysis using MrBayes 3.0 (Ronquist and Huelsenbeck, 2003) (four chains, 1,000,000 generations, 10% of burn-in) and a maximum likelihood analysis with PHYML (500 bootstrap pseudo-replicates) with the same evolutionary model described above (JTT+I).

Finally, we obtained a concatenate tree using all available sequences (‘579-gene’ data set) by setting as “missing character” those amino acids that corresponded to a protein not present in any of the 21 genomes. It was analyzed with TREEPUZZLE using the JTT model of substitution with invariants estimated form the data set.
Alternative, more complete methods were not feasible due to computational limitations.

**Consensus Trees.**—A majority rule consensus tree was obtained with the program CONSENSE from the PHYLIP package (Felsenstein, 2005) with the gene trees from the ‘200-gene’ data set.

**Supertrees.**—For the ‘200-gene’ data set and for the set of gene trees with an incomplete number of genomes represented we used a supertree approach (Sanderson et al., 1998). A supertree integrates all the topological information present in the source gene trees. In this case, we applied supertree reconstruction to three different data sets: (i) genes present in all the genomes (‘200-gene’ data set, the same set for which we obtained a consensus tree); (ii) genes missing in at least one genome (‘379-genes’ data set); and (iii) to the complete phylome (‘579-gene’ data set). We applied the different optimization algorithms implemented in program CLANN (Phillips et al., 2004; Creevey and McInerney, 2005) to our data set. The most commonly applied method in supertree reconstruction is matrix representation using parsimony (MRP) (Loomis and Smith, 1990; Baum, 1992; Ragan, 1992). This method considers each node of the source trees as a binary character, assigning a “1” to the taxa contained in the clade, defined by the internal node, a “0” for the taxa not present in this clade, and a “?” for the taxa not present in the tree. The resulting matrix is analyzed by parsimony. The other methods implemented and used in our work were most similar supertree method (dfit), maximum quartet fit (qfit), and maximum split fit (sfit) (Creevey and McInerney, 2005). They are based in the optimization of distances between nodes, shared quartets, or shared partitions between the supertrees proposed by the heuristic search and the source trees, respectively.

Nonparametric bootstrap by sampling with replacement over gene trees was used to obtain an indicative measure of the support of the derived clades in each supertree from the corresponding data. Support for each node in the supertrees was evaluated from 100 pseudoreplicate samples for each data set used to derive the supertrees.

**Presumed Species Tree versus Gene Trees: Likelihood-Based Test of Topologies**

Likelihood-based tests of competing phylogenetic hypotheses were carried out only once we obtained the complete phylome of *B. floridanus*. For each alignment we compared the topologies of the presumed species tree and the corresponding gene tree by means of three different tests implemented in TREEPUZZLE (Schmidt et al., 2002): the one-side Kishino-Hasegawa test (1sKH) (Kishino and Hasegawa, 1989; Goldman et al., 2000), Shimodaira-Hasegawa test (SH) (Shimodaira and Hasegawa, 1999), and the expected likelihood weight test (ELW) (Strimmer and Rambaut, 2002). For the two former tests, the null hypothesis assumes that the difference between the likelihoods associated to each topology is not significantly different from zero. The first test to be described was the KH test, but its validity when one of the competing hypotheses was derived from the alignment has been questioned (Goldman et al., 2000). In this study we used the one-sided version of the test because the gene tree is the maximum likelihood tree and therefore we expected its likelihood to always be higher or equal than the likelihood associated to the species tree. In any case we have used it for comparative purposes and not as the reference test because of its problems with multiple comparisons corrections. The SH test is a multiple hypotheses contrast. Even though it is not free from errors (Goldman et al., 2000), it overcomes some of the problems associated with the KH test. We used it as our reference test for examining the rejection of the presumed species tree by each gene alignment. For both tests the resampling-estimated log likelihood (RELL) bootstrapping method with 1000 replicates was used. The ELW test is an alternative way of topology testing. It creates a confidence set of phylogenies that could include (acceptance) or not (rejection) our presumed species tree topology. The usual $\alpha = 0.05$ level was used to delimit the acceptance/rejection region.

**Results**

**Reference Tree and Search for Homologs**

The construction of a phylogenetic tree for a set of species from their genome sequences needs, as a starting point, a phylogeny on which to accept or reject which gene should be used for this kind of analysis on the basis of their true orthology. This is a circular question that we have approached by obtaining an initial or reference tree, testing its reliability, and then using it as benchmark for further decisions. For this, we started by obtaining the phylogenetic tree from a subset of conserved protein coding gene usually accepted as providing the most robust, reliable phylogenies when concatenated (Jain et al., 1999).

The concatenation of 60 informational gene after their independent alignment and trimming of residues of uncertain homology resulted in an alignment with 8067 amino acid positions. Figure 2 shows the phylogenetic tree obtained by maximum likelihood and Bayesian inference methods. This reference tree was identical to the one reported in Gil et al. (2003) except for the addition of four genomes from Proteobacteria in the $\alpha$- and $\beta$-divisions. The tree reflected monophyly not only for the three *Buchnera* species but also for a wider group including the other insect endosymbionts, *Blochmannia* and *Wigglesworthia*. This group is a sister clade to the YESS (*Yersinia-Escherichia-Shigella-Salmonella*) cluster. Due to its concordance with the accepted taxonomy, the high support values for the nodes and the fact that it recovers the monophyly of *Buchnera* species, we considered this as a good reference tree for the ensuing BLAST searches. We further verified that this tree was obtained in most cases after concatenation of 60 random gene chosen among those present in the 21 genomes (results not shown).
Figure 3. Histogram summarizing the distribution of putative orthologs for each protein coding gene (579 proteins of known function) in the Blochmannia floridanus genome among the 21 genomes considered in this study. Note that 15 gene from last column (those present in all the species) were considered of dubious orthology and were analyzed in the 379-gene data set.

The similarity searches for homologs in the 21 genomes resulted in 215 protein coding gene present in all of them. After their initial multiple alignment and phylogenetic tree reconstruction, we proceeded to verify their reliability as putative orthologs according to the criteria indicated above. This resulted in the exclusion of 15 proteins with doubtful orthology for at least one γ-Proteobacteria species different from the insect endosymbionts, thus conforming a 200-gene data set. However, in order to analyze the complete B. floridanus phylome, we included these 15 proteins present in 21 taxa but with doubtful orthology for all of them into the data set of proteins absent from at least one taxon. Therefore, this set was composed of 379 gene (‘379-gene’ data set) with a heterogeneous distribution of presence in the different genomes, summarized in Figure 3. Each protein was present on average in 16 of the 21 genomes. Nearly 60% of the proteins were present in at least 16 genomes and only 14% in less than 10. A table with the distribution of B. floridanus protein coding gene in the other 20 genomes is available as supplementary information from the Systematic Biology website at http://systematicbiology.org.

The Phylome of B. floridanus and the 16S rDNA Tree

We obtained all the phylogenetic trees derived from the set of protein coding gene of B. floridanus as explained in Materials and Methods. In the analysis of the ‘200-gene’ data set, we identified only three gene (rpoC, pheT, and mopA) with the same topology (Robinson-Foulds distance = 0) than the reference tree (Table 2). In the ‘379-gene’ data set, we found 40 gene whose topology was fully congruent with that of the reference tree, although this number reduced to 29 when only those present in at least 11 species were considered.

Apart from trees based on protein sequences, we obtained the tree based on the 16S DNA. The four methods used (see Materials and Methods) resulted in placing Xanthomonadales in the β-Proteobacteria branch. For the endosymbiotic bacteria almost all methods retrieved a monophyletic group (see Fig. 4 ML tree), although for the Galliker and Gouy distance this relationship was paraphyletic (see Fig. 4 NJ tree). Other taxa like Vibrio cholerae or Pseudomonas aeruginosa also had a variable position depending on the model or method used (Fig. 4 and data not shown).

Phylogenomic Analysis

Concatenation Analysis.—We obtained an alignment of 52,029 amino acid positions, 40,576 of which were variable and 32921 parsimony informative, from the concatenation of the ‘200-gene’ data set. The phylogenetic analysis resulted in a common topology for all the reconstruction methods used, with high support values for most of the nodes and methods (Fig. 5). This topology agreed with the reference tree established by the preliminary analysis (see above) and the phylogenetic reconstruction in Gil et al. (2003) as well as the most commonly accepted phylogenetic history for these species. In consequence, we adopted this topology as the

Table 2. Summary of tests for topologies. For each test (1sKH, one-sided Kishino-Hasegawa; SH, Shimodaira-Hasegawa; EWL, expected-weighted likelihoods) percentage of cases in which the species tree topology was rejected is presented. The last column indicates the number of cases in which the species and gene tree topologies are identical.

<table>
<thead>
<tr>
<th>Group</th>
<th>1sKH (%)</th>
<th>SH (%)</th>
<th>EWL (%)</th>
<th>No. of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL (579 gene)</td>
<td>30.4</td>
<td>29.5</td>
<td>31.4</td>
<td>43</td>
</tr>
<tr>
<td>21 spp. (200 gene)</td>
<td>30.5</td>
<td>29</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>&lt;21 spp.(379 gene)</td>
<td>30.3</td>
<td>29.8</td>
<td>31.1</td>
<td>40</td>
</tr>
</tbody>
</table>
presumed species tree on which to base further comparisons. This topology included a monophyletic clade with the insect endosymbionts, which received a high support by the three different methods used (bootstrap analysis for maximum likelihood and parsimony and quartet-puzzling maximum likelihood). This high support was common for most nodes in this tree and in fact there were only two cases in which one method provided a support lower than 70%. One corresponded to a large group of \( \gamma \)-Proteobacteria, including the Enterobacteriales and \( H. \) \textit{influenzae} and \( P. \) \textit{multocida}, which was recovered in only 69% by quartet puzzling analysis. The second case was the bootstrap support (59%) received by the three \( \beta \)-Proteobacteria using parsimony analysis. In both cases, as well as for all other nodes, the bootstrap support of maximum likelihood analysis equaled 100%. It is also remarkable that the Xanthomonadales were placed at the base of \( \gamma \)-Proteobacteria with very high support with the three methods used.

We also obtained the same topology in two other concatenate analyses. Firstly, we used the same concatenate alignment of the ‘200-gene’ data set but removing all the positions with those amino acids most influenced by a biased GC content (Singer and Hickey, 2000). After removal of positions with amino acids FYIMNK, there were 16,772 positions left for analysis. We retrieved the same topology with even higher support for the nodes (data not shown). Secondly, we also carried out the analysis of the concatenated alignment of the whole data set (‘579-gene’ data set), incorporating “missing data” as necessary, and obtained again the same topology of the presumed species tree. The resulting concatenate included 137,301 positions with an average of 14.59% sites encoded as “missing” per genome (excluding \( B. \) \textit{floridanus}).

Consensus and Supertree Analyses.—Two kinds of methods dealing with the gene trees were used. Trees derived from the ‘200-gene’ data set were analyzed by traditional majority rule consensus and supertree approaches. Both analyses recovered the same topology (Fig. 6) characterized by the deviation from the presumed species tree in one of the nodes. This divergence placed the Xanthomonadales group and \( P. \) \textit{aeruginosa} with the \( \beta \)-Proteobacteria. These alternative groupings received relatively low bootstrap support and, in the most frequent alternative to the majority rule consensus tree, \( P. \) \textit{aeruginosa} returned to its position in the presumed species tree (data not shown). Although the remaining clusters in the consensus were coincident with those in the presumed species tree, including the monophyly for insect endosymbionts, the frequency of each clade in the source trees was generally low. Even well-defined groups such as free-living Enterobacteria received lower support than expected.

For the remaining gene trees (‘379-gene’ data set) from the \( B. \) \textit{floridanus} genome, we applied only the supertree
FIGURE 5. Common topology obtained from the analysis of the 200-gene concatenate and the 579-gene supertree. This topology was adopted as the presumed species tree. Values above the nodes indicate support values for the 200-gene concatenate obtained by maximum likelihood (bootstrapping), maximum parsimony (bootstrapping), and maximum likelihood (quartet-puzzling). Values below the nodes indicate bootstrap support for the 579-gene supertree. The asterisk indicates the node that appears in a different position in the 579-gene bootstrap consensus tree. Divisions of Proteobacteria different from the gamma are indicated next to the corresponding species.

approach due to the heterogeneity in the number of taxa represented in each tree. The four algorithms used resulted in the same topology (Fig. 6). This topology was very close to the consensus and supertree topologies obtained from the '200-gene' data set (Fig. 6). The Xanthomonadales group was also outside the γ-Proteobacteria but P. aeruginosa was recovered in its “natural” group.

Finally, we also obtained a supertree from the complete phylome (579 gene trees) of B. floridanus. Interestingly, the topology obtained by heuristic search using the MRP algorithm agreed with the presumed species tree (Fig. 5). Therefore, the complete phylome analysis recovered the position of Xanthomonadales as γ-Proteobacteria although with low bootstrap support. In fact, the bootstrap consensus tree (obtained after 100 pseudoreplicates of 579 gene trees analyzed by MRP) again showed Xanthomonadales grouping with β-Proteobacteria. This is a clear indication of the presence of conflicting phylogenetic signals in this data set.

Presumed Species Tree versus Gene Trees: Likelihood-Based Test of Topologies

We next tested whether the presumed species tree was significantly worse than each gene tree in the B. floridanus phylome. Table 2 summarizes the proportion of
cases in which the former was rejected for the three likelihood based tests used. As expected, the SH test was more conservative than the other two although the results of the three tests were quite similar. Globally we observed a 30% rate of rejections with no noticeable difference between gene present in all (‘200-gene’ data set) or absent from some of the 21 genomes (‘379-gene’ data set). We also found slight differences in the rate of rejection among informational and operational gene (21.4% and 32.5%, respectively).

**DISCUSSION**

**Comparison of Methods: From Phylogenetics to Phylogenomics**

We have used complete genome sequences of several Proteobacteria to analyze the phylogenetic relationships of a group of Bacteria characterized by their endosymbiotic relationship with different insects. Starting from the genome of one of these bacteria, *Blochmannia floridanus*, the endosymbiont of carpenter ants, we have obtained
the phylogenetic trees for every protein coding gene in this species, by using single-gene, standard phylogenetic methods, and we have also analyzed this data set with phylogenomic approaches.

Single-gene phylogenies are representative of traditional phylogenetic analyses. They have several advantages over multigene approaches. Taxon sampling is widest and the acquisition of raw data in the laboratory and from public databases is easy and cheap. However, our analysis with 16S rDNA exemplifies the low robustness of this marker for phylogenetic reconstruction, because different topologies were obtained when the evolutionary model or reconstruction method were changed (see also Herbeck et al., 2004). Furthermore, the analysis of the B. floridanus phylome has revealed the presence of a large number of different topologies from single gene trees in agreement with other studies (Canbäck et al., 2004). As a case in point, only three of the gene present in all genomes had exactly the same topology than the presumed species tree obtained by the concatenate analysis (Fig. 5). This high variability in reconstructed topologies is the result of limitations in the phylogenetic methods, sampling error due to the limited length of single-gene alignments, the action of evolutionary forces, and model misspecification that may result in single-gene phylogenies not coincident with the presumed species tree. Therefore, the topological heterogeneity obtained by using a single-gene approach, both for the 16S rDNA and the phylome, does not justify in most cases their use in the inference of the evolutionary relationships among species, especially when whole genome data are available. However, still very often these are the only kind of data available for phylogenetic analysis. In this case, congruence with the known taxonomy and convergence with other phylogenetic markers available are reassuring but incongruence does not necessarily mean that a whole new evolutionary history has to be envisioned. Furthermore, the joint analysis of many single-gene phylogenies provides a better, more accurate picture of the evolutionary history of the whole genome, in which different evolutionary trajectories may be revealed once noisy and/or unreliable signals are identified and considered appropriately.

On the other hand, phylogenomic methods, which can be divided in supermatrix, consensus, and supertree approaches, have their own problems (Bininda-Emonds and Sanderson, 2001; Bininda-Emonds et al., 2002; Bininda-Emonds, 2004b). Although the concatenates are most useful for retrieving a main phylogenetic signal, techniques based on the analysis of underlying gene phylogenies may reveal the divergent signals encoded in the single gene that are usually ignored by concatenate analysis (see Gatesy and Baker, 2005).

In this study we have used the traditional consensus (with the ‘200-gene’ data set) and supertree (with the ‘200-gene,’ ‘379-gene,’ and ‘579-gene’ data sets) approaches. Except for the complete phylome supertree, all the methods agreed in placing the Xanthomonadales with the β-Proteobacteria due to the high incidence of gene phylogenies with this placement instead of the expected gamma placement. This case has been studied in detail elsewhere (Comas et al., 2006a). Deviations in supertrees from the presumed species tree could arise from two methodological sources: errors due to heterogeneity in the taxa used for analysis or from differences in their most common topological position in the individual trees (Creevey, 2004). However, the former problem can be excluded in this case because we have worked with a very homogeneous data set and the conflicting result appears even in the two analyses based on the ‘200-gene’ trees that are common to all the genomes considered. Thus, the three different supertree analyses (‘200-gene,’ ‘379-gene,’ and ‘579-gene’) have revealed a large amount of conflicting signals in these gene about the phylogenetic positioning of Xanthomonadales. This group is also the one responsible for most discrepancies between the consensus and supertree approaches, especially over the support values of the nodes. It has to be noted that consensus trees are summaries of the observed clades in the underlying gene trees, all obtained from the complete set of the same number of taxa, whereas support values in a supertree analysis are not a mere recount of observed clades but a measure of the support, derived from bootstrap resampling (Burleigh et al., 2006), that the source gene trees provide to their most compatible topology.

Additionally, the use of supertrees is suitable for many situations where supermatrix or consensus may be limited or nonapplicable. For example, in cases where the evolutionary distance among taxa is very high and the amount of shared genetic information is very low (Sanderson and Driskell, 2003), it is highly likely that sparse data matrices are obtained, leading to concatenate alignments with many missing characters and individual trees with incomplete sets of taxa. Furthermore, supertrees are useful in cases, such as this work, where more than one phylogenetic signal is present because they can reveal alternative signals appearing in single gene phylogenies. Finally, supertrees stand as the best choice method when phylogenetic data sources are heterogeneous; for instance, when combining morphological and molecular data based phylogenies.

The other phylogenomic method commonly used is concatenation of shared sequences. With this approach we derived a phylogenetic tree that was identical to the reference tree and in agreement with the most commonly accepted phylogeny for the taxa considered. This led us to adopt it as the “presumed species tree,” a hypothesis that received further support when alternative reconstruction methods were applied. But the concatenation approach cannot be considered free from errors and/or biases, the most important one being the assumption that a single, main phylogenetic signal exists and that it will be revealed by the simultaneous consideration of as many data as possible, regardless of their (evolutionary) congruence in terms of history or mode of evolution (Daubin et al., 2002).

Further problems arise in the concatenation approach because the number of gene shared in all the genomes decreases when the evolutionary distance between taxa
increases (Charlebois and Doolittle, 2004). In this case we have used only a subset of 200 gene from the 579 present in the *B. floridanus* phylome. Although we have analyzed the concatenate of the complete genome, this approach also presents some problems, mainly the large number of missing characters in the supermatrix and the heavy computational load it imposes, with serious limitations on the methods and programs that can be used for its analysis. At least for the taxa we have studied, the subset of 200 common gene seems to be enough for recovering the main phylogenetic signal. In fact, even smaller subsets are still able to recover it, but in these cases there is more uncertainty in the inferred trees since these depend on the specific gene, and their evolutionary histories, chosen for analysis (Comas et al., 2006b).

Contrary to the approaches based on gene trees, the concatenation methodology seems to favor mainly one signal, the most common or less incongruent with all the gene partitions in the alignment. The presence in the concatenation of gene that, when analyzed separately, rejects the reference tree implies that incongruent or conflicting phylogenetic signals are hidden at least in alignments corresponding to a large fraction of the genomes (Gatesy and Baker, 2005). Despite these potential problems, particularly important in bacterial phylogenetics, concatenation is the most popular phylogenetic method. Tools like the Microbial Genome Database for Comparative Analysis (Uchiyama, 2003) allow recovering the shared gene content from the growing number of completely sequenced bacterial genomes, thus enabling an easier and faster analysis of concatenates. Similarly, recent works propose algorithms for obtaining maximum concatenated sequences from databases (Driskell et al., 2004) and the analysis of the required number of gene (Rokas et al., 2003) or the impact of missing data to obtain a good phylogeny (Philippe et al., 2004). These works are allowing the formalization of a concatenation methodology.

As we have shown for this particular data set, supertree and supermatrix methods allow exploring all the phylogenetic signals contained in a bacterial genome. Both analyses are valid, complementary starting points in the evaluation of the incidence of events such as vertical transmission, horizontal gene transfers, duplications, or phylogenetic noise. Supermatrix methods are a double-edged sword, because they allow recovering the strongest phylogenetic signal even if the supermatrix is composed by alignments where this signal is hidden; i.e., it is not the strongest one (Gatesy 1999, Gatesy and Baker, 2005), but, at the same time, this could result in the masking of other alternative signals. The presence of these alternative signals is more easily revealed by analyses based on the underlying gene topologies. As shown in this work, these signals usually arise in a supertree or consensus framework in the form of incongruence around some taxa. For instance, the supermatrix approach failed to reveal the phylogenetic incongruence around *Xanthomonadales* position in the base of the γ-Proteobacteria tree (Fig. 2), whereas both the 200-gene’ tree consensus and the supertree of the 379- and 579-gene’ trees placed them as β-Proteobacteria or as a low-supported γ-Proteobacteria clade (Comas et al., 2006a).

Many factors have to be considered in selecting a methodology for whole genome phylogenetic analyses. We have shown that for intermediate phylogenetic depths, it is possible to recover the primary phylogenetic signal of the genomes through the concatenation of shared genetic content, although caution must be taken because usually a single model of evolution is applied to all the alignment; only recently the use of mixed models is being evaluated (Pagel and Meade, 2004). Meanwhile, the supertree approach seems to be better suited to reveal conflicting gene tree signals (i.e., the positioning of *Xanthomonadales*) and when data are sparse (Sanderson and Driskell, 2003), which is usually the case for large evolutionary distances. One such example is the supertree obtained from 730 gene trees of 45 organisms from the three domains of life (Daubin et al., 2002). The gene tree approach is suitable when working with species without complete genome sequences or for fast, exploratory phylogenetic analyses. These techniques are not incompatible but complementary, allowing the detection of possible sources of error and pointing towards interesting evolutionary problems. As we have shown, the use of different methodologies, from single-gene to genome-scale phylogenies, allows the identification of the different phylogenetic signals encoded in bacterial genomes.

The Phylogenetic Landscape of Proteobacteria: The Placement of *Xanthomonadales* and Insect Endosymbionts

Previous reports and this phylogenomic analysis of Proteobacteria have identified two main problems in the phylogeny of these bacteria. On one hand, the monophyletic or polyphyletic origin of insect endosymbionts has generated a debate. Most studies using a multigene approach have placed this group as a monophyletic clade sister to the YESS cluster (Lerat et al., 2003; Canbäck et al., 2004). However, studies with single-gene phylogenies have pointed out that the group is not monophyletic (Charles et al., 2001; Herbeck et al., 2004), with the *Buchnera* clade as sister of the cluster YESS and suggesting that the other two endosymbionts (*B. floridanus* and *Wigglesworthia brevipalpis*) could have an independent origin related with secondary endosymbionts. On the other hand, the concatenate analysis, although presenting a strong primary phylogenetic signal, revealed some secondary topologies with incongruence in the placement of *Xanthomonadales*. In consequence, the topologies obtained in this study have been analyzed and discussed based on the position and evolutionary relatedness of the *Xanthomonadales* and insect endosymbionts groups.

The most frequently used phylogenetic marker in bacterial systematics is 16S rDNA (Woese, 1987). In this particular case, the 16S rDNA topology was not robust to different methods of phylogenetic reconstruction as the result obtained depended on the method and model of evolution used for analysis. However, despite these discrepancies, all 16S rDNA topologies clustered
the Xanthomonadales with the β-Proteobacteria. A screening of the gene trees reveals that the three Xanthomonadales usually conform to a monophyletic group whose placement in the Proteobacteria tree is unclear (see also Comas et al., 2006a).

As we have described, all the concatenate analyses revealed a main phylogenetic signal in which the Xanthomonadales are placed at the base of γ-Proteobacteria. However, a secondary phylogenetic signal was also observed. This signal also appeared in some single gene phylogenies in which the Xanthomonadales were excluded from the gamma clade. Therefore, although the use of concatenates of different numbers of gene (60, 200, and 579) has allowed us to establish the main phylogenetic signal, this approach also indicated the existence of alternative signals through different analyses evaluating the support for the main topology. Obviously, the larger the number of gene incorporating the alternative signal, the higher the support for conflicting phylogenies will be obtained.

Moreover, this secondary signal appeared more strongly in the consensus and supertree analyses. The majority rule consensus tree and the supertree of the 200 common gene as well as the supertree obtained from the other 379 genes placed again the Xanthomonadales with the β-Proteobacteria. These methods are only indirectly related to the sequence data because they are based on the gene trees derived from them. Because of this, they reflect what was already hinted in the single-gene analysis, that an important fraction of gene exclude the Xanthomonadales from the gamma subtree, thus revealing the existence of either nonvertical evolutionary events, or methodological problems, or both, in the placement of Xanthomonadales in the source trees (but see Comas et al., 2006a).

The evolutionary relationship of γ-Proteobacteria endosymbionts remains a controversial issue. The first attempts to address this problem using single-gene phylogenies produced conflicting results. For example, Heddi et al. (1998) and Schroder et al. (1996) supported the common ancestor hypothesis whereas Charles et al. (2001) defended a paraphyletic origin for insect endosymbionts. More recently, even with the availability of genomic data, the same conflicting results have been reproduced. Three other phylogenomic analyses support the common ancestor hypothesis: Gil et al. (2003) obtained this result after maximum likelihood and Bayesian analyses of 61 conserved proteins involved in translation; Lerat et al. (2003) obtained it from the common core of proteins, those present in the common ancestor and not involved in lateral transfers, for 13 γ-Proteobacteria species including one Buchnera and one Wigglesworthia species; and Canbäck et al. (2004) also examined the relationship of Proteobacteria including two Buchnera and W. brevisalpis using three different methods. These three studies agree with ours in postulating a common ancestor for the endosymbionts and placing them as sister group to the YESS cluster.

However, two other studies have reached different conclusions: Herbeck et al. (2004) analyzed the 16S rDNA of endosymbionts and concluded that GC bias has influenced most published phylogenies. Using Galtier and Gouy’s (1995) maximum likelihood method, they evaluated an array of possible phylogenies obtained by standard methods and models as well as other alternatives derived from the permutation of the Buchnera position in these source trees. The topologies preferred by Galtier and Gouy’s method were those showing a paraphyletic origin of endosymbionts. This same strategy was used with the groEL gene and they reached the same conclusion. Although the two single-gene analyses supported the paraphyly of α-Proteobacteria insect endosymbionts, the corresponding topologies were incongruent and some taxa appeared in unlikely positions. A similar result has been obtained by Belda et al. (2005) using an alternative method for phylogeny reconstruction based on whole genome analysis. These authors have analyzed the number of pairwise genome rearrangements and gene order from 244 gene common to 30 γ-Proteobacterial genomes. Their analysis concluded that insect endosymbionts do not conform a monophyletic cluster, with B. floridanus being closer to Escherichia coli and separate from Wigglesworthia and Buchnera, genera that do not show a common origin themselves. Nevertheless, the lack of monophyly for insect endosymbionts is the only common result from these two studies, because the retrieved topologies are not compatible.

All our phylogenetic and phylogenomic analyses recover a single clade for endosymbionts, with the notable exception of the 16s rDNA analysis with the Galtier and Gouy model, which corrects for GC content biases. In consequence, we have addressed the possible influence of the GC content in the topologies recovered with the aim of discarding a convergence artifact in the rest of the analyses. Although we have based our phylome analysis on amino acid sequences, these are not free from base composition artifacts (Foster and Hickey, 2002). From the concatenate alignment of the 200 proteins common to all the genomes, we removed those positions most likely to be influenced by GC bias. The trees obtained with two different methods were coincident and identical to the species tree shown in (Fig. 5). Furthermore, the support values obtained for each node were even higher. Consequently, our results are robust to GC bias effects and they support the existence of a common ancestor for insect endosymbionts.

The Genomic Core and the Interpretation of Incongruence in Bacterial Genomes

One controversial issue in current evolutionary genomics studies of bacteria is the existence of a core of gene that share a common (i.e., vertical) evolutionary history (Makarova et al., 1999; Nesbo et al., 2001; Bapteste et al., 2005; Susko et al., 2006). It has been argued that the vast majority of components of this core belong to the informational category of gene. Most of these gene are important because of their central role in essential metabolic pathways or because of their high connectivity in terms of macromolecular interactions. The consequence of this
division is that informational gene are less prone to horizontal gene transfer (HGT) than operational ones (Jain et al., 1999).

One remarkable result from our phylogenetic analyses is that a substantial fraction (21.4%) of informational gene reject the presumed species tree, whereas for operational ones the corresponding value is not too much larger (32.5%). To prove the existence of a genomic core, Daubin et al. (2002) also computed the Robinson-Foulds distances for all of gene trees pairs from an initial pool of 310 trees. Next, they represented these distances in a two-dimensional space by means of principal coordinates analysis and chose as a core those gene trees in the highest density zone. Using this method they selected 121 gene to establish their set of core gene. Their most relevant result is that nearly half of these gene were operational. Therefore, it seems that the division between operational and informational gene is not adequate to explain the existence of incongruences and of a core of gene in bacterial genomes. These results call for some caution when selecting gene for phylogenetic analysis on the basis of their functional category assignment (see also Comas et al., 2006b). In fact, in our concatenate analysis we have shown that the presence of gene that reject the presumed species tree and the mixture of operational and informational gene are not sufficient obstacles to recover the main evolutionary history of these species.

It is also relevant to note that most of the gene we have analyzed are present in the genomes as single-copy gene. This means that possible incongruences not related with methodological problems between gene and presumed species trees for these gene must be classified as cases of nonorthologous gene displacement (horizontal gene transfer, hidden paralogies). Lerat et al. (2003) carried out a similar approach to estimate the incidence of incongruence, which is analyzed in more detail elsewhere (Comas et al., 2006a) and is aimed only at revealing the importance of considering the different phylogenetic signals present in bacterial genomes. This study, which is mainly centred on analysing the evolutionary relationships of endosymbionts, has not been designed to specifically detect horizontal gene transfers. Endosymbionts have not undergone gene exchange since they established their close intracellular relationship with the insect hosts (Akman et al., 2002; Tamas et al., 2002; Gil et al., 2003; van Ham et al., 2003). However, our analysis has pointed towards taxa where phylogenetic incongruence is particularly pervasive. A closer look to Xanthomonadales (Comas et al. 2006a) has allowed distinguishing between phylogenetic noise and nonvertical evolutionary events, and similar analyses (Beiko et al., 2005) might also be revealing of the complexities of the evolutionary processes in bacterial lineages.

ACKNOWLEDGEMENTS

We express our gratitude to the associate editor, Ron DeBry, and the reviewers, Amy Briskell and James McInerney, for their many valuable comments and suggestions that have greatly helped us. This work has been supported by projects BMC2003-00305 from Spanish Ministerio de Ciencia y Tecnología and Groups 03/2004 from Generalitat Valencia. IC is recipient of a predoctoral fellowship from the Spanish Ministerio de Educación y Ciencia.

REFERENCES


Camponotus ants on bamboo. Photo courtesy of Dr. Heike Feldhaar.