Parasitism and Mutualism in Wolbachia: What the Phylogenomic Trees Can and Cannot Say

Seth R. Bordenstein,*†1 Charalampos Paraskevopoulos,‡1 Julie C. Dunning Hotopp,§1 Panagiotis Sapountzis,‡† Nathan Lo,‡ Claudio Bandi,‡ Hervé Tettelin,§§ John H. Werren,** and Kostas Bourtzis‡‡

*Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA; †Department of Biological Sciences, Vanderbilt University; ‡Department of Environmental and Natural Resources Management, University of Ioannina, Agrinio, Greece; §J. Craig Venter Institute, Rockville, MD; §§Institute for Genome Sciences, University of Maryland School of Medicine; ‡‡School of Biological Sciences, The University of Sydney, New South Wales, Australia; #Dipartimento di Patologia Animale, Igiene e Sanità Pubblica Veterinaria, Università di Milano, Milano, Italy; and **Department of Biology, University of Rochester, Rochester, NY

Ecological and evolutionary theories predict that parasitism and mutualism are not fixed endpoints of the symbiotic spectrum. Rather, parasitism and mutualism may be host or environment dependent, induced by the same genetic machinery, and shifted due to selection. These models presume the existence of genetic or environmental variation that can spur incipient changes in symbiotic lifestyle. However, for obligate intracellular bacteria whose genomes are highly reduced, studies specify that discrete symbiotic associations can be evolutionarily stable for hundreds of millions of years. Wolbachia is an inherited obligate, intracellular infection of invertebrates containing taxa that act broadly as both parasites in arthropods and mutualists in certain roundworms. Here, we analyze the ancestry of mutualism and parasitism in Wolbachia and the evolutionary trajectory of this variation in symbiotic lifestyle with a comprehensive, phylogenomic analysis. Contrary to previous claims, we show unequivocally that the transition in lifestyle cannot be reconstructed with current methods due to long-branch attraction (LBA) artifacts of the distant Anaplasma and Ehrlichia outgroups. Despite the use of 1) site-heterogenous phylogenomic methods that can overcome systematic error, 2) a taxonomically rich set of taxa, and 3) statistical assessments of the genes, tree topologies, and models of evolution, we conclude that the LBA artifact is serious enough to afflicct past and recent claims including the root lies in the middle of the Wolbachia mutualists and parasites. We show that different inference methods yield different results and high bootstrap support did not equal phylogenetic accuracy. Recombination was rare among this taxonomically diverse data set, indicating that elevated levels of recombination in Wolbachia are restricted to specific coinfecting groups. In conclusion, we attribute the inability to root the tree to rate heterogeneity between the ingroup and outgroup. Site-heterogenous models of evolution did improve the placement of aberrant taxa in the ingroup phylogeny. Finally, in the unrooted topology, the distribution of parasitism and mutualism across the tree suggests that at least two interphylum transfers shaped the origins of nematode mutualism and arthropod parasitism. We suggest that the ancestry of mutualism and parasitism is not resolvable without more suitable outgroups or complete genome sequences from all Wolbachia supergroups.

Introduction

Anton de Bary coined the term “symbiosis” in his 1879 monograph as the living together of dissimilar organisms (de Bary 1879). By not specifically referring to beneficial or parasitic associations, this definition set the stage for understanding species interactions without constraint. Symbiosis as defined today encompasses all forms of species interactions because there are no universal principles that differentiate the mechanisms of mutualism and parasitism (Hentschel et al. 2000). Symbionts and hosts experience frequent transitions between different lifestyles, and mutualists and parasites can use similar genetic machinery for mediating parasitism and mutualism (Gargas et al. 1995; Ewald 2004; Sachs and Simms 2006).

One possible exception to this plastic view of symbiosis occurs in obligate intracellular (or endosymbiotic) bacteria that exclusively replicate inside host cells. These bacteria occur in diverse eukaryotic hosts and form parasitic and mutualistic interactions that can be evolutionarily stable for hundreds of millions of years. A prevailing view is that such endosymbiotic lifestyles become irreversible as the bacteria streamline their gene content, thereby limiting the evolutionary potential for encoding alternative lifestyles (Moran and Wernegreen 2000).

Here, we present an evolutionary analysis of mutualism and parasitism in Wolbachia piipiensis endosymbionts, widespread intracellular bacteria of arthropods, and filarial nematodes. Wolbachia evolved from a ~400-Mo-old clade of gram-negative, aerobic, α proteobacteria that encompass obligatory intracellular, vertebrate pathogens and arthropod infections of the genera Rickettsia, Ehrlichia, Anaplasma, Orientia, Neorickettsia, and Midichloria. Despite the clade’s ancient intracellular association, small genome size (0.9–1.6 Mb) and dependence on intracellar replication, the genus Wolbachia evolved labile lifestyles, primarily as reproductive parasites in arthropods and mutualists in filarial nematodes.

In arthropods, the reproductive parasites distort sex ratios and sexual reproduction strategies to gain a maternal transmission advantage (Werren 1997; Stouthamer et al. 1999). These sexual alterations include parthenogenesis, feminization, male killing, and cytoplasmic incompatibility, some of which are implicated in driving the evolution of new mechanisms of host sex determination (Rousset et al. 1992; Normark 2003; Negri et al. 2006), alternative modes of sexual selection (Jiggins et al. 2000), and incipient species (Bordenstein et al. 2001; Jang et al. 2006; Koukou et al. 2006). In rare cases, arthropod hosts have evolved...
codependencies with reproductive parasites to the point where the Wolbachia are essential to host fertility (Starr and Cline 2002; Pannebakker et al. 2007). In contrast to the arthropods, antibiotic curing experiments suggest that in nematodes, Wolbachia infections are primarily beneficial to nematode fertility and larval development (Taylor et al. 2005). Further, the Wolbachia genome sequence from the filarial Brugia malayi suggests that these mutualists contribute essential compounds such as nucleotides, heme, and riboflavin to the host nematodes (Foster et al. 2005).

The major lifestyle differences in Wolbachia notably associate with discrete phylogenetic supergroups that differ at bacterial protein-coding genes and typically adhere to the criteria of greater than 3% divergence at the 16S rDNA gene (Lo et al. 2007). Thus, these lifestyle transitions within the Wolbachia-invertebrate endosymbiosis have occurred in a time frame potentially amenable to assessing the evolutionary trajectory of mutualism and parasitism through molecular phylogenetic analysis. The distinct lineages include the arthropod reproductive parasites in supergroups A and B and the nematode mutualists of supergroups C and D. Other diagnostic differences include a 200-kb smaller genome (Foster et al. 2005), complete vertical transmission (Casiraghi et al. 2001), and no recombination in the nematode C and D Wolbachia, whereas the arthropod A and B Wolbachia contain higher fractions of mobile DNA (Wu et al. 2005), the filarial nematode 

E. malayi (Wu et al. 2004) (AE017196) and 

B. malayi (Foster et al. 2005) (AE017321), Ehrlichia chaffeensis Arkansas (Hotopp et al. 2006) (CP000236), Ehrlichia ruminantium Welgevonden (Collins et al. 2005) (CR767821), Anaplasma phagocytophilum HZ (Hotopp et al. 2006) (CP000235), and Anaplasma marginale St Maryes (Brayton et al. 2005) (CP000030), and 3) direct sequencing of polymerase chain reaction (PCR) products for which methods have been previously published (Baldo et al. 2006). All new sequences were deposited in GenBank under accession numbers: FJ390143–FJ390370. The average length of each trimmed sequence used in the alignments was 542 bp, and the average taxon representation per ortholog was 2.54%.

Materials and Methods
Gene and Taxon Sampling

Primary sequences were obtained from three sources: 1) published gene sequences (Baldo et al. 2006; Paraskevopoulos et al. 2006), 2) genome sequences of Wolbachia strains wMel (Wu et al. 2004) (AE017196) and wBm (Foster et al. 2005) (AE017321), Ehrlichia chaffeensis Arkansas (Hotopp et al. 2006) (CP000236), Ehrlichia ruminantium Welgevonden (Collins et al. 2005) (CR767821), Anaplasma phagocytophilum HZ (Hotopp et al. 2006) (CP000235), and Anaplasma marginale St Maryes (Brayton et al. 2005) (CP000030), and 3) direct sequencing of polymerase chain reaction (PCR) products for which methods have been previously published (Baldo et al. 2006). All new sequences were deposited in GenBank under accession numbers: FJ390143–FJ390370. The average length of each trimmed sequence used in the alignments was 542 bp, and the average taxon representation per ortholog was 15 of 18 Wolbachia taxa. The 21 loci include WD_0001 (dnaA), WD_0160 (ruoG), WD_0183, WD_0198, WD_0203 (atpD), WD_0237, WD_0301 (cosA), WD_0307 (groEL), WD_0359, WD_0413 (aspS),
WD_0473 (pdhB), WD_0484 (hcpA), WD_0544 (sucB), WD_0560 (nuoD), WD_0723 (ftsZ), WD_0791, WD_0976, WD_1005, WD_1029 (aspC), WD_1151 (gltA), and WD_1170 (fabK). A second data set was derived from previous analyses (Fenn and Blaxter 2006).

**Sequence Nomenclature and Alignments**

*Wolbachia* are named according to their host species and are labeled with a capital letter denoting the supergroup association. Primary and translated amino acid sequences were aligned in ClustalX and edited manually in MacClade 4.08. All indels and hypervariable regions were removed for a final alignment length of 11.9 kb per taxon.

**Phylogenomic Analyses**

PhyloBayes, MrBayes, and ML methods were used to infer phylogenetic relationships. PhyloBayes 2.1c and 2.3 analyses were performed with the Jones-Taylor-Thorton (JTT), Whelan and Goldman (WAG), general time reversible (GTR) models, and the category amino acid site-heterogeneous mixture model to suppress tree artifacts associated with LBA (Lartillot and Philippe 2004; Lartillot et al. 2007). For all PhyloBayes analyses, at least two independent runs were performed with default PhyloBayes conditions until a maxdiff value < 0.15 was achieved to ensure chain equilibration. The first 100 points were discarded as burn-in, and the posterior consensus was computed on the remaining trees. Data sets were also analyzed with WAG amino acid model in MrBayes. Runs were one million generations with tree sampling every 100 generations, and a consensus tree was built on the last 7,000 trees. Posterior probabilities were determined by constructing a 50% majority rule tree of all the sampled trees. Finally, concatenated nucleotide sequences were analyzed with ML and MrBayes. Trees were reconstructed based on all nucleotides and only the first two codon positions to examine potential artifacts associated with third-codon positions. Prior to ML analyses, a DNA substitution model for each data set was selected using Modeltest v3.06 and the Akaike information criterion. The models selected for the first two codon positions and all nucleotides were the transversional model with a gamma distribution + invariant sites (TVM + I + G) and the GTR model with a gamma distribution + invariant sites (GTR + I + G), respectively. ML heuristic searches were performed using 100 random taxon addition replicates with tree bisection and reconnection branch swapping. ML bootstrap support was determined using 100 bootstrap replicates, each using 10 random taxon addition replicates. Searches were performed in parallel on a Beowulf cluster using a clusterpaup program and PAUP version 4.0b10. Bayesian nucleotide analyses were performed with the same running conditions as the protein analyses.

**Statistical Assessment of the Root**

We tested the significance of topological differences in phylogenetic trees with different rooting positions using the Shimodaira–Hasegawa (SH) test (Shimodaira and Hasegawa 1999). The SH test compares the likelihood score (− ln L) of a given data set across its ML tree with the − ln L of that data set across alternative topologies, which in this case are the ML phylogenies for other data sets. The differences in the − ln L values are evaluated.
for statistical significance using bootstrap (1,000 replicates) based on RELL sampling and the more extensive full optimization (PAUP version 4.0b10). These two approaches yielded similar results.

Posterior Predictive Analyses

We used the predefined, posterior predictive tests, -div and -comp, as implemented in PhyloBayes. These statistics measure the site-specific amino acid diversity and the compositional heterogeneity among taxa, respectively. In the latter case, a compositional $Z$ score is produced along with a $P$ value to determine significance. All genes with taxon showing a $Z$ score greater than 2 always had $P$ values less than 0.05. In these cases, the genes were excluded to make a new data set when noted.

Detection of Recombination ($\Phi_W$)

Alignments of individual and concatenated genes both with and without outgroups were screened for significant levels of recombination using the pairwise homoplasy index (PHI) or PHItest (Bruen et al. 2006) implemented in SplitsTree4 under default conditions (Huson and Bryant 2006). As a positive control, an alignment from the highly recombinating Wolbachia surface protein (wsp) gene was included in this analysis. In comparison to other tests, Max $\chi^2$ and neighbor similarity score (NSS), the PHItest reduces the frequency of false positives when confounding processes such as substitution rate heterogeneity affect the data set, especially under conditions of large sample sizes. It has been applied to varied data sets including viruses, mitochondria, fungi, and Wolbachia endosymbionts (Bruen et al. 2006).

Results and Discussion

Rooting Position Using Standard Reconstruction Methods

We obtained primary sequences directly from the major Wolbachia supergroups A–F, H, and several divergent lineages from PCR products or public databases. These taxa cover the major supergroups/lifestyles and, when possible, two to three lineages within each of those supergroups. The final data set included 18 Wolbachia ingroup taxa and 4 outgroup species. The supermatrix consisted of a total of 21 protein-coding genes and up to 261 kb of sequence with indels or hypervariable regions removed to reduce nonphylogenetic signal in the data.

MP (DNA and Amino Acid)

Figure 1A shows the MP reconstruction of the rooted Wolbachia tree, based on the final concatenated matrix of 11,919 nt. The root is positioned at the base of supergroup E from the host springtail Folsomia candida. The symbiotic function of this lineage has not yet been characterized. The next earliest branching lineage in this tree is the mutualistic Wolbachia from the filarial nematodes B. malayi and Litotyphlus sigmodontis. Although there is 100% bootstrap support for the root’s placement, there is reason to doubt the supergroup E rooting in the MP analysis. MP inference methods as mentioned above can be inferior to probabilistic models such as ML. In particular, the susceptibility of MP to LBA artifacts can artificially lead to the placement of the divergent outgroup branch next to the longest branch in the unrooted, ingroup tree. Consistent with this artifact, the longest ingroup branch of the Wolbachia tree is supergroup E from F. candida (see fig. 1). It is the single, divergent representative of supergroup E, and it has a different positioning in the unrooted MP phylogeny. The different evolutionary relationships in the rooted and unrooted MP
trees are a further indicator of LBA artifacts. Finally, a separate MP analysis of the translated amino acid sequences also places the root in supergroup E with weak posterior probability support of 67% and a highly unresolved tree (data not shown).

Bayes and ML (DNA and Amino Acid)

ML and Bayesian inference methods are generally preferred over MP for analyzing real sequence data because they incorporate models of evolution and are less sensitive to LBA artifacts (Felsenstein 1981). Figure 1B shows the Bayesian reconstruction of the rooted Wolbachia tree based on the 3,972 amino acid translation using the WAG model of protein sequence evolution. The most striking difference between this Bayesian tree and the MP tree is that the rooting position changed from the base of supergroup E to supergroup B—the reproductive parasites of arthropods (100% posterior probability support). The JTT and GTR models of protein-sequence evolution also yielded the identical topology with support for the root of supergroup B at 92% and 98%, respectively. Finally, the topology from the Bayesian amino acid analysis parallels the ML topology based on the entire nucleotide data set, which yielded 77% bootstrap support for the root at the base of the parasitic supergroup B.

When taken together, these analyses support reproductive parasitism as the ancestral association of the Wolbachia symbiosis. Although the probabilistic inferences models appear to be more robust than MP against LBA artifacts and it would be tempting to conclude that supergroup B is the root, a deeper assessment suggests this conclusion would be erroneous.

We also note caution in interpreting the unusual placement of the D. gracile nematode and C. scorpioi diodes arthropod branches. Dipetalonema gracile is surprisingly positioned as a sister taxa to the arthropod supergroup A, and similarly C. scorpioi diodes is positioned next to the filarial nematode supergroup C. These phylogenetic relationships between arthropod and nematode Wolbachia would suggest recurrent instances of interphylum host transfers of Wolbachia. In fact, some of these peculiar evolutionary relationships are contradicted by analyses under more complex inference models used below. However, to ensure that these patterns could not be explained by template contamination between arthropod and nematode DNAs, diagnostic PCRs using conserved rDNA primers confirmed the absence of contaminating arthropod DNA in the D. gracile template and the absence of nematode DNA in the C. scorpioi diodes template.

Assessment of the Support for the Supergroup B Root
Codon Position

The ML bootstrap support for the rooting in the reproductive parasites is 77% when the nucleotide data set is analyzed with all three codon positions but 53% when the first and second codon positions are analyzed. This reduction is interesting because the removal of rapidly evolving, third-codon positions could potentially reduce the nonphylogenetic signal in the analysis and thereby increase support for the root. Alternatively, the reduction could be due to a decrease in the number of base pairs analyzed from 11,919 to 7,946 bp.

Single-Gene Trees

Of the 21 single-gene nucleotide alignments analyzed by ML with all three codon positions, only 6 produced trees with >50% bootstrap support for any rooting position. Thus, there is an overall lack of support concerning the rooting when each gene is taken separately. Of the six with >50% bootstrap support, only three placed the root at the base of supergroup B (67%, 58%, and 57% support), and the other three showed similarly weak support for alternative rootings, including the bases of supergroup C (53%), supergroup E (56%), and lineage C. scorpioi diodes (52%). Finally, of the remaining 15 gene trees with less than 50% bootstrap support for the root, only 4 of these placed the root basally in supergroup B parasites. These variable results indicate a significant nonphylogenetic signal among the individual genes that could be reinforced as systematic error when using concatenated data sets.

Likelihood-Based Test of Topologies

If supergroup B is the earliest branching lineage of Wolbachia, then we predict that there will be significant topological differences between this rooting position and topologies with alternative rooting positions. In contrast, if the supergroup B rooting is not strongly supported, then we predict that there will be no topological differences to trees with different outgroup positions. Using an SH test (Shimodaira and Hasegawa 1999), we compared the relative support of the concatenated ML best tree (i.e., root in supergroup B, fig. 1B) with itself and 11 other topologies in which the root of the tree (leading to the four outgroups) was placed as a sister group to each of the different supergroups as well as to Wolbachia spp. from D. gracile, C. canis, and C. scorpioi diodes. All 12 topologies were statically indistinguishable after a Bonferroni correction (fig. 2, corrected P value = 0.004); 8/12 were statistically indistinguishable without a Bonferroni correction, with no overall clustering of significant P values and taxonomic supergroups. Therefore, even with a framework of a taxonomically rich, multigenic data set and likelihood-based inference models, reproductive parasitism in supergroup B is not significantly more likely to be the root (fig. 1B) than other competing positions. Therefore, we are still left with the question of whether the analyses can accurately resolve the root of the Wolbachia clade using the closest known outgroups.

An Unresolved and Unresolvable Rooting Position

Multigenic analyses are predicated on the assumption that the concatenation of many genes can lead to an improved species tree. Systematic artifacts are the biggest potential problems associated with large genetic data sets owing to model violations and/or poor gene and taxon
sampling. For instance, the use of site-homogenous substitution models under an ML and Bayesian framework may not be resourceful enough to overcome the LBA phenomena associated with the Wolbachia tree. Further, extensive levels of gene exchange could also lead to artifacts and an unresolved species tree (Schierup and Hein 2000; Posada and Crandall 2002). To clarify the roles of recombination and rate heterogeneity on the rooting position, we took two approaches.

Recombination

First, we used the program PHItest (Bruen et al. 2006) to determine the presence or absence of recombination within each of the nucleotide alignments of the 21 gene data set. An evaluation of the performance of several recombination programs using both simulated and empirical data found that PHItest effectively determines recombination under diverse conditions and performs markedly better than Max\textsuperscript{v}\textsubscript{2} and NSS at avoiding false positives of recombination under models of substitution rate heterogeneity (Bruen et al. 2006). As the different Wolbachia and outgroup lineages can vary in their branch lengths, it is especially important to control for artifacts associated with rate heterogeneity. Recombination within and between the A and B Wolbachia supergroups has been reported previously using Max\textsuperscript{v}\textsubscript{2}, and the A and B Wolbachia frequently coinfect the same arthropod hosts (Baldo et al. 2006). However, recombination and superinfection appear to be absent or much less frequent in other Wolbachia supergroups (Jiggins 2002; Bordenstein and Wernegreen 2004).

PHItest inferred recombination in only 1 of the 21 genes following a Bonferroni correction (WD0359, helicase, $P = 0.027$). As a positive control, we included an alignment of the highly recombinating wsp gene from Wolbachia and found a highly significant level of recombination (WD1063, $P < 0.0001$), thus confirming the usefulness of the program. We conclude that recombination does not significantly affect our data set and that the genes can be concatenated for phylogenomic analysis.

As a positive control, we included an alignment of the highly recombinating wsp gene from Wolbachia and found a highly significant level of recombination (WD1063, $P < 0.0001$), thus confirming the usefulness of the program. We conclude that recombination does not significantly affect our data set and that the genes can be concatenated for phylogenomic analysis.

At least two reasons explain why recombination is absent in the same genes previously reported to recombine: 1) taxonomic sampling from Wolbachia supergroups other than A and B, many of which are not known to undergo recombination and 2) a test statistic, $\Phi_W$, that reduces false-positive detections of recombination. Using PHItest, we inferred significant recombination in the same two data sets previously used to infer recombination with Max\textsuperscript{v}\textsubscript{2}, gltA, and groEL (Baldo et al. 2006). This result indicates that diverse taxon sampling schemes or the use of different taxa from the same supergroups can mitigate the signal of recombination in certain portions of the Wolbachia phylogeny.

PhyloBayes CAT

Second, we used the CAT inference model as implemented in the PhyloBayes software package to attempt to overcome at least some of the risks of systematic errors in our analyses (Lartillot and Philippe 2004). Briefly, in contrast to models using a single empirical substitution matrix, CAT incorporates site-specific features of protein evolution by using a mixture model in which the amino acid replacement pattern at each position of the alignment can be described by a different substitution process and thus a unique subset of the 20 amino acids. Recent studies
emphasize that accounting for across-site heterogeneities in
the amino acid replacement process is crucial for a better
model fit and importantly can alleviate LBA (Lartillot
et al. 2007; Rodriguez-Ezpeleta et al. 2007). Thus, the
CAT model is an ideal method to assess the robustness of the
Wolbachia phylogeny, including those trees from
previous analyses claiming to have positioned the root between
the mutualists and parasites (Fenn et al. 2006).

The CAT model was applied to the amino acid align-
ment previously studied in figure 1B using WAG, JTT, and
GTR site-homogenous models. A less resolved tree was re-
covered that included an exceedingly long branch to the
outgroup, which in turn did not resolve the evolutionary
relationships of the base of the Wolbachia ingroup
(fig. 3A). Although the tips of the ingroup tree were well
resolved with high posterior probability support, the unre-
solved base of the ingroup seems to reflect the model’s am-
biguity in positioning the root. This conclusion is supported
by the observation that the uncertainty in the tree disappears
when the CAT model was applied to the same data set with-
out the divergent outgroups. Posterior probabilities for each
node of the unrooted tree then ranged confidently between
93% and 100%, with the exception of the branch from su-
pergroup C leading to C. scorpionides (posterior probability
is 57%, data not shown). This branch is also weakly sup-
ported in the rooted tree of figure 3A.

Analysis of Model Violations

The CAT model was confirmed to be a better statistical
fit to the data set than the other site-homogenous models by
examining two statistics that account for systematic arti-
facts, the mean number of substitutions per site (diversity)
and the level of convergences and reversions (homoplas). Diveristy accounts for how well a model predicts the site-
specific biochemical pattern of the data set. For our data set,
the mean number of substitutions per site was 2.06. Under
WAG, JTT, and GTR models, however, the posterior pre-
predicted number of substitutions per site was higher at 2.18,
2.19, and 2.16, respectively. In contrast, under CAT, the
value was 2.05, a much closer prediction to the observed
value. If the posterior predictive values are significantly dif-
ferent from the observed ones from the data, then the model
of amino acid evolution underestimates diversity and can be
more prone to false reconstructions. Because CAT and
other models such as the site-homogenous WAG model
principally differ in their amino acid replacement process,
differences in the posterior predictive analyses indicate
whether the amino acid replacement process directly alters
the tree reconstructions. As expected, there were univer-
sal significant differences between the observed and pre-
predicted diversity values for WAG, JTT, and GTR (P <
0.0001), whereas those for CAT were not significantly dif-
ferent (P = 0.59). Further, the posterior predicted mean
number of homoplasies per site for CAT, 2.01, does not deviate from the observed number, 1.98. In contrast, the
other three models underestimate the level of saturation
as measured by homoplasy. For example, WAG has a much
lower posterior number, 0.80, and a more deviating ob-
served value of 0.95. These findings indicate that the
CAT site-heterogenous model correctly accommodates the saturation patterns of the amino acid sequences. Con-
sequently, it is more likely to offer accurate phylogenetic
predictions, especially in the case where the root is not re-
solvable with current data (fig. 3A).

Data Exclusion to Reduce Systematic Error

An additional approach to reduce ambiguity and
LBA artifacts in the tree reconstruction is to remove the
sequences that most clearly violate some hypotheses of
the model. This can have the effect of increasing the ratio
of phylogenetic to nonphylogenetic signal in the data set.
Such an approach is complementary to using the more so-
plicated inference models of sequence evolution. Be-
cause fast-evolving genes by definition are more likely
to accumulate multiple substitutions, we used the Phylo-
Bayes software package to identify and remove the six
genes with taxa that show significant compositional devia-
tion (Z score > 2.0, P value < 0.05) and are excluded in downstream analyses. Labels in each
plot correspond to the respective gene labels used in each data set.

FIG. 4.—Summary plot of Z scores for each gene’s amino acid
alignment in ascending order for (A) the current data set and (B) a previously
published data set (Fenn et al. 2006). Protein alignments with Z scores
greater than 2 have taxa that show significant compositional deviation
(P value < 0.05) and are excluded in downstream analyses. Labels in each
plot correspond to the respective gene labels used in each data set.

Wageningen Academic Publishers, 2007
mutualistic supergroup D Wolbachia to the later originating reproductive parasites, the root in the mutualists remains weakly supported at 57% posterior probability support, and the reconstruction cannot be interpreted as a more accurate tree. In contrast, under WAG, JTT, and GTR, the reconstructions yield strong support (99–100%) for the same rooting placement as CAT; this marked difference in support is a good example of when site-homogenous models can overpredict support for these basal relationships. WAG, JTT, and GTR models are more sensitive to saturation and attraction artifacts, further indicating their lack of reliability for which Wolbachia lifestyle is the earliest originating.

Despite the ambiguity in the rooting, the increase in the phylogenetic signal through data exclusion notably improved the ingroup tree by placing the D. gracile nematode branch from a sister taxa to the arthropod supergroup A (fig. 1B) to a sister taxa to the nematode supergroup C (fig. 3B). The same change is also evident in the unrooted CAT tree (fig. 5). Knowledge about the relationships of diverse Wolbachia strains is desirable to guide future nomenclature decisions and to provide clues to the evolutionary origins of Wolbachia functions. Based on the unrooted topology in figure 5, we can hypothesize that the nematode mutualistic Wolbachia either arose once in the ancestor of all supergroups containing nematode hosts [C, D, F, D. gracile] or arose twice by evolving independently in the ancestors of [D, F] and [C, D. gracile], which are separated by a strain from the arthropod cat flea. We can also hypothesize that arthropod reproductive parasitism arose once at the ancestor of the A and B supergroups. Because supergroups E and H cluster in this major group of the tree [A, B, E, H], we can further hypothesize that these two taxa are or were ancestrally reproductive parasites. Based on this analysis and excluding the C. scorpioides lineage due to poor support in the tree, we conclude that there were at least two interphylum transfers of Wolbachia between nematodes and arthropods to explain the radiation of this endosymbiosis.

Reconsideration of the Previously Published Root

The results above demonstrate that caution is warranted concerning the placement of the root, even under cases of strong root support. A recent study using 42 protein-coding genes derived from five Wolbachia genomes and three outgroups (Fenn et al. 2006) gave a ML phylogeny where the root was placed between the reproductive parasites (supergroup A) and the mutualists (supergroups C and D) with 100% posterior probability support. However, the data set was noted to be missing the other taxonomic supergroups now included in our analysis. Further, support values assess sampling effects but cannot indicate whether the tree is accurate. Thus, the previous published root may appear correct, as had several of our own reconstructions, but could be subject to systematic artifacts. If the previously published rooting is due to nonphylogenetic signal in some of the data set, then we predict that there will be a significant fraction of genes with compositionally deviant taxa and that removing these genes would have a marked effect on the tree accuracy.

Using the same taxa from that study (Fenn et al. 2006), we constructed protein-sequence alignments for each of

FIG. 5.—Unrooted phylogenomic tree inferred using the site-heterogenous CAT mixture model after excluding 6 compositionally deviant protein sequences of the original 21 sequences. Support values at the nodes indicate Bayesian posterior probabilities. The scale bar indicates the distance in substitutions per nucleotide.
the 42 genes. To expand the taxon diversity in this data set, we also added newly available genes from the supergroup B *Wolbachia* genome sequence to the alignments (Klasson et al. 2008). Using the PhyloBayes Z-score analysis for compositional heterogeneity, we identified that the overwhelming majority of genes had compositionally deviant taxa that could mislead the phylogenomic inferences. Indeed, only 10 of the original 42 genes in that data set were found to have Z scores below the default cutoff of 2.0, *P* < 0.05 (fig. 4B). Following data removal and tree reconstruction under CAT, we demonstrate in figure 6A that there is no defined, rooting position and the branch to the three outgroups is fatally long. Thus, composition variability has severely misled the tree accuracy reported previously. These results further illustrate the importance of examining nonphylogenetic signal that can mislead phylogenomic analyses.

To help increase the number of genes analyzed under the same criteria established above, we combined the compositionally homogenous genes from our data set with those of the previously published data set (Fenn et al. 2006) for a total of 25 protein-coding genes. We continued to use the same ingroup and outgroup taxa, with the exception that we added a fourth outgroup, *A. phagocytophilum*, because the whole-genome sequence was available. Following a new Z-score analysis, four genes were discarded: 2B, 2F, 2G, and WD0359. Under the CAT inference method, these 21 genes yielded a tree that polarizes the ancestor of *Wolbachia* in the reproductive parasites, supergroup A (fig. 6B). As demonstrated in all our analyses, the CAT model continues to caution the rooting placement in this particular clade or for that matter any clade. The root has a 55% posterior probability support. Therefore, the benefit of using the data and models that can recover the most phylogenetic signal is interestingly the warning they send against confidently placing the root anywhere in the *Wolbachia* tree.

### Conclusion

We presented an extensive, evolutionary analysis of mutualism and parasitism in *W. pipientis* endosymbionts, widespread intracellular bacteria of arthropods and filarial nematodes. The utility of the *Wolbachia* endosymbiosis to assess transitional steps between parasitism and mutualism in obligate intracellular bacteria has been well recognized (Lo et al. 2002, 2007; Bordenstein and Rosengaus 2005; Fenn and Blaxter 2006). Within the genus, there are four genetically defined lineages that can be discretely clustered into clades of mutualists that enhance filarial nematode fertility or viability (supergroups C and D) and clades of mostly reproductive parasites that distort sexual reproduction of their arthropod hosts to enhance their own transmission through the matriline (supergroups A and B). Despite attempts to reconstruct the ancestry of mutualism and parasitism in this bacterium, only one large-scale phylogenomic data set has been performed and claimed to position the root between the arthropod A and B parasites and the nematode C and D mutualists (Fenn et al. 2006).

We demonstrated in the analyses here that exploiting an amino acid mixture model and excluding mutationally saturated genes are crucial aspects of correctly accounting for the systematic error associated with the exceedingly long branch lengths between the last common ancestor of the ingroup and that of the outgroup genera. Indeed, a sequential look at the MP, ML/Bayesian, and CAT trees.
shows a progressively, elongating outgroup branch. As the quality of the inference methods improve, the systematic biases become more evident in the length of the branch to the outgroup. Empirical matrices for protein sequence evolution including WAG, JTT, and GTR and probabilistic inference methods such as ML and Bayesian ML all failed to resolve the root upon a deeper assessment. In some cases, high statistical ML and Bayesian support for the rooting placement could have been interpreted as discovering the accurate tree and closing the book on the transition in symbiotic lifestyle. However, each of these models and implementations can be subject to systematic errors because they assume that the underlying model of sequence evolution is the same across the whole sequence and that there is no or weak mutational saturation. This assumption can lead to erroneous, but highly supported reconstructions, especially in genome-scale data sets with poor gene or taxon sampling. Phylogenomics must emphasize quality rather than quantity. Phylogenetic artifacts that lurk in the background of single-gene analyses can be compounded in large data sets such as the ones used here. Indeed, our subsequent analyses revealed that LBA artifacts grossly misled the phylogenetic accuracy of our data set and a previously published one (Fenn et al. 2006). After correctly accounting for the source of these LBA artifacts, the rooting placement returned to its rightful, ambiguous placement. In addition, recombination did not affect the global analysis and the accuracy of the ingroup tree improved once compositionally deviant genes were excluded. The D. gracile nematode lineage did improve from an aberrant positioning near the supergroup A Wolbachia in arthropods to a more accurate placement with strong support near the other nematode Wolbachia in supergroup C.

Even with the most taxonomically rich, phylogenomic data set of the Wolbachia clade, careful gene selection, and site-heterogeneous inference models, LBA artifacts preclude an accurate rooting of the Wolbachia tree. The findings suggest that all past reconstructions using the Anaplasmataceae outgroups are fatally flawed from findings suggest that all past reconstructions using the Wolbachia and site-heterogenous inference models, LBA artifacts near the supergroup A tode lineage did improve from an aberrant positioning. In addition, the source of these LBA artifacts, the rooting placement preceded other hybrid incompatibilities in Nasonia. Nature. 409:707–710.


Acknowledgments

The authors thank Dr Nicholas Lartillot for helpful advice on PhyloBayes and the manuscript and Dr Laura Baldo, Sarah Biber, and Rahul Nene for technical assistance. This work was supported by grants from EU FP7 CSA_SA_REGPROT-2007-1 (Number 203590—MicrobeGR) and University of Ioannina to K.B.; NSF EF-0328363 to J.H.W., S.R.B, J.C.D.H, and H.T.; and National Science Foundation IOS-0852344, National Institute of Health (NIH) R01 GM085163-01, and NASA Astrobiology Institute NNA04CC04A to S.R.B. N.L. was supported by an Australian Research Council Postdoctoral Fellowship.

Literature Cited


de Bary A. 1879. The phenomenon of symbiosis. Strassburg: Verlag Von Karl J. Trubner.


