Comparison of Intracellular "*Ca.* Endomicrobium Trichonymphae" Genomovars Illuminates the Requirement and Decay of Defense Systems against Foreign DNA

Kazuki Izawa[1], Hirokazu Kuwahara[1], Kumiko Kihara[1,2], Masahiro Yuki[3], Nathan Lo[4], Takehiko Itoh[5], Moriya Ohkuma[3,6], and Yuichi Hongoh[1,6,*]

[1]Department of Biological Sciences, Tokyo Institute of Technology, Tokyo, Japan

[2]Department of Biological and Chemical Systems Engineering, National Institute of Technology Kumamoto College, Yatsushiro, Japan

[3]Biomass Research Platform Team, RIKEN Biomass Engineering Program Cooperation Division, RIKEN Center for Sustainable Resource Science, Tsukuba, Japan

[4]School of Biological Sciences, University of Sydney, Sydney, NSW, Australia

[5]Department of Biological Information, Tokyo Institute of Technology, Tokyo, Japan

[6]Japan Collection of Microorganisms, RIKEN BioResource Center, Tsukuba, Japan

*Corresponding author: E-mail: yhongo@bio.titech.ac.jp.

Abstract

"*Candidatus* Endomicrobium trichonymphae" (*Bacteria*; *Elusimicrobia*) is an obligate intracellular symbiont of the cellulolytic protist genus *Trichonympha* in the termite gut. A previous genome analysis of "*Ca.* Endomicrobium trichonymphae" phylotype Rs-D17 (genomovar Ri2008), obtained from a *Trichonympha agilis* cell in the gut of the termite *Reticulitermes speratus*, revealed that its genome is small (1.1 Mb) and contains many pseudogenes; it is in the course of reductive genome evolution. Here we report the complete genome sequence of another Rs-D17 genomovar, Ti2015, obtained from a different *T. agilis* cell present in an *R. speratus* gut. These two genomovars share most intact protein-coding genes and pseudogenes, showing 98.6% chromosome sequence similarity. However, characteristic differences were found in their defense systems, which comprised restriction-modification and CRISPR/Cas systems. The repertoire of intact restriction-modification systems differed between the genomovars, and two of the three CRISPR/Cas loci in genomovar Ri2008 are pseudogenized or missing in genomovar Ti2015. These results suggest relaxed selection pressure for maintaining these defense systems. Nevertheless, the remaining CRISPR/Cas system in each genomovar appears to be active; none of the "spacer" sequences (112 in Ri2008 and 128 in Ti2015) were shared whereas the "repeat" sequences were identical. Furthermore, we obtained draft genomes of three additional endosymbiotic *Endomicrobium* phylotypes from different host protist species, and discovered multiple, intact CRISPR/Cas systems in each genome. Collectively, unlike bacteriome endosymbionts in insects, the *Endomicrobium* endosymbionts of termite-gut protists appear to require defense against foreign DNA, although the required level of defense has likely been reduced during their intracellular lives.

Key words: insect, termite, endosymbiosis, gut bacteria, restriction–modification system, CRISPR.

Introduction

Many invertebrates and single-celled eukaryotes (protists) harbor intracellular symbionts, which generally contribute to the host's nutrition and/or defense systems (Moran et al. 2008; Hongoh 2011; Nakabachi et al. 2013). The genus *Endomicrobium* (phylum *Elusimicrobia*, formerly Termite Group 1; class *Endomicrobia*) is one of the predominant bacterial groups in the gut of phylogenetically basal ("lower")

termites (Ohkuma and Kudo 1996; Hongoh et al. 2003, 2005; Yang et al. 2005; Boucias et al. 2013; Dietrich et al. 2014; Abdul Rahman et al. 2015; Tai et al. 2015). The majority of *Endomicrobium* species are found as intracellular symbionts of various protist species in the termite gut (Stingl et al. 2005; Ikeda-Ohtsubo et al. 2007; Ohkuma et al. 2007). No free-living phase has been observed for these endosymbionts, and the relationship with the protist host is strictly species-specific (Ikeda-Ohtsubo et al. 2007; Ohkuma et al. 2007; Zheng et al. 2015). A sole cultured representative of this genus is a free-living species, *Endomicrobium proavitum*, which has been isolated from a termite gut (Zheng, Dietrich, Radek, et al. 2016). No endosymbiotic *Endomicrobium* species has been cultivated thus far.

"*Candidatus* Endomicrobium trichonymphae" is an intracellular symbiont of the protist genus *Trichonympha* (phylum *Parabasalia*) (Stingl et al. 2005; Ohkuma et al. 2007). *Trichonympha* species are widespread in the gut of lower termites and play a pivotal role in cellulose digestion (Yamin 1979, 1981). Previous phylogenetic analysis showed that *Trichonympha* species and their *Endomicrobium* endosymbionts have strictly cospeciated, and it was estimated that their symbiosis commenced 73–38 million years ago (Ikeda-Ohtsubo and Brune 2009). Among the 16S rRNA phylotypes of "*Ca.* Endomicrobium trichonymphae", phylotype Rs-D17 specifically inhabits the cytoplasm of *Trichonympha agilis* in the gut of the termite *Reticulitermes speratus* (Hongoh et al. 2003; Ohkuma et al. 2007). The complete genome sequence of phylotype Rs-D17 was previously obtained using a whole genome amplification (WGA) technique from a single host *T. agilis* cell (Hongoh et al. 2008a). The genome analysis suggested that the endosymbiont imports glucose 6-phosphate as a carbon and energy source, which is expected to be abundant in the cellulolytic host cytoplasm. Phylotype Rs-D17, in turn, likely provides amino acids and cofactors, which are deficient in dead wood, to their host protists and termites. The Rs-D17 genome is small (1.1 Mb) and corresponds to about two-thirds of the genome size of the free-living isolate *Endomicrobium proavitum* (1.6 Mb) (Hongoh et al. 2008a; Zheng and Brune 2015). In addition, the Rs-D17 genome contains 121 pseudogenes, which account for approximately 15% of the predicted protein-coding sequences (CDSs) (Hongoh et al. 2008a). These suggest that phylotype Rs-D17 is in the course of reductive genome evolution, as seen in many other intracellular bacterial symbionts (McCutcheon and Moran 2012).

In the present study, we reconstructed the complete genome sequence of another genomovar, designated as Ti2015, of phylotype Rs-D17 from a different *T. agilis* cell present in an *R. speratus* worker's gut. To discriminate the two genomovars, we designated the Rs-D17 genome obtained in the previous study (Hongoh et al. 2008a) as genomovar Ri2008. By comparing these two closely related genomovars,

we aimed to elucidate the evolutionary process of the endosymbionts within a short time scale.

Comparison of General Features of Genomovars Ti2015 and Ri2008

The Ti2015 genome comprised a circular chromosome and three circular plasmids (table 1), which showed very high sequence similarities to the respective genome components of genomovar Ri2008 (fig. 1 and supplementary fig. S1, Supplementary Material online). The two genomovars had identical 16S rRNA gene sequences and shared 97.0% sequence identity in the internal transcribed spacer (ITS) regions of the rRNA operon. The overall chromosome nucleotide sequence identity was 98.6%, and the nucleotide sequence identity in the CDSs was 99.6% on average. The chromosome size of Ti2015 is 11.3 kb smaller than that of Ri2008 (table 1). This difference in genome size is largely attributed to indels longer than 1 kb that are distributed around the entire genomes (fig. 1). The number of intact CDSs on the chromosome of Ti2015 was lower (719 vs. 761) and the number of pseudogenes was higher (150 vs. 121) (table 1). All of the 719 CDSs in Ti2015 are present in Ri2008. The frequency of small indels < 3 bp in the pseudogenes was as high as that in the intergenic regions (supplementary table S1, Supplementary Material online). The two genomes showed a high level of synteny except for certain regions in the plasmids and a large inversion (*ca.* 200 kb) in the chromosome (fig. 1 and supplementary fig. S1, Supplementary Material online). This inversion occurred around the predicted chromosome-dimer resolution site (*dif*) near the replication terminus site (supplementary fig. S2, Supplementary Material online). Inversions around the terminus site have occasionally been reported

Table 1

Comparison of Genome Features of Genomovars Ti2015 (This Study) and Ri2008 (Hongoh et al. 2008a)

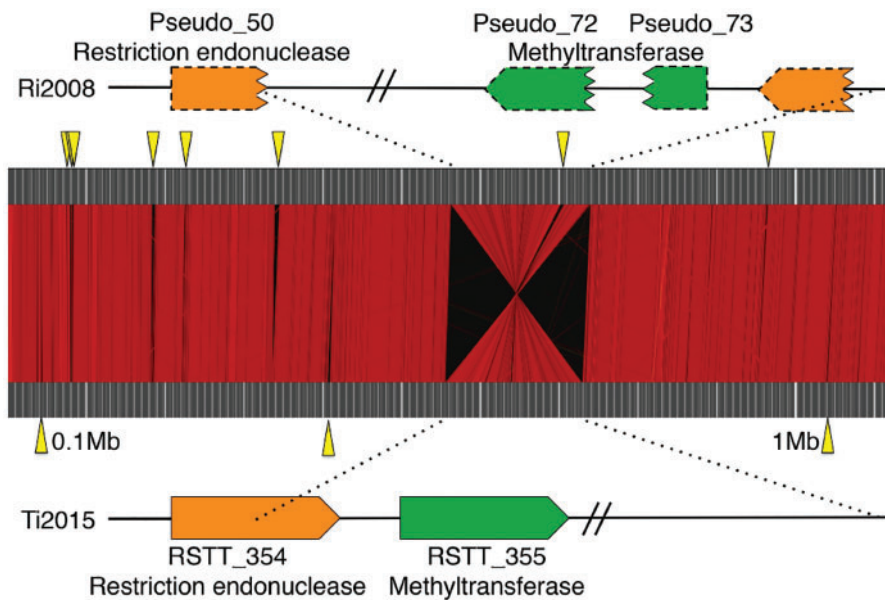|	**Ti2015**	**Ri2008**
Chromosome	1,114,532 bp	1,125,857 bp
G + C content	35.3%	35.2%
Predicted CDSs	719	761
Pseudogenes	150	121
tRNA genes	45	45
rRNA operon	1	1
Plasmid 1	5,518 bp	11,650 bp
G + C content	33.3%	34.3%
Predicted CDSs	6	9
Plasmid 2	5,752 bp	5,701 bp
G + C content	33.0%	35.4%
Predicted CDSs	10	3
Plasmid 3	5,038 bp	5,362 bp
G + C content	32.1%	32.6%
Predicted CDSs	4	3

Fɪɢ. 1.—Synteny of chromosomes between genomovars Ri2008 and Ti2015. Upper and lower columns indicate the chromosomes of genomovars Ri2008 and Ti2015, respectively. Red lines show the regions with ≥90% nucleotide sequence identity between the genomovars. Yellow wedges indicate the positions of deletions ≥1kb length. A schematic view of the pseudogenization of an R–M system caused by a large inversion is also shown. A restriction endonuclease gene in Ti2015 (RSTT_354: orange box) was split into two parts by the inversion and found as pseudogene fragments in Ri2008 (two dashed orange boxes). An adjacent methyltransferase gene (RSTT_355: green box) was also split into two and found as pseudogene fragments in Ri2008 (two dashed green boxes). The length of the boxes indicates the relative sequence length. Wavy ends of CDSs indicate split sites. Other genes around the inverted region are not shown, and omitted regions are indicated by double slashes.

between closely related bacterial strains, although the mechanism is unknown (Eisen et al. 2000; Tillier and Collins 2000). Recently, it has been suggested that restriction–modification (R–M) systems might have caused the massive genome rearrangements present between "Ca. Endomicrobium trichonymphae" and Endomicrobium proavitum (Zheng, Dietrich, Hongoh, et al. 2016). We could not find evidence for involvement of R–M systems in this large inversion or other rearrangements between the Ti2015 and Ri2008 genomes.

Most of the genes and pseudogenes were shared by the two genomovars, including the pseudogenes of the chromosome replication initiator protein DnaA, ribosomal protein S1 (RpsA) and glutamine synthetase (GlnA) (supplementary table S2, Supplementary Material online) (Hongoh et al. 2008a). On the other hand, clear differences were found in the defense systems comprising restriction–modification systems and clustered regularly interspaced short palindromic repeat (CRISPR) regions with its associated genes (cas genes).

Decay and Conservation of Defense Systems

Genomovars Ti2015 and Ri2008 possess three or two intact R–M systems, respectively, in addition to numerous

pseudogenized R–M systems (supplementary fig. S3 and table S2, Supplementary Material online) (Hongoh et al. 2008a). Interestingly, the repertoire of the intact R–M systems completely differs between the genomovars (table 2 and supplementary fig. S3, Supplementary Material online). For example, a type II R–M system (RSTT_354 and RSTT_355) of genomovar Ti2015 is pseudogenized in genomovar Ri2008, where the restriction endonuclease gene was split by the large inversion mentioned above (fig. 1). This gene corruption likely allowed the pseudogenization of the adjacent DNA methyltransferase (RSTT_355) in Ri2008. Likewise, many R–M system-related genes are pseudogenized in only one of the two genomovars (table 2), whereas most other intact CDSs and pseudogenes are shared. This suggests that the decay of R–M systems has occurred randomly and has proceeded rapidly, compared with other genes.

The CRISPR/Cas system is an adaptive and heritable defense mechanism: foreign DNA fragmentized by the action of Cas proteins are recorded as "spacers" flanked by "repeat" sequences in a CRISPR region (Karginov and Hannon 2010). Genomovar Ri2008 has three CRISPR/Cas systems (fig. 2). According to the classification proposed by Makarova et al. (2011), one system belongs to type I-C and the other two belong to type II-C. The two sets of the type II-C cas genes are phylogenetically distinct and designated here as type II-C-a

Table 2
List of Intact Genes Present in Only One of the Two Genomovars Ti2015 and Ri2008

Gene ID in Ti2015*	Gene ID in Ri2008*	Predicted functions
RSTT_pseudo_010	TGRD_038	Type I R–M system, substrate recognition subunit
RSTT_pseudo_011	TGRD_039	Type I R–M system, restriction endonuclease
RSTT_pseudo_013	TGRD_041	Type II R–M system, methyltransferase
RSTT_pseudo_015	TGRD_051	Putative phosphatase/phosphohexomutase
RSTT_pseudo_016	TGRD_052	Chromosome partitioning protein (ParA-like)
RSTT_pseudo_017	TGRD_055 + 056	CRISPR-associated protein Csn1
RSTT_pseudo_018	TGRD_057	CRISPR-associated protein Cas1
RSTT_pseudo_019	TGRD_058	CRISPR-associated protein Cas2
RSTT_pseudo_021	TGRD_061	Putative ABC transporter, ATPase component
RSTT_pseudo_022	TGRD_063**	tRNA CCA-adding nucleotidyltransferase
RSTT_pseudo_025	TGRD_126	Type III R–M system, methyltransferase
RSTT_pseudo_027	TGRD_127	Type III R–M system, restriction endonuclease
RSTT_pseudo_037	TGRD_173	M23B family peptidase
RSTT_171	Pseudo_23	Type I R–M system, methyltransferase
RSTT_173	Pseudo_24	Type I R–M system, restriction endonuclease
–	TGRD_222	CRISPR-associated protein Csn1
–	TGRD_223	CRISPR-associated protein Cas1
–	TGRD_224	CRISPR-associated protein Cas2
RSTT_249	Pseudo_29,30	Site-specific DNA methyltransferase
RSTT_pseudo_056	TGRD_307	Putative lipoprotein containing TPR domain
RSTT_354	Pseudo_50	Type II restriction endonuclease
RSTT_355	Pseudo_72 + 73	Type II modification methyltransferase
RSTT_368	Pseudo_69	DNA methyltransferase
RSTT_pseudo_072	TGRD_493	N-(5'-phosphoribosyl) anthranilate isomerase
RSTT_pseudo_074	TGRD_490**	OstA-like outer membrane protein
RSTT_381	Pseudo_67	Type III R–M system, DNA methyltransferase
RSTT_382	Pseudo_66	Type III R–M system, restriction endonuclease
RSTT_385	Pseudo_64	DNA helicase II
RSTT_pseudo_085	TGRD_414	Adenine-specific DNA methyltransferase
RSTT_458	Pseudo_54	Adenine-specific DNA methyltransferase
RSTT_pseudo_088	TGRD_389	ABC transporter, ATPase
RSTT_pseudo_090	TGRD_508	Multidrug-resistance ABC transporter, ATPase
RSTT_pseudo_092	TGRD_515	Adenine/cytosine-specific DNA methyltransferase
RSTT_pseudo_094	TGRD_516	DNA repair protein RecN
RSTT_pseudo_097	TGRD_557**	3-Oxoacyl-[acyl-carrier-protein] synthase II
RSTT_pseudo_126	TGRD_647	Lipoprotein ABC transporter LolD
RSTT_610	Pseudo_101	Flp pillus assembly protein CpaB-like
RSTT_625	Pseudo_105	Adenine-specific DNA methyltransferase
RSTT_pseudo_141	TGRD_719	ADP-heptose synthase
RSTT_pseudo_149	TGRD_745	DNA modification methyltransferase

*Pseudogenes are shown in red. Genes related to R–M systems or CRISPR/Cas systems are highlighted in green and yellow, respectively. Hypothetical genes are not shown in this list.

**Genes redundant in the Ri2008 genome.

and b, respectively, for convenience (fig. 3 and supplementary fig. S4, Supplementary Material online). Genomovar Ti2015 retained the type I-C system, but the type II-C-a is pseudogenized and the type II-C-b is completely missing (fig. 2). The *cas9* gene of the type II-C-a in Ri2008 has been split into two parts, which probably caused the loss of the adjacent CRISPR region as in Ti2015. Taken together with the ongoing decay of the R–M systems, these results suggest that the defense systems of these bacteria have been under relaxed selection pressure,

likely because of their intracellular lifestyle, which affords protection within host eukaryotic cells.

Nevertheless, none of the 112 and 128 "spacer" sequences of the type I-C CRISPR/Cas system in Ri2008 and Ti2015, respectively, was shared, whereas the "repeat" sequences are identical (supplementary table S3, Supplementary Material online). This indicates that at least the type I-C CRISPR/Cas system is actively working in "*Ca.* Endomicrobium trichonymphae" phylotype Rs-D17. Indeed,
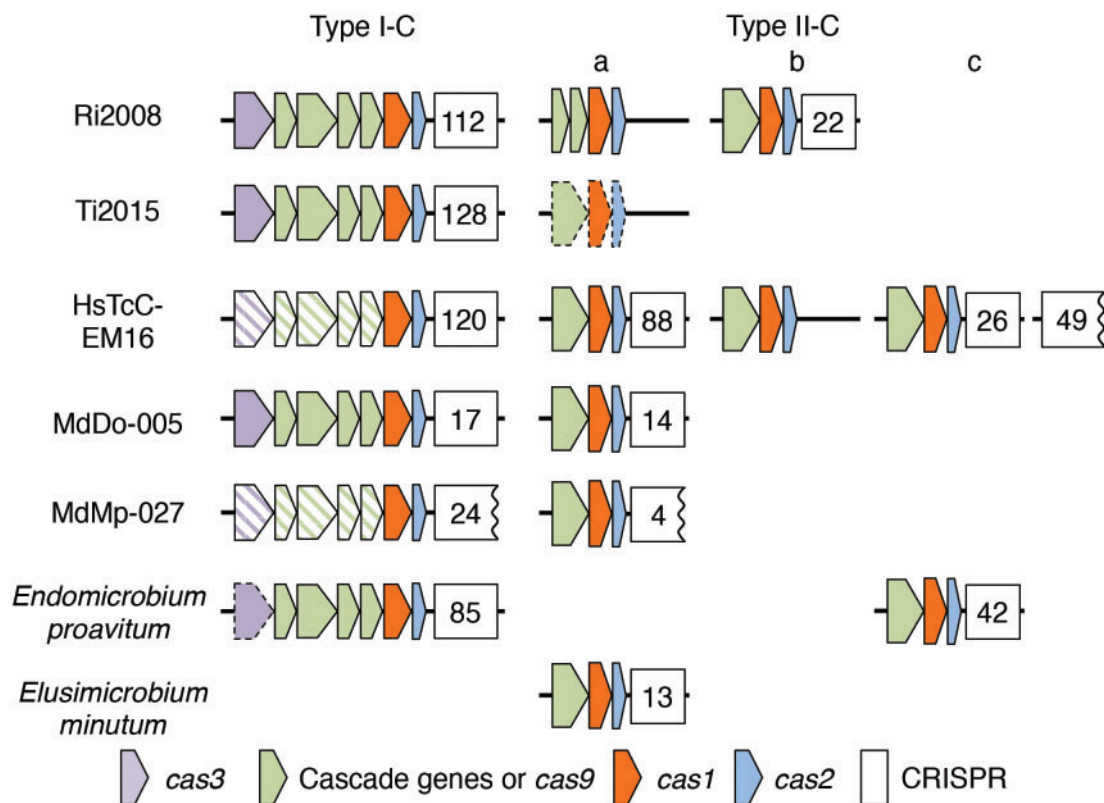
Fig. 2.—CRISPR/Cas systems in the phylum *Elusimicrobia*. Boxes indicate *cas* genes and CRISPR regions, whereas dashed boxes indicate *cas* pseudo-genes. Number of "spacers" is shown in the CRISPR regions. White boxes with a wavy end are partial CRISPR loci on contigs in draft genomes. The boxes representing genes are depicted according to the relative sequence length. Genes shown with shaded boxes in HsTcC-EM16 and MdMp-027 have been replaced by homologous genes with a different phylogenetic origin (see also fig. 4*b*).

transcripts of the *cas7/csd2* and *cas1* genes of the type I-C system were detected by the reverse transcription polymerase chain reaction (RT-PCR) (data not shown). Thus, the "*Ca.* Endomicrobium trichonymphae" phylotype Rs-D17 appears to have been subjected to invading foreign DNA even after taking up residence inside *T. agilis* cells.

R–M and CRISPR/Cas Systems in Other *Endomicrobium* Endosymbionts

To examine whether a genome with intact CRISPR/Cas systems and numerous intact or pseudogenized R–M systems is unique to phylotype Rs-D17 or common in *Endomicrobium* endosymbionts, we additionally reconstructed draft genomes of three endosymbiotic *Endomicrobium* phylotypes associated with different host protist species: phylotype HsTcC-EM16 from *Trichonympha* sp. HsjTcC in the gut of the termite *Hodotermopsis sjostedti*; phylotype MdDo-005 from the protist *Deltotrichonympha operculata* and phylotype MdMp-027 from the protist *Mixotricha paradoxa* in the gut of the termite *Mastotermes darwiniensis*. These three *Endomicrobium* phylotypes indeed possess numerous R–M systems including

many pseudogenes (supplementary table S4, Supplementary Material online) and also apparently intact CRISPR/Cas systems (fig. 2). Many of the R–M systems are shared by two or more of the endosymbiotic phylotypes, whereas the free-living isolates in the phylum *Elusimicrobia*, *Elusimicrobium minutum*, and *Endomicrobium proavitum*, possess only few R–M systems (supplementary table S4, Supplementary Material online) (Herlemann et al. 2009; Zheng and Brune 2015; Zheng, Dietrich, Hongoh, et al. 2016). It is therefore likely that most of the numerous R–M systems were acquired by common ancestors of the endosymbionts after divergence from the *Endomicrobium proavitum* lineage, as suggested previously (Zheng, Dietrich, Hongoh, et al. 2016). Although the lifestyles of those ancestors are unknown, they likely needed an array of defense systems for protection against phages and other invading DNA elements.

Endosymbiotic and free-living *Endomicrobium* species commonly have an intact type I-C CRISPR/Cas system (fig. 2), and each of the *cas1, cas2,* and *cas4* genes of the *Endomicrobium* species constituted a monophyletic cluster (fig. 4*a* and supplementary fig. S5*a,b*, Supplementary Material online). However, the remaining gene components, *cas3, cas5, cas7/csd2,* and
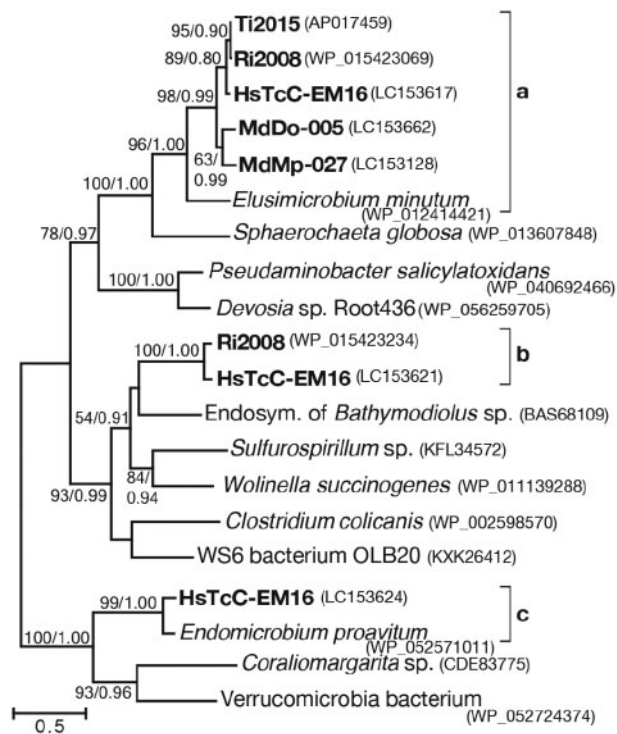
FIG. 3.—Phylogenetic positions of the Cas1 proteins in the type II-C CRISPR/Cas systems of the phylum *Elusimicrobia*. A maximum-likelihood tree was constructed using the LG + Γ amino acid substitution model. Bootstrap confidence values (left) and posterior probabilities in Bayesian statistics (right) are shown for the internal branches. Unambiguously aligned 237 sites were used. The pseudogene of *cas1* in Ti2015 was included by adjusting the codon frame. Sequences obtained in this study are shown in bold.

cas8/csd1, each formed two distinct phylogenetic clusters (fig. 4b and supplementary fig. S5c,d,e, Supplementary Material online). This indicates that a part of the *cas* gene cluster has been replaced by a phylogenetically distinct *cas* gene set in phylotypes HsTcC-EM16 and MdMp-027 (figs. 2 and 4b, and supplementary fig. S5, Supplementary Material online). Because these two *Endomicrobium* phylotypes are not monophyletic based on the 16S and 23S rRNA gene sequences (supplementary fig. S6, Supplementary Material online), the replacements probably occurred independently.

The type II-C-a system is commonly found in *Endomicrobium* species, although the system appears to have lost its function in phylotype Rs-D17 as discussed above (fig. 2), and the system is absent from *Endomicrobium proavitum*. Each of the *cas* genes of this system in the *Elusimicrobia* formed a monophyletic cluster (fig. 3, and supplementary fig. S4a,b, Supplementary Material online), and the tree topology is congruent with that of the rRNA gene-based tree (supplementary fig. S6, Supplementary Material online). The type II-C-b system in

Ri2008 was found also in phylotype HsTcC-EM16, although the CRISPR region of the latter is missing. Phylotype HsTcC-EM16 possesses another phylogenetically distinct gene set of the type II-C system, designated here as II-C-c, accompanying two separate CRISPR regions, which have identical "repeat" sequences. The type-II-C-c system was also found in *Endomicrobium proavitum* (fig. 2). These results indicate that at least the type I-C and type II-C-a systems of the endosymbiotic *Endomicrobium* phylotypes have been inherited from a common ancestor, even though the former system experienced the replacement of a part of the *cas* gene cluster.

Requirement of Defense against Foreign DNA in Intracellular Symbionts

Most reported intracellular symbionts of insects are housed in specialized host cells called bacteriocytes, allowing sequestration away from the insect body or gut lumen. The necessity of the defense systems against foreign DNA for these organisms is therefore expected to be low. Endosymbiotic *Endomicrobium* species also inhabit the cytoplasm of eukaryotic cells; however, their protist hosts are exposed to the gut luminal fluid that contains a dense community of prokaryotes and presumably phages. Because the protist hosts phagocytose wood particles, other gut luminal components such as smaller protists, prokaryotic cells, phages, and extracellular nucleic acids could simultaneously be endocytosed (Yamaoka and Nagatani 1977). It is conceivable that the endosymbionts are exposed to phages and DNA released from digested microbial cells, as discussed previously (Zheng, Dietrich, Hongoh, et al 2016), even though the frequency of exposure is likely to be much lower than for gut bacteria living outside the protist cells (i.e., in the gut lumen). Thus, it might be expected that endosymbiotic *Endomicrobium* species require defense systems against foreign DNA. On the other hand, "*Candidatus* Azobacteroides pseudotrichonymphae" (phylum *Bacteroidetes*; order *Bacteroidales*), an intracellular symbiont of the protist *Pseudotrichonympha grassii* in the gut of the termite *Coptotermes formosanus*, has neither R–M systems nor CRISPR/Cas systems (Hongoh et al. 2008b), despite its genome size and the number of CDSs being comparable to "*Ca*. Endomicrobium trichonymphae". Thus, the requirement of such defense systems may also depend on the bacterial taxonomic groups and/or other unknown factors.

Materials and Methods

Sample Collection

The termite *R. speratus* was collected at Ogose in Saitama Prefecture, Japan in 2012. This is the same site where the *R. speratus* individual used for reconstructing the Ri2008 genome was collected in 2006. *H. sjostedti* was collected in Kagoshima Prefecture, Japan, and *M. darwiniensis* was collected in Darwin, Australia. Worker termites reared with
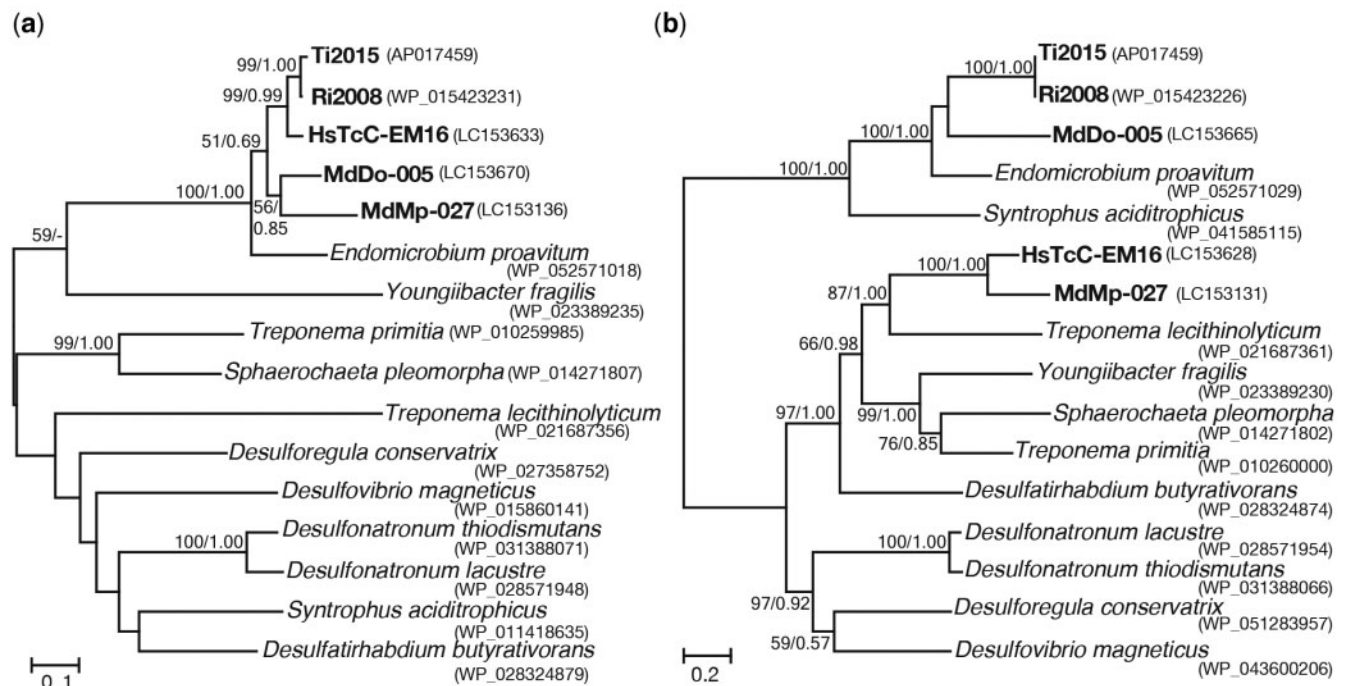
Fig. 4.—Phylogenetic positions of the Cas1 and Cas3 proteins in the type I-C CRIPSR/Cas systems of the phylum *Elusimicrobia*. Maximum-likelihood trees were constructed using the LG + Γ and LG + Γ + I amino acid substitution models, for Cas1 and Cas3, respectively. Bootstrap confidence values (left) and posterior probabilities in Bayesian statistics (right) are shown for the internal branches. Only values >50% are shown. Sequences obtained in this study are shown in bold. (*a*) Cas1 protein. 341 unambiguously aligned sites were used. (*b*) Cas3 protein. 541 unambiguously aligned sites were used. The pseudogene of *cas3* in *Endomicrobium proavitum* was included by adjusting codon frames.

cellulose powder for 3 days were subjected to the experiments. Single-cell isolation of gut protists and collection of their endosymbiotic *Endomicrobium* cells were performed as described previously (Hongoh et al. 2008a), with an addition of 0.1% Tween 20 (Sigma-Aldrich) (Sato et al. 2014). WGA was performed with the illustra GenomiPhi HY DNA Amplification Kit (GE Healthcare Life Sciences) for 8 h, as described previously (Hongoh et al. 2008b).

Although we did not analyze the phylogenetic relationship between the *T. agilis* cells that hosted the Ti2015 and Ri2008 genomovars, there has been no report that there are multiple phylotypes of *T. agilis* in *R. speratus* guts. Considering the very high genome sequence identity of the endosymbionts Ti2015 and Ri2008 genomovars and the strict host-specificity of *Endomicrobium* endosymbionts (Zheng et al. 2015), the host *T. agilis* cells were most likely highly similar.

Genome Sequencing and Bioinformatics

Libraries for paired-end and mate-pair sequencing were prepared using the TruSeq DNA PCR-free Sample Prep Kit and the Nextera Mate Pair Sample Prep Kit (Illumina), respectively. Sequencing was performed on an Illumina MiSeq platform using the MiSeq Reagent Kit v3 (600 cycles). The generated reads were quality-filtered and assembled into contigs using

SPAdes 3.1 (Bankevich et al. 2012), which were further assembled to scaffolds using SCARPA 0.241 (Donmez and Brudno 2013). Gaps in scaffolds were closed by PCR amplification and Sanger sequencing on ABI3730 Genetic Analyzers. Detailed procedures are described in the supplementary methods (Supplementary Material online). CDSs were predicted using the BLASTn algorithm on the basis of the annotation for genomovar Ri2008 (Hongoh et al. 2008a). The sequence alignment with genomovar Ri2008 was manually checked. Pseudogenes were identified as described previously (Hongoh et al. 2008a), and single nucleotide polymorphisms were counted manually. Synteny of the genome regions was analyzed using GenomeMatcher 2.0 (Ohtsubo et al. 2008). Chromosome sequence identity was calculated using Genome-to-Genome Distance Calculator 2.1 (Meier-Kolthoff et al. 2013). The *dif* sequence was predicted on the basis of the deduced amino acid sequence of the site-specific recombinase XerCD, using a prediction system in the web site "Database of bacterial replication terminus" (http://www.g-language.org/data/repter/) (Kono et al. 2012). The "repeat" and "spacer" sequences of CRISPRs were identified using CRISPRFinder (Grissa et al. 2007). The *cas* genes and R–M system-related genes in the draft genomes of the *Endomicrobium* phylotypes were identified by BLASTx

searches against the NCBI non-redundant protein sequence database with default settings.

Phylogenetic Analysis

Nucleotide or deduced amino acid sequences were aligned using the ClustalW algorithm implemented in MEGA6.0 (Tamura et al. 2013). The best-fit substitution models were selected using Model test and used for constructing maximum-likelihood trees in MEGA6.0. For bootstrap analysis, 1,000 resamplings were performed. Bayesian trees were constructed using MrBayes 3.2.5 (Ronquist et al. 2012). Four Markov chains (three heated and one cold) were run simultaneously for 1,000,000 generations.

RNA Extraction and RT-PCR

Gut contents of 10 *R. speratus* workers were suspended in the Trager U solution (Trager 1934) in a 1.5 ml tube. The suspension was centrifuged with a low speed for collection of large protists including *T. agilis*, and the precipitate was resuspended in sterilized double-distilled water. Total RNA was extracted using the PowerViral® Environmental RNA/DNA Isolation Kit (MO BIO Laboratories). The TURBO DNA-free™ Kit (Ambion) was used to remove DNA. Reverse transcription was performed with random hexamers using the SuperScript™ III First-Strand Synthesis System for RT-PCR (Invitrogen). Primers specific to each of *cas1* and *csd2* genes were designed (supplementary table S5, Supplementary Material online). RT-PCR was performed using *Ex-Taq* polymerase (TaKaRa) on a C1000 Thermal Cycler (Bio-Rad). The following program was used: 3-min initial denaturation at 95 °C, 30 cycles of denaturation (15 s at 95 °C), annealing (30 s at 60 °C), extension (30 s at 72 °C), and a final 3-min extension at 72 °C. No specific amplification was detected from the RNA samples without reverse transcription. The amplicons were directly sequenced with the same primers on an ABI3730 Genetic Analyzer.

Supplementary Material

Supplementary figures S1–S7 and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

Literature Cited

Abdul Rahman N, et al. 2015. A molecular survey of Australian and north American termite genera indicates that vertical inheritance is the primary force shaping termite gut microbiomes. Microbiome 3:5.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 5:455–477.

Boucias DG, et al. 2013. The hindgut lumen prokaryotic microbiota of the termite *Reticulitermes flavipes* and its responses to dietary lignocellulose composition. Mol. Ecol. 22:1836–1853.

Dietrich C, Köhler T, Brune A. 2014. The cockroach origin of the termite gut microbiota: patterns in bacterial community structure reflect major evolutionary events. Appl. Environ. Microbiol. 80:2261–2269.

Donmez N, Brudno M. 2013. SCARPA: scaffolding reads with practical algorithms. Bioinformatics 29:428–434.

Eisen AJ, Heidelberg JF, White O, Salzberg SL. 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. 1:research0011–research0011.9.

Grissa I, Vergnaud G, Pourcel C. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. 35:W52–W57.

Herlemann DP, et al. 2009. Genomic analysis of "*Elusimicrobium minutum*," the first cultivated representative of the phylum "*Elusimicrobia*" (formerly Termite Group 1). Appl. Environ. Microbiol. 75:2841–2849.

Hongoh Y, Ohkuma M, Kudo T. 2003. Molecular analysis of bacterial microbiota in the gut of the termite *Reticulitermes speratus* (Isoptera; Rhinotermitidae). FEMS Microbiol. Ecol. 44:231–242.

Hongoh Y, et al. 2005. Intra- and interspecific comparisons of bacterial diversity and community structure support coevolution of gut microbiota and termite host. Appl. Environ. Microbiol. 71:6590–6599.

Hongoh Y, et al. 2008a. Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. Proc. Natl. Acad. Sci. USA. 105:5555–5560.

Hongoh Y, et al. 2008b. Genome of an endosymbiont coupling N$_2$ fixation to cellulolysis within protist cells in termite gut. Science 322:1108–1109.

Hongoh Y. 2011. Toward the functional analysis of uncultivable, symbiotic microorganisms in the termite gut. Cell Mol. Life Sci. 68:1311–1325.

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.

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.

Karginov FV, Hannon GJ. 2010. The CRISPR system: small RNA-guided defense in bacteria and archaea. Mol. Cell 37:7–19.

Kono N, Arakawa K, Tomita M. 2012. Validation of bacterial replication termination models using simulation of genomic mutations. PLoS One 7:e34526.

Makarova KS, et al. 2011. Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. 9:467–477.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10:13–26.

Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14:60.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. Annu. Rev. Genet. 42:165–190.

Nakabachi A, et al. 2013. Defensive bacteriome symbiont with a drastically reduced genome. Curr. Biol. 23:1478–1484.

Ohtsubo Y, Ikeda-Ohtsubo W, Nagata Y, Tsuda M. 2008. GenomeMatcher: a graphical user interface for DNA sequence comparison. BMC Bioinformatics 9:376.

Ohkuma M, Kudo T. 1996. Phylogenetic diversity of the intestinal bacterial community in the termite *Reticulitermes speratus*. Appl. Environ. Microbiol. 62:461–468.

Ohkuma M, et al. 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.

Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Sato T, et al. 2014. Intranuclear verrucomicrobial symbionts and evidence of lateral gene transfer to the host protist in the termite gut. ISME J. 8:1008–1019.

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.

Tai V, et al. 2015. The role of host phylogeny varies in shaping microbial diversity in the hindguts of lower termites. Appl. Environ. Microbiol. 81:1059–1070.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30:2725–2729.

Tillier RME, Collins AR. 2000. Genome rearrangement by replication-directed translocation. Nat. Genet. 26:195–197.

Trager W. 1934. The cultivation of a cellulose-digesting flagellate, *Trichomonas termopsidis*, and of certain other termite protozoa. Biol. Bull. 66:182–190.

Yamaoka I, Nagatani Y. 1977. Cellulose digestion system in the termite, *Reticulitermes speratus* (Kolbe). II. Ultra-structural changes related to the ingestion and digestion of cellulose by the flagellate, *Trichonympha agilis*. Zool. Mag. 86:34–42.

Yamin MA. 1979. Flagellates of the orders Trichomonadida Kirby, Oxymonadida Grassé, and Hypermastigida Grassi and Foà reported from lower termites (Isoptera families Mastotermitidae, Kalotermitidae, Hodotermitidae, Termopsidae, Rhinotermitidae, and Serritermitidae) and from the wood-feeding roach *Cryptocercus* (Dictyoptera: Cryptocercidae). Sociobiology 4:1–120.

Yamin MA. 1981. Cellulose metabolism by the flagellate *Trichonympha* from a termite is independent of endosymbiotic bacteria. Science 211:58–59.

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.

Zheng H, Brune A. 2015. Complete genome sequence of *Endomicrobium proavitum*, a free-living relative of the intracellular symbionts of termite gut flagellates (Phylum *Elusimicrobia*). Genome Announc. 3:e00679–e00615.

Zheng H, Dietrich C, Thompson CL, Meuser K, Brune A. 2015. Population structure of Endomicrobia in single host cells of termite gut flagellates (*Trichonympha* spp.). Microbes Environ. 30:92–98.

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.

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.

Associate editor: Rachel Whitaker