Simultaneous Single-Cell Genome and Transcriptome Sequencing of Termite Hindgut Protists Reveals Metabolic and Evolutionary Traits of Their Endosymbionts

Michael E. Stephens[1,2#], Jacquelynn Benjamino[1,3#], Joerg Graf[1], Daniel J. Gage[1*]

[1]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT USA
[2]Department of Entomology, Cornell University, Ithaca, NY USA
[3]The Jackson Laboratory for Genomic Medicine, Farmington, CT USA

#These two authors share first authorship. Both contributed equally to work described in the manuscript. Order was determined by which did the majority of the manuscript writing.

* corresponding author

Abstract

Different protist species which colonize the hindguts of wood feeding *Reticulitermes* termites are associated with endosymbiotic bacteria belonging to the genus *Endomicrobium*. In this study, we focused on the endosymbionts of three protist species from *Reticulitermes flavipes*, which included *Pyrsonympha vertens*, *Trichonympha agilis*, and *Dinenympha* species II. Since these protist hosts represented members of difference taxa which colonize different niches within the hindguts of their termite hosts, we investigated if these differences translated to differential gene content and expression in their endosymbionts. Following assembly and comparative genome and transcriptome analyses, we discovered that these endosymbionts differed with respect to possible niche specific traits such carbon metabolism. Our analyses supported that genes related to carbon metabolism were acquired by horizontal gene transfer (HGT) from donor taxa which are present in termite's hindgut community. In addition, our analyses supported that these endosymbionts have retained and expressed genes related to natural transformation (competence) and recombination. Taken together, the presence of genes acquired by HGT and a putative competence pathway supported that these endosymbionts are not cut-off from gene flow and that competence may be a mechanism by which members of the *Endomicrobium* can acquire new traits.

Importance

The composition and structure of wood, which contains cellulose, hemicellulose and lignin, prevents most organisms from using this common food source. Termites are a rare exception among animals, and they rely on a complex microbiome housed in their hindguts to use wood as a source of food. The lower termite *R. flavipes* houses a variety of protist and prokaryotes that are the key players the disassembly of lignocellulose. In this paper we describe the genomes and the gene expression profiles of five *Endomicrobium* endosymbionts living inside three different protist species from *R. flavipes*. Data from these genomes suggest that these *Endomicrobium* species have different mechanisms for using both carbon and nitrogen. In addition, they harbor genes that may be used to import free DNA from their environment. This process of DNA-uptake may contribute to the high levels of horizontal gene transfer often seen in the *Endomicrobium* species.

Introduction

Among the wood-feeding lower termites, symbiotic protists which reside in the hindgut are often colonized by endosymbionts (1–4). In *Reticulitermes* spp. termites both Oxymonadida (order) and Parabasilia (class) protists associate with endosymbiotic bacteria belonging to the genus *Endomicrobium* (phylum Elusimicrobia, class Endomicrobia) (2, 5–7). Members of *Endomicrobium* have been shown to comprise a significant portion of the core bacterial community in wood-feeding termites such as *R. flavipes* (8, 9). These endosymbiotic lineages are thought to have initiated their associations with hindgut protists approximately 70 - 40 million years ago (10) and arose from free-living relatives during multiple independent acquisition events (11). Vertical passage from one protist cell to its progeny, has resulted in co-speciation as inferred from congruent ribosomal RNA (rRNA) phylogenies (7, 10, 12, 13).

In addition to colonizing the cytoplasm of certain hindgut protist species, *Endomicrobium* spp. are ectosymbionts of protists (14) can be free-living as well (15)(11, 16, 17). Because of their distribution across these different niches, they provide an opportunity for studying bacterial genome evolution across different association lifestyles: free-living, endosymbiotic, and ectosymbiotic.

To determine differences two *Endomicrobium* species that are closely related but with distinct lifestyles, a previous study compared genomes of a free-living *Endomicrobium*, *E. proavitum* strain Rsa215 (16) isolated from *R. santonensis* (*R. flavipes*), and '*Candidatus* Endomicrobium trichonymphae' strain Rs-D17 (3), an endosymbiont isolated from the cytoplasm of a *Trichonympha* from the termite *R. speratus* (3, 18). The findings suggested that the transition from the free-living state to an intracellular lifestyle involved genome reduction, similar to that of endosymbionts of sap-feeding insects and many obligate intracellular pathogens. However, the intracellular strain Rs-D17 also incorporated genes, possibly from other termite gut inhabitants, by horizontal gene transfer (HGT) (18). For example, the genome of '*Ca*. E. trichonymphae' Rs-D17 appeared to have acquired several pathways including those that encode sugar and amino acid transporters and genes involved in amino acids biosynthesis (18). These findings suggested that, unlike the endosymbionts of sap-feeding insects, *Endomicrobium* species may not be completely cut-off from gene flow (18).

We expand upon these studies by presenting and comparing near-complete draft genomes and transcriptomes of three different *Endomicrobium* organisms, that were assembled from single protists cells of three different species that inhabit the hindgut of *R. flavipes*. One of these protists species, *Pyrsonympha vertens* , lives attached to the oxic gut wall (19, 20), while the other two *Trichonympha agilis* and *Dinenympha* species II, are found in the more anoxic hindgut lumen. In addition, *P. vertens* and *D.* species II are both Oxymonads while *T. agilis* is a Parabasalid.

The analyses indicate that these *Endomicrobium* have differences in gene content and expression, related to carbon usage and metabolism. And as seen previously in '*Ca*. E. trichonymphae' Rs-D17, they have likely acquired genes from putative donor taxa that are commonly associated with termites. In addition, we describe data suggesting that these *Endomicrobium* have retained competence genes which may allow them to import exogenous DNA and that perhaps have contributed to HGT. Genes involved in this pathway are conserved across several *Endomicrobium* species and were expressed in the endosymbionts examined in this study.

Methods

Termite collection and species identification. *R. flavipes* termites were collected using cardboard traps placed under logs for 2 to 4 weeks at the UConn Campus at Storrs, Connecticut (Longitude -72.262216, Latitude 41.806543) and their identity was verified as previously described (7) by amplifying and sequencing the mitochondrial cytochrome oxidase II gene. Termites were maintained in the lab with moistened sand and spruce wood that were initially sterilized.

Single protist cell isolation. Termites from the worker caste were brought into an anaerobic chamber and their hindguts were dissected with sterile forceps. Hindguts were ruptured in ice-cold Trager's Solution U (TU) (21) and washed three times by centrifuging in 500 ul of TU at 3,000 rpm in an Eppendorf microcentrifuge for 90 seconds. This washed cell suspension was then diluted 10-fold in TU buffer on ice. A 1 µl aliquot of the washed and diluted cell suspension was added to a 9 µl droplet on a glass slide treated with RNase AWAY® Reagent (Life Technologies) and UV light. Individual protist cells were isolated using a micromanipulator (Eppendorf CellTram® Vario) equipped with a hand-drawn glass capillary. Individual cells were washed three times in 10 µl droplets of TU via micromanipulation, transferring approximately 0.1 µl each time, and finally placed in 10µl molecular grade phosphate buffered saline (PBS), flash frozen on dry ice, and immediately stored at -80ºC. Meta-data regarding these protist cell samples can be found as Supplementary Table 1 in S1 File.

Whole genome and transcriptome amplification and sequencing. The metagenome (DNA) and metatranscriptome (cDNA) from individual protist cells and their associated bacteria were simultaneously amplified 12-24 hours after isolation. Cell lysis and amplification was performed using the Repli-g WGA/WTA kit (Qiagen). Cells were lysed using a Qiagen lysis buffer followed immediately by incubation on ice. Two samples from each lysed cell were taken and used in for whole genome amplification and whole transcriptome amplification. These were carried using the manufacturer's standard protocol with exception that random hexamer primers were used to amplify DNA and cDNA. DNA and cDNA were sheared using a Covaris M220 ultra-sonicator™ according to the manufacturer's protocol. WGA samples were sheared to a 550 bp insert size using 200 ng of DNA. WTA samples were sheared to a 350 bp insert size using 100 ng of cDNA. Sequencing libraries were prepared using the TruSeq Nano DNA Library Prep kit from Illumina® according to the manufacturer's protocol. Each sample was prepared with a forward and reverse barcode such that samples could be multiplexed on the same sequencing run. The samples were sequenced using an Illumina® NextSeq 1x150 mid-output run and two NextSeq 1x150 high-output runs. Meta-data regarding amplicon yields can be found in Supplementary Table 1 in S1 File.

Genomic read processing and assembly. Reads were preprocessed before assembly using BBmap (22). Reads were filtered for contaminating sequences by mapping reads to reference genomes of potential contamination sources such as human DNA, human associated microbiota, and organisms commonly used in our research laboratories. A list of references genomes used for contamination filtering is provided in Supplementary Table 2 in S1 File. Using BBmap scripts, adaptor sequences were trimmed from reads and last base pair of 151 bp reads was removed. Reads were then trimmed at both ends using a quality score cutoff of Q15. Homopolymers were removed by setting an entropy cutoff of 0.2, a max G+C cutoff of 90%, and by removing reads which possessed stretches of G's equal to or greater than 23 bases long. In addition, reads which were below a minimum average quality of Q15 and/or 50 bases long were removed. Genomic reads were then normalized to a minimum coverage of 2X and a maximum coverage of 50X and then deduplicated using BBnorm. Genomic reads were assembled using the A5 assembly pipeline (23) on the KBase web server (24). Meta-data regarding metagenome and metatranscriptome reads numbers can be found in Supplementary Table 3 in S1 File.

Genomic binning, draft genome assessment and annotation. Metagenomic assemblies from single protist host cells and their bacterial symbionts were binned using either 4mer or 6mer frequencies with VizBin (25) and scaffolds at least 1Kb in size. Clustered scaffolds in genomic bins of interest (low GC content) were selected in VizBin. Each scaffold from these bins were used in a blastn (26) search against previously sequenced Elusimicrobia genomes (Supplementary Table 4 in S1 File). Scaffolds which had a positive hit to other Elusimicrobia (at least 70% identity over a 1kb alignment) were retained in the draft genomes and scaffolds which did not have a significant hit to other Elusimicrobia genomes were used in a second blastn search against the non-redundant (NR) database. Scaffolds which had positive hits to other Elusimicrobia in the NR database were retained in the draft genomes. Draft genomes were iteratively polished with the program Pilon (27). These draft genomes were then assessed for contamination and completeness using CheckM which uses lineage specific marker genes to perform analyses (28). The resulting near-complete draft genomes were then annotated on the RAST Server using a customized RASTtk workflow with options selected to call insertion sequences and prophages (29, 30). Metabolic pathways pertaining to carbon metabolism, amino acid biosynthesis, vitamin biosynthesis, and peptidoglycan biosynthesis were reconstructed from the annotated genomes using pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (31).

Analysis of ribosomal gene phylogeny and average nucleotide identities. Ribosomal 16S genes from each of the *Endomicrobium* spp. draft genomes were trimmed and aligned to references using MUSCLE (32), evolutionary models were tested and a Maximum likelihood (ML) phylogenetic tree was made using IQ-TREE (33). JSpeciesWS (34) was used for determining the genomic average nucleotide identities based on BLAST+ searches (ANIb) between the *Endomicrobium* spp. draft genomes and the genome of '*Ca*. Endomicrobium trichonymphae' Rs-D17, which is a close relative (3).

Assembled 18S rRNA genes were retrieved from metagenome assemblies by performing a BLAST+ search using previously published 18S rRNA reference sequences for each protist species as queries (7). When possible, protist 18S rRNA genes were amplified using leftover DNA from WGA samples using universal primers 18SFU; 5'-ATGCTTGTCTCAAAGGRYTAAGCCATGC-3' and 18SRU; 5'-CWGGTTCACCWACGGAAACCTTGTTACG-3' (35) as previously described (7) and sequenced by Sanger sequencing. This confirmation PCR was done on samples TA21, TA26, and DS12. Assembled 18S rRNA genes were aligned to references using MUSCLE and a Maximum likelihood (ML) phylogenetic tree was generated using IQ-TREE with model testing.

Detection of horizontally acquired genes. Genes that may have been acquired by horizontal gene transfer were identified using phylogenetic methods. Initially, protein sequences of genes of interest that were not shared across our draft genomes were aligned to references that spanned eight different bacterial phyla (including Acidobacteria, Bacteroidetes, Nitrospiraceae, Spirochaetes, Firmicutes, Actinobacteria, Proteobacteria, and group PVC) using MUSCLE and phylogenetic trees were generated using IQ-TREE with model testing. Gene trees were then compared to the 16S rRNA gene tree phylogeny (Supplemental Figure 3) to determine evolutionary incongruence.

Analysis of genes involved in competence and recombination. Genes known to be involved in DNA uptake, competence, and recombination were identified in each *Endomicrobium* spp. draft genome based on their RAST annotations and homology to reference sequences. The distribution of these genes was then compared across draft genomes and references which included free-living relatives and other endosymbionts. To asses if these genes were complete and if the encoded proteins likely retained their putative functions, homologs of each gene were obtained from genomes of bacteria belonging to the phylum Elusimicrobia, aligned with MUSCLE, and phylogenetic trees were generated using IQ-TREE (33) with model testing, and support values generated using the "–abayes" and "–bb 1000" commands. The resulting phylogenetic trees were used along with the MUSCLE alignments to perform a dN/dS analysis using the program Codeml which is a part of the PAML and PAMLX packages (36, 37).

Mapping transcriptome reads to draft genomes. RNA-seq metatranscriptome reads were quality trimmed and filtered as described above and error corrected in Geneious R11 (38) using BBNorm with default settings. To remove rRNA reads before mapping, rRNA sequences were identified from each metagenome assembly using RNAmmer (39) and reads were mapped to these, as well as, rRNA references from refseq (40), SILVA (41) , and DictDb (42) databases using BBmap (22). Remaining metatranscriptome reads were then mapped to their respective *Endomicrobium* spp. draft genome in Geneious R11 using Bowtie 2 (43) with alignment type set to "End to End" and using the "Medium Sensitivity" preset. Expression levels were then calculated in Geneious R11, ambiguously mapped reads were excluded from the calculations. RPKM values for genes in each genome are given in Supplementary Tables 5 and 6 in S1 File

Verification of *comEC* expression by RT-PCR. Primers were designed to amplify *comEC* from '*Ca*. Endomicrobium agilae' in Geneious R11 using Primer3 (44) (Primers: endo_comec_F: 5'-ATTTGCCTGTGTTTGAGAGT-3' and endo_comec_R: 5'-CCTGTTCCTGTGCTTTCAG-3'). Twenty termites were used to prepare RNA and cDNA samples for RT-PCR analysis. Termite hindguts were dissected and ruptured in TU on ice in an anaerobic chamber. Hindgut contents were washed with ice-cold TU three times at 3,000 rpm in an Eppendorf microcentrifuge for 90 seconds and then lysed in 1mL of TRIzol™ Reagent (Thermo Fisher Scientific). RNA was isolated per the manufacturer's protocol and treated with TURBO™ DNase (Thermo Fisher Scientific) following the manufacturer's protocol for 50µl reactions using ~5.6 mg of total RNA in a 20 ul volume. 20 ul of the DNase-treated RNA was then used as template for cDNA synthesis using SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific) following the manufacturer's protocol for first strand synthesis primed with random hexamers. The resulting cDNA was treated with *E. coli* RNaseH for 20 minutes at 37°C.

RT-PCR reactions were performed using the *Endomicrobium comEC* primers with RNaseH-treated cDNA serving as template and the no-RT control consisting of DNase-treated RNA that did not undergo cDNA synthesis. RT-RCR was performed using Phusion Polymerase (Thermo Fisher Scientific) with HF buffer and DMSO. Cycling conditions were: initial denaturation at 94°C for 3 minutes, followed by 35 cycles of 94°C for 45 seconds, annealing at 59°C for 30 seconds, and extension at 72°C for 45 seconds. Final extension was done 72°C for 10 minutes. Hindgut DNA (washed protists cell fractions from five hindguts in molecular grade Tris EDTA buffer) was used as a positive PCR control. RT-PCR products were visualized using a 1% agarose gel with ethidium bromide. Products were purified using the Monarch DNA gel purification kit (New England Biolabs), and Sanger sequenced.

Data availability. Raw reads and assemblies will be submitted to NCBI GenBank under BioProject PRJNA644342.

Results

Phylogeny of protist hosts. Protist 18S rRNA genes were retrieved from metagenome assemblies and confirmed (when possible) independently by PCR and Sanger sequencing. A Maximum likelihood (ML) phylogenetic tree was made that indicated species of the protist cells used in this study were *Trichonympha agilis* (cells TA21 and TA26), *Pyrsonympha vertens* (cells PV1 and PV7), and *Dinenympha* species II (cell DS12) (Figure 1). These protist species have been previously confirmed to live associated with *R. flavipes*, the termite species used in this study.

***Endomicrobium* genome statistics**. Five near-complete *Endomicrobium* genomes were obtained from single protist cell metagenomic assemblies. The five genomes ranged from 1.12 - 1.37 Mb in size, 35.3 – 36.6 % G+C, and 93.3 - 96.6% completeness (Figure 2A). To determine if these genomes were from the same or different *Endomicrobium* species, we calculated pairwise genomic s using an ANI score of 95% or greater as a marker for species-level cutoff (Figure 2B) (45).

From *T. agilis* samples, we assembled two Endomicrobium draft genomes which had and ANI score of greater than 96% to one another but less than 90% to '*Ca*. E. trichonymphae' Rs-D17 indicating that they are likely different species. Based on this analysis, we refer to the draft genomes as coming from '*Candidatus* Endomicrobium agilae' TA21 and '*Candidatus* Endomicrobium agilae' TA26. We also assembled two *Endomicrobium* genomes from *P. vertens* samples which had an ANI score greater than 97% identity to each another (Figure 2B) and whose 16S rRNA genes were greater than 98% identical to a previously described species, '*Candidatus* Endomicrobium pyrsonymphae' (6), which is the Candidatus species designation that we use for PV1 and PV7. One additional *Endomicrobium* genome was assembled from

Dinenympha species II. This genome did not share an ANI score greater than 95% to other *Endomicrobium* genomes and was thus given a new Candidatus species designation '*Candidatus* Endomicrobium dinenymphae' DS12 (Figure 2B).

Individually, these *Endomicrobium* genomes contained between 1005 – 1230 orthologous gene clusters (OGCs), of which 717 were found in all five genomes (Figure 2C). Additionally, 409 OCGs were unique to '*Ca*. E. agilae' TA21 and TA26 and another 183 OGCs were unique to '*Ca*. E. pyrsonymphae' PV1 and PV7 (Figure 2C). Although the genome of '*Ca*. E. dinenymphae' DS12 only had 24 unique OGCs, it shared 153 with '*Ca*. E. pyrsonymphae' PV1 and PV7 (Figure 2C) which may reflect similar selective pressures for gene retention in their Oxymonad hosts *Dinenympha* and *Pyrsonympha*, or their more recent shared history, compared with the *Endomicrobium* (TA21 and TA26) that associated with the Parabasalid *T. agilis*.

Biosynthesis of amino acids, vitamins and peptidoglycan. The presence of genes for the various function discussed below were detected by tblastn using the queries listed in Supplementary File 1 (Tables 8 and 9). When genes were not detected in this manner, read-mapping using Geneious and read-mapping using Megan (63) were done as well to identify reads from genes that may not have assembled into contigs. In general, each of the five *Endomicrobium* genomes assembled in this study had similar gene content for processes involved in the biosynthesis of amino acids (Supplementary Figure 1A), vitamins (Supplementary Figure 1B), and peptidoglycan (Supplementary Figure 1D). Each genome possessed complete pathways for alanine (from cysteine) aspartate, arginine, glutamine, glutamate, glycine (from imported serine), isoleucine, leucine, valine, lysine, tyrosine, phenylalanine and tryptophan biosynthesis (Supplementary Figure 1A). Interestingly, the

Endomicrobium symbionts of Oxymonad protists (PV1, PV7 and DS12) lacked at least one gene in the biosynthesis pathway for histidine (*hisG*) (Supplementary Figures 1A and 2). The histidine biosynthetic pathway was complete in the genomes of '*Ca.* E. agilae' TA21 and TA26 (Supplementary Figure 1A). Conversely, it is likely that the *Endomicrobium* symbionts of Oxymonad protists (PV1, PV7, and DS12) can make proline, while the symbionts represented by genomes TA21, TA26 and RsD17 cannot (Supplementary Figure 1A). The five genomes encoded incomplete pathways for the synthesis of cysteine and methionine. The three genomes isolated from Oxymonad protists encoded a methionine transporter (MetT) and all contained a gene encoding aB12-dependent methionine synthase system comprised of MetH and an activation protein MetH2 (Supplementary Figure 1A). Also incomplete in all five genomes were pathways for the synthesis of serine and asparagine. Each genome encoded a serine transporter SdaC, a proline transporter (ProT) and PV1, PV7, and DS12 each encoded a glutamate transporter, GltP (Supplementary Figure 1A).

The five *Endomicrobium* genomes also had similar gene content for processes involved in the biosynthesis of vitamins and co-factors, with the pathways to pantothenate, CoA, NAD and NADP being complete and other pathways being incomplete (Supplementary Figure 1B). Interestingly, the biotin biosynthesis pathways in the five genomes are missing just a single gene (*bioW*) needed to convert pimelate to pimelate-CoA suggesting that pimelate-CoA may be synthesized by another enzyme or imported (Supplementary Figure 1B). Several genes in the thiamine biosynthesis pathway were also missing in each of these genomes (Supplementary Figure 1B). As noted previously for *E. proavitum* and "*Candidatus* Endomicrobium trichonymphae" strain Rs-D17 the five genomes described here were also missing the steps in the folate pathway needed to make 4-aminobenzoate, which may be transported into the cells

(18). The pathways for pyridoxine (B6) and vitamin B12 were also incomplete, though each of the five *Endomicrobium* genomes appeared to encode ABC transport systems for vitamin B12 and heme.

Regarding peptidoglycan synthesis, each *Endomicrobium* genome was missing an enzyme (BacA) that typically dephosphorylates undecaprenyl pyrophosphate. Since these different *Endomicrobium* species, including the free-living *E. proavitum*, are missing the same gene it may be that these bacteria utilize an alternate phosphatase to carry out the same function as BacA.

Differences in Carbon Metabolism. Some of the more interesting differences between these *Endomicrobium* genomes pertained to their carbon metabolisms. Each of these five *Endomicrobium* genomes encoded relatively simple pathways for importing and using different wood-derived carbon sources. Each had a complete phosphotransferase system (PTS) for importing sugars. Present were two EIIA genes encoding sugar specific phosphorylation proteins most closely related to those of the mannose and fructose type EIIA proteins (Supplementary Figure 1C). Zheng et al. reported that *E. proavitum,* which contains a very similar PTS pathway, did not grow on mannose or fructose, but did grow on glucose, suggesting that glucose may be the carbohydrate transported by the PTS in that *Endomicrobium* species, and perhaps in the one described here as well (16, 18).

Based on the gene content in the five genomes analyzed here, carbon sources capable of being catabolized by endosymbiotic *Endomicrobium* species may often differ from each other and from their free-living relatives. For example, '*Ca.* E. agilae' TA21 and TA26 encoded all the genes necessary to import and use both glucuronate and glucose-6-phosphate (Figure 3A & 3B).

The closely-related ‘*Ca*. E. trichonymphae’ Rs-D17 (3) also contained these genes. Interestingly, genome analyses suggest that these two carbon sources cannot be used by the other *Endomicrobium* species studied here which lack the glucuronate transporter ExuT, the glucuronate isomerase UxaA and the glucose-6-phosphate transporter UhcP. The other *Endomicrobium* genomes encoded either arabinose (‘*Ca*. E. pyrsonymphae’ PV1 and PV7) or xylose (‘*Ca*. E. dinenymphae’ DS12) import and catabolism proteins that were not encoded in the TA21, TA26, *E. proavitum* or ‘*Ca*. E. trichonymphae’ Rs-D17 genomes. (Figures 4A & 4B; Figures 5A & 5B, respectively). Transcriptome data indicated that each of the genes involved in these carbon usage pathways were expressed in the respective *Endomicrobium* while they resided in their protist hosts (Figures 3C, 4C, & 5C). Metabolites from these carbon sources are typically fed into both the non-oxidative pentose phosphate pathway and glycolysis, both of which are complete in the five genomes described here (Figures 3B, 4B, 5B, & Supplementary Figure 1C).

Other likely differences in carbon metabolism of these *Endomicrobium* species, included the production of fermentation end products (Supplementary Figure 1C). Analysis of the five *Endomicrobium* genomes suggested that following glycolysis, pyruvate can be fermented to acetate, however only the genomes of ‘*Ca.* E. agilae’ and ‘*Ca*. E. dinenymphae’, encoded AdhE which can convert acetate to ethanol. (Figures 3B, 4B, & 5B). In addition, genes encoding lactate dehyrodenase (LdH) were in the genomes of both ‘*Ca*. E. dinenymphae’, and ‘*Ca*. E. pyrsonymphae’, but not ‘*Ca.* E. agilae’ or ‘*Ca*. E. trichonymphae’ Rs-D17 (Figures 3B, 4B, & 5B). Differences in these fermentation pathways between free-living *E. proavitum* and ‘*Ca*. E. trichonymphae’ Rs-D17 were described earlier by Zheng et al. (18).

Previous studies identified genes acquired by horizontal gene transfer (HGT) in other *Endomicrobium* species (18), therefore, we tested whether HGT could, at least in part, explain the differences seen in carbon metabolism across the genomes presented in this study. Phylogenetic trees were made for each of the transport and isomerase proteins in the glucuronate, arabinose and xylose degradation pathways (Figures 3D, 4D, & 5D) and the phylogenies compared to the *Endomicrobium* 16S rRNA gene phylogeny to determine if they were congruent (Supplementary Figure 3). In each case, these phylogenies were not congruent, suggesting that these genes were acquired by HGT (Figures 3D, 4D, & 5D). Likely donor taxa include Bacteroidetes, Actinobacteria, and Firmicutes (Figures 3D, 4D, & 5D), which are all part of the hindgut community of *R. flavipes* (8). Similar data supporting HGT in '*Ca*. E. trichonymphae' Rs-D17 have been reported and suggest that *Endomicrobium symbionts* are not cut off from gene flow and HGT (18). This is in contrast to the older endosymbionts of sap-feeding insects, which are traditionally thought to experience little to no gene flow, however recent analyses suggested that HGT may occur more frequently than previously thought in these symbionts (46).

Natural transformation and competence as a possible mechanism for acquiring genes

Analyses of sequenced genomes of endosymbiotic *Endomicrobium* lineages indicate that acquisition of genes by HGT is relatively common. Thus, their genomes could reveal insights into the mechanisms by such genes were acquired. Interestingly, compared to other endosymbionts, the *Endomicrobium* genomes were enriched in genes related to the uptake of exogenous DNA and recombination (natural transformation/competence) (Supplementary Figures 4A and 4C). Of special interest are the *Endomicrobium* genes *comEC*, *comEB*, *comF*, *comM*, *ssb*, *drpA*, and *recA* which are all involved in natural transformation in bacteria such as *Vibrio cholerae* (47).

The dN/dS analyses of these genes supported the hypothesis that selection was acting to maintain the amino acid sequences of their corresponding gene products (dN/dS < 1.0) with the exception of *ssb* from TA21 (Figure 6A). In addition, transcriptome analysis indicated that these genes were expressed (Figure 6B). Expression of *comEC*, which encodes a transporter that imports single stranded DNA across the inner-membrane and into the cytoplasm of Gram-negative bacteria (47, 48), was verified by RT-PCR and sequencing of *Ca*. Endomicrobium agilae using *comEC* specific primers on a protist cell fraction sample prepared from 20 worker termite hindguts (Figure 6C). Together these data support the hypothesis that genes involved in this competence pathway are both conserved and expressed in these *Endomicrobium* symbionts of hindgut protists of *R. flavipes*.

The competence genes discussed above are involved in the translocation of single-stranded DNA across the inner membrane of Gram-negative bacteria and subsequent recombination. Also present in the genomes of all five *Endomicrobium* species analyzed in this study are genes which encode proteins that are similar to Type IV pilins. The TA21 and TA26 genomes contained a large chromosomal region devoted to Type IV Tad-like pilus synthesis as does E. *proavitum* and *E. minutum*. The bacterium *Ca*. E. trichonymphae' Rs-D17 has a similar region, but it appears that many of the genes have become pseudogenes. The *P. vertens* and *Dinenympha* species II *Endomicrobium* symbionts had genes encoding pilins similar to the Type II PulG pilins. Some pilins from classes Type-IV and Type-II can bind and import double-stranded DNA across the outer membrane and have been shown to work in conjunction with ComEC-type proteins (47, 49). In addition, all five genomes possessed a pre-pilin peptidase

(PilD). A graphical summary of these findings along with a model of how competence may work in these organisms is provided as Supplementary Figure 6. A list of genes and their putative function can be found in Supplementary Table 7 in S1 File.

Transcriptome analysis of *Endomicrobium* populations inside single protist cells.

Transcriptome analyses of the *Endomicrobium* populations inside single protist cells (from which the five genomes were derived) revealed similar gene expression profiles with a few notable exceptions (Figure 7). While our sample size for this work was necessarily small and while we were unable to do time-resolved sampling of the hindgut community from single termites, some general trends did appear in the transcriptomic data. Among COG categories, which are quite broad, the expression by each endosymbiont population was relatively similar with one exception being that there was higher expression of genes related to carbohydrate transport and metabolism in *Trichonympha* hosts (TA) compared to the Oxymonad hosts (PV and DS12) (Figure 7A).

Analyses which focused on narrower categories such genes in related biosynthetic pathways, carbohydrate transport and break down, peptidoglycan synthesis and DNA uptake and repair revealed further differences not only between the endosymbionts of different protist species but even between the populations of endosymbionts of individual protist cells of the same type (Figure 7B). This is demonstrated by the differences related to the expression of genes in the glutamine and glutamate biosynthesis pathway (Figure 7B). Overall this pathway is more highly expressed by the endosymbionts of *Pyrsonympha* hosts compared to those in other protist species but there was variation in the expression of this pathway between individual protist host cells. For example, in host cell PV1 this pathway represented 4.2% of the total transcriptome reads, mostly from the gene *glnN* encoding a glutamine synthase, whereas in host cell PV7 it comprised only 0.2% (Figure 7B). Similar variation in this pathway was seen in the TA21 and TA26 transcriptomes. This data suggests that even populations of the bacterium residing in different host cells may not be expressing the same functions at any given point in time.

Core genes, which were shared between all five of the *Endomicrobium* species, represented an average of 30% – 36% of the transcription of each endosymbiont population (Figure 7C). The expression of genes that were specific to each *Endomicrobium* species ranged from 11% in 'Ca. E. pyrsonymphae' to 22% in 'Ca. E. agilae' indicating that there is likely differential gene content and gene expression endosymbionts of different protist host species (Figure 7C). Transcriptomic data are in Supplementary Tables 5 and 6 in S1 File

Discussion

Single-cell protist metagenomics has enabled the assembly of genomes from several protist-associated bacterial symbionts from termites (1, 3, 4, 50, 51). In this study, we present near-complete draft genomes and transcriptomes of five endosymbiotic *Endomicrobium* samples, from three different protist species. These endosymbionts displayed differences with regards to their gene content and expression. For example, these organisms possessed different carbon usage pathways. One hypothesis that may explain such differences in carbon utilization is that different carbon sources may be provided to the *Endomicrobium* endosymbionts as by-products of the protist hosts' hydrolysis and fermentation of the polysaccharides that originated in wood. For example, glucuronate may be present in the cytoplasm of *Trichonympha* spp. because they possessed the enzymes needed to cleave those monomers from polysaccharides found in wood, whereas the other protists, *P. vertens* and *D.* species II, may not be able to generate such monomers, or they may use them for other purposes (see below). If true, this suggests that there may be specialization among the protists with regards to polysaccharide hydrolysis in the hindgut of wood-feeding termites. A recent study demonstrated a division of labor among symbiotic protist species in a different termite, *Coptotermes formosanus*, where certain protist species produce different hydrolytic enzymes to degrade polysaccharides found in wood (52). These differences in protist functions may also explain why there was higher expression of genes related to carbohydrate transport and metabolism in the endosymbionts of *Trichonympha* hosts compared to the other protist species (Figure 7).

However, an alternative hypothesis is that metabolites can be partitioned within the host and some are specifically provided to certain symbionts. This, if true, may allow the host to control endosymbiont population densities through selective carbon source provision. Such host control of carbon provisioning is thought to operate in nitrogen-fixing root nodule symbioses ensuring that bacterial symbionts continue to provide fixed nitrogen in return for plant-provided carbon. In support of the second hypothesis, the membrane-embedded symbiont '*Ca*. Desulfovibrio trichonymphae' which co-colonizes the same *Trichonympha* host as '*Ca*. E. trichonymphae' Rs-D17, uses malate and citrate as carbon sources whereas its co-inhabitant likely uses glucuronate and glucose-6-phosphate (51).

Evidence suggests that *Endomicrobium* species have acquired genes by HGT and some of the donor taxa may include termite-associated bacteria. Endosymbiotic lineages of *Endomicrobium*, and their free-living relatives, possess many genes involved in DNA uptake, repair, and recombination (Supplementary Figure 4). Our analyses showed that the genes *comEC*, *comEB*, *comF*, *comM*, *ssb*, *drpA*, and *recA* are usually conserved within the Elusimicrobia phylum and were expressed in the endosymbiotic *Endomicrobium* species characterized in this study (Figure 6). Collectively these genes have been shown to be involved in the translocation of single stranded DNA across the inner membrane of other Gram-negative bacteria and in homologous recombination (47). The gene *comEC*, in particular, has an important function in this process as it encodes an essential part of the DNA transporter (47, 48). Using both transcriptome data, RT-PCR, and sequencing we were able to show that '*Ca*. Endomicrobium agilae' *comEC* is expressed in these endosymbionts (Figure 6B and 6C). Collectively the genes involved in the competence pathway comprised between 0.4% to 1.3% of the total transcriptome reads of each endosymbiont population (Figure 7). These data suggest that *Endomicrobium* species may have the ability to become competent which may allow them to acquired DNA from the wider termite gut community and could result in HGT.

It is not clear how these organisms transport DNA across their outer membranes. None of these *Endomicrobium* species possessed all the components of a Type IV pili-based DNA-translocation system, but '*Ca*. E. agilae' TA21 and TA26 contained a near-complete Type IV *tad* system which may allow DNA uptake (Supplemental Figures 4 and 5). It is also puzzling why each of the five genomes have retained *pilD,* a prepilin peptidase as well as genes encoding Type II and Type IV pilins. These may carry out some function in DNA-uptake or they may be non-functional and are in the process of being lost. If competence is a common trait among the *Endomicrobium*, this could explain why these organisms have many genes acquired through HGT and this capability may allow for rapid adaptation to new and diverse niches. Because hindgut protists phagocytize wood, wood-associated bacteria and perhaps free-living bacteria (53), this may be a route through which endosymbiotic *Endomicrobium* could be exposed to exogenous DNA.

However, it is worth noting that competence is not the only plausible avenue for DNA acquisition in these endosymbionts. HGT could also occur by bacteriophage transduction, conjugation, or other routes. Several lines of evidence indicate that these endosymbionts are susceptible to molecular parasites, such as bacteriophages and plasmids. Previous studies have reported that *Endomicrobium* species possessed several intact defense mechanisms to combat molecular parasites such as CRISPR-Cas and restriction-modification systems (54, 55). The *Endomicrobium* species sequenced in this study also contained those defense systems. The complete genome sequence of a bacteriophage of an endosymbiont ('*Candidatus* Azobacteroides pseudotrichonymphae') of a termite hindgut protist has previously been published, indicating that phage infection is not limited to *Endomicrobium* endosymbionts and may be common in termite hindguts (56).

Our analysis of *Endomicrobium* genomes and transcriptomes obtained from single protist cell metagenomes, highlighted several important differences across protist hosts which have led to hypotheses that warrant further investigation. In each case, the major hurdle of testing these is the current inability to culture the protists hosts which restricts their experimental tractability. However, the use of additional and new -omics approaches should further our understanding of these symbioses by focusing on the hosts' and their symbionts' genes, mRNA, metabolic and protein contents (57–62).

Acknowledgments. This research was funded by the National Science Foundation (NSF) division of Emerging Frontiers in Research and Innovation in Multicellular and Inter-kingdom Signaling. Award number 1137249 (DJG and JG).

References

1. Hongoh Y, Sharma VK, Prakash T, Noda S, Toh H, Taylor TD, Kudo T, Sakaki Y, Toyoda A, Hattori M, Ohkuma M. 2008. Genome of an endosymbiont coupling N2 fixation to cellulolysis within protist cells in termite gut. Science 322:1108–1109.

2. Ohkuma M. 2008. Symbioses of flagellates and prokaryotes in the gut of lower termites. Trends Microbiol 16:345–352.

3. Hongoh Y, Sharma VK, Prakash T, Noda S, Taylor TD, Kudo T, Sakaki Y, Toyoda A, Hattori M, Ohkuma M. 2008. Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. PNAS 105:5555–5560.

4. Strassert JFH, Mikaelyan A, Woyke T, Brune A. 2016. Genome analysis of ‘ Candidatus Ancillula trichonymphae’’, first representative of a deep-branching clade of Bifidobacteriales , strengthens evidence for convergent evolution in flagellate endosymbionts.’ Environ Microbiol Rep 8:865–873.

5. Ohkuma M, Sato T, Noda S, Ui S, Kudo T, Hongoh Y. 2007. The candidate phylum “Termite Group 1” of bacteria: Phylogenetic diversity, distribution, and endosymbiont members of various gut flagellated protists. FEMS Microbiol Ecol 60:467–476.

6. Stingl U, Radek R, Yang H, Brune A. 2005. “Endomicrobia”: Cytoplasmic symbionts of termite gut protozoa form a separate phylum of prokaryotes. Appl Environ Microbiol 71:1473–1479.

7. Stephens ME, Gage DJ. 2020. Single-cell amplicon sequencing reveals community structures and transmission trends of protist-associated bacteria in a termite host. PLoS One 15:1–19.

8. Benjamino J, Graf J. 2016. Characterization of the Core and Caste-Specific Microbiota in

the Termite, Reticulitermes flavipes. Front Microbiol 7:171.

9. Boucias DG, Cai Y, Sun Y, Lietze VU, Sen R, Raychoudhury R, Scharf ME. 2013. The hindgut lumen prokaryotic microbiota of the termite Reticulitermes flavipes and its responses to dietary lignocellulose composition. Mol Ecol 22:1836–1853.

10. Ikeda-Ohtsubo W, Brune A. 2009. Cospeciation of termite gut flagellates and their bacterial endosymbionts: Trichonympha species and “Candidatus Endomicrobium trichonymphae.” Mol Ecol 18:332–342.

11. Mikaelyan A, Thompson CL, Meuser K, Zheng H, Rani P, Plarre R, Brune A. 2017. High-resolution phylogenetic analysis of Endomicrobia reveals multiple acquisitions of endosymbiotic lineages by termite gut flagellates. Environ Microbiol Rep 9:477–483.

12. Ikeda-Ohtsubo W, Desai M, Stingl U, Brune A. 2007. Phylogenetic diversity of “Endomicrobia” and their specific affiliation with termite gut flagellates. Microbiology 153:3458–3465.

13. Zheng H, Dietrich C, L. Thompson C, Meuser K, Brune A. 2015. Population Structure of Endomicrobia in Single Host Cells of Termite Gut Flagellates (*Trichonympha* spp.). Microbes Environ 30:92–98.

14. Izawa K, Kuwahara H, Sugaya K, Lo N, Ohkuma M, Hongoh Y. 2017. Discovery of ectosymbiotic Endomicrobium lineages associated with protists in the gut of stolotermitid termites. Environ Microbiol Rep 9:411–418.

15. Herlemann DPR, Geissinger O, Ikeda-Ohtsubo W, Kunin V, Sun H, Lapidus A, Hugenholtz P, Brune A. 2009. Genomic analysis of “Elusimicrobium minutum” the first cultivated representative of the phylum “Elusimicrobia” (formerly termite group 1). Appl Environ Microbiol 75:2841–2849.

16. Zheng H, Dietrich C, Radek R, Brune A. 2016. Endomicrobium proavitum , the first isolate of Endomicrobia class . nov . (phylum Elusimicrobia) – an ultramicrobacterium with an unusual cell cycle that fixes nitrogen with a Group IV nitrogenase. Environ Microbiol 18:191–204.

17. Ikeda-Ohtsubo W, Faivre N, Brune A. 2010. Putatively free-living 'Endomicrobia'- ancestors of the intracellular symbionts of termite gut flagellates? Environ Microbiol Rep 2:554–9.

18. Zheng H, Dietrich C, Brune A. 2017. Genome analysis of Endomicrobium proavitum suggests loss and gain of relevant functions during the evolution of intracellular symbionts. Appl Environ Microbiol 83:1–14.

19. Brune A, Emerson D, Breznak JA. 1995. The termite gut microflora as an oxygen sink: Microelectrode determination of oxygen and pH gradients in guts of lower and higher termites. Appl Environ Microbiol 61:2681–2687.

20. Yang H, Schmitt-Wagner D, Stingl U, Brune A. 2005. Niche heterogeneity determines bacterial community structure in the termite gut (Reticulitermes santonensis). Environ Microbiol 7:916–932.

21. Trager W. 1934. The Cultivation of a Cellulose-Digesting Flagellate, Trichomonas Termopsidis, and of Certain other Termite Protozoa. Biol Bull 66:182–190.

22. Bushnell B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner.

23. Coil D, Jospin G, Darling AE. 2015. A5-miseq: An updated pipeline to assemble microbial genomes from Illumina MiSeq data. Bioinformatics 31:587–589.

24. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, Sneddon MW, Henderson ML, Riehl WJ, Murphy-Olson D, Chan

SY, Kamimura RT, Kumari S, Drake MM, Brettin TS, Glass EM, Chivian D, Gunter D, Weston DJ, Allen BH, Baumohl J, Best AA, Bowen B, Brenner SE, Bun CC, Chandonia J-M, Chia J-M, Colasanti R, Conrad N, Davis JJ, Davison BH, DeJongh M, Devoid S, Dietrich E, Dubchak I, Edirisinghe JN, Fang G, Faria JP, Frybarger PM, Gerlach W, Gerstein M, Greiner A, Gurtowski J, Haun HL, He F, Jain R, Joachimiak MP, Keegan KP, Kondo S, Kumar V, Land ML, Meyer F, Mills M, Novichkov PS, Oh T, Olsen GJ, Olson R, Parrello B, Pasternak S, Pearson E, Poon SS, Price GA, Ramakrishnan S, Ranjan P, Ronald PC, Schatz MC, Seaver SMD, Shukla M, Sutormin RA, Syed MH, Thomason J, Tintle NL, Wang D, Xia F, Yoo H, Yoo S, Yu D. 2018. KBase: The United States Department of Energy Systems Biology Knowledgebase. Nat Biotechnol 36:566–569.

25. Laczny CC, Sternal T, Plugaru V, Gawron P, Atashpendar A, Margossian HH, Coronado S, der Maaten L V., Vlassis N, Wilmes P. 2015. VizBin - An application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome 3:1–7.

26. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

27. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9.

28. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from. Cold Spring Harb Lab Press Method 1:1–31.

29. Aziz RK, Bartels D, Best A, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S,

Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST Server: Rapid annotations using subsystems technology. BMC Genomics 9:1–15.

30. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason JA, Stevens R, Vonstein V, Wattam AR, Xia F. 2015. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep 5.

31. Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27–30.

32. Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

33. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274.

34. Richter M, Rosselló-Móra R, Oliver Glöckner F, Peplies J. 2016. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinformatics 32:929–931.

35. Tikhonenkov D V., Janouškovec J, Keeling PJ, Mylnikov AP. 2016. The Morphology, Ultrastructure and SSU rRNA Gene Sequence of a New Freshwater Flagellate, Neobodo borokensis n. sp. (Kinetoplastea, Excavata). J Eukaryot Microbiol 63:220–232.

36. Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics 13:555–556.

37. Xu B, Yang Z. 2013. PamlX: A graphical user interface for PAML. Mol Biol Evol 30:2723–2724.

38. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

39. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. 2007. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108.

40. Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:61–65.

41. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res 41:590–596.

42. Mikaelyan A, Köhler T, Lampert N, Rohland J, Boga H, Meuser K, Brune A. 2015. Classifying the bacterial gut microbiota of termites and cockroaches: a curated phylogenetic reference database (DictDb). Syst Appl Microbiol 38:472–482.

43. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

44. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3-new capabilities and interfaces. Nucleic Acids Res 40:1–12.

45. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High

throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:1–8.

46. López-Madrigal S, Gil R. 2017. Et tu, brute? Not even intracellular mutualistic symbionts escape horizontal gene transfer. Genes (Basel) 8:1–16.

47. Seitz P, Blokesch M. 2013. DNA-uptake machinery of naturally competent Vibrio cholerae. Proc Natl Acad Sci 110:17987–17992.

48. Pimentel ZT, Zhang Y. 2018. Evolution of the natural transformation protein, ComEC, in Bacteria. Front Microbiol 9:1–10.

49. Ellison CK, Dalia TN, Vidal Ceballos A, Wang JCY, Biais N, Brun Y V., Dalia AB. 2018. Retraction of DNA-bound type IV competence pili initiates DNA uptake during natural transformation in Vibrio cholerae. Nat Microbiol 3:773–780.

50. Utami YD, Kuwahara H, Igai K, Murakami T, Sugaya K, Morikawa T, Nagura Y, Yuki M, Deevong P, Inoue T, Kihara K, Lo N, Yamada A, Ohkuma M, Hongoh Y. 2019. Genome analyses of uncultured TG2/ZB3 bacteria in 'Margulisbacteria' specifically attached to ectosymbiotic spirochetes of protists in the termite gut. ISME J 13:455–467.

51. Sato T, Hongoh Y, Noda S, Hattori S, Ui S, Ohkuma M. 2009. Candidatus Desulfovibrio trichonymphae, a novel intracellular symbiont of the flagellate Trichonympha agilis in termite gut. Environ Microbiol 11:1007–15.

52. Nishimura Y, Otagiri M, Yuki M, Shimizu M, Inoue J ichi, Moriya S, Ohkuma M. 2020. Division of functional roles for termite gut protists revealed by single-cell transcriptomes. ISME J.

53. Brune A. 2014. Symbiotic digestion of lignocellulose in termite guts. Nat Rev Microbiol 12:168–80.

54. Izawa K, Kuwahara H, Kihara K, Yuki M, Lo N, Itoh T, Ohkuma M, Hongoh Y. 2016. Comparison of intracellular “ca. endomicrobium trichonymphae” genomovars illuminates the requirement and decay of defense systems against foreign DNA. Genome Biol Evol 8:3099–3107.

55. Zheng H, Dietrich C, Hongoh Y, Brune A. 2016. Restriction-modification systems as mobile genetic elements in the evolution of an intracellular symbiont. Mol Biol Evol 33:721–725.

56. Pramono AK, Kuwahara H, Itoh T, Toyoda A, Yamada A, Hongoh Y. 2017. Discovery and Complete Genome Sequence of a Bacteriophage from an Obligate Intracellular Symbiont of a Cellulolytic Protist in the Termite Gut. Microbes Environ Environ 32:112–117.

57. Bhattacharya D, Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, Yang EC, Duffy S. 2011. Single-Cell Genomics Reveals Organismal Interactions in Uncultivated Marine Protists. Science (80-) 332:714–717.

58. Karnkowska A, Vacek V, Zubáčová Z, Treitli SC, Petrželková R, Eme L, Novák L, Žárský V, Barlow LD, Herman EK, Soukal P, Hroudová M, Doležal P, Stairs CW, Roger AJ, Eliáš M, Dacks JB, Vlček Č, Hampl V. 2016. A eukaryote without a mitochondrial organelle. Curr Biol 26:1274–1284.

59. Mangot JF, Logares R, Sánchez P, Latorre F, Seeleuthner Y, Mondy S, Sieracki ME, Jaillon O, Wincker P, Vargas C De, Massana R. 2017. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. Sci Rep 7:1–12.

60. Vacek V, Novák LVF, Treitli SC, Táborský P, Cepicka I, Kolísko M, Keeling PJ, Hampl

V. 2018. Fe-S Cluster Assembly in Oxymonads and Related Protists. Mol Biol Evol 35:2712–2718.

61. Kolisko M, Boscaro V, Burki F, Lynn DH, Keeling PJ. 2014. Single-cell transcriptomics for microbial eukaryotes. Curr Biol 24:R1081–R1082.

62. Hamann E, Tegetmeyer HE, Riedel Di, Littmann S, Ahmerkamp S, Chen J, Hach PF, Strous M. 2017. Syntrophic linkage between predatory Carpediemonas and specific prokaryotic populations. ISME J 11:1205–1217.

63. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, et al. (2016) MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS Computational Biology 12(6): e1004957. https://doi.org/10.1371/journal.pcbi.1004957

Figures and Legends

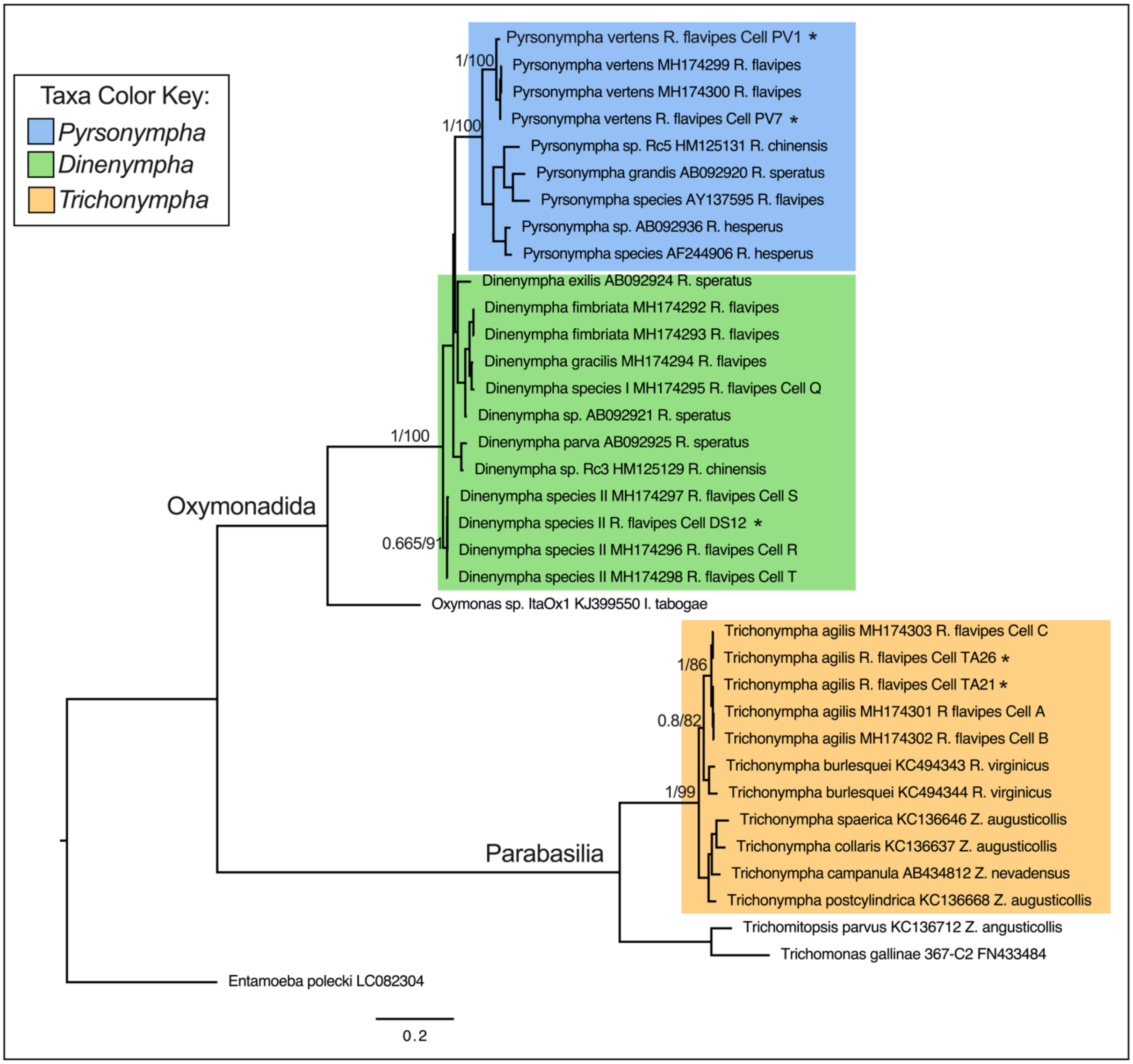
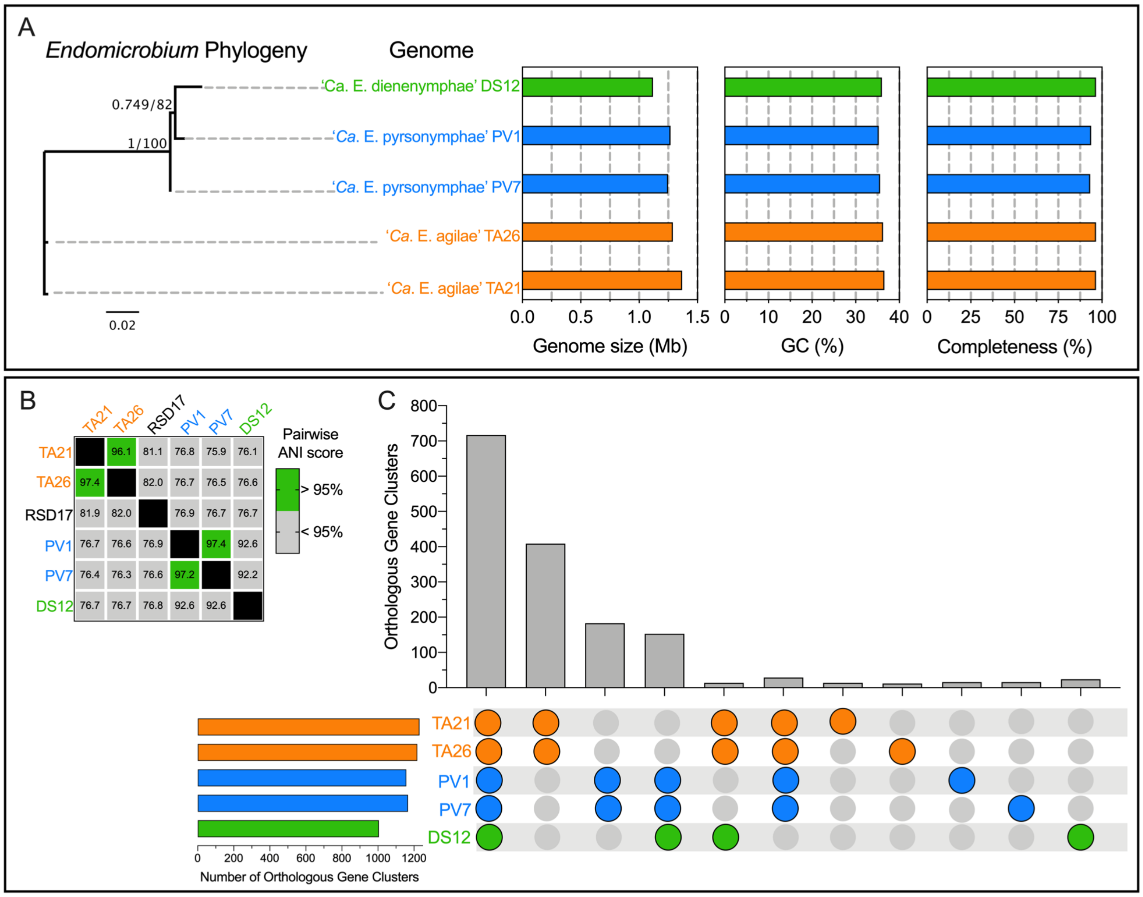

Taxa Color Key:
🟦 Pyrsonympha
🟩 Dinenympha
🟧 Trichonympha
Pyrsonympha vertens R. flavipes Cell PV1 *
1/100
Pyrsonympha vertens MH174299 R. flavipes
Pyrsonympha vertens MH174300 R. flavipes
1/100
Pyrsonympha vertens R. flavipes Cell PV7 *
Pyrsonympha sp. Rc5 HM125131 R. chinensis
Pyrsonympha grandis AB092920 R. speratus
Pyrsonympha species AY137595 R. flavipes
Pyrsonympha sp. AB092936 R. hesperus
Pyrsonympha species AF244906 R. hesperus
Dinenympha exilis AB092924 R. speratus
Dinenympha fimbriata MH174292 R. flavipes
Dinenympha fimbriata MH174293 R. flavipes
Dinenympha gracilis MH174294 R. flavipes
Dinenympha species I MH174295 R. flavipes Cell Q
Dinenympha sp. AB092921 R. speratus
1/100
Dinenympha parva AB092925 R. speratus
Dinenympha sp. Rc3 HM125129 R. chinensis
Oxymonadida
Dinenympha species II MH174297 R. flavipes Cell S
Dinenympha species II R. flavipes Cell DS12 *
0.665/91
Dinenympha species II MH174296 R. flavipes Cell R
Dinenympha species II MH174298 R. flavipes Cell T
Oxymonas sp. ItaOx1 KJ399550 I. tabogae
Trichonympha agilis MH174303 R. flavipes Cell C
Trichonympha agilis R. flavipes Cell TA26 *
1/86
Trichonympha agilis R. flavipes Cell TA21 *
0.8/82
Trichonympha agilis MH174301 R flavipes Cell A
Trichonympha agilis MH174302 R. flavipes Cell B
Trichonympha burlesquei KC494343 R. virginicus
1/99
Trichonympha burlesquei KC494344 R. virginicus
Trichonympha spaerica KC136646 Z. augusticollis
Parabasilia
Trichonympha collaris KC136637 Z. augusticollis
Trichonympha campanula AB434812 Z. nevadensus
Trichonympha postcylindrica KC136668 Z. augusticollis
Trichomitopsis parvus KC136712 Z. angusticollis
Trichomonas gallinae 367-C2 FN433484
Entamoeba polecki LC082304
0.2

Figure 1. Protist 18S rRNA gene phylogeny. 18S rRNA genes were retrieved from single protist cell metagenome assemblies, aligned to references, and a Maximum likelihood (ML) phylogenetic tree was made using IQ-Tree using substitution model TIM2+G4. All 18S rRNA gene sequences obtained in this study (denoted by *) are shown grouped with their respective references. Branch support values represent the Bayesian posterior probability and Bootstrap support values respectively.



	TA21	TA26	RSD17	PV1	PV7	DS12
TA21	■	96.1	81.1	76.8	75.9	76.1
TA26	97.4	■	82.0	76.7	76.5	76.6
RSD17	81.9	82.0	■	76.9	76.7	76.7
PV1	76.7	76.6	76.9	■	97.4	92.6
PV7	76.4	76.3	76.6	97.2	■	92.2
DS12	76.7	76.7	76.8	92.6	92.6	■

Figure 2. *Endomicrobium* draft genomes statistics, speciation, and shared gene content. (A) 16S rRNA gene Maximum likelihood tree (unrooted) of the three *Endomicrobium* species, genome sizes, percent G+C content, and estimated percent genome completeness. (B) Pairwise genomic ANI scores of *Endomicrobium* genomes obtained by this study and a previously sequenced relative Rs-D17. (C) UpSet graph of the number of orthologous gene clusters (OGCs) of protein coding sequences within and across each of the *Endomicrobium* draft genomes.

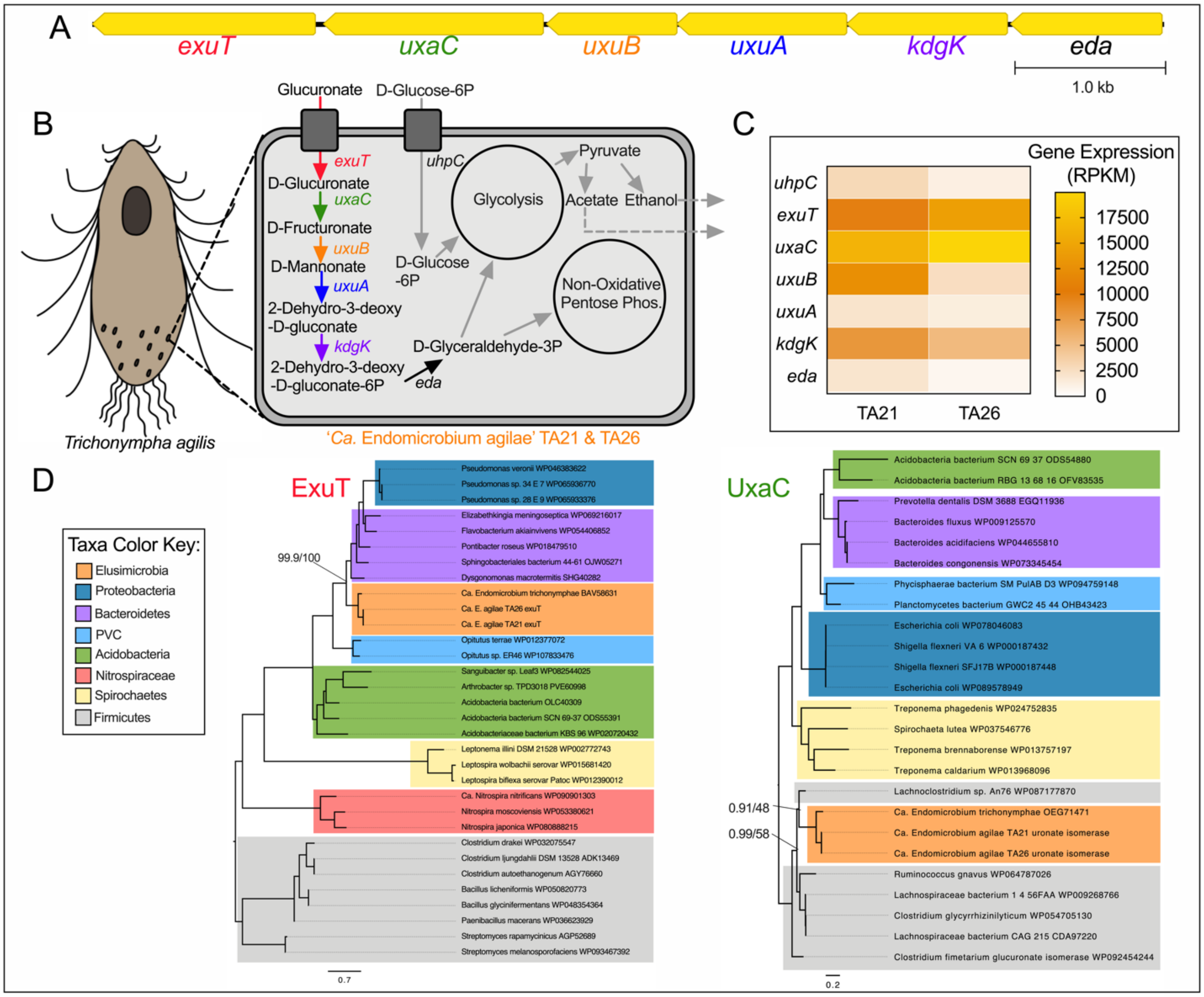

A
exuT
uxaC
uxuB
uxuA
kdgK
eda
1.0 kb
B
Glucuronate
D-Glucose-6P
exuT
uhpC
D-Glucuronate
uxaC
D-Fructuronate
uxuB
D-Mannonate
uxuA
2-Dehydro-3-deoxy
-D-gluconate
kdgK
2-Dehydro-3-deoxy
-D-gluconate-6P
eda
D-Glucose
-6P
Glycolysis
Pyruvate
Acetate
Ethanol
Non-Oxidative
Pentose Phos.
D-Glyceraldehyde-3P
Trichonympha agilis
'Ca. Endomicrobium agilae' TA21 & TA26
C
Gene Expression
(RPKM)
uhpC
exuT
uxaC
uxuB
uxuA
kdgK
eda
TA21
TA26
17500
15000
12500
10000
7500
5000
2500
0
D
Taxa Color Key:
Elusimicrobia
Proteobacteria
Bacteroidetes
PVC
Acidobacteria
Nitrospiraceae
Spirochaetes
Firmicutes
ExuT
99.9/100
Pseudomonas veronii WP046383622
Pseudomonas sp. 34 E 7 WP065936770
Pseudomonas sp. 28 E 9 WP065933376
Elizabethkingia meningoseptica WP069216017
Flavobacterium akiainvivens WP054406852
Pontibacter roseus WP018479510
Sphingobacteriales bacterium 44-61 OJW05271
Dysgonomonas macrotermitis SHG40282
Ca. Endomicrobium trichonymphae BAV58631
Ca. E. agilae TA26 exuT
Ca. E. agilae TA21 exuT
Opitutus terrae WP012377072
Opitutus sp. ER46 WP107833476
Sanguibacter sp. Leaf3 WP082544025
Arthrobacter sp. TPD3018 PVE60998
Acidobacteria bacterium OLC40309
Acidobacteria bacterium SCN 69-37 ODS55391
Acidobacteriaceae bacterium KBS 96 WP020720432
Leptonema illini DSM 21528 WP002772743
Leptospira wolbachii serovar WP015681420
Leptospira biflexa serovar Patoc WP012390012
Ca. Nitrospira nitrificans WP090901303
Nitrospira moscoviensis WP053380621
Nitrospira japonica WP080888215
Clostridium drakei WP032075547
Clostridium ljungdahlii DSM 13528 ADK13469
Clostridium autoethanogenum AGY76660
Bacillus licheniformis WP050820773
Bacillus glycinifermentans WP048354364
Paenibacillus macerans WP036623929
Streptomyces rapamycinicus AGP52689
Streptomyces melanosporofaciens WP093467392
0.7
UxaC
Acidobacteria bacterium SCN 69 37 ODS54880
Acidobacteria bacterium RBG 13 68 16 OFV83535
Prevotella dentalis DSM 3688 EGQ11936
Bacteroides fluxus WP009125570
Bacteroides acidifaciens WP044655810
Bacteroides congonensis WP073345454
Phycisphaerae bacterium SM PulAB D3 WP094759148
Planctomycetes bacterium GWC2 45 44 OHB43423
Escherichia coli WP078046083
Shigella flexneri VA 6 WP000187432
Shigella flexneri SFJ17B WP000187448
Escherichia coli WP089578949
Treponema phagedenis WP024752835
Spirochaeta lutea WP037546776
Treponema brennaborense WP013757197
Treponema caldarium WP013968096
Lachnoclostridium sp. An76 WP087177870
0.91/48
0.99/58
Ca. Endomicrobium trichonymphae OEG71471
Ca. Endomicrobium agilae TA21 uronate isomerase
Ca. Endomicrobium agilae TA26 uronate isomerase
Ruminococcus gnavus WP064787026
Lachnospiraceae bacterium 1 4 56FAA WP009268766
Clostridium glycyrrhizinilyticum WP054705130
Lachnospiraceae bacterium CAG 215 CDA97220
Clostridium fimetarium glucuronate isomerase WP092454244
0.2

Figure 3. Carbon metabolism and HGT in '*Ca*. Endomicrobium agilae'. (A) Gene neighborhood of the genes involved in the metabolism of glucuronate in the '*Ca*. Endomicrobium agilae' TA21 and TA26 genomes. (B) Diagram of a protist host and an *Endomicrobium* cell showing the inferred metabolic conversions of carbon sources based on gene content data. (C) Gene expression data of genes of interest (rows) pertaining to carbon metabolism in '*Ca*. Endomicrobium agilae' TA21 and TA26 (columns). (D) Maximum likelihood phylogenetic trees of amino acid sequences of the transporter (ExuT, using substitution model LG+F+G4) and isomerase (UxaC, using substitution model LG+I+G4) in the glucuronate metabolism pathway. Support values represent the Bayesian posterior probability and Bootstrap support values respectively.

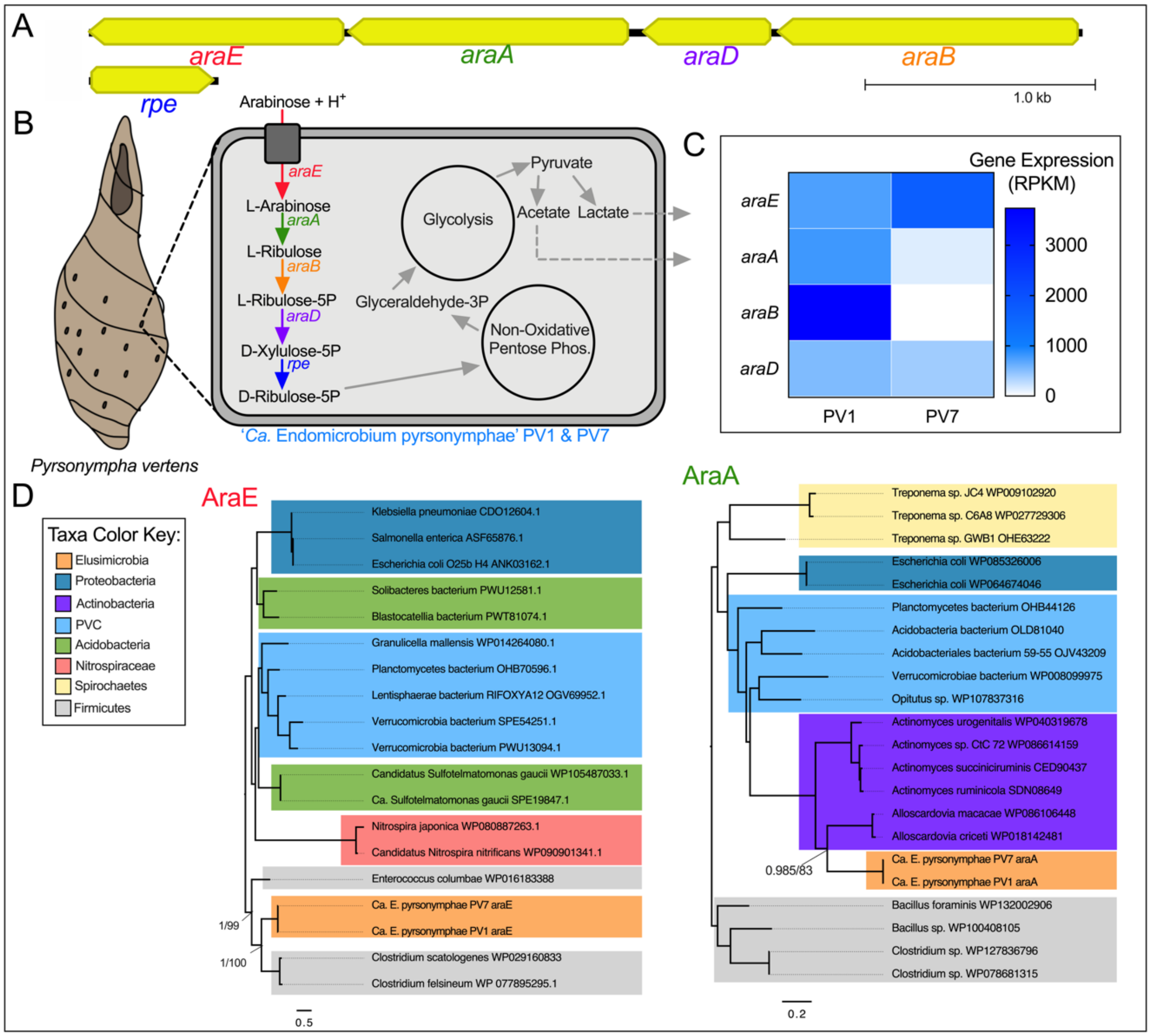

A
araE
araA
araD
araB
rpe
1.0 kb
B
Arabinose + H+
araE
L-Arabinose
araA
L-Ribulose
araB
L-Ribulose-5P
araD
D-Xylulose-5P
rpe
D-Ribulose-5P
Glycolysis
Pyruvate
Acetate
Lactate
Glyceraldehyde-3P
Non-Oxidative Pentose Phos.
'Ca. Endomicrobium pyrsonymphae' PV1 & PV7
Pyrsonympha vertens
C
Gene Expression (RPKM)
araE
araA
araB
araD
PV1
PV7
3000
2000
1000
0
D
Taxa Color Key:
Elusimicrobia
Proteobacteria
Actinobacteria
PVC
Acidobacteria
Nitrospiraceae
Spirochaetes
Firmicutes
AraE
Klebsiella pneumoniae CDO12604.1
Salmonella enterica ASF65876.1
Escherichia coli O25b H4 ANK03162.1
Solibacteres bacterium PWU12581.1
Blastocatellia bacterium PWT81074.1
Granulicella mallensis WP014264080.1
Planctomycetes bacterium OHB70596.1
Lentisphaerae bacterium RIFOXYA12 OGV69952.1
Verrucomicrobia bacterium SPE54251.1
Verrucomicrobia bacterium PWU13094.1
Candidatus Sulfotelmatomonas gaucii WP105487033.1
Ca. Sulfotelmatomonas gaucii SPE19847.1
Nitrospira japonica WP080887263.1
Candidatus Nitrospira nitrificans WP090901341.1
Enterococcus columbae WP016183388
Ca. E. pyrsonymphae PV7 araE
Ca. E. pyrsonymphae PV1 araE
Clostridium scatologenes WP029160833
Clostridium felsineum WP 077895295.1
1/99
1/100
0.5
AraA
Treponema sp. JC4 WP009102920
Treponema sp. C6A8 WP027729306
Treponema sp. GWB1 OHE63222
Escherichia coli WP085326006
Escherichia coli WP064674046
Planctomycetes bacterium OHB44126
Acidobacteria bacterium OLD81040
Acidobacteriales bacterium 59-55 OJV43209
Verrucomicrobiae bacterium WP008099975
Opitutus sp. WP107837316
Actinomyces urogenitalis WP040319678
Actinomyces sp. CtC 72 WP086614159
Actinomyces succiniciruminis CED90437
Actinomyces ruminicola SDN08649
Alloscardovia macacae WP086106448
Alloscardovia criceti WP018142481
Ca. E. pyrsonymphae PV7 araA
Ca. E. pyrsonymphae PV1 araA
0.985/83
Bacillus foraminis WP132002906
Bacillus sp. WP100408105
Clostridium sp. WP127836796
Clostridium sp. WP078681315
0.2

Figure 4. Carbon metabolism and HGT in '*Ca*. Endomicrobium pyrsonymphae'. (A) Gene neighborhood of the genes involved in the metabolism of arabinose in the '*Ca*. Endomicrobium pyrsonymphae' PV1 and PV7 genomes. (B) Diagram of a protist host and an *Endomicrobium* cell showing the inferred metabolic conversions of carbon sources based on gene content data. (C) Gene expression data of genes of interest (rows) pertaining to carbon metabolism in '*Ca*. Endomicrobium pyrsonymphae' PV1 and PV7 (columns). (D) Maximum likelihood phylogenetic trees of amino acid sequences from the transporter (AraE, using substitution model LG+F+G4) and isomerase (AraA, using substitution model LG+I+G4) in the arabinose metabolism pathway. Support values represent the Bayesian posterior probability and Bootstrap support values respectively.

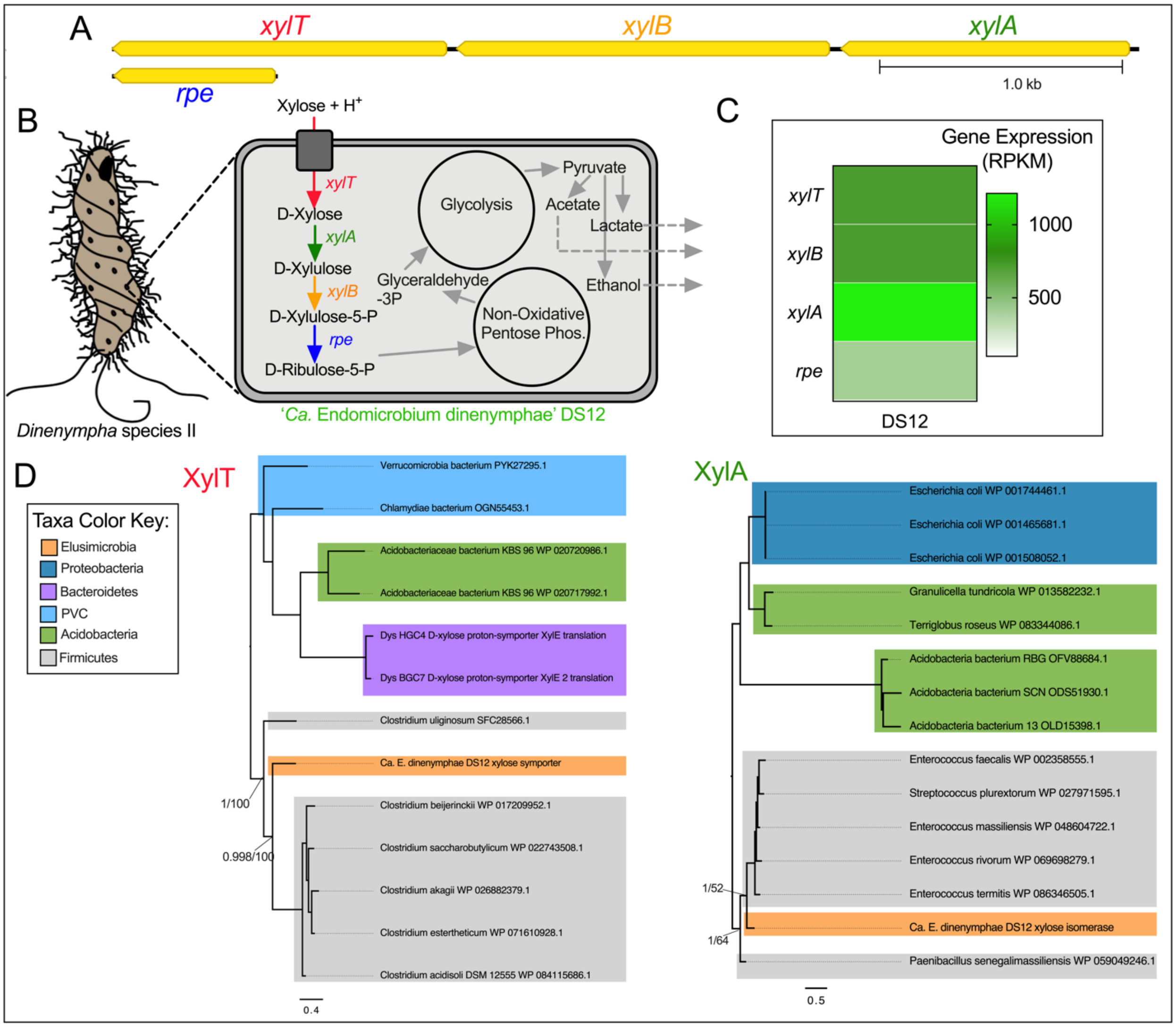

A
xylT
xylB
xylA
rpe
1.0 kb
B
Xylose + H+
xylT
D-Xylose
xylA
D-Xylulose
xylB
D-Xylulose-5-P
rpe
D-Ribulose-5-P
Glycolysis
Pyruvate
Acetate
Lactate
Ethanol
Glyceraldehyde
-3P
Non-Oxidative
Pentose Phos.
'Ca. Endomicrobium dinenymphae' DS12
Dinenympha species II
C
Gene Expression
(RPKM)
xylT
xylB
xylA
rpe
1000
500
DS12
D
XylT
Taxa Color Key:
🟧 Elusimicrobia
🟦 Proteobacteria
🟪 Bacteroidetes
🟦 PVC
🟩 Acidobacteria
⬜ Firmicutes
Verrucomicrobia bacterium PYK27295.1
Chlamydiae bacterium OGN55453.1
Acidobacteriaceae bacterium KBS 96 WP 020720986.1
Acidobacteriaceae bacterium KBS 96 WP 020717992.1
Dys HGC4 D-xylose proton-symporter XylE translation
Dys BGC7 D-xylose proton-symporter XylE 2 translation
Clostridium uliginosum SFC28566.1
Ca. E. dinenymphae DS12 xylose symporter
1/100
0.998/100
Clostridium beijerinckii WP 017209952.1
Clostridium saccharobutylicum WP 022743508.1
Clostridium akagii WP 026882379.1
Clostridium esterthеticum WP 071610928.1
Clostridium acidisoli DSM 12555 WP 084115686.1
0.4
XylA
Escherichia coli WP 001744461.1
Escherichia coli WP 001465681.1
Escherichia coli WP 001508052.1
Granulicella tundricola WP 013582232.1
Terriglobus roseus WP 083344086.1
Acidobacteria bacterium RBG OFV88684.1
Acidobacteria bacterium SCN ODS51930.1
Acidobacteria bacterium 13 OLD15398.1
Enterococcus faecalis WP 002358555.1
Streptococcus plurextorum WP 027971595.1
Enterococcus massiliensis WP 048604722.1
Enterococcus rivorum WP 069698279.1
Enterococcus termitis WP 086346505.1
1/52
Ca. E. dinenymphae DS12 xylose isomerase
1/64
Paenibacillus senegalimassiliensis WP 059049246.1
0.5

Figure 5. Carbon metabolism and HGT in '*Ca*. Endomicrobium dinenymphae'. (A) Gene neighborhood of the genes involved in the metabolism of xylose in the '*Ca*. Endomicrobium dinenymphae' DS12 genome. (B) Diagram of a protist host and an *Endomicrobium* cell showing the inferred metabolic conversions of carbon sources based on gene content data. (C) Gene expression data of genes of interest (rows) pertaining to carbon metabolism in '*Ca*. Endomicrobium dinenymphae' DS12 (column). (D) Maximum likelihood phylogenetic trees of amino acid sequences from the transporter (XylT, using substitution model LG+F+G4) and isomerase (XylA, using substitution model LG+G4) in the xylose metabolism pathway. Support values represent the Bayesian posterior probability and Bootstrap support values respectively.

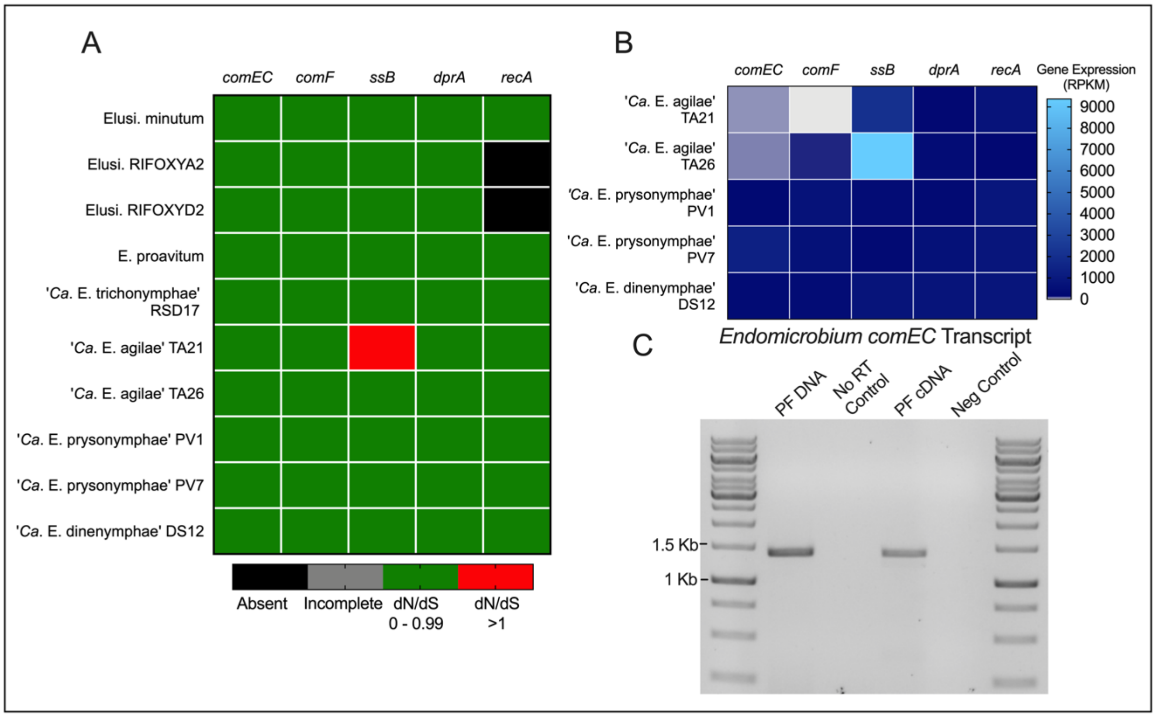


Figure 6. Analysis of genes involved in a putative competence pathway in *Endomicrobium* spp. (A) Heatmap showing the results of dN/dS analyses of genes involved in competence and recombination (columns) from *Endomicrobium* spp. and *Elusimicrobium* relatives (rows). (B) Gene expression data of those genes (columns) in the *Endomicrobium* spp. (rows) presented in this study. (C) RT-PCR gel image of *Endomicrobium comEC* transcript. Samples consisted of protist fraction (PF) DNA (positive control), No RT Control, PF cDNA, and molecular grade water (Negative control). Accession numbers for reference genomes used can be found in Supplementary Table 2.

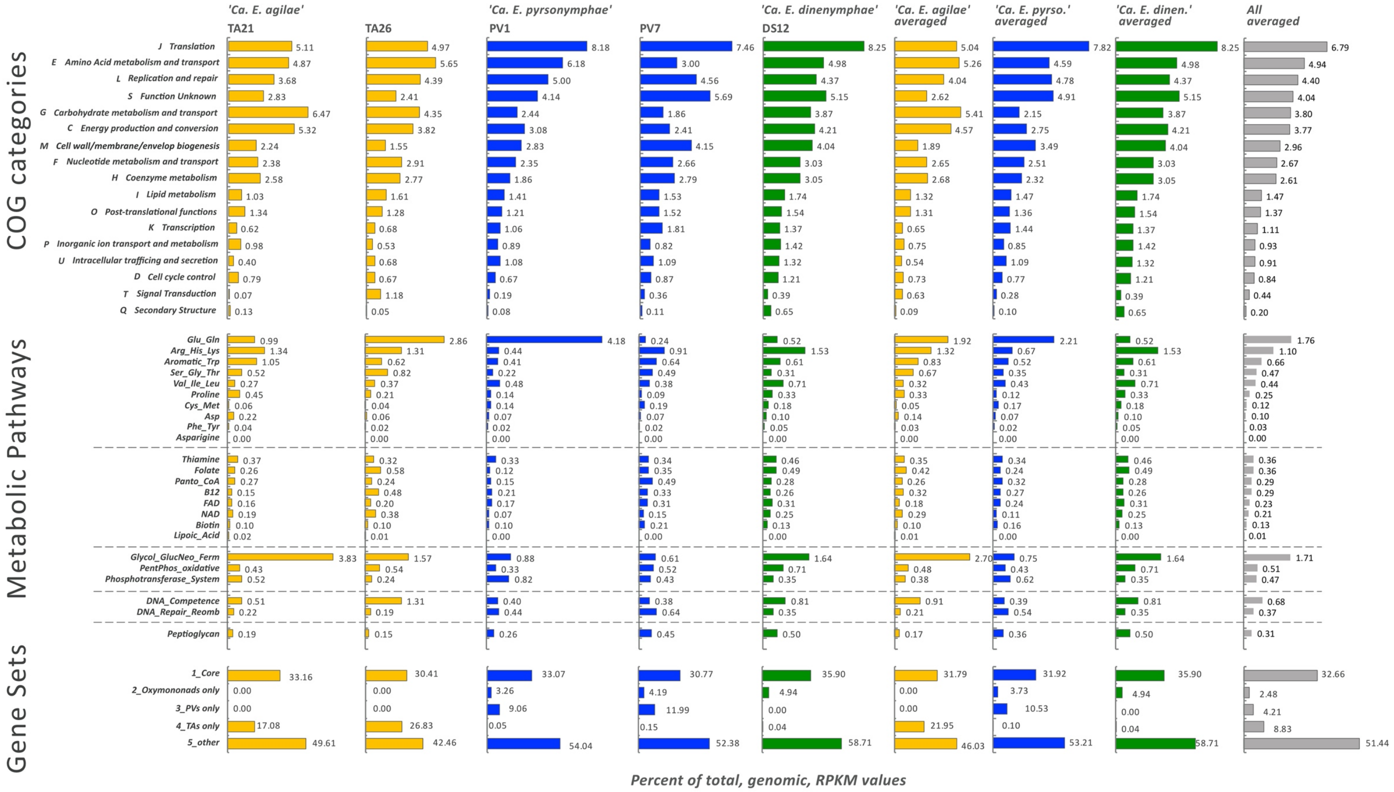
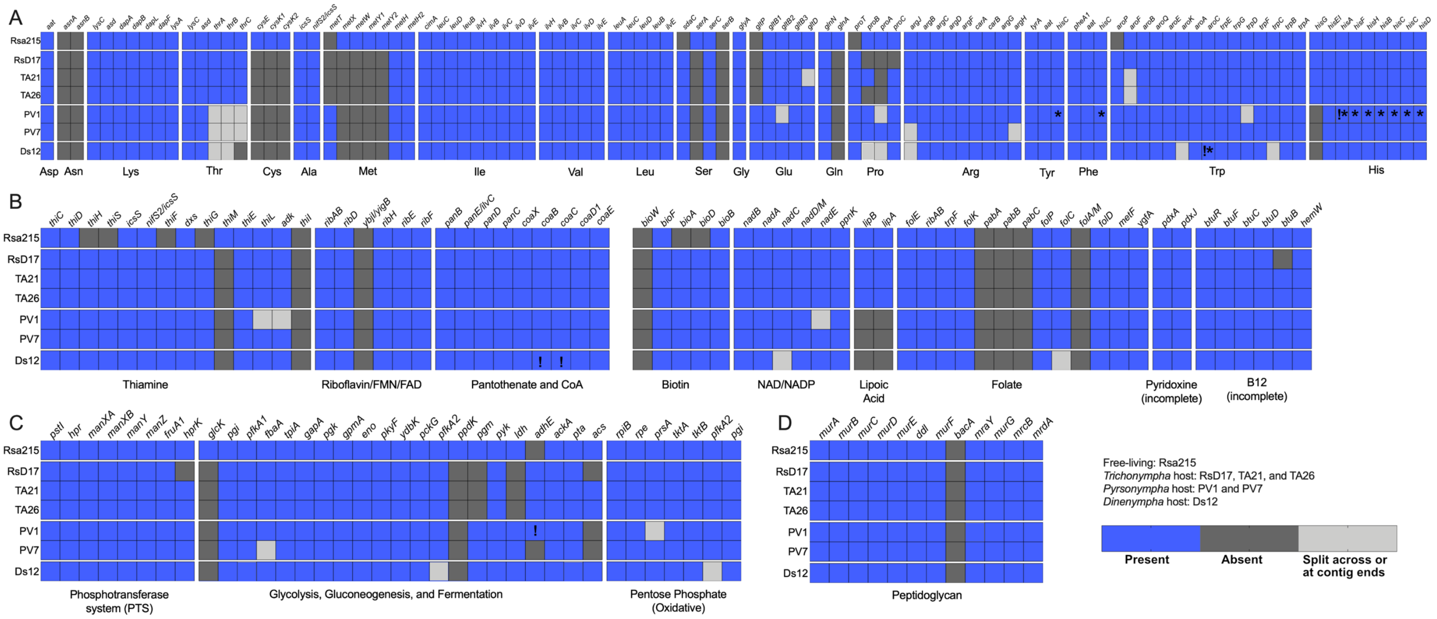
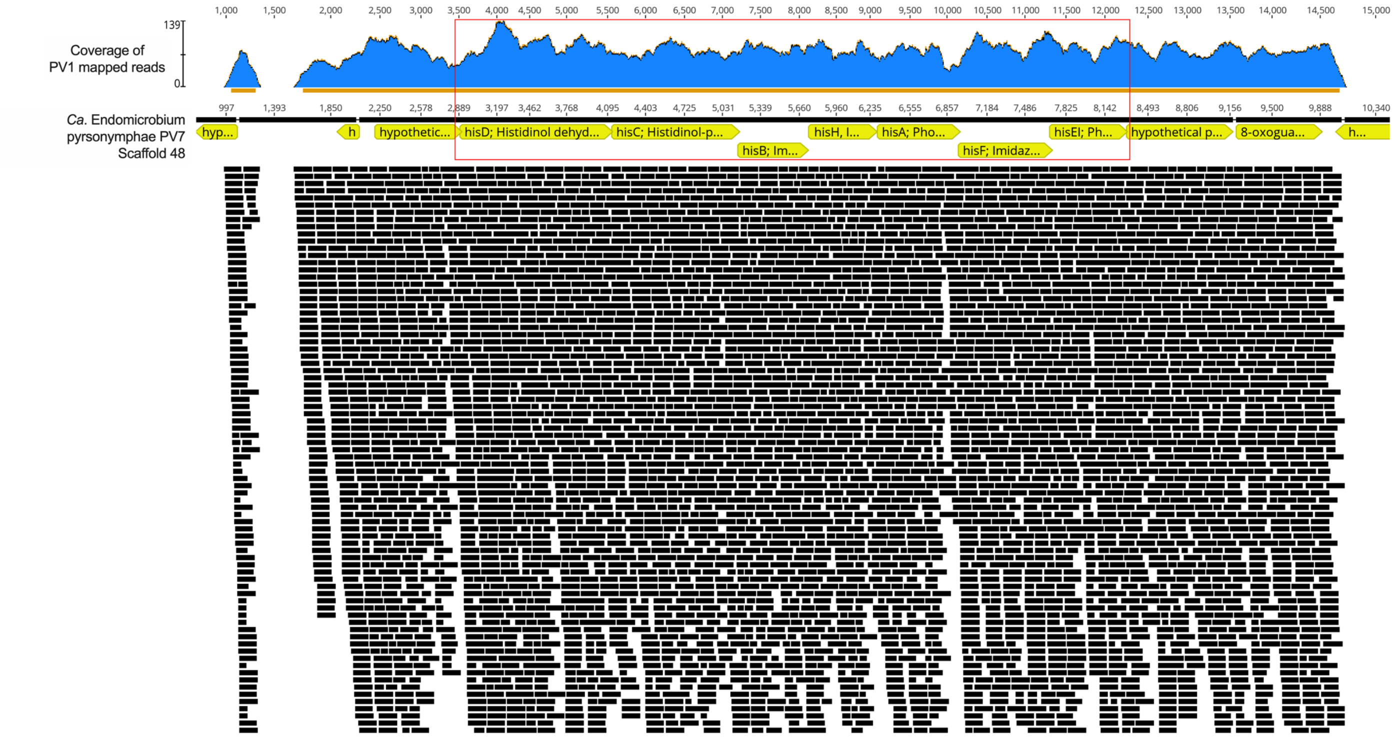


Figure 7. Transcriptome analysis of *Endomicrobium* populations from individual protist cells. (Top) Expression analysis of genes grouped into clusters of orthologous groups (COG) functional categories. (Middle) Expression analysis of genes in certain metabolic pathways pertaining to amino acid and cofactor biosynthesis, DNA processing, and peptidoglycan biosynthesis. (Bottom) Expression of genes grouped by their distributions among the different *Endomicrobium* species. RPKM values for individual cells, and averages of genera and all cells are in the various columns.

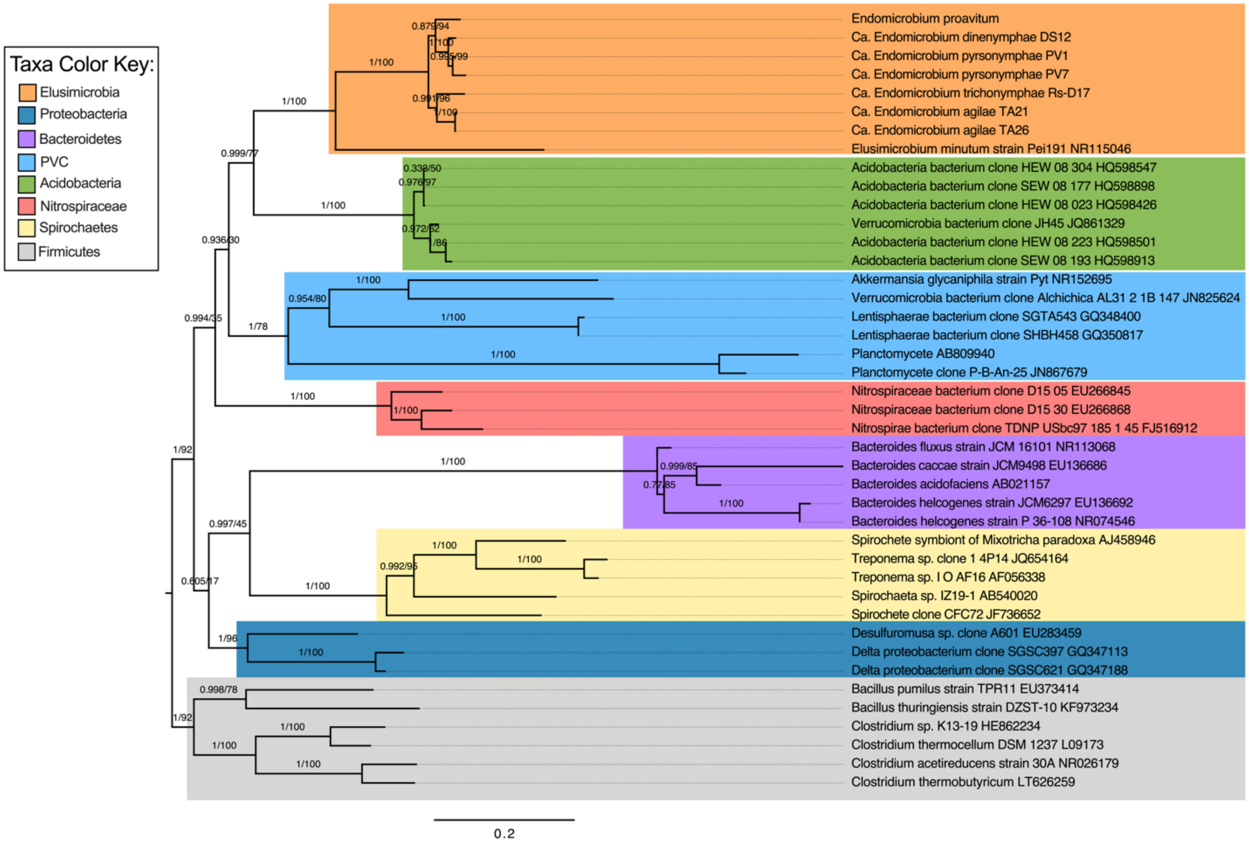
783 ## Supplementary Figures and Legends



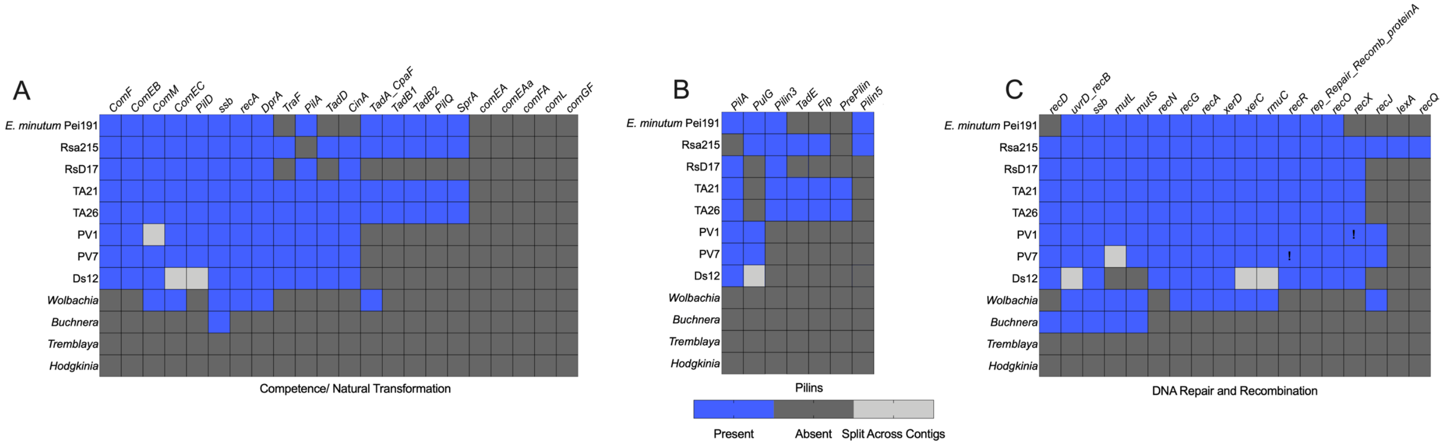
784 **Supplementary Figure 1. Gene content in *Endomicrobium* spp. genomes regarding metabolic functions and cell wall biosynthesis.** (A) Gene content of amino acid biosynthesis pathways. (B) Gene content of vitamins and co-factor biosynthesis pathways. (C) Genes involved in central metabolism. (D) Gene content of peptidoglycan biosynthesis. Note that the genes marked with a "*" or a "!" were not in the final assemblies, but their reads were detected by either Megan (!) or by read mapping onto the same gene from another closely related organisms (*) or both (!*) (see Supplementary Figure 2 as an example).



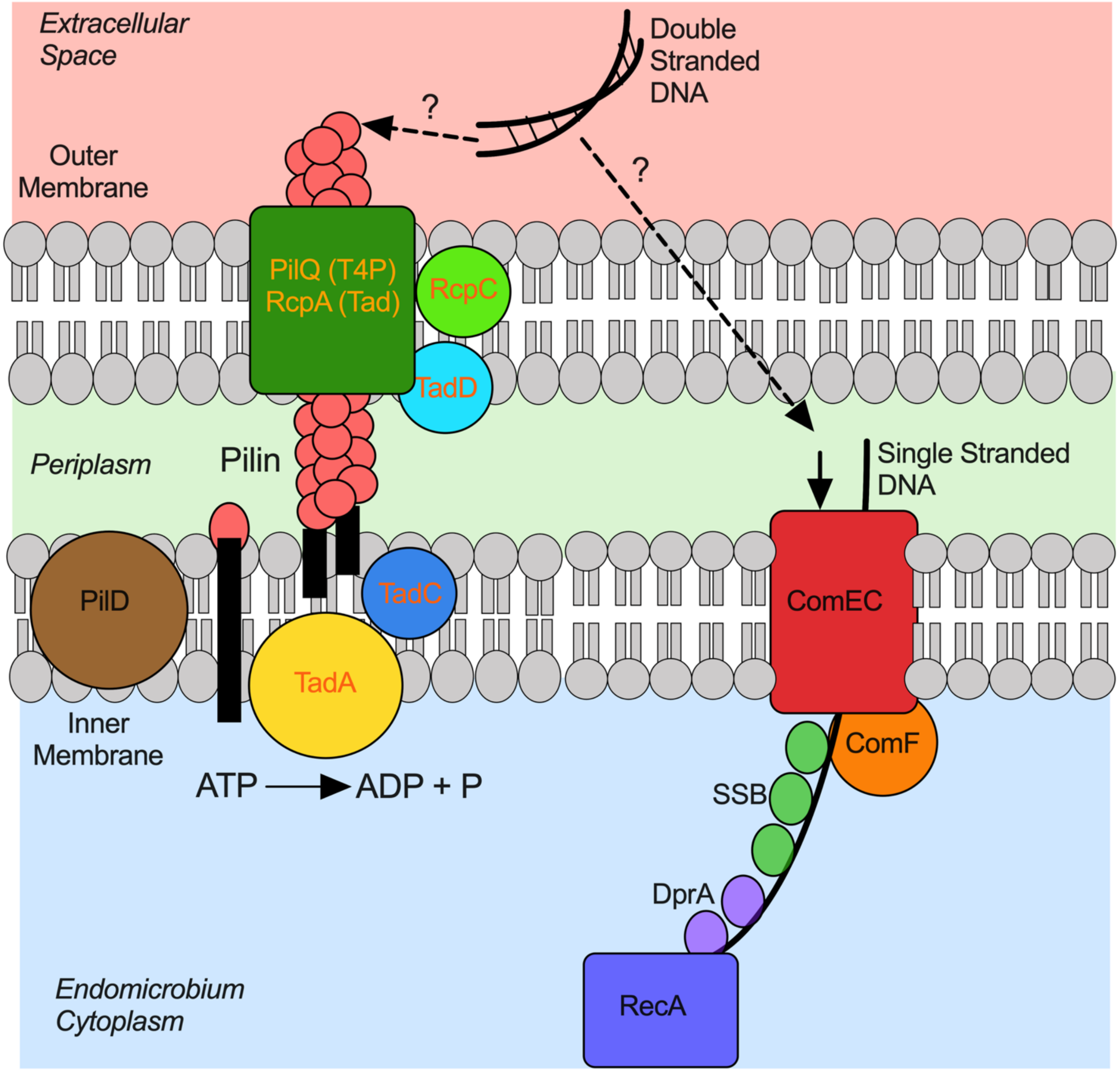
Supplementary Figure 2. Mapping coverage of the histidine biosynthesis pathway in '*Ca.* Endomicrobium pyrsonymphae' PV1. Metagenomic reads from sample PV1 were mapped to the draft genome of '*Ca*. Endomicrobium pyrsonymphae PV7'. The resulting coverage of these genes indicate that '*Ca*. Endomicrobium pyrsonymphae PV1' encoded the histidine biosynthesis pathway and that the reason those genes are missing from the draft genome is likely due to an artifact of assembly or binning.


Taxa Color Key:
🟧 Elusimicrobia
🟦 Proteobacteria
🟪 Bacteroidetes
🟦 PVC
🟩 Acidobacteria
🟥 Nitrospiraceae
🟨 Spirochaetes
⬜ Firmicutes
Endomicrobium proavitum
Ca. Endomicrobium dinenymphae DS12
Ca. Endomicrobium pyrsonymphae PV1
Ca. Endomicrobium pyrsonymphae PV7
Ca. Endomicrobium trichonymphae Rs-D17
Ca. Endomicrobium agilae TA21
Ca. Endomicrobium agilae TA26
Elusimicrobium minutum strain Pei191 NR115046
Acidobacteria bacterium clone HEW 08 304 HQ598547
Acidobacteria bacterium clone SEW 08 177 HQ598898
Acidobacteria bacterium clone HEW 08 023 HQ598426
Verrucomicrobia bacterium clone JH45 JQ861329
Acidobacteria bacterium clone HEW 08 223 HQ598501
Acidobacteria bacterium clone SEW 08 193 HQ598913
Akkermansia glycaniphila strain Pyt NR152695
Verrucomicrobia bacterium clone Alchichica AL31 2 1B 147 JN825624
Lentisphaerae bacterium clone SGTA543 GQ348400
Lentisphaerae bacterium clone SHBH458 GQ350817
Planctomycete AB809940
Planctomycete clone P-B-An-25 JN867679
Nitrospiraceae bacterium clone D15 05 EU266845
Nitrospiraceae bacterium clone D15 30 EU266868
Nitrospirae bacterium clone TDNP USbc97 185 1 45 FJ516912
Bacteroides fluxus strain JCM 16101 NR113068
Bacteroides caccae strain JCM9498 EU136686
Bacteroides acidofaciens AB021157
Bacteroides helcogenes strain JCM6297 EU136692
Bacteroides helcogenes strain P 36-108 NR074546
Spirochete symbiont of Mixotricha paradoxa AJ458946
Treponema sp. clone 1 4P14 JQ654164
Treponema sp. I O AF16 AF056338
Spirochaeta sp. IZ19-1 AB540020
Spirochete clone CFC72 JF736652
Desulfuromusa sp. clone A601 EU283459
Delta proteobacterium clone SGSC397 GQ347113
Delta proteobacterium clone SGSC621 GQ347188
Bacillus pumilus strain TPR11 EU373414
Bacillus thuringiensis strain DZST-10 KF973234
Clostridium sp. K13-19 HE862234
Clostridium thermocellum DSM 1237 L09173
Clostridium acetireducens strain 30A NR026179
Clostridium thermobutyricum LT626259
0.879/94
1/100
0.995/99
1/100
1/100
0.991/96
1/100
0.999/77
0.333/50
0.976/97
1/100
0.972/62
1/86
0.936/30
1/100
0.954/80
1/100
0.994/35
1/78
1/100
1/100
1/100
1/92
1/100
0.999/85
0.77/85
1/100
0.997/45
1/100
1/100
0.992/95
0.605/17
1/100
1/96
1/100
0.998/78
1/92
1/100
1/100
1/100
0.2

Supplemental Figure 3. 16S rRNA phylogeny of Elusimicrobia with respect to other phyla. Maximum likelihood phylogenetic tree of 16S rRNA genes (using substitution model TPM3u + G4) which was used a marker-gene to establish an organismal phylogeny of the Elusimicrobia phylum. This phylogeny was used to determine incongruence between the 16S rRNA gene and other genes of interest that may have been acquired via HGT by the *Endomicrobium* species.



Supplementary Figure 4. Gene content of regarding genes involved in natural transformation, pilus assembly, and DNA recombination/repair. (A) Presence and absence matrices of genes involved in natural transformation, (B) pilus assembly, and (C) DNA repair and recombination found the endosymbiotic *Endomicrobium* spp. genomes. The presence of these genes was investigated in their free-living relatives (Rsa215 and *E. minutum* Pei191) and other endosymbiotic bacteria. Accession numbers for references genomes used in this analysis are provided in Supplementary Table 4. Note that the genes marked with a "*" or a "!" were not in the final assemblies, but their reads were detected by either Megan (!) or by read mapping onto the same gene from another closely related organisms (*) or both (!*) (see Supplementary Figure 2 as an example).



Supplemental Figure 5. Graphical summary of a putative competence pathway and proteins involved in pilus assembly in *Endomicrobium* species. Proteins shared by all *Endomicrobium* species are in black font while those that are only retained in '*Ca*. E. agilae' are colored in orange. All *Endomicrobium* possessed the pre-pilin peptidase (PilD) and one or more genes that encode pilins. '*Ca*. E. agilae' possessed a near-complete *tad* locus/ T2SS as well the T4P secretin (PilQ).

Supplementary File S1 spreadsheet contains the following:
Sup. Table 1 Meta-Data
Sup. Table 2 Filtering ref.
Sup. Table 3 Read Numbers
Sup. Table 4 Genome refs.
Sup. Table 5 RPKM values
Sup. Table 6 RPKM charts
Sup. Table 7 Type IV tad genes
Sup. Table 8 Biosynthetic genes
Sup. Table 9 Repair_Compet genes