The Evolutionary Origin of Xanthomonadales Genomes and the Nature of the Horizontal Gene Transfer Process

Iñaki Comas,† Andrés Moya,* Rajeev K. Azad,† Jeffrey G. Lawrence,† and Fernando Gonzalez-Candelas**

*Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universidad de Valencia, Valencia, Spain and
†Department of Biological Sciences, University of Pittsburgh

Determining the influence of horizontal gene transfer (HGT) on phylogenomic analyses and the retrieval of a tree of life is relevant for our understanding of microbial genome evolution. It is particularly difficult to differentiate between phylogenetic incongruence due to noise and that resulting from HGT. We have performed a large-scale, detailed evolutionary analysis of the different phylogenetic signals present in the genomes of Xanthomonadales, a group of Proteobacteria. We show that the presence of phylogenetic noise is not an obstacle to infer past and present HGTs during their evolution. The scenario derived from this analysis and other recently published reports reflect the confounding effects on bacterial phylogenomics of past and present HGT. Although transfers between closely related species are difficult to detect in genome-scale phylogenetic analyses, past transfers to the ancestor of extant groups appear as conflicting signals that occasionally might make impossible to determine the evolutionary origin of the whole genome.

Introduction

The evolution of gene content of bacterial species is strongly influenced by their ability to incorporate DNA from other species in a process known as horizontal gene transfer (HGT) (Koonin et al. 2001; Boucher et al. 2003). The study of HGT events has shifted from reports of individual cases to genome-scale analyses taking advantage of the growing number of microbial genomes sequenced (Koonin and Galperin 1997; Koonin et al. 2001).

Although the importance of HGT as generator of evolutionary novelty is widely recognized (Ochman et al. 2000), its impact on the inference of organismal phylogenies is still hotly debated. The availability of a large number of microbial genome sequences has allowed the construction of genome phylogenies using different types of information (Snel et al. 2005), with raw sequences, gene trees, shared gene content, and shared gene order being most widely used. Typically, the impact of HGT on these genome phylogenies has been neglected and considered as mere phylogenetic noise in favor of a vertical signal resulting from the transmission of information from ancestors to descendants. Yet several authors (Doolittle 1999; Gogarten et al. 2002; Kunin et al. 2005) have claimed that retrieving a tree of life for bacteria is impossible, noting that 1) every gene has been transferred at least once during its evolutionary history, and therefore 2) the phylogenetic signal associated to HGTs opposes, and often overcomes, the vertical signal, hence obscuring the deep phylogenetic relationships among current genomes.

The analysis of whole genomes has shown that incongruence between gene trees and organismal phylogenies is a pervasive feature of a significant fraction of genes from almost every bacterial genome except cases of obligate intracellular associations (Tamas et al. 2002). This incongruence could arise from 2 main sources: phylogenetic noise and HGT. Phylogenetic noise primarily results from sequences with poor phylogenetic signal, high evolutionary rates for certain genes or lineages, or long-branch attraction problems. On the contrary, the signals derived from HGT differ from noise because they usually reflect a robust and systematic incongruence toward the donors. Therefore, conflicting phylogenetic signals coexist in bacterial genomes due to the vertical and horizontal histories of genes as well as phylogenetic noise. Divergent signals appear in core gene sets (Bapteste et al. 2005; Susko et al. 2006) indicating that incongruence, often interpreted as HGT, could affect any gene and cellular function.

Gogarten et al. (2002) proposed a model that assumes that the likelihood of transfers is higher among related organisms thus generating a phylogenetic signal indistinguishable from the vertical one. Therefore, the treelike evolution in bacteria is merely showing preferred paths for transfer events. In a recent analysis (Beiko et al. 2005), 144 genomes were screened looking for the phylogenetic origin of all possible HGT events. This analysis suggested the preference of gene sharing among relatively closely related taxa, thus reaffirming the hypothesis that there are some constraints on HGT related to the compatibility of genome architectures and/or their phylogenetic distance (Gogarten et al. 2002; Hendrickson and Lawrence 2006).

In this work, we have focused on a particular group of bacteria, the Xanthomonadales, which seem to have been especially affected by HGT. This is the most basal group of the gamma-Proteobacteria clade, and it is composed by phytopathogens ranging from obligate associations, like *Xylella* species, to nonobligate associations, like those belonging to the *Xanthomonas* genus (Van Sluys et al. 2002). Previous works have revealed an unstable position of Xanthomonadales in the Proteobacteria tree (Van Sluys et al. 2002; Beiko et al. 2005). Both individual gene phylogenies and new genome phylogenies have placed them with the same frequency as beta-, gamma-, or alpha-Proteobacteria or as an external clade to the 3 groups (Hauck et al. 1999; Van Sluys et al. 2002; Omelchenko et al. 2003; Martins-Pinheiro et al. 2004; Bern and Goldberg 2005; Dutilh et al. 2005). Indeed a non–gamma-Proteobacteria position for the Xanthomonadales is increasingly common in recently published phylogenies (Van Sluys et al. 2002; Omelchenko et al. 2003; Creevey et al. 2004; Martins-Pinheiro et al. 2004; Dutilh et al. 2005; Studholme et al.

Key words: phylogenomics, microbial evolution, conflicting phylogenetic signals, horizontal gene transfer.

E-mail: inaki.comas@uv.es.

Mol. Biol. Evol. 23(11):2049–2057. 2006
doi:10.1093/molbev/msl075
Advance Access publication August 1, 2006

2005). These reports point toward 2 most probable explanations: phylogenetic noise or HGT.

On the one hand, phylogenetic noise is expected to affect up to certain degree gene phylogenies particularly for basal groups, whose position may change due to limitations of phylogenetic reconstruction methods. On the other hand, following the hypothesis of Gogarten and coworkers (2002), it is expected that transfers between Proteobacteria and Xanthomonadales genomes were more likely in the past when the divergence among the major groups was relatively recent, whereas recent transfer is expected to be among more closely related Xanthomonadales species. From a phylogenomic perspective, a high amount of ancient transfers from different donors to the Xanthomonadales ancestor might result in an unstable or unresolved position of these species in the proteobacterial tree.

We have assessed whether the origin of the observed conflicting reports for Xanthomonadales in the proteobacterial tree is the result of phylogenetic noise due to convergence and/or loss of signal or to recent or ancient HGT events. We have considered as phylogenetic noise the results of those processes unrelated to the form of transmission of a gene (horizontal or vertical), which violate the assumptions of phylogenetic reconstruction methods. Alternatively, we have considered as phylogenetic signal that derived from the vertical or horizontal transmission of genes. To separate these 2 components in the genomes of Xanthomonadales, we first identified all possible phylogenetic signals encoded therein. Next, we investigated the affinity of the genes for gamma-, beta-, or alpha-proteobacterial clades. Our results indicate the existence of different, robust phylogenetic signals on the genomes of Xanthomonadales with origins in the 3 groups considered. We show that, unlike phylogenetic noise, these signals are not randomly distributed among genes; adjacent genes with the same phylogenetic signal appear more often than expected by chance in Xanthomonadales genomes, indicating that the signal detected is not due to selecting a conservative significance threshold for noise. These results are analyzed in light of proposed models for the impact of HGT and preferred gene-sharing paths in bacteria.

Methods

Selection of Homologs, Gene Alignments, and Gene Trees

Initially 18 Proteobacteria genomes were selected for analysis (see table S1, Supplementary Material online). The data set included a balanced number of representatives from the 3 major Proteobacteria groups and 3 Xanthomonadales genomes. We used *Xanthomonas citri* as the base genome for selecting putative orthologs from the 17 additional genomes analyzed. For each protein-coding gene, we used a Reciprocal Best Hit Blast strategy (Altschul et al. 1997). We accepted as possible orthologs those genes that were reciprocal best hit between 2 genomes. Blast searches were performed at the NeuroGadgets Inc. Bioinformatics Web Service server (Charlebois et al. 2003) with a very stringent criterion (*e* value = 1×10^{-10}) to minimize problems associated to Blast identifications. The annotation of each sequence and the corresponding multiple alignments were revised individually to discard wrongly identified putative orthologs.

As a first step, we analyzed possible incongruence not due to the evolutionary position of Xanthomonadales. We selected those proteins common to the 18 genomes and obtained their gene tree as explained below. With these gene trees, we obtained a consensus topology. As the consensus reflects the most frequent position of each taxon in the tree, we could identify non-Xanthomonadales species with an unresolved phylogenetic position and which, therefore, might affect future phylogenetic congruence analyses. Once these problematic taxa, *Nitrosomonas europaea* and *Legionella pneumophila*, were removed, we repeated a Blast search for the remaining genomes as explained above, and finally, only those sequences present in at least 10 species were considered for further analysis. This resulted in a set of 1,051 genes of which 207 were present in the 16 genomes finally considered.

Each selected protein from the set of putative orthologs was aligned with ClustalW (Altschul et al. 1997) using default parameters. Phylogenetic trees were inferred by maximum likelihood with PHYML (Guindon and Gascuel 2003), using JTT model (Jones et al. 1994) as the model of amino acid evolution with a gamma distribution with 8 categories for modeling substitution rate heterogeneity among sites and an additional category of invariant sites estimated from the data set.

In order to explore the evolution of the genes identified in the previous step, we performed 2 different phylogenetic analyses. For assessing the different phylogenetic signals embedded in the Xanthomonadales genomes, we carried out firstly a "congruence map" analysis. Additionally, we addressed directly the question of the proteobacterial origin of each gene by analyzing the support for different plausible evolutionary scenarios for each gene.

Congruence Map Analysis

We applied a previously described procedure (Bapteste et al. 2005) to construct a congruence map for the 207-gene set. Each gene alignment was tested against every gene tree (207 × 207 comparisons) by means of the expected likelihood test (ELW) test (Strimmer and Rambaut 2002). For the congruence map, we only considered "acceptance" or "rejection" of the corresponding topology. The topologies (columns) were identified as alpha, beta, or gamma by visual inspection.

Phylogenetic Origin Analysis

For the 1,051 genes found in the search for putative orthologs, we tested their congruence to 5 possible phylogenetic hypotheses (fig. 1) using the ELW test of topologies. The 5 topologies were "artificially" generated and can be divided into 2 groups. Firstly, we tested 2 control phylogenies: a star phylogeny (STAR1) with no resolved nodes and a star phylogeny (STAR2) in which the tips of the main phylogenetic groups (alpha-, beta-, and gamma-Proteobacteria and Xanthomonadales) were resolved. Secondly, we generated 3 phylogenies for which the only difference was the placement of Xanthomonadales at the base of gamma- (G-tree) or beta-Proteobacteria (B-tree) and between alpha and the beta–gamma split (A-tree).

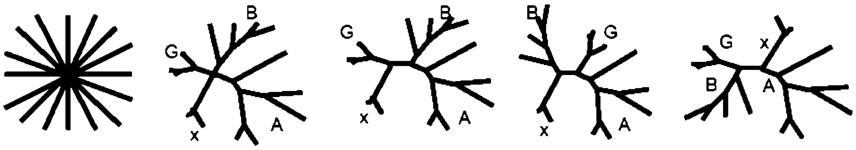

FIG. 1.—The 5 topologies used for the phylogenetic preference test. The first 2 are named STAR1 and STAR2 and are used as phylogenetic signal control phylogenies. The other 3 topologies assume a placement of Xanthomonadales nearest to gamma (G-tree), beta (B-tree), or alpha (A-tree) species.

For most genes, the ELW test could not reject all but one phylogeny. In consequence, we eliminated from further analyses those genes that could not reject both star phylogenies, hence ensuring the presence of some phylogenetic signal. The preferential phylogenetic origins of the genes were mapped into the 3 Xanthomonadales genomes considered.

Testing for Long-Branch Attraction Artifacts

Once the most plausible phylogenetic origin of each gene had been assessed, we tested for possible convergences due to shared high rates of substitutions and not to common origins. We used the program RRTree 1.0 (Robinson-Rechavi and Huchon 2000) with the 207 common genes' data set. We divided the species into 4 groups according to their taxonomic assignment (alpha-, beta-, and gamma-Proteobacteria and Xanthomonadales) and performed all possible pairwise comparisons of substitution rates between the 4 groups. As a common outgroup, we chose the corresponding gene of *Ricketssia prowazekii*. The significance of the difference in number of substitutions for each comparison was assessed at the 0.0083 level (Bonferroni corrected significance level [α], 0.05) to take into account the 6 non-independent comparisons performed for each gene.

Testing for Functional Association and Clustering Along the Genome

We have studied the relationship between phylogenetic origin and functional assignment of the genes. We used the functional categories described in the clusters of orthologous groups of proteins (COG) database (Tatusov et al. 2000) to classify the genes into 4 general or 21 more detailed categories. We have also compared our results with a list of virulence-associated genes of *X. citri*.

We used a subset of the previous data set to analyze whether genes with the same phylogenetic origin tended to cluster in the 3 Xanthomonadales genomes. We selected those genes with at least one adjacent gene with phylogenetic information. We computed the number of the 6 possible combinations of alpha, beta, and gamma assignments for all pairs of adjacent genes in each genome. We calculated the expected number of pairs in each category under the assumption of independent origin for each gene in a pair. Let AA, BB, GG, AB, AG, and BG denote the possible observed pairs. We obtained the observed frequencies of each phylogenetic assignment (pA, pB, and pG) and calculated the expected frequency of each pair as the product of the individual frequencies of its components. For example, the expected number of alpha–beta pairs is given by $2 \cdot pA \cdot pB \cdot N$, where N is the total number of pairs considered. The observed and the expected number of pairs were compared by means of a chi-square test with 3 degrees of freedom (df). A whole-genome alignment of the 3 Xanthomonadales genomes was carried out with MAUVE (Darling et al. 2004) in order to study the possible influence of rearrangements in the results of the clustering test.

Lastly, we tried to determine whether adjacent pairs of genes with the same phylogenetic origin were also in a potential operon. Pairs of genes used in the clustering analysis were assigned to 2 groups, adjacent and nonadjacent pairs. For each group, we counted the number of times the members were in the same direction of transcription, the number of divergently transcribed gene pairs, and the number of convergently transcribed gene pairs. A chi-square test was used for assessing the significance of the possible differences between adjacent and nonadjacent pairs with 2 df.

Detection of Atypical Genes

We have used a new technique (Azad and Lawrence 2005) that takes advantage of common parametric measures for HGT detection and akaike information criterion (AIC) as a criterion for clustering. This technique allows the identification not only of clusters of native genes in the *X. citri* genome but also of different clusters of atypical genes. The 2 parametric measures used were nucleotide composition and codon bias. Briefly, these measures were computed for all the genes in the *X. citri* genome, and the AIC criterion was used for deciding when the addition of more genes to a cluster was not significant. Frequently, this procedure retrieves a large gene cluster that usually corresponds to the native or typical genes of the genome analyzed and smaller clusters of genes with atypical features. Genes shorter than 300 bp were excluded because their low information content could result in GC content or codon usage biases (Lawrence and Ochman 2002). A summary of the analysis pipeline followed in this work is shown in figure 2.

Results

We have analyzed a set of 18 Proteobacteria genomes in order to study the phylogenetic origin of Xanthomonadales genes. The data set included a balanced number of representatives from the 3 major Proteobacteria groups and 3 Xanthomonadales genomes, *Xanthomonas axonopodis* pv. *citri* str. 306 (*X. citri*, Xci), *Xanthomonas campestris* pv. *campestris* str. ATCC 33913 (Xca), and *Xylella fastidiosa* 9a5c (*Xy. fastidiosa*, Xy) (fig. 3 and table S1, Supplementary Material online). Our goal was to analyze the different phylogenetic signals present in their genomes as well as to determine the phylogenetic origin of their genes from gamma-, beta-, or alpha-Proteobacteria ancestors.

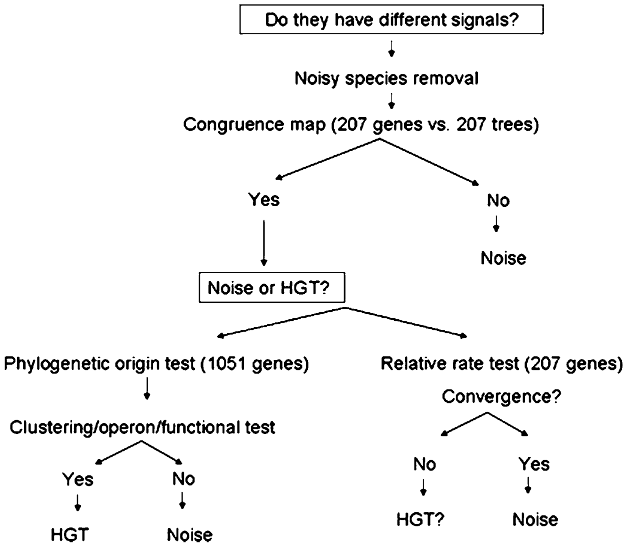
Do they have different signals?
Noisy species removal
Congruence map (207 genes vs. 207 trees)
Yes
No
Noise
Noise or HGT?
Phylogenetic origin test (1051 genes)
Clustering/operon/functional test
Yes
No
HGT
Noise
Relative rate test (207 genes)
Convergence?
No
Yes
HGT?
Noise

FIG. 2.—General pipeline of the methodology followed to analyze phylogenetic signal in the genomes of 3 Xanthomonadales species when compared with other proteobacterial genomes.

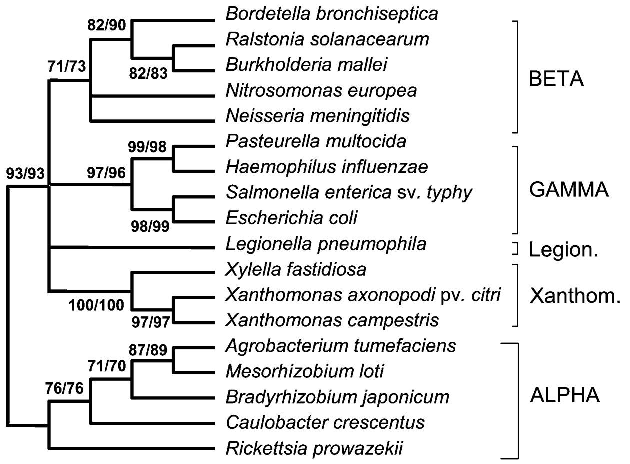

FIG. 3.—Majority-rule consensus for the initial set of 18 genomes of the 207 gene trees used in the congruence map analysis. The tree is arbitrarily rooted with the alpha branch. The nodes show the frequency of appearance of the corresponding group before (left) and after (right) removal of *Legionella pneumophila* and *Nitrosomonas europaea*. Species names and taxonomy groups according to National Center for Biotechnology Information.

Our initial search for putative orthologs retrieved 207 genes common to all the genomes. Our first goal was to identify "noisy" phylogenetic signals. As our analysis was intended to detect only incongruence related with Xanthomonadales, we looked to eliminate other species that could introduce noise in the global phylogenetic analysis. We constructed the majority-rule consensus tree of the trees obtained from these 207 common genes set as a way of summarizing the degree of incongruence present in each species (fig. 3). The tree identified 3 nodes with low resolution, those corresponding to *N. europaea*, *L. pneumophila*, and the Xanthomonadales group. Because we were only interested in the later, we discarded *N. europaea* and *L. pneumophila* from the ensuing analyses.

With these 207 common genes of the remaining 16 genomes, we tested for the presence of different phylogenetic signals. We performed a congruence map analysis in which each gene was tested for congruence against all the other gene trees (fig. 4); here, each row corresponds to a gene and each column to a gene tree. The analysis identified numerous genes, whose phylogenetic reconstructions provide clear and robust support for other alpha, beta, and gamma gene tree topologies. These were defined on the basis of the monophyletic grouping of Xanthomonadales sequences with the remaining sequences of each group. In addition, many tests were able to reject some, but not all, alternative phylogenies. The cases ranged from genes that were compatible only with their own gene tree to those that could not distinguish between alpha, beta, or gamma topologies. Overall, the gamma topologies were the most frequently accepted, considering both cases in which this was the only topology selected and those with other topologies being also accepted. This analysis showed a mixture of noisy phylogenetic signals, which corresponded to genes only congruent with their own gene tree and genes that were congruent with almost any topology. But, in addition to noisy signal, robust but divergent phylogenetic signals were detected in terms of acceptance or rejection of groups of topologies corresponding to different positions of Xanthomonadales with respect to other Proteobacteria.

The 207-gene analysis strongly suggested that HGT might have played an important role in the evolution of Xanthomonadales, but other alternatives could also be considered. The HGT hypothesis was tested by selecting a larger gene data set that would result in more robust statistics. This allowed us to analyze whether the cause that some genes were unable to reject alternative, incompatible hypothesis in the congruence map was the result of phylogenetic noise or the hallmark of past HGT events. Consequently, we extended the initial data set to incorporate genes from *X. citri* with orthologs in at least 10 genomes (see fig. S1, Supplementary Material online). The extended set of 1,051 genes was composed of quasiuniversal genes in Proteobacteria with functions not drectly related to the virulence of *X. citri*. A comparison with a list of known virulence-related genes identified only 19 candidates.

To determine if the genes contributing conflicting phylogenetic signals were recently introduced in the Xanthomonadales genomes, we identified atypical genes by a clustering methodology based on the AIC criterion using both codon usage bias and nucleotide composition as discriminating criteria (Azad and Lawrence 2005). We found a main cluster of typical genes and several clusters of atypical ones. The analysis revealed that only 0.2% for codon usage, and 13.61% for nucleotide composition of the genes used in this study were atypical. Meanwhile, the frequencies of atypical genes in the whole genome were 2.21% and 22.23%, respectively. Therefore, the incongruence observed in the congruence map analysis cannot be attributed to the confounding influence of recent HGT events. Although some of these cannot be ruled out, this is an unlikely scenario for the whole set of conflicting genes because these phylogenetically incongruent genes were found in all 3 Xanthomonadales genomes, with appropriate branching

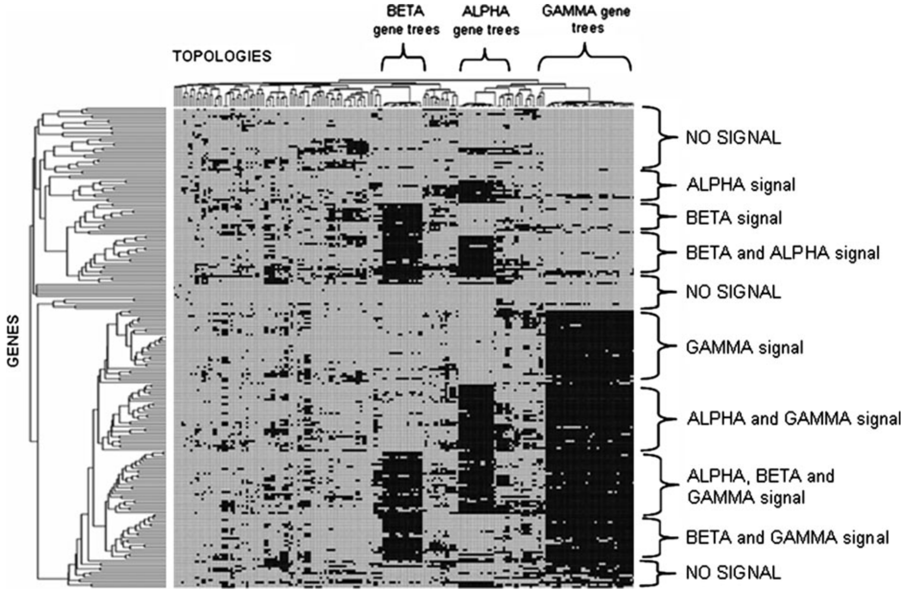

FIG. 4.—Congruence map of the patterns of acceptance/rejection of each gene tree (columns) by each gene (rows). Rows are arranged following the Euclidean distance and posterior unweighted pair group method with arithmetic mean clustering between the acceptance/rejection vectors of each gene (left-side tree). Topologies are arranged following the order derived from the single linkage clustering of the Euclidean distances of the corresponding column vectors of acceptance/rejection (upper tree). Dark gray dots indicate acceptance of a particular gene tree by a gene, whereas light gray dots indicate rejection. The right column indicates the nature of the topologies accepted in each region (alpha, beta, and gamma topologies, only few gene trees [no signal], and almost all gene trees [no signal]).

orders. These results suggest that the transfer events were old.

To evaluate the possible incidence of long-branch attraction artifacts in our data sets, we also carried out relative rate tests for the 207 genes common to all the genomes. The analyses revealed only one case of possible convergence due to shared high rates of substitutions (corresponding to the *hisS* gene). The remaining genes showed no evidence of grouping due to shared accelerated evolutionary rates and, in consequence, we excluded this phylogenetic artifact as responsible for apparently incongruent groupings.

To verify that ancient transfers to Xanthomonadales genomes resulted in phylogenetic incongruence among quasiuniversal genes, we examined the compatibility of each gene with 5 phylogenetic hypotheses (fig. 1). The STAR1 and STAR2 topologies are unresolved topologies with the difference that in the latter the tips of the major Proteobacteria groups are resolved; genes with strong phylogenetic signal should reject these topologies. The other 3 topologies placed the Xanthomonadales clade as the most basal group of the gamma (G-tree), beta (B-tree), or alpha (A-tree) groups. All of the 1,051 genes analyzed rejected the STAR1 topology, but 51 genes could not reject the STAR2 topology; these were removed from the ensuing analyses. The distribution of the most likely phylogenetic origin of the remaining genes is shown in figure 5. A majority of genes preferred the A-tree. However, most of these genes were unable to reject some or all the other topologies. In consequence, posterior analyses were based on the most likely assignment regardless of their compatibility with other alternatives. In any case, an analysis of those genes selecting only one of the topologies revealed the same pattern, with alpha topologies as the most preferred and beta topologies as the least.

As the phylogenetic origin test was not enough to distinguish between phylogenetic noise and phylogenetic signal, we tested the noise threshold by analyzing the distribution of the genes and their possible origins in the Xanthomonadales genomes. An adjacency analysis was carried out with a reduced subset of the 1,051-gene set. We selected only those genes that were adjacent at least to another gene from this subset in the genomes of the Xanthomonadales. The number of pairs analyzed was 430 in *X. citri*, 438 in *X. campestris*, and 377 in *Xy. fastidosa*. This allowed us to test 2 alternative predictions. If incongruence was merely due to noise, then pairs of adjacent genes would show no association with respect to their phylogenetic origins. On the other hand, at least under certain models for HGT (Lawrence and Roth 1996), pairs of genes adjacent in the recipient genome should tend to share the same phylogenetic origin. Our statistical tests revealed that the number of adjacent pairs of genes with the same phylogenetic

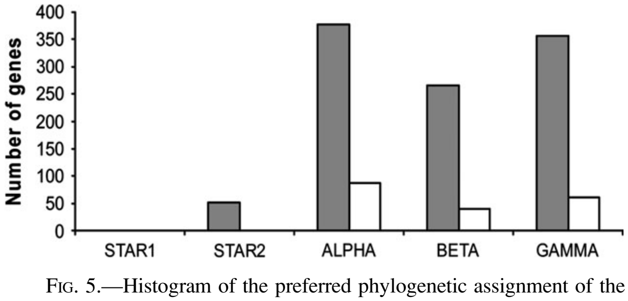

FIG. 5.—Histogram of the preferred phylogenetic assignment of the 1,051 genes analyzed. As most genes are compatible with more than one topology, the figure shows the results for the complete data set (black bars) and that for the 187 genes that are congruent with only 1 of the 5 hypotheses (white bars).

origin in the Xanthomonadales genomes was higher than expected (fig. 6). Furthermore, such clustering was evident and significant for all 3 Xanthomonadales genomes and the 3 possible phylogenetic origins considered. This evidence of clustering highlighted 2 aspects. On the one hand, it rejected the possibility that most of these results were simply the product of phylogenetic noise although this was evidently present in some degree. On the other hand, although the adjacency test reduced the analysis to the observed and expected number of pairs, the actual size of the clusters identified tended to be larger than 2 genes, with examples including as many as 8 genes. The mean size of the clusters was 2.35 genes, pointing toward HGT of operons as a possible mechanism of evolution.

If operons were being transferred between gamma-, alpha-, and beta-proteobacterial lineages, resulting in the phylogenetic incongruence seen in Xanthomonadales genomes, then the overabundance of adjacent genes with common phylogenetic signals should be biased toward genes transcribed in the same direction. We studied the transcription direction of the genes present in the adjacent pairs analyzed above; as expected under the hypothesis of horizontal transfer of complete or partial operons, most of the genes identified were present in the same strand. Furthermore, the frequency in which adjacent genes were in the same direction of transcription was higher than that of nonadjacent cases, with significant statistical support for pairs of alpha ($P < 0.0002$) and gamma origin ($P < 0.002$), but not for those of beta origin ($P = 0.3079$). Thus, it is likely that most of the clusters identified in our study, mainly those involving gamma and alpha origins, were operons.

We also investigated the relationship between functional assignment of the genes and their established phylogenetic origin to test whether informational genes were less prone to be transferred than noninformational ones (Jain et al. 1999). We did not detect any association between functional classes and putative phylogenetic origin, but some patterns could be distinguished. For example, 7 of the 8 flagellar genes analyzed showed the same phylogenetic origin, B-tree. Alternatively, the informational category was richer in G-tree topologies, whereas the predominant topology for metabolic genes was the A-tree. As a consequence, the different composition in functional categories of the 207-genes set (richer in informational genes) and the 1,051-genes set might explain the differences in the most frequent origin of the genes in each set (G-trees and A-trees, respectively). Nevertheless, most of categories presented a mixture of topologies including some excellent phylogenetic markers such as genes related to transcription (fig. S2, Supplementary Material online).

Discussion

Xanthomonadales Evolution Illustrates the Nature of the HGT Process

The influence of HGT on the reconstruction of bacterial phylogenetic relationships and also on its relevance to shape their genomes has been a hotly debated issue. Different models have been proposed to explain patterns of transfers derived from complete microbial genomes (Jain et al. 1999; Gogarten et al. 2002; Kunin et al. 2005). Gogarten et al. (2002) proposed that if there is a negative correlation between the likelihood of transfers and the evolutionary distance separating 2 taxa, then HGTs are more likely among closely related taxa, thus generating a cohesive signal for the clade that agrees with that of vertical transmission.

One of the clues for testing the different proposals could be the barely studied subject of constraints to HGT. It is clear that there are some restrictions to transfers, but most studies have focused on functional analyses (Jain et al. 1999; Nakamura et al. 2004; Pal et al. 2005). However, 2 recent works have shed light on the probabilities of successful transfers from a phylogenetic point of view. The results of Beiko et al. (2005) are consistent with an uneven phyletic distribution of the transfers, whereas Lawrence and coworkers (Lawrence and Hendrickson 2003, 2004; Hendrickson and Lawrence 2006) have pointed out a possible molecular mechanism that limits the equally probable sharing of genes among all bacteria.

The analysis of 220,240 proteins from 144 genomes (Beiko et al. 2005) has revealed a consistent vertical signal but also the relevance of horizontal transfer in shaping bacterial genomes. The gene tree for each protein was reconstructed and a reference supertree that is supposed to represent the vertical history of the species was derived.

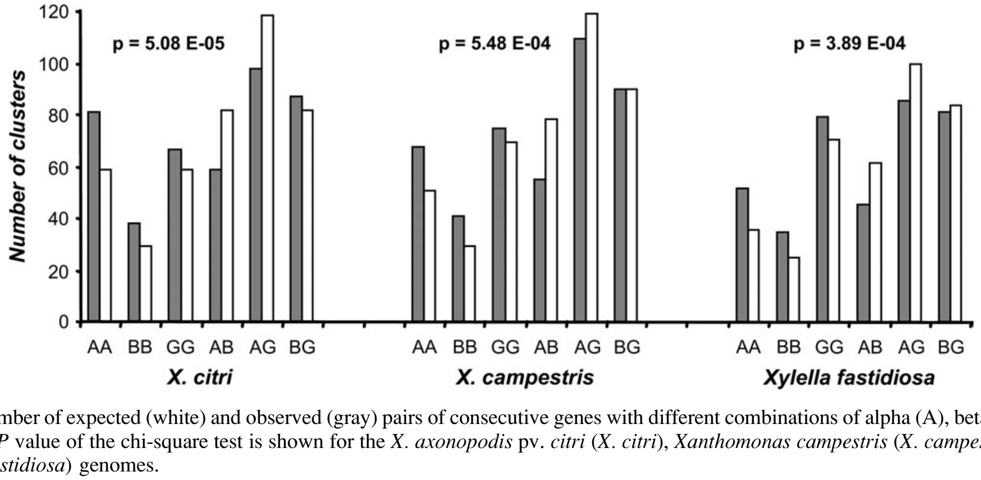

FIG. 6.—Number of expected (white) and observed (gray) pairs of consecutive genes with different combinations of alpha (A), beta (B), and gamma (G) origins. The *P* value of the chi-square test is shown for the *X. axonopodis* pv. *citri* (*X. citri*), *Xanthomonas campestris* (*X. campestris*), and *Xylella fastidiosa* (*Xy. fastidiosa*) genomes.

In this analysis, the mean number of steps for reconciling the trees was surprisingly low for a phylogeny in which all the taxonomic groups of Bacteria and some Archaea were represented. From a biological point of view, these translate into preferential sharing of genes among bacterial species from the same group or from close divisions such as among Proteobacteria clades. These preferences for gene sharing also reveal the presence of limitations to random transfers.

These results suggest that there are constraints in the genomic architecture that make gene acquisition from distantly related taxa unlikely. A possible mechanism for such constraints has been reported (Lawrence and Hendrickson 2004; Hendrickson and Lawrence 2006). These authors analyzed the distribution of octomers along the bacterial chromosome. Some octomers are preferentially found on leading strands and increase in abundance toward the replication terminus; here, selection would be maximal for their role in efficient chromosome segregation (Hendrickson and Lawrence 2006). This selection at the level of chromosome structure could have important implications for bacterial chromosome dynamics. For instance, DNA compatible with the octomer distribution of a recipient genome would have a higher chance of being transferred successfully. This implies that successful transfers will be more likely between closely related and, therefore, compatible genomes, although they are most difficult to identify with current methodologies (Lawrence and Hendrickson 2003, 2004).

Our results are compatible with the predictions of Gogarten and coworkers (2002) as reflected in figure 7. We have found HGT events to the ancestor of Xanthomonadales, therefore prior to their diversification and, most likely, to the diversification of nascent Proteobacteria lineages. The low number of atypical genes detected among them also reveals the old age of these transfers. Ongoing, or recent, transfers to Xanthomonadales genomes are likely, as reflected by the proportion of atypical genes in the *X. citri* genome but are not those detected by our phylogenetic origin test because a much lower proportion of atypical genes was identified among the genes causing conflicting phylogenetic signal in this genome. Because these recent transfers do not result in an incongruent position of the involved Xanthomonadales taxa out from this group, it seems reasonable to assume that most of these recent transfers have occurred among members of the Xanthomonadales clade and not with other Proteobacteria or more external groups. Therefore, the age of transfers reveals the more likely partner(s) throughout the evolutionary history of the group: other Proteobacteria lineages in the past, when they had not diverged much yet, or other Xanthomonadales lineages in recent times, when the divergence between Xanthomonadales and other Proteobacteria groups is significant but not so within the Xanthomonadales group.

Additionally, this result reveals the different effects of HGT events on phylogenetic reconstructions depending on the time since the transfers. Those methods based on gene trees, such as supertrees and consensus (fig. 3), are appropriate to pinpoint (in the form of unresolved nodes) high amounts of transfers in the distant past. The higher the number of past transfer events on the ancestor of a clade, the higher the likelihood of retrieving its corresponding branch as an unresolved position in bacterial genome phylogenies. On the other hand, all gene trees derived from the 207 common genes' set support the monophyly of the Xanthomonadales clade. This strong signal may be powered not only by the vertical transmission of its members but also by current HGTs inside the group (Tettelin et al. 2005). Recombination between closely related strains and HGT between close, intragenera species could result in the observation of a clear, vertical signal in genome phylogenies because there is not enough divergence to be detected at a genome-scale analysis.

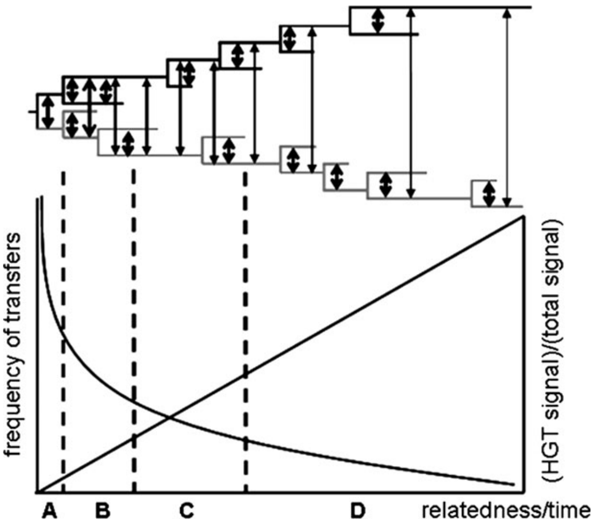

FIG. 7.—The figure reflects changes in the frequency of HGT events (curve line) and amount of accumulated HGT signal along time (straight line) in Xanthomonadales evolution. The frequency of transfers is also represented by the thickness of the arrows inside the genealogy. The figure can be interpreted in 2 alternative ways. Firstly, the *x* axis could reflect the evolutionary relatedness of extant lineages. Then, regions A and B reflect exchanges within the Xanthomonadales clade, and therefore, between very close taxa. These transfers correspond to recombination within populations (region A) and to relatively recent transfers detected in our atypical gene analysis (region B). Regions C and D are those corresponding to current transfers from more distant groups, for example, from other Proteobacteria (region C) or even more distant taxa (non-Proteobacteria and Archaea) (region D). Secondly, the *x* axis could be interpreted as time, therefore, reflecting the evolutionary history of a lineage from its origin to the present. In this case, the gray lineage represents the genealogy of current Xanthomonadales and the black lineage the genealogy of a hypothetical proteobacterial lineage. In region A, because these lineages had not diverged much, the frequency of transfers between them was very high. This and region B reflect ancient transfers to the Xanthomonadales ancestor as the ones detected in our test of phylogenetic origin. As nascent Proteobacteria lineages started to diverge, the frequency of exchange between Xanthomonadales and other Proteobacteria began to decrease (region C). Finally, the number of transfer events between Proteobacteria and Xanthomonadales has disappeared almost completely and only within group events, such as those revealed by our atypical gene analyses and the consistent monophyly of the Xanthomonadales clade, remain (region D).

To sum up, 2 conclusions have been outlined. On one hand, Xanthomonadales genomes have approximately the same number of genes with beta-, gamma-, or alpha-proteobacterial affinity, making them an extreme case of mosaicism and preventing us from conclusively assigning them to one of the major proteobacterial clades. On the other hand, we have shown that it is possible to disentangle noise from signal through exhaustive and careful analysis even in the most complex cases like Xanthomonadales evolution. We have shown the existence of ancient and recent transfers despite possible phylogenetic artifacts. The effect of these transfers in phylogenomics and the resolution of ancestral nodes will depend on the vertical/horizontal signal ratio in the branches leading to a node. These ratios will determine which parts of the genome trees are treelike and which are not. Obviously, Xanthomonadales seem to fit this model because they appear as a monophyletic group recovered in almost all gene phylogenies. Meanwhile, their correct position in the Proteobacteria tree is obscured by the presence of a high amount of ancient transfer events to their common ancestor as shown in this analysis. Other groups such as Pseudomonadales and some Cyanobacteria also seem to present high ancient HGT rates (Beiko et al. 2005) and may follow the same pattern. Alternatively, the evolutionary scenario of resolved tips and poorly resolved deep nodes should not apply to genomes with less promiscuity or susceptibility to HGT in ancient times, in which case their vertical phylogenetic signal in the past had more weight than the horizontal one, thus allowing a better resolution of their deep phylogenetic relationships.

Supplementary Material

Supplementary table S1 and figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work has been supported by grants BMC2003-00305 from Spanish Ministerio de Ciencia y Tecnología and Grupos 03/2004 from Generalitat Valenciana to A.M. I.C. is recipient of a University Personnel Formation Fellowship Predoctoral Fellowship from the Spanish Ministerio de Educación y Ciencia.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–402.

Azad RK, Lawrence JG. 2005. Use of artificial genomes in assessing methods for atypical gene detection. PLoS Comput Biol 1:e56.

Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF. 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evol Biol 5:33.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci USA 102:14332–7.

Bern M, Goldberg D. 2005. Automatic selection of representative proteins for bacterial phylogeny. BMC Evol Biol 5:34.

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau MER, Nesbo CL, Case RJ, Doolittle WF. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet 37:283–328.

Charlebois RL, Clarke GDP, Beiko RG, Jean A. 2003. Characterization of species-specific genes using a flexible, web-based querying system. FEMS Microbiol Lett 225:213–20.

Creevey CJ, Fitzpatrick DA, Philip GK, Kinsella RJ, O'Connell MJ, Pentony MM, Travers SA, Wilkinson M, McInerney JO. 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proc R Soc Lond B Biol Sci 271:2551–8.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–403.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science 284:2124–8.

Dutilh BE, Huynen MA, Bruno WJ, Snel B. 2005. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J Mol Evol 58:527–39.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19: 2226–38.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704.

Hauck B, Gehring WJ, Walldorf U. 1999. Functional analysis of an eye specific enhancer of the eyeless gene in Drosophila. Proc Natl Acad Sci USA 96:564–9.

Hendrickson H, Lawrence JG. 2006. Selection for chromosome architecture in bacteria. J Mol Evol 62:615–29.

Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 96:3801–6.

Jones DT, Taylor WR, Thornton JM. 1994. A mutation data matrix for transmembrane proteins. FEBS Lett 339:269–75.

Koonin EV, Galperin MY. 1997. Prokaryotic genomes: the emerging paradigm of genome-based microbiology. Curr Opin Genet Dev 7:757–63.

Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709–42.

Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. 2005. The net of life: reconstructing the microbial phylogenetic network. Genome Res 15:954–9.

Lawrence JG, Hendrickson H. 2003. Lateral gene transfer: when will adolescence end? Mol Microbiol 50:739–49.

Lawrence JG, Hendrickson H. 2004. Chromosome structure and constraints on lateral gene transfer. Dynamical Genet 2004:319–36.

Lawrence JG, Ochman H. 2002. Reconciling the many faces of lateral gene transfer. Trends Microbiol 10:1–4.

Lawrence JG, Roth JR. 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143: 1843–60.

Martins-Pinheiro M, Galhardo RS, Aires KA, Lima-Bessa KM, Menck CFM. 2004. Different patterns of evolution for duplicated DNA repair genes in bacteria of the Xanthomonadales group. BMC Evol Biol 4:29.

Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet 36:760–6.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.

Omelchenko M, Makarova K, Wolf Y, Rogozin I, Koonin E. 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biol 4:R55.

Pal C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37:1372–5.

Robinson-Rechavi M, Huchon D. 2000. RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. Bioinformatics 16:296–7.

Snel B, Huynen MA, Dutilh BE. 2005. Genome trees and the nature of genome evolution. Annu Rev Microbiol 59: 191–209.

Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc R Soc Lond B Biol Sci 269:137–42.

Studholme DJ, Downie JA, Preston GM. 2005. Protein domains and architectural innovation in plant-associated Proteobacteria. BMC Genomics 6:17.

Susko E, Leigh J, Doolittle WF, Bapteste E. 2006. Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the γ-Proteobacteria. Mol Biol Evol 23:1019–30.

Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SGE. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science 296:2376–9.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–6.

Tettelin H, Masignani V, Cieslewicz MJ, et al. (36 co-authors). 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." Proc Natl Acad Sci USA 102:13950–5.

Van Sluys MA, Monteiro-Vitorello CB, Camargo LEA, Menck CFM, da Silva ACR, Ferro JA, Oliveira MC, Setubal JC, Kitajima JP, Simpson AJ. 2002. Comparative genomic analysis of plant-associated bacteria. Annu Rev Phytopathol 40: 169–89.

William Martin, Associate Editor

Accepted July 25, 2006