Bacterial DNA Sifted from the *Trichoplax adhaerens* (Animalia: Placozoa) Genome Project Reveals a Putative Rickettsial Endosymbiont

Timothy Driscoll[1,†], Joseph J. Gillespie[1,2,*,†], Eric K. Nordberg[1], Abdu F. Azad[2], and Bruno W. Sobral[1,3]

[1]Virginia Bioinformatics Institute at Virginia Polytechnic Institute and State University

[2]Department of Microbiology and Immunology, University of Maryland School of Medicine

[3]Present address: Nestlé Institute of Health Sciences SA, Campus EPFL, Quartier de L'innovation, Lausanne, Switzerland

*Corresponding author: E-mail: jgille@vbi.vt.edu.

[†]These authors contributed equally to this work.

Accepted: March 1, 2013

Abstract

Eukaryotic genome sequencing projects often yield bacterial DNA sequences, data typically considered as microbial contamination. However, these sequences may also indicate either symbiont genes or lateral gene transfer (LGT) to host genomes. These bacterial sequences can provide clues about eukaryote–microbe interactions. Here, we used the genome of the primitive animal *Trichoplax adhaerens* (Metazoa: Placozoa), which is known to harbor an uncharacterized Gram-negative endosymbiont, to search for the presence of bacterial DNA sequences. Bioinformatic and phylogenomic analyses of extracted data from the genome assembly (181 bacterial coding sequences [CDS]) and trace read archive (16S rDNA) revealed a dominant proteobacterial profile strongly skewed to Rickettsiales (*Alphaproteobacteria*) genomes. By way of phylogenetic analysis of 16S rDNA and 113 proteins conserved across proteobacterial genomes, as well as identification of 27 rickettsial signature genes, we propose a Rickettsiales endosymbiont of *T. adhaerens* (RETA). The majority (93%) of the identified bacterial CDS belongs to small scaffolds containing prokaryotic-like genes; however, 12 CDS were identified on large scaffolds comprised of eukaryotic-like genes, suggesting that *T. adhaerens* might have recently acquired bacterial genes. These putative LGTs may coincide with the placozoan's aquatic niche and symbiosis with RETA. This work underscores the rich, and relatively untapped, resource of eukaryotic genome projects for harboring data pertinent to host–microbial interactions. The nature of unknown (or poorly characterized) bacterial species may only emerge via analysis of host genome sequencing projects, particularly if these species are resistant to cell culturing, as are many obligate intracellular microbes. Our work provides methodological insight for such an approach.

Key word: endosymbiont, genome-mining, host–microbe interactions, intracellular bacteria, Rickettsiales, symbiosis.

Introduction

Bacterial DNA sequences may be generated as a byproduct of eukaryotic genome sequencing. The source of this bacterial DNA can be 1) contamination (failure to separate incidental bacterial species from eukaryotic cell preparation, or even failure to completely eliminate sequencing adapters or cloning vector sequences); 2) environmental (extracellular bacteria sequenced as a consequence of occupying the same niche as the eukaryote); 3) symbiotic (extracellular or facultative/obligate intracellular bacterial species directly associated with the eukaryotic host); or 4) LGT [*l*ateral *g*ene *t*ransfer of bacterial sequences to the genome of the eukaryote]. Aside from contamination, the remaining sources of bacterial DNA sequences generated by eukaryotic genome projects provide clues about the biological relationships between eukaryotes and their associated bacterial species. Thus, the creation of methods for the detection, extraction, and characterization of microbial sequences generated by eukaryotic genome sequencing studies are of critical importance, particularly for gaining insight on poorly characterized bacterial species, some of which may be recalcitrant to cultivation. Furthermore, annotation of such bacterial reads or genomes and deposition

into appropriate public databases is paramount for facilitating this approach.

Studies identifying endosymbiotic bacterial genomes within the data generated from eukaryotic sequencing projects are growing. Inspection by Salzberg et al. (2005a, 2005b) of disparate fruit fly (Arthropoda: Diptera: *Drosophila ananassae*, *D. simulans*, and *D. willistoni*) genome trace file archives resulted in the identification of three novel species of *Wolbachia* (*Alphaproteobacteria*: Rickettsiales: Anaplasmataceae). The genome sequence of another *Wolbachia* strain was discovered within the whole-genome sequencing data for the mosquito *Culex quinquefasciatus* strain JHB (Salzberg et al. 2009). Sequencing of the *Hydra magnipapillata* (Cnidaria: Hydrozoa) genome revealed the presence of an endosymbiont most closely related to species of *Curvibacter* (*Betaproteobacteria*: Burkholderiales: Comamonadaceae) (Chapman et al. 2010). Most recently, the genome of a *R*ickettsiales endosymbiont of *Ixodes scapularis* (Rickettsiales: Rickettsiaceae: REIS) was assembled from mining the initial data generated from the deer tick sequencing effort (Gillespie et al. 2012). All of these studies have revealed genomic data essential for furthering the knowledge of bacterial endosymbioses within animal species. In the case of REIS, important characteristics of a nonpathogen came to light when compared with the genomes of closely related pathogenic-spotted fever group rickettsiae (Gillespie et al. 2012).

Genomic analyses of several eukaryotes, such as the rotifers *Adineta vaga* and *A. ricciae* (Rotifera; Bdelloidea) (Gladyshev et al. 2008; Boschetti et al. 2012), *H. magnipapillata* (Chapman et al. 2010), the silkmoth *Bombyx mori* (Arthropoda: Lepidoptera) (Li et al. 2011), and the spider mite *Tetranychus uriticae* (Arthropoda: Acari) (Grbic et al. 2011), have revealed the presence of many genes originating from diverse bacterial species, illustrating the role of LGT in the diversification of eukaryotic genomes. For instance, a bacterial mannanase gene from *Bacillus* spp. (Firmicutes: Bacilliales) was recently reported in the genome of the coffee berry borer beetle, *Hypothenemus hampei* (Arthropoda: Coleoptera), and demonstrated to metabolize galactomannan, the major storage polysaccharide of coffee (Acuna et al. 2012). Large portions of *Wolbachia* genomes have been identified in several arthropod host genomes, including the bean beetle *Callosobruchus chinensis* (Kondo et al. 2002; Nikoh et al. 2008), the longicorn beetle *Monochamus alternatus* (Aikawa et al. 2009), *D. ananassae*, (Dunning Hotopp et al. 2007), the parasitoid wasp *Nasonia vitripennis* (Arthropoda: Hymenoptera) (Werren et al. 2010), as well as several filarial nematode genomes (Dunning Hotopp et al. 2007; McNulty et al. 2010), underscoring the prevalence of LGT between obligate intracellular bacterial species and their eukaryotic hosts. Intriguingly, several bacterial genes encoded in the genome of the pea aphid, *Acyrthosiphon pisum* (Arthropoda: Hemiptera), presumably foster its well-characterized mutualism with *Buchnera aphidicola* (*Gammaproteobacteria*: Enterobacteriales), potentially relegating the symbiont to aphid bacteriocytes (Nikoh and Nakabachi 2009; Nikoh et al. 2010). These studies of bacterial gene incorporation into eukaryotic genomes illustrate the need to develop tools to distinguish congener bacterial genes serendipitously captured in eukaryotic sequencing projects from true LGT events.

In this study, we analyzed the genome project (reads and assembly) of the primitive metazoan *Trichoplax adhaerens* (Animalia: Placozoa) for the presence of bacterial DNA sequences. Published in 2008, the *T. adhaerens* genome revealed "cryptic complexity," as most genes encoding transcription factors and signaling pathways underpinning eumetazoan cellular differentiation and development are present in this simple animal (Srivastava et al. 2008). *Trichoplax adhaerens* lacks nerves, sensory cells, and muscle cells, with only four cell types previously described (Grell 1971; Schierwater 2005). Morphologically, the animal resembles a flat disc of cells with two epithelial layers sandwiching a region of multinucleate fiber cells (Grell and Ruthmann 1991; Guidi et al. 2011). *Trichoplax adhaerens* is known to harbor a Gram-negative endosymbiont within fiber cells (Grell 1972; Grell and Benwitz 1974), with bacteria passed to developing oocytes via fiber cell extensions (Eitel et al. 2011). Our motivation for analyzing the *T. adhaerens* genomic data for sequences belonging to this symbiont was generated by previous studies that included bacterial-like genes from *T. adhaerens* in phylogeny estimations (Felsheim et al. 2009; Baldridge et al. 2010; Gillespie et al. 2010; Nikoh et al. 2010). As two of these genes are rickettsial signatures (*virD4* and plasmid-like *parA*), we considered it likely that the *T. adhaerens* fiber cell symbiont is a member of the obligate intracellular Rickettsiales.

We report an in-depth analysis of the *T. adhaerens* genome assembly and trace read archive, which divulged bacterial 16S rDNA sequences, 181 bacterial-like coding sequences (CDS) and many additional partial gene fragments of probable bacterial nature. Robust phylogenomic analyses grouped the *T. adhaerens* bacterium with the mitochondria invader "*Candidatus* Midichloria mitochondrii" (*Alphaproteobacteria*: Rickettsiales), albeit with only 53% conservation across the core proteins of these two species. Using this substantial molecular evidence, we name a Rickettsiales endosymbiont of *T. adhaerens* (RETA) and provide adjusted annotation and related genomic information for its genes deposited in the Pathosystems Resource Integration Center (PATRIC) (www.patricbrc.org, last accessed March 2013). This work illustrates the rich resource of eukaryotic genome projects for data pertinent to diverse host–microbial interactions, and also demonstrates that highly divergent, poorly known microbial species can be characterized via in-depth mining and phylogenomic analyses of even minimal genetic information captured from these broad-scale eukaryotic-sequencing efforts.

Materials and Methods

Small Subunit rDNA Analyses

Read Analysis

To assess the taxonomic distribution of bacterial species sequenced concomitantly with *T. adhaerens*, 1,230,612 WGS sequencing reads from the *T. adhaerens* genome project (Joint Genome Institute) were downloaded from the NCBI Trace Archive for analysis. Reads were cleaned of vector contamination using cross_match (Ewing et al. 1998) and screened for quality using the fastqc program from the Babraham Bioinformatics group (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, last accessed March 2013). All reads with a Phred quality score greater than 20 were subsequently mapped against a library of small subunit (SSU) rDNA sequences, including full-length bacterial 16S rDNA sequences from the Greengenes database (DeSantis et al. 2006) and the full-length 18S rDNA sequence for *T. adhaerens* (NCBI acc. no. Z22783). *Trichoplax adhaerens* sequencing reads were aligned to this SSU library using the Burrows–Wheeler Aligner (Li and Durbin 2010) with the options bwasw -t 4 -T 37. Reads that had at least one successful match in the SSU library were binned according to the taxonomic classification of their matches, and subsequently visualized using Krona v.2.0 (Ondov et al. 2011).

Phylogeny Estimation

The rickettsial-like 16S rDNA sequences retrieved from the *T. adhaerens* read archive were used as subjects in BLASTN searches of the NCBI NR database for the closest bacterial sequences (of greater or equal length). Seven rickettsial-like sequences, from various environmental studies (Revetta et al. 2010, 2011; Sunagawa et al. 2010), were retrieved and combined with a data set ($n = 47$) recently used to estimate Rickettsiales phylogeny (Gillespie, Nordberg, et al. 2012). Additional mitochondrial SSU rDNA sequences and outgroup sequences (*Betaproteobacteria*, *Gammaproteobacteria*, *Alphaproteobacteria*: Rhodospirillales, Parvularculales, Rhizobiales) were included based on previous phylogenetic studies of *Alphaproteobacteria* (Williams et al. 2007; Thrash et al. 2011; Viklund et al. 2012). Further rickettsial 16S rDNA sequences from recent studies (Kawafune et al. 2012; Matsuura et al. 2012) were added to entail robust sampling within the major Rickettsiales groups, bringing the data set to 93 SSU rDNA sequences. Information pertaining to all analyzed SSU rDNA sequences, and the consensus rickettsial 16S rDNA sequence mined from the *T. adhaerens* read archive, are provided in supplementary table S1, Supplementary Material online.

All sequences, plus a second set excluding the mitochondrial SSU rDNA sequences ($n = 83$), were aligned using MUSCLE v3.6 (Edgar 2004a, 2004b) with default parameters. Ambiguously aligned positions, the majority being present within the variable regions of the SSU rRNA structure, were culled using Gblocks (Castresana 2000; Talavera and Castresana 2007). Phylogenies were estimated under maximum likelihood using RAxML (Stamatakis et al. 2008). The GTR substitution model was used with estimation of GAMMA and the proportion of invariable sites. Branch support was measured with bootstrapping (1,000 replications).

CDS Analyses

Assembly Analysis

A BLAST-based pipeline was used to identify candidate bacterial CDS within the *T. adhaerens* genome assembly. Each of the 11,540 predicted proteins of the Triad1 assembly was used as a query in BLASTP searches against three databases: 1) A scoping database consisting of all available Rickettsiales proteins (NCBI taxonomy id 766); 2) all bacteria proteins (NCBI taxonomy id 2) excluding those in the Rickettsiales database; and 3) all eukaryotic proteins (NCBI taxonomy id 2759) excluding *T. adhaerens*. The choice of Rickettsiales for the scoping database was informed by preliminary results from our SSU taxonomic distribution analysis and phylogeny estimation, as well as by previous studies that included *T. adhaerens* bacterial-like genes in phylogeny estimations (Felsheim et al. 2009; Baldridge et al. 2010; Gillespie et al. 2010; Nikoh et al. 2010). For each *T. adhaerens* protein, the top 50 matches (based on *E* value) in each database were pooled and ranked according to a comparative sequence similarity match score, S_m (eq. 1):

$$S_m = b \times I \times Q \quad (1)$$

where b is the bitscore of the match, I is the percent identity (%ID), and Q is the percent length of the query that aligned. By incorporating %ID and match length, S_m is intended to de-emphasize highly significant matches to short stretches of query (i.e., conserved domains) in favor of longer stretches of similarity.

The top five scoring matches from the pooled lists of subjects were retained and grouped according to hit number (1–5) and organism taxonomy. *Trichoplax adhaerens* proteins with no top 5 scoring matches to bacteria were excluded from further analyses ($n = 9,843$, or 85.3% of the total *T. adhaerens* proteins). The remaining proteins ($n = 1,697$) were then subjected to cursory inspection of *T. adhaerens* annotation, as well as targeted BLASTP searches against various databases (*Alphaproteobacteria*, individual Rickettsiales genera, mitochondria, etc.) with manual inspection of functional annotations from top hits. A final data set of probable bacterial CDS ($n = 181$) was constructed with annotations derived primarily from Uniprot Consortium (2012), PATRIC (Aziz et al. 2008; Gillespie et al. 2011), and the NCBI-conserved domains database (Marchler-Bauer et al. 2011). In some cases (e.g., rickettsial signature proteins), annotations from the literature were selected. We named a hypothetical

organism, RETA, based on the hypothesis that these proteins define one single bacterial species. Each protein was assigned a unique identifier RETA0001–RETA0181. A complete list of the RETA proteins is available at PATRIC (http://enews.patricbrc.org/rickettsial-endosymbiont-of-trichoplax-adhaerens/, last accessed March 2013) and provided in supplementary table S2, Supplementary Material online.

Data Set Classification

The 181 bacterial-like CDS extracted from the *T. adhaerens* assembly were divided into two groups based on manual inspection of BLASTP results. A core data set of proteins with conserved domains (functions) that are generally vertically inherited, and hence not typical constituents of the bacterial mobilome, was constructed ($n = 119$). These bacterial-like proteins had one of three characteristics: 1) Top BLASTP hits to Rickettsiales with the next closest homologs in *Alphaproteobacteria*; 2) top BLASTP hits to *Alphaproteobacteria* with rickettsial homologs present or absent; or 3) top BLASTP hits to other *Proteobacteria* but with highly similar rickettsial homologs. This relaxed criterion permitted the capture of putative rickettsial-like genes that may not be known from the available rickettsial (or even alphaproteobacterial) sequenced genomes. Further, it allowed for identifying CDS that may be difficult to detect due to extreme divergence of the symbiont genome. Finally, this approach also provided flexibility with interpretation of BLASTP results, which may be biased due to a number of characteristics in the query and/or subject sequences (e.g., truncated sequences, length heterogeneity across matches due to insertions and deletions, *b*ase *c*ompositional *b*ias [BCB], etc.). Three instances of split open reading frames (ORFs) were detected (*secA*, *mnmA,* and *GlmS*), as well as three fused gene models (*tolC-sppA*, *rmuC-uvrD,* and *kdsA-smpA*), bringing the number of core data set genes to 116.

The remaining 62 bacterial-like CDS, or the accessory data set, mostly encompassed proteins of the bacterial mobilome, especially those typically encoded by intracellular species. Aside from lacking a phylogenetic signal typical of conserved alphaproteobacterial proteins, CDS of the accessory data set had one of the following characteristics: 1) highly similar to Rickettsiales signature proteins, 2) present in some (or all) Rickettsiales genomes yet divergent in sequence and phylogenetic signal, or 3) unknown from Rickettsiales genomes. Proteins of the accessory data set were analyzed separately (discusses later, accessory data set analyses) since, while they could all depict proteins encoded by one putative rickettsial symbiont (RETA), it is also possible that some may be from additional microbes captured in the *T. adhaerens* genome sequencing (particularly those species for which 16S rDNA sequences were mined).

Genome Comparison

Careful observation of the BLASTP profiles and preliminary phylogeny estimations revealed the mitochondria-associated rickettsial species "*Candidatus* Midichloria mitochondrii" (hereafter *M. mitochondrii*), (Lo et al. 2004; Sacchi et al. 2004; Sassera et al. 2006) as the closest relative (with available genome sequence data) to the majority of the mined bacterial-like CDS. Accordingly, an all-against-all BLASTP analysis was executed between *M. mitochondrii* ($n = 1{,}211$) and *T. adhaerens* ($n = 11{,}540$). The BLASTP results for 347 matches, including S_m scores and *E* values, were mapped over a circular plot of the *M. mitochondrii* genome using Circos (Krzywinski et al. 2009), with manual adjustment. Proteins of both the core and accessory data sets with homologs in *M. mitochondrii* ($n = 138$) were highlighted, and regions of synteny between the *M. mitochondrii* genome and CDS from several *T. adhaerens* scaffolds were superimposed on the plot.

Core Data Set Analyses

Genome-Based Phylogeny

The RETA core data set proteins were combined under the assumption that they were vertically inherited from an alphaproteobacterial ancestor. To better understand the systematic position of RETA, a total of 176 genomes were used for robust phylogeny estimation (supplementary table S3, Supplementary Material online). Aside from the RETA core data set proteins, the analysis included genomes from 80 Rickettsiales, 82 non-Rickettsiales *Alphaproteobacteria*, 12 mitochondria, and two outgroup taxa (*Betaproteobacteria* and *Gammaproteobacteria*). The *T. adhaerens* mitochondrial genome, which was generated separately from the whole genome sequencing project (Dellaporta et al. 2006), was used. Taxon sampling was modeled after several previous studies on Rickettsiales phylogeny (Sassera et al. 2011; Rodriguez-Ezpeleta and Embley 2012; Viklund et al. 2012) for the purpose of presenting a comparable hypothesis. Trees from these previous studies are summarized in supplementary figure S1, Supplementary Material online, to assist the interpretation of our current hypothesis.

For genome-based phylogeny estimation, an automated pipeline for protein family selection and tree building was implemented in Java. Bacterial protein sequences were downloaded from PATRIC (Gillespie et al. 2011). The RETA and mitochondria proteins were extracted from NCBI, as were an additional 27 *Rickettsia* genomes not annotated at PATRIC at the time of analysis. BLAT (refined BLAST algorithm) (Kent 2002) searches were performed to identify similar protein sequences between all genomes, including the outgroup taxa. To predict initial homologous protein sets, mcl (Van Dongen 2008) was used to cluster BLAT results, with subsequent refinement of these sets using hidden Markov models

as previously described (Durbin et al. 1998). These protein families were then filtered to include only those with membership in more than 80% of the analyzed genomes (141 or more taxa included per protein family, excluding the mitochondrial genomes). Multiple sequence alignment of each protein family was performed using MUSCLE (default parameters) (Edgar 2004a, 2004b), and regions of poor alignment (length heterogeneous regions) were masked using Gblocks (Castresana 2000; Talavera and Castresana 2007). All modified alignments were concatenated into a single data set for phylogeny estimation.

Tree-building was initially performed using FastTree (Price et al. 2010). Support for generated lineages was estimated using a modified bootstrapping procedure, with 100 pseudoreplications sampling only half of the aligned protein sets per replication (note: standard bootstrapping tends to produce inflated support values for very large alignments). Local refinements to tree topology were attempted in instances where highly supported nodes have subnodes with low support. This refinement is executed by running the entire pipeline using only those genomes represented by the node being refined (with additional sister taxa for rooting purposes). The refined subtree is then spliced back into the full tree.

Using PhyloBayes v3.3 (Lartillot et al. 2009), we also analyzed the data set with the CAT model of substitution, which is a nonparametric method for modeling site-specific features of sequence evolution (Lartillot and Philippe 2004, 2006). Given the nature of the BCB of Rickettsiales and mitochondrial genomes, and the ability of the CAT model to accommodate saturation due to convergences and reversions (Lartillot et al. 2007), this approach is of substantial importance for estimating Rickettsiales phylogeny (Rodriguez-Ezpeleta and Embley 2012; Viklund et al. 2012). Two independent Markov chains were run in parallel using PhyloBayes MPI v.1.2e under the CAT-GTR model, with the bipartition frequencies analyzed at various time points using the bpcomp program. For tree-building, appropriate burn-in values were determined by plotting the log likelihoods for each chain over sampled generations (time). Analyses were considered complete when the maximum difference in bipartition frequencies between the two chains was less than 0.1. Ultimately, a burn-in value of 1,000, with sampling every 2 trees, was used to build a consensus tree.

Finally, to further evaluate the rickettsial nature of the core data set, all proteins were used as queries in BLASTP searches against three databases: Rickettsiales, Bacteria (excluding Rickettsiales), and Eukaryota (excluding *T. adhaerens*). The proteins were then binned into three "sub-data sets" (Ric, Bac, or Euk) based on the highest S_m score against each database (supplementary fig. S2, Supplementary Material online). The resulting three sub-data sets (Ric-78, Bac-26, and Euk-9) were then run through the procedure described earlier for phylogeny estimation, resulting in one FastTree-based and one PhyloBayes-based tree for each sub-data set (six total trees).

Genome Divergence

To determine whether the degree of divergence between the RETA core data set proteins and their homologs in *M. mitochondrii* is typical for major rickettsial lineages, an approximation of genome divergence across the genera of Rickettsiales and the RETA core data set was calculated. The final alignment of the core data set was processed to include only one representative species from each Rickettsiales genus (*Odyssella*, *Midichloria*, *Neorickettsia*, *Wolbachia*, *Anaplasma*, *Ehrlichia*, *Orientia*, and *Rickettsia*) plus the RETA core data set proteins. All positions of the alignment containing missing data (?s) were removed, resulting in 8,327 aa sites (8,319 informative). The program DIVEIN (Deng et al. 2010) was used to estimate percent protein divergence using both the Blosum62 and WAG amino acid substitution models.

Accessory Data Set Analyses

The 62 *T. adhaerens* bacterial-like sequences lacking a typical alphaproteobacterial signal (accessory data set) were separated from the core data set proteins using BLASTP searches. The NR (all GenBank + RefSeq Nucleotides + EMBL + DDBJ + PDB) database was used, coupled with a search against the Conserved Domains Database (Marchler-Bauer et al. 2011). Searches were performed across "all organisms", as well as "Rickettsiales" with composition-based statistics. No filter was used. Default matrix parameters (BLOSUM62) and gap costs (Existence: 11 Extension: 1) were implemented, with an inclusion threshold of 0.005. This process facilitated the division of the RETA accessory data set into three groups: 1) proteins with closest homologs to Rickettsiales ($n = 27$); 2) proteins present in (some or all) Rickettsiales genomes but divergent from their rickettsial counterparts ($n = 18$); and 3) proteins unknown from Rickettsiales ($n = 17$). The two groups containing rickettsial homologs were then used in subsequent BLASTP searches against the following five databases: 1) "Rickettsiales", 2) "*Alphaproteobacteria* (minus Rickettsiales)", 3) "*Proteobacteria* (minus *Alphaproteobacteria*)", 4) "Bacteria (minus *Proteobacteria*)", and 5) "minus Bacteria." The top 20–50 (query-dependent) subjects from each search resulting in significant (>40 bits) alignments were retrieved, compiled, and aligned using MUSCLE v3.6 (default parameters). Full alignments were used for subsequent analyses. In some instances, all sequences within alignments were screened for possible signal peptides using SignalP v.4.0 (Petersen et al. 2011), LipoP v.1.0 (Juncker et al. 2003), and Phobius (Kall et al. 2007). Potential transmembrane spanning regions were predicted using transmembrane hidden Markov model v.2.0 (Krogh et al. 2001).

Phylogenetic trees were estimated using PAUP* v4.0b10 (Altivec) (Wilgenbusch and Swofford 2003) under parsimony and implemented heuristic searches with 500 random sequence additions holding 50 trees per replicate. Single most parsimonious trees or consensus trees of equally parsimonious topologies were generated, with branch support assessed using bootstrapping (1,000 pseudoreplications). Phylogenies were also estimated under maximum likelihood using RAxML v.7.2.8 (Stamatakis et al. 2008). A gamma model of rate heterogeneity was used with estimation of the proportion of invariable sites. Branch support was assessed with 1,000 bootstrap pseudoreplications. Finally, for the analyses of flagella (FlgG and FliG) and T4SS proteins (RvhD4 and RvhB6) alignments were combined and analyzed together using both RAxML and PhyloBayes (as described earlier).

Evaluating Bacterial Gene Transfer to the *T. adhaerens* Genome

Several approaches were made to determine whether any of the 181 bacterial-like genes of the core and accessory data sets have strong evidence for being a part of the *T. adhaerens* genome (as opposed to belonging to the genomes of RETA or other microbes). We first evaluated the scaffold properties that contain each bacterial-like gene, judging that bacteria-to-host LGTs could only be demonstrated on scaffolds greater than one gene and containing eukaryotic-like genes. The 181 RETA genes were divided into four categories: 1) genes present on large (>40 genes) scaffolds with predominately eukaryotic-like genes ($n = 18$); 2) genes present on small (<7 genes) "hybrid" scaffolds with both bacterial- and eukaryotic-like genes ($n = 19$), 3) genes present on small (<5 genes) scaffolds comprised entirely of bacterial-like genes ($n = 59$), and 4) singleton-gene scaffolds ($n = 85$). Next, CDS within each category were split into single- and multi-exon genes. All multi-exon genes were then subjected to BLASTX searches using the entire gene models (exons + introns) as queries. These entire gene models were also analyzed with the bacterial gene prediction program fgenesb (Tyson et al. 2004), using the "generic BACTERIAL" model, to determine discrepancies with the original eukaryotic gene predictions within the *T. adhaerens* assembly. This intron evaluation allowed for the distinction between true eukaryotic genes inadvertently included within the RETA data sets (e.g., nuclear genes encoding mitochondrial proteins) and bacterial LGTs undergoing a transformation to eukaryotic-like gene structures (i.e., accruement of introns, gain of eukaryotic signal sequences, etc.). Importantly, the approach also revealed evidence against predicted introns due to 1) chimeric gene models comprised of two or more genes (or gene fragments) that were "stitched" together by the eukaryotic gene calling algorithms, 2) bacterial genes that were divided into fragments due to multiple start sites called by eukaryotic gene calling algorithms, and 3) gene models that were fused with additional short (and likely spurious) ORFs. Finally, for the 18 RETA genes found on large scaffolds that are dominated by eukaryotic-like genes, individual protein phylogenies were estimated (discussed earlier, accessory data set analyses) to lend an additional level of support for discerning between true eukaryotic genes and LGTs to the *T. adhaerens* genome.

Results

An overview of the methodology applied to the analysis of *T. adhaerens* genome project (sequence read archive and assembly) illustrates the various approaches implemented to identify bacterial DNA sequences (fig. 1). Totals for CDS and scaffolds are given for extracted data that suggest the presence of a rickettsial species, with more detailed information provided in the various sections below.

Bacterial DNA Mined from *Trichoplax*

SSU rDNA

Within the *T. adhaerens* genome project sequence read archive, a total of 289 SSU rDNA sequences were mined for analysis (fig. 2*a*). The majority (88.2%) of these sequences were identified as eukaryotic 18S rDNA genes belonging to the *T. adhaerens* genome. The remaining 34 SSU rDNA sequences were determined to have highest similarity with bacterial or plastid 16S rDNA-like sequences. Using the prokaryotic 16S rDNA sequences from the Greengenes database (DeSantis et al. 2006), these sequences received the most accurate taxonomic assignment possible (supplementary table S4, Supplementary Material online). Three major groups comprised 76.5% of the sequences: *Alphaproteobacteria* ($n = 9$), *Gammaproteobacteria* ($n = 4$), and eukaryotic chloroplasts ($n = 13$). The remaining eight sequences were grouped into a diverse array of taxa (*Betaproteobacteria*, *Deltaproteobacteria*, Spirochaetes, Firmicutes, and Plantomycetes). Importantly, ten of the 16 taxonomic assignments were made for one individual 16S rDNA operational taxonomic unit, with the bacterial assignments for *Marivita* spp. (Rhodobacterales), *Limnobacter* spp. (Burkholderiales), and *Borrelia* spp. (Spirochaetes) possibly representing a single organism with multiple rDNA operons (Kembel et al. 2012). The 13 cyanobacterial-like sequences had the best matches to chloroplasts of marine eukaryotes, such as haptophyte and cryptomonad algae, as well as heterokonts. These rDNA sequences may also be inflated due to a high copy number of plastid genomes. Finally, the two Rickettsiales sequences were determined to depict partial fragments of the same molecule, and thus were concatenated into one rDNA sequence and classified as RETA.

Bacterial CDS

Of the 11,540 predicted CDS within the *T. adhaerens* assembly, 14.7% ($n = 1,697$) had at least one prokaryotic

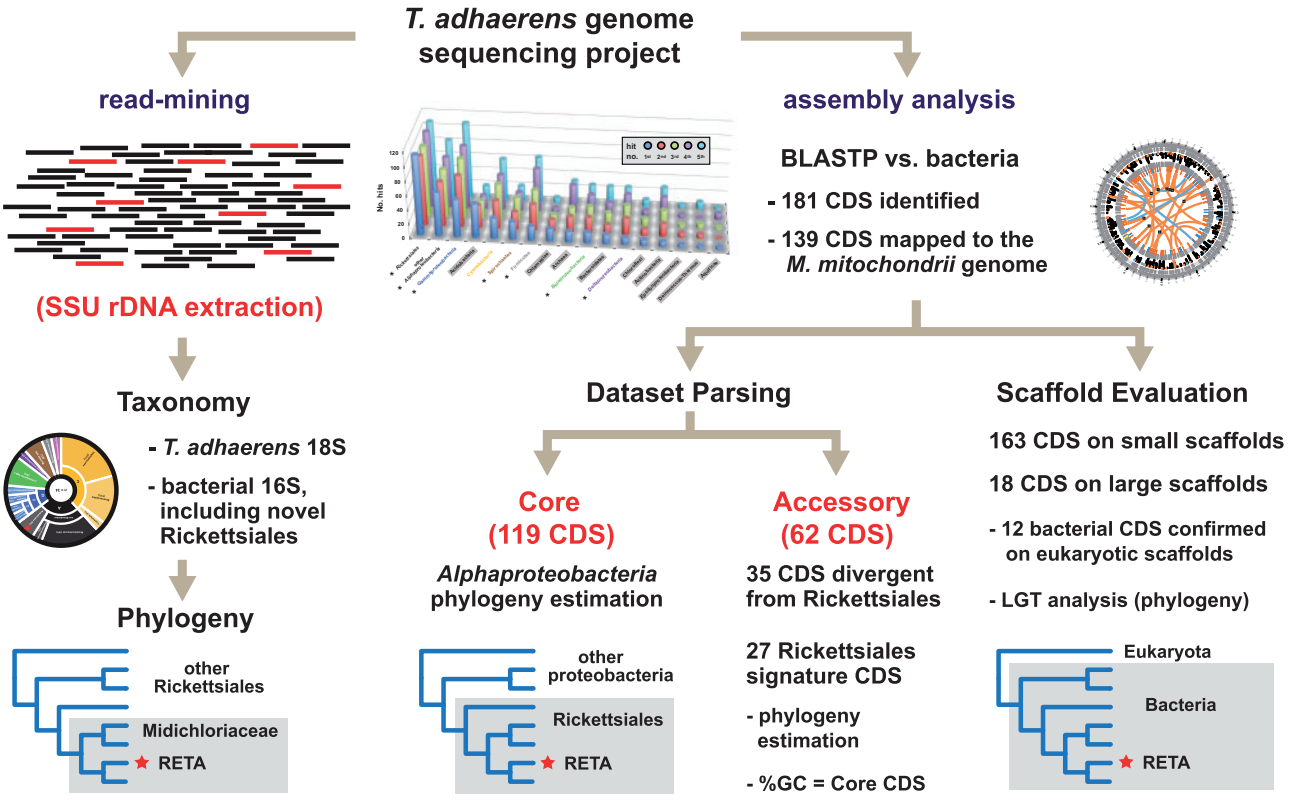


Fig. 1.—Overview of the methodology used to identify bacterial DNA sequences within the *Trichoplax adhaerens* genome project. Bacterial SSU rDNA sequences were mined from the trace read archive (left), with rickettsial sequences further analyzed via phylogeny estimation. Bacterial CDS identified in the assembly (right) were determined to be primarily rickettsial-like based on phylogeny estimation and identification of rickettsial signature genes. The distribution of bacterial CDS on small (bacterial-like) and large (eukaryotic-like) scaffolds is shown, with CDS on the latter further evaluated for LGT via phylogeny estimation.

analog within the top five scoring hits in a reciprocal BLASTP analysis against the NR database (fig. 2*b*). Pooling the bacterial hits according to higher-level taxonomy illustrated a bias towards Rickettsiales, other *Alphaproteobacteria* and *Gammaproteobacteria*, with those three groups showing at least one representative taxon within the top five hits in 163, 175, and 211 BLASTP matches, respectively. The remaining higher level taxa with more than 15 total hits (1–5) comprised a diverse group of prokaryotes. Importantly, aside from potential symbiont DNA and LGT products, many of these *T. adhaerens* proteins are nuclear genes encoding proteins that are trafficked to and imported by the mitochondria, with the diversity of bacterial groups in the BLASTP matches consistent with the genetic mosaicism of nuclear-encoded mitochondrial genes (Thiergart et al. 2012). Along with Rickettsiales, other *Alphaproteobacteria*, and *Gammaproteobacteria*, four other higher level taxonomic groups (*Betaproteobacteria*, *Deltaproteobacteria*, Spirochaetes, and Firmicutes) within this analysis had a corresponding 16S rDNA mined from the *T. adhaerens* sequence read archive (fig. 2*a*). Finally, within each taxonomic group containing scores to over 75 of the 1,697 *T. adhaerens* proteins, only the Rickettsiales showed a consistent representation across hits 1–5 ($R^2 = 0.006$), with other groups showing increasing or decreasing representation across hits 2–5 (avg. $R^2 = 0.805$ for *Gammaproteobacteria*, Other *Alphaproteobacteria*, Firmicutes, Acidobacteria, Cyanobacteria, Deinococcus-Thermus, Archaea, and Actinobacteria). Thus, in most cases where Rickettsiales was the top scoring hit to a *T. adhaerens* protein, hits 2–5 were also occupied by Rickettsiales, suggesting a strong rickettsial signal within these proteins.

Rickettsiales 16S rDNA Phylogeny

Despite mining a diverse set of 16S rDNA sequences from the *T. adhaerens* sequence read archive, we only estimated a phylogeny of the RETA 16S rDNA sequence with a diverse group of Rickettsiales for three primary reasons: 1) a match between the retrieved rickettsial 16S rDNA (fig. 2*a*) and CDS (fig. 2*b*); 2) the long-known presence of an intracellular bacterial symbiont associated with *T. adhaerens* (Grell 1972; Grell and Benwitz 1974; Eitel et al. 2011); and 3) evidence from other studies suggesting the presence of rickettsial CDS within the *T. adhaerens* assembly (Felsheim et al. 2009; Baldridge et al. 2010; Gillespie et al. 2010; Nikoh et al. 2010). Phylogeny estimation

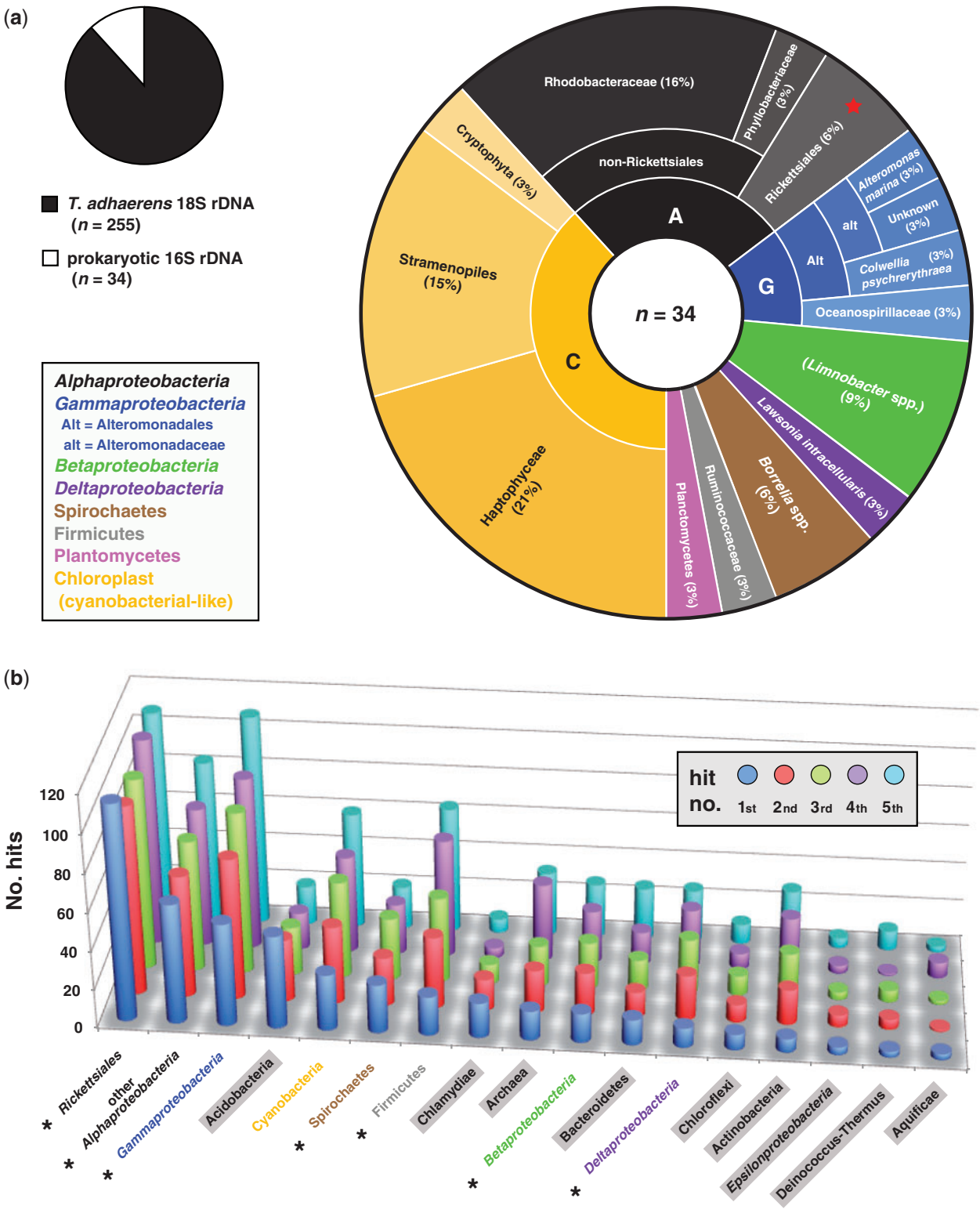


Fig. 2.—Identification of bacterial DNA sequences within the *Trichoplax adhaerens* genome trace read archive and assembly. (*a*) Illustration of 289 SSU rDNA sequences identified in the *T. adhaerens* trace read archive (http://genome.jgi-psf.org/Triad1/Triad1.download.ftp.html, last accessed March 2013). The pie chart at top left illustrates the 34 prokaryotic 16S rDNA sequences detected among 255 *T. adhaerens* 18S rDNA sequences. Larger graph at right illustrates the taxonomic distribution of the 34 prokaryotic 16S rDNA sequences (see text for details on taxonomic assignment). Sequences are grouped into nested sectors according to hierarchical taxonomy, progressing from the interior to exterior of the plot. Color scheme is explained in box at bottom left.

(continued)

of the SSU rDNA data set grouped RETA in a clade of diverse rickettsial species that is sister to the traditional Anaplasmataceae sensu stricto (Anaplasmataceae *s. s.*) (fig. 3). Previous studies have also recovered this large clade within Rickettsiales (Beninati et al. 2004; Davis et al. 2009; Vannini et al. 2010; Kawafune et al. 2012; Boscaro et al. 2013), which includes many species with diverse eukaryotic hosts, and we recently proposed the name "Midichloriaceae" as a sister family within the Anaplasmataceae sensu lato (Gillespie, Nordberg, et al. 2012). Here, we determined RETA to be part of a clade comprising poorly described bacteria from species of coral (*Gorgonia ventalina* and *Montastraea faveolata*) and sponge (*Cymbastela concentrica*) (Revetta et al. 2010, 2011; Sunagawa et al. 2010), consistent with the aquatic habitat of *T. adhaerens*. Importantly, this lineage is well diverged from the group comprising *M. mitochondrii* (89% identity between RETA and *M. mitochondrii* str. IricVA 16S rDNA sequences), which is comprised predominantly of bacteria identified in various arthropod species. The basal lineages of "Midichloriaceae" are comprised mostly of protist-associated rickettsial species and uncharacterized species collected via environmental sampling. Collectively, our analysis of the RETA SSU rDNA sequence retrieved from the *T. adhaerens* sequence reads is consistent with the presence of a rickettsial bacterial symbiont associated with Placozoa.

A Rickettsiales Genome Associated with *Trichoplax*

The 1,697 "bacterial-like" proteins identified in our CDS mining of the *T. adhaerens* (fig. 2*b*) were further evaluated via manual inspection of annotation, as well as a series of BLASTP analyses against specific databases (bacterial groups and mitochondria), to yield a data set of 181 probable bacterial CDS (supplementary table S2, Supplementary Material online). The proteins were divided into RETA core ($n = 119$) and accessory ($n = 62$) data sets and assigned unique identifiers (RETA0001-RETA0181). Given that the majority of subjects from all-against-all BLASTP analyses (between *T. adhaerens* and "all bacteria") were from Rickettsiales genomes (fig. 2*b*), we compared the RETA proteins directly with the taxon containing the most top BLASTP subjects, *M. mitochondrii* (fig. 4*a*). An all-against-all BLASTP analysis between *T. adhaerens* and *M. mitochondrii* yielded 347 hits above a set threshold ($S_m > 20$), with 124 of these matches illustrating homologous proteins from the *M. mitochondrii* genome and the *T. adhaerens* assembly. Additional CDS ($n = 14$) were later identified as the best RETA-*M. mitochondrii* matches based on manual BLASTP analyses (supplementary fig. S3, Supplementary Material online). Thus, a total of 138 RETA proteins were mapped to the *M. mitochondrii* genome, comprising 93.2% ($n = 111$) of the core and 41.5% ($n = 27$) of the accessory RETA data sets (fig. 4*b*). Despite being present on predominantly small scaffolds within the *T. adhaerens* (discussed later, Bacterial Genes in the *Trichoplax* Genome), seven regions of synteny were identified across RETA and *M. mitochondrii*, an understandable result given the lack of genome synteny across genera of Rickettsiales (Gillespie, Nordberg, et al. 2012). Collectively, 76.2% of the RETA CDS were found to have highly similar homologs in the *M. mitochondrii* genome, including seven syntenic regions, suggesting that these bacterial CDS from the *T. adhaerens* genome project comprise a potential Rickettsiales bacterium.

Genome-Based Phylogenetic Position of RETA

An estimated phylogeny of the core data set, which accommodated the strong inherent BCB in some of the data, unambiguously placed RETA with *M. mitochondrii* in a clade ("Midichloriaceae") within the Rickettsiales (fig. 5). The sampled mitochondrial genomes formed a lineage within the Rickettsiales, diverging after the Holosporaceae ("*Candidatus* Odyssella thessalonicensis," hereafter *O. thessalonicensis*) but prior to the derived rickettsial families (Anaplasmataceae, Rickettsiaceae). Thus, the ancestral positions of the Holosporaceae and branching point for the mitochondrial ancestor are consistent across trees estimated from SSU rDNA (fig. 3) and multiple proteins (fig. 5). The discrepancy in the placement of "Midichloriaceae" as ancestral to the Rickettsiaceae in the genome-based tree versus its sister relationship to the Anaplasmataceae *s. s.* in the SSU rDNA-based tree is explained by the lack of genomic data available for other members of the "Midichloriaceae."

Although the tree generated without accommodating BCB also grouped RETA and *M. mitochondrii* together, the position of the SAR11 group of *Alphaproteobacteria* within the Rickettsiales was recovered (supplementary fig. S4, Supplementary Material online). Importantly, previous genome-based phylogenies estimated with methods that accommodate BCB unambiguously show that SAR11 are not

Fig. 2.— Continued

Cyanobacterial hits correspond to chloroplast rDNA sequences of cyanobacterial origin. Plot made with Krona v.2.0 (Ondov et al. 2011) with manual adjustment. (*b*) Illustration of *T. adhaerens* proteins that have strong similarity to their prokaryotic counterparts. All proteins encoded within the *T. adhaerens* assembly (NCBI, ASM15027v1, $n = 11,540$) were used as queries in BLASTP searches against prokaryotic and eukaryotic proteins within the nr database (NCBI), with subjects ranked according to S_m score (see text for details). Graph depicts the taxonomic distribution of the top five scores per *T. adhaerens* protein that included a prokaryotic protein ($n = 1,697$). The taxa are arrayed along the *x* axis in decreasing order according to the number of top hits (blue). Prokaryotic groups with less than 15 total hits per group (sum 1–5) are not shown. Asterisks depict taxonomic groups that also have a 16S rDNA sequence illustrated in panel *a*.

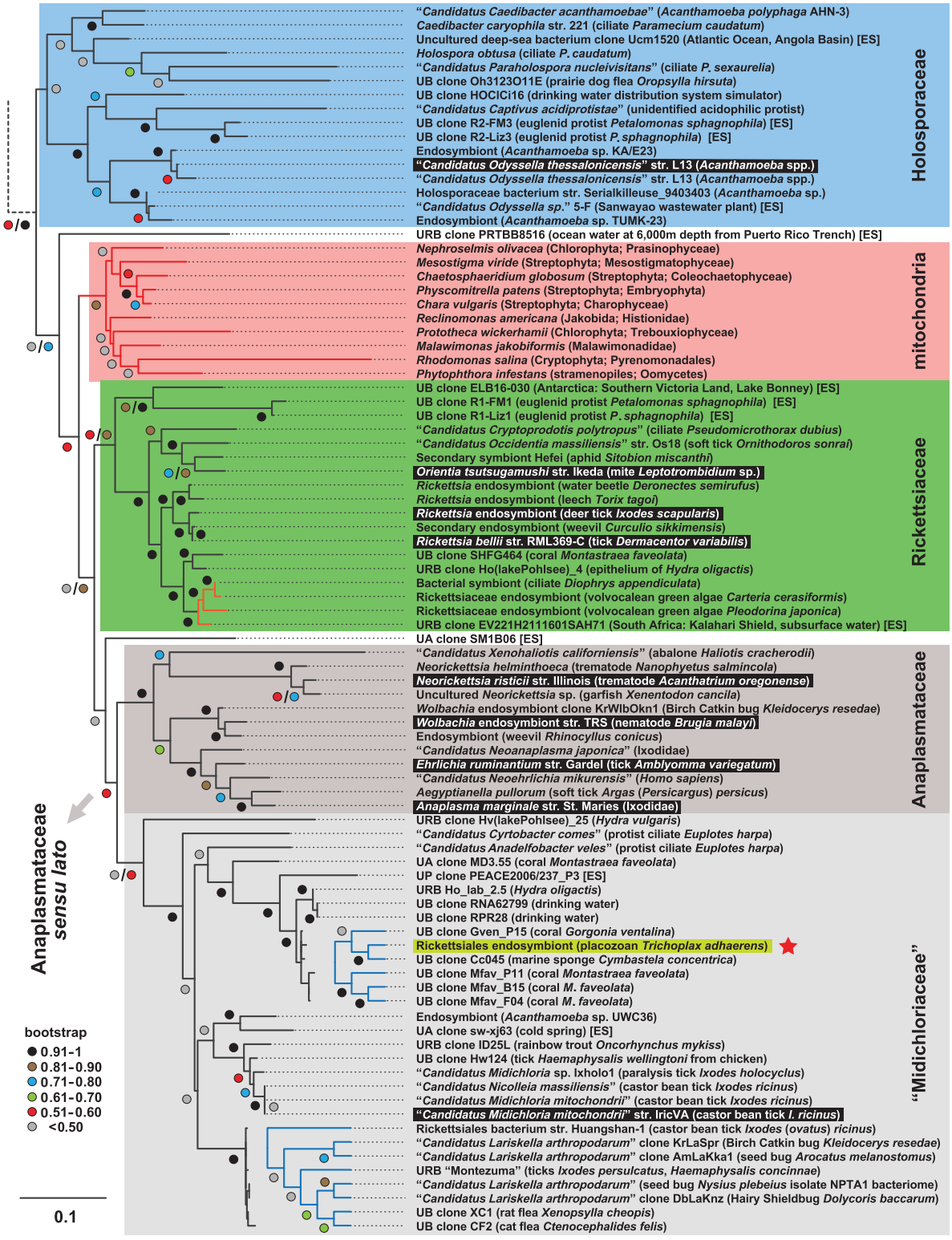


Fig. 3.—Phylogeny of SSU rDNA sequences estimated for 78 Rickettsiales taxa, 10 mitochondria, and 5 outgroup taxa. See text for alignment and tree-building methods. Tree is final optimization likelihood: (−22042.321923) using GTR substitution model with GAMMA and proportion of invariant sites estimated. Brach support is from 1,000 bootstrap pseudoreplications. For nodes represented by 2 bootstrap values, the left is from the analysis that included 10 mitochondrial sequences, with the right from the analysis without the mitochondrial sequences. All nodes with single bootstrap values had similar support in both analyses. Red (mitochondria) and orange (within Rickettsiaceae) branches are reduced 75% and increased 50%, respectively. Blue cladograms depict

(continued)

closely related to mitochondria, yet are more derived in the *Alphaproteobacteria* (Rodriguez-Ezpeleta and Embley 2012; Viklund et al. 2012) (supplementary fig. S1, Supplementary Material online). Of note, our tree estimated without accommodating BCB placed *O. thessalonicensis* outside of the Rickettsiales as an early branching lineage of the remaining *Alphaproteobacteria*. Although originally described as a member of the Holosporaceae (Birtles et al. 2000), a study presenting genome-based phylogenies failed to group *O. thessalonicensis* within Rickettsiales (Georgiades et al. 2011). This disparity with our hypothesis is discussed later in light of rickettsial signatures and the nature of Holosporaceae (see Discussion).

The robustness of the phylogenetic signal within the core data set was determined by estimating trees from the three sub-data sets (Ric-78, Bac-26, and Euk-9), which were grouped based on top S_m score against three databases (Rickettsiales, Bacteria excluding Rickettsiales, and Eukaryota). Whether generated using RAxML (GTR model) or PhyloBayes (CAT model), estimated trees based on all three sub-data sets unambiguously grouped RETA within the Rickettsiales (supplementary fig. S5, Supplementary Material online). This suggests that, despite differences in top BLASTP hits using RETA proteins as queries (as revealed by S_m scores in supplementary fig. S2, Supplementary Material online), the proteins of the core data set contain a phylogenetic signal that consistently places RETA within the Rickettsiales, distinct from the Rickettsiaceae and Anaplasmataceae.

RETA and Rickettsiales Genome Divergence

Although some RETA CDS have a clear rickettsial signature (i.e., *rvhD4*, *parA,* and *ampD* described previously (Felsheim et al. 2009; Baldridge et al. 2010; Gillespie et al. 2010; Nikoh et al. 2010), many of the mined CDS had limited %ID to any counterparts in the NR database, making distinction even between eukaryotic and prokaryotic proteins often difficult to assess by sequence comparison alone. Despite this, a predominant rickettsial signal emerged from the mined CDS (fig. 5), yet it is clear from the 16S rDNA tree that RETA is a member of a diverse rickettsial lineage with minimal genomic information available (*M. mitochondrii*). To determine whether RETA is typical in its degree of divergence from other rickettsial lineages, we calculated genome divergence across the most conserved protein regions within the core data set, sampling one representative member of each rickettsial genus, plus RETA (table 1). As expected, RETA was most similar to *M. mitochondrii*, with these genomes being 45% divergent from one another. The mean genome divergence across the major Rickettsiales genera was 51%, and ranged from 30% (*Anaplasma* vs. *Ehrlichia*) to 63% (RETA vs. *Neorickettsia*). Importantly, RETA had a mean divergence of 56% compared with other rickettsial genomes, just below *Neorickettsia* (57%) and not atypical from other rickettsial lineages. These results are consistent with our phylogeny estimation based on 16S rDNA (fig. 3) and the conserved proteins of the core data set (fig. 5).

Variable Bacterial CDS Mined from *Trichoplax*

Of the total bacterial-like CDS identified from the *T. adhaerens* genome assembly, 32% ($n = 62$) were determined to lack a conserved alphaproteobacterial signal, meaning there is little support for the vertical inheritance of these genes from a common alphaproteobacterial ancestor. These 62 RETA accessory data set proteins were further divided into three groups based on BLASTP analyses: 1) proteins with closest homologs to Rickettsiales ($n = 27$); 2) proteins present in (some or all) Rickettsiales genomes but divergent from their rickettsial counterparts ($n = 18$); and 3) proteins unknown from Rickettsiales ($n = 17$). Plotting these CDS by their %GC revealed that the Rickettsiales-like proteins differed from the genes within the other two categories, possessing a mean %GC consistent with Rickettsiales genomes (fig. 6). Importantly, the mean %GC of these rickettsial-like CDS is the same as the mean %GC of the RETA core data set CDS (29%), suggesting that at least these 27 CDS of the accessory data set likely belong to RETA.

Phylogeny estimations substantiated the rickettsial nature of the entire 27 rickettsial-like CDS (supplementary fig. S6, Supplementary Material online). Most of these "rickettsial signature" CDS encode secretion system proteins (*rvh* T4SS, *apr*-like T1SS), variable transporters involved in metabolite scavenging from the host (*gltP*, *tlc1*, 2A0306, and *citT*), and proteins involved in intracellular growth and survival (*spoT*, *sodA*, *proP*, patatins [cd07199], *ampG*, and *mdlB*) (supplementary table S5, Supplementary Material online). In addition, six other CDS in this set provided overwhelming evidence for the presence of a rickettsial symbiont associated with *T. adhaerens*. The *parA* gene, which is ubiquitous on rickettsial plasmids (Baldridge et al. 2010; Gillespie et al. 2012), hints at a

Fig. 3.— Continued
minimally resolved lineages within the "Midichloriaceae." The divergence point of five outgroup taxa from *Betaproteobacteria* ($n = 1$), *Gammaproteobacteria* ($n = 1$), and other *Alphaproteobacteria* ($n = 3$) is shown with a dashed branch. For each taxon, associated hosts are within parentheses, with ES depicting an environmental sample. Other abbreviations: UB, uncultured bacterium; UP, uncultured proteobacterium; UA, uncultured alphaproteobacterium; URB, uncultured Rickettsiales bacterium. Taxa within black boxes have available genome sequence data. The 16S rDNA sequence mined from the *T. adhaerens* trace archive is boxed green and noted with a red star. Accession numbers for all sequences are provided in supplementary table S1, Supplementary Material online.

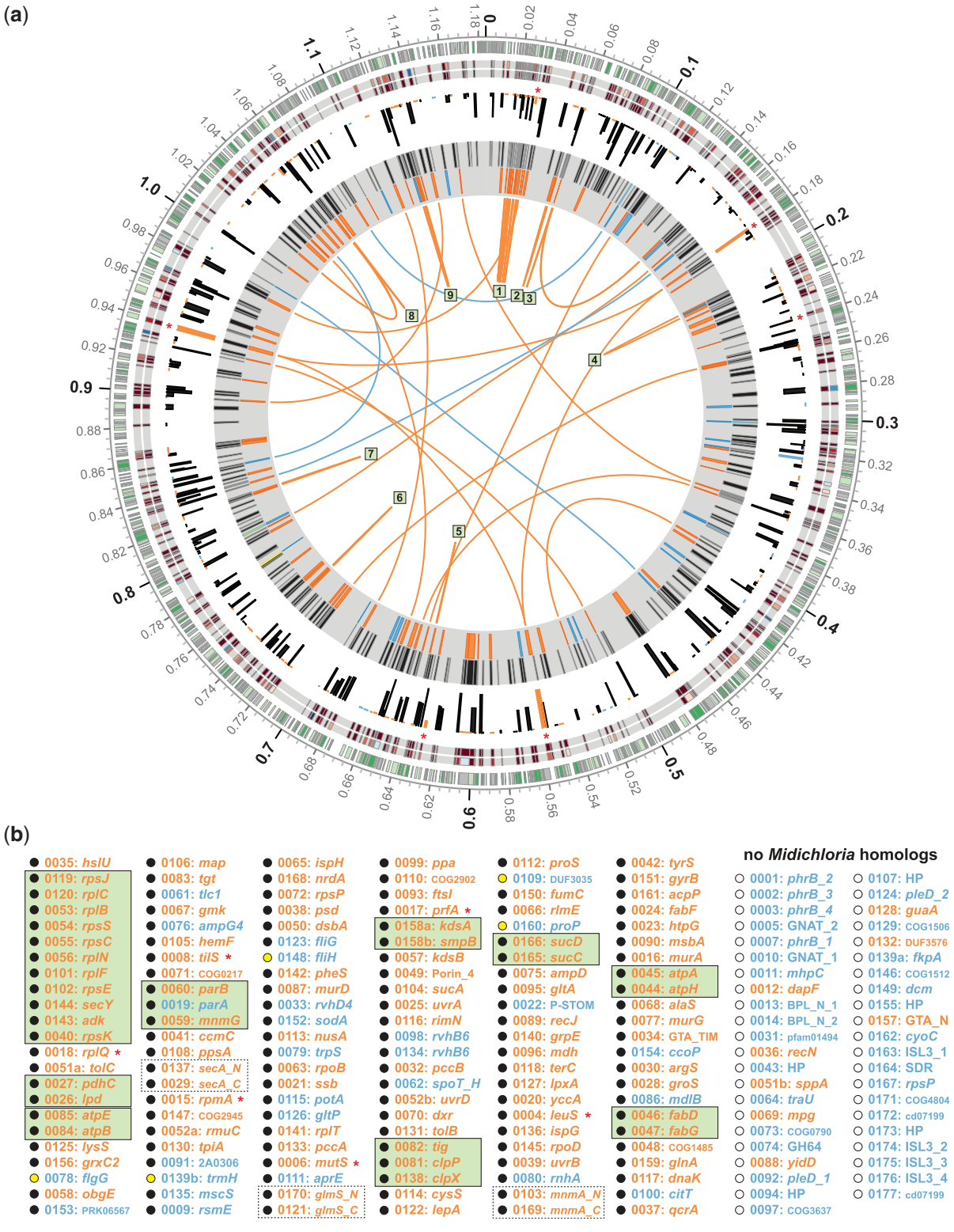


Fig. 4.—Bacterial CDS identified within the *Trichoplax adhaerens* genome assembly. (*a*) Results of an all-against-all BLASTP analysis between the genomes of *T. adhaerens* Grell-BS-1999 ($n = 11,540$) and "*Candidatus* Midichloria mitochondrii" str. IricVA ($n = 1,211$), hereafter *M. mitochondrii*. Outer black circle is a scale with coordinates (in Mb) for the *M. mitochondrii* genome, with the putative origin of replication positioned at 12 o'clock as previously determined (Sassera et al. 2011). Four rings inside the scale as follows: 1) 1,211 CDS of the *M. mitochondrii* genome, with operons and transcriptional units (predicted using fgenesb (Tyson et al. 2004)) colored green and gray, respectively; 2) heat maps for S_m scores >20 (outer) and corresponding *E* values (inner)

(continued)

possible plasmid associated with RETA. Furthermore, this CDS is fused with a short ORF encoding the antitoxin HicB, and *parA* and *hicB* are adjacent on the plasmids carried by "*Candidatus* Rickettsia amblyommii" strains, *R. massiliae* strains, and *R. rhipicephali*. PRK06567 encodes a putative bifunctional protein with both glutamate synthase subunit β (GltD) and 2-polyprenylphenol hydroxylase (UbiB) domains. Chimeric GltD-UbiB proteins are encoded in nearly every Rickettsiales genome (save *O. tsutsugamushi*), but sparsely encoded in other proteobacterial genomes and unknown from other bacterial genomes. The estimated phylogeny of p-stomatins, members of the SPFH (Stomatin-Prohibitin-Flotillin-HflC/K superfamily (Hinderhofer et al. 2009), grouped the RETA protein with *M. mitochondrii* and Rickettsiaceae in a clade that is the likely origin of eukaryotic SPFH members (Thiergart et al. 2012). Finally, three components of the bacterial flagella system (FlgG, FliG, and FliH) grouped with homologs from *M. mitochondrii* in the estimated tree, with this clade ancestral to all other alphaproteobacterial lineages containing flagella. Importantly, flagella were unknown from Rickettsiales until the recent genomes of *M. mitochondrii* and *O. thessalonicensis* revealed their presence (Georgiades et al. 2011; Sassera et al. 2011), with a recent study demonstrating the expression of several *M. mitochondrii* flagellar components (Mariconti et al. 2012). The putative presence of these three genes in RETA suggests that flagella may be common among rickettsial species outside of the traditional Rickettsiaceae and Anaplasmataceae.

The 18 CDS with divergent rickettsial homologs have a mean %GC higher than rickettsial genomes (39%) (fig. 6). Oddly, two (*trpS* and *rnhA*) were determined to be identical to sequences from the genome of the anaerobic thermohalophile *Halothermothrix orenii* H 168 (Firmicutes: Haloanaerobiales) (Mavromatis et al. 2009). As both of these CDS are on the same *T. adhaerens* scaffold (NW_002061683) and the genes in the *Hal. orenii* genome are neighbors, these CDS are likely contamination. For the other 16 CDS, phylogeny estimations placed them in clades distinct from the Rickettsiales homologs (supplementary fig. S7, Supplementary Material online). Although no dominant phylogenetic signal from any taxonomic group was observed, two genes (*cox3* and *rsmE*) and one gene fusion (*fkpA-trmH*) have clear chlamydial origins. Despite *fkpA* and *trmH* being adjacent in many chlamydial genomes, no chlamydial 16S rDNA was found in the *T. adhaerens* trace reads, casting doubt on the presence of a chlamydial species within the assembly. Encoding a chloroplast-targeted RpsP, *rpsP* was determined to be of algal (Haptophyceae) origin, consistent with the mining of cyanobacterial-like (chloroplast) 16S rDNA sequences (fig. 2*a*). Thus, it is likely that a small amount of algal contamination was present in the material used for sequencing, possibly originating from the food source of the cultured *T. adhaerens* laboratory colony. RETA proteins for the remaining 11 CDS (CcoP, PleD_2, MhpC, GNAT_1–2, PhrB_1–4, TraU, and PotA) grouped with various different bacterial species in the estimated trees, none of which had a clear association with the 16S rDNA sequences mined from the *T. adhaerens* trace reads.

Finally, the 17 CDS unknown from Rickettsiales had best scoring BLASTP hits to a wide range of bacterial species, most of which had no had clear associations with the extracted 16S rDNA sequences. However, three CDS (COG1506, hypothetical protein RETA0107 and PleD_1) had high similarity to sequences from aquatic Alteromonadales genomes, with the latter two proteins having 99% identity with sequences from the genome of *Alteromonas macleodii* ATCC 27126, a marine planktonic copiotroph (Ivars-Martinez et al. 2008). This suggests likely contamination of the *T. adhaerens* assembly with at least one species of aquatic Alteromonadales, given the mining of three 16S rDNA sequences from this group (fig. 2*a*). Of note in this set, four sequences of a transposase (ISL3) of cyanobacterial origin, and a protein-encoding SEL1 and other tetratricopeptide (TPR) motifs (COG0790), are strong candidates for being components of the RETA accessory genome. Several larger rickettsial genomes (e.g., REIS, *R. massiliae*, *R. felis*, and *R. bellii*) have accessory genomes comprised of transposases of diverse origins, including Cyanobacteria (Ogata et al. 2005, 2006; Blanc et al. 2007; Gillespie et al. 2012). Furthermore, bacterial proteins encoding eukaryotic domains (i.e., SEL1 and TPR motifs) are

Fig. 4.— Continued

for 347 *T. adhaerens-M. mitochondrii* protein matches, with S_m scores from 20 (dark blue) to 576 (burgundy) and *E* values from 1 (dark blue) to 1.00E−500 (burgundy); 3) histograms depicting the number of contigs on each scaffold that contain the identified *T. adhaerens* gene: eukaryotic-like CDS (black), RETA CDS of core data set (orange), RETA CDS of accessory data set (blue); 4) all 347 *T. adhaerens* CDS (outer, black) and 138 RETA CDS (inner, orange, blue) (supplementary fig. S3, Supplementary Material online, for linear histogram and further information). NOTE: five yellow CDS (outer) were below the S_m 20 cutoff but were determined to be RETA CDS via manual inspection. RETA CDS present on the same *T. adhaerens* scaffold are linked in the interior of the plot, with boxes (1–7) depicting syntenic regions across *M. mitochondrii* and RETA. Plot made using Circos (Krzywinski et al. 2009) with manual adjustment. (*b*) List of 181 RETA CDS identified within the *T. adhaerens* assembly. RETA identifier (0001–0181) followed by gene symbol or predicted product description (complete annotations in supplementary table S2, Supplementary Material online). Core data set CDS (orange) comprise 119 ORFs corresponding to 116 genes, with three split genes (dashed boxes). Accessory data set CDS (blue) comprise 62 genes. Black circles depict RETA CDS with homologs present in the *M. mitochondrii* genome ($n = 138$), and are listed according to their clockwise arrangement in ring 4 of the plot in (*a*). Yellow circles depict the five genes added manually ($S_m < 20$) to ring 4. Green boxes enclose the seven syntenic regions illustrated in the interior on the plot. Open circles depict the 42 RETA CDS that do not have significant homologs in the *M. mitochondrii* genome. Red asterisks denote six CDS that were subsequently determined to be likely nuclear encoded mitochondrial genes (see text). Red asterisks also mark the location of these CDS in (*a*) between rings 2 and 3.

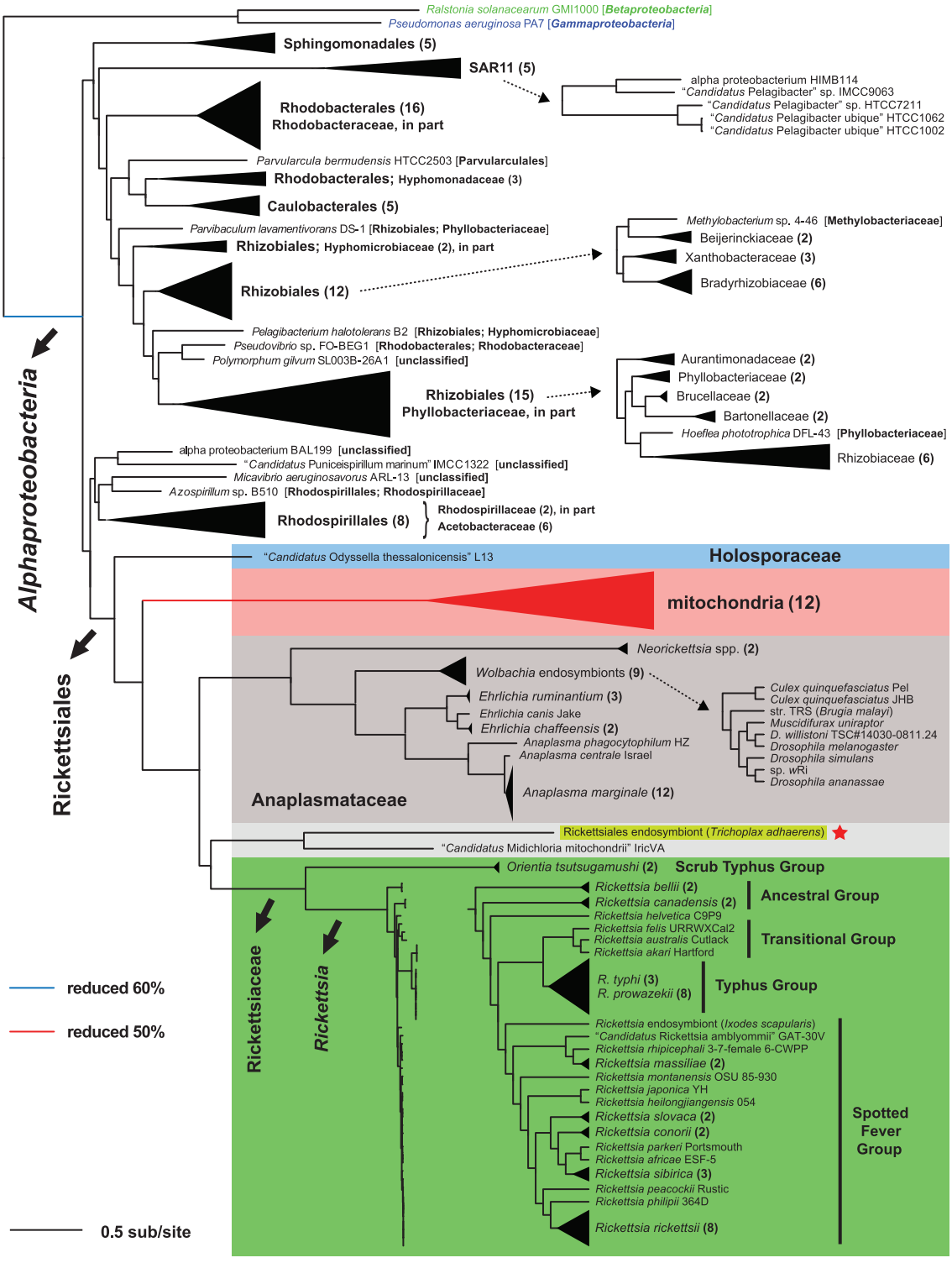


Fig. 5.—Genome-based phylogeny estimated for RETA, 162 alphaproteobacterial taxa, 12 mitochondria, and 2 outgroup taxa. RETA core proteins ($n = 113$) were included in the phylogenetic pipeline that entails ortholog group (OG) generation, OG alignment (and masking of less conserved positions), and concatenation of aligned OGs (see text). Tree was estimated using the CAT-GTR model of substitution as implemented in PhyloBayes v3.3 (Lartillot and Philippe 2004, 2006). Tree is a consensus of 1,522 trees (post burn-in) pooled from two independent Markov chains run in parallel. Branch support was measured via posterior probabilities, which reflect frequencies of clades among the pooled trees. RETA is boxed green and noted with a red star. Classification scheme for *Rickettsia* spp. follows previous studies (Gillespie et al. 2007, 2008). Taxon names, PATRIC genome IDs (bacteria) and NCBI accession numbers (mitochondria) for the 176 genomes are provided in supplementary table S3, Supplementary Material online.

Table 1

Comparison of Sequence Divergence across RETA and Rickettsiales Genera[a]

	RETA	Midi	Odys	Neor	Wolb	Ehrl	Anap	Orie	Rick
RETA		*0.47*	0.56	0.66	0.61	0.60	0.64	0.60	0.54
Midi	*0.45*		0.47	0.60	0.53	0.53	0.55	0.53	0.46
Odys	0.54	0.45		0.58	0.52	0.53	0.54	0.51	0.44
Neor	0.63	0.58	0.56		0.55	0.55	0.56	0.63	0.59
Wolb	0.58	0.51	0.50	0.53		0.36	0.40	0.55	0.52
Ehrl	0.58	0.51	0.51	0.53	0.36		0.30	0.54	0.51
Anap	0.61	0.53	0.52	0.54	0.39	0.30		0.58	0.54
Orie	0.57	0.51	0.50	0.60	0.53	0.52	0.56		0.39
Rick	0.52	0.45	0.43	0.56	0.50	0.49	0.52	0.39	

NOTE.—RETA, Rickettsiales endosymbiont of *Trichoplax adhaerens*; Midi, "*Candidatus* Midichloria mitochondrii" str. IricVA; Odys, "*Candidatus* Odyssella thessalonicensis" str. L13; Neor, *Neorickettsia risticii* str. Illinois; Wolb, *Wolbachia* endosymbiont str, TRS of *Brugia malayi*; Ehrl, *Ehrlichia ruminantium* str. Gardel; Anap, *Anaplasma phagocytophilum* str. HZ; Orie, *Orientia tsutsugamushi* str. Ikeda; Rick, *Rickettsia bellii* str. RML369-C.

[a]Calculated % divergence (8,327 aa sites of core data set) with DIVEIN (Deng et al. 2010), using the WAG (top right of matrix) and Blosum62 (bottom left of matrix) amino acid substitution models. Color scheme as follows: light yellow, <35 (% divergence); yellow, 36–40; tan, 41–45; light orange, 46–50; orange, 51–55; dark orange, 56–60; red, >60. Values for the closest taxon to RETA (*Midichloria*) are in italics.

characteristic of a variety of intracellular species, some of which show evidence for extensive LGT with rickettsial genomes (Schmitz-Esser et al. 2010; Penz et al. 2012). Importantly, aside from the probable contamination from algal, Alteromonadales and *H. orenii* genomes, most of the CDS either lacking or having divergent rickettsial homologs should be considered as possible components of the RETA accessory genome, given that diverse elements are known to be encoded within rickettsial genomes (e.g., the aminoglycoside antibiotic biosynthesis cluster of REIS (Gillespie et al. 2012)).

Bacterial Genes in the *Trichoplax* Genome

To determine whether any of the 181 RETA CDS might instead represent lateral transfers into the *T. adhaerens* genome, we evaluated the assembly scaffolds that contained these CDS (fig. 7*a*). A total of 79.5% ($n = 144$) of the RETA CDS were present on either all-bacteria ($n = 59$) or singleton ($n = 85$) scaffolds. These CDS were removed from consideration, because LGT events could only be predicted on scaffolds with eukaryotic-like genes. Assessment of the non-RETA genes on small hybrid scaffolds did not reveal any strong candidates for complete eukaryotic-like genes, being comprised mostly of small *T. adhaerens* ORFans and partial sequences with little or no similarity to sequences in the NR database (data not shown). Thus, these CDS ($n = 19$) were likewise removed from consideration as LGT events. In further support of the earlier mentioned, none of the intron-containing RETA genes within these three small scaffold categories were determined to contain plausible introns using a bacterial gene prediction program (supplementary table S2, Supplementary Material online). Of the four scaffold categories we created, only 10% ($n = 18$) of the total RETA CDS were found on large scaffolds dominated by eukaryotic-like genes and were thereby amenable to LGT analysis.

The eight large eukaryotic-like scaffolds that include the 18 RETA CDS that we analyzed for LGT properties are major components of the *T. adhaerens* assembly, containing from 41 to 1,541 genes (fig. 7*b*). Phylogeny estimations for all 18 RETA proteins (supplementary fig. S8, Supplementary Material online) clearly indicate that six of the CDS (*leuS*, *mutS*, *tilS*, *rpmA*, *prfA*, and *rplQ*) are eukaryotic genes that were erroneously obtained by our pipeline. Five of these genes (*leuS*, *mutS*, *tilS*, *rpmA*, and *prfA*) grouped with eukaryotic genes encoding counterparts that are predicted to be imported by the mitochondria, with the latter four having rickettsial origins. This explains their initial characterization as RETA genes. Bacterial gene predictions across the scaffold regions encoding all of these genes, coupled with manual assessment, determined that the introns within all six genes are valid. In addition, homologs to all of these genes were identified in MitoCarta, a compendium of mammalian nuclear-encoded genes with strong support for mitochondrial localization (Pagliarini et al. 2008). When we excluded these six genes from the RETA core data set and re-estimated the genome phylogeny (as described earlier), the resultant tree (data not shown) had an identical topology to the one including these genes (fig. 5), suggesting the minimal eukaryotic signal within the RETA core data set did not override the predominantly rickettsial signal.

Phylogeny estimations for the remaining 12 RETA CDS present on large eukaryotic-like scaffolds strongly implied that all of these genes are bacterial-like and present among eukaryotic genes within the *T. adhaerens* genome (supplementary fig. S8, Supplementary Material online). If they are indeed LGTs, 10 of these genes are likely recent transfers, as they

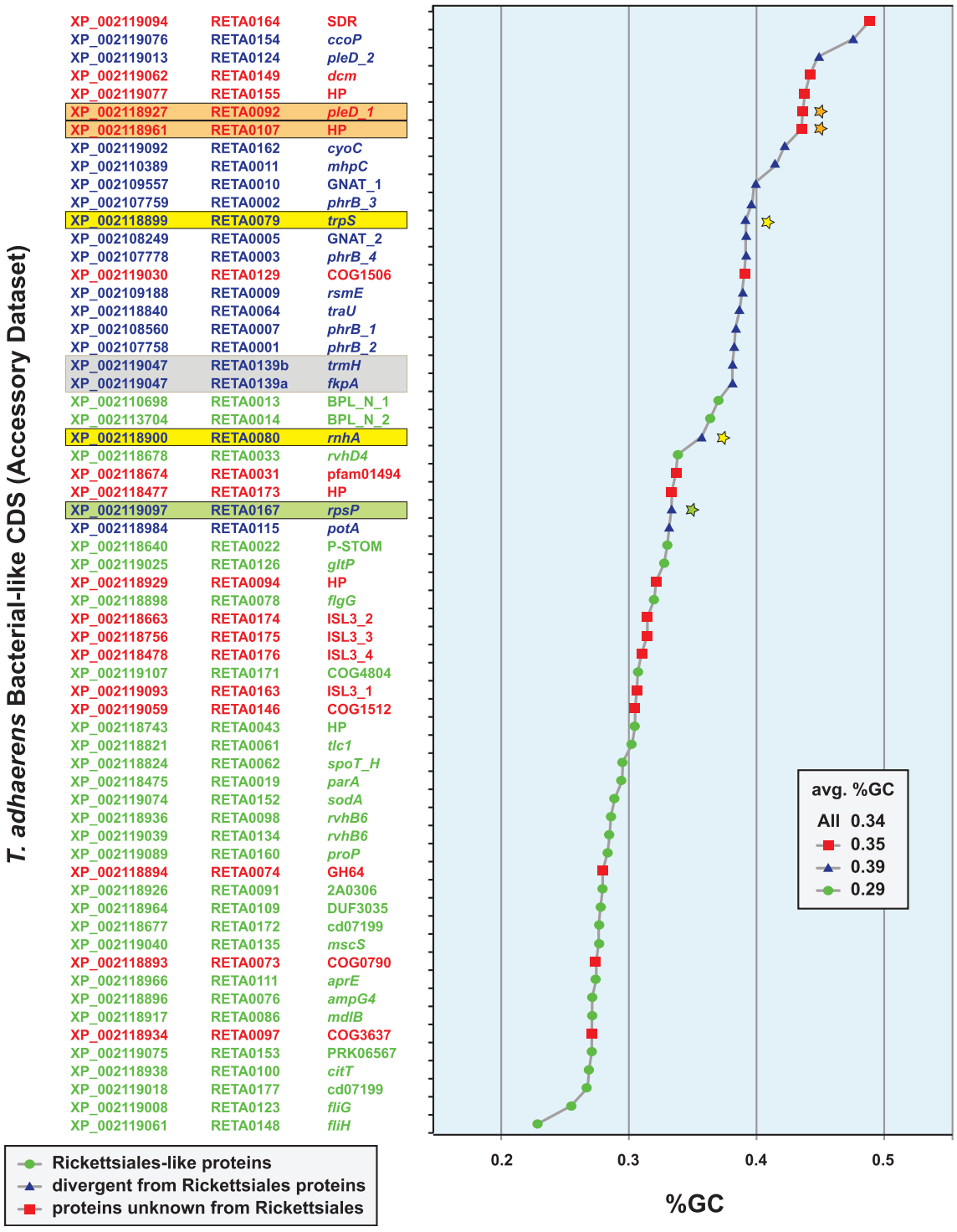
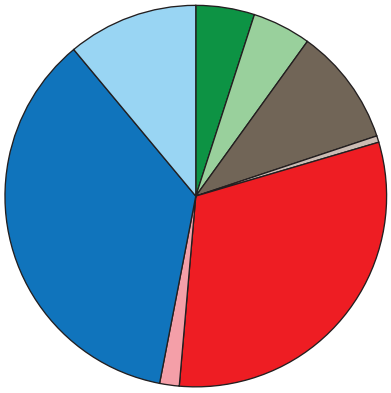


Fig. 6.—Bacterial CDS (accessory data set) identified within the *Trichoplax adhaerens* genome assembly. These 62 CDS were determined to lack the profile of typical Rickettsiales genes inherited vertically from an alphaproteobacterial ancestor (see text). RETA CDS are plotted by %GC (*x* axis). *Trichoplax adhaerens* protein accession numbers (NCBI), RETA IDs and gene/protein names are listed on the *y* axis, with color scheme as follows: green, highly similar to Rickettsiales signature proteins (trees shown in supplementary fig. S6, Supplementary Material online); blue, present in some (or all) Rickettsiales genomes yet divergent in sequence and phylogenetic signal (trees shown in supplementary fig. S7, Supplementary Material online); red, unknown from Rickettsiales genomes. Inset shows the average %GC for all 62 CDS, as well as for the three groups. Stars depict the following: yellow, identical to sequences from the genome of *Halothermothrix orenii* H 168 (Firmicutes: Haloanaerobiales); orange, 99% aa identity with sequences from the genome of *Alteromonas macleodii* ATCC 27126 (*Gammaproteobacteria*: Alteromonadales); green, most similar to chloroplast sequences of haptophytic algae (Eukaryota; Haptophyceae). Colored boxes on the *y*-axis correspond with stars on the plot, with the gray box illustrating a fused gene model (*trmH-fkpA*).

(a)



Assembly features	1 exon	introns
Eukaryotic scaffolds (*n* = 18)	9	9
Small hybrid scaffolds (*n* = 19)	18	1
All-bacteria scaffolds (*n* = 59)	56	3
Singleton scaffolds (*n* = 85)	65	20

(b)

Scaffold ID/ No. genes	CDS (RETA ID)	Protein	Sub-dataset	No. exons	Coverage (*T. adhaerens*)	Discrepancy	Coverage (RETA)	Phylogeny
NW_002060943 1541	0001	PhrB_2	AD-B	1	1.00	-	0.93	A; Nitrospirae/Legionellales
	0002	PhrB_3	AD-B	1	1.00	-	0.95	A; Nitrospirae/Legionellales
	0003	PhrB_4	AD-B	1	0.80	N	0.96	A; Nitrospirae/Legionellales
	0004	LeuS	CD-B	12	0.06	N	0.56	B; mitochondria
	0005	GNAT_2	AD-B	1	1.00	-	0.98	C; Spirochaetes/Odyssella
	0006	MutS	CD-B	3	0.83	C	0.88	D; mitochondria
	0007	PhrB_1	AD-B	1	1.00	-	0.95	A; Nitrospirae/Legionellales
	0008	TilS	CD-R	7	0.15	C	0.44	E; mitochondria
NW_002060944 1091	0009	RsmE	AD-B	1	1.00	-	0.97	F; Chlamydiae
	0010	GNAT_1	AD-B	1	1.00	-	0.94	C; Spirochaetes/*Odyssella*
NW_002060945 1071	0011	MhpC	AD-B	1	1.00	-	0.98	G; Cyanobacteria
	0012	DapF	CD-R	3	0.66	C	0.99	H; Rickettsiales
	0013	BPL_N_1	AD-R	2	0.39	C	0.90	I; *Rickettsia* spp./Chlamydiae
NW_002060948 745	0014	BPL_N_2	AD-R	2	0.07	C	0.85	I; *Rickettsia* spp./Chlamydiae
NW_002060955 173	0015	RpmA	CD-B	3	0.00	-	0.00	J; Rickettsiales/mitochondria
NW_002060959 152	0017	PrfA	CD-E	7	0.00	-	0.00	K; mitochondria
NW_002060958 145	0016	MurA	CD-R	1	0.88	N	0.98	L; Rickettsiales
NW_002060975 41	0018	RplQ	CD-B	3	0.07	N	0.52	M; Eukaryota/Chloroflexi

Fig. 7.—Evidence for bacterial-like genes encoded in the *Trichoplax adhaerens* genome. (*a*) Division of the 181 RETA CDS into four categories based on the composition of their scaffolds: eukaryotic scaffolds, CDS present on large (>40 genes) scaffolds with predominately eukaryotic-like genes ($n = 18$); small hybrid scaffolds, CDS present on small (<7 genes) scaffolds with both bacterial- and eukaryotic-like genes ($n = 19$); all-bacteria scaffolds, CDS present on small (<5 genes) scaffolds comprised entirely of bacterial-like genes ($n = 59$); and singleton-gene scaffolds ($n = 85$). Each category is further divided into single exon genes and genes possessing one or more introns (as predicted within the original *T. adhaerens* assembly). (*b*) Eight large, eukaryotic-like *T. adhaerens* scaffolds contain 18 RETA CDS. Scaffold IDs and number of encoded genes are from the *T. adhaerens* assembly (see text). RETA IDs and protein names are further described in supplementary table S2, Supplementary Material online. For core data set CDS: CD-R, CD-B, and CD-E correspond to the sub-data sets Ric-78, Bac-26, and Euk-9, respectively (supplementary fig. S2, Supplementary Material online). For accessory data set CDS: AD-R, highly similar to Rickettsiales signature proteins; AD-B, present in some (or all) Rickettsiales genomes yet divergent in sequence and phylogenetic signal (fig. 6). The number of

(continued)

do not contain introns or other eukaryotic-like features (e.g., eukaryotic secretion signals). *dapF* was determined to encode a complete DapF-like protein when predicted as a bacterial gene, ruling out its two predicted introns. However, the single introns splitting the two coding regions within BPL_N_1 and BPL_N_2 are localized within conserved sites as compared with closely related BPL_N sequences (supplementary fig. S9, Supplementary Material online). The two BPL_N proteins are divergent from one another (45% amino acid identity), yet form a clade together with proteins encoded in *Rickettsia* and Chlamydiae genomes (supplementary fig. S8, Supplementary Material online). Furthermore, the sizes of both proteins (BPL_N_1, 267 aa; BPL_N_2, 264 aa) are consistent with bacterial BPL_N proteins rather than the larger eukaryotic proteins that include a BirA-like domain in conjunction with the N-terminal region solely encoded by prokaryotic BPL_N proteins. Thus, the genes encoding BPL_N_1 and BPL_N_2 appear to be LGT products from bacteria that are undergoing transitions to eukaryotic-like gene structures.

Aside from the BPL_N proteins, genes encoding DapF and MurA also appear to be LGT products with rickettsial origins. Thus, all four of these *T. adhaerens* genes may be transfers from RETA and involved in maintenance of the symbiosis. The gene encoding rRNA small subunit methyltransferase E (RsmE) is clearly of chlamydial origin, and a strong candidate for LGT since the only eukaryotic-like *rsmE* genes encode products shipped to chloroplasts. The remaining candidate LGT products encode DNA photolyase repair enzymes (PhrB_1–4), GCN5-related N-Acetyltransferases (GNAT_1–2) and a bacterial lysophospholipase (MhpC). The phylogeny estimations for all of these proteins suggest a substantial degree of LGT underlying their distribution across bacteria, with several intracellular species grouping close to the RETA proteins. Collectively, our analysis of potential bacteria-to-*T. adhaerens* LGT events yielded 12 RETA genes that are clearly bacterial in origin and apparently encoded *within* the *T. adhaerens* genome.

Discussion

For a variety of reasons (e.g., contamination, failure to purify the target organism from environmental microbiota, low-level capture of endosymbionts, LGT), eukaryotic genome sequencing projects often include DNA sequences from the non-target organism. Contigs and scaffolds are typically filtered from the assembly if there is evidence for organellar contamination, sequencing artifacts (e.g., bacterial and phage cloning vectors), and/or prokaryotic contamination. These sequences remain available within the trace read archives, and several studies have utilized these resources to assemble bacterial genomes and associate eukaryotes with their resident microbes (Salzberg et al. 2005b, 2009). Still, other studies have capitalized on the concomitant sequencing of host and associated microbes and reported the presence of these microbes in conjunction with the eukaryotic sequencing project (Chapman et al. 2010; Gillespie et al. 2012). Whatever the route for identification of microbial DNA generated via eukaryotic genome sequencing projects, it is clear that it is no longer a rare event, and that methods are needed to facilitate the process of sifting out "who is who" amongst the generated sequence data. These methodological approaches will not only be applicable for identifying and analyzing microbial species from eukaryotic genome sequencing projects but also be practical for effectively processing data from metagenome and microbiotic studies (Iverson et al. 2012).

Mining Bacterial Genes from the *Trichoplax* Genome

As an aquatic animal, *T. adhaerens* is known to feed on green algae (Chlorophyta), cryptomonad (Cryptophyta) species of the genera *Cryptomonas* and *Rhodomonas*, Cyanobacteria, and detritus from other organisms (Schierwater and Kuhn 1998; Schierwater 2005). The "Grell" strain of *T. adhaerens*, which was fed a monoculture of the cryptophyte alga *Pyrenomonas helgolandii*, was the source for genome sequencing (Srivastava et al. 2008). Despite the effort to purify tissues prior to genome sequencing, the results of our study show that a small, yet detectable portion of the generated sequence data was from a different source(s) than *T. adhaerens*. This is not surprising, as the animals were not cultured axenically and may be associated with other free-living organisms in culture.

Our analysis of the trace read archive and genome assembly revealed considerable microbial diversity associated with the *T. adhaerens* sequencing project (fig. 2). According to the mined 16S rDNA sequences, many of the organisms with the highest sequence similarity are aquatic and probably inhabit similar niches as *T. adhaerens* (e.g., *Marivita* spp., *Alteromonas* spp, haptophyte and cryptomonad algae, heterokonts). Other 16S rDNA sequences extracted from the trace read archive (e.g., *Borrelia* spp., *Lawsonia intracellularis*) are harder to explain as environmental, and probably do not reflect any direct

Fig. 7.— Continued

exons for each CDS is shown. The results of gene predictions by fgenesb (Tyson et al. 2004) (headings for three columns colored green) are described as follows: "Coverage (*T. adhaerens*)," percentage of bps in the eukaryotic gene prediction matching those in the fgenesb prediction; "Discrepancy," differences at either the N- or C-terminus across eukaryotic and fgenesb predictions; "Coverage (RETA)," percentage of bps in the fgenesb prediction matching those in the eukaryotic gene prediction. The most related sequences as determined by phylogeny estimation are listed, with letters referring to individual phylogeny estimations (supplementary fig. S8, Supplementary Material online). Potential bacteria-to-*T. adhaerens* LGT products are highlighted in yellow.

biological associations with *T. adhaerens*. Their presence suggests a low level of microbial contamination, as we did not find any substantial evidence for matching CDS in the assembly or trace read archive for these organisms. Notwithstanding, this minimal information may prove useful in future studies of placozoan biology, particular regarding ecological interactions, and it may also be important for testing the interpretations that we have presented here.

Importantly, the 181 bacterial-like CDS we obtained from the *T. adhaerens* sequencing project were identified from the published assembly. According to the inclusion criteria from the original study (Srivastava et al. 2008), scaffolds were removed from the assembly if they 1) were shorter than 1 kb in total length, 2) were suspected to be organellar contaminants, and/or 3) possessed a distinct GC content or a prevalence of BLASTN alignments to prokaryotic genomes. In our analysis, 88.4% ($n = 160$) of bacterial-like CDS mined from the assembly were on small scaffolds (mean length of 2.7 kb) just over the exclusion cutoff, with the remaining 21 CDS on much larger scaffolds. Only one CDS (a second copy of *rpsP*) was determined likely to originate from chloroplasts of haptophytic algae (Eukaryota; Haptophyceae). Finally, the mean %GC of all 181 CDS (31.7) was highly similar to that of the *T. adhaerens* genome (32.7), masking their detection as distinct non-*T. adhaerens* genes. Collectively, these characteristics of the identified bacterial-like CDS likely account for their inclusion in the *T. adhaerens* assembly.

Evidence for a Rickettsiales Endosymbiont of *T. adhaerens*

The rickettsial 16S rDNA sequence mined from the *T. adhaerens* sequence read archive was determined to be most similar to sequences isolated from the marine sponge *Cymbastela concentrica* (99% nt identity). Other highly similar 16S rDNA sequences have been reported from *H. oligactis*, diverse coral species and unknown hosts from environmental samplings (Fraune and Bosch 2007; Revetta et al. 2010, 2011; Sunagawa et al. 2010), and all of these sequences grouped in a clade outside of the well-studied rickettsial families Rickettsiaceae and Anaplasmataceae in our estimated phylogeny (fig. 3). A recent study on the microbiome of the euglenoid alga *Eutreptiella* sp. revealed the presence of a rickettsial symbiont (Kuo and Lin 2013) that is also a member of this clade (tree not shown; 97% nt identity to RETA 16S rDNA, NCBI accession no. JQ337869). Thus, RETA belongs to a poorly known rickettsial lineage comprised of species associated with eukaryotic organisms from aquatic environments, consistent with the marine niche of *T. adhaerens*. Importantly, this lineage is divergent from other members of the "Midichloriaceae," including *M. mitochondrii*, and suggests that a substantial amount of diversity and host range underlay this large rickettsial assemblage (Gillespie, Nordberg, et al. 2012).

The identification of a rickettsial 16S rDNA sequence was important for corroborating both the long-known presence of a Gram-negative intracellular symbiont in *T. adhaerens* (Grell 1972; Grell and Benwitz 1974; Eitel et al. 2011) and the previous detection of rickettsial-like CDS associated with this genome (Felsheim et al. 2009; Baldridge et al. 2010; Gillespie et al. 2010; Nikoh et al. 2010). It also steered the approach for analyzing the *T. adhaerens* assembly with the understanding that this rickettsial species was likely divergent from those with available genome sequences. The majority (81%, $n = 146$) of the 181 CDS mined from the *T. adhaerens* assembly indeed showed a consistent rickettsial signal. We observed that *M. mitochondrii* was the most similar rickettsial species to RETA with an available genome sequence, and subsequently mapped 138 of the total RETA CDS to the *M. mitochondrii* chromosome (fig. 4). In fact, the publication of the *M. mitochondrii* genome (Sassera et al. 2011) was an invaluable resource for effectively mining the RETA genes from the *T. adhaerens* assembly, as attempts prior to the availability of this genome yielded far fewer genes of interest. Principally, the strategy to mine rickettsial-like CDS from the *T. adhaerens* assembly was plagued by the divergent nature of the bacterial-like CDS, and in many instances (e.g., universal proteins and nuclear-encoded organellar proteins) it was difficult to discern between top BLASTP subjects in the eukaryotic and prokaryotic databases. Thus, a thorough manual component of our methodology was invoked, coupled with rigorous phylogeny estimation of all the CDS. Calculating an approximation of the genome divergence across RETA and the major Rickettsiales lineages illustrated that the observed divergence of RETA genes, in relation to other lineages, is consistent for rickettsial genomes (table 1). Collectively, the divergent nature of the mined RETA genes is in agreement with the SSU rDNA-based phylogeny that suggests the RETA-containing group of aquatic Rickettsiales is a well-diverged clade within the "Midichloriaceae."

A majority (66%, $n = 119$) of the RETA CDS, named the core data set, contained characteristics of vertically inherited alphaproteobacterial genes; as such, their predicted proteins were used to estimate a robust genome-based phylogeny across *Alphaproteobacteria* (fig. 5). This phylogeny estimation grouped RETA with *M. mitochondrii*, to the exclusion of the derived rickettsial families Anaplasmataceae and Rickettsiaceae. Without genome sequences for any other members of the "Midichloriaceae," coupled with the limited data identified for RETA, it is difficult to make conclusions regarding the factors that distinguish this group from the other derived Rickettsiales lineages. However, our genome-based phylogeny estimation did agree with the overall higher level divergences within Rickettsiales based on the SSU rDNA tree (fig. 3), especially regarding the ancestral position of Holosporaceae and the branching point of the mitochondrial ancestor prior to the diversification of the Anaplasmataceae, Rickettsiaceae, and "Midichloriaceae."

The uniqueness of Rickettsiales within the *Alphaproteobacteria*, particularly regarding the relatedness of Holosporaceae, will become clearer with the generation of more genome sequences from this poorly understood taxon. Notwithstanding, the clade containing RETA and its aquatic relatives will provide useful genomic information regarding the diversification of the three derived rickettsial families, especially regarding the evolution of vertebrate pathogenicity from a seemingly vast array of obligate intracellular symbionts of virtually every major eukaryotic lineage (Gillespie, Nordberg, et al. 2012).

The remaining 62 identified CDS, named the accessory data set (fig. 6), provided additional evidence for a rickettsial symbiont associated with *T. adhaerens*. Nearly half (43.5%, $n = 27$) of these genes are signatures of all or some rickettsial genera (supplementary table S5, Supplementary Material online), and phylogeny estimation unambiguously supports the rickettsial nature of these RETA CDS (supplementary fig. S6, Supplementary Material online). In conjunction with the genes of the core data set, a profile emerged implicating the probable metabolic dependency of RETA on eukaryotic cells; for example, the presence of genes for the uptake of host ATP (*tlc1*), carbohydrates (*citT*) and amino acids (2A0306, *proP*, *gltP*). Other rickettsial signature genes encode proteins involved in the establishment of osmoregulation (*proP*), antioxidant defense (*sodA*), regulation of the stringent response (*spoT*), and peptidoglycan recycling/β-lactamase induction (*ampG*). Aside from the six rickettsial hallmark genes described earlier that encode ParA-HicB, PRK06567, p-stomatin, and several flagellar proteins (FlgG, FliG, and FliH), two additional rickettsial signatures are noteworthy. First, a gene encoding a patatin phospholipase (cd07199) was identified that is highly similar to the rickettsial Pat1 proteins recently demonstrated to function in the invasion of host cells (Rahman MS, Gillespie JJ, Kaur S, Sears KT, Ceraul SM, Beier-Sexton M, Azad AF, unpublished data). Second, a gene was identified that encodes an MdlB-like transporter with unknown specificity that is present within the genomes of many intracellular bacterial species (Gillespie et al. 2012). We have previously identified many of these rickettsial signatures (e.g., Tlc, ProP, GltP, SpoT, and MdlB) as components of integrative conjugative elements (Gillespie et al. 2012). Their presence in RETA and phylogeny estimations provided in this study (supplementary fig. S6, Supplementary Material online) suggests that these genes are widely dispersed across Rickettsiales and are likely critical factors that orchestrate the obligate intracellular life cycle of these bacteria. Aside from broadening our perspective on rickettsial biology and genomics, particularly the range and nature of the rickettsial mobilome, these signatures will be useful for identifying Rickettsiales in future metagenomic, microbiome and environmental studies.

In relation to the phylogeny of Rickettsiales, two notable observations were made from the analysis of the accessory data set. First, several RETA genes were identified that encode components of the *rvh* T4SS (Gillespie et al. 2009, 2010); however, we did not find any of the *rvh* genes in the genome of *O. thessalonicensis*, suggesting that either the rickettsial lineages branching off after Holosporaceae acquired a P-T4SS, or the Holosporaceae have secondarily lost the P-T4SS. Second, regarding the flagella system genes identified for RETA, only *M. mitochondrii* and *O. thessalonicensis* are known to contain flagellar genes among Rickettsiales. This suggests that the two well-studied groups Anaplasmataceae and Rickettsiaceae have secondarily lost the requirement for flagella, moving effectively within and across eukaryotic cells without them. More genome sequences from the Holosporaceae and "Midichloriaceae" are needed to test these evolutionary scenarios for the gain and loss of P-T4SSs and flagella across Rickettsiales.

Finally, 56.5% ($n = 35$) of the accessory data set is comprised of CDS that are not rickettsial in origin, with 17 of these genes not having any significant homologs in Rickettsiales genomes. Some of these CDS, which are highly similar to sequences from algal, Alteromonadales and *H. orenii* genomes, hint at a low level of contamination within the *T. adhaerens* assembly. However, the majority of these mined CDS may depict genes of the RETA accessory genome, especially considering that none of them have corresponding 16S rDNA sequences extracted from the trace read archive. Additionally, these CDS do not encode proteins that expand the typical metabolic capacity of Rickettsiales genomes. The complete sequencing of the RETA genome will be essential for determining whether these sequences are indeed encoded within its genome; regardless, even excluding these tenuous CDS of the accessory data set, the rickettsial-like CDS of the accessory data set together with all CDS of the core data set provide substantial evidence for RETA as the intracellular denizen of *T. adhaerens* fiber cells.

Bacterial Genes Encoded in the *T. adhaerens* Genome

Our analysis of the 181 mined bacterial-like CDS revealed that 12 of these genes are present on large scaffolds primarily composed of eukaryotic-like genes (fig. 7). Phylogeny estimation of these LGT products suggests that four of these genes encode proteins (BPL_N_1, BPL_N_2, DapF, and MurA) with strong homology to rickettsial counterparts, possibly reflecting RETA transfers to the *T. adhaerens* genome (supplementary fig. S8, Supplementary Material online). Although the two BPL_N encoding genes have acquired introns (supplementary fig. S9, Supplementary Material online), they are likely not complementing any host deficiency related to biotin ligation that is associated with eukaryotic BirA proteins that carry the BPL_N domain. Three lines of evidence support this hypothesis. First, the genes only encode the N-terminal domain of BirA proteins, which is not the region responsible for ligating biotin to biotin-dependent enzymes. Instead, this domain has homology to type 1 glutamine amidotransferases (cd03144),

which function in the transfer of the ammonia groups of Gln residues to other substrates. Second, despite the presence of introns, both BPL_N genes encode predicted proteins similar in size (~250 aa) to bacterial-like BPL_N proteins, none of which contain a BirA domain. Finally, the *T. adhaerens* genome encodes a separate BirA gene (XP_002116544) located on a scaffold (NW_002060958) distinct from those encoding the BPL_N genes, with the encoded protein 44% identical to the human holocarboxylase synthetase. This suggests that *T. adhaerens* is capable of ligating biotin to biotin-dependent enzymes without the need for BPL_N proteins.

The presence in the *T. adhaerens* genome of two rickettsial genes encoding cell envelope synthesis enzymes (DapF and MurA) is reminiscent of aphid (LdcA, RlpA, and AmiD) and rotifer (Ddl) genomes, wherein such bacterial-like genes are functional (Gladyshev et al. 2008; Nikoh and Nakabachi 2009; Nikoh et al. 2010). In the case of the aphid, *ldcA*, *rlpA,* and *amiD* have acquired introns and (in some instances) eukaryotic signal sequences and are highly expressed in bacteriocytes that house its obligate symbiont *B. aphidicola*. This remarkable phenomenon likely restricts symbiont growth to these specific tissues (Nikoh et al. 2010). It is tempting to speculate that putative transfers of RETA *dapF* and *murA* genes to the *T. adhaerens* genome may operate in a similar fashion. Recent images of the *T. adhaerens* symbiont in oocytes (transferred from fiber cell extensions) (Eitel et al. 2011) show a Gram-negative cell envelope structure, suggesting the presence of peptidoglycan. Thus, tissue-specific expression of *murA* and *dapF* by *T. adhaerens* could limit the growth of RETA to the fiber cells, which are primarily where the symbiotic bacteria are observed (Grell 1972; Grell and Benwitz 1974).

Four of the eight remaining putative LGT products encode copies of a bacterial-like photolyase (PhrB_1–4) that may play a functional role in *T. adhaerens*, or the maintenance of RETA, or both. PhrB is a photoreactivation enzyme, functioning in the repair of pyrimidine dimers (Schul et al. 2002). Genes encoding these enzymes are frequently detected in marine metagenomic studies and are likely important factors of aquatic microbes inhabiting surface waters (DeLong et al. 2006; Frias-Lopez et al. 2008; Singh et al. 2009). Photolyase genes are known to be differentially gained and lost throughout bacterial evolution (Lucas-Lledo and Lynch 2009), and our phylogeny estimation suggests a substantial degree of LGT shaping the distribution of these genes in bacteria (supplementary fig. S8, Supplementary Material online). Furthermore, it was recently demonstrated that species of the SAR11 clade of *Alphaproteobacteria* encode some PhrB- and PhrB-like genes that are most similar to those encoded within cynaobacterial genomes (Viklund et al. 2012). Thus, transfer to the *T. adhaerens* genome of these genes from a non-rickettsial source could function to repair DNA damage to its resident symbiont, particularly if there is limited symbiont gene flow due to its vertical transmission. Alternatively, the PhrB proteins may function to benefit *T. adhaerens* in its aquatic niche if it is vulnerable to substantial UV light. It has been proposed that loss of photolyases in eukaryotic species would enhance deleterious mutation rates (Lucas-Lledo and Lynch 2009), and given that we did not detect any PhrB or related cryptochrome genes within the *T. adhaerens* genome (data not shown), the four bacterial-like *phrB* genes may be under strong selective pressure to repair damage to *T. adhaerens* genes.

The remaining four putative LGT products (GNAT_1, GNAT_2, MhpC, and RsmE) are not of rickettsial origin, and encode enzymes known solely from prokaryotes or plastids; consequently, it is difficult to envision these bacterial-like genes complementing functions lost in the *T. adhaerens* genome. They may, however, function in the maintenance of RETA, despite their various evolutionary sources, possibly representing parts of the RETA accessory genome that were laterally acquired and subsequently transferred to the *T. adhaerens* genome. There is precedence for such an event; for instance, most of the identified gene transfers to the aphid genome that support its gammaproteobacterial symbiont (*B. aphidicola*) are of rickettsial and not gammaproteobacterial origin (Nikoh et al. 2010).

Whatever the source of the 12 identified bacterial-like genes encoded in the *T. adhaerens* genome, their significance in placozoan biology and possible role they may play in fostering its symbiosis with RETA is an exciting area of future research. Importantly, if LGTs to the *T. adhaerens* genome support the maintenance of a rickettsial symbiont, it may be that growth of RETA in other eukaryotic cells or cell lines may not be possible without these *T. adhaerens* bacterial genes.

Conclusion

From the genomic sequence data generated for a placozoan (*T. adhaerens*), we identified and analyzed bacterial DNA sequences and compiled evidence for a rickettsial endosymbiont of *T. adhaerens*. This genomic profile of RETA potentially confirms the long suspected presence of a bacterial symbiont associated primarily with *T. adhaerens* fiber cells. Based on available genome sequences for Rickettsiales, all of which are either obligate intracellular symbionts or pathogens of various metazoan species, the data we present here likely depicts approximately 20% of the entire RETA genome. Despite the lack of a complete genome, our phylogeny estimations and other analyses place RETA solidly within the "Midichloriaceae" clade. This rickettsial lineage is poorly understood, but based on the closest relative with an available genome sequence (*M. mitochondrii*), is quite different than the more than 80 genome sequences currently available for the well-studied species of the Anaplasmataceae and Rickettsiaceae. Thus, a better understanding of the "Midichloriaceae" clade of Rickettsiales is needed, particularly for 1) highlighting features involved in vertebrate pathogenicity in species of Anaplasmataceae and

Rickettsiaceae, 2) determining the diversifying factors that define obligate intracellular life cycles of a wide range of eukaryotic hosts, and 3) deciphering the processes that shaped the transition of a rickettsial symbiont to an organelle (mitochondria) of eukaryotic cells.

Placozoans are an early branching metazoan lineage (Miller and Ball 2008; Schierwater et al. 2009; Schierwater and Kamm 2010), and publication of the *T. adhaerens* genome sequence has proven invaluable to the fields of evolutionary genetics, developmental biology and animal phylogenetics, among others. However, given previous morphological evidence, as well as our data presented for RETA, the relevance of a symbiotic genome must be included in discussions and analyses of placozoan biology. Of primary importance is understanding whether RETA is primitive in gene repertoire compared with other rickettsial species that invade more complicated metazoan species with much more diverse cellular networks. Also critical is determining the function of bacterial-like genes encoded in the *T. adhaerens* genome, and whether these functions pertain to the placozoan itself or symbiosis with RETA (or both). Given the complex genetics underlying the aphid-*Buchnera* mutualism (discussed earlier) and other metazoan-bacterial symbioses (Woyke et al. 2006; Wu et al. 2006; McCutcheon and Moran 2007, 2010; McCutcheon et al. 2009; McCutcheon and von Dohlen 2011; Engel et al. 2012), as well as the recent demonstration of a tripartite metabolic codependency across genomes of an obligate biotrophic fungus (*Gigaspora margarita*), its plant hosts, and its beta-proteobacterial endosymbiont (*Candidatus* Glomeribacter gigasporarum) (Ghignone et al. 2012), it is clear that symbioses pose challenges for high-throughput genomics. Complete metabolic and cellular pathways can only be garnered by obtaining the genes from all participant genomes that underlay the complete "organism," particularly if some (or all) of the players are inextricably tied together. Thus, isolating RETA from *T. adhaerens* and determining the nature of this symbiosis is critical for understanding this aspect of placozoan biology.

Supplementary Material

Supplementary tables S1–S5 and figures S1–S9 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank members of the Cyberinfrastructure Division (Virginia Bioinformatics Institute at Virginia Tech) and the Azad Laboratory (School of Medicine, University of Maryland) for invaluable feedback and discussion throughout the duration of this project. They are grateful to Shrinivasrao Mane (Dow AgroSciences) for bioinformatics assistance in the preliminary phase of this work. They thank Bernd Schierwater and Tina Herzog (University of Veterinary Medicine Hannover Foundation) for providing literature on *T. adhaerens*, and Leo Buss (Yale University) for fielding inquiries about placozoan biology. This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), and Department of Health and Human Services grants HHSN272200900040C to B.W.S. and R01AI017828 and R01AI59118 to A.F.A. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAID or the NIH.

Note Added in Proof

During the production of this work, a recent publication (Montagna et al. 2013) formally classified a novel Rickettsiales family, Candidatus Midichloriaceae, for which the Rickettsiales endosymbiont of *Trichoplax adhaerens* is a member.

Literature Cited

Acuna R, et al. 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. Proc Natl Acad Sci U S A. 109: 4197–4202.

Aikawa T, et al. 2009. Longicorn beetle that vectors pinewood nematode carries many *Wolbachia* genes on an autosome. Proc Biol Sci. 276: 3791–3798.

Aziz RK, et al. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Baldridge GD, et al. 2010. Wide dispersal and possible multiple origins of low-copy-number plasmids in *Rickettsia* species associated with blood-feeding arthropods. Appl Environ Microbiol. 76:1718–1731.

Beninati T, et al. 2004. A novel alpha-Proteobacterium resides in the mitochondria of ovarian cells of the tick *Ixodes ricinus*. Appl Environ Microbiol. 70:2596–2602.

Birtles RJ, et al. 2000. "*Candidatus* Odyssella thessalonicensis" gen. nov., sp. nov., an obligate intracellular parasite of *Acanthamoeba* species. Int J Syst Evol Microbiol. 50(Pt 1):63–72.

Blanc G, et al. 2007. Lateral gene transfer between obligate intracellular bacteria: evidence from the *Rickettsia massiliae* genome. Genome Res. 17:1657–1664.

Boscaro V, Fokin SI, Schrallhammer M, Schweikert M, Petroni G. 2013. Revised systematics of *Holospora*-like bacteria and characterization of "*Candidatus* Gortzia infectiva," a novel macronuclear symbiont of *Paramecium jenningsi*. Microb Ecol. 65:255–267.

Boschetti C, et al. 2012. Biochemical diversification through foreign gene expression in bdelloid rotifers. PLoS Genet. 8:e1003035.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17: 540–552.

Chapman JA, et al. 2010. The dynamic genome of *Hydra*. Nature 464: 592–596.

Davis AK, et al. 2009. New findings from an old pathogen: intraerythrocytic bacteria (family Anaplasmatacea) in red-backed salamanders *Plethodon cinereus*. Ecohealth 6:219–228.

Dellaporta SL, et al. 2006. Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal lower metazoan phylum. Proc Natl Acad Sci U S A. 103:8751–8756.

DeLong EF, et al. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503.

Deng W, et al. 2010. DIVEIN: a web server to analyze phylogenies, sequence divergence, diversity, and informative sites. Biotechniques 48: 405–408.

DeSantis TZ, et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 72:5069–5072.
Dunning Hotopp JC, et al. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317: 1753–1756.
Durbin R, Eddy S, Krogh A, Mitchison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press.
Edgar RC. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.
Edgar RC. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Eitel M, Guidi L, Hadrys H, Balsamo M, Schierwater B. 2011. New insights into placozoan sexual reproduction and development. PLoS One 6: e19639.
Engel P, Martinson VG, Moran NA. 2012. Functional diversity within the simple gut microbiota of the honey bee. Proc Natl Acad Sci U S A. 109: 11002–11007.
Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8: 175–185.
Felsheim RF, Kurtti TJ, Munderloh UG. 2009. Genome sequence of the endosymbiont *Rickettsia peacockii* and comparison with virulent *Rickettsia rickettsii*: identification of virulence factors. PLoS One 4: e8361.
Fraune S, Bosch TC. 2007. Long-term maintenance of species-specific bacterial microbiota in the basal metazoan *Hydra*. Proc Natl Acad Sci U S A. 104:13146–13151.
Frias-Lopez J, et al. 2008. Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A. 105:3805–3810.
Georgiades K, Madoui MA, Le P, Robert C, Raoult D. 2011. Phylogenomic analysis of *Odyssella thessalonicensis* fortifies the common origin of Rickettsiales, *Pelagibacter ubique* and *Reclimonas americana* mitochondrion. PLoS One 6:e24857.
Ghignone S, et al. 2012. The genome of the obligate endobacterium of an AM fungus reveals an interphylum network of nutritional interactions. ISME J. 6:136–145.
Gillespie JJ, et al. 2007. Plasmids and rickettsial evolution: insight from *Rickettsia felis*. PLoS One 2:e266.
Gillespie JJ, et al. 2008. *Rickettsia* phylogenomics:unwinding the intricacies of obligate intracellular life. PLoS One 3:e2018.
Gillespie JJ, et al. 2009. An anomalous type IV secretion system in *Rickettsia* is evolutionarily conserved. PLoS One 4:e4833.
Gillespie JJ, et al. 2010. Phylogenomics reveals a diverse Rickettsiales type IV secretion system. Infect Immun. 78:1809–1823.
Gillespie JJ, et al. 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect Immun. 79:4286–4298.
Gillespie JJ, et al. 2012. A *Rickettsia* genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. J Bacteriol. 194:376–394.
Gillespie JJ, Nordberg EK, Azad AF, Sobral BW. 2012. Phylogeny and comparative genomics: the shifting landscape in the genomics era. In: Azad AF, Palmer GH, editors. Intracellular pathogens II: rickettsiales. Boston (MA): American Society of Microbiology. p. 84–141.
Gladyshev EA, Meselson M, Arkhipova IR. 2008. Massive horizontal gene transfer in *Bdelloid rotifers*. Science 320:1210–1213.
Grbic M, et al. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature 479:487–492.
Grell KG. 1971. *Trichoplax adhaerens*, F.E. Schulze und die Entstehung der Metazoen. Naturwiss Rundsch. 24:160–161.
Grell KG. 1972. Eibildung und Furchung von *Trichoplax adhaerens* F.E. Schulze (Placozoa). Z Morph Tiere. 73:297–314.
Grell KG, Benwitz G. 1974. Spezifische Verbindungsstruktnren der Faserzellen yon *Trichoplax adhaerens* F.Schulze E. Naturforsch Z. 29e:790.
Grell KG, Ruthmann A. 1991. Placozoa. In: Harrison FW, Westfall JA, editors. Microscopic anatomy of invertebrates. New York: Wiley-Liss. p. 13–28.
Guidi L, Eitel M, Cesarini E, Schierwater B, Balsamo M. 2011. Ultrastructural analyses support different morphological lineages in the phylum Placozoa Grell, 1971. J Morphol. 272:371–378.
Hinderhofer M, et al. 2009. Evolution of prokaryotic SPFH proteins. BMC Evol Biol. 9:10.
Ivars-Martinez E, et al. 2008. Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles associated with different kinds of particulate organic matter. ISME J. 2:1194–1212.
Iverson V, et al. 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335: 587–590.
Juncker AS, et al. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. 12:1652–1662.
Kall L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res. 35:W429–W432.
Kawafune K, Hongoh Y, Hamaji T, Nozaki H. 2012. Molecular identification of rickettsial endosymbionts in the non-phagotrophic volvocalean green algae. PLoS One 7:e31749.
Kembel SW, Wiu M, Eisen JA, Green JL. 2012. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS Comp Biol. 8:e1002743.
Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12: 656–664.
Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T. 2002. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. Proc Natl Acad Sci U S A. 99:14280–14285.
Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305: 567–580.
Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639–1645.
Kuo RC, Lin S. 2013. Ectobiotic and endobiotic bacteria associated with *Eutreptiella* sp. isolated from Long Island Sound. Protist 164: 60–74.
Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 7(1 Suppl):S4.
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 21:1095–1109.
Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst Biol. 55:195–207.
Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595.
Li ZW, Shen YH, Xiang ZH, Zhang Z. 2011. Pathogen-origin horizontally transferred genes contribute to the evolution of Lepidopteran insects. BMC Evol Biol. 11:356.
Lo N, Beninati T, Sacchi L, Genchi C, Bandi C. 2004. Emerging rickettsioses. Parassitologia 46:123–126.
Lucas-Lledo JI, Lynch M. 2009. Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. Mol Biol Evol. 26: 1143–1153.

Marchler-Bauer A, et al. 2011. CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res. 39: D225–D229.
Mariconti M, et al. 2012. On the presence of flagella in the Rickettsiales: the case of *Midichloria mitochondrii*. Microbiology 158(Pt 7): 1677–1683.
Matsuura Y, Kikuchi Y, Meng XY, Koga R, Fukatsu T. 2012. Novel clade of alphaproteobacterial endosymbionts associated with stinkbugs and other arthropods. Appl Environ Microbiol. 78:4149–4156.
Mavromatis K, et al. 2009. Genome analysis of the anaerobic thermohalophilic bacterium *Halothermothrix orenii*. PLoS One 4:e4192.
McCutcheon JP, McDonald BR, Moran NA. 2009. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc Natl Acad Sci U S A. 106:15394–15399.
McCutcheon JP, Moran NA. 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc Natl Acad Sci U S A. 104:19392–19397.
McCutcheon JP, Moran NA. 2010. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. Genome Biol Evol. 2:708–718.
McCutcheon JP, von Dohlen CD. 2011. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. Curr Biol. 21: 1366–1372.
McNulty SN, et al. 2010. Endosymbiont DNA in endobacteria-free filarial nematodes indicates ancient horizontal genetic transfer. PLoS One 5: e11029.
Miller DJ, Ball EE. 2008. Animal evolution: *Trichoplax*, trees, and taxonomic turmoil. Curr Biol. 18:R1003–R1005.
Montagna M, et al. 2013. "Candidatus Midichloriaceae" fam. n 1 ov. (Rickettsiales), an ecologically widespread clade of intracellular alphaproteobacteria. Appl Environ Microbiol. Advance Access published March 15, 2013, doi:10.1128/AEM.03971-12.
Nikoh N, Nakabachi A. 2009. Aphids acquired symbiotic genes via lateral gene transfer. BMC Biol. 7:12.
Nikoh N, et al. 2008. *Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. Genome Res. 18:272–280.
Nikoh N, et al. 2010. Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. PLoS Genet. 6: e1000827.
Ogata H, et al. 2005. The genome sequence of *Rickettsia felis* identifies the first putative conjugative plasmid in an obligate intracellular parasite. PLoS Biol. 3:e248.
Ogata H, et al. 2006. Genome sequence of *Rickettsia bellii* illuminates the role of amoebae in gene exchanges between intracellular pathogens. PLoS Genet. 2:e76.
Ondov BD, Bergman NH, Phillippy AM. 2011. Interactive metagenomic visualization in a web browser. BMC Bioinformatics 12:385.
Pagliarini DJ, et al. 2008. A mitochondrial protein compendium elucidates complex I disease biology. Cell 134:112–123.
Penz T, et al. 2012. Comparative genomics suggests an independent origin of cytoplasmic incompatibility in *Cardinium hertigii*. PLoS Genet. 8: e1003012.
Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8:785–786.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.
Revetta RP, Matlib RS, Santo Domingo JW. 2011. 16S rRNA gene sequence analysis of drinking water using RNA and DNA extracts as targets for clone library development. Curr Microbiol. 63:50–59.
Revetta RP, Pemberton A, Lamendella R, Iker B, Santo Domingo JW. 2010. Identification of bacterial populations in drinking water using 16S rRNA-based sequence analyses. Water Res. 44:1353–1360.
Rodriguez-Ezpeleta N, Embley TM. 2012. The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. PLoS One 7: e30520.
Sacchi L, et al. 2004. A symbiont of the tick *Ixodes ricinus* invades and consumes mitochondria in a mode similar to that of the parasitic bacterium *Bdellovibrio bacteriovorus*. Tissue Cell 36: 43–53.
Salzberg SL, et al. 2005a. Correction: serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila species*. Genome Biol. 6:402.
Salzberg SL, et al. 2005b. Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila species*. Genome Biol. 6:R23.
Salzberg SL, Puiu D, Sommer DD, Nene V, Lee NH. 2009. Genome sequence of the *Wolbachia* endosymbiont of *Culex quinquefasciatus* JHB. J Bacteriol. 191:1725.
Sassera D, et al. 2006. '*Candidatus* Midichloria mitochondrii', an endosymbiont of the tick *Ixodes ricinus* with a unique intramitochondrial lifestyle. Int J Syst Evol Microbiol. 56:2535–2540.
Sassera D, et al. 2011. Phylogenomic evidence for the presence of a flagellum and cbb(3) oxidase in the free-living mitochondrial ancestor. Mol Biol Evol. 28:3285–3296.
Schierwater B. 2005. My favorite animal, *Trichoplax adhaerens*. Bioessays 27:1294–1302.
Schierwater B, et al. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol. 7:e20.
Schierwater B, Kamm K. 2010. The early evolution of Hox genes: a battle of belief? Adv Exp Med Biol. 689:81–90.
Schierwater B, Kuhn K. 1998. Homology of Hox genes and the zootype concept in early metazoan evolution. Mol Phylogenet Evol. 9:375–381.
Schmitz-Esser S, et al. 2010. The genome of the amoeba symbiont "*Candidatus* Amoebophilus asiaticus" reveals common mechanisms for host cell interaction among amoeba-associated bacteria. J Bacteriol. 192:1045–1057.
Schul W, et al. 2002. Enhanced repair of cyclobutane pyrimidine dimers and improved UV resistance in photolyase transgenic mice. EMBO J. 21:4719–4729.
Singh AH, Doerks T, Letunic I, Raes J, Bork P. 2009. Discovering functional novelty in metagenomes: examples from light-mediated processes. J Bacteriol. 191:32–41.
Srivastava M, et al. 2008. The *Trichoplax* genome and the nature of placozoans. Nature 454:955–960.
Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst Biol. 57:758–771.
Sunagawa S, Woodley CM, Medina M. 2010. Threatened corals provide underexplored microbial habitats. PLoS One 5:e9554.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56:564–577.
Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. 2012. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol. 4:466–485.
Thrash JC, et al. 2011. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Sci Rep. 1:13.
Tyson GW, et al. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43.
UniProt Consortium. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 40: D71–D75.
Van Dongen S. 2008. Graph clustering via a discrete uncoupling process. SIAM J Matrix Anal Appl. 30:121–141.

Vannini C, et al. 2010. "*Candidatus* anadelfobacter veles" and "*Candidatus* cyrtobacter comes," two new rickettsiales species hosted by the protist ciliate *Euplotes harpa* (Ciliophora, Spirotrichea). Appl Environ Microbiol. 76:4047–4054.

Viklund J, Ettema TJ, Andersson SG. 2012. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. Mol Biol Evol. 29:599–615.

Werren JH, et al. 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science 327: 343–348.

Wilgenbusch JC, Swofford D. 2003. Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics. Chapter 6:Unit 6.4.

Williams KP, Sobral BW, Dickerman AW. 2007. A robust species tree for the alphaproteobacteria. J Bacteriol. 189:4578–4586.

Woyke T, et al. 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443:950–955.

Wu D, et al. 2006. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol. 4:e188.

Associate editor: Shu-Miaw Chaw