Thermosipho spp. Immune System Differences Affect Variation in Genome Size and Geographical Distributions

Thomas H.A. Haverkamp[1,2,*], Claire Geslin[3,4,5], Julien Lossouarn[3,4,5], Olga A. Podosokorskaya[6], Ilya Kublanov[6,7], and Camilla L. Nesbø[1,8]

[1]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Norway

[2]Norwegian Veterinary Institute, Oslo, Norway

[3]Laboratoire de Microbiologie des Environnements Extrêmes (LMEE), Institut Universitaire Européen de la Mer (IUEM) – UMR 6197, Université de Bretagne Occidentale (UBO), Plouzané, France

[4]CNRS, IUEM – UMR 6197, Laboratoire de Microbiologie des Environnements Extrêmes (LMEE), Plouzané, France

[5]Ifremer, UMR 6197, Laboratoire de Microbiologie des Environnements Extrêmes (LMEE), Technopôle Pointe du diable, Plouzané, France

[6]Winogradsky Institute of Microbiology, Research Center of Biotechnology, Russian Academy of Sciences, Moscow, Russia

[7]Immanuel Kant Baltic Federal University, Kaliningrad, Russia

[8]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

*Corresponding author: E-mail: thhaverk@gmail.com.

Abstract

Thermosipho species inhabit thermal environments such as marine hydrothermal vents, petroleum reservoirs, and terrestrial hot springs. A 16S rRNA phylogeny of available *Thermosipho* spp. sequences suggested habitat specialists adapted to living in hydrothermal vents only, and habitat generalists inhabiting oil reservoirs, hydrothermal vents, and hotsprings. Comparative genomics of 15 *Thermosipho* genomes separated them into three distinct species with different habitat distributions: The widely distributed *T. africanus* and the more specialized, *T. melanesiensis* and *T. affectus*. Moreover, the species can be differentiated on the basis of genome size (GS), genome content, and immune system composition. For instance, the *T. africanus* genomes are largest and contained the most carbohydrate metabolism genes, which could explain why these isolates were obtained from ecologically more divergent habitats. Nonetheless, all the *Thermosipho* genomes, like other Thermotogae genomes, show evidence of genome streamlining. GS differences between the species could further be correlated to differences in defense capacities against foreign DNA, which influence recombination via HGT. The smallest genomes are found in *T. affectus* that contain both CRISPR-cas Type I and III systems, but no RM system genes. We suggest that this has caused these genomes to be almost devoid of mobile elements, contrasting the two other species genomes that contain a higher abundance of mobile elements combined with different immune system configurations. Taken together, the comparative genomic analyses of *Thermosipho* spp. revealed genetic variation allowing habitat differentiation within the genus as well as differentiation with respect to invading mobile DNA.

Key words: extremophiles, hydrothermal vents, mobile DNA, speciation, Thermotogae, Vitamine B$_{12}$.

Introduction

Bacteria of the genus *Thermosipho* are found in high temperature environments such as deep-sea hydrothermal vents and subsurface oil reservoirs. The genus belongs to the phylum Thermotogae and there are currently eight described *Thermosipho* species: *T. africanus*, *T. activus*, *T. affectus*, *T. atlanticus*, *T. geolei*, *T. globiformans*, *T. japonicus*, and *T. melanesiensis*. These are thermophilic, organotrophic, sulfur-reducing anaerobes, fermenting various sugars and peptides to H$_2$, CO$_2$, and acetate (Podosokorskaya et al. 2014).

Comparative genomic analyses of Thermotogae, including *Thermosipho*, have revealed complex evolutionary histories with extensive horizontal gene transfer (HGT), particularly

GBE

Table 1

Overview of *Thermosipho* Genomes Used in This Study

T. genome	NCBI Accession Number	Nr of Contigs[e]	Genome Size (bp)	GC Content (%)	Intergenic Region Size (% GS)	Sequence Technology	Isolation Source	References
T. melanesiensis BI429[a,b]	NC_009616	1	1.915.238	31.4	7.31	454/Sanger	Gills of mussel, Deep-sea H. vent, Pacific Ocean, Lau Basin	Antoine et al. (1997), Zhaxybayeva et al. (2009)
T. melanesiensis 430[b]	JYCX00000000	4	1.906.053	31.5	5.98	MiSeq	Gills of gastropod Deep-sea H. vent, pacific ocean, Lau Basin Deep-sea H. vent, pacific ocean, Lau Basin	
T. melanesiensis 431[b]	CP007389	1	1.915.344	31.4	6.06	Pacbio /MiSeq	Sample of active chimney	This study
T. melanesiensis 432[b]	JYCY00000000	21	1.890.294	32.6	6.70	MiSeq	Sample of active chimney	This study
T. melanesiensis 433[b]	JYCZ00000000	21	1.890.575	32.6	6.51	MiSeq	Gills of a mussel	
T. melanesiensis 434[b]	JYDA00000000	22	1.892.506	32.7	6.31	MiSeq		
T. melanesiensis 487[b]	JYDB00000000	21	1.890.479	32.6	6.39	MiSeq	Sample of active chimney	
T. affectus 1063[c]	CP007223	1	1.766.633	31.5	7.30	Pacbio /MiSeq	Deep sea H. vent, Atlantic ocean	Wery et al. (2002)
T. affectus 1070[c]	CP007121	1	1.781.851	31.5	6.60	Pacbio /MiSeq	Menez-Gwen hydrothermal field	This study
T. affectus 1074[c]	LBEX00000000	20	1.788.307	33.3	7.02	MiSeq	Menez-Gwen hydrothermal field	This study
T. affectus ik275mar	LBFC00000000	27	1.771.018	32.6	6.80	MiSeq	Rainbow H.vent field	Podosokorskaya et al. (2011)
T. affectus sp. 1223[d]	LBEY00000000	39	1.767.555	32.1	6.88	MiSeq	Rainbow H.vent field	This study
T. africanus ob7	NKRG00000000	23	1.933.435	30.6	8.25	Ion Torrent	Tadjoura gulf Hydrothermal springs, Djibouti	Huber et al. (1989)
T. africanus H17ap60334	AJIP0100000	49	2.083.551	29.9	7.64	454/Sanger	Produced water from Hibernia oil platform, North-west Atlantic Ocean.	This study
T. africanus TCF52B[a]	NC_011653	1	2.016.657	30.8	8.98	Sanger	Produced water from Troll C oil platform, North Sea	Dahle et al. (2008), Nesbo et al. (2009)

[a]Not sequenced during this study.

[b]Strains 429–487 isolated during Oceanographic cruise Biolau: Pacific Ocean, Lau Basin, Deep-sea Hydrothermal vent.

[c]Strains 1063–1074 isolated during Oceanographic cruise Marvel: Atlantic ocean, Menez Gwen, Deep-sea Hydrothermal vent.

[d]Isolated during Oceanographic cruise Atos, Atlantic ocean, Rainbow, Deep sea Hydrothermal vent.

[e]Genomes with 1 contig were closed.

Thermosipho spp. Genome Size Variation and Geographical Distributions

GBE

involving members of the bacterial phylum Firmicutes and the Archaea domain (Zhaxybayeva et al. 2009). HGT allows organisms to acquire novel metabolic capacities and adaptations that may lead to species differentiation and innovation (Koonin 2015). In addition, HGT and homologous recombination, may be important in the long-term maintenance and repair of genomic information, by replacing inactivated genes (Takeuchi et al. 2014).

Several observations from *Thermosipho* genomes imply that they are influenced by recent HGT. First, a remnant prophage element was detected in the *T. melanesiensis* BI429 genome (Zhaxybayeva et al. 2009). Second, in *T. africanus* TCF52B seventy-eight open-reading frames (ORFs) are annotated as transposases or integrases, and their high intragenomic similarity suggests recent Insertion Sequence (IS) element activity (Nesbø et al. 2009). Both IS elements and prophages are mobile DNA elements and might thus play a role in HGT (Darmon and Leach 2014). And finally, *Thermosipho* genomes carry horizontally acquired genes encoding the vitamin B_{12} synthesis pathway which is absent from other Thermotogae (Swithers et al. 2012).

Intriguingly, thermophile genomes tend to have a higher proportion of HGT acquired genes than mesophile genomes (Anderson et al. 2014; Fuchsman et al. 2017). This might be due to either adaptation to hot environments, DNA repair or dissimilarities of mobile DNA activity rates between hot and mesophilic environments (van Wolferen et al. 2013; Fuchsman et al. 2017). Counteracting this trend is the observation that prokaryotic genome size (GS) decreases with an increase of the optimal growth temperature (OGT). This causes thermophiles to have smaller genomes than mesophiles (Zhaxybayeva et al. 2012; Sabath et al. 2013). Sabath et al. (2013) proposed that temperature-induced-genome streamlining was due to a selection for decreased cell and GS at elevated temperatures. They also showed that many characteristics of streamlined genomes are identified in thermophiles, but it is unclear why thermophiles then have a higher abundance of foreign genes.

Therefore, increasing OGT might increase the selective pressure for mechanisms controlling HGT to prevent unwanted genome expansion. Known vectors of "foreign" DNA, such as phages can have deleterious effects on their host, and most, if not all, prokaryotes have acquired mechanisms that limit their activity. Among these are mechanisms that regulate cellular mobile DNA activity: Restriction–modification (RM); abortive infection (ABi); toxin–antitoxin (TA) systems and the clustered regulatory interspaced short palindromic repeats (CRISPR-cas) system (van Houte et al. 2016). CRISPR-cas systems are notably enriched in prokaryotic thermophiles compared with psychrophiles and mesophiles, which implies that HGT control mechanisms are indeed important for thermophile genome maintenance (Koonin et al. 2017).

The understanding of the interaction between GS, HGT and immune system genes is limited for the phylum Thermotogae and the genus *Thermosipho* in particular. For instance, the RM system is well studied and almost universally found in prokaryotes (Roberts et al. 2010), but only poorly characterized among described Thermotogae species (Xu et al. 2011). In contrast, CRISPR-cas systems have been identified in all Thermotogae genomes characterized to date (Zhaxybayeva et al. 2009). Interestingly, CRISPR-cas genes can also be part of mobile DNA itself (Seed et al. 2013), and many show evidence of HGT when comparing for instance *T. africanus* to *Thermotoga maritima* (Nesbø et al. 2009).

Here, we present a comparative analysis of fifteen *Thermosipho* genomes, thirteen of which were generated in this study. The isolates included here were obtained from deep-sea hydrothermal vents and produced fluids from oil reservoirs. The genomes fall into three well-defined lineages (or species) with isolates from the same sample sites showing very high similarities to each other. We show that the three species differ with regard to GS, metabolic repertoire, presence of mobile elements and their immune system gene content, and discuss how the repertoire of immune system genes may influence the observed genomic differences between the three species.

Materials and Methods

DNA Extraction and Genome Sequencing

Thirteen *Thermosipho* strains were cultured in a modified Ravot medium as previously described in Lossouarn et al. (2015) and used for total DNA extraction as described in Charbonnier et al. (1992) and Geslin et al. (2003). All strains, except *T. africanus* Ob7 were submitted for genome sequencing at the Norwegian Sequencing Centre (NSC, Oslo, Norway). A 300 bp insert paired end library was prepared using the genomic DNA sample prep kit (Illumina, San Diego, CA, USA) for each DNA sample. The Library was spiked with phiX DNA. Subsequently, all strains except Ob7 were sequenced on an Illumina MiSeq generating a data set of 250 bp paired end reads.

Strain Ob7 was sequenced at University of Alberta, Canada, using an Iontorrent (Thermo Fisher Scientific, Waltham, MA, USA) approach. DNA was enzymatically sheared using the Ion Shear Plus kit (Life Technologies, Carlsbad, CA, USA) then cloned using the Ion Plus Fragment Library kit (Life Technologies) following the manufacturer's instructions. The library was then sequenced on an Ion Torrent PGM using a 316D chip and 500 flows.

Four isolates (*Thermosipho* sp. 430, 431, 1063, 1070) were selected for long read sequencing with the goal to complete the genomes and assess methylation status of the DNA. DNA's were prepared following the Pacific Biosciences instructions and sequenced at the NSC on a Pacbio RSII (Pacific Biosciences, Menlo Park, CA, USA).

Quality Control and Genome Assembly

Pacbio sequencing of the strains 431, 1063, and 1070 produced libraries suitable for genome assembly without the addition of Illumina paired-end reads. For all strains Pacbio read filtering and error-correction were performed using version 2.1 of the Pacbio smrtportal (http://www.pacb.com/devnet/; last accessed October 3, 2018). Assembly and base modification detection were performed using the RS_HGAP_Assembly.2 and RS_Modification_and_Motif_analysis protocols.

Overlapping regions between contigs were manually detected using a combination of read mapping and subsequent visualization with IGV (v. 2.3.23) (Thorvaldsdottir et al. 2013). This allowed us to inspect regions with low quality mappings (due to the presence of duplicated regions), and similar gene annotations on both sides of a contig gap. Preliminary annotations were prepared using Glimmer at the RAST server and inspected in the CLC Bio main workbench (Overbeek et al. 2014). Bam files were generated by mapping MiSeq PE reads using bwa version (v. 0.7.5a) to the annotated contigs (Li and Durbin 2009) and by converting sam files with Samtools (v. 0.1.19). Bam files were imported into the IGV viewer together with the annotated contigs. This approach resulted in closed chromosomal sequences.

For strain 430, we generated a combined assembly using high quality illumina MiSeq reads and a smaller set of Pacbio reads (15594 subreads) in an assembly using Spades (v. 3.5.0.) (Bankevich et al. 2012) with kmer size set to 127. Low coverage contigs (<2.0) were discarded. The remaining contigs were checked in IGV for miss-assemblies using mapped MiSeq and Pacbio reads and if detected contigs were discarded. Finally, the contigs were checked for unambiguously overlapping ends as described above and if detected the contigs were combined.

The _T. africanus_ Ob7 genome was assembled using NEWBLER v. 2.6 (Margulies et al. 2005). The remaining strains, were assembled using CLC assembler v.7, Spades (v. 3.5.0.) and Velvet (v 1.2.10) (Zerbino and Birney 2008) (table 1). Genome assemblies were compared using REAPR (v1.0.16) and the most optimal assembly for each strain was selected for genome annotation with the PGAP pipeline from NCBI (Hunt et al. 2013).

16S rRNA Phylogeny of Thermosipho Species

Full-length 16S rRNA sequences were manually extracted from the annotated _Thermosipho_ genome assemblies generated in this study as well as publically available genomes. Next we searched the SILVA database version SSU r122 for _Thermosipho_ spp. 16S rRNA sequences with the following setting: Sequence length >1,300 bp, alignment quality >80, pintail quality >80 (Quast et al. 2012). Matching sequences were downloaded and added to extracted 16S rRNA sequences from the de novo sequenced genomes.

Multiple identical sequences from the same strain with different accession numbers (e.g., the four 16S rRNA sequences of _T. melanesiensis_ BI429, five 16S rRNA sequences _T. africanus_ TCF52B) are represented by a single sequence in the data set.

The NCBI nr database was then queried using BlastN (blast+ v 2.2.26, Camacho et al. 2009) with the 16S rRNA gene data set described above, to identify _Thermosipho_ sp. like 16S rRNA sequences missed with the above methods. Additional full-length sequences were downloaded from Genbank and combined with the de novo/SILVA 16S rRNA sequences.

The final 16S rRNA gene alignment consisted of _Thermosipho_ sequences combined with outgroup sequences from other Thermotogales species. The resulting alignment was used to build phylogenies using Maximum Likelihood as implemented in MEGA 6.0 (Tamura et al. 2013) and the General Time Reversible (GTR) model (G + I, four categories) as suggested by JModelTest v. 2.1.7 (Darriba et al. 2012).

Genome Properties and Selection Analysis

The %GC of the three positions in each codon of all genes was using the software codonW version 1.4.4 (http://codonw.sourceforge.net/; last accessed October 3, 2018). The N/C-ARSC values were calculated with the python scripts from the publication (Mende et al. 2017) after modifying them for usage on our genomes.

Pangenome Analysis

Blast Ring genome plots were generated using BRIG version 0.95 (Alikhan et al. 2011) running blast+ version 2.2.28 (Camacho et al. 2009) with the following set up. The nucleotide sequence of one of the fifteen _Thermosipho_ genomes was used to create the blast-database with default settings. Nucleotide sequences of all coding genes were extracted from each genome (NCBI Genbank annotation) using CLC Main workbench Version 6.8.3 and were used for BlastN analysis in BRIG with the following settings: max_target_seqs: 1; max e-value cut-off: 1.0^{-4}. Alignments with a minimum of 70% similarity were visualized with BRIG. The same procedure was used for all genomes.

Complete chromosome sequences or contigs were uploaded for all fifteen genomes to the Panseq 2.0 server (https://lfz.corefacility.ca/panseq/; last accessed October 3, 2018) for pangenome analysis (Laing et al. 2010). Sequence similarity cut-off (SC) was set at 70% to identify core-genome segments and SNP's. We used the standard settings with Panseq except: Percent sequence identity cut-off set to 70%, core genome threshold: 15, BLAST word size: 11. The final alignment of the single nucleotide polymorphisms (SNPs) was loaded into splitstree and visualized as an unrooted phylogeny using the neighbor network algorithm (Huson and Bryant 2006).

Thermosipho spp. Genome Size Variation and Geographical Distributions

GBE

Ten genome sequences (none-closed genomes of the *T. melanesiensis* cluster were excluded, due to high sequence similarity) were aligned with progressiveMauve (Mauve (v. 2.3.1) Darling et al. 2010) using *T. africanus* H17ap60344 set as the reference and automatically calculated seed weights and minimum Locally Colinear Blocks (LCB) scores. Gaps were removed and the edited LCBs were concatenated in Geneious 8 (www.geneious.com). Recombination analysis of the concatenated alignment was done in LikeWind (Archibald and Roger 2002) using the maximum likelihood tree calculated in PAUP* version 4.0b10 (Swofford 2002) under a GTR+Γ+I model as the reference tree.

Pairwise Average Nucleotide Identity Blast scores (ANIb) were calculated using the JspeciesWS webtool (http://jspecies.ribohost.com/jspeciesws/; last accessed October 3, 2018) (Richter and Rosselló 2009; Goris et al. 2007). The results were visualized with R (version 3.2.1) using the heatmap.2 function (Package gplots) and the *Thermosipho* genomes were clustered using Jaccard distances based on pairwise ANIb values (Package Vegan. version 2.3-0).

The IMG/ER Pairwise ANI calculation was used to determine the number of shared genes between each genome separately (Markowitz et al. 2009). The method uses pairwise bidirectional best nSimScan hits where similar genes share a minimum of 70% sequence identity with 70% coverage of the smaller gene. For each genome we calculated the fraction of shared genes with all other genomes. Unique genes per genome were determined using the IMG Phylogenetic profiler tool for single genes, where each genome was analyzed for the presence of genes without any homologs in all other 14 genomes. The settings were: Minimum e-value of 1.0^{-5}; min percent identity: 70%; pseudogenes were excluded; the algorithm was present/absent homologs. The same tool was used to identify the presence of homologs shared with all genomes, and with only the genomes from the same cluster.

Functional Comparison of Genome

The fifteen genomes were compared using the Clusters of Orthologous Genes (COGs) annotations. A COG reference database (version 10) was downloaded from the STRING database (http://string-db.org/; last accessed October 3, 2018). For each genome all protein sequences were aligned using BlastP, with the settings: 1 target sequence; maximum e-value: $1.0e^{-20}$; database size 1.0^7; tabular output. Only hits to COG database sequences were retained when the alignment was ≥70% of the length of the longest protein sequence and were used to build a protein COG classification table. The COG IDs were summarized by classifying them to any of the available COG categories (ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/COG0303/cogs.csv; last accessed October 3, 2018). For the COG analysis of the species-specific genes, we used the results from the IMG phylogenetic profiler tool to identify cluster specific genes. The COG classification table was screened for cluster specific genes and summarized into COG categories. Next, we searched for paralogous genes by examining how many orthologous genes, classified as COGs, were assigned per COG. This identified COGs only present once or multiple times in the genomes, and those results were summarized.

In order to statistically compare species differences of COG categories we normalized counts by total COG gene annotations, giving relative abundances per genome. R (version 3.3.0) was used to identify COG categories that were significantly different between the three species using the nonparametric Kruskal–Wallis test (P-value ≤ 0.01). The R-package ggplot2 (version 2.1.0) was used to generate the comparisons graphically.

HGT Detection

Genes putatively acquired by HGT were identified using HGTector (Zhu et al. 2014) with BLASTp (blast+ v 2.2.26). The databaser.py script was used on December 6, 2015 to download per species one representative proteome of all microorganisms from the NCBI refseq database. We compared the predicted *Thermosipho* spp. protein sequences to the reference database using the following BLAST cut-offs: E-value: 1.0^{-5}, Percentage identity: 70%, percentage coverage: 50%, and a maximum of 100 hits were returned. To determine which genes were putatively acquired by any of the strains we set the HGTector self group to the genus *Thermosipho* (NCBI taxonomy ID: 2420) and the close group to either the family Fervidobacteriaceae (NCBI taxonomy ID: 1643950), or the order Thermotogales (NCBI taxonomy ID: 2419). HGTector then analyzes the BLAST output from each protein for hits matching taxa belonging to either the self or close groups, or more distantly related taxa, which is used to determine which genes have likely been acquired from taxa more distantly related than the close group. When the close group was set to Fervidobacteriaceae, we identified putative HGT genes specific for *Thermosipo* sp. but not found in Fervidobacteriaceae species. This setting however, did not indicate putative HGT genes derived from other families within the Thermotogales order or beyond.

Defense Genes

Using the defense genes list created by Makarova et al. (2011), we screened the IMG genome COGs annotations for the presence of any COGs involved in any of the mobile DNA defences (Makarova et al. 2011). We summarized the identified COGs and distinguished between CRISPR-cas associated or RM system genes. CRISPR-cas genes were manually inspected to determine completeness/disruption of the genes. In addition, did we manually inspect IMG gene annotations for the presence of CRISPR-cas genes not identified with the Makarova et al. (2011) database. This was used to identify the CRISPR-cas gene clusters and determine the CRISPR type encoded by the genes in a cluster.

Crispr-Spacer Analysis

All genomes were uploaded to CRISPR finder (http://crispr.i2bc.paris-saclay.fr/; last accessed October 3, 2018) to detect CRISPR-arrays (Grissa et al. 2007). CRISPR spacer sequences from each genome were compared using BlastN against all *Thermosipho* sp. spacer sequences using the following settings: e-value cut-off: 1.0^{-5}, database size: 1.0^{7}, dust: no. The tabular BLAST results were visualized with R-statistics using the Markov Cluster Algorithm (MCL) (v1.0) and Igraph (v1.0.1) packages. Igraph was used for matrix construction. MCL was run using the matrix with the inflation set to 1.4 and max iterations set to 100 (Enright et al. 2002).

In order to search for matching sequences within the genome but outside the CRISPR arrays, for example, target genes, we masked all CRISPR arrays using maskfeat (EMBOSS v. 6.5.7, Rice et al. 2000). Next, we ran BlastN (v2.2.26+) using the spacers of each genome against the own genome using the following settings: e-value cut-off: 1.0^{-5}, database size: 1.0^{7}, dust: no. CRISPR array spacers were also compared against the NCBI nucleotide database to find other species with similar sequences. BlastN was run with the settings: e-value cut-off: 1.0^{-5}, dust: no. Each genome was screened for the presence of prophages (Supplementary Material for details).

Vitamine B$_{12}$ Pathway Analysis

The genes involved in the Vitamine B$_{12}$ metabolism are found in four different gene clusters (BtuFCD, Corriniod, Cobalamin, and SucCoA) in *Thermosipho* and can be regulated by B$_{12}$ riboswitches (Swithers et al. 2012). All 15 genomes were screened for the presence of Cobalamin specific riboswitches using Riboswitch scanner (Mukherjee and Sengupta 2016). This information was used to confirm the presence of the four gene clusters in each genome. Next, we extracted the protein sequences from the *T. melanesiensis* BI429 genome involved in B$_{12}$ metabolism (Swithers et al. 2012) and used them to identify homologous genes in all *Thermosipho* genomes using tBLASTn with a maximum e-value 1.0^{-20}.

Data Deposition

All genomes were deposited in the Genbank database and their accession numbers are found in table 1. In addition, all genomes were deposited in the IMG databases and are linked to the NCBI accession numbers. The 16S rRNA alignment is available upon request from the corresponding author.

Results and Discussion

Comparative Genomic Analysis of Thermosipho Isolates

We screened the NCBI-nonredundant (nr)- and IMG 16S rRNA metagenome databases using BlastN and *Thermosipho* spp. 16S rRNA genes to assess how our genomes span the available environmental diversity of *Thermosipho*. This indicated that most cultivated lineages or species, for example, *T. melanesiensis*, *T. affectus*, and *T. africanus*, are well covered by our analysis (fig. 1, supplementary fig. S1, Supplementary Material online). Nonetheless, there are several lineages, including *T. geolei*, *T. ferriphilus*, and *T. activus*, for which we do not have genomic data. An IMG database search identified three sequences closely related to *T. africanus* from oil reservoir and hydrothermal vent metagenomes (data not shown). Two recent studies of subsurface crustal fluids also contained *T. africanus* sequences (Nakagawa et al. 2006; Smith et al. 2017), indicating that *T. africanus* is widespread in the subsurface. Moreover, the absence of *T. melanesiensis* and *T. affectus* sequences from other environments than hydrothermal vents, suggests a more specialized lifestyle with limited geographic and ecological ranges.

The current study added 13 new *Thermosipho* genomes, with three closed and ten draft genomes, to the two existing ones (table 1). The overall genomic similarity and phylogenetic relationship of the isolates was calculated by identifying core-genome SNPs (Panseq) (Laing et al. 2010), ANI (Average Nucleotide Identity) (Richter and Rosselló 2009) and the fraction of shared genes between genomes (fig. 2). These analyses revealed the presence of three distinct lineages. The diversity within each lineage differed and correlated with the geographic range of sampling sites. For example, among *T. melanesiensis* isolates, which were all isolated from the same hydrothermal vent system (table 1), we found extremely low diversity (intracluster ANI values ≥ 99.98%), suggesting they are closely related clones. Thus, we kept only the two closed genomes, *T. melanesiensis* BI429 and BI431, in the further analysis. The *T. affectus* genomes, originating from two geographically close sites, show intermediate variation levels. The largest intraspecies variation is found among *T. africanus* genomes, which were sampled from three geographically and ecologically separated populations (table 1). Altogether, these results are consistent with the 16S rRNA findings and suggest that isolation by distance is important for *Thermosipho* spp. differentiation, as observed for other thermophiles (Mino et al. 2017).

Detection of genomic recombination between prokaryotic isolates can be used to assess species boundaries (Nesbø et al. 2015). LikeWind analysis (Archibald and Roger 2002) detected no recombination events between the lineages suggesting that they indeed correspond to three distinct species (supplementary fig. S2, Supplementary Material online). These results were further supported by pangenome analysis (fig. 2C). Each lineage contained species-specific genes that were at best only very distantly related to genes in the other lineages. These genes mark the *Thermosipho* species boundaries and possibly contribute to niche differentiation between them (Moldovan and Gelfand 2018).
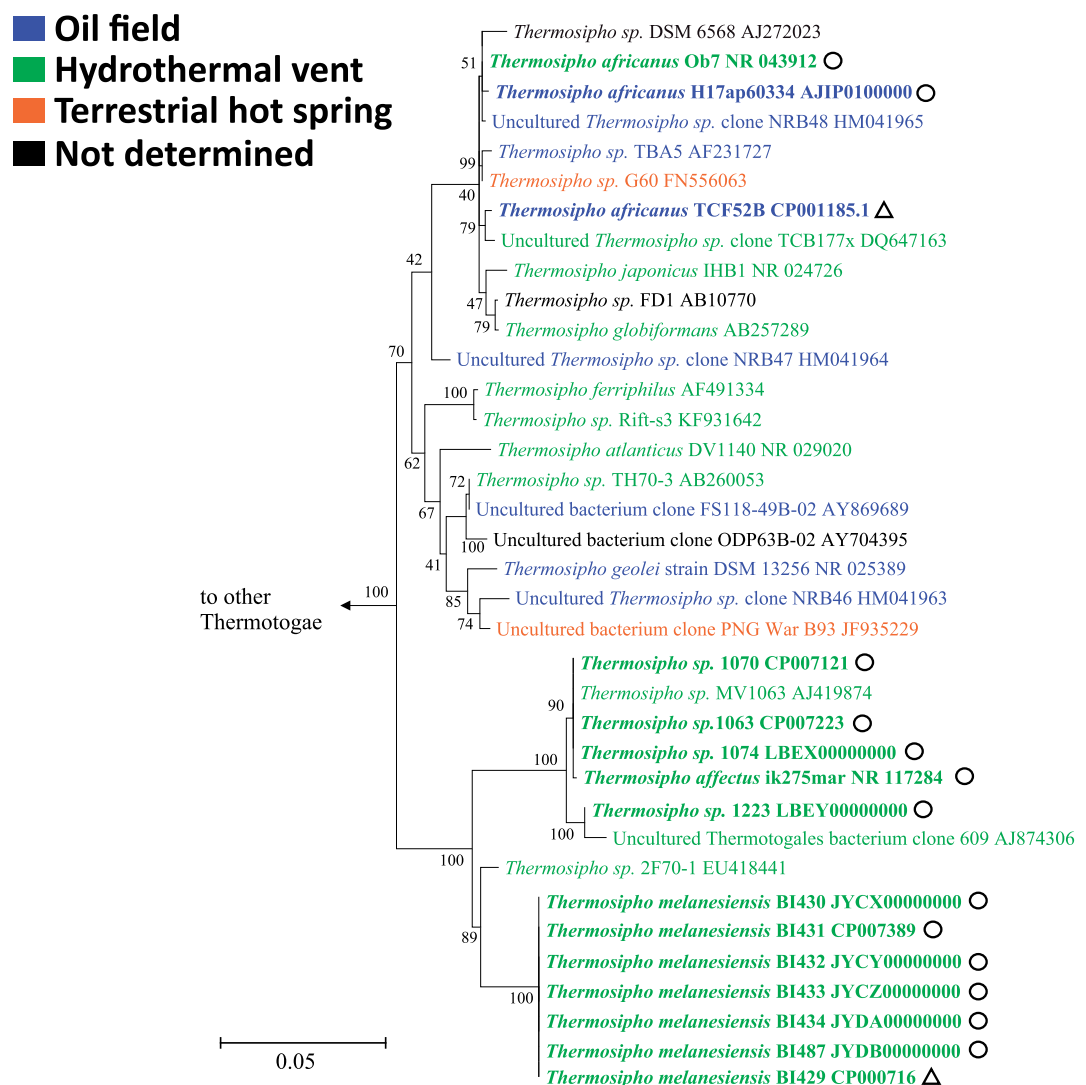
Fig. 1.—Maximum Likelihood phylogeny of *Thermosipho* 16S rRNA sequences. The tree was constructed in MEGA6 (Tamura et al. 2013) using the General Time Reversible (GTR) model ($\Gamma + I$, four categories, Nei and Kumar 2000). The tree with the highest log likelihood is shown. The percentage of trees in 100 bootstrap replicates in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site, the bar indicates 0.05 substitutions per site. The analysis involved 61 nucleotide sequences. Only alignment positions with fewer than 5% gaps, missing data, and ambiguous bases were included, resulting in a total of 1,245 positions in the final data set. The phylogeny was rooted using 16S rRNA gene sequences of representative sequences from other Thermotogae genera (*Defluviitoga*, *Fervidobacterium*, *Geotoga*, *Kosmotoga*, *Marinitoga*, *Mesoaciditoga*, *Mesotoga*, *Oceanotoga*, *Petrotoga*, *Pseudothermotoga*, and *Thermotoga*). All non-*Thermosipho* Thermotogales formed a distinct clade and were collapsed (supplementary fig. S1, Supplementary Material online for the full phylogeny). *Thermosipho* sequences are colored based on the environment of isolation. Sequences with bold fonts are whole genome sequences. Triangles behind the sequence ID indicate genome not from this study. Circles behind sequence ID indicate genome from this study.

The genomic properties of each strain follow these species boundaries (table 1; fig. 3A and B). *Thermosipho* genomes have a GC-content (%GC) ranging from 29.9% (*T. africanus* H17ap60334) to 32.7% (*T. melanesiensis* 434) with the largest genomes having lowest %GC. GS ranges from 1.77 (*T. affectus* BI1063) to 2.08 Mbp (*T. africanus* H17ap60334), with *T. melanesiensis* strains showing intermediate values (fig. 3A). The *T. affectus* GS places them among the smallest

genomes found within the Thermotogae phylum (fig. 3A). Like in other prokaryotic taxa, Thermotogae GS is negatively correlated with OGT, with thermophiles having smaller genomes than mesophilic organisms (Sabath et al. 2013). However, within the *Thermosipho* genus there is a deviation from this trend where *T. africanus* has a larger genome and a higher OGT than the other species (fig. 3A). Additionally, *T. africanus* genomes have more ORFs compared with the other
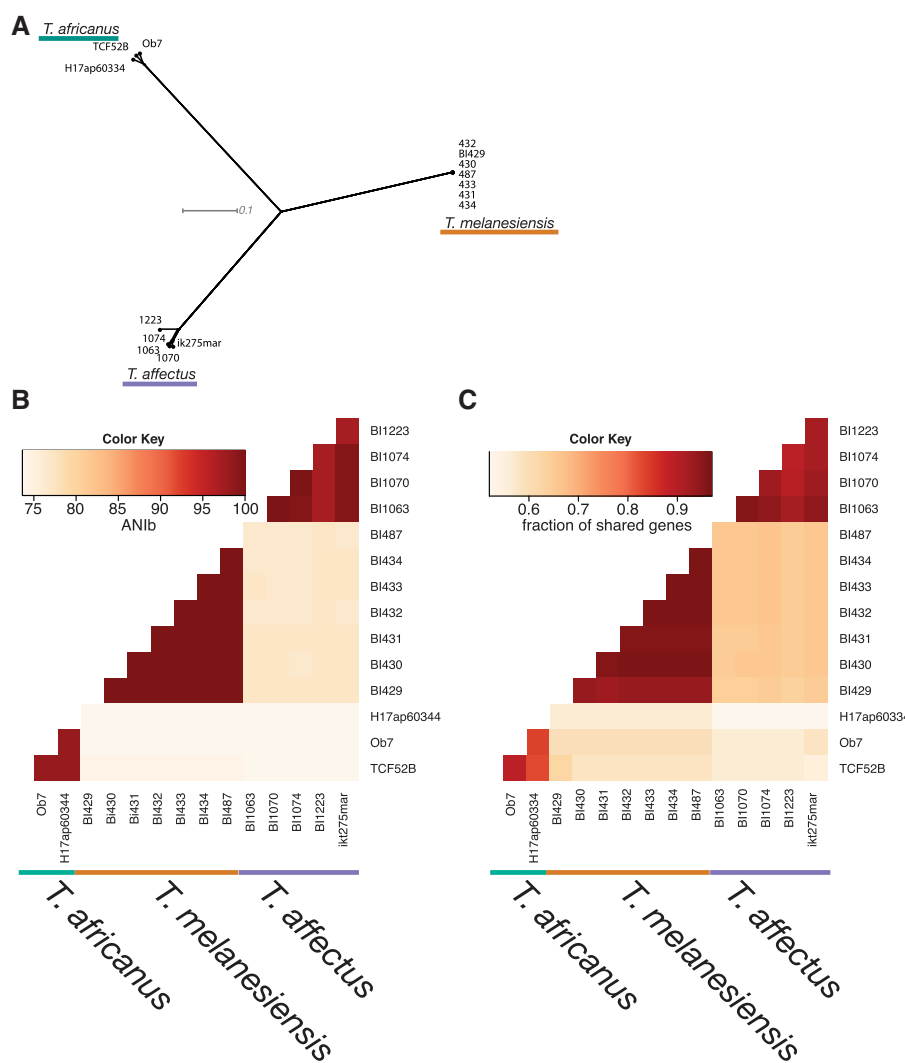
FIG. 2.—Phylogenomic comparison of 15 *Thermosipho* genomes using core SNPs, ANIb, and the number of shared genes. (*A*) Neighbor network of the 15 *Thermosipho* strains. The network is based SNP's in core genome fragments that were present in all genomes with a minimum of 70% similarity. The network was visualized in Splitstree using the NeighborNet algorithm from uncorrected distances (Huson and Bryant 2006). (*B*) Heatmap visualization of pairwise ANIb distances. (*C*) Fraction of genes shared between 15 *Thermosipho* genomes. Data were generated at the IMG-database using pairwise Bidirectional Best nSimScan Hits, with genes sharing 70% sequence identity and at least 70% coverage. Percentages were calculated by dividing the number shared genes by the total number of genes for the genome on the *y*-axis. The colored lines (black, green and blue) indicated which strains belong to which *Thermosipho* lineage/species.

two species (table 2) and larger relative intergenic region size (RIRS; 8.29%) compare to *T. melanesiensis* ($\approx 6.47\%$) and *T. affectus* ($\approx 6.92\%$) (table 1, fig. 3*C*). Finally, detection of recently acquired genes showed that the larger *Thermosipho* genomes have more HGT acquired genes than the smaller genomes (table 2; supplementary table S1, Supplementary Material online). These results indicate that temperature is not the only factor responsible for the observed *Thermosipho* GS dissimilarities, but that other unidentified processes also influence GS.

Environmental nutrient status was shown to have a positive correlation with prokaryotic GS and %GC, with low nutrient levels selecting for reduced GS and %GC via genome streamlining or by increasing the mutational load (Giovannoni et al. 2014; Gupta et al. 2016; Mende et al. 2017). Thus, under oligotrophic conditions prokaryotes experience selective pressure against the usage of high nitrogen content amino acids (aa). Accordingly, preference for aa with lower nitrogen and higher carbon content could be an indicator for nitrogen limited environments (Grzymski and Dussaq 2012; Mende et al. 2017). An analysis of average nitrogen and carbon molecules per aa-residue-side-chain (N-ARSC and C-ARSC) showed that *Thermosipho* genes have a median N-ARSC of 0.32 and a C-ARSC of 3.2 (fig. 3*C* and *D*). Such N/C-ARSC values are similar

Thermosipho spp. Genome Size Variation and Geographical Distributions

GBE
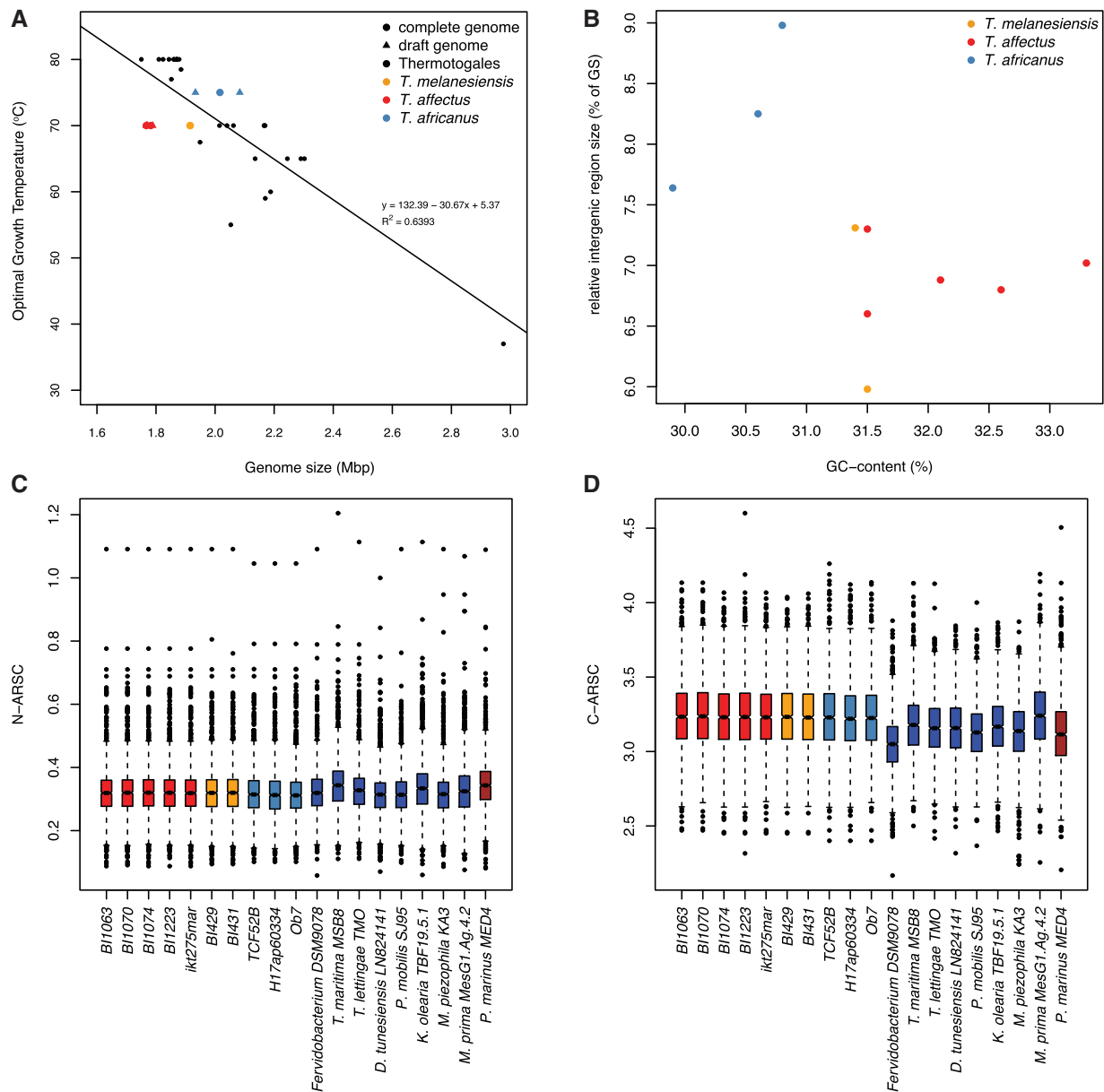
Fig. 3.—Genomic properties of the *Thermosipho* genomes. (*A*) Genome size versus optimal growth temperature for all *Thermosipho* genomes compared with all Thermotogales genomes, for which OGT is known. (*B*) %GC versus relative intergenic region size (% of total GS). (*C* and *D*) **Boxplot** comparison of N-ARSC & C-ARSC values per gene found in *Thermosipho* genomes, one representative genome of each Thermotogales species and *Prochlorococcus marinus* MED4.

to what is found in other Thermotogae and also in streamlined bacteria living in oligotrophic environments (e.g., *Prochlorococcus marinus* strain MED4) (fig. 3*C* and *D*) (Grzymski and Dussaq 2012), and are consistent with Thermotogae lineages being members of the nutrient poor and nitrogen limited subsurface biosphere (Head et al. 2014). These results could explain the generally small Thermotogae GSs.

A characteristic of small genomes is the presence of few paralogous genes. When considering all *Thermosipho* genomes, on average 14.7% of the COGs have paralogs

while 85.3% (± 0.7%) of the COGs are single-copy genes. These numbers are similar to those of other streamlined genomes (Grote et al. 2012). A comparison of COG annotations (fig. 4*A*; supplementary table S2, Supplementary Material online) revealed several categories with significant differences in relative abundance between the three species ($P \leq 0.01$ for B, C, H, I, J, O, and R and $P \leq 0.05$ for G and L). Two COG categories show large relative abundance differences between species that together may explain a significant fraction of the GS differences.

Table 2

Overview of Genome Content with a Focus on Mobile DNA Defence Systems and Mobile Elements

Genome	Closed	ORFs	Species-specific ORFs[a]	Strain Specific ORFs	Putative HGT ORFs	Defence Genes	CRISPR Arrays	Transposase Genes[b]	Prophage Present	Prophage ORFs
T. melanesiensis BI429*	Y	1,879	424	33	80	34	5	1	Yes	49
T. melanesiensis 431	Y	1,868	422	4	83	37	5	2 (2)[b]	Yes	53
T. sp. 1063	Y	1,706	349	11	66	31	4	0 (1)		
T. sp. 1070	Y	1,765	351	43	71	31	4	0 (1)		
T. sp. 1074		1,756	353	47	73	31	5	1	Yes	77
T. affectus strain ikt275mar		1,720	348	17	71	33	5	1 (1)		
T. sp. 1223		1,726	349	68	76	30	5	0 (1)		
T. africanus ob7		1,902	638	63	83	24	6	10		
T. africanus H17ap60334		1,982	656	169	116	19	6	0 (4)	Yes	48
T. africanus TCF52B*	Y	1,954	657	64	77	32	12	43		

[a]Each genome is used as a reference therefore species-specific orfs can show variations.
[b]Number in brackets indicates genes with disrupted open reading frames.

For the first category, COG G (Carbohydrate transport and metabolism), *T. africanus* genomes have more genes than the other two species (fig. 4A). Similarly, PFAM based comparisons using the carbohydrate database dbCAN (Yin et al. 2012) (Supplementary Material; supplementary table S3, Supplementary Material online), identified several enzyme families involved in breakdown of various beta-linked oligo- and polysaccharides and some of these were exclusively found among *T. africanus* representatives. Taken together with phenotypic differences (Podosokorskaya et al. 2014), this suggests that *T. africanus* spp. might be more versatile with regard to carbohydrate uptake and metabolism than the other two species and this could be at least partly responsible for its wider ecological distribution.

For the second category, COG H (coenzyme transport and metabolism), the large relative abundance difference is due to *T. affectus* genomes lacking most (20 out of 22) genes needed for corrinoid synthesis. Corrinoid synthesis is an essential part of de novo vitamin B_{12} synthesis and the lack of these genes suggests that *T. affectus* only rely on vitamin B12 precursor molecule recovery (supplementary fig. S3, Supplementary Material online). In contrast, a complete set of corrinoid synthesis genes is found in *T. africanus* and *T. melanesiensis* genomes, which are able to synthesize vitamin B_{12} (Swithers et al. 2012) (supplementary fig. S3, Supplementary Material online). It is unclear how this vitamin B_{12} pathway difference affects *Thermosipho* spp. ecology, but the loss of part of this system could be interpreted as a way to reduce metabolic costs under nutrient limited conditions.

GS in prokaryotes can be influenced by the presence of genomic islands, which often encode IS elements or prophages (Hacker and Kaper 2000). The *T. africanus* genomes have the highest abundance of transposase genes, while the two other species have few or only disrupted transposases (table 2). Besides IS elements each isolate contains blocks of colocalized genes unique for that genome (cluster size range: 4–31 genes). Interestingly, several of these regions encode

likely integrated prophages (*T. melanesiensis* BI429/BI431, *T. affectus* BI1074, *T. africanus* H17ap60334) as they contain typical phage genes (Supplementary Material, supplementary table S4, Supplementary Material online). Recent reports have identified three temperate siphoviruses in Thermotogae genomes, and together with the above findings, it becomes clear that prophages are not unusual genomic features within this phylum (Lossouarn et al. 2015; Mercier et al. 2018). The remaining clusters are dominated by hypothetical proteins, CRISPR-cas genes or genes involved in various types of cellular activities (supplementary table S4, Supplementary Material online). These findings indicate higher overall mobile element abundance in *T. africanus* genomes compare to both other species, which partially explains the larger GS of this lineage.

The *Thermosipho* spp. Immune Systems

The mobile element differences among *Thermosipho* spp. suggest dissimilarities between species in their foreign-DNA defense mechanisms. The *Thermosipho* genomes contain 44 COGs involved in such defense systems (Makarova et al. 2011); RM system (12), CRISPR-cas genes (18), and genes with other COG annotations (14) (fig. 4B).

The function of CRISPR-cas genes is to recognize and degrade mobile elements using CRISPR spacers. Each genome contains mostly unique spacers with only a few shared within species (supplementary fig. S4, Supplementary Material online). Considering CRISPR-cas genes, most are found in *T. affectus* isolates and *T. africanus* TCF52B, while *T. africanus* Ob7 and H17ap60334 genomes have the least. All genomes contain CRISPR-cas Type III systems (A & B, targeting RNA) while some have also Type I (targeting DNA) (Tamulaitis et al. 2017). The Type III system requires *cas10* (COG1353) and is detected at multiple loci in all genomes, except for *T. africanus* that have one gene (Type III-B). CRISPR-cas Type I requires *cas3* (COG1203) (Makarova et al. 2015). This gene is present in *T. affectus* BI1223 and *T. africanus* TCF52B, but disrupted in

Thermosipho spp. Genome Size Variation and Geographical Distributions

GBE
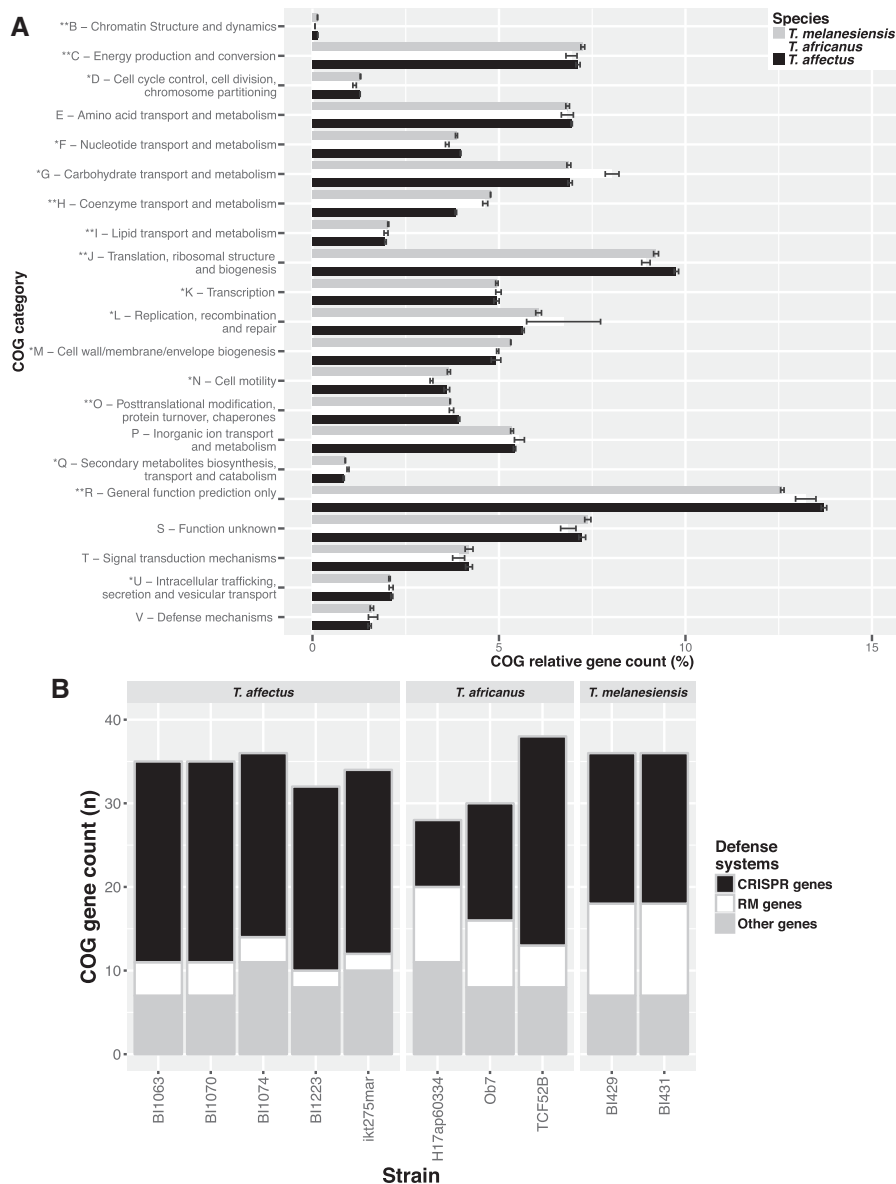
Fig. 4.—Comparison of COG functional annotations for three *Thermosipho* species. (*A*) Complete genome COG category annotations averaged for three *Thermosipo* species. COG category counts are relative to indicate proportional differences in gene content between the three species. The three species are indicated in the textbox in the figure. (*B*) Mobile DNA defense related COG annotation counts for 10 *Thermosipho* isolates. A total of 38 COGs were identified using a list of COGs related to defense genes (Makarova et al. 2011). The counts of the identified COGs were summarized into three groups: Restriction–modification systems COGs, CRISPR-cas associated COGs, and other COGs. The strains are grouped per *Thermosipho* species.

T. melanesiensis, *T. africanus* Ob7, and absent in *T. africanus* H17ap60334 (supplementary tables S5 and S6, Supplementary Material online). Finally, the remaining *T. affectus* genomes have a divergent *cas3* (NOG212552) (supplementary tables S5 and S6, Supplementary Material online). Overall, we found that CRISPR-cas systems differ across the genomes, which likely results in different CRISPR-cas immune responses among *Thermosipho* spp.

The RM system utilizes genome methylation and restriction enzymes to identify and attack foreign DNA.

T. affectus genomes lack restriction enzymes, suggesting the absence of a functional RM system (fig. 4*B*). Pacbio sequencing of two *T. affectus* isolates (BI1063/BI1070) indicated absence of genome methylation. In contrast, a methylated genome was detected in *T. melanesiensis* BI431 together with type I, II, and III restriction enzymes and DNA methylases. All the *T. africanus* genomes have a complete type III system, while type II is only present in strains Ob7 and H17ap60334 (supplementary tables S5 and S6, Supplementary Material online). It is unclear if

these genomes are methylated, since Pacbio sequencing was not performed.

The above analysis shows differences in foreign-DNA defenses between three *Thermosipho* species. Since each defense system can impose different fitness costs on the host (van Houte et al. 2016), our results suggest that each *Thermosipho* spp. experiences different phage predation pressures. For instance, when phage diversity is high, a non-specific defense system like the RM system might give a selective advantage despite increased fitness costs due to autoimmune reactions affecting host genome integrity. In contrast, the CRISPR-cas system is an inducible and highly precise response offering resistance to specific phages, and is optimal at low phage diversity and under small infection risk. Such immune systems differences may lead to long-term evolutionary implications for acquiring novel DNA via HGT or recombination. Several studies showed that: 1) organisms that rely on HGT lack CRISPR-cas systems since it can block HGT; 2) genomes without CRISPR-cas tend to be larger than closely related isolates with CRISPR-cas present; and 3) phase-variation of RM systems, switching of gene activity and phenotype, can lead to uptake of beneficial DNA which allows for adaptive speciation (reviewed in van Houte et al. 2016).

Therefore, based on these fitness effects of their immune systems, we propose the following interpretation of the observed GS differences. The presence of different CRISPR-cas types in *T. affectus* genomes has likely limited mobile element occurrence in those isolates. In addition, these systems may have induced genetic isolation and speciation by also limiting DNA recombination. In *T. melanesiensis* we find a complete CRISPR-cas Type III system and a full scale RM-system. This is able to tolerate the presence of a prophage in the genome, which would likely not be possible if an active Type I system was targeting the specific phage (Pyenson and Marraffini 2017). In contrast, the *T. africanus* defense systems with a more reduced CRISPR-cas and RM systems set-up, may have allowed genomic expansion by increased recombination via HGT and by higher mobile element activity. This could further have induced, or maintained, a larger metabolic potential as observed from the carbohydrate genes, permitting this species to thrive in a wider variety of ecosystems than the metabolically less diverse *T. melanesiensis* and *T. affectus* species.

Conclusions

Here we described three clearly separated species-level genomic lineages within the genus *Thermosipho*: *T. affectus*, *T. africanus*, and *T. melanesiensis*. Recombination detection showed very little interspecies but high intraspecies gene flow. This finding was supported by differences in GS, RIRS, %GC, and the presence of large species-specific gene sets, reflecting the metabolic differences between species. The species-specific genes sets probably influence distribution

differences of *Thermosipho* species in various subsurface ecosystems. *Thermosipho* bacteria frequently interacted with phages that were identified in several genomes. The differences in abundance of mobile elements among the three species are likely caused by immune system differences. Interestingly, the nature of the immune system could play a large role in genome maintenance and ecological adaptation, by differential limitation of genomic recombination via HGT. Taken together, our results suggests that genomic and phenotypic differences between the three *Thermosipho* species can be interpreted as the combinatorial result between genome streamlining and immune system composition.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

Literature Cited

Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 12(1):402.

Anderson RE, Sogin ML, Baross JA. 2014. Evolutionary strategies of viruses, bacteria and archaea in hydrothermal vent ecosystems revealed through metagenomics. PLoS One 9(10):e109696.

Antoine E, et al. 1997. *Thermosipho melanesiensis sp.* nov., a new thermophilic anaerobic bacterium belonging to the order Thermotogales, isolated from deep-sea hydrothermal vents in the southwestern Pacific Ocean. Int. J. Syst. Bacteriol. 47:1118–1123.

Archibald JM, Roger AJ. 2002. Gene duplication and gene conversion shape the evolution of archaeal chaperonins. J Mol Biol. 316(5):1041–1050.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10(1):421.

Charbonnier F, Erauso G, Barbeyron T, Prieur D, Forterre P. 1992. Evidence that a plasmid from a hyperthermophilic archaebacterium is relaxed at physiological temperatures. J Bacteriol. 174(19):6103–6108.

Dahle Dahle H, Garshol F, Madsen M, Birkeland N-K. 2008. Microbial community structure analysis of produced water from a high-temperature North Sea oil-field. Antonie van Leeuwenhoek. 93:37–49.

Thermosipho spp. Genome Size Variation and Geographical Distributions

GBE

Darmon E, Leach DRF. 2014. Bacterial genome instability. Microbiol Mol Biol Rev. 78(1):1–39.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Meth. 9(8):772.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE. 5:e11147.

Enright AJ, Van DS, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30(7):1575–1584.

Fuchsman CA, Collins RE, Rocap G, Brazelton WJ. 2017. Effect of the environment on horizontal gene transfer between bacteria and archaea. PeerJ 5:e3865.

Geslin C, et al. 2003. PAV1, the first virus-like particle isolated from a hyperthermophilic euryarchaeote, "*Pyrococcus abyssi*". J Bacteriol. 185(13):3888–3894.

Giovannoni SJ, Cameron Thrash J, Temperton B. 2014. Implications of streamlining theory for microbial ecology. ISME J. 8(8):1553–1565.

Goris J, et al. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 57(1):81–91.

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

Grote J, et al. 2012. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. mBio 3(5):e002525-12.

Grzymski JJ, Dussaq AM. 2012. The significance of nitrogen cost minimization in proteomes of marine microorganisms. ISME J. 6(1):71–80.

Gupta A, LaBar T, Miyagi M, Adami C. 2016. Evolution of genome size in asexual digital organisms. Sci Rep. 6(1):25786.

Hacker J, Kaper JB. 2000. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol. 54(1):641–679.

Head IM, Gray ND, Larter SR. 2014. Life in the slow lane; biogeochemistry of biodegraded petroleum containing reservoirs and implications for energy recovery and carbon management. Front Microbiol. 5:566.

Huber R, Woese CR, Langworthy TA, Fricke H, Stetter KO. 1989. *Thermosipho africanus* gen. nov., Represents a New Genus of Thermophilic Eubacteria within the '*Thermotogales*'. Syst Appl Microbiol. 12:32–37.

Hunt M, et al. 2013. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 14(5):R47.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23(2):254–267.

Koonin EV, Makarova KS, Wolf YI. 2017. Evolutionary genomics of defense systems in archaea and bacteria. Annu Rev Microbiol. 71(1):233–261.

Koonin EV. 2015. The turbulent network dynamics of microbial evolution and the statistical tree of life. J Mol Evol. 80(5–6):244–250.

Laing C, et al. 2010. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics. 11(1):461.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760.

Lossouarn J, et al. 2015. Ménage à trois': a selfish genetic element uses a virus to propagate within Thermotogales. Environ Microbiol. 17(9):3278–3288.

Makarova KS, et al. 2015. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 13(11):722–736.

Makarova KS, Wolf YI, Snir S, Koonin EV. 2011. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J Bacteriol. 193(21):6039–6056.

Margulies M, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380.

Markowitz VM, et al. 2009. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics 25(17):2271–2278.

Mende DR, et al. 2017. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nat Microbiol. 2(10):1367–1373.

Mercier C, et al. 2018. Two viruses, MCV1 and MCV2, which infect Marinitoga bacteria isolated from deep-sea hydrothermal vents: functional and genomic analysis. Environ Microbiol. 20(2):577–587.

Mino S, et al. 2017. Endemicity of the cosmopolitan mesophilic chemolithoautotroph Sulfurimonas at deep-sea hydrothermal vents. ISME J. 11:909–919.

Moldovan MA, Gelfand MS. 2018. Pangenomic definition of prokaryotic species and the phylogenetic structure of Prochlorococcus spp. Front Microbiol. 9:428.

Mukherjee S, Sengupta S. 2016. Riboswitch Scanner: an efficient pHMM-based web-server to detect riboswitches in genomic sequences. Bioinformatics 32(5):776–778.

Nakagawa S, et al. 2006. Microbial community in black rust exposed to hot ridge flank crustal fluids. Appl Environ Microbiol. 72(10):6789–6799.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Nesbø CL, et al. 2009. The genome of *Thermosipho africanus* TCF52B: lateral genetic connections to the *Firmicutes* and *Archaea*. J Bacteriol. 191(6):1974–1978.

Nesbø CL, et al. 2015. Evidence for extensive gene flow and *Thermotoga* subpopulations in subsurface and marine environments. ISME J. 9(7):1532–1542.

Overbeek R, et al. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 42(D1):D206–D214.

Podosokorskaya OA, et al. 2014. *Thermosipho activus* sp. nov., a thermophilic, anaerobic, hydrolytic bacterium isolated from a deep-sea sample. Int J Syst Evol Microbiol. 64(Pt 9):3307–3313.

Podosokorskaya OA, Kublanov IV, Reysenbach A-L, Kolganova TV, Bonch-Osmolovskaya EA. 2011. Thermosipho affectus sp. nov., a thermophilic, anaerobic, cellulolytic bacterium isolated from a Mid-Atlantic Ridge hydrothermal vent. Int J Syst Evol Microbiol. 61:1160–1164.

Pyenson NC, Marraffini LA. 2017. Type III CRISPR-Cas systems: when DNA cleavage just isn't enough. Curr Opin Microbiol. 37:150–154.

Quast C, et al. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41(D1):D590–D596.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16(6):276–277.

Richter M, Rosselló MR. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 106(45):19126–19131.

Roberts RJ, Vincze T, Posfai J, Macelis D. 2010. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 38(Suppl 1):D234–D236.

Sabath N, Ferrada E, Barve A, Wagner A. 2013. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. Genome. Biol Evol. 5(5):966–977.

Seed KD, Lazinski DW, Calderwood SB, Camilli A. 2013. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature 494(7438):489–491.

Smith AR, et al. 2017. Deep crustal communities of the Juan de Fuca ridge are governed by mineralogy. Geomicrobiol J. 34(2):147–156.

Swithers KS, et al. 2012. Vitamin B(12) synthesis and salvage pathways were acquired by horizontal gene transfer to the *Thermotogales*. Genome Biol Evol. 4(8):730–739.

Swofford DL. 2002. PAUP*: Phylogenetic analysis using parsimony (*and other methods). Sunderland (MA): Sinauer & Associates.

Takeuchi N, Kaneko K, Koonin EV. 2014. Horizontal gene transfer can rescue prokaryotes from muller's ratchet: benefit of DNA from dead cells and population subdivision. G3 (Bethesda) 4:325–339.

Tamulaitis G, Venclovas Č, Siksnys V. 2017. Type III CRISPR-Cas immunity: major differences brushed aside. Trends Microbiol. 25(1):49–61.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 30(12):2725–2729.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14(2):178–192.

van Houte S, Buckling A, Westra ER. 2016. Evolutionary ecology of prokaryotic immune mechanisms. Microbiol Mol Biol Rev. 80(3): 745–763.

van Wolferen M, Ajon M, Driessen AJM, Albers S-V. 2013. How hyperthermophiles adapt to change their lives: DNA exchange in extreme conditions. Extremophiles 17(4):545–563.

Wery N, Cambon Bonavita M-A, Lesongeur F, Barbier G. 2002. Diversity of anaerobic heterotrophic thermophiles isolated from deep-sea hydrothermal vents of the Mid-Atlantic Ridge. FEMS Microbiol Ecol. 41:105–114.

Xu Z, Han D, Cao J, Saini U. 2011. Cloning and characterization of the TneDI restriction: modification system of *Thermotoga neapolitana*. Extremophiles 15(6):665–672.

Yin Y, et al. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 40(W1):W445–W451.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5):821–829.

Zhaxybayeva O, et al. 2009. On the chimeric nature, thermophilic origin, and phylogenetic placement of the *Thermotogales*. Proc Natl Acad Ssci U S A. 106(14):5865–5870.

Zhaxybayeva O, et al. 2012. Genome sequence of the mesophilic Thermotogales bacterium Mesotoga prima MesG1.Ag.4.2 reveals the largest Thermotogales genome to date. Genome. Biol Evol. 4(8):812–708.

Zhu Q, Kosoy M, Dittmar K. 2014. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. BMC Genomics. 15(1):717.

Associate editor: Eugene Koonin