Single-cell genomics of *Treponema* and other microbiota isolated from xylophagous termites

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

David Edward Starns

September 2016

RIKEN

UNIVERSITY OF LIVERPOOL

僕達はこの長い旅路の
果に何を思う
誰も皆愛求め彷徨う
旅人なんだろう
共に行こう飽きる程に

浜崎あゆみ

Acknowledgments

This body of work would not have been made possible without my friends and family as support.

Firstly, I'd like to thank everyone at the Japanese collection of Microorganisms in Tsukuba "Koiin yanagatoshi", Masahiro and Jun-Ichi both of which helped me along the way, one constructively the other fuelling my alcoholism, my second mother Kayo who always knew that beer was extra delicious after sports, Yumi my wife and Suzu-neko (chome chome ninko). Anna Martinez, for keeping the age-old Anglo-Franco feud going on, and Kato-san is Mulan! Ohkuma-sensei, Iwaki-san and Kobe-san (Wakasugi-san, Numata-san), for being a great team. Kinjo, Tung-Ting and the Thai dream team (Melon, Ing, Nai, Dream, Yon and Pocky). The riken team Maya, Saito and Mihoko san, Nagamoriotakanomori and Abiko-chan. The Hongoh lab and Moriya lab.

Alistair, for being a great supervisor.

Fran for being someone I look-up to, someone going through the same, I love you. You were my motivation, friend, comrade, and crutch.

Ian Goodhead and the awesomeness of the Darby group

My friends Gwen, Sarah and Jen, I feel so much brotherly love. To Richard, Sam, Fatima and Linda. To the whole of the CGR - Nicky, Charlotte, Marg, Anita, Lisa, Richie and Lucille.

Chris for reminding me that no matter how shit I felt it could have been worse, you are an amazing friend! Don't surrender to the system!

My sister Josephane, you are so precious to me, I'd walk through the fire, my best friend Leisyen, there's enough fire for me to walk through for you too.

My mum and brother for funnelling me money near the end, my Dad and my sister, I love you all so much and Ayu.

Thank you all for the support.

Table of contents

Abstract

The nutritional symbiosis between the wood feeding termite and its microbiota is an important model of biomass conversion. Recalcitrant wood is broken down to its basic components and is transformed into energy for the sustenance of the host termite. The microbial community is responsible for cellulose degradation, reductive acetogenesis and nitrogen fixation activities within the gut, and is composed of diverse bacteria, archaea and protists. A prominent taxon within the community is *Treponema*, a genus of helical motile bacteria.

In the higher termite *Nasutitermes takasagoensis* treponemes are the dominant taxa with the potential of reductive acetogenesis, nitrogen fixation and cellulose degradation, they are joined by members of TG3 and Fibrobacteres and work synergistically in the utilisation of recalcitrant wood components. The genomes of these organisms were analysed using single cell genomics to assign putative functions to these uncultured bacteria and metatranscriptomics to uncover the contributions within the whole community.

In the lower termite *Hodotermopsis sjoestedti*, treponemes are dominant taxa within the cellulolytic protist Eucomonympha. Here the treponeme has evolved an endosymbiotic lifestyle to sustain the host by reductive acetogenesis from H_2 and CO_2 to supply acetate and nitrogen fixation, and has lost its helical and motile characteristics. This elucidation was made possible again by use of single cell genomics and biochemical analyses.

Treponemes from the lower termite *Reticulitermes speratus* were also sequenced and used in conjunction with other treponeme sequences from diverse environments to classify important genes associated with the environments they were isolated from. Virulence factors from human periodontal strains were the most important at distinguishing the environment they were associated with, whereas termite associated treponemes were categorised by central metabolic genes.

List of Abbreviations

16S rRNA	Small subunit bacterial RNA
ABC	Adenosine tri phosphate binding cassette
BLAST	Basic local alignment search tool
CAZY	Carbohydrate active enzymes database
CMFDA	5-chloromethylfluorescein diacetate
CoA	Coenzyme A
DNA	Deoxyribonucleic acid
FACS	Fluorescence activated cell sorting
FBP	Fructose bisphosphotase
FISH	Fluorescence in-situ hybridisation
GH	Glycosyl hydrolase family
GTR	Generalised Time Reversible
KEGG	Kyoto encyclopedia of genes and genomes
MALBAC	Multiple annealing and loop based amplification
MCL	Markov clustering algorithm
MCP	Methyl accepting chemotaxis protein
MDA	Multiple displacement amplification
MEGA	Molecular evolution genetics analysis
MIDAS	Microwell displacement amplification system
ML	Maximum likelihood algorithm
MUSCLE	Multiple sequence comparison by Log-cxpectation
N50	Shortest sequence length at 50% of the genome
NGS	Next generation sequencing
PCR	Polymerase chain reaction
PROKKA	Prokaryote annotation
RAST	Rapid annotation using subsystem technology
RFLP	Restriction fragment length polymorphism
RIKEN	National Institute of Physical and Chemical Research
RIN	RNA integrity number
RNA	Ribonucleic acid
RPKM	Reads per kilobase of transcript per million mapped reads

RT-PCR	Reverse transcriptase polymerase chain reaction
SCG	Single cell genomics
SRA	Short read archive
TG3	Termite group 3
WAG	Whelan and Goldman
WGA	Whole genome amplification

Chapter I

General Introduction

Wood feeding termites

1.1.1 Termite biology and evolution

Termites are insects from the order Blattodea (infraorder Isoptera); they are colloquially known as eusocial cockroaches due to their evolutionary history and development of an insect caste system. Termites diversified from the last common ancestor of wood feeding cockroaches (Cryptocercoidae) and Termitoidae between 170-150 MYA during the Jurassic period (Bourguignon *et al*., 2015, Figure 1). They inhabit diverse environments and are currently found on every continent except Antarctica.

Termites unlike cockroaches have a tripartite caste system, with worker, soldier and winged alate forms, however not all species exhibit all these forms. The caste system has been shown to be present in fossil represented species from the early Cretaceous period circa 100MYA (Engel *et al*., 2016). There have been over 3,000 species of termite classified worldwide (Krishna *et al*., 2013) with most species described from the apical family Termitidae and also show the most diversity.

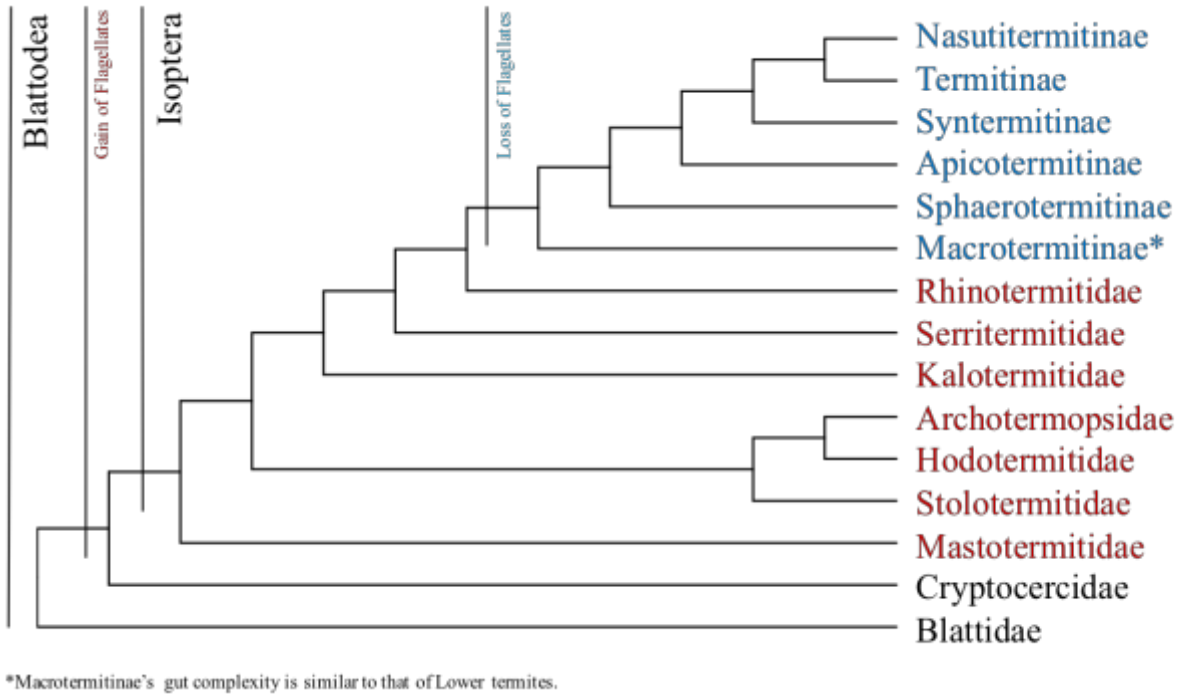

Figure 1. Cladogram representing evolution of the Blattodea order. The apical higher termites are represented in blue as sub-families (family Termitidae), lower termites in red, cockroaches in black, adapted from Brune & Dietrich 2015.

1.1.2 Lower and higher termites

Termites are categorised as either higher or lower, based on the presence of flagellated gut protists. Lower termites are the evolutionarily basal lineages (Figure 1) that harbour cellulolytic protists within their guts, they tend to be wood feeding and sharing highest similarity to that of wood feeding cockroaches. Higher termites are apical lineages that have evolved and lost their gut protists, they are much more variable in terms of microbial diversity, diet and gut complexity (Brune and Dietrich, 2015). Higher termites can feed on fungi, soil, wood, grass and derivatives (Donovan *et al*., 2001) and the gut usually includes a mixed segment and proctodeal segment 1, an alkaline compartment (Bignell *et al*., 1983) (Figure 2). Soil feeding higher termites such as *Cubitermes* species are shown to have even higher pH levels within this gut segment, hypothesised for the proficient breakdown of humic substances (Brune and Kuhl, 1996). The fungal growing higher termites such as *Macrotermes* species tend a fungal garden of the associated *Termitomyces* fungal species (Poulsen *et al*., 2014) which acts as a food source and decomposer of recalcitrant lignocellulose. Within this family the structure of the gut is less complex and comparable to lower termite families.

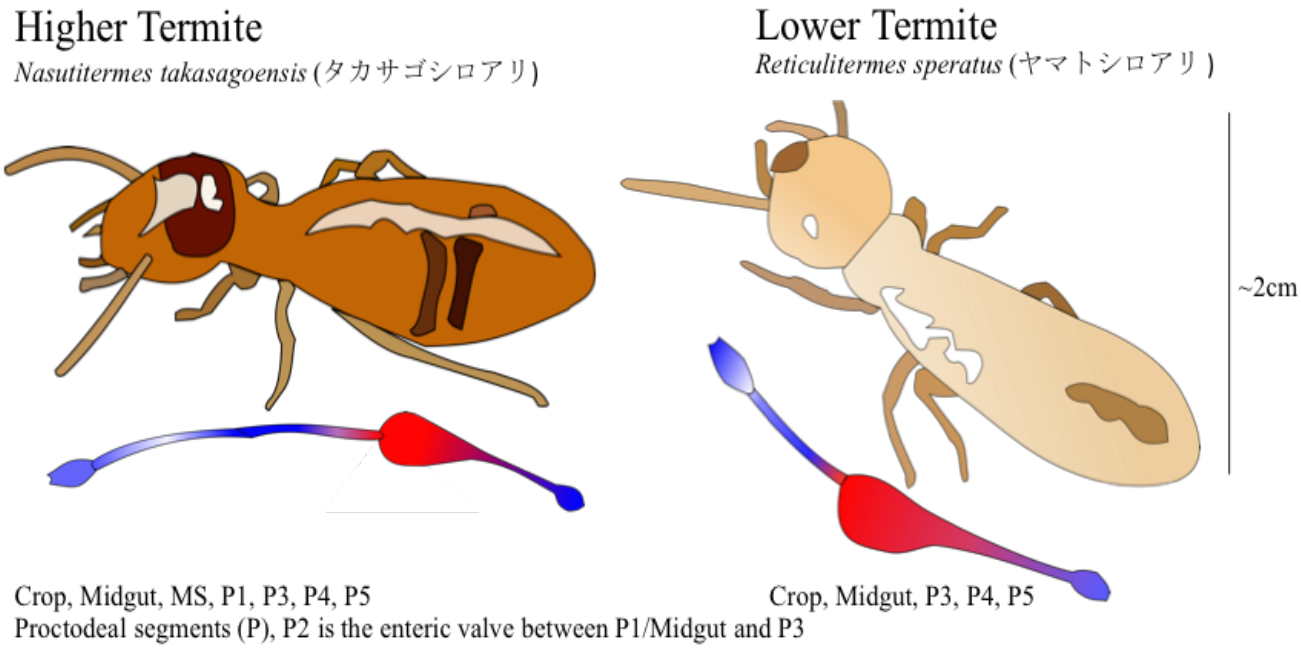

Figure 2. Two examples of wood eating termites, the higher termite has a greater complex gut structure that includes an alkaline mixed segment and first proctodeal segment, the lower termite is simpler in comparison.

1.1.3 Termite microbiology and diversity of microbiota

Termites are most widely known as detrimental economic pests from the ability of many species to survive on wood. This in part is due to the nature of their microbiota, which is highly conserved between species because of the practice of proctodeal trophylaxis, coprophagy and continued social interactions (Diouf *et al*., 2015). Protists, Archaea, Fungi and Bacteria reside in the guts of termites. The Protists observed in termite guts are oxymonad and parabasilid flagellates and are unique to termites and wood feeding cockroaches; these are known to be

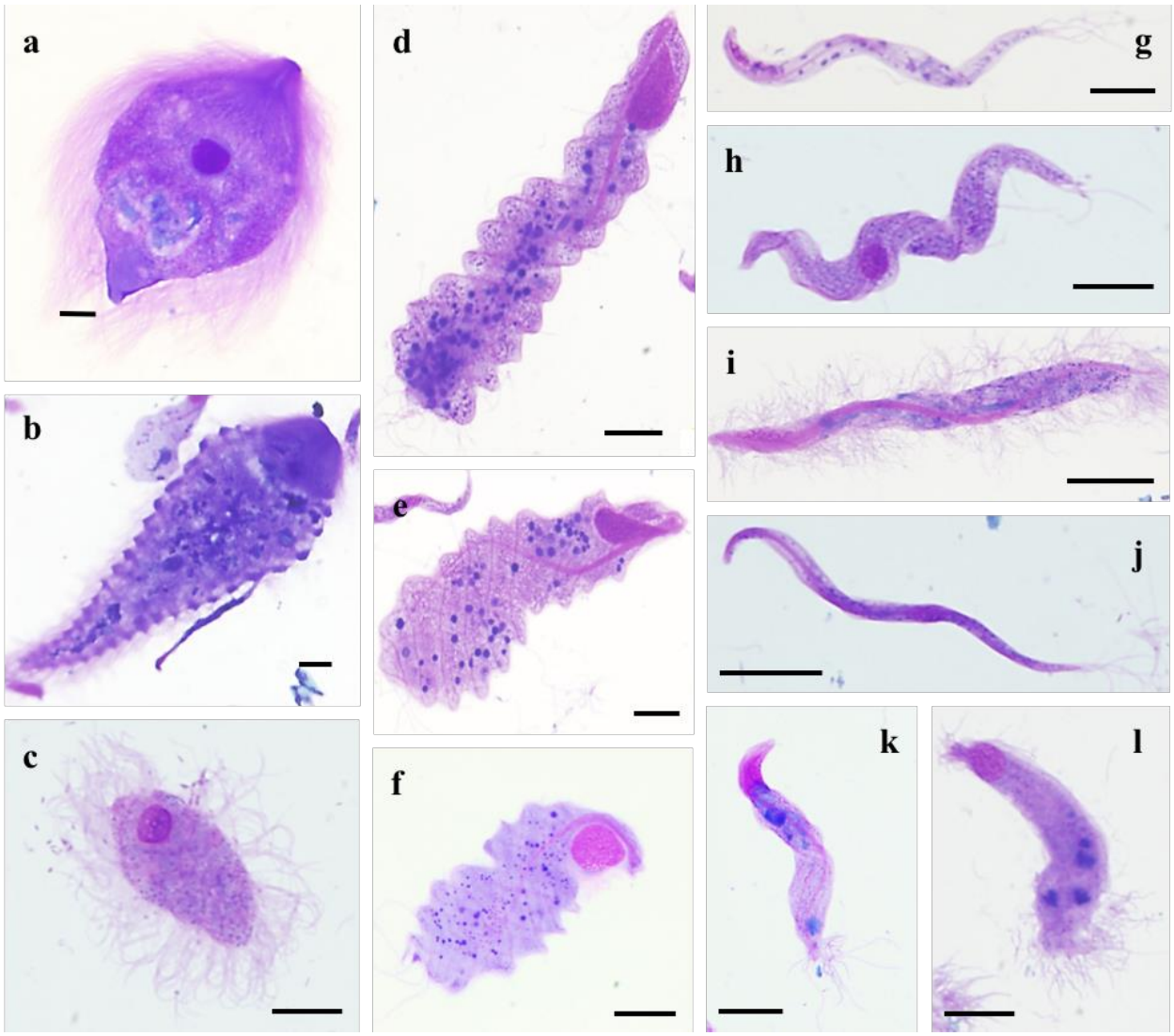

Figure 3. Protists isolated from the gut of *R. speratus*, morphological diversity of protists isolated from the gut of one lower termite consisting of both parabasasilid species (a) *Teranympha mirabilis* (b) *Trichonympha agilis* and (c) *Holomastigotes elongatum*, and oxymonad species (d-f) *Pyrsonympha sp*. (h-l) *Dinenympha sp*.
Protists are stained with a Giemsa like stain RAL-555. The scale bar indicates 10 µm. Pictures courtesy of Anna Martinez (unpublished) (first described in Koidzumi, 1921).

essential to termites in the breakdown of cellulose (Cleveland, 1923). Many species of Protists can inhabit one single termite and are morphologically distinctive (Figure 3).

At a greater diversity are the prokaryotes that inhabit the guts. In higher termites, the bacterial gut microbiota is also essential in maintaining survival (Eutick *et al*., 1978). The bacteria are further highly diversified in phyla, and many species are characteristic of termite species and of diet in levels of abundance; it is estimated that 6,000 different bacteria can exist within one millilitre of gut fluid that differ in 97% sequence identity of their 16S RNA gene (Hongoh *et al*., 2003). The most common phyla found are Bacteroidetes, Firmicutes, Spirochaetes, Proteobacteria, Fibrobacteres and Elusimicrobia and Figure 4 illustrates the phyla diversity in *Nasutitermes takasagoensis* gut microbiota (data taken from chapter II). These bacteria can be free living symbionts within the termite gut or associated with Protists creating tripartite matryoshka-like symbioses as ecto- or endosymbionts and hosts can harbour one or more of these symbionts for example in *Trichonympha* spp.

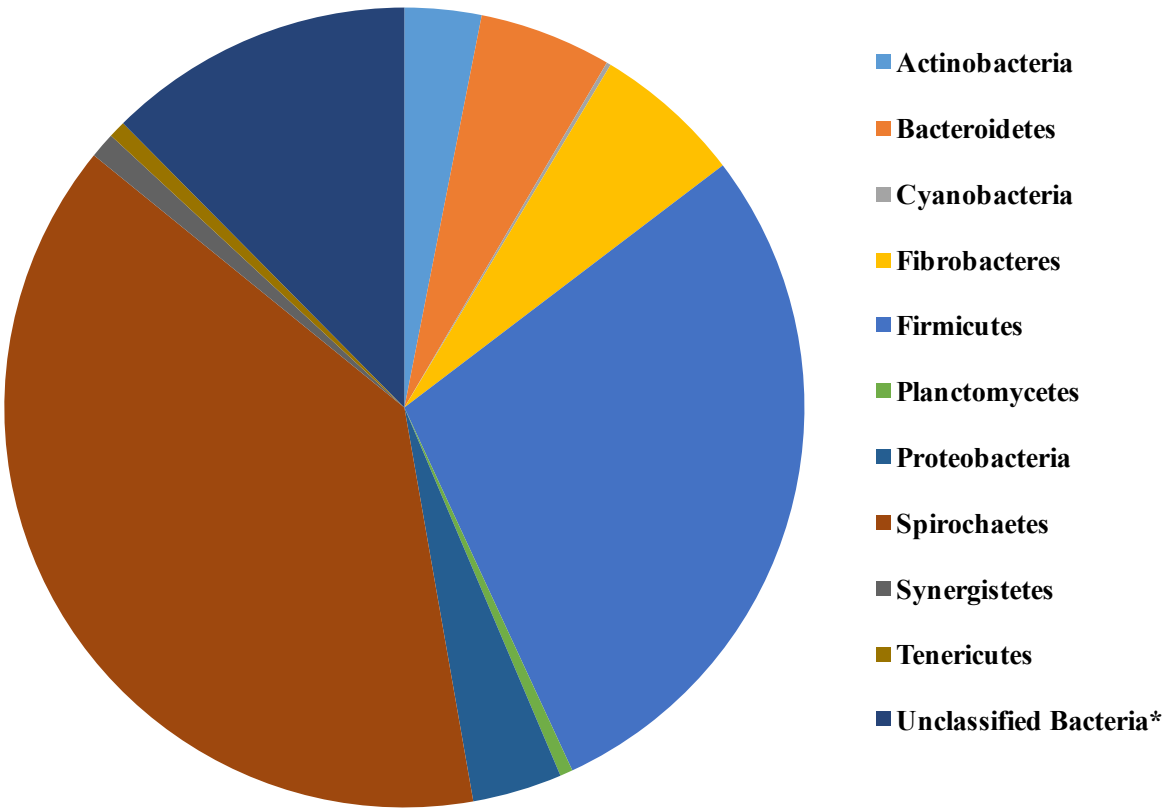

Figure 4. Pie chart representing proportion of 16S rRNA sequences at phylum level derived from isolated RNA from gut content of *Nasutitermes takasagoensis*. Data obtained from 16S rRNA sequences from the sequenced gut metatranscriptome of *N. takasagoensis* in chapter II. *Unclassified Bacteria represent sequences not identified at 60% percentage identity.

1.1.4 Functions of the microbial gut community

The gut community make up a nutritional symbiotic system, where the consortia are there to deal with the dietary requirements of the host. Wood feeding termites, both higher and lower, require their microbiota to convert recalcitrant wood into an energy source; in the termite this is the short chain fatty acid acetate that can support 100% of respiratory requirements (Odelson and Breznak, 1983). The conversion of wood into acetate relies on glycosyl hydrolase action as well as glycolytic pathway processing. In the lower termite, this process is primarily localised within the protists, where ingested and masticated wood particles undergo enzymatic hydrolysis (Brune and Dietrich, 2015).

Some species of termite harbour endogenous glycosyl hydrolases, however the significance of these is not fully understood (Lo *et al*., 2011). Protistan and Prokaryotic glycosyl hydrolase provisions are required for the breakdown of recalcitrant wood components of cellulose and hemicellulose, lignin is not utilised and excreted in the frass. A full complement of different glycosyl hydrolases are required to utilise all the components of cellulose and highly substituted hemicelluloses. Recalcitrant wood is also deficient in nitrogen and is essential for growth, the microbiota also performs the essential processes of nitrogen fixation and recycling, and also thought to upgrade these into amino acids and cofactors. It is thought that these essential processes have led to selective pressures to co-speciate host with symbiont (Noda *et al*., 2007). In fungal garden cultivating species, the symbiosis relies on a higher number of chitin/fungal polysaccharide degrading enzymes to utilise the fungal comb as a food source (Poulsen *et al*., 2014). The diet has also been shown to affect the composition of the gut microbiota and is the primary determinant for bacteria community structure (Miyata *et al*., 2007; Mikaelyan *et al*., 2015).

1.1.5 Treponemes as wood feeding termite gut community members

In certain species of termites, members of the phylum Spirochaete form the bulk of the bacterial community, as in the wood feeding lower termite *Reticulitermes speratus* and the higher termite *Nasutitermes takasagoensis*. These Spirochaetes are diverse members of the genus *Treponema* (Hongoh *et al*., 2003; 2006). Only three species of *Treponema* have been successfully isolated and cultured from lower termites, and none from higher termites; *Treponema primitia* and *Treponema azotonutricium* (*Zootermopsis angusticolis*; Graber *et al*., 2004) and *Treponema isoptericolens* (*Incisitermes tabogae*; Dröge *et al*., 2008). Utilising 16S

rRNA gene sequences of uncultured bacteria and environmental samples, the phylogeny of treponemes has been refined from the original two termite associated clusters (Ohkuma *et al*., 1999), to multiple sub clusters within Treponema cluster I (Mikaelyan *et al*., 2015).

Treponemes have evolved multiple functions and efficiencies for essential processes within the termite gut, *T. primitia* (strain ZAS-2) is an efficient homoacetogen utilising the Wood-Ljungdahl pathway to fully convert H_2/CO_2 to acetate (Leadbetter *et al*., 1999) and *T. azotonutricium* is proficient in fixing nitrogen (Graber *et al*., 2004).

The *Treponema* genus is very diverse and more widely known for their incidences of causing disease than symbiosis. The genus includes the human sexually transmitted infection syphilis causative pathogen *Treponema pallidum* that was first described in 1905 by Schaudinn and Hoffman with the disease being known for over 500 years, other pathogenic species are responsible for periodontitis in humans and digital dermatitis in cattle. Some species have been isolated from bovine rumen that are thought to be symbiotic (*Treponema bryantii*) and even non-host associated from a thermophilic mat (*Treponema caldarium*). Both species are thought to potentiate the growth and metabolism of secondary microorganisms similar to those seen in the termite gut (Stanton and Canale-Parola, 1980, Pohlschroeder *et al*., 1994, Rosenthal *et al*., 2011).

1.2.1 The structure of wood

Wood is the structural fibrous biomass component of trees and shrubs. Its composition is of lignocellulose compacted into plant cell fibers, as primary and secondary cell walls. The three major components of lignocellulose are that of cellulose, hemicellulose and lignin, in different proportions. Hardwoods, those of eudicot plants have a composition of 43±2% cellulose, 20 to 35% hemicellulose and 18 to 25% lignin (Timell, 1967). The arrangements of these components produce its recalcitrance. Cellulose is a biopolymer of β-D-glucopyranose, arranged in crystalline microfibrils or amorphously throughout the plant cell walls. Hemicellulose unlike cellulose, is a heteropolymer and usually comprised of long chains of xylose, glucose or mannose and highly substituted with glycosidic residues. Eudicots usually possess different proportions of hemicelluloses, xyloglucan at 20 to 25% and glucuronoarabinoxylan at 5% in their primary cell walls, glucuronoxylan at 20 to 30% in their secondary cell walls and galactogluconmannan throughout both (Scheller and Ulvskov, 2010). The last component is lignin, a large heteropolyphenolic polymer that encrusts both the cellulose framework and hemicellulose components of wood.

1.2.1 Lignocellulose degradation

The utilisation of lignocellulose by microbial communities and other organisms comes from enzymatic hydrolysis of the carbohydrate components. The recalcitrant nature of wood means that usually physical processing i.e. mastication by the insect needs to be completed to expose and increase the surface area of the wood components for downstream processing. Exposure to other environmental factors can increase the decomposition of the structure of wood e.g. alkalinity, but full decomposition requires the action of glycosyl hydrolases. There are many families of glycosyl hydrolases (GH), that have the potential to degrade the many components of wood – cellulose and hemicellulose. These have been categorised within the Carbohydrate Active enZyme (CAZy) database and are important in the bioprocessing of plant biomass as one of the futures only sustainable energy sources (Lynd *et al*., 1999). CAZy enzymes (CAZymes) do not just cover glycosyl hydrolases but also transferases, lyases, esterases, carbohydrate binding modules and some redox enzymes associated with other CAZymes. Lignin's degradation is not covered by the CAZy database but rather is processed by laccases, lignin modifying enzymes and the most important lignin degrading enzymes – the peroxidases. To date there are 136 glycosyl hydrolase families, four families of which are further divided into subfamilies, based on amino acid sequence homology (Lombard *et al*., 2014; http://www.cazy.org/). For the degradation of lignocellulose from woody biomass many of these families are not required but the major families required for this are those that are active on cellulose (GH1, GH3, GH5, GH6, GH8, GH9, GH12, GH45, GH48, GH51 and GH74) and those acting on hemicellulose components (GH2, GH10, GH11, GH16, GH26, GH30, GH31, GH39, GH42, GH43 and GH53).

1.2.1 Symbionts and the nature of the termite symbiosis

In higher termites, treponemes exist as free-living gut bacteria. Within the lower termites treponemes exist as free-living gut bacteria, and as both endosymbionts (Chapter IV) and as ectobionts attached to various gut flagellates. This has been shown by fluorescence in-situ hybridisation (FISH) (Iida *et al*., 2000, Inoue *et al*., 2008), but the function of these ectosymbiotic treponemes remains unclear. Non-treponemal ectosymbionts have been shown to provide motility for their host protist; the Synergiste, 'Candidatus *Tamella caducaeia*' (Hongoh *et al*., 2007a). The Bacteroidales ectosymbiont 'Candidatus *Symbiothrix*

dinenymphae' (Hongoh *et al*., 2007b, Yuki *et al*., 2015) is proposed to support the host protist by the fermentation of lignocellulose components and the upgrade nitrogenous compounds.
There are many examples of these symbioses in the termite gut that have recently been studied and pertain to diverse bacteria. The localisations of these endo-ecto symbionts vary among hosts, from residing in the cytoplasm near glycogen granules 'Candidatus *Ancillula trichonymphae*' (Strassert *et al*., 2016), and near hydrogenosomes 'Candidatus *Adiutrix intracellularis*' which is also co-localised with the ectosymbiont 'Candidatus *Desulfovibrio trichonymphae*' (Ikeda-Ohtsubo *et al*., 2016) in the host *Trichonympha collaris*, to even localising in the nucleus 'Candidatus *Nucleococcus trichonymphae*' (Sato *et al*., 2014).

1.2.2 Genomic techniques to study termite gut microbiota

Termite guts are highly compartmentalised with environmental conditions changing from segment to segment. Many of these organisms are completely anaerobic, swimming in the anoxic lumen, whereas others scavenge oxygen as micro-aerophiles attached to the outer most boundary of the gut. The complexity of this environment often means that the production of pure cultures is near impossible and without culture, these organisms' functions remain elusive. However, comparative genomics and other multi-omic approaches can explore these organisms without the need to culture. Using single cell genomics or metagenomic approaches can give insights into the diversity and function of these communities and with the reducing costs of genome sequencing, these techniques are available to many more researchers.
The key diversity and stability of termite gut microbial communities has been profiled in many studies by exploring 16S rRNA gene sequencing from initially clone and restriction fragment length polymorphism (RFLP) analyses (e.g. Ohkuma and Kudo, 1996) to next generation microbial profiling (e.g. Otani *et al*., 2014, Rahman *et al*., 2015, Reid *et al*., 2014). These studies underline what taxa are present in these communities, but they do not show the functional contributions of individual members to the community. Metagenomics has given some insight to both the taxa present and the overall community function, but functions could only be assigned putatively to taxa (Warnecke *et al*., 2007). Single cell genomic investigations have been used to clarify roles of symbiotic relationships and attribute functions to individual taxa.

1.2.3 Single cell genomics

Single cell genomics (SCG) is the isolation of a single cell, the amplification of the cell's genome/DNA and the sequencing of this information to establish the organism's genome. Considerations should be given to sample collection and preservation as unculturable samples are usually collected from diverse environments and these processes may affect the downstream applications (Clingenpeel *et al*., 2014). There are many techniques employed in this field of science for single cell isolation, from fluorescence activated single cell sorting (FACS), micromanipulation, microdissection to microfluidics. These methods each have different strengths and limitations, micromanipulation allows for accurate visual confirmation of single cells, but is low throughput and suitable for low cell numbers. Microfluidic and microencapsulation techniques offer higher throughput but there are challenges with administering the downstream reactions and in drop selection (Blainey, 2013). FACS is a widely-used tool for the isolation of single cells. It allows a large throughput of single cells to be sorted and advances in technology have greatly refined the precision of this platform. The basis of FACS sorting is the staining of a sample; the instruments' laser excites the fluorophore and this is detected, the particle is then electrically charged and sorted into collection. All techniques are subject to availability of facilities, cost considerations and dependant on the desired target material.

After individual cells are isolated whole genome amplification (WGA) is required to generate enough DNA for sequencing, single bacteria possess only femtograms of DNA (10^{-15} grams). Next generation sequencing (NGS) platforms require inputs of at least nanograms (10^{-9} grams) of input sample. The most commonly used method of WGA uses the highly processive Phi29 DNA polymerase in multiple displacement amplification (MDA (Dean *et al*., 2002)) a method that amplifies DNA a billion fold; other methods include using similar enzymes or techniques such as multiple annealing and loop based amplification cycles (MALBAC (Zong *et al*., 2012)) or microwell displacement amplification system MIDAS (Gole *et al*., 2013). These methods generate enough DNA for next generation sequencing, no method is yet gold standard (de Bourcy *et al*., 2014), but other methods used in conjunction such as using gel microdroplets (Dichosa *et al*., 2014) as a pre-isolation technique have been shown to improve genome recovery.

In comparison to SCG, metagenomics is a technique where heterogenous populations or unknown mixed environmental samples are sequenced. Contigs that are assembled are

putatively assigned taxonomy however their cellular origins are questionable. Improved bioinformatics and deep sequencing can allow for the extrapolation of whole single organism recovery in the binning of metagenomic sequences (Albertsen *et al*., 2013).

1.2.4 Previous SCG work with termites and beyond

The study of termite gut genomic endosymbioses has benefitted with the utilisation of single cell genomics. The first such study was applied to the endosymbiont of the parabasalid *Pseudotrichonympha grassi,* whose genome was recovered and sequenced – 'Candidatus *Azobacteroides pseudotrichonymphae*' (Hongoh *et al*., 2008a). Genomic information established this endosymbiont could fix nitrogen and shortly after, another endosymbiont isolated from *Trichonympha agilis* – 'Candidatus *Endomicrobium trichonymphae*' was also thought to be involved in the provision of nitrogenous resources such as co-factors and amino acids, to the host protist (Hongoh *et al*., 2008b).

Single cell genomics was applied in understanding the putative roles of the Bacteriodetes ectosymbiont 'Candidatus *Symbiothrix dinenymphae*' (Yuki *et al*., 2015), however SCG usage has not only be used to research the unculturable bacteria residing in the termite gut nutritional symbiosis but in many other studies where classical microbiological techniques have failed to shed light. In the case of the honey bee (*Apis melifera*) SCG was used to study diversification of two nutritional symbionts and heterogeneity within their populations (Engel *et al*., 2014).

The first application of SCG on an individual bacterial sample came in 2005 by Raghunathan and colleagues in sequencing the model organism *Escherichia coli*, however the potential at looking at unculturable bacteria wasn't realised until a few years later. The expansion of SCG usage came about after the publication of insights into 'microbial dark matter' by Rinke and colleagues in 2013. This was the largest scale single cell genome study helping to resolve tree of life and sampled multiple environments for unculturable bacteria. The resulting 201 uncultured bacteria and archaeal genomes expanded our knowledge of the microscopic world (GEBA -genome encyclopaedia of bacteria and archaea). The research also helped in identifying genes ubiquitous within bacteria that could be used in the determination of genome recovery using SCG.

More recently, microbial SCG studies have focused on exploring the diversity of varied uncultured environments, these include hot springs (Dodsworth *et al*., 2013), marine (Zhang *et al*., 2016), volcanic (Field *et al*., 2015), oceanic trench (Fullerton and Moyer, 2016), lake (Martinez-Garcia *et al*., 2012a; Martinez-Garcia *et al*., 2012b) and those focusing on other

symbiotic systems (Siegl *et al*., 2011). Many of these studies unveiled novel processes and unique metabolic activities of uncultured organisms for example active polyketide and peptide production by a filamentous sponge symbiont (Wilson *et al*., 2014) and pathways associated with sulphur oxidation in the *Thiovulum* genus (Marshall *et al*., 2012). There are still large holes within our knowledge of the microscopic world however SCG and other culture independent analyses will aid us in furthering this information.

1.3 Research objectives

The aim of this body of work is to explore the functions and diversity of the genus *Treponema* in species of wood-eating higher and lower termites, to understand what roles they play within

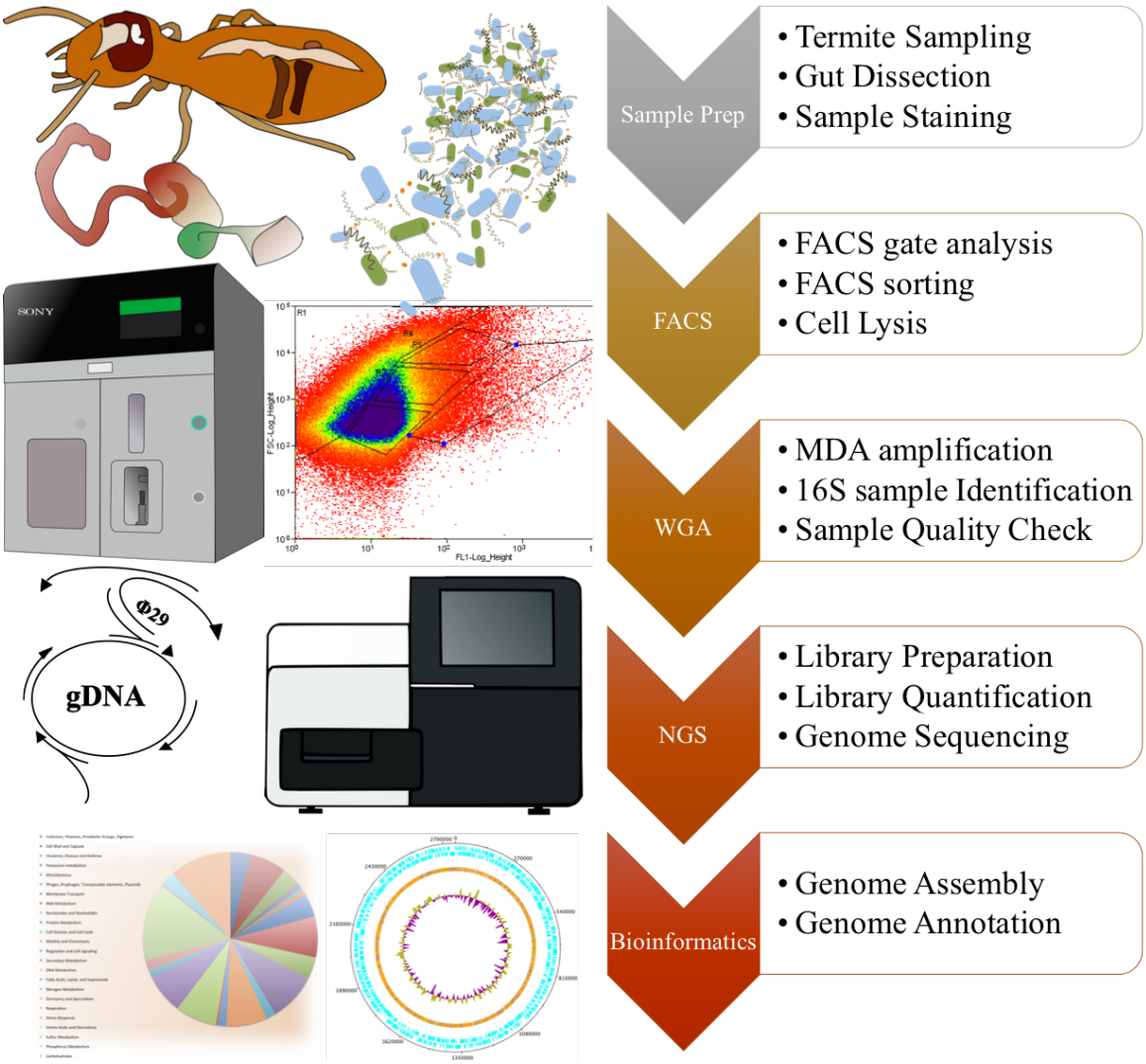

Figure 5. Single cell microbe exploration workflow. The standard approach which was taken in obtaining the genomes of unculturable organisms. The sample quality check is based on the phred scores of sanger sequencing the 16S rRNA gene sequence, to eliminate the chance of multiple cell contamination through the sorting step.

this community using single cell genomics and methods independent of microbiological culturing. The genus *Treponema* was selected due to their documented abundance in the guts of wood-feeding termite species, at over 45% of the gut community in *Nasutitermes* and *Reticulitermes* termites (Brune, 2014).

The workflow used to analyse these treponemes is illustrated in figure 5. Firstly, for sample preparation the termites were collected from different locations around Japan (sampling sites are shown in Figure 6), and in the laboratory, the termites guts were dissected out and contents exuded. Termite gut exudate is then prepared for FACS, and sorted into 96/384 well plates followed by MDA of individual sample wells. 16S rRNA gene sequence integrity was checked for each sample after sequencing using the phred quality score for sanger sequencing, this can

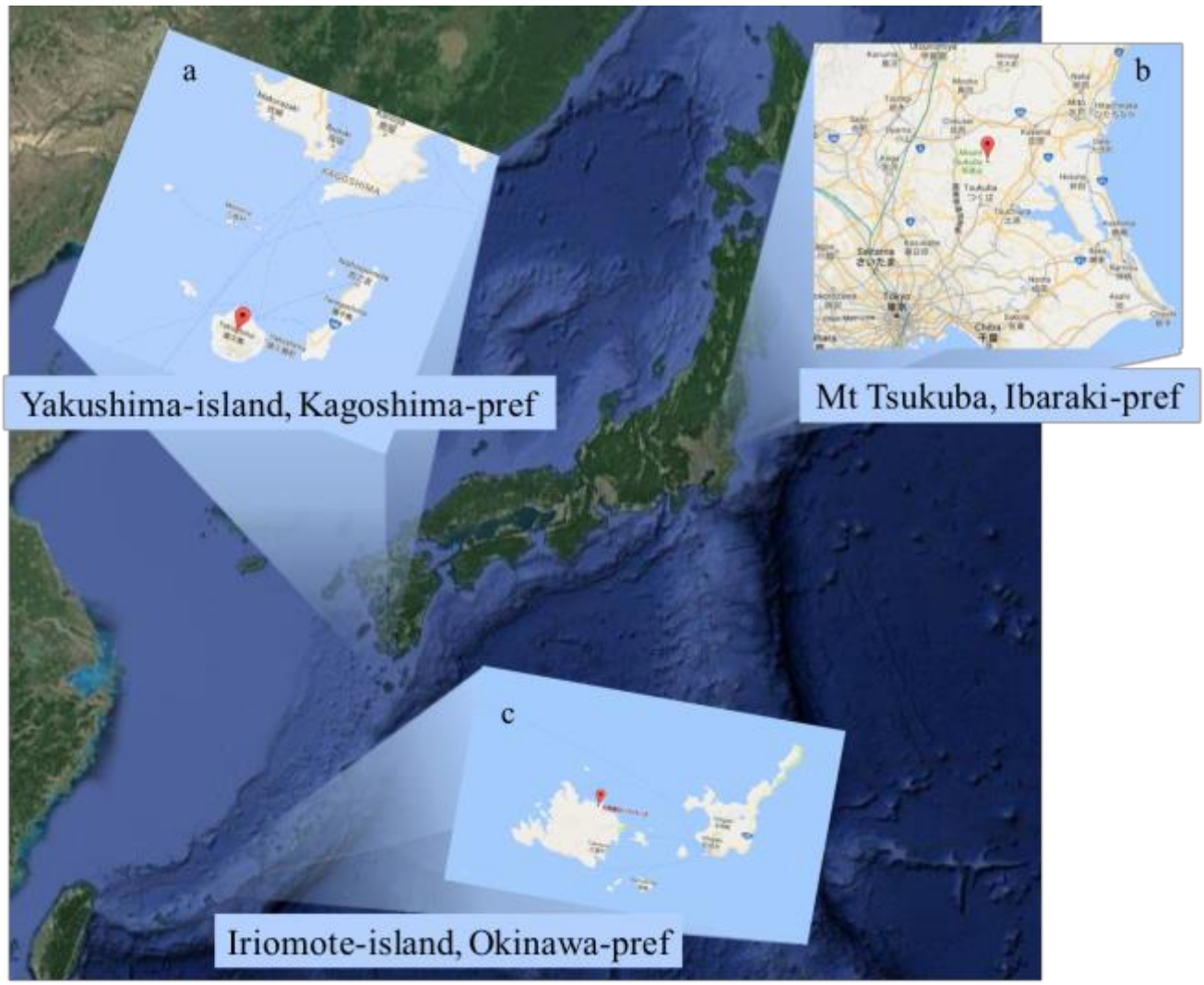

Figure 6. Sampling sites of termites around the Japanese Islands, a) *H. sjoestedti* sampled from the humid subtropical island Yakushima in Kagoshima prefecture. b) *R. speratus* sampled from Mount Tsukuba in Ibaraki prefecture on the main island of Honshu and c) *N. takasagoensis* sampled from the tropical rainforest island Iriomote in Okinawa prefecture.

determine possible contamination and the efficiency of the FACS in sorting individual bacteria. MDA genomic DNA was selected for NGS library preparation and sequencing. Sequencing genomic reads generated from the NGS platform are quality checked and assembled into genomes. Genome assemblies can then be annotated by database and bioinformatic software to determine gene function and overall bacteria metabolism with the presence of genes coding for glycosyl hydrolases and others involved in key metabolic pathways.

The first termite explored was the apical wood feeding higher termite *Nasutitermes takasagoensis* (タカサゴシロアリ) from the family Termitidae, sampled from Iriomote Island (Okinawa pref; Figure 6c). The functions and novelty in treponemes and two other key genera are explored from FACS isolated and genome sequenced samples (Chapters II & III). The second termite was from the basal lower termite lineage *Hodotermopsis sjoestedti* (オオシロアリ), from the family Hodotermitidae, sampled from the Japanese island Yakushima (Kagoshima pref; Figure 6a). *H. sjoestedti* gut contains the protist *Eucomonympha* sp. whose treponeme endosymbiont 'Candidatus *Treponema intracellularis*' was classified and whose contributions to the lower termite symbiosis were explored (Chapter IV). The last termite explored was the lower termite *Reticulitermes speratus* (ヤマトシロアリ), from the family Rhinotermitidae, sampled from mount Tsukuba (Ibaraki pref; Figure 6b). Treponemes were isolated and sequenced, as in previous chapters, but also key orthologous groups were used to compare with other treponemes in the genus, to describe the differences between those treponemes that evolved in the higher termite and those in the lower termite (Chapter V).

1.4 References

Albertsen M, Hugenholtz P, Skarshewski A, *et al.* (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**, 533-538.

Bignell DE, Oskarsson JM, Anderson JM, Ineson P (1983) Structure, microbial associations and function of the so-called "mixed segment" of the gut in two soil-feeding termites, *Procubitermes aburiensis* and *Cubitermes severus* (Termitidae, Termitinae). *J Zool Lond* **201**, 445-480

Blainey PC (2013) The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev* **37**, 407-427.

Bourguignon T, Lo N, Cameron S L, *et al*. (2015) The evolutionary history of termites as inferred from 66 mitochondrial genomes. *Mol Biol Evol* **32**, 406-421.

Brune A (2014) Symbiotic digestion of lignocellulose in termite guts. *Nature Rev Microbiol* **12**, 168-180

Brune A, Dietrich C (2015) The gut microbiota of termites: Digesting the diversity in the light of ecology and evolution. *Annu Rev Microbiol* **69**, 145-166.

Brune A, Kuhl M (1996) pH profiles of the extremely alkaline hindguts of soil-feeding termites (Isoptera: Termitidae) determined with microelectrodes. *J Insect Physiol* **42**, 1121-1127.

Cleveland LR (1923) Symbiosis between termites and their intestinal protozoa. *Proc Natl Acad Sci U S A* **9**, 424-428.

Clingenpeel S, Schwientek P, Hugenholtz P, Woyke T (2014) Effects of sample treatments on genome recovery via single-cell genomics. *ISME J* **8**, 2546-2549.

de Bourcy CF, De Vlaminck I, Kanbar JN*, et al.* (2014) A quantitative comparison of single-cell whole genome amplification methods. *PLoS One* **9**, e105585.

Dean FB, Hosono S, Fang L, et al. (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* **99**, 5261-5266.

Dichosa AE, Daughton AR, Reitenga KG, *et al*. (2014) Capturing and cultivating single bacterial cells in gel microdroplets to obtain near-complete genomes. *Nat Protoc* **9**, 608-621.

Diouf M, Roy V, Mora P*, et al.* (2015) Profiling the succession of bacterial communities throughout the life stages of a higher termite *Nasutitermes arborum* (Termitidae, Nasutitermitinae) using 16S rRNA gene pyrosequencing. *PLoS One* **10**, e0140014.

Dodsworth JA, Blainey PC, Murugapiran SK*, et al.* (2013) Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat Commun* **4**, 1854.

Donovan, S. E., P. Eggleton, and D. E. Bignell. 2001. Gut content analysis and a new feeding group classification of termites. *Ecol Entomol.* **26**: 356-366.

Dröge S, Rachel R, Radek R, König H (2008) *Treponema isoptericolens* sp. nov., a novel spirochaete from the hindgut of the termite *Incisitermes tabogae*. *Int J Syst Evol Microbiol* **58**, 1079-1083.

Engel MS, Barden P, Riccio ML, Grimaldi DA (2016) Morphologically specialized termite castes and advanced sociality in the Early Cretaceous. *Curr Biol* **26**, 522-530.

Engel P, Stepanauskas R, Moran NA (2014) Hidden diversity in honey bee gut symbionts detected by single-cell genomics. *PLoS Genet* **10**, e1004596.

Eutick ML, O'Brien RW, Slaytor M (1978) Bacteria from the gut of Australian termites. *Appl Environ Microbiol* **35**, 823-828.

Field EK, Sczyrba A, Lyman AE*, et al.* (2015) Genomic insights into the uncultivated marine Zetaproteobacteria at Loihi Seamount. *ISME J* **9**, 857-870.

Fullerton H, Moyer CL (2016) Comparative single-cell genomics of Chloroflexi from the Okinawa trough deep-subsurface biosphere. *Appl Environ Microbiol* **82**, 3000-3008.

Gole J, Gore A, Richards A*, et al.* (2013) Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat Biotechnol* **31**, 1126-1132.

Graber JR, Leadbetter JR, Breznak JA (2004) Description of *Treponema azotonutricium* sp nov and *Treponema primitia* sp nov., the first Spirochetes isolated from termite guts. *Appl Environ Microbiol* **70**, 1315-1320.

Hongoh Y, Deevong P, Hattori S*, et al.* (2006) Phylogenetic diversity, localization, and cell morphologies of members of the candidate phylum TG3 and a subphylum in the phylum Fibrobacteres, recently discovered bacterial groups dominant in termite guts. *Appl Environ Microbiol* **72**, 6780-6788.

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, Sato T, Dolan MF*, et al.* (2007a) The motility symbiont of the termite gut flagellate *Caduceia versatilis* is a member of the "Synergistes" group. *Appl Environ Microbiol* **73**, 6270-6276.

Hongoh Y, Sato T, Noda S*, et al.* (2007b) Candidatus *Symbiothrix dinenymphae*: bristle-like Bacteroidales ectosymbionts of termite gut protists. *Environ Microbiol* **9**, 2631-2635.

Hongoh Y, Sharma VK, Prakash T*, et al.* (2008a) Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Proc Natl Acad Sci U S A* **105**, 5555-5560.

Hongoh Y, Sharma VK, Prakash T*, et al.* (2008b) Genome of an endosymbiont coupling N2 fixation to cellulolysis within protist cells in termite gut. *Science* **322**.

Iida T, Ohkuma M, Ohtoko K, Kudo T (2000) Symbiotic spirochetes in the termite hindgut: phylogenetic identification of ectosymbiotic spirochetes of oxymonad protists. *FEMS Microbiol Ecol* **34**, 17-26.

Ikeda-Ohtsubo W, Strassert JF, Köhler T*, et al.* (2016) 'Candidatus Adiutrix intracellularis', an endosymbiont of termite gut flagellates, is the first representative of a deep-branching clade of Deltaproteobacteria and a putative homoacetogen. *Environ Microbiol* **18**, 2548-2564.

Inoue J, Noda S, Hongoh Y, Ui S, Ohkuma M (2008) Identification of endosymbiotic methanogen and ectosymbiotic spirochetes of gut protists of the termite *Coptotermes formosanus*. *Microbes Environ* **23**, 94-97.

Koidzumi, M. (1921) Studies on the intestinal protozoa found in the termites of Japan. *Parasitology*, **13**, 235-309.

Krishna K, Grimaldi D, Krishna V, Engel M (2013) Treatise on the Isoptera of the world. *Bull Am Mus Nat Hist* **377,** 2704-2904.

Leadbetter JR, Schmidt TM, Graber JR, Breznak JA (1999) Acetogenesis from H-2 plus CO2 by spirochetes from termite guts. *Science* **283**, 686-689.

Lo N, Tokuda G, Watanabe H (2011) Evolution and function of endogenous termite cellulases In: Bignell DE, Roisin Y, Lo N, (eds). Biology of Termites: A Modern Synthesis. Springer-Verlag: 51-67.

Lombard V, Golaconda Ramulu H, Drula E, *et al*. (2014) The Carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res*, **42**, 490–495.

Lynd LR, Wyman CE, Gerngross TU (1999) Biocommodity engineering. *Biotechnol Prog* **15**, 777-793.

Marshall IP, Blainey PC, Spormann AM, Quake SR (2012) A Single-cell genome for Thiovulum sp. *Appl Environ Microbiol* **78**, 8555-8563.

Martinez-Garcia M, Brazel D, Poulton NJ*, et al.* (2012a) Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *ISME J* **6**, 703-707.

Martinez-Garcia M, Brazel DM, Swan BK*, et al.* (2012b) Capturing single cell genomes of active polysaccharide degraders: an unexpected contribution of Verrucomicrobia. *PLoS One* **7**, e35314.

Mikaelyan A, Dietrich C, Köhler T, *et al.* (2015) Diet is the primary determinant of bacterial community structure in the guts of higher termites. *Mol Ecol* **24**, 5284-5295.

Miyata R, Noda N, Tamaki H*, et al.* (2007) Influence of feed components on symbiotic bacterial community structure in the gut of the wood-feeding higher termite *Nasutitermes takasagoensis*. *Biosci Biotechnol Biochem* **71**, 1244-1251.

Noda S, Kitade O, Inoue T*, et al.* (2007) Cospeciation in the triplex symbiosis of termite gut protists (*Pseudotrichonympha* spp.), their hosts, and their bacterial endosymbionts. *Mol Ecol* **16**, 1257-1266.

Odelson DA, Breznak JA (1983) Volatile Fatty Acid production by the hindgut microbiota of xylophagous termites. *Appl Environ Microbiol* **45**, 1602-1613.

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

Otani S, Mikaelyan A, Nobre T*, et al.* (2014) Identifying the core microbial community in the gut of fungus-growing termites. *Mol Ecol* **23**, 4631-4644.

Pohlschroeder M, Leschine S, Canaleparola E, (1994) *Spirochaeta calderia* Sp-nov, A thermophic bacterium that enhances cellulose degradation by *Clostridium thermocellum*. *Arch Microbiol*, **161**, 17-24.

Poulsen M, Hu H, Li C*, et al.* (2014) Complementary symbiont contributions to plant decomposition in a fungus-farming termite. *Proc Natl Acad Sci U S A* **111**, 14500-14505.

Raghunathan A, Jr, Ferguson HR, Bornarth J, *et al*. (2005) Genomic DNA amplification from a single bacterium genomic DNA amplification from a single bacterium. *Appl Environ Microbiol* **71**, 3342–3347

Rahman NA, Parks DH, Willner DL, *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.

Reid NM, Addison SL, West MA, Lloyd-Jones G (2014) The bacterial microbiota of *Stolotermes ruficeps* (Stolotermitidae), a phylogenetically basal termite endemic to New Zealand. *FEMS Microbiol Ecol* **90**, 678-88

Rinke C, Schwientek P, Sczyrba A*, et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431-437.

Rosenthal AZ, Matson EG, Eldar A, Leadbetter JR (2011) RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. *ISME J* **5**, 1133-1142.

Sato T, Kuwahara H, Fujita K*, 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.

Schaudinn F, Hoffman E (1905) Vorläufiger bericht über das Vorkommen für Spirochaeten in syphilitischen Krankheitsprodukten und bei Papillomen. *Arb Gesundh Amt Berlin* **22**, 528-534.

Scheller HV, Ulvskov P (2010) Hemicelluloses. *Annu Rev Plant Biol* **61**, 263–89.

Siegl A, Kamke J, Hochmuth T*, et al.* (2011) Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges. *ISME J* **5**, 61-70.

Stanton TB, Canale-Parola E (1980) *Treponema bryantii* sp. nov., a rumen spirochete that interacts with cellulolytic bacteria. *Arch Microbiol* **127**, 145-156.

Strassert JF, Mikaelyan A, Woyke T, Brune A (2016) Genome analysis of 'Candidatus *Ancillula trichonymphae*', first representative of a deep-branching clade of Bifidobacteriales, strengthens evidence for convergent evolution in flagellate endosymbionts. *Environ Microbiol Rep* **8**, 865–873.

Timell T (1967) Recent progress in the chemistry of wood hemicelluloses. *Wood Sci Technol* **1**, 45-70.

Warnecke F, Luginbuhl P, Ivanova N*, et al.* (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**, 560-U517.

Wilson MC, Mori T, Rückert C*, et al.* (2014) An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58-62.

Yuki M, Kuwahara H, Shintani M*, et al.* (2015) Dominant ectosymbiotic bacteria of cellulolytic protists in the termite gut also have the potential to digest lignocellulose. *Environ Microbiol* **17**, 4942-4953

Zhang Y, Sun Y, Jiao N, *et al*. (2016) Ecological genomics of the uncultivated marine Roseobacter lineage CHAB-I-5. *Appl Environ Microbiol* **82**, 2100-2111.

Zong C, Lu S, Chapman AR, Xie XS (2012) Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622-1626.

Chapter II

Single cell genomic investigations into the diversity of treponemes isolated from the higher termite gut

My roles in this study were in designing the experiment, manuscript writing, in the assembly, annotation, and genome analysis of single cells and in performing the RNA-seq and subsequent bioinformatic analyses.

Single cell genomic investigations into the diversity of treponemes isolated from the higher termite gut

David Starns[1,2], Masahiro Yuki[3], Yuichi Hongoh[4], Alistair Darby[2], Moriya Ohkuma[1,3],

[1]Japan Collection of Microorganisms, RIKEN BioResource Center, Tsukuba, Japan
[2]Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom
[3]Biomass Research Platform Team, RIKEN Biomass Engineering Program Cooperation Division, RIKEN Center for Sustainable Resource Science, Tsukuba, Japan
[4]Department of Life Science and Technology, Tokyo Institute of Technology, Tokyo, Japan
*** For correspondence:**
Moriya Ohkuma
Japan Collection of Microorganisms, RIKEN BioResource Center, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan
Email: mohkuma@riken.jp
Tel. (+81) 29 829 9101; Fax (+81) 29 829 9102

Working Title : Treponemes of Nasutitermes show niche speciation
Keywords ; Termite, Symbiosis, Treponema, niche-partitioning

Abstract

Wood-feeding higher termites are supported by a diverse and populous microbiota that efficiently sustains the host on a recalcitrant wood diet. One abundant group of bacteria from this environment, providing this essential role are of the genus *Treponema*. These are morphologically distinct spiral bacteria that navigate the anaerobic viscous core of the hindgut in termites. Few *Treponema* strains have been cultured, and only those isolated from lower termites (basal lineages, cohabiting the gut with protists) have been studied in depth.

Here the functions of individual members are explained within this diverse bacterial group using the culture independent technique of single cell genomics, through the physical isolation of individual cells by fluorescence activated cell sorting on the gut contents of the tropical arboreal termite *Nasutitermes takasagoensis*. Individual isolated cells from four sub clusters of Treponema cluster I based on 16S rRNA gene sequences were genome sequenced and the roles in nitrogen sequestration and lignocellulose digestion inferred.

The assembled single cell genomes contain various glycosyl hydrolases involved in lignocellulose digestion, iron hydrogenases in the reversible conversion of H^+ to H_2 and genes involved in reductive acetogenesis and nitrogen utilisation. We also show, through the use of RNA-seq, highly transcribed genes localised within this diverse group. The communities of *Treponema* are inferred to be partitioned based on the utilisation of different lignocellulosic derived nutritional components ensuring a conserved yet diverse whole microbial community sustaining the host termite in this essential symbiotic relationship.

Introduction

The microorganisms in the gut of wood feeding higher termites are essential for their hosts survival (Eutick *et al*., 1978), these bacterial communities are able to utilise recalcitrant wood and supply their host with the energy source acetate (Tokuda *et al*., 2007; Warnecke *et al*., 2007; He *et al*., 2013; Mikaelyan *et al*., 2014; Poulsen *et al*., 2014). This symbiotic system is unlike wood feeding lower termites where the majority of this role is filled by Protists, who themselves can harbour ecto- and endosymbionts (Iida *et al*., 2000; Ohkuma *et al.,* 2015). In higher termites the bacterial diversity has been characterised in previous studies (Miyata et al., 2007; Warnecke *et al*., 2007; Kohler *et al*., 2012) but members of one particular genus ultimately dominates the guts of wood eating species – *Treponema*. This has also been demonstrated when changing the termite's food source between simple wood components (Miyata *et al*., 2007).

The functions of the diverse *Treponema* in higher termites are speculative, due to a lack of cultured representatives, however two species have been cultured and isolated from the lower termite *Zootermopsis angusticollis*, *Treponema primitia* and *Treponema azotonutricium* (Graber *et al*., 2004). *T. primitia* is a homoacetogen, producing acetate by chemoautotrophic growth, which is in turn used by the host termite, whereas *T. azotonutricium* is diazotrophic; these two exhibit two distinct phenotypes and co-culturing has shown to potentiate their growth (Rosenthal *et al*., 2011). The metagenome, metatranscriptome and metaproteome of the gut contents of a wood feeding termite *Nasutitermes sp.* were published in 2007, and here the authors classified the overall functions of the members of the gut community however could not allocate them to specific representatives in this gut. Furthermore, the absence of formate dehydrogenase, a key enzyme in homo-acetogenesis, suggests alternative energy cycling within treponemes. The metagenome placed the treponeme community into ten different groups (termite treponeme groups) with 103 phylotypes observed in *Nasutitermes* sp. (Warnecke *et al*., 2007). Recently the phylogeny of *Treponema* from most termite lineages has become more refined (Mikaelyan *et al*., 2015) and in the higher termites wood associated genus-level clusters have been categorized into this phylogeny in *Treponema* cluster I, with two of these subclusters, Ic and If associated with (hemi)cellulose degradation activities (Mikaelyan *et al*., 2014).

Studies on rumen *Treponema* such as the non-cellulolytic *Treponema bryantii* have shown that interaction with cellulolytic bacteria (*Fibrobacter succinogenes* and *Ruminococcus albus*) promotes cellulolysis (Kudo *et al*., 1987). *Treponema caldarium*, a free-living treponeme

isolated from cyanobacteria mats has also been shown to potentiate cellulose degradation in co-cultures with *Clostridium thermocellum* (Pohlschroeder *et al*., 1993). It is probable that similar mechanisms have evolved in the termite gut symbioses of higher termites concerning similar class candidates - Fibrobacter subphylum 2, candidate phylum TG3 and cellulolytic firmicutes (Hongoh *et al*., 2006) that make up some individuals in the rest of the community.

So far the inability to culture these individuals has hindered our understanding of the community and thus culture independent methods have become the gold standard in exploring hidden diversity. Metagenomics provides insight into how whole communities potential function but using single cell genomics we can categorize functions to niche specific taxa (Marshall *et al*., 2012, Dodsworth *et al*., 2013, Rinke *et al*., 2013) within these communities of non-culture viable organisms. Functions are categorized by genome encoding enzymes such as glycosyl hydrolases in lignocellulose degradation or in nitrogen metabolism.

In this study we classify the putative functions of this diverse group of treponemes using single cell genomics to understand the individual functions these organisms contribute to this nutritional symbiosis. Functions are classified by analysing genes encoding glycosyl hydrolases that degrade components of lignocellulose and genes encoding nitrogen metabolism associated proteins. We couple the single cell data with metatranscriptomic data to understand the contribution of each genome in which genes are expressed from these individuals within the whole community.

Materials and methods

Sampling

The subtropical arboreal termite *Nasutitermes takasagoensis* was collected from three *Castanopsis cuspidata* trunk cuttings from a coastal forest termite colony on Iriomote island, Okinawa prefecture, Japan (coordinates 24°23'40.2"N 123°51'38.0"E). Termites were collected from this area in 2013 for the single cell genome analysis and again in 2015 for the RNA-seq analysis.

Single cell isolation, genome amplification and identification

Nasutitermes termite worker guts were prepared as previously described (Yuki *et al*., 2015). The gut was removed using sterile forceps and content expelled into isotonic Solution U (Trager, 1934), gut tissue was removed or homogenised. The guts of ten workers were pooled and subsequently washed, filtered, and stained with Cell Tracker CMFDA. A MoFlo XDP (Beckman Coulter) cell sorter was used to isolate individual bacteria based on fluorescence using previous optimised conditions (Shintani *et al*., 2014).

Cell lysis and multiple displacement amplification was conducted as previously described (Yuki *et al*., 2015) using Qiagen REPLI-g UltraFast mini kit and samples were identified using 16S rRNA gene sequence PCR as previously described (Yuki *et al*., 2015). 16S rRNA gene sequences were sanger sequenced by RIKEN's Bio-materials analysis unit and those with a phred score of 40 or above were utilised further downstream. Sequences were taxonomically identified with BLAST searches against the NCBI-non redundant database.

Treponema genome analysis

DNA libraries were constructed using the Illumina Nextera XT and Tru-Seq library protocols and sequenced on the MiSeq platform. Reads generated were initially checked with FASTQC (Andrews, 2010) and subsequently quality and adapter trimmed using Sickle and Cutadapt (Joshi and Fass, 2011; Martin, 2011). Quality trimmed reads were assembled using SPAdes 3.1 assembler (Bankevich *et al.,* 2012) and assembly quality inferred using QUAST (Gurevich *et al*., 2013). Contamination was screened using the Blobology bash (Kumar *et al*., 2013).

The completeness of the *Treponema* genomes was assessed using 139 established marker genes (Rinke *et al*., 2013) and querying the genomes using HMMER3 against a hidden markov model (HMM) database. To increase the overall completeness of phylogenetic cluster reference

samples, seven pairs of sample read files were combined in assembly to create hybrid assemblies; these were based on 16S gene sequence identity, orthoMCL (Li *et al*., 2003) ortholog analysis and nucmer whole genome alignments (Delcher *et al*., 2002).

Assembled contigs were annotated using Prokka version 1.6.2 (Seemann, 2014) and generated genbank files uploaded to the RAST server (Aziz *et al*., 2008) for manual screening of genes of interest. RAST integrated SEED and KEGG maps were used to assign putative functions. OrthoMCL (Li *et al.,* 2003) was used to cluster orthologs for comparative analyses of the genomes. Diamond (Buchfink *et al*., 2015) was used to query a Carbohydrate-Active enZYmes (CAZy) database for glycosyl hydrolase analysis and Interproscan V (Jones *et al*., 2014) was used for overviews of protein domain structures. Non-supervised orthologous groups were determined using Diamond searches against the eggNOG database (Powell *et al*., 2014).

Phylogenetic analysis

Nucleotide and amino acid alignments were constructed using the MUSCLE algorithm and substitution model discerned using Jmodeltest and Prottest (Darriba *et al*., 2011, 2012) modelling programs. Subsequent best models were used in final tree construction using Maximum Likelihood method in Fasttree (version 2.1.8; Price *et al*., 2010), MEGA (version 7; Tamura *et al*., 2011) and RaxML (version 8.0.23; Stamatakis, 2014) tree packages. Phylosift (Darling *et al*., 2014) was used to extract marker ribosomal proteins, that were then concatenated and uploaded into phylogenetic software. Samples were assigned to termite *Treponema* subclusters using classify.seqs in Mothur (Schloss *et al*., 2009) against the DictDb (vers. 3.4; Mikaeylan *et al*., 2015) reference database and Termite *Treponema* groups using additional sequences (Warnecke *et al*., 2007) with 16S phylogenetic tree comparisons.

RNA-seq analysis.

Whole guts of ten *Nasutitermes takasagoensis* workers were exuded into RNA later and stored for 2 days at room temperature from near the site of sampling. Samples were washed and total RNA was extracted with Qiagen's RNAeasy kit and quantity and RIN quality checked using a Qubit and Agilent Bioanalyser 2100 respectively. Ribosomal RNA was depleted using RiboZero gold epidemiology kit and the final library was constructed using ScriptSeq Complete gold (Illumina). The library was sequenced on the Illumina MiSeq platform using V3 chemistry and Tru-Seq HT protocol. The generated reads were filtered with both Sickle and Cutadapt. Quality filtered reads underwent two analyses, the first was assembling the whole

metatranscriptome using Trinity (Grabherr *et al*., 2011) and the second was the mapping of the reads to the single cell genomes. The Trinity derived assembly was uploaded to MG-RAST for annotation. Bowtie-2 (Langmead and Salzburg, 2012) was used to map the reads to the single cell genomes and HTSeq (Anders *et al*., 2015) was used to count the number of reads mapped to sample genes. Reads per kilobase of transcript per million mapped reads (RPKM) were calculated to generate normalised counts for the transcriptome data.

Results and Discussion

Single cell sorting

The *Nasutitermes* worker gut sample was sorted over nine 96 well plates at a total of 864 wells, a putative 864 single cells. Of these, 216 generated 16S rRNA DNA amplicons and passed the phred score threshold of 40 for sequence quality. 148 of these cells were identified from the *Treponema* genus, other bacteria isolated were from classes Bacteroidales, Clostridia, Deltaproteobacteria, TG3 and Fibrobacteres; the latter two were used for a further study (Yuki *et al*., in prep).

From the 148 *Treponema* isolated cells, the 16S rRNA DNA sequences were used to generate a phylogenetic tree, 30 samples were initially chosen for sequencing that covered a good range of diversity, and those that could be paired with others for hybrid assembly due to the fragmental nature of single cell genomics, 95 cells were identified as sub cluster Ic, seven were sub cluster 1a, 35 were sub cluster If and 11 were identified as sub cluster Ih.

Phylogeny

The sequenced *Treponema* assemblies span four sub clusters of the established Termite *Treponema* Cluster I (Ohkuma *et al*., 1999; Mikaelyan *et al*., 2015) and nine of the ten termite *Treponema* groups (Warnecke *et al*., 2007). FACS sorting verified the most abundant group detected within *Nasutitermes* as sub cluster Ic that includes termite *Treponema* groups 1-6 and the most abundant group seen in the fibre fraction of the higher termite gut community (Mikaelyan *et al*., 2014). We discern that *Treponema* groups 5 and 6 (samples R83H11 and R80F12) are a separate clade from sub cluster Ic and If using the concatenated ribosomal protein tree (Figure 1b). Phylogenetic analysis of the single cell genomes using whole genomic 16S rRNA gene sequences clustered the *Treponema* similarly to those in the metagenome study with strong bootstrap support (Figure 1a). A 38 concatenated ribosomal protein maximum likelihood tree was generated in this study that better clarifies the positions of these taxa and ingraining the positions of other classical treponemes (Figure 1b).

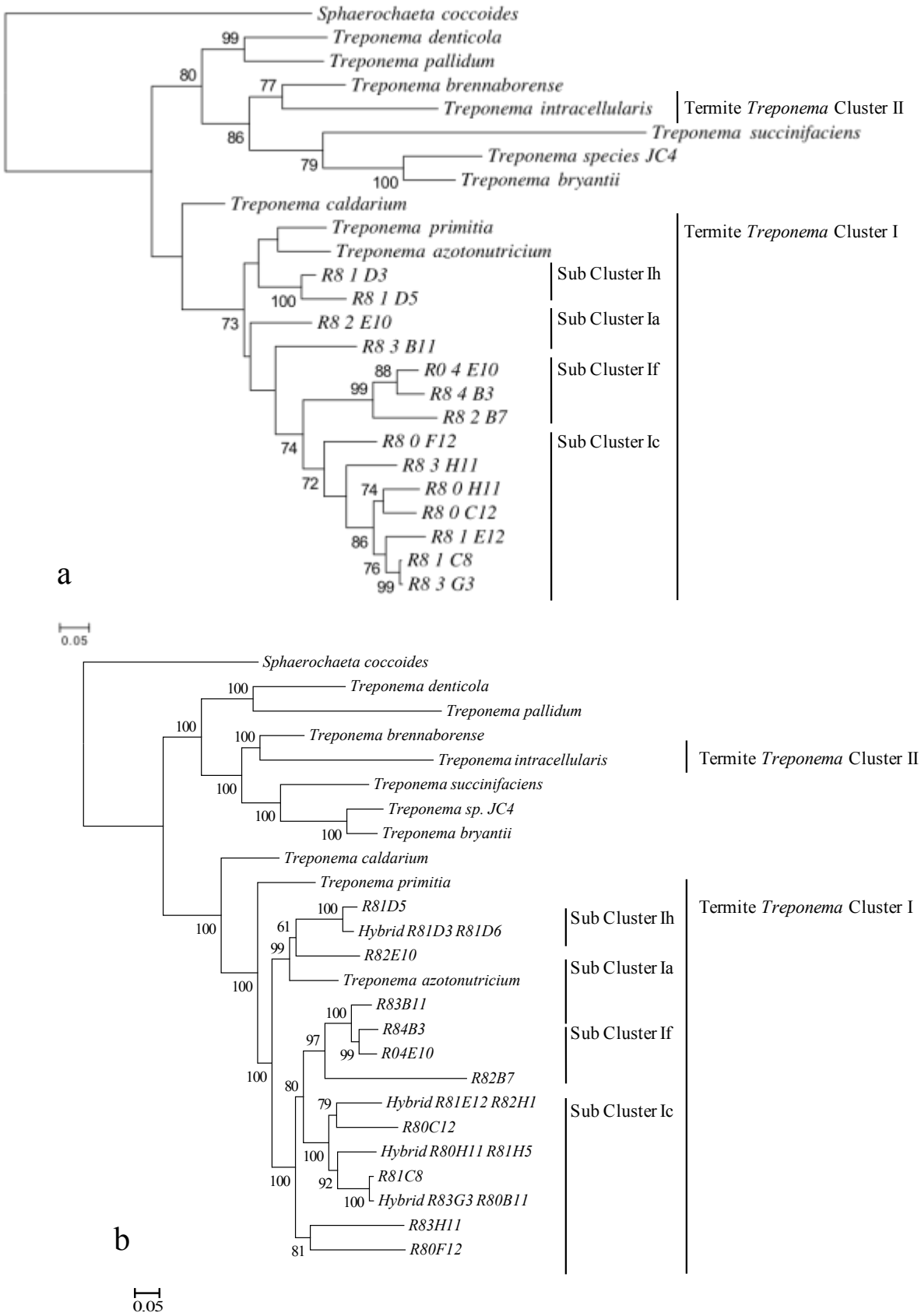

Figure 1 Phylogenetic positions of treponeme samples showing treponeme clusters and sub clusters.
1a. 16S rRNA gene sequence phylogenetic tree showing positions of samples isolated from *N. takasagoensis*. The tree was created using maximum likelihood method using the generalised time reversible nucleic acid substitution model with 1000 bootstrap replicates. Samples were selected to avoid redundancy, and generate the strongest bootstrap support. **1b.** Concatenated 38 ribosomal protein phylogenetic positions of sub clusters from *N. takasagoensis* with relevant attributes for termite symbiosis. The phylogenetic tree was made using maximum likelihood method using the Whelan and Goldman amino acid substitution model with 1000 bootstrap replicates.

Genome summary

30 different single cell treponeme assemblies were obtained, which ranged in completeness from 83.5% to 16% (Table 1. average 55%; Rinke *et al*., 2013). Hybrid assembly of pairs of individual isolates were shown to improve genomic completeness and contiguity overall (Table 2), hybrid assembly has been used in previous studies (Marshall *et al*., 2012) to create completer reference genomes. The greatest sample completeness of representatives from the different phylogenetic sub clusters of genomes were Ia 76%, Ic 92%, If 59% and Ih 65%. The most near complete genomes were from sub cluster Ic with multiple samples reaching over 70% completeness. In proportion to this the *Treponema* genomes are reported to be around 4.23Mb (using hybrid), comparable with *T. primitia* and *T. azotonutricium* (3.8-4Mb). GC content of the individual assemblies ranges from 40 to 50% and follows a similarly profile to the phylogenetic clustering. At least one phylotype affiliated with 9 out of the 10 groups illustrated in *Nasutitermes spp.* (Warnecke *et al*., 2007) was sequenced.

RNA-seq data also provided evidence of the more abundant species in the greater number of reads aligned to the genomes of those in subcluster Ic, The genomes with the highest percentage of mapped reads were R80H11(6%) and R81H5(8.5%) here, both samples sharing the same 16S rRNA gene sequences (Table 1). The total alignment rate for the entirety of the treponeme single cell genomic data was 23% of the RNA-seq library.

The termite *Treponema* sub cluster Ic are the most abundant *Treponema* found in the guts of the genus *Nasutitermes*, and covers groups 1-6 defined in the metagenome (Warnecke *et al*., 2007). These are the first genomes sequenced for this representative cluster. From this sub cluster, we assembled three phylotypes using hybrid assembly (Table 2), within 85-100% completeness, these belong to termite treponeme groups 1 and 2. This cluster is reported to be loosely associated with wood particles in the gut lumen and associated with higher cellulase activity (Mikaelyan *et al*., 2014). Sub cluster If is the second most abundant cluster within the genus *Nasutitermes*, encompassing termite *Treponema* group 7 in the 2007 metagenomic study. Two genomes from sub cluster Ia (TTG-8 and 10) were sequenced, this sub cluster also contains *T. primitia* and *T. azotonuticium*, is the most widely abundant sub cluster found in lower termite lineages, (Dietrich *et al*., 2014). Three samples from sub cluster Ih were sequenced, one single sample assembly (65% completeness) and two used in a hybrid assembly (37%).

Table 1. Genome statistics of single cell assemblies, including comparative genomes from *T. caldarium*, *T. primitia* and *T. azotonutricium*.

Single cell genome	Chromosome Size (Gb)	# Contigs	Longest Contig (bp)	N50 (bp)	GC content (%)	tRNA	CDS	% Completeness	SubCluster /TTG	RNA-Seq % reads mapped
R80H11	2.6	608	120,714	18,605	45.21	39	2,421	69.78	C 1	6.05
R81H5	3.9	1,281	126,125	17,744	44.29	59	3,483	80.58	C 1	8.55
R81F1	3.1	1,296	61,466	8,397	48	32	3,208	58.99	C 1	2.95
R83B6	1.8	895	47,995	6,126	45.3	33	1,701	64.03	C 1	4.36
R81C6	1	475	57,914	12,383	45.55	7	904	16.55	C 1	2.17
R6D11	2.9	1,064	83,207	10,272	42.55	54	2,818	78.42	C 2	5.71
R80C12	2.1	966	52,863	11,686	40.5	24	1,978	64.03	C 3	1.52
R81E12	2.2	504	102,135	12,808	42.38	32	2,122	70.50	C 1/2	0.98
R82H1	2.6	612	129,404	12,734	42.6	41	2,572	64.75	C 1/2	1.12
R81C8	1.8	649	121,251	10,021	43.71	19	1,730	61.15	C 2	2.49
R80B11	2.2	761	64,272	12,019	42.54	29	2,088	63.31	C 2	3.04
R83G3	2.8	781	86,362	11,660	42.61	33	2,520	77.70	C 2	4.07
R83G4	2.1	1,260	41,148	4,559	48.84	23	2,334	27.34	C 5	0.77
R83H11	3.6	1,312	49,096	10,248	47.1	34	3,995	57.55	C 5	1.12
R80F12	2.6	635	121,314	21,345	48.89	28	2,441	83.45	C 6	0.96
R84B3	1.6	465	72,062	10,842	48.61	17	1,576	59.71	F 7	2.26
R80B2	2.4	923	54,089	8,713	48.8	19	2,329	59.71	F 7	2.21
R5G12	4	2,031	58,389	5,263	47.2	39	4,342	56.83	F 7	2.28
R6C12	4.5	2,178	45,263	4,720	46.9	41	4,887	58.99	F 7	2.54
R04E10	2.6	1,232	58,863	6,047	48.25	20	2,665	59.71	F 7	1.84
R84B11	1.8	565	69,924	18,013	48	27	1,724	43.88	F 7	1.37
R82B7	2	867	52,723	9,600	49.99	25	1,873	35.97	F 7	0.67
R83B11	4.8	1,474	85,253	19,622	47.4	52	4,657	76.26	A 8	2.71
R82E10	1.8	694	93,524	12,159	46.69	26	1,770	51.80	A 10	0.4
R81D3	1.4	518	77,390	13,596	46.36	15	1,358	20.86	H 9	0.45
R81D5	3.1	885	73,142	19,598	48.58	22	3,022	65.47	H 9	0.82
R81C1	1.3	474	53,386	13,437	47.62	11	1,116	45.32	F 7	0.79
R83A5	2.8	1,079	61,764	11,288	47.37	27	2,739	49.64	F 7	2.08
R83G8	2.9	1,262	66,614	8,002	46.93	34	2,977	63.31	F 7	1.71
R84F11	1.5	706	40,038	8,301	49.4	16	1,525	48.92	F 7	1.25
T. azotonutricium	3.8	1	3,855,671	N/A	49.8	48	3,403	97.12	A N/A	N/A
T. primitia	4	1	4,059,867	N/A	50.8	53	3,499	96.40	A N/A	N/A
T. caldarium	3.2	1	3,239,340	N/A	45.6	51	2,873	96.40	N/A N/A	N/A

Table 2. Genome statistics of hybrid single cell assemblies.

Sample 1	Sample2	Chromosome Size (bp)	# Contigs	Longest Contig (bp)	N50 (bp)	GC content (%)	16S	tRNA	CDS	Completeness (%)	SubCluster /TTG	RNA-Seq reads mapped (%)
R80B11	R83G3	3,351,006	612	166,562	39,804	42.4	1	38	3020	92.09	Ic 2	4.63
R82H1	R81E12	2,993,691	413	145,053	34,639	42.3	1	44	2890	89.21	Ic 2	1.24
R80H11	R81H5	3,682,397	808	116,114	20,220	44.89	1	58	3312	87.05	Ic 1	7.70
R81F1	R83B6	2,395,569	843	88,402	10,492	45.4	1	40	2,253	70.50	Ic 1	5.36
R6C12	R5G12	5,802,316	2,069	61,092	7405	47	1	51	6,189	71.22	If 7	2.93
R83E8	R84B11	2,823,583	183	173,481	40858	47.8	1	41	2585	64.75	If 7	1.81
R81D3	R81D6	2,020,811	558	81,129	18,322	46.58	1	21	1971	37.41	Ih 9	0.52

Lignocellulose digestion.

Different genes coding for potential lignocellulose degrading glycosyl hydrolases across the treponeme sample groups were analysed to better understand their roles in this nutritional symbiosis. The termite hosts diet of hardwood trees, where the composition of hemicellulose is primarily xylans and substituted derivatives (Scheller and Ulvskov, 2010), provides the generation of niche resource partitioning. All samples varied with the content of GH families with figure 2 providing a summary outline of the proportions of glycosyl hydrolase families associated with the breakdown of cellulose and hemicellulose components between the sub clusters. Cellulases include endoglucanases and β-glucosidases, and hemicellulases include xylanases, xylosidases, and arabinofuranosidases.

GH10 was the most abundant GH family present in the treponeme single cell genomes. The most commonly known function of this group is the cleavage of endo-β-xylan glycosidic bonds. Families 3, 5 and 43 were the second most abundant glycosyl hydrolases within these genomes. These encode enzymes for β-D-glucosidases, α-L-arabinofuranosidases, cellulases and xylanases. There were five different clusters of GH10's present in the samples; the diversity was not group specific and multiple samples encoded enzymes that spanned the GH10 diversity. One group was annotated as secreted thermostable xylanases able to break down amorphous cellulose and xylan by endohydrolysis of 1,4-β-glycosidic bonds, this group included one protein with a galactose binding module. Sub cluster Ic showed the greatest diversity of these enzymes. The highest expression levels of GH10 enzymes were from TTG2 (Ic). Three GH10's share homology with similar enzymes isolated from a bovine rumen treponeme (SAG7).

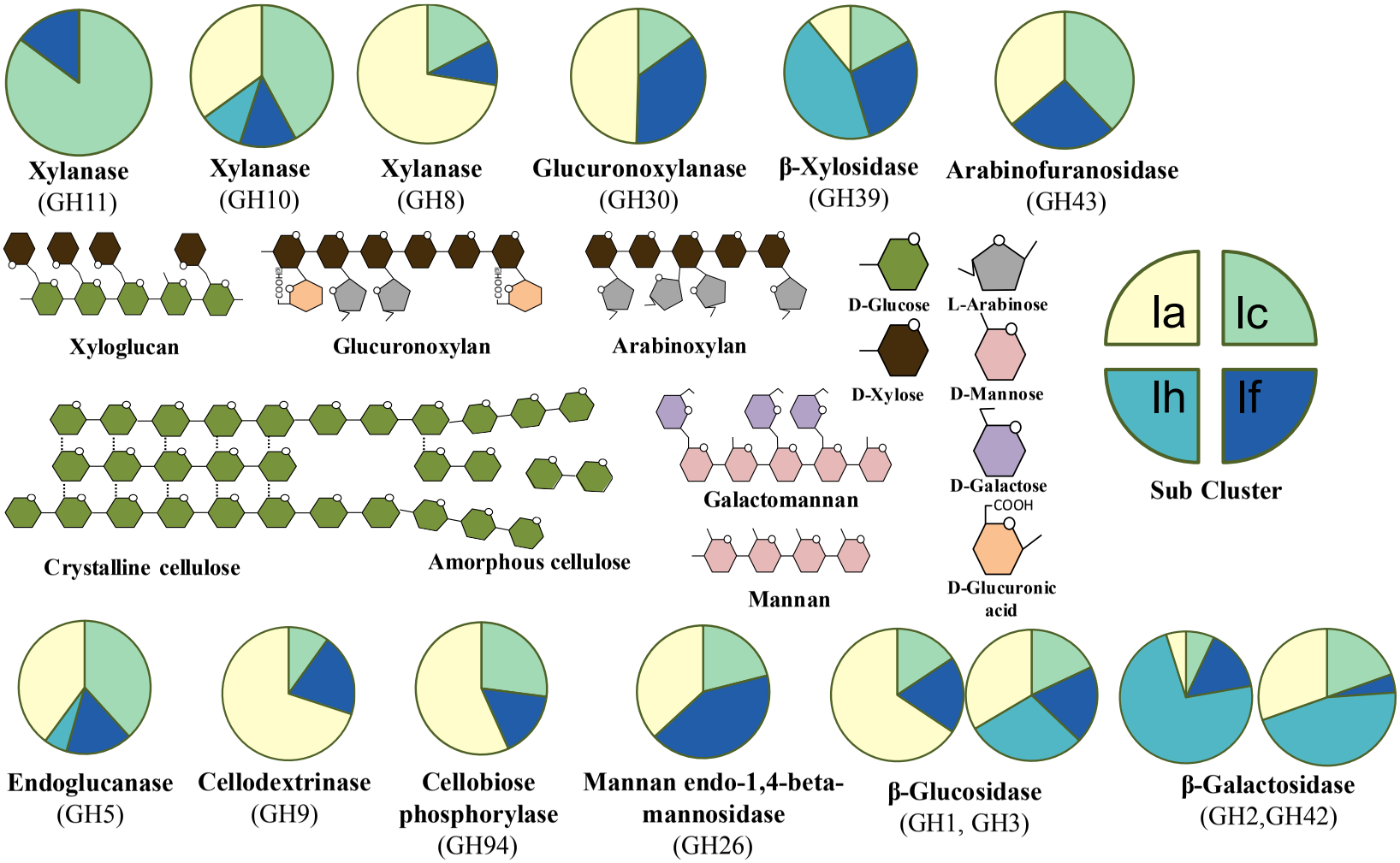

Figure 2. Proportions of glycosyl hydrolases allocated to sub clusters of treponeme samples. Proportions are representative of number of GH enzymes per genome as an average among sub cluster samples. Sub cluster Ia is indicated by cream, Ic by green, Ih by turquoise and If by blue. Chemical structures of cellulose (green), hemicellulose (xyloglucan, glucuronoxylan, arabinoxylan, galactomannan and mannan) and individual monosaccharide components shown.

Two genes coding for GH11 enzymes featured bacterial immunoglobulin domains and signal peptides near their N termini denoting secretion, both originating from sub cluster Ic. In the sample R80H11, this enzyme is flanked by very highly transcribed sugar binding and transporting proteins putatively transporting large polysaccharides and sharing greatest homology with that of an equivalent transport protein in *T. caldarium* (Figure 3). One sample from sub cluster If contained an enzyme from the GH11 family, however all the rest were annotated from sub cluster Ic genomes, where they formed two groups and both appear to be non-cytoplasmic secreted proteins.

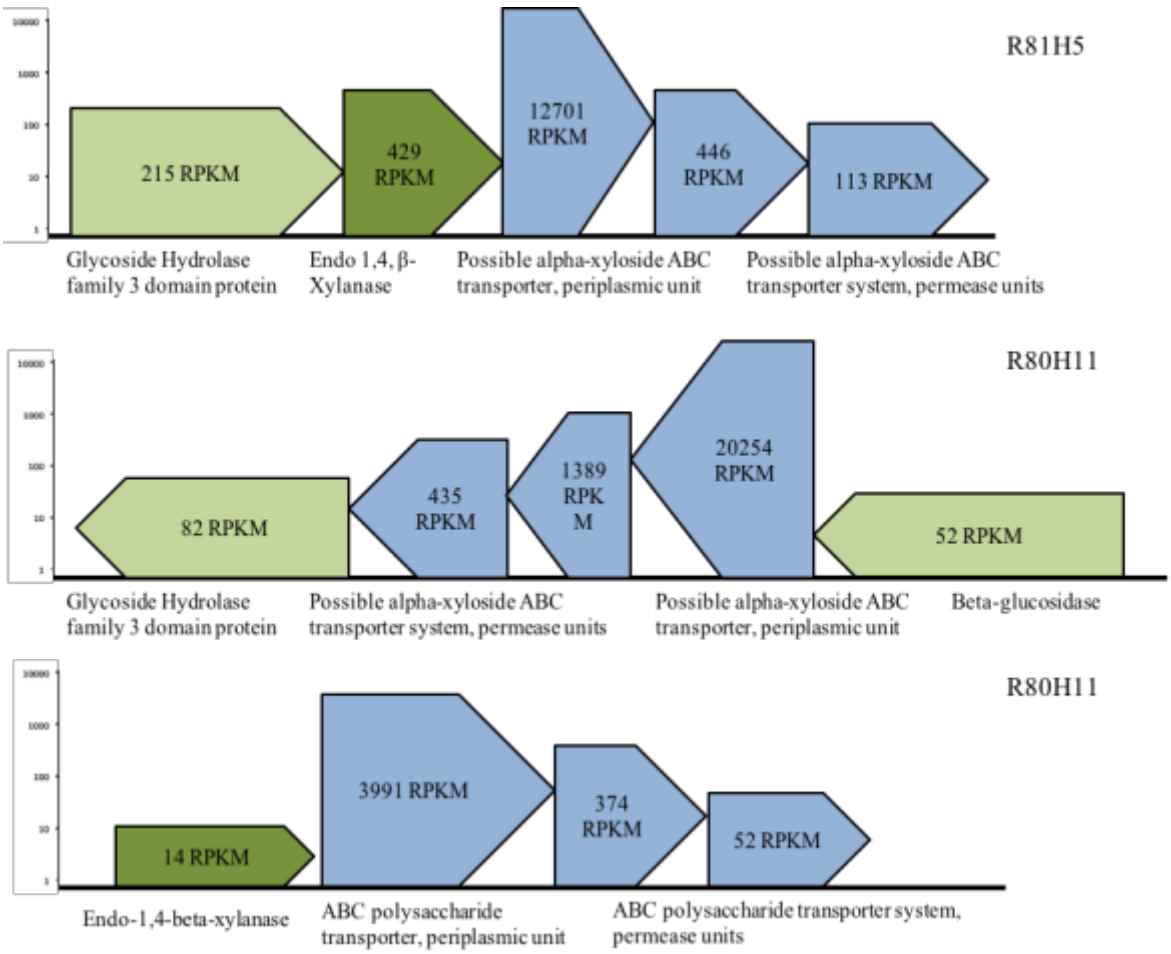

Figure 3. Expression of Glycosyl hydrolase transporters and operons from Sub Cluster Ic genomes. The scale is as log RPKM. Blue indicates the ABC transporter modules and light and dark green indicate adjacent glycosyl hydrolase enzymes. R81H5's endo-1,4-beta-xylanase (GH5) contains signal peptides and shares homology to the crystallised *Bacillus agaradherans* enzyme, whereas the GH3 shares homology with a *T. azotonutricium* enzyme that is not putatively excreted. RPKM reads per kilobase of transcript per million mapped reads

Glycosyl hydrolases of family 5 encode cellulolytic endoglucanase enzymes; nine different clusters were assigned to encode GH 5 family enzymes. These were analysed to check their sub

cluster positions within the GH5 sub families (Figure 4; Aspeborg *et al*., 2012). Here they were allocated to eight different subfamilies (2, 4, 12, 36, 39, 44, 47, 52). The majority of sequences were allocated to sub family 4 that encode xyloglucan specific endo-glucanases and endo 1,4 – β- xylanases. Here they clustered into three groups, with greatest homology to cellulolytic Bacteroidetes and Ruminococci, some possessing carbohydrate binding modules of family 2 known to bind crystalline cellulose (Gilbert *et al*., 2013). Samples clustering with Bacteroidetes also exhibited interesting ABC transporting genes at high abundancies flanking them but sharing homology with ruminal clostridial genes. Sub family 39, was only comprised of

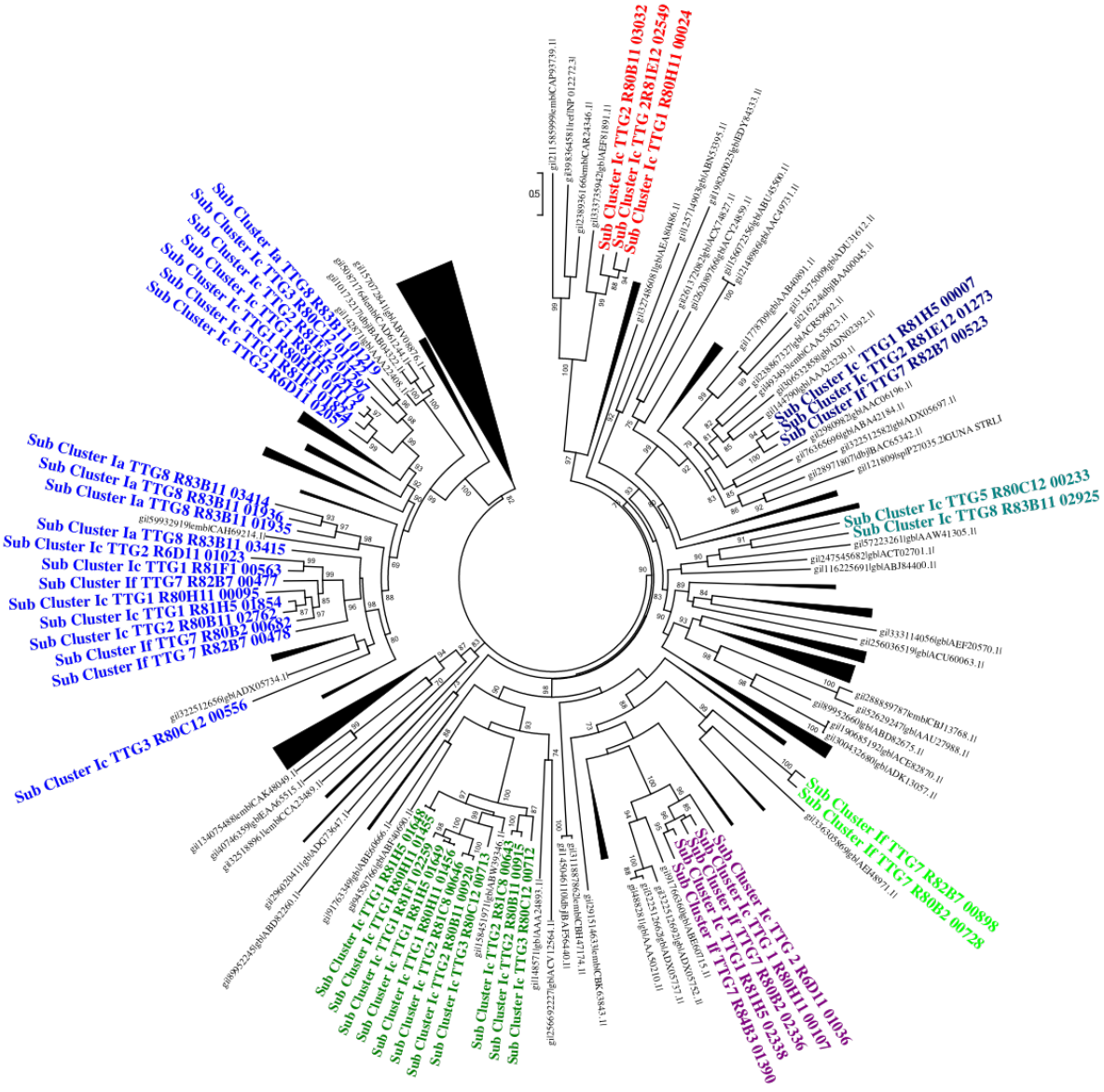

Figure 4. Glycosyl Hydrolase Family 5 phylogenetic tree unrooted, diversity of Treponema GH5 sub family enzymes, (Aspeborg *et al*., 2012), largest sub family 4 in blue, sub family 39 in green, sub family 52 shown in purple, sub family 36 in lime green, sub family 47 in teal, sub family 2 in navy and finally sub family 12 in red. Sub family 44 not shown. The tree was completed using 462 GH5 sequences, Whelan and Goldman (WAG) amino acid substitution model, supported with 500 bootstrap replicates.

subcluster Ic genomes and share homology with other classical treponemes and those extracted from the metagenome (Warnecke *et al*., 2007), this sub family have been shown to only have Endo-β-1,4- glucanase activity (Aspeborg *et al*., 2012). Some sub family 52 encoded enzymes were annotated as cellodextrinases that have been shown to degrade cellobiose oligosaccharides by the endohydrolysis of 1,4 –β- glycosidic bonds. Here, these show greatest homology to *Fibrobacter succinogenes* and clostridial sequences. Two sub cluster If samples encode a GH5 with most similarity in length to that of *Rhodopirellula* species. Three samples encoded subfamily 2 and are reported to be xylanases and endoglucanases (Aspeborg *et al*., 2012).

Glycosyl hydrolase family 43 split into three groups; the least divergent group among the samples was a conserved alpha-N-arabinofuranosidase, this was closely homologous to arabinofuranosidases, almost exclusively from the phylum Bacteroidetes, at ~70% sequence similarity, that degrades hemicellulose containing arabinoxylan, and arabinogalactan residues, suggesting that it may have originated as a horizontal gene transfer. The largest group split into two clusters with those containing the GH43 domain and the plant cell wall binding moiety CBM6 and a signal peptide and then those containing the GH43 domain with a concanavalin(lectin)/glucanase domain.

GH family 8 formed three groups, one sample in TTG1 contained examples of all three groups with one gene exhibiting high expression levels, this enzyme shares closest homology to a glycosyl hydrolase derived from an uncultured moth gut bacterium, which is said to bind and degrade large oligoxylans (Brennan *et al*., 2004). GH family 53 enzymes (arabinogalactan endo-beta-1,4-galactanases) for the digestion of arabinogalactans, another component of hemicellulose were found in the genomes of sub cluster If (R6G12 and R5C12)

Genes predicted to encode cellobiohydrolases and exo-glucanases were absent from the treponeme genomes, which coincides with what was discovered in the metagenome (Warnecke *et al*., 2007). These enzymes breakdown the crystalline structure of cellulose, without this hydrolysis, components of lignocellulose are not available, however, lignocellulose degradation may also benefit from the alkaline treatment within the first proctodeal section (P1) in higher termites, as its shown to separate hemicellulosic components and alter the crystalline cellulose structure (Mittal *et al*., 2011). The effect of the alkaline PI on the overall structure of digested and masticated lignocellulose has not been currently researched. Cellulose has also known to be degraded more rapidly in co-cultures of cellulolytic bacteria (*T. caldarium* and *C. thermocellum*), where in conjunction with other bacteria in the termite gut, the treponemes may have the potential of enhancing cellulose breakdown (Leschine, 1995).

The different proportions of glycosyl hydrolases (Figure 2) among the single cell genomes suggests that this symbiosis is under constant dynamic shifting, that these treponemes may function in total mutualism with each other and other dominant members of this community (Yuki *et al*., in prep). The niche partitioning most likely developed through the horizontal gene transfer (HGT) of GH enzymes as shown by the diverse homology of treponeme derived sequences. HGT is common in mesophilic anaerobic environments where there is a diverse assemblage of microorgansims, however the reasons for this remain unclear (Caro-Quintero and Konstantinidis, 2014).

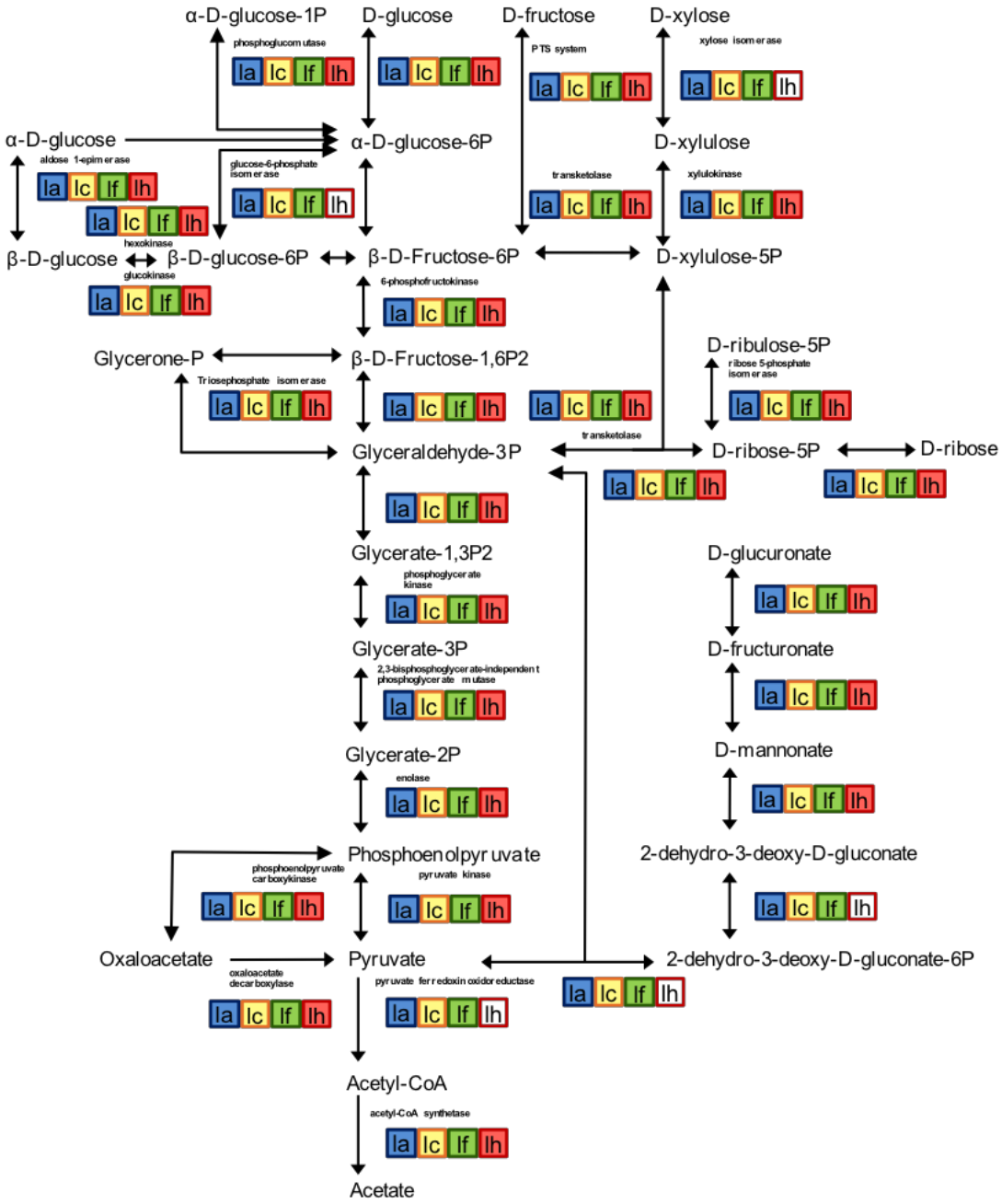

Figure 8. Summary of the general central metabolic pathways possessed by *Treponema* subclusters within *N. takasagoensis* gut. Blue corresponds to subcluster Ia, Yellow to subcluster Ic, Green to subcluster If, Red to subcluster Ih and white infers absence.

Central metabolism

All the treponeme groups seem to be capable of central metabolic processes (Figure 8. and Table 5). Akin to many anaerobic bacteria the TCA cycle is incomplete, and therefore not essential for the growth of these organisms. The enzyme Fructose-1,6-bisphosphatase I (EC3.1.3.11) from the glycolytic pathway was absent in all samples including termite derived cultured species and free-living *T. caldarium*, suggesting this enzyme is non-essential in Treponema cluster I. Instead the presence of pyrophosphate--fructose 6-phosphate 1-phosphotransferase (Kemp and Tripathi, 1993) complements the lack of FBP.

Table 5. Summary of central metabolic pathway enzymes within the Treponema groups

	TTG1	TTG2	TTG3	TTG5	TTG6	TTG7	TTG8	TTG9	TTG10	*T. azotonutricium*	*T.primitia*	*T.caldarium*
Glycolysis												
PTS system (EC2.7.1.69)	+	+	+	+	+	+	+	+	+	+	+	+
Glucose-6-phosphate isomerase (EC5.3.1.9)	+	+	+	+	–	+	–	–	+	+	+	+
Aldose 1-epimerase (EC5.1.3.3)	+	+	+	–	–	+	–	+	+	+	+	+
Hexokinase (EC2.7.1.1)	+	+	+	+	+	+	+	+	+	+	+	+
Glucokinase (EC2.7.1.2)	–	–	+	–	–	–	–	+	+	+	–	+
6-phosphofructokinase 1 (EC2.7.1.11)	+	+	+	+	+	+	+	+	+	+	+	+
Fructose-1,6-bisphosphatase I (EC3.1.3.11)	–	–	–	–	–	–	–	–	–	–	–	–
Pyrophosphate--fructose 6-phosphate 1-phosphotransferase, alpha subunit (EC 2.7.1.90)	+	+	+	+	+	+	–	–	+	+	+	+
Fructose-bisphosphate aldolase, class II (EC4.1.2.13)	+	+	+	+	+	+	+	+	+	+	+	+
Triosephosphate isomerase (EC5.3.1.1)	+	+	–	+	–	+	+	+	–	–	+	+
Glyceraldehyde 3-phosphate dehydrogenase(EC1.2.1.12)	+	+	+	+	–	+	+	+	+	+	+	+
Phosphoglycerate kinase (EC2.7.2.3)	+	+	+	+	–	+	+	+	+	+	+	+
2,3-bisphosphoglycerate-independent phosphoglycerate mutase (EC5.4.2.12)	+	+	–	+	–	+	–	+	+	+	+	+
Enolase (EC4.2.1.11)	+	+	–	–	+	+	–	+	+	+	+	+
Pyruvate kinase (EC2.7.1.40)	+	+	–	+	–	+	+	+	+	+	+	+
Pentose and glucuronate interconversions, Pentose phosphate pathway												
Xylose isomerase (EC5.3.1.5)	+	+	–	+	–	+	+	–	+	–	+	+
Xylulokinase (EC2.7.1.17)	+	+	–	+	–	+	+	+	+	+	+	+
Transketolase (EC2.2.1.1)	+	+	–	+	+	+	+	+	+	+	+	+
Ribulose-phosphate 3-epimerase (EC5.1.3.1)	+	+	–	+	+	+	+	+	+	+	+	+
Ribose 5-phosphate isomerase A (EC5.3.1.6)	+	+	+	+	–	+	+	+	+	+	B	+
Ribokinase (EC2.7.1.15)	+	+	+	+	–	+	+	+	+	+	+	+
Deoxyribose-phosphate aldolase (EC4.1.2.4)	+	+	+	+	+	+	+	+	+	+	+	+
Pyruvate metabolism												
Pyruvate-flavodoxin oxidoreductase (EC1.2.7.-)	+	+	+	+	+	+	+	–	–	+	+	+
Acetyl-CoA synthetase (EC6.2.1.1)	+	+	+	+	–	+	+	+	+	+	+	+
Phosphate acetyltransferase (EC2.3.1.8)	+	+	+	–	–	+	–	–	–	+	+	+

Acetate kinase (EC2.7.2.1)	+	+	+	–	+	+	+	+	+	+	+	+
Acylphosphatase (EC 3.6.1.7)	–	–	–	–	–	–	–	–	+	+	+	+
Phosphoenolpyruvate carboxykinase [GTP] (EC4.1.1.32)	+	+	+	+	+	+	+	+	+	+	+	+
Malate dehydrogenase (EC 1.1.1.37)	+	+	+	–	–	–	–	–	–	+	–	–
NADP-dependent malic enzyme (EC 1.1.1.40)	–	+	–	–	+	+	+	–	+	+	+	+
Oxaloacetate decarboxylase (EC 4.1.1.3)	+	+	+	–	+	+	+	+	+	+	+	+
Alcohol dehydrogenase, Acetaldehyde dehydrogenase	+	+	+	+	–	+	+	+	+	+	+	+
Sugar and uronate utilisation												
Mannose-6-phosphate isomerase (EC5.3.1.8)	+	+	+	+	+	+	+	+	+	+	+	+
4-Alpha-glucanotransferase (amylomaltase) (EC2.4.1.25)	+	+	+	+	+	+	+	+	–	+	+	+
Trehalase (EC3.2.1.28)	+	+	–	+	–	+	–	+	+	+	+	+
Uronate isomerase (EC5.3.1.12)	+	+	+	–	+	+	+	+	+	+	+	+
D-mannonate oxidoreductase (EC1.1.1.57)	+	+	+	+	–	+	+	+	+	+	+	+
Mannonate dehydratase (EC4.2.1.8)	+	+	+	+	–	+	+	+	–	–	+	+
2-Dehydro-3-deoxygluconate kinase (EC4.2.1.8)	+	+	–	+	+	+	+	–	+	+	+	+
2-Dehydro-3-deoxyphosphogluconate aldolase (EC4.1.2.14)	+	+	–	+	+	+	+	–	+	+	+	+
Glycogen metabolism												
Glucose-1-phosphate adenylyltransferase (EC2.7.7.27)	+	+	–	–	+	+	+	+	+	+	+	+
Glycogen synthase, ADP-glucose transglucosylase (EC2.4.1.21)	+	+	+	+	+	+	+	+	+	+	+	+
Glycogen phosphorylase (EC2.4.1.1)	+	+	–	+	+	+	+	–	+	+	+	+
Glycogen branching enzyme, GH-57-type, archaeal (EC2.4.1.18)	+	+	–	+	–	+	–	+	+	+	+	+
Phosphoglucomutase (EC5.4.2.2)	+	+	+	+	–	+	–	+	+	+	+	+

Acetogenesis.

In termite guts the provision of acetate to the host termite is essential for the host termites' survival, acetogenesis depends on the Wood-Ljundahl pathway from the combined fixing of formate and carbon dioxide to form acetate. The complete pathway allows organisms to fix two molecules of carbon dioxide into acetate autotrophically, by two branches, methyl and carbonyl (Schuchmann and Müller, 2014). Genes coding for Wood-Ljungdahl pathway proteins were present intermittently throughout the genomes. This was attributed to the partial nature of single cell genome recovery, whole sub clusters were summarised in Table 3. Within the carbonyl branch, the enzyme Acetyl-coA synthase/Carbon monoxide dehydrogenase (CO/ACS) was found in many samples, and three distinct phylogenies for this enzyme were apparent (Figure 5). Here, one cluster is comprised of samples from sub clusters Ih, Ia and If and shares homology with sequences from other acetogenic termite gut treponemes, 'Candidatus *T. intracellularis*', *T. primitia* and *T. azotonutricium*. Sub cluster Ic samples were phylogenetically dissimilar to other treponeme sequences clustering with Firmicutes and Deltaproteobacteria (Figure 5) and showed a distinctly different gene structure where the acsD

subunit of the complex is fused to the main CO/ACS, possibly inferring secondary or novel function.

The methyl branch of the Wood-Ljundahl pathway begins with formate dehydrogenase (FDH), we confirmed this enzyme is absent within the treponeme genomes, supporting findings of Warnecke and colleagues in 2007. The FDH-linked iron hydrogenase was also absent from treponeme genomes after conserved region searching (Ballor *et al*., 2012). The lack of FDH or a FDH linked hydrogenase suggests either another mechanism, as many homoacetogens have various mechanisms and different enzyme complexes involved in the production/fixing the H_2 CO_2 of formate (Schuchmann and Müller, 2014) or FDH is just not required in this environment. The second gene on the methyl branch of reductive acetogenesis is formate-tetrahydrofolate synthase/ligase (fths), within all samples that harbour this gene, the enzyme shares the greatest sequence homology to those in *T. primitia*, 'Ca *T. intracellularis*' and acetogenic Firmicutes.

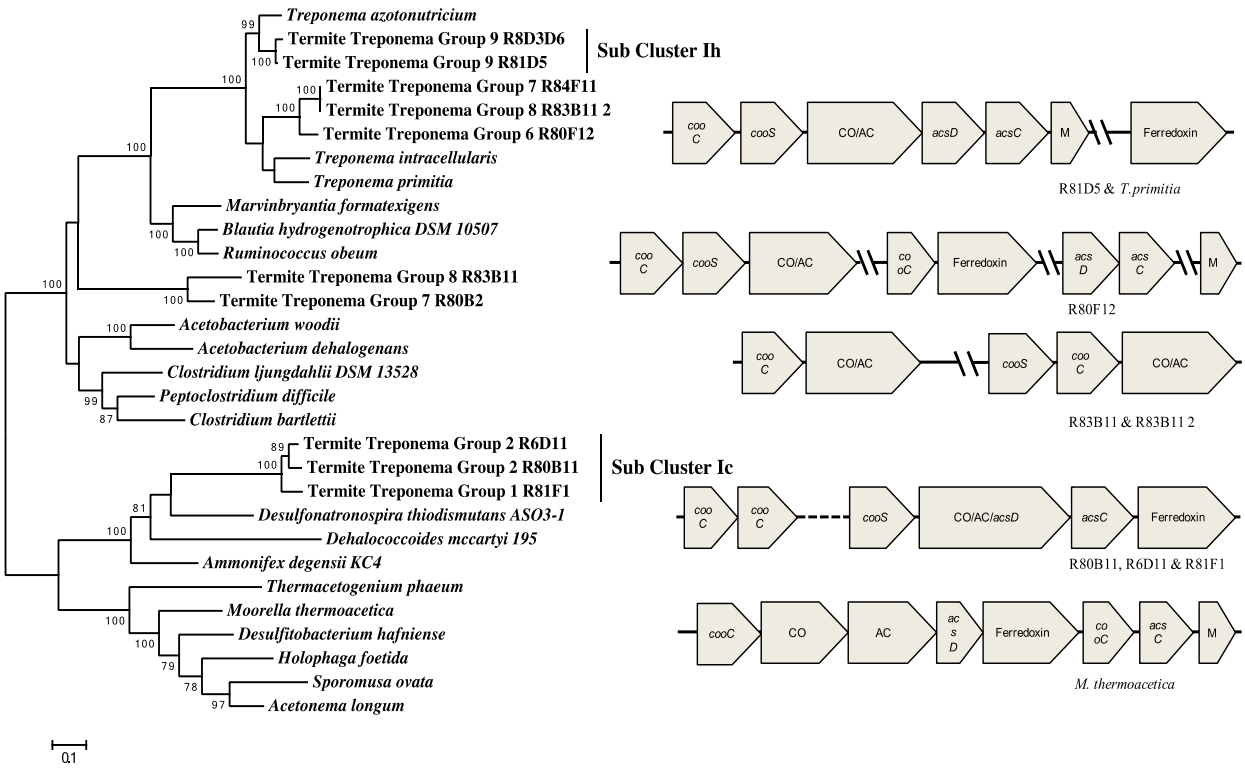

Figure 5. Acetyl-CoA synthase/CO dehydrogenase amino acid tree and genomic syntenic profiles of selected samples showing associated genes. CooC, CO dehydrogenase accessory protein; CooS, CO dehydrogenase subunit; COAC, CO dehydrogenase/acetyl-CoA synthase; acsC, Acetyl-CoA synthase corrinoid iron-sulphur large protein; acsD, Acetyl-CoA synthase corrinoid iron-sulphur small protein; M, 5-Methyltetrahydrofolate corrinoid iron-sulphur methyltransferase and ferredoxin

Table 3. Summary of Wood-Ljungdahl gene components and presence in the sub clusters.

Carbonyl Branch	Sub Cluster Ia	Sub Cluster Ic	Sub Cluster If	Sub Cluster Ih	Protein
acsA	○	○	○	○	Carbon monoxide dehydrogenase subunit /cooS
acsB	○	○	○	○	Acetyl-CoA synthase subunit
acsC	○	○	○	○	large and small subunits of Fe-S-Co protein
acsD	○	○	○	○	large and small subunits of Fe-S-Co protein
acsE	○	○	○	○	5-methyltetrahydrofolate:corrinoid iron-sulphur protein methyltransferase
acsF	○	○	○	○	Ni-insertion protein/cooC
Methyl branch					
fths	○	○	○	x	Formate--tetrahydrofolate ligase (EC 6.3.4.3)
folD	○	○	○	○	Methylenetetrahydrofolate dehydrogenase (NADP+) (EC 1.5.1.5) / Methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9)
metF	○	○	○	x	5,10-methylenetetrahydrofolate reductase (EC 1.5.1.20)

There were no samples whose fths genes shared homology to *T. azotonutricium* or other non-acetogenic treponemes, *T. caldarium* and *Spirocheta alcalica* (Figure 6).

Unlike other cultured treponemes, these single cell genomes encode carbonic anhydrase genes, relating to the Beta clade, class C, these are thought to be involved in acetogenesis (Smith and Ferry., 2000), the treponeme *T. caldarium* contains a Beta clade enzyme but is a separate class B. Expression of these Carbonic anhydrases were slightly elevated in Termite Treponema group 3 (704 RPKM) than in TTG2 (221 RPKM). The role of the beta clade carbonic anhydrase in these treponemes, has been suggested to concentrate carbon dioxide levels within the cell for utilisation by the Wood-Ljundahl pathway, similar to the elevation of CO_2 in photosynthetic cyanobacteria (Smith and Ferry., 2000) and where growth is also enhanced by CO_2-bicarbonate as in ruminal xylan fermenting bacteria *Roseburia intestinalis* and *Eubacterium xylanophilum* (van Gylswyk and van der Toorn, 1985).

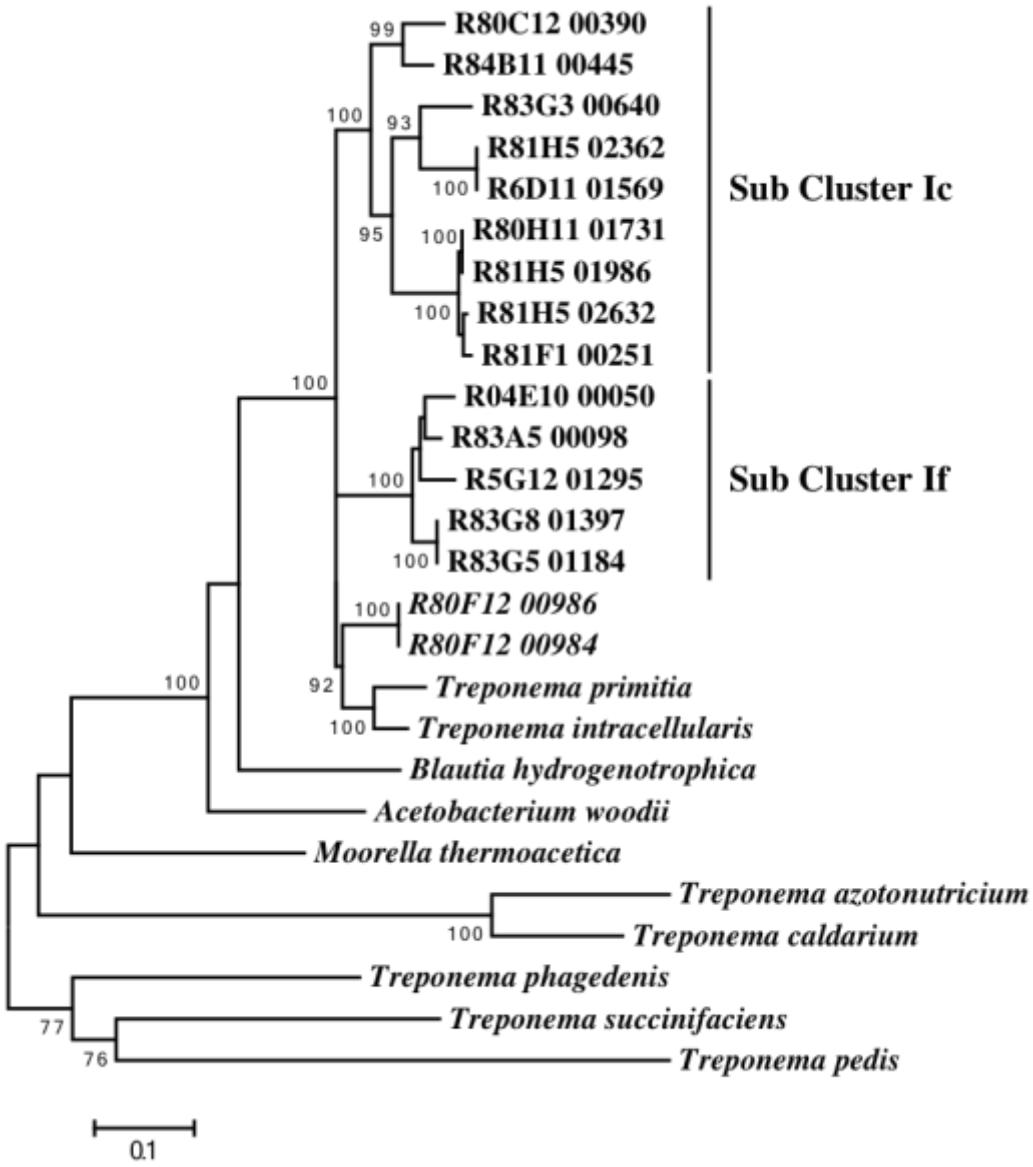

Figure 6. Formate tetrahydrofolate synthase phylogenetic tree of related homoacetogens and those in other *Treponema* species. Maximum likelihood phylogenetic tree, based on amino acid alignment and WAG + G+I model

Iron Hydrogenases

Hydrogenases are important metabolic enzymes utilised in anaerobic microorganisms, where they catalyse the reversible reaction of hydrogen ions (H^+) to H^2, generated by fermentation. The genomes yielded multiple iron hydrogenase and hydrogenase-like genes, many lacking key cysteine residues in their active sites. These were similar in structure to those found in the metagenomic study (Warnecke *et al*., 2007). Here they clustered with similar families. Family

three are group A Fe hydrogenases and these were found in almost all groups except TTG3, TTG8 and TTG9 samples. TTG9 had fragmented associated proteins syntenic in full genomic structures with the NAD-reducing Iron hydrogenase. Samples in sub cluster Ic contain family 7 of iron hydrogenases, those with a thioredoxin domain near the C-terminal, shared greatest homology with Desulfovibrio, Endomicrobia, *Spirochaeta bajacaliforniensis* and *Spirochaeta smaragdinae* sequences (Zheng *et al*., 2013).

Traditional FeFe hydrogenases cluster into 4 different groups, 3 groups were affiliated into the FeFe-hydrogenase group A and 1 group into group B, the genomes yielded many iron hydrogenases. Almost all active (cysteine residing) group A sequences clustered with sequences of ZAS-9 and Elusimicrobia. The second biggest group clustered with other known treponemes, Bacteroides, Desulfovibrio and anaerobic protists including those from oxymonad species from lower termites. Similarly, the TTG 2 included separated phylogenetic positions from the majority and settling between termite group 1 in FE group A and *Pelosinus fermentens* in group B, little is known of the activity of group B FE hydrogenases. The operon structure includes both group B and A hydrogenases coupled to a NADH dehydrogenase with closest homology with *S. bajacaliforniensis* and *S. smaragdinae*.

Nitrogen metabolism

Nitrogen fixation is an important process for the incorporation of atmospheric nitrogen into biomass where general sources are low, and depends on the presence of a nitrogenase enzyme and assembly proteins, usually localised in operons. Nitrogenase operons were discovered in multiple samples and cluster into pseudo-nif, anaerobe cluster III-I and anaerobe cluster I. A concatenated NifHDK tree (Figure 7) clarified the positions of these and they all seem to fit into pseudo-nif category. However recently in the organism *Endomicrobium proavitum* (Zheng *et al*., 2016) a free living relative of the Elusimicrobia endosymbiont *Endomicrobium trichonymphae* and a Bacteroides ectosymbiont of Barbulanympha from *Cryptocercus punctulatus* (Tai *et al*., 2016) both expressed a pseudo-*nif* gene that was shown to be responsible for nitrogen fixation in pure cultures of the bacteria indicating that the treponemes within *N. takasagoensis* are potentially capable of nitrogen fixation.

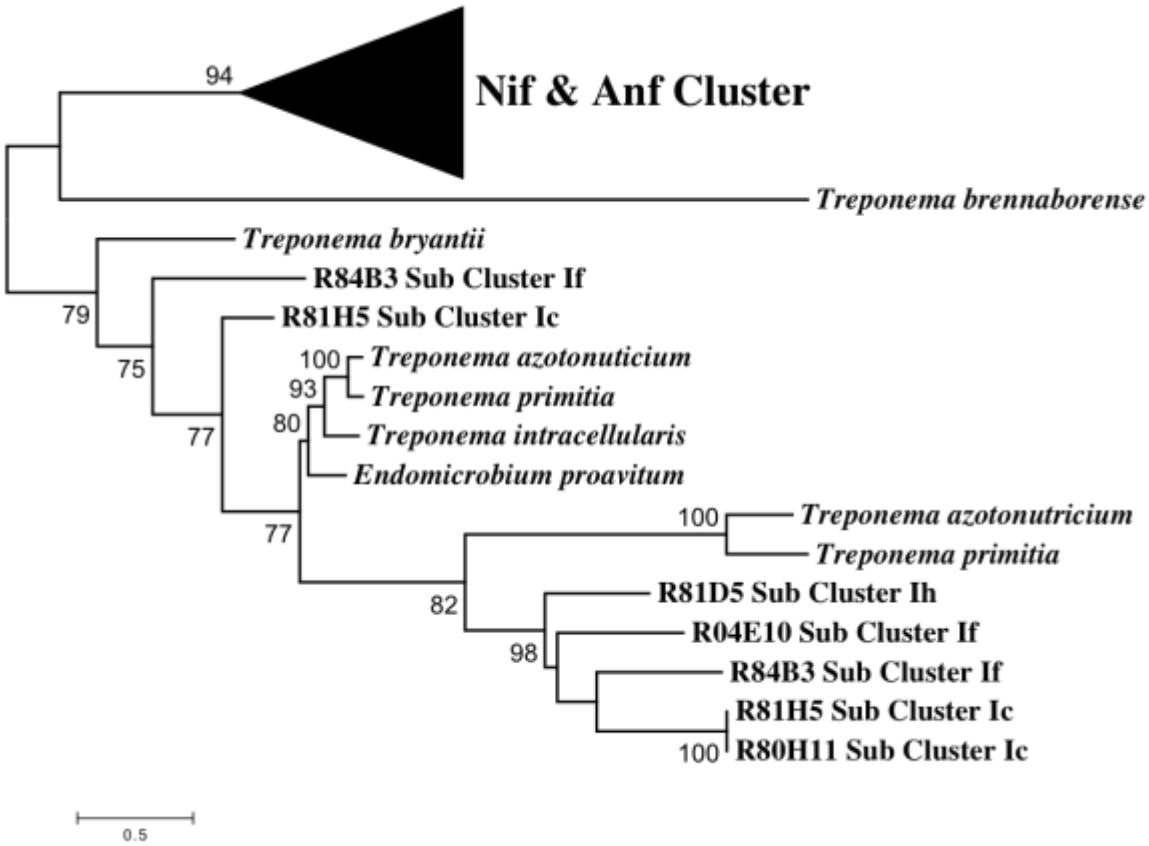

Figure 7. Concatenated NifHDK amino acid phylogenetic tree showing placement of *Nasutitermes Treponema* samples. The cluster is believed to be Cluster IV with unknown function, however there is evidence to support its nitrogen fixation capability. The maximum likelihood tree is based on an alignment of 1030 amino acids masked to the R80H11 sample. The amino acid substitution was LG with gamma and invariable sites. The tree topology was reanalysed with 500 bootstrap resamplings and rooted at midpoint.

Recycling of nitrogenous compounds can also liberate useful waste products, for example urea and uric acid. Genes associated with the urease operon were present in three samples. They phylogenetically cluster with 'Ca *T. intracellularis*' and clostridial related sequences (74%), with a high sequence identity between all sample genes (95%), and one near complete operon in R81D5 with high expression. Indicating that some members of sub clusters C, F and H are capable of recycling nitrogen using waste urea and ammonia and that nitrogen recycling was not definitive of a sub cluster, and was shared between members of this community. Alternative ways of recycling and upscaling Nitrogen within the deficient environment were present and include nitrite reductase (Ih) and ammonia transporters (amt). From these findings, treponemes have the potential to provide metabolic nitrogen recycling and fixing methods to sustain the host despite the lack of nitrogen resources in the diet.

Amino acids

All amino acid biosynthesis pathways were present throughout samples, individual samples may have had integral synthesis enzymes missing due to incompleteness of individual sample genomes, however others in the same cluster/group would have been present (Table 4).

There were many incidences of possible HGT's within genomes regarding tRNA synthases based on comparative work completed on *Spherochaeta globusa* (Caro-Quintero *et al*., 2012). The enzyme aspartyl/glutamyl-tRNA amidotransferase has two separate clusters in the treponeme samples, one group share sequence homology to the cultured termite treponemes and the other clustering with ruminal Clostridia (Roseburia) sequences. Both R80F12 and R6C12 samples in sub cluster Ic and If respectively, share different phenylalanyl tRNA synthetases phylogenies to the rest of the samples, related to clostridia and the Spirochete genus Leptospira. Ferredoxin-dependent Glutamate synthases that share high sequence identities with thermophilic cellulose degrading Firmicutes (*Herbinix hemicellulosilytica)*. These all indicate that horizontal gene transfers (HGT's) are common between these sub clusters and other gut microbiota.

Table 4. Presence and absence of enzymes involved in amino acid synthesis pathways, summarised for the Termite *Treponema* groups and including close species.

Amino Acid Synthesis Pathway	Enzyme (EC. Number)	TTG1	TTG2	TTG3	TTG5	TTG6	TTG7	TTG8	TTG9	TTG10	*T. azotonutricium*	*T. primitia*	*T. caldarium*
Histidine Pathway	Ribose 5-phosphate isomerase (EC 5.3.1.6)	+	+	+	+	–	+	+	+	+	+	+	+
	Ribose-phosphate pyrophosphokinase (EC 2.7.6.1)	+	+	–	+	–	+	+	–	+	+	+	+
	ATP phosphoribosyltransferase (EC 2.4.2.17)	+	+	+	–	+	+	+	+	+	+	+	+
	Phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / Phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31)	+	+	+	+	–	+	+	+	–	+	+	+
	Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)	–	–	–	–	–	–	–	–	–	–	–	–
	Imidazole glycerol phosphate synthase cyclase subunit (EC 4.1.3.-)	+	+	+	+	–	+	+	+	+	+	+	+
	Imidazole glycerol phosphate synthase amidotransferase subunit (EC 2.4.2.-)	+	+	+	–	+	+	+	+	+	+	+	+
	Imidazoleglycerol-phosphate dehydratase (EC 4.2.1.19)	+	+	+	–	–	+	+	+	–	+	+	+
	Histidinol-phosphate aminotransferase (EC 2.6.1.9)	+	+	+	+	–	+	+	+	–	+	+	+
	Histidinol-phosphatase (EC 3.1.3.15)	+	+	+	+	+	+	+	–	+	+	+	+
	Histidinol dehydrogenase (EC 1.1.1.23)	+	+	+	+	–	+	+	+	+	+	+	+
Aromatic Amino Acid Pathway	2-keto-3-deoxy-D-arabino-heptulosonate-7-phosphate synthase I beta (EC 2.5.1.54)	+	+	–	–	–	+	–	+	–	+	+	+
	3-dehydroquinate synthase (EC 4.2.3.4)	+	+	–	–	–	+	+	–	–	+	+	+
	3-dehydroquinate dehydratase I (EC 4.2.1.10)	+	+	+	–	–	+	+	+	–	–	–	–

	Shikimate 5-dehydrogenase I alpha (EC 1.1.1.25)	+	+	+	+	–	+	+	+	+	+	+	+
	Shikimate kinase I (EC 2.7.1.71)	+	+	–	–	–	+	–	+	+	+	+	+
	5-Enolpyruvylshikimate-3-phosphate synthase (EC 2.5.1.19)	+	+	–	–	–	+	+	–	+	+	+	+
	Chorismate synthase (EC 4.2.3.5)	+	+	–	–	–	+	–	–	+	+	+	+
	Chorismate mutase I (EC 5.4.99.5)	+	+	–	–	–	+	+	+	–	+	+	+
	Prephenate and/or arogenate dehydrogenase (EC 1.3.1.12)(EC 1.3.1.43)	+	+	–	–	–	+	–	–	–	+	+	+
	Biosynthetic Aromatic amino acid aminotransferase alpha (EC 2.6.1.57) / Aspartate aminotransferase (EC 2.6.1.1)	+	+	+	–	+	+	+	+	+	+	+	+
	Aspartate transaminase/Aspartate aminotransferase (EC 2.6.1.1)	+	+	–	+	+	+	+	+	+	+	+	+
	Prephenate dehydratase (EC 4.2.1.51)	+	+	–	–	–	+	–	+	–	–	–	–
	Anthranilate synthase, aminase component (EC 4.1.3.27)	+	+	+	–	–	+	+	+	–	–	+	+
	Anthranilate phosphoribosyltransferase (EC 2.4.2.18)	+	+	–	–	–	+	–	+	–	–	+	+
	Phosphoribosylanthranilate isomerase (EC 5.3.1.24)	+	+	–	+	–	+	–	+	–	–	+	–
	Indole-3-glycerol phosphate synthase	+	+	–	–	–	+	–	+	–	–	+	+
	Tryptophan synthase (EC 4.2.1.20)	+	+	+	+	+	+	+	+	–	+	+	+
Serine Cysteine Glycine Alanine Pathways	D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95)	+	+	–	+	+	+	–	+	+	+	+	+
	Phosphoserine aminotransferase (EC 2.6.1.52)	+	+	–	+	+	+	–	+	+	+	+	+
	Phosphoserine phosphatase (EC 3.1.3.3)	+	+	–	+	+	–	–	–	–	–	–	–
	Serine hydroxymethyltransferase (EC 2.1.2.1)	+	+	+	+	–	+	+	+	–	+	–	+
	Serine acetyltransferase (EC 2.3.1.30)	+	+	+	+	+	+	–	+	–	+	+	+
	Cysteine synthase (EC 2.5.1.47)	+	+	+	+	+	+	+	+	+	+	+	+
	Cysteine desulfurase (EC 2.8.1.7)	+	+	+	+	+	+	+	+	+	+	+	+
Threonine Methionine Pathways	Threonine synthase (EC 4.2.3.1)	+	+	–	+	+	+	+	–	+	+	+	+
	Homoserine kinase (EC 2.7.1.39)	+	+	–	+	+	+	+	+	+	+	+	+
	Homoserine O-succinyltransferase (EC 2.3.1.46)	+	+	–	+	+	+	+	+	–	+	+	+
	O-acetylhomoserine sulfhydrylase (EC 2.5.1.49) / O-succinylhomoserine sulfhydrylase/Cystathionine gamma-synthase (EC 2.5.1.48)	+	+	+	–	+	+	+	+	–	+	+	+
	Cystathionine beta-lyase, Bsu PatB (EC 4.4.1.8)	+	+	–	+	–	–	–	+	–	+	+	–
	5-methyltetrahydrofolate--homocysteine methyltransferase (EC 2.1.1.13)	+	+	+	+	+	+	+	+	+	+	+	+
	Aspartokinase (EC 2.7.2.4) / Homoserine dehydrogenase (EC 1.1.1.3)	+	+	+	+	+	+	+	+	–	+	+	+
	Aspartate aminotransferase (EC 2.6.1.1)	+	+	+	+	+	+	+	+	+	+	+	+
Aspartate Pathway	Aspartate-ammonia ligase (EC 6.3.1.1)	+	+	–	–	–	–	–	+	+	–	–	+
	Aspartokinase (EC 2.7.2.4)	+	+	+	+	+	+	+	+	–	+	+	+
	Argininosuccinate synthase (EC 6.3.4.5)	+	+	+	+	+	+	+	+	+	+	+	+
	Argininosuccinate lyase (EC 4.3.2.1)	+	+	+	+	+	+	+	+	+	+	+	+
	Aspartate-semialdehyde dehydrogenase (EC 1.2.1.11)	+	+	+	+	+	+	+	+	–	+	+	+
	4-hydroxy-tetrahydrodipicolinate synthase (EC 4.3.3.7)	+	+	+	+	–	+	+	+	–	+	–	+
	4-hydroxy-tetrahydrodipicolinate reductase (EC 1.17.1.8)	+	+	–	+	–	+	+	+	–	+	–	+
	L,L-diaminopimelate aminotransferase (EC 2.6.1.83)	+	+	+	+	+	+	+	+	–	+	+	+
	Carbamoyl-phosphate synthase large chain (EC 6.3.5.5)	+	+	+	+	–	–	–	+	+	–	–	–
	Diaminopimelate decarboxylase (EC 4.1.1.20)	+	+	+	+	+	+	+	+	+	+	+	+
Glutamate Pathway	Glutamate synthase [NADPH] large chain (EC 1.4.1.13)	+	+	+	+	+	+	+	+	+	+	+	+
	Glutamine synthetase type III, GlnN (EC 6.3.1.2)	+	+	–	+	–	+	+	+	+	+	+	+
	Glutamate 5-kinase (EC 2.7.2.11)	+	+	+	+	+	+	+	+	+	+	+	+
	Gamma-glutamyl phosphate reductase (EC 1.2.1.41)	+	+	+	+	+	+	+	+	+	+	+	+

	Pyrroline-5-carboxylate reductase (EC 1.5.1.2)	+	+	+	–	–	+	+	+	+	+	+	+
Branched-Chain	Ketol-acid reductoisomerase (EC 1.1.1.86)	+	+	+	–	–	+	–	+	–	+	+	+
Amino Acid Pathway	Dihydroxy-acid dehydratase (EC 4.2.1.9)	+	+	+	–	+	+	–	+	–	+	+	+
	Branched-chain amino acid aminotransferase (EC 2.6.1.42)	+	+	+	–	+	+	+	+	+	+	+	+
	2-isopropylmalate synthase (EC 2.3.3.13)	+	+	–	–	–	+	–	+	–	+	+	+
	3-isopropylmalate dehydratase large subunit (EC 4.2.1.33)	+	+	–	–	+	+	+	+	–	+	+	+
	3-isopropylmalate dehydrogenase (EC 1.1.1.85)	+	+	+	–	–	+	+	+	–	+	+	+

Motility genes

The ability of treponemes to move within the anaerobic termite gut lumen is important in responding to environmental stimuli and to locate potential resources. The termite gut luminal fluid is highly viscous, which poses a challenge to non-helical bacteria whose movement is reduced. Spirochetes helical morphology allows fast propulsion through the viscous gut due to periplasmic flagella rotation (Greenberg and Canale-Parola, 1977). The RNA-seq data revealed that transcripts of components of flagellin (subunit B3), and of related chemotaxis proteins (cheY) were highly abundant and ortholog analysis revealed the most common orthologous group found in the genomes were the methyl accepting chemotaxis proteins (MCP). This family of proteins are used chemosensory to detect attractive environmental compounds. Novel MCP's were detected featuring REC domains, also found in the iron hydrogenases, this domain is known as a signal receiver domain homologous to cheY (Galperin, 2006). The diversity of these proteins suggests that the treponemes are responsive to many external stimuli, and potentially important to avoid environmental stressors (O_2) and chemo-attraction to partially hydrolysed lignocellulose.

The abundance of motility related genes infers their potential reliance on motion in this environment, in contrast to the termite endosymbiont Ca *T. intracellularis* that has lost all of its' flagella/motility orthologous genes and thus their spiral morphology and serves another function within the protist host.

CRISPR analysis

The CRISPR/CAS system is a method for the removal of foreign DNA from the host genome, CRISPR analysis was completed to explore the presence of these sequences within the treponeme genomes. From the entire dataset 16 samples contained CRISPR related sequences within their genomes. There were two different types shown in the annotation as samples from TTG1, 6 and 7 contain CAS 1-5, whereas four samples in TTG7 feature a CRISPR type III-A system. The TTG7 CAS-system, CAS I and II shared homology *T. azotonutricium* and

associated proteins to *T. primitia*, whereas in TTG1 they shared greatest homology to *T. primitia* and the other CRISPR associated proteins sharing sequence homology to that of Firmicutes. In TTG 6, all CRISPR-CAS associated proteins shared greatest homology to *T. primitia*. Two general genotypes exist; one that shares homology with other Spirochaetes and the other affiliated with Firmicutes, potentially indicating further HGT between these two phyla.

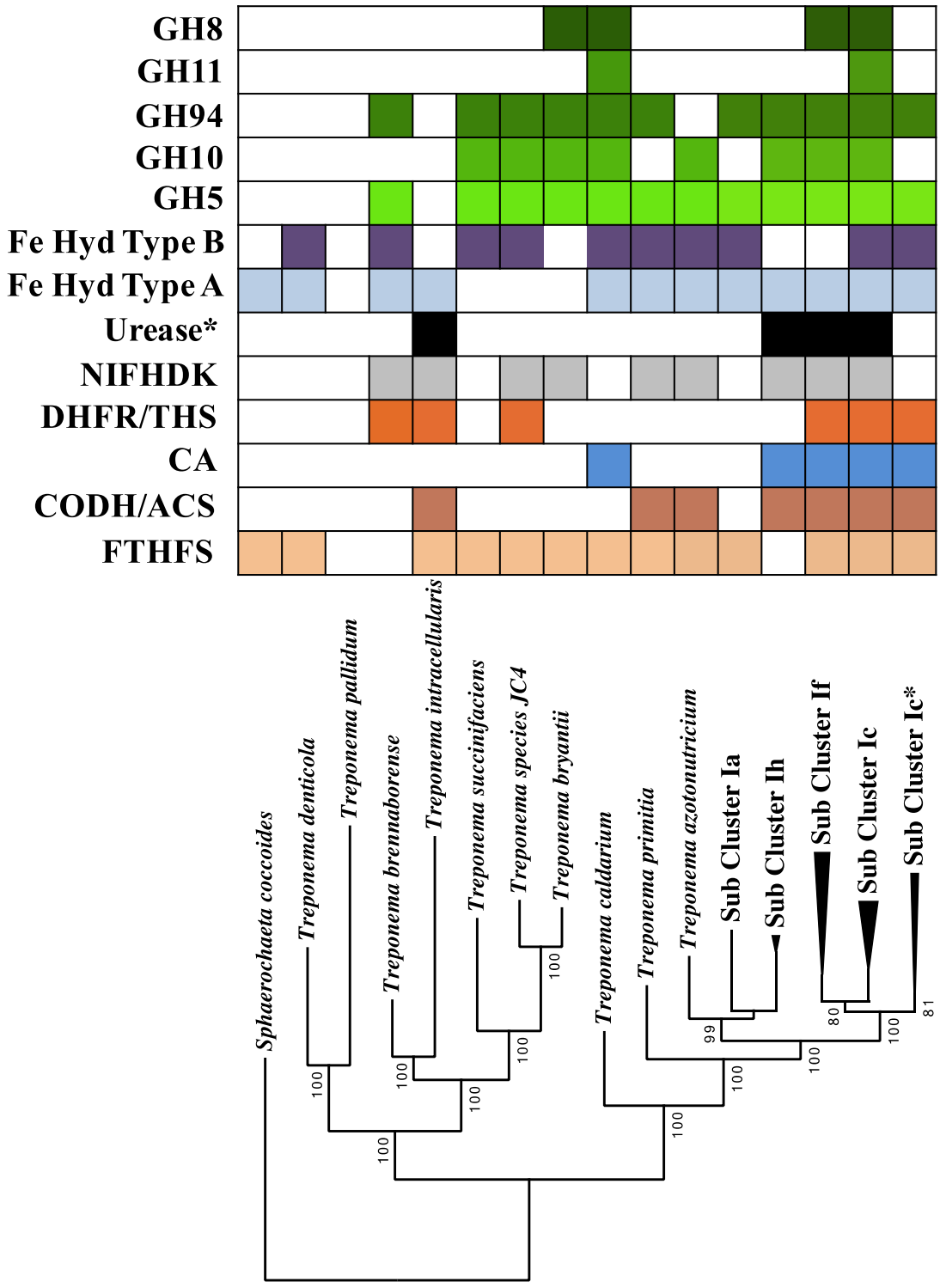

Figure 9. Concatenated 38 ribosomal protein phylogenetic positions of sub clusters from *Nasutitermes takasagoensis* with relevant attributes for termite symbiosis. The phylogenetic tree was made using maximum likelihood method with Whelan and Goldman amino acid substitution model with 1000 bootstrap replicates. FTHS Formate tetrahydrofolate synthase; CODH/ACS, Carbon monoxide dehydrogenase/Acetyl CoA Synthase; CA, Carbonic Anhydrase; DHFR/THS, Dihydrofolate synthase/Thymidylate synthase; NIFHDK, Nitrogenase complex; Urease *presence of UreC protein; Fe Hyd Type A/B, Iron Hydrogenase Type A/B; GH Glycosyl Hydrolase families 5, 10, 94, 11 and 8.

Conclusion

These are the first genomes sequenced for treponemes isolated from the gut of a higher termite, and the first genomic representations of individuals of the phylogenetic sub clusters Ic, If and Ih. The genomes demonstrate the putative roles of these treponemes within the gut of *N. takasagoensis* in lignocellulose digestion, nitrogen fixation and recycling and acetate production.

All four sub clusters have the genomic potential to utilise and breakdown components of both cellulose and hemicellulose properties of wood. The fibre-associated sub clusters Ic and If would potentially utilise secreted glycosyl hydrolases and highly expressed mono/oligosaccharide ABC type transport systems to release and capture lignocellulosic components for heterotrophic growth. These systems would likely work in conjunction with other fibre-associated community members, Fibrobacteres and TG3 forming the nutritional symbiosis. Sub cluster Ic genomes share a similar complement of GH families important to lignocellulose degradation as the thermophilic *T. caldarium* (Figure 9), which also in co-cultures with *C. thermocellum* enhances cellulose breakdown (Leschine, 1995).

Sub cluster representatives of Ic, If and Ih have the molecular machinery required for nitrogen fixation, however their capacity as functioning nitrogenases is unknown, and within these same sub clusters urease is also present for the recycling of nitrogen (Figure 9). In an environment that is deficient in accessible nitrogen the treponeme samples have the potential to provision the host and other community members.

The termite gut provides a species rich and dynamically fluid environment for genetic transfer, within the treponeme genomes horizontal gene transfer is evident in multiple metabolic pathways. Two bacterial sources of HGT were phylogenetically inferred, these were Firmicutes and Bacteroidetes, both phyla include species previously cultured from the termite gut environment (*Clostridium termitidis*, *Bacteroides reticulotermitis*).

All termite isolated samples and sub clusters including sub cluster Ia (*T. azotonutricium*) possess the Carbon monoxide dehydrogenase/Acetyl-CoA synthase enzyme (Figure 9) in contrast to non-termite associated treponeme genomes. This indicates the importance of these enzymes within the termite gut environment. The presence of both lignocellulose degrading glycosyl hydrolases and acetogenesis associated enzymes indicates that these treponemes are mixotrophic. The presence of iron hydrogenases in the treponeme sub clusters can help

counterpoise the release of hydrogen in the fermentation of cellulosic compounds and enhance acetogenesis.

The complexity and diversity of the termite gut community means that potentially there are thousands of interactions between residing microbes and the termite gut wall and the actions of the gut treponemes intrinsically prime the nutritional symbiosis on their native wood feeding diet.

This study supports the usage of single cell genomics and metatranscriptomics in exploring the termite gut symbiosis of these diverse phenotypes, and in understanding individual members of this complex microbial community.

Acknowledgements

Hirokazu Kuwahara, Kazuki Izawa and Kaito Sugaya for help with the sequencing, Jun-ichi Inoue for help with Nitrogen fixation and hydrogenase activity.

Author contributions

D.S., M.Y., Y.H., A.D., and M.O. designed the research. M.Y. sorted single cells. D.S. and M.Y., sequenced amplified DNA, assembled and annotated single-cell genomes. D.S., and M.Y., performed RNA-seq. D.S., M.Y., A.D., and M.O. wrote the paper.

Conflict of interest

The authors declare no conflict of interest.

References

Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166-169.

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol Biol* **12**, 186.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75.

Ballor NR, Paulsen I, Leadbetter JR (2012). Genomic analysis reveals multiple [FeFe] hydrogenases and hydrogen sensors encoded by treponemes from the H2-rich termite gut. *Microb Ecol* **63,** 282-294.

Bankevich A, Nurk S, Antipov D, Gurevich AA *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477.

Brennan Y, Callen WN, Christoffersen L, Dupree P *et al.* (2004) Unusual microbial xylanases from insect guts. *Appl Environ Microbiol* **70**, 3609-3617.

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60.

Caro-Quintero A, Konstantinidis K (2014) Inter-phylum HGT has shaped the metabolism of many mesophilic and anaerobic bacteria. *ISME J* **14**, 1751-7362.

Caro-Quintero A, Ritalahti KM, Cusick KD, Löffler FE, Konstantinidis KT (2012) The chimeric genome of Sphaerochaeta: nonspiral spirochetes that break with the prevalent dogma in spirochete biology. *MBio* **3**. e00025-12

Darling AE, Jospin G, Lowe E, Matsen FA *et al*. (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243.

Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165.

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

Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale Genome alignment and comparison. *Nucleic Acids Res* **30**, 2478-2483.

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.

Dodsworth JA, Blainey PC, Murugapiran SK, Swingley WD *et al.* (2013) Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat Commun* **4**, 1854.

Eutick ML, O'Brien RW, Slaytor M (1978) Bacteria from the gut of Australian termites. *Appl Environ Microbiol* **35**, 823-828.

Galperin MY (2006) Structural Classification of Bacterial Response Regulators: Diversity of Output Domains and Domain Combinations. *J Bacteriol* **188**, 4169-4182.

Gilbert HJ, Knox JP, Boraston AB (2013) Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Curr Opin Struct Biol* **23**, 669-677.

Graber JR, Leadbetter JR, Breznak JA (2004) Description of *Treponema azotonutricium* sp. nov. and *Treponema primitia* sp. nov., the first spirochetes isolated from termite guts. *Appl Environ Microbiol* **70**, 1315-1320.

Grabherr MG, Haas BJ, Yassour M, Levin JZ *et al*. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652.

Greenberg EP, Canale-Parola E (1977) Relationship between cell coiling and motility of spirochetes in viscous environments. *J Bacteriol* **131**, 960-969.

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075.

He SM, Ivanova N, Kirton E, Allgaier M *et al.* (2013) Comparative Metagenomic and Metatranscriptomic Analysis of Hindgut Paunch Microbiota in Wood- and Dung-Feeding Higher Termites. *Plos One* **8**, 14.

Hongoh Y, Deevong P, Hattori S, Inoue T *et al.* (2006) Phylogenetic diversity, localization, and cell morphologies of members of the candidate phylum TG3 and a subphylum in the phylum Fibrobacteres, recently discovered bacterial groups dominant in termite guts. *Appl Environ Microbiol* **72**, 6780-6788.

Iida T, Ohkuma M, Ohtoko K, Kudo T (2000) Symbiotic spirochetes in the termite hindgut: phylogenetic identification of ectosymbiotic spirochetes of oxymonad protists. *FEMS Microbiol Ecol* **34**, 17-26.

Jones P, Binns D, Chang H-Y, Fraser M *et al*. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240.

Joshi NA, Fass JN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle

Kemp RG, Tripathi RL (1993) Pyrophosphate-dependent phosphofructo-1-kinase complements fructose 1,6-bisphosphatase but not phosphofructokinase deficiency in *Escherichia coli*. *J Bacteriol* **175**, 5723-5724.

Kohler T, Dietrich C, Scheffrahn RH, Brune A (2012) High-Resolution Analysis of Gut Environment and Bacterial Microbiota Reveals Functional Compartmentation of the Gut in Wood-Feeding Higher Termites (Nasutitermes spp.). *Appl Environ Microbiol* **78**, 4691-4701.

Kudo H, Cheng KJ, Costerton JW (1987) Interactions between *Treponema bryantii* and cellulolytic bacteria in the in vitro degradation of straw cellulose. *Can J Microbiol* **33**, 244-248.

Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M (2013) Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front Genet* **4**, 237.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359.

Leschine SB (1995) Cellulose degradation in anaerobic environments. *Annu Rev Microbiol* **49**, 399-426.

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189.

Marshall IP, Blainey PC, Spormann AM, Quake SR (2012) A Single-cell genome for Thiovulum sp. *Appl Environ Microbiol* **78**, 8555-8563.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12

Mikaelyan A, Strassert JF, Tokuda G, Brune A (2014) The fibre-associated cellulolytic bacterial community in the hindgut of wood-feeding higher termites (*Nasutitermes spp*.). *Environ Microbiol* **16**, 2711-2722.

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

Mittal A, Katahira R, Himmel ME, Johnson DK (2011) Effects of alkaline or liquid-ammonia treatment on crystalline cellulose: changes in crystalline structure and effects on enzymatic digestibility. *Biotechnol Biofuels* **4**, 41.

Miyata R, Noda N, Tamaki H, Kinjyo K, Aoyagi H, Uchiyama H *et al.* (2007) Influence of feed components on symbiotic bacterial community structure in the gut of the wood-feeding higher termite *Nasutitermes takasagoensis*. *Biosci Biotechnol Biochem* **71**, 1244-1251.

Ohkuma M, Iida T, Kudo T (1999) Phylogenetic relationships of symbiotic spirochetes in the gut of diverse termites. *FEMS Microbiol Lett* **181**, 123-129.

Ohkuma, M, Noda S, Hattori S, Iida T *et al.* (2015) Acetogenesis from H2 plus CO2 and nitrogen fixation by an endosymbiotic spirochete of a termite-gut cellulolytic protist. *Proc Natl Acad Sci U S A* **112**, 10224-10230

Pohlschroeder M, Leschine S, Canaleparola E (1994) *Spirochaeta calderia* Sp-nov, A thermophic bacterium that enhances cellulose degradation by *Clostridium thermocellum*. *Arch Microbiol* **161**, 17-24.

Poulsen M, Hu H, Li C, Chen Z, Xu L, Otani S *et al.* (2014) Complementary symbiont contributions to plant decomposition in a fungus-farming termite. *Proc Natl Acad Sci U S A* **111**, 14500-14505.

Powell S, Forslund K, Szklarczyk D, Trachana K *et al.* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* **42**, 231-239.

Price MN, Dehal PS, Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431-437.

Rosenthal AZ, Matson EG, Eldar A, Leadbetter JR (2011) RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. *ISME J* **5**, 1133-1142.

Scheller HV, Ulvskov P (2010) Hemicelluloses. *Annu Rev Plant Biol* **61**, 263–89.

Schloss PD, Westcott SL, Ryabin T, Hall JR *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537-7541.

Schuchmann K, Müller V (2014) Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria. *Nat Rev Microbiol* **12**, 809-821.

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069.

Shintani M, Matsui K, Inoue J-i, Hosoyama A *et al.* (2014) Single-Cell Analyses Revealed Transfer Ranges of IncP-1, IncP-7, and IncP-9 Plasmids in a Soil Bacterial Community. *Appl Environ Microbiol* **80**, 138-145.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313.

Tai V, Carpenter KJ, Weber PK*, et al.* (2016) Genome Evolution and Nitrogen Fixation in Bacterial Ectosymbionts of a Protist Inhabiting Wood-Feeding Cockroaches. *Appl Environ Microbiol* **82**, 4682-4695.

Tamura K, Peterson D, Peterson N, Stecher G, *et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739.

Tokuda G, Watanabe H, Lo N (2007) Does correlation of cellulase gene expression and cellulolytic activity in the gut of termite suggest synergistic collaboration of cellulases? *Gene* **401**, 131-134.

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

van Gylswyk NO, van der Toorn JJTK (1985) *Eubacterium uniforme* sp. nov. and *Eubacterium xylanophilum* sp. nov., fibre digesting bacteria from the ruminal of sheep fed stover. *Int J Syst Bacteriol* **35**, 323-326.

Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M *et al.* (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**, 560-U517.

Yuki M, Kuwahara H, Shintani M, Izawa K *et al.* (2015) Dominant ectosymbiotic bacteria of cellulolytic protists in the termite gut also have the potential to digest lignocellulose. *Environ Microbiol* **17**, 4942-4953.

Zheng H, Bodington D, Zhang C*, et al.* (2013) Comprehensive phylogenetic diversity of [FeFe]-hydrogenase genes in termite gut microbiota. *Microbes Environ* **28**, 491-494.

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.

Chapter III

Intertwining
symbiotic roles of
three dominant
bacteria in the gut of
a wood-feeding
higher termite

Chapter III

This Chapter has been submitted for publishing and is submitted here in its original format. My roles in this study were in designing the experiment, manuscript writing, in the assembly, annotation, and genome analysis of single cells and in performing the RNA-seq and subsequent bioinformatic analyses.

Intertwining symbiotic roles of three dominant bacteria in the gut of a wood-feeding higher termite

Masahiro Yuki[1], David Starns[2], Hirokazu Kuwahara[3], Masaki Shintani[2,4], Yuichi Hongoh[2,3], Moriya Ohkuma[1,2*]

[1]Biomass Research Platform Team, Biomass Engineering Program Cooperation Division, RIKEN Center for Sustainable Resource Science, Tsukuba, Japan; [2]Japan Collection of Microorganisms, RIKEN BioResource Center, Tsukuba, Japan; [3]Department of Life Science and Technology, Tokyo Institute of Technology, Tokyo, Japan; [4]Applied Chemistry and Biochemical Engineering Course, Department of Engineering, Graduate School of Integrated Science and Technology, Shizuoka University, Hamamatsu, Japan

* Corresponding author: mohkuma@riken.jp

Abstract

Gut bacteria generally comprise a complex community but the underlying mechanisms for this complexity remain unclear owing to the lack of our understanding of detailed individual bacterial functions. Here, single-cell genome analyses of three dominant bacterial species, each representing candidate phylum TG3, phylum Fibrobacteres and genus *Treponema* (phylum Spirochaetes) in the gut of a wood-feeding higher termite, *Nasutitermes takasagoensis* are reported. The compositions of encoded glycoside hydrolase genes suggest that the three bacteria digest different lignocellulose components. The TG3 and Fibrobacteres bacteria are nitrogen fixers, while the Fibrobacteres and *Treponema* bacteria play reductive acetogenesis roles with methyl-group donors being different. Surprisingly, the genes responsible for the nitrogen fixation and reductive acetogenesis have been laterally transferred between each pair of the bacteria. RNA-seq analyses indicate substantial contributions of these three bacteria to the gut ecosystem. The complexity of the gut community can be attributed to these intertwining symbiotic roles of individual bacteria.

Introduction

The gut microbiome is one of the most complex microbial communities comprising diverse yet-uncharacterized species. Functional characterizations of its individual members are important to understand how the microbiome develops such complexity. A fascinating example of the gut microbiome is found in the termite gut where symbiotic relationships occur between termites and their gut microorganisms responsible for digestion of woody biomass (Brune & Ohkuma., 2011; Hongoh., 2011; Brune., 2014). Based on the social behaviour of termites exchanging their nutrition in the gut, the gut microbiome is shared among the nestmates and vertically inherited across their generations. Despite the presence of a bottleneck effect during the inheritance, termites maintain their complex gut microbiomes probably owing to the need for symbiotic digestion. Termites are divided into so-called lower and higher termites. The lower termites harbour flagellated protists, bacteria and archaea in their gut, and protists play a central role in lignocellulose digestion. These cellulolytic protists generally harbour specific bacteria intracellularly and/or on their cell surface, and in some cases, the associated bacteria are crucial for the nutrition of the host protists and/or termites such as in nitrogen fixation and reductive acetogenesis from H_2 and CO_2 as well as even aiding in lignocellulose digestion (Hongoh *et al*., 2008a; Hongoh *et al*., 2008b; Desai & Brune., 2012; Ohkuma *et al*., 2015; Yuki *et al*., 2015; Ikeda-Ohtsubo *et al*., 2016).

The higher termites typically lack flagellated protists, and bacteria play an essential role in lignocellulose digestion (Tokuda & Watanabe., 2007). Previous studies indicate that higher termites harbour hundreds of mostly yet-uncultured bacteria in the gut. In the wood-feeding higher termite, *Nasutitermes takasagoensis,* the gut community comprises three predominant groups of bacteria: genus *Treponema* (phylum Spirochaetes, approximately 60% of the gut bacteria), candidate phylum TG3 (originally, candidate phylum Termite Group 3), and *Fibrobacteres* subphylum 2 (each approximately 10% (see Supplementary Fig. 1 (Hongoh *et al*., 2006)). These three groups of bacteria are commonly abundant in many wood-feeding higher termites (Hongoh *et al*., 2005; Warnecke *et al*., 2007; Köhler *et al*., 2012; Mikaelyan *et al*., 2015; Abdul Rahman *et al*., 2015). Only two cultured representatives of the candidate TG3 phylum have been described, which were isolated from a hypersaline soda lake and named as *Chitinivibrio*

alkaliphilus and *Chitinispirillum alkaliphilum* (Sorokin *et al*., 2014; Sorokin *et al*., 2016). Cultured species in the *Fibrobacteres* phylum are also few: *Fibrobacter succinogenes* and *Fibrobacter intestinalis* have been isolated from mammalian intestinal tracts (Jewell *et al*., 2013). The genomes of these TG3 and Fibrobacteres isolates possess a diverse array of glycoside hydrolases (GHs)(Sorokin *et al*., 2014; Sorokin *et al*., 2016; White *et al*., 2014). In the gut of *Nasutitermes* species, it was observed that bacteria belonging to these two groups or *Treponema* adhered onto wood fibers, and a high cellulase activity was detected in this wood-fiber fraction (Mikaelyan *et al*., 2014). In metagenomic studies of wood-feeding higher termites, many genes involved in lignocellulose digestion were reported, and members of TG3, Fibrobacteres, and *Treponema* are predicted to possess the majority of the genes for the digestion (Warnecke *et al*., 2007; Abdul Rahman *et al*., 2015; He *et al*., 2013). However, information on individual bacteria species in the higher termite gut is still limited and their roles in the symbioses remain unclear. Therefore, the fundamental question why the termite gut microbial community is so complex remains unanswered. Recently, genomes of members of TG3 and Fibrobacteres were reconstructed from metagenomic sequences of the gut communities of the higher termites *Nasutitermes* sp. and *Microcerotermes* sp., respectively, and based on the genome-wide phylogenetic analyses, TG3 is proposed to be assigned to the phylum Fibrobacteres (Abdul Rahman *et al*., 2015). However, the genome of *Treponema* species, a member of the other predominant group in the gut of higher termites, has not yet been reconstructed. Furthermore, genome sequences of multiple individual species in the gut community of a single higher termite are needed in order to understand functional interactions among the community members. For the predicted functions based on genome analyses, it is important to evaluate contributions of individual bacteria in the gut community.

In this study, in order to elucidate the functions of individual symbiotic bacteria in the higher termite gut, we analysed the genomes of three dominant species, each representing the major bacterial groups TG3, Fibrobacteres (subphylum 2), and *Treponema* in the gut of *N. takasagoensis* by single-cell genome sequencing. To clarify the contributions of these individual bacteria and their related species to this symbiotic system, we also analysed the native expression levels of their genes involved in symbiotic functions in the gut community by RNA-seq. We report partially overlapping but distinct roles of these dominant bacterial species in crucial aspects of the symbioses, e.g.

utilisation of lignocellulose, reductive acetogenesis, and nitrogen fixation.

Materials and Methods

Sample collection, single-cell sorting, and whole genome amplification

The wood-feeding higher termite *N. takasagoensis* was collected from Iriomote Island, Okinawa Prefecture, Japan. The gut of a worker termite was removed using sterile forceps, and the gut contents were suspended in Solution U (Trager 1934). The sample was centrifuged, washed, filtered and then stained using CellTracker™ Green CMFDA (Life Technologies) as previously described (Yuki *et al*., 2015). The single-cell sorting, multiple displacement amplification (MDA), phylogenetic identification, and checking the degree of contamination were conducted as previously described[8].

Genome sequencing and assembly

For genome sequencing of the selected single-cell MDA sample, second MDA and treatment of the amplified genome were performed as described previously (Yuki *et al*., 2015). The paired-end library and mate-pair library were prepared using the TruSeq DNA Sample Prep kit (Illumina) and the Nextera Mate Pair Sample Prep kit (Illumina), respectively. Genome sequencing was performed on an Illumina MiSeq using the Illumina Reagent Kit V3 (600 cycles). The generated reads were assembled using the program SPAdes 3.0.0 (Bankevich *et al*., 2012).

Whole genome amplification using specific-primers

In order to increase the genome completeness in the cases of the TG3 R8-0-B4 and Fibrobacteres R8-3-H12 samples, we conducted whole genome amplification using specific-primers instead of the standard random hexamer primers accompanying the reagent kit. The specific-primers of 20 mers on average were designed from 500 bp upstream from the 3′-end of the assembled contigs of top 100 in the length and all primers were phosphorothioated for protecting the 3′–5′ exonuclease activity of Phi29 DNA polymerase. To remove the random hexamer, the first MDA samples were purified using the MultiScreen HTS (Millipore). The purified first MDA samples were denatured at 95℃ for 3 minutes then put on ice. The first MDA samples were re-amplified with

RepliPHI™ Phi29 DNA polymerase (Epicentre) using a mixture of the specific-primers instead of random hexamers. Library construction and sequencing were conducted as described above. The generated reads were combined with those obtained by the first MDA using random hexamers and assembled using SPAdes.

Genome sequence analyses

To remove suspected contaminating sequences, the blobology bash (Kumar *et al*., 2013) was used to classify contigs based on GC content, coverage and putatively assigned taxonomy; this allowed for visualisation in R (https://www.R-project.org) and subsequent contig removal. Gene predictions were carried out using Prokka (Seemann., 2014), uploaded to the RAST server (Aziz et al., 2008), and the results were subjected to BLAST searches (BlastP) against the NCBI non-redundant database (NCBI-nr), and CAZy database, for manual curation. The predicted genes were assigned to functional categories, using the non-supervised orthologous groups database from eggNOG (Powell *et al*., 2014). Genome completeness was estimated based on the presence of single-copy genes common in bacteria (Rinke *et al*., 2013) and estimated genome size was calculated based on this value.

Phylogenetic analysis

To determine the phylogenetic positions of the genome-sequenced TG3 and Fibrobacteres isolates, a maximum likelihood (ML) tree was constructed based on concatenated amino acid sequences of 16 ribosomal proteins (RplA, RplC, RplD, RplE, RplJ, RplK, RplN, RplO, RplW, RpsG, RpsJ, RpsK, RpsL, RpsM, RpsO and RpsQ) using RAxML (Stamatakis., 2006) with bootstrap analysis (500 re-samplings). Phylogenetic analyses based on 16S rRNA gene sequences were also conducted to determine their placement within the previously defined termite-specific clusters of TG3, Fibrobacteres and *Treponema* in the DictDB database (Mikaelyan *et al*., 2015). The 16S rRNA gene sequences were aligned using MUSCLE (Edgar., 2004) and Neighbor-joining (NJ) trees were constructed using MEGA 5.0 (Tamura *et al*., 2011) with bootstrap analyses (500 re-samplings).

For the phylogenetic analyses of other protein-coding genes, deduced amino acid sequences were aligned using MUSCLE, concatenated, and ML trees were constructed

using RAxML with bootstrap analyses (1,000 re-samples).

RNA-seq

The whole-gut content of *N. takasagoensis* (10 workers) were dissected into RNAlater (Thermo Fisher Scientific) at the sampling area to protect the RNA. In the laboratory, the samples were washed, filtered and RNA extracted using the RNAeasy kit (Qiagen). The Agilent bioanalyser 2100 was used to assess the quality of the extracted RNA comparatively against RNA integrity numbers and total RNA was quantified using the Qubit fluorometer (Thermo Fisher Scientific).

The ScriptSeq complete gold kit (Epidemiology) was used to construct stranded libraries and RiboZero (Illumina) was used to reduce the rRNA in samples. Sequencing was conducted on the Illumina MiSeq platform using Reagent Kit V3 and the TruSeq HT protocol. Reads were quality- and adapter-trimmed and assembled using the Trinity pipeline (Haas *et al*., 2013). Quality-trimmed reads were also mapped to the respective single-cell genome assemblies, using bowtie2 (Langmead & Saizberg., 2012). The mapped reads were counted with HTSeq-count (Anders *et al*., 2015) and the results were normalised to generate the reads per kilobase of transcript per million mapped reads (RPKM) of the genes. The glycoside hydrolase genes, nitrogenase genes and genes involved in reductive acetogenesis and folate biosynthesis were selected from the Trinity-assembled sequences, using DIAMOND (Buchfink *et al*., 2014; BlastX alignment mode) with the NCBI, CAZy protein databases and the genes of the TG3, Fibrobacteres and *Treponema* single-cell genomes obtained in this study (e-value; < 1e-50).

Results

General features of the TG3, Fibrobacteres and Treponema *single-cell genomes*

We examined genome sequences of several representative single-cell isolates in each of the three predominant bacterial groups and selected single-cell isolates, TG3 (R8-0-B4), Fibrobacteres (R8-3-H12), and *Treponema* (R8-4-B8). Because the N50 value of the contig size and the largest contig length were considerably smaller in the TG3 and Fibrobacteres single-cell genomes than in the *Treponema* genome (Table 1), we further sequenced re-amplified genomic DNA of the TG3 and Fibrobacteres single-cell isolates with custom specific primers, designed based on the first genome assemblies, instead of standard random hexamer primers (for details, see Materials and Methods). The co-assembly of sequences obtained from both genomic DNA samples amplified with random hexamers and specific primers improved total length of the assembled contigs and completeness of the genome, estimated using a single-copy gene set (Rinke *et al*., 2013) in each sample (Table 1, Figure 1). We selected confident contigs using the blobology bash (Kumar *et al*., 2013), and finally obtained genome sequences that showed the genome completeness ranging from 77.0% to 84.9% (Table 2).

Phylogenetic analysis based on the concatenated sequence of 16 ribosomal proteins revealed that TG3 R8-0-B4 was sister to *Chitinispirillum alkaliphilum* and *Chitinivibrio alkaliphilus*, and Fibrobacteres R8-3-H12 was sister to *Fibrobacter succinogenes* (Figure 2a). These five bacteria formed a monophyletic clade, supporting the previous reports that the candidate phylum TG3 is phylogenetically related to and thus should be assigned to the phylum Fibrobacteres (Hongoh *et al*., 2006; Sorokin *et al*., 2014; Sorokin *et al*., 2016; Abdul Rahman *et al*., 2015). The phylogenetic trees based on the 16S rRNA gene revealed that TG3 R8-0-B4, Fibrobacteres R8-3-H12, and *Treponema* R8-4-B8 belong to three dominant clusters in the termite gut, Termite cluster IIa, Termite cluster Ib, and *Treponema* Ic, respectively (Fig. 2b–d, Figure 4), and these three single-cell isolates corresponded to major OTUs (operational taxonomic units, defined with 5% sequence divergence) of 17.5%, 7.0%, and 10.5% abundance, respectively, in the clone analysis of bacterial 16S rRNA gene sequences in the same termite species (Hongoh *et al*., 2006). Mikaelyan *et al*. 2014 reported that bacteria belonging to these three clusters bind to wood particles and thus likely play crucial roles in lignocellulose decomposition.

Table 1. Summary of the assembly of single-cell genomes.

Sample	WGA	Number of Contigs	Total length (bp)	Largest Contig (bp)	N50 (bp)	Completeness
TG3 R8-0-B4	random*[1]	728	3,289,882	172,115	28,320	79.9%
	random, random and specific*[2]	580	3,644,966	156,705	44,408	86.3%
Fibrobacteres R8-3-H12	random*[1]	978	3,208,886	147,834	41,252	77.0%
	random, random and specific*[2]	1025	3,420,509	175,722	32,772	89.2%
Treponema R8-4-B8	random*[1]	669	3,775,198	347,258	80,331	84.2%

The assembly data shown in this table are those before removing suspected contaminating sequences.
*1 The data from paired-end and mate-pair libraries of the WGA sample with random hexamers were used for genome assembly. *2 In addition, the data from a paired-end library of the WGA sample with specific primers were also used for genome assembly.

Table 2. Comparisons of genome features of the TG3, Fibrobacteres and *Treponema* single-cell isolates with related cultured bacteria.

*1. Size of the draft genome. Estimated genome size is shown in parentheses.

Species	Habitat	Number of contigs	Total length	GC content	tRNAs	CDS	Completeness	
TG3 R8-0-B4	*Nasutitermes* gut	536	3.58 Mb[*1] (4.22 Mb)	53.4%	39	3,264	84.9%	This study
Chitinivibrio alkaliphilus	Hypersaline soda lake	-	2.59 Mb	46.2%	39	2,304	-	Sorokin *et al*., 2014
Fibrobacteres R8-3-H12	*Nasutitermes* gut	724	3.07 Mb[*1] (3.99 Mb)	41.9%	39	3,146	77.0%	This study
Fibrobacter succinogenes S85	The rumen of herbivores	-	3.84 Mb	48.0%	58	3,159	-	Suen *et al*., 2011
Treponema R8-4-B8	*Nasutitermes* gut	529	3.62 Mb[*1] (4.41 Mb)	42.2%	42	3,236	82.0%	This study
Treponema azotonutricium ZAS-9	*Zootermopsis* gut	-	3.86 Mb	49.8%	47	3,474	-	Rosenthal *et al*., 2011
Treponema primitia ZAS-2	*Zootermopsis* gut	-	4.06 Mb	50.8%	50	3,523	-	Rosenthal *et al*., 2011

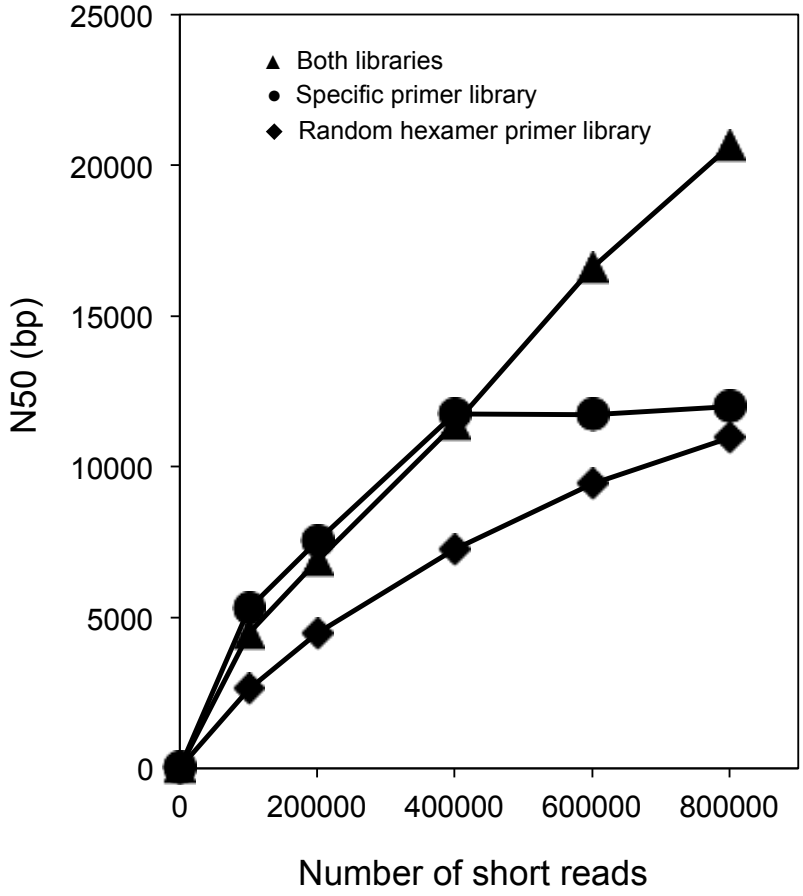

Figure 1. *De novo* genome assembly of the TG3 R8-0-B4 genome using WGA samples with specific primers and random hexamers. The short sequence reads with paired-end libraries in fixed numbers (100,000 to 800,000) were randomly selected from each WGA sample and both samples and assembled using SPAdes. The vertical axis indicates the N50 of assembled contigs. The horizontal axis indicates the number of short reads used for assembly. Note that the assembly was greatly improved when the sequence reads of both samples were combined and assembled together.

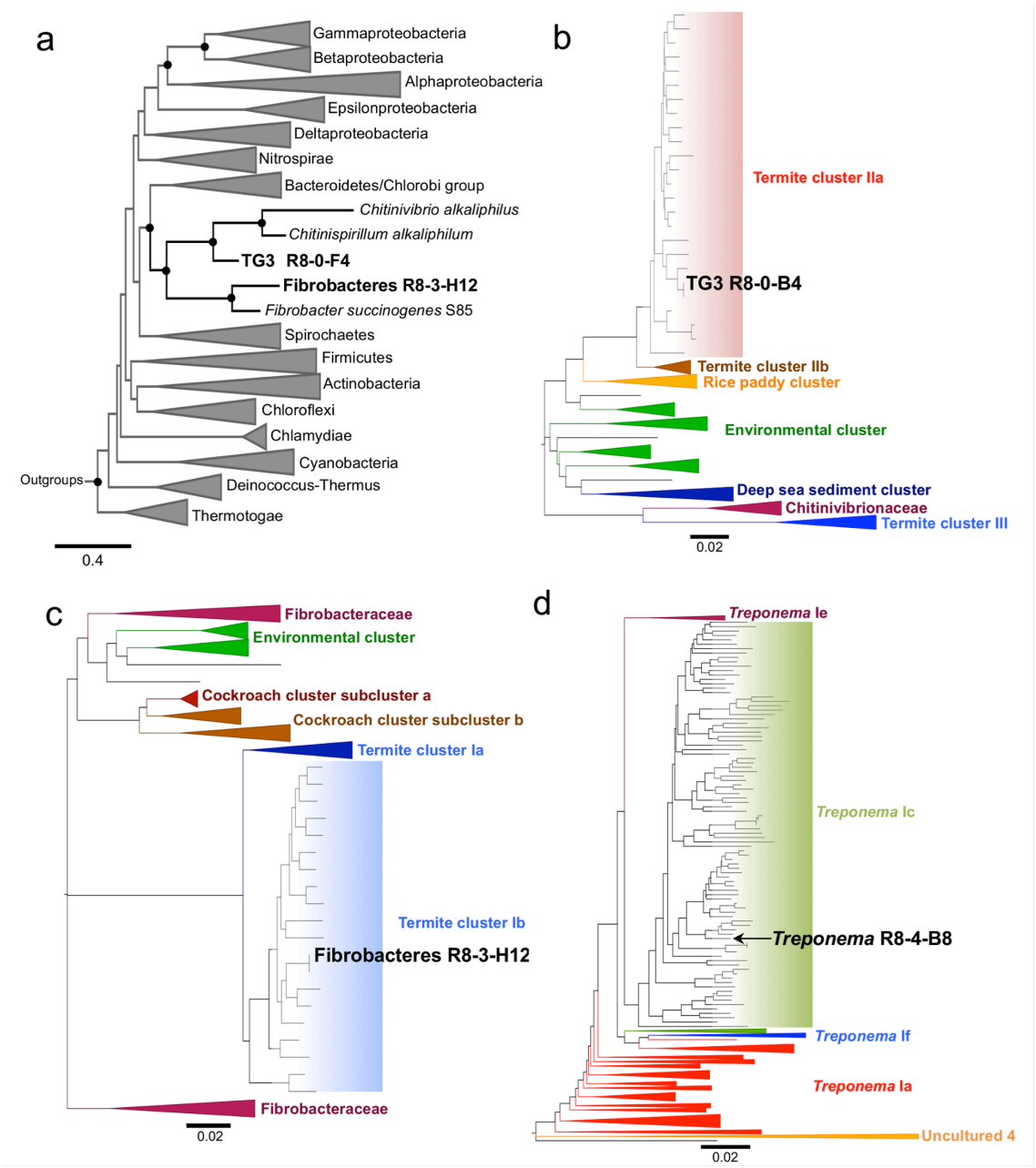

Figure 2. Phylogenetic position of single-cell genome sequencing TG3, Fibrobacteres and *Treponema* bacteria. (a) A maximum likelihood tree showing phylogenetic positions of bacteria represented by TG3 R8-0-B4 and Fiborobacteres R8-3-H12. The tree was inferred based on the concatenated sequence of 16 ribosomal proteins. Bootstrap values of 90% to 100% are indicated with closed circles. Archaeal sequences were used as outgroups. (b-d) Neighbor-joining trees based on the 16S rRNA gene sequences showing the positions of bacteria represented by (b) TG3 R8-0-B4, (c) Fibrobacteres R8-3-H12, and (d) *Treponema* R8-4-B8 in the phylogenetic clusters comprised of termite-gut symbionts. The taxa in each termite cluster are indicated in Figure 3. Scales correspond to 0.5 (a) or 0.02 (b-d) substitutions per aligned sequence site.

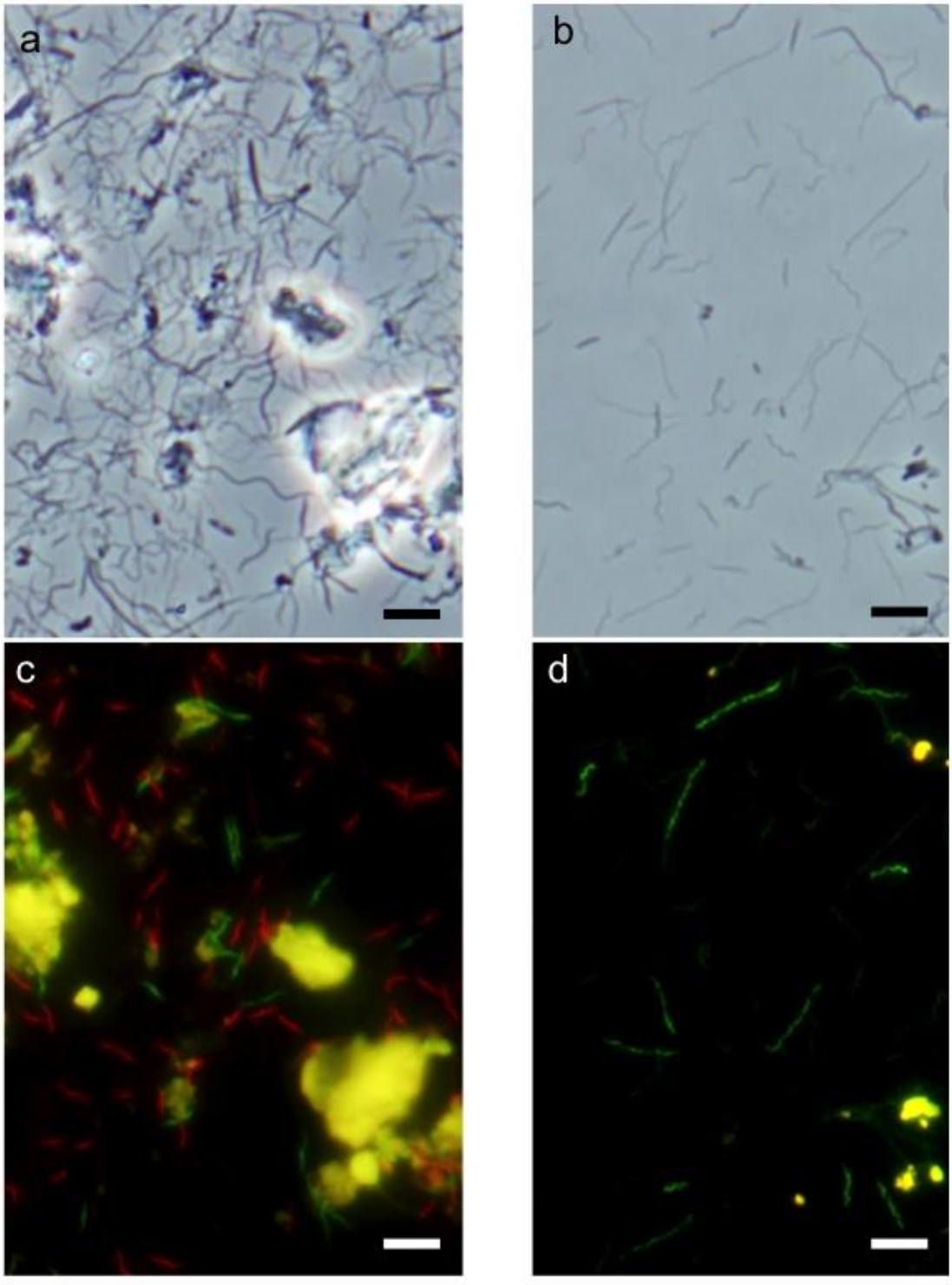

Figure 3. Bacterial microbiota in the gut of *N. takasagoensis*. Phase-contrast microscopy of gut bacteria in (a, b). Fluorescence *in situ* hybridizations using Fibrobacteres-specific (Texas Red) and TG3-specific (6FAM, green) probes described previously (Hongoh *et al*., 2006) (c), and sequence specific probes for the Treponema R8-4-B8 and its related species (6FAM: 5′-aacagctatccccatcct, designed in the present study) (d). Fluorescence *in situ* hybridizations were performed as described previously (Hongoh *et al*., 2006). Bar = 5 μl.

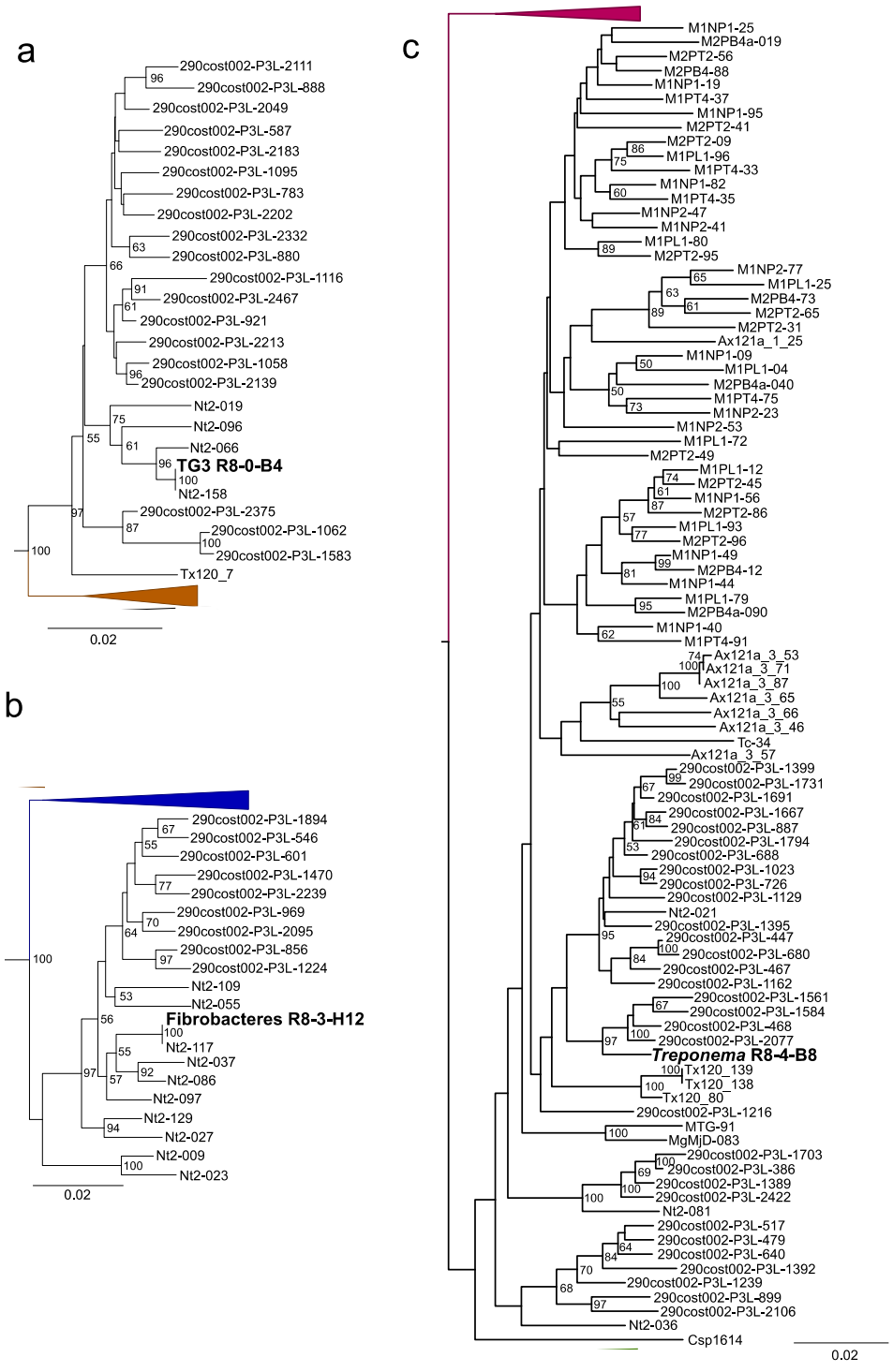

Figure 4. Neighbor-joining trees based on the16S rRNA gene sequences showing relationship of TG3 R8-0-B4 with termite-gut symbionts in the Termite cluster IIa (a), that of Fibrobacteres R8-3-H12 in the Termite cluster IIB (b), and that of *Treponema* R8-4-B8 in the *Treponema* Ic cluster (c). The trees correspond to those shown in Figure 2 b-d in the main text, respectively. Bootstrap values in percentages are indicated at the nodes when the values were over 50%.

Lignocellulose-digesting system

The potential roles of these three dominant bacterial species in lignocellulose digestion were assessed by screening the putative protein-coding sequences (CDSs) against the Carbohydrate-Active enZYmes (CAZy) database (http://www.cazy.org). The TG3 R8-0-B4, Fibrobacteres R8-3-H12 and *Treponema* R8-4-B8 genomes possess at least 20, 24 and 26 genes coding for lignocellulose degrading enzymes, and they were classified into 13, 11 and 14 different GH families, respectively (Figure 5a, Table 3). There were little differences in the frequencies of the number of GH genes among the TG3 (5.58/Mb), Fibrobacteres (7.82/Mb), and *Treponema* (7.19/Mb) single-cell genomes; however the ratio of cellulolytic to hemicellulolytic enzyme genes per genome were quite different, i.e. 0.82, 3.00 and 0.37, respectively (Figure 5a).

We performed RNA-seq analysis of the entire gut microbial community and assessed expression levels of genes of the three dominant bacteria and their closely related species. As a result, the expression levels of GH genes of TG3, Fibrobacteres and *Treponema* bacteria measured as sum totals of reads per gene length were 2.16, 1.73 and 0.36, respectively (Figure 5b). The ratios of gene expression levels of cellulolytic to hemicellulolytic enzymes in the three bacterial groups were similar to the ratios of the number of corresponding genes in the three single-cell genomes (compare Figures 5a and 5b).

The TG3 R8-0-B4 genome possesses genes for one beta-glucosidase (GH1) and six endoglucanases (GH5 and GH9) as cellulolytic enzymes and additionally 11 genes for hemicellulolytic enzymes, e.g. endo-beta-mannanase (GH5), alpha-N-arabinofuranosidase (GH51), endo-1,4-xylanase (GH10 and GH11) and xyloglucanase (GH74) (Table 3). Almost all of the GH10 xylanase genes expressed in the gut community were closely related to the TG3 R8-0-B4 gene, but the total expression level of the GH10 xylanase genes was lower than that of the GH11 xylanase genes (Figure 6). The majority of expressed GH11 xylanase genes were homologous to the genes of Firmicutes bacteria.

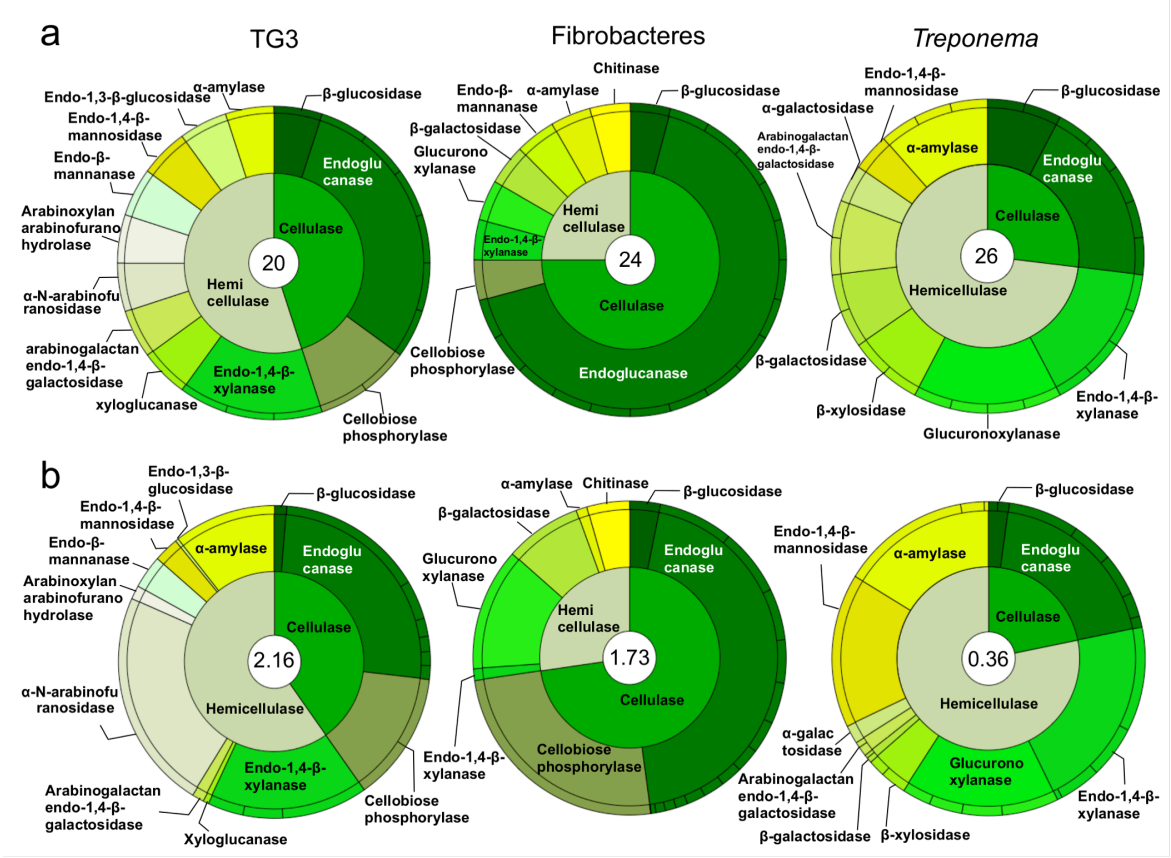

Figure 5. The glycoside hydrolase genes of the TG3, Fibrobacteres and *Treponema* genomes and the expression levels of the related genes in the gut community. (a) The ratio of cellulase and hemicellulase genes present in each genome and the ratio of the predicted encoding enzymes are shown in inside and outside circles in each pie chart, respectively. The number in the centre of each pie chart indicates the total number of glycoside hydrolase genes. (b) The ratios of expressed cellulase and hemicellulase genes and their encoding enzymes are shown in inside and outside circles in each pie chart, respectively. The expression levels of these genes were calculated by the number of short reads per gene length. The number in the centre of each pie chart indicates the sum of the numbers of short reads per length of all genes.

The Fibrobacteres R8-3-H12 genome encoded a total of 16 genes for endoglucanases in four different GH families (GH5, 8, 9 and 45), with GH8 and GH45 found exclusively in this genome. Almost all the expressed genes for endoglucanases of GH8 and GH45 families were closely related to the Fibrobacteres R8-3-H12 genes. The expression of the Fibrobacteres GH5 and GH9 endoglucanase genes was approximately the same level as that of the TG3 genes (Figure 6). On the other hand, the number of genes encoding hemicellulolytic enzymes (six genes) was the least amongst the three single-cell genomes. Genes encoding cellobiose phosphorylase, which catalyses phosphorolysis of cellobiose into glucose-1-phosphate and glucose, were found in both TG3 and Fibrobacteres single-cell genomes.

The *Treponema* R8-4-B8 genome encoded five genes for GH5 endoglucanases belonging to subfamily 4. Members of this subfamily are assigned to xyloglucan-specific endoglucanases that cleave xyloglucans (Aspeborg *et al*., 2012). The RNA-seq results showed that 83% of the expressed genes of this subfamily were closely related to the *Treponema* R8-4-B8 genes (Figure 6). This *Treponema* genome possessed 19 genes for hemicellulolytic enzymes classified into 12 families. Genes for beta-xylosidases, which cleave xylose, and a gene for alpha-galactosidase, which cleaves galactose, were exclusively found in this *Treponema* genome (Table 3). In the functional categorisation analysis of gene content (Figure 7), the number of genes related to 'carbohydrate transport and metabolism' in this genome was higher than the other two genomes.

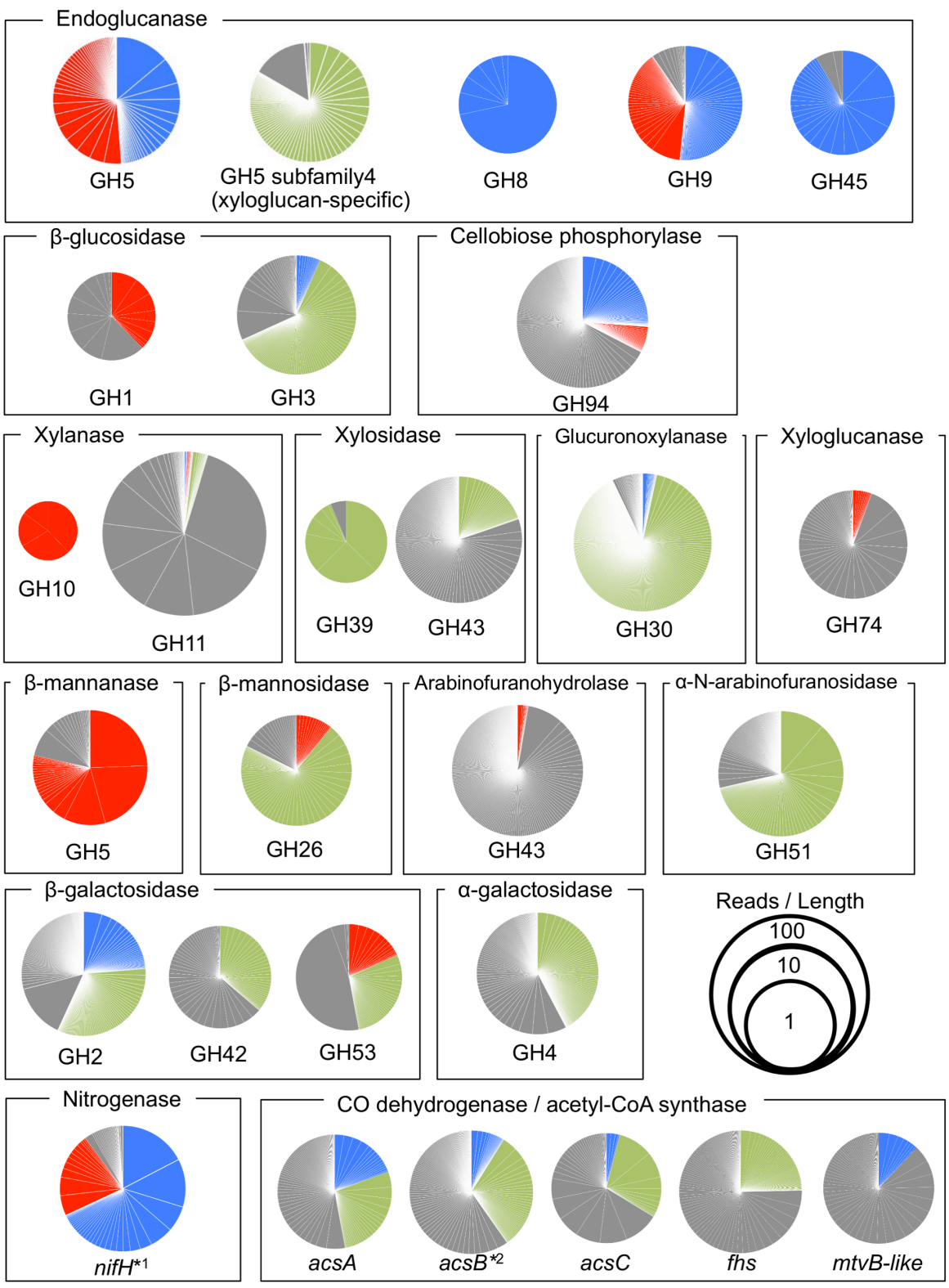

Figure 6. Gene expression levels of glycoside hydrolases, nitrogenase and enzymes for reductive acetogenesis in the gut community. The gene expression level is shown as the number of sequence reads per gene length. The pie charts indicate total gene expression in the termite gut. The size of the pie chart is proportional to the gene expression level. The colours indicate the expressed gene homologous to the genes encoded by the TG3 (red), Fibrobacteres (blue), *Treponema* (green) genomes and other bacteria (grey). *[1] The full-length *nifH* gene found in another TG3 single-cell genome R8-4-B1 that showed 100 % sequence identity of the 16S rRNA gene to the TG3 R8-0-B4 was used because the use of the fragmented *nifH* gene in the TG3 R8-0-B4 genome underestimated its expression levels. *[2] *acsB* and *acsC* of another *Treponema* single-cell genome R80B11-R83G3 showing 99.7 % sequence identity of the 16S rRNA gene to the *Treponema* R8-4-B8 was used because of the fragmentation of these genes in the *Treponema* R8-4-B8 genome

Table 3. Glycoside hydrolase families encoded by the TG3 R8-0-B4, Fibrobacteres R8-3-H12, and *Treponema* R8-4-B8 genomes.

Cellulase		TG3	Fibrobacteres	*Treponema*
GH1	Beta-glucosidase	1	0	0
GH3	Periplasmic beta-glucosidase precursor	0	1	2
GH5	Endoglucanase	2	6	5
GH8	Endoglucanase	0	3	0
GH9	Endoglucanase	4	3	0
GH45	Endoglucanase	0	4	0
GH94	Cellobiose phosphorylase	2	1	0
	Total	9	18	7
Hemicellulase		TG3	Fibrobacteres	*Treponema*
GH2	Beta-galactosidase	0	1	1
GH4	Alpha-galactosidase	0	0	1
GH5	Endo-beta-mannanase	1	1	0
GH10	Endo-1,4-beta-xylanase	1	0	3
GH11	Endo-1,4-beta-xylanase	2	1	1
GH13	Alpha-amylase	1	0	2
GH16	Glucan endo-1,3-beta-glucosidase	1	0	0
GH18	Chitinase	0	1	0
GH26	Mannan endo-1,4-beta-mannosidase	1	0	1
GH30	Glucuronoxylanase xynC	0	1	4
GH39	Beta-xylosidase	0	0	1
GH42	Beta-galactosidase	0	0	1
GH43	Arabinoxylan arabinofuranohydrolase	1	0	0
GH43	Beta-xylosidase	0	0	1
GH51	Alpha-N-arabinofuranosidase	1	0	0
GH53	Arabinogalactan endo-1,4-beta-galactosidase	1	0	2
GH57	Alpha-amylase	0	1	1
GH74	Xyloglucanase	1	0	0
	Total	11	6	19

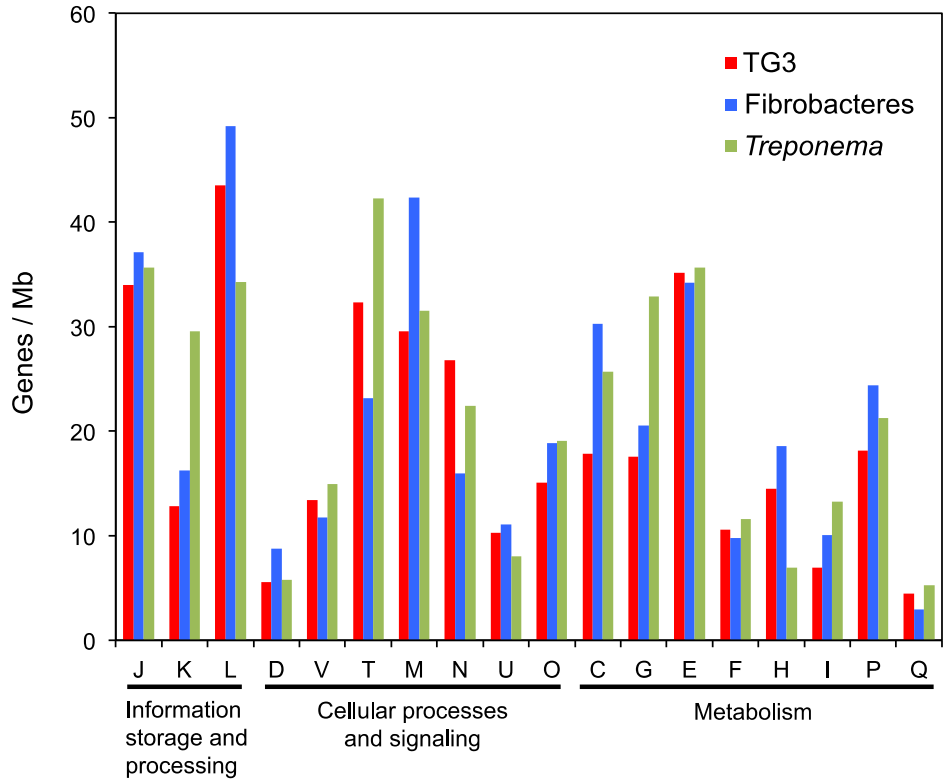

Figure 7. Comparisons of the TG3 R8-0-B4, Fibrobacteres R8-3-H12 and *Treponema* R8-4-B8 genomes based on gene functional categories. Abbreviations of the functional categories are: J, Translation, ribosomal structure and biogenesis; K, Transcription; L, Replication, recombination and repair; D, Cell cycle control, cell division, chromosome partitioning; V, Defense mechanism; T, Signal transduction mechanisms; M, Cell wall/membrane/envelope biogenesis; N, Cell motility; U, Intracellular trafficking, secretion, and vesicular transport; O, Posttranslational modification, protein turnover, chaperones; C, Energy production and conversion; G, Carbohydrate transport and metabolism; E, Amino acid transport and metabolism; F, Nucleotide transport and metabolism; H, Coenzyme transport and metabolism; I, Lipid transport and metabolism; P, Inorganic ion transport and metabolism; Q, Secondary metabolism biosynthesis, transport and catabolism.

Energy metabolism

The predicted metabolic pathways based on the TG3, Fibrobacteres, and *Treponema* single-cell genomes are shown in Figure 8 and Table 4. Most genes for the glycolytic pathway were present in the Fibrobacteres R8-3-H12 genome, while we could not find several genes in the TG3 R8-0-B4 and *Treponema* R8-4-B8 genomes. In accordance with the presence of a cellobiose phosphorylase gene, a phosphoglucomutase gene responsible for utilisation of glucose-1-phosphate was detected in these TG3 and Fibrobacteres genomes, respectively. Genes for the tricarboxylic acid cycle were only partially found in all the three genomes. The presence of genes for sugar transporters in all the three genomes indicates their potential to import mono- and oligosaccharides derived from lignocellulose (Table 3). These TG3 and *Treponema* genomes further encoded the genes for phosphotransferase (PTS) system IIA components.

Interestingly, the genes encoding xylose isomerase and xylulokinase, which are required for the utilisation of D-xylose derived from xylan, were identified only in the *Treponema* R8-4-B8 genome (Figure 8), which also possessed genes for the non-oxidative pentose phosphate pathway that metabolises the product D-xylulose 5-phosphate into the glycolytic pathway. In accordance, this *Treponema* genome possessed at least two xylose transporter genes, strongly suggesting the utilisation of xylose, while no genes for xylose transporters were identified in the other two genomes. The differences in the utilisation of monosaccharaides derived from hemicellulose were further predicted from these genomes. The *Treponema* R8-4-B8 genome possessed genes involved in the utilisation of other sugars such as galactose, mannose, and fructose. The TG3 R8-0-B4 genome also possessed genes involved in the mannose metabolic pathway except for hexokinase, but a PTS system may replace the hexokinase function. In contrast, the Fibrobacteres R8-3-H12 genome did not encode any of these genes. In addition, the *Treponema* R8-4-B8 genome possessed genes for the utilisation of D-glucuronate to produce pyruvate via the Entner-Doudoroff pathway (Figure 8). The expressions of these identified genes in the gut community were confirmed in the RNA-seq analysis (Figure 8). All three genomes revealed the potential to ferment these monosaccharaides to acetate.

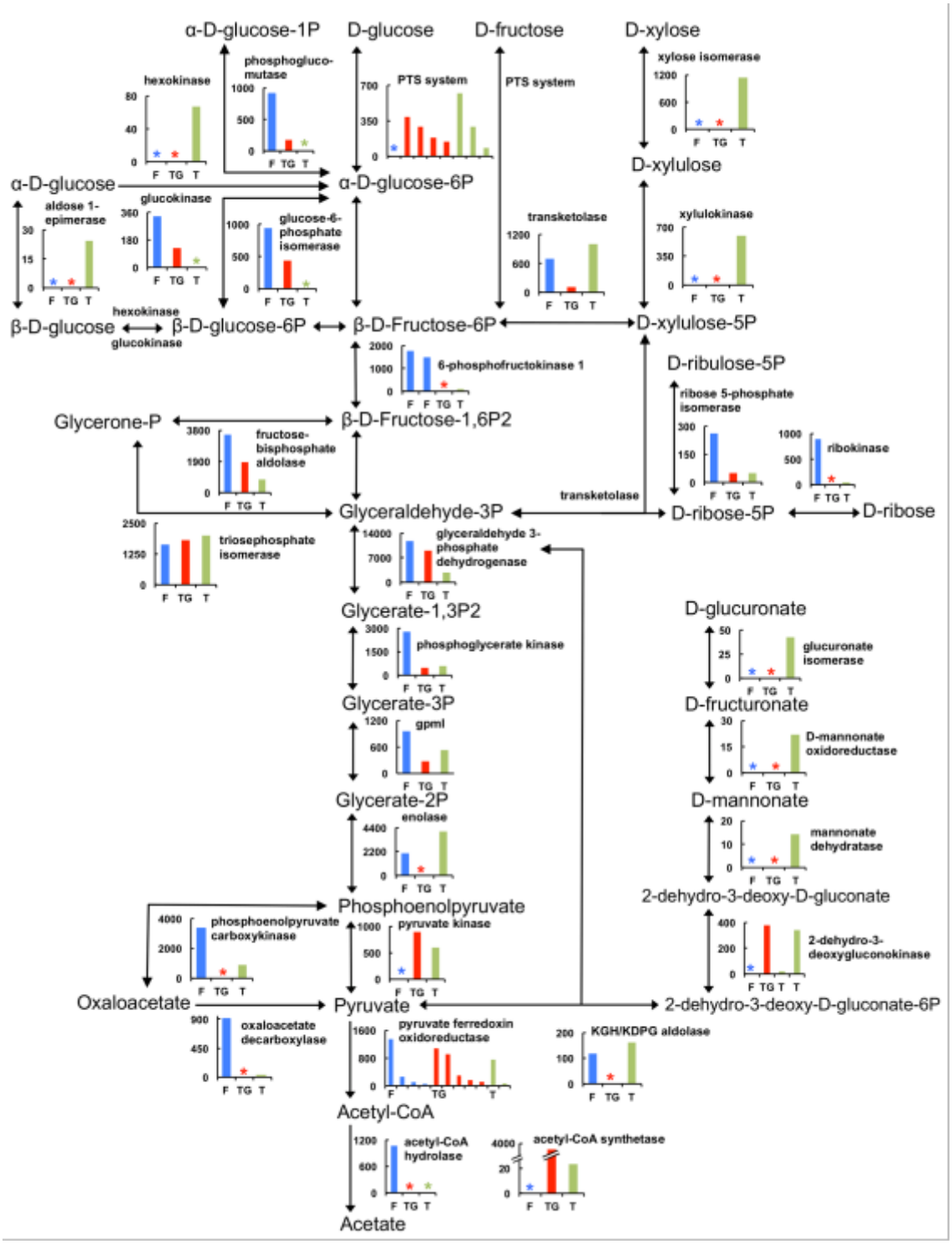

Figure 8. Central metabolic pathways of the dominant TG3, Fibrobacter and *Treponema* bacterial species and their gene expression levels in each step of the pathways. The gene expression levels are shown as RPKMs in the RNA-seq data. The vertical axis in each graph represents RPKM. F: Fibrobacteres R8-3-H12 (blue), TG: TG3 R8-0-B4 (red), and T: *Treponema* R8-4-B8 (green). The asterisks in the graph indicate that the genes were not found in the genomes studied here.

Table 4. Presence of genes for key metabolic functions in the TG3 R8-0-B4, Fibrobacteres R8-3-H12, and *Treponema* R8-4-B8 genomes

	TG3	Fibrobacteres	*Treponema*
Glycolysis			
PTS system (EC 2.7.1.69)	+	-	+
Aldose 1-epimerase (EC 5.1.3.3)	-	-	+
Hexokinase (EC 2.7.1.1)	-	-	+
Glucokinase (EC 2.7.1.2)	+	+	-
Glucose-6-phosphate isomerase (EC 5.3.1.9)	+	+	-
6-phosphofructokinase 1 (EC 2.7.1.11)	-	+	+
Fructose-1,6-bisphosphatase I (EC 3.1.3.11)	+	-	-
Fructose-bisphosphate aldolase, class I (EC 4.1.2.13)	+	+	+
Triosephosphate isomerase (EC 5.3.1.1)	+	+	+
Glyceraldehyde 3-phosphate dehydrogenase (EC 1.2.1.12)	+	+	+
Phosphoglycerate kinase (EC 2.7.2.3)	+	+	+
2,3-bisphosphoglycerate-independent phosphoglycerate mutase (EC 5.4.2.12)	+	+	+
Enolase (EC 4.2.1.11)	-	+	+
Pyruvate kinase (EC 2.7.1.40)	+	-	+
Pentose and glucuronate interconversions, Pentose phosphate pathway			
Xylose isomerase (EC 5.3.1.5)	-	-	+
Xylulokinase (EC 2.7.1.17)	-	-	+
Transketolase (EC 2.2.1.1)	+	+	+
Ribulose-phosphate 3-epimerase (EC 5.1.3.1)	-	-	+
Ribose 5-phosphate isomerase A (EC 5.3.1.6)	+	+	+
Ribokinase (EC 2.7.1.15)	-	+	+
Deoxyribose-phosphate aldolase (EC 4.1.2.4)	-	-	+
Pyruvate metabolism			
Pyruvate-flavodoxin oxidoreductase (EC 1.2.7.-)	+	+	+
Acetyl-CoA synthetase (EC 6.2.1.1)	+	-	+
Acetyl-CoA hydrolase (EC 3.1.2.1)	-	+	-
Phosphate acetyltransferase (EC 2.3.1.8)	-	-	-
Acetate kinase (EC 2.7.2.1)	-	+	-
Acylphosphatase (EC 3.6.1.7)	+	-	-
Phosphoenolpyruvate carboxykinase [GTP] (EC 4.1.1.32)	-	+	+
Malate dehydrogenase (EC 1.1.1.37)	-	+	+
NADP-dependent malic enzyme (EC 1.1.1.40)	+	-	+
Oxaloacetate decarboxylase (EC 4.1.1.3)	-	+	+
Alcohol dehydrogenase (EC 1.1.1.1); Acetaldehyde dehydrogenase (EC 1.2.1.10)	+	-	+
Lactate dehydrogenase (EC 1.1.1.27)	-	-	-
Sugar and uronate utilisation			
Mannose-6-phosphate isomerase (EC 5.3.1.8)	+	-	+
4-Alpha-glucanotransferase (amylomaltase) (EC 2.4.1.25)	+	+	+
Trehalase (EC 3.2.1.28)	-	-	+
Uronate isomerase (EC 5.3.1.12)	-	-	+

D-mannonate oxidoreductase (EC 1.1.1.57)	-	-	+
Mannonate dehydratase (EC 4.2.1.8)	-	-	+
2-dehydro-3-deoxygluconokinase (EC 2.7.1.45)	+	-	+
2-dehydro-3-deoxyphosphogluconate aldolase (EC 4.1.2.14)	-	+	+
Glycogen metabolism			
Glucose-1-phosphate adenylyltransferase (EC 2.7.7.27)	+	-	+
Glycogen synthase, ADP-glucose transglucosylase (EC 2.4.1.21)	-	+	+
Glycogen phosphorylase (EC 2.4.1.1)	+	+	+
Glycogen branching enzyme, GH-57-type, archaeal (EC 2.4.1.18)	+	+	+
Phosphoglucomutase (EC 5.4.2.2)	+	+	-
Sugar transport			
Putative D-xylose transporter	-	-	+
Putative maltose/maltodextrin transporter	-	-	+
Sugar transporter (not assigned substrate)	+	+	+

Reductive acetogenesis

We assessed the potential for reductive acetogenesis by searching the genes involved in the acetyl-CoA (Wood-Ljungdahl) pathway in the single-cell genomes (Figure 9). The key enzyme of this pathway is the acetyl-CoA synthase/carbon monoxide dehydrogenase complex (ACS/CODH) encoded by *acsA*, *acsB*, *acsC acsD*, and *acsF*. The Fibrobacteres R8-3-H12 and *Treponema* R8-4-B8 genomes possessed these genes, except for *acsB* in the *Treponema* genome, although another single-cell genome of *Treponema* R80B11-R83G3 showing 16S rRNA gene sequence of 99.7% identity to *Treponema* R8-4-B8 possessed *ascB*. However, none of these genes were present in the TG3 R8-0-B4 genome. Interestingly, the *acsB* gene encoding acetyl-CoA synthase subunit in these Fibrobacteres and *Treponema* genomes were fused in frame with the *acsD* gene encoding small subunit of Fe-S-Co protein, implying that the gene encodes a bifunctional enzyme. The expression levels of these genes of *Treponema* were higher than those of Fibrobacteres (Figure 6 and Figure 9). The phylogenetic analysis based on concatenated sequences of five ACS/CODH subunits indicated that the ACS/CODH of Fibrobacteres R8-3-H12 and *Treponema* R8-4-B8 were closely related to each other and somewhat to *Desulfonatronospira thiodismutans* (Figure 11a).

Traditional reductive acetogenesis utilises formate dehydrogenase (FDH) to synthesise formate from hydrogen and CO_2, the first step of the methyl branch in the Wood-Ljungdahl pathway (Figure 9). We were unable to find *fdh* genes in these single-cell

genomes; however, a few *fdh* gene sequences were detected in the RNA-seq data and they were primarily homologous to those in Clostridia and other *Treponema* species. The gene encoding pyruvate formate-lyase (*pflD*) was not found in these genomes, although the gene for pyruvate formate-lyase activating enzyme (*pflA*) was detected in the *Treponema* R8-4-B8 genome. The gene of formate-tetrahydrofolate ligase (*fhs*) was present in this *Treponema* genome but not in the Fibrobacteres R8-3-H12 and TG3 R8-0-B4 genomes. The gene for methylene-tetrahydrofolate reductase (*metF*) was found in the TG3 genome but not in the Fibrobacteres and *Treponema* genomes; however, another single-cell genome of *Treponema* R6D11 showing 16S rRNA gene sequence of 96% identity to *Treponema* R8-4-B8 possessed *metF*. The gene for 5-methyltetrahydrofolate corrinoid/iron sulfur protein methyltransferase (*acsE*) was not found in the three genomes; however, *acsE* genes were found in the RNA-seq data and about 85% of expressed *acsE* genes were homologous to *acsE* genes of cultured *Treponema* species (Figure 9). These results indicate that the typical methyl branch of Wood-Ljungdahl pathway of the *Treponema* is potentially functional but that of Fibrobacteres is likely incomplete.

Many homoacetogens utilise *O*-methyl group of methoxylated aromatic compounds, such as vanillate and syringate, as a methyl-group donor for reductive acetogenesis. In an acetogenic bacterium, *Moorella thermoacetica*, the *mtvABC* genes encode *O*-demethylase, which transfers the methyl group to tetrahydrofolate (THF) to produce methyl-THF (Naidu & Ragsdale., 2001). In the Fibrobacteres R8-3-H12 genome, we found homologous genes to all the three *mtvABC* genes, although we also detected *mtvA*- and *mtvC*-like genes in the TG3 R8-0-B4 and *Treponema* R8-4-B8 genomes (Figure 9). The transcripts of the *mtvABC*-like genes of Fibrobacteres R8-3-H12 were detected in the RNA-seq data in a moderate amount (Figure 6 and Figure 9). The results imply that the Fibrobacteres bacterium utilises methoxylated aromatic compounds as a donor of methyl group for reductive acetogenesis although the *acsE* gene has not yet been identified in this genome.

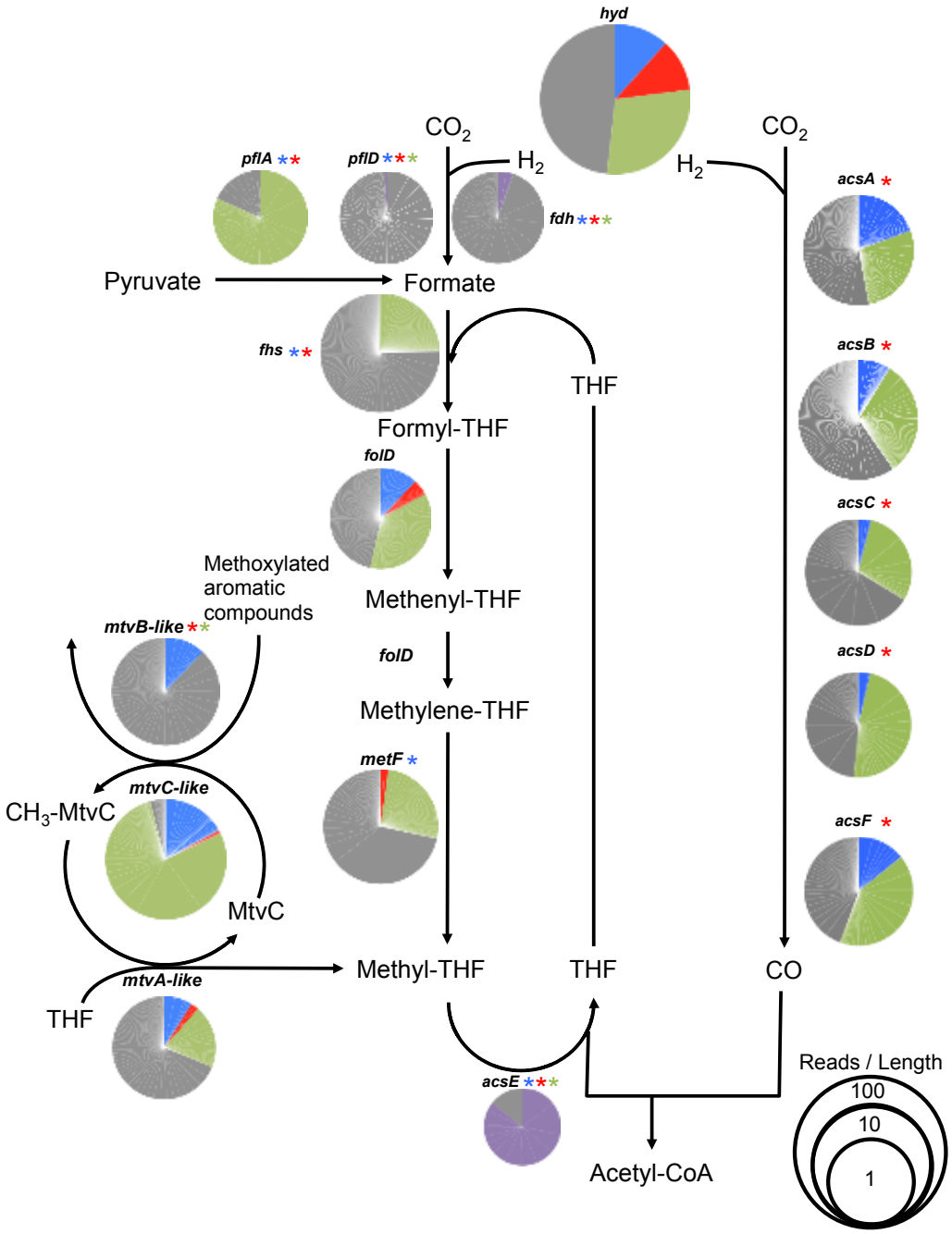

Figure 9. Expression levels of genes related to reductive acetogenesis. The gene expression level is shown as the number of reads in the RNA-seq data per gene length. *hyd*, FeFe-hydrogenase; *pflA*, pyruvate-formate lyase activating enzyme; *pflD*, pyruvate-formate lyase; *fhs*, formate-tetrahydrofolate ligase; *folD*, methylenetetrahydrofolate dehydrogenase (NADP$^+$) / methenyltetrahydrofolate cyclohydrolase; *metF*, methylenetetrahydrofolate reductase; *acsE*, 5-methyltetrahydrofolate corrinoid/iron sulfur protein methyltransferase; *mtvABC*, *O*-demethylase of methoxylated aromatic compounds and *mtvB, mtvC* and *mtvA* correspond to uroporphyriongen decarboxylase domain, cobalamin-binding domain and pterin-binding domain, respectively. Genes encoding subunits of ACS/CODH are: *acsA*, carbon monoxide dehydrogenase subunit (also known as *cooS*); *acsB*, acetyl-CoA synthase subunit; *acsC*, large subunit of Fe-S-Co protein; *acsD*, small subunit of Fe-S-Co protein; and *acsF,* Ni-insertion protein. The pie charts indicate total gene expression levels in the gut community and their size is proportional to the expression level of each gene. The colours indicate that expressed gene homologous to that encoded in the genomes of Fibrobacteres (blue), TG3 (red), *Treponema* (green), and other termite-gut *Treponema* genomes (purple) and other bacteria (grey). The asterisks in colours indicate that the gene was not found in the respective genomes examined in this study. The *acsB, metF* and *mtvC* in other single-cell genomes of the dominant *Treponema* species mentioned in the main text were used to evaluate the expression levels. The *acsD* of the Fibrobacteres and *Treponema* bacteria were analysed using the *acsD* domain of their fused *acsBD* gene.

All three single-cell genomes possessed the genes for FeFe-hydrogenase, which is likely involved in reductive acetogenesis (Figure 9). In addition, the *Treponema* R8-4-B8 genome possessed the gene for carbonic anhydrase, which reversibly catalyses the conversion of CO_2 to bicarbonate and water. The function of carbonic anhydrase in homoacetogenic bacteria is considered to increase the intracellular CO_2 level for CO_2 fixation (Smith & Ferry., 2000).

Biosynthesis of folate

Folate is an essential metabolite for producing THF, which is necessary in reductive acetogenesis. The Fibrobacteres R8-3-H12 genome possessed almost all genes involved in the biosynthesis of folate, except for the dihydroneopterin aldolase gene (*folB*) (Figure 10). The *folB* gene was not found at all in the TG3 and *Treponema* single-cell genomes nor the entire RNA-seq data. The absence of *folB* gene was prevalent in diverse bacterial genomes (de Crézy-Lagard *et al*., 2007), and instead, an unusual paralog of 6-pyruvoyltetrahydropterin synthase (PTPS-III) that can functionally replace FolB was found in almost all of these *folB*-deficient genomes (Pribat *et al*., 2009). The gene encoding PTPS-III with conserved glutamate and cysteine residues in the active site was found in the TG3 R8-0-B4 genome, although its folate biosynthetic pathway was incomplete because of the lack of genes for 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase (*folK)* and dihydrofolate reductase (*folA*) in this genome. The RNA-seq results showed that 34% of the expressed genes of PTPS-III were closely related to the TG3 R8-0-B4 gene. The *Treponema* R8-4-B8 genome possessed *folC* and *folA* genes for only the final two steps catalysed by dihydrofolate synthase/folylpolyglutamate synthase, indicating that the dominant *Treponema* bacterium requires folate or the precursor of its biosynthesis for growth as in many *Treponema* species (Graber & Breznak., 2005). These results imply that none of the three dominant bacteria is able to synthesise folate and they depend on one another or other community members for its biosynthesis.

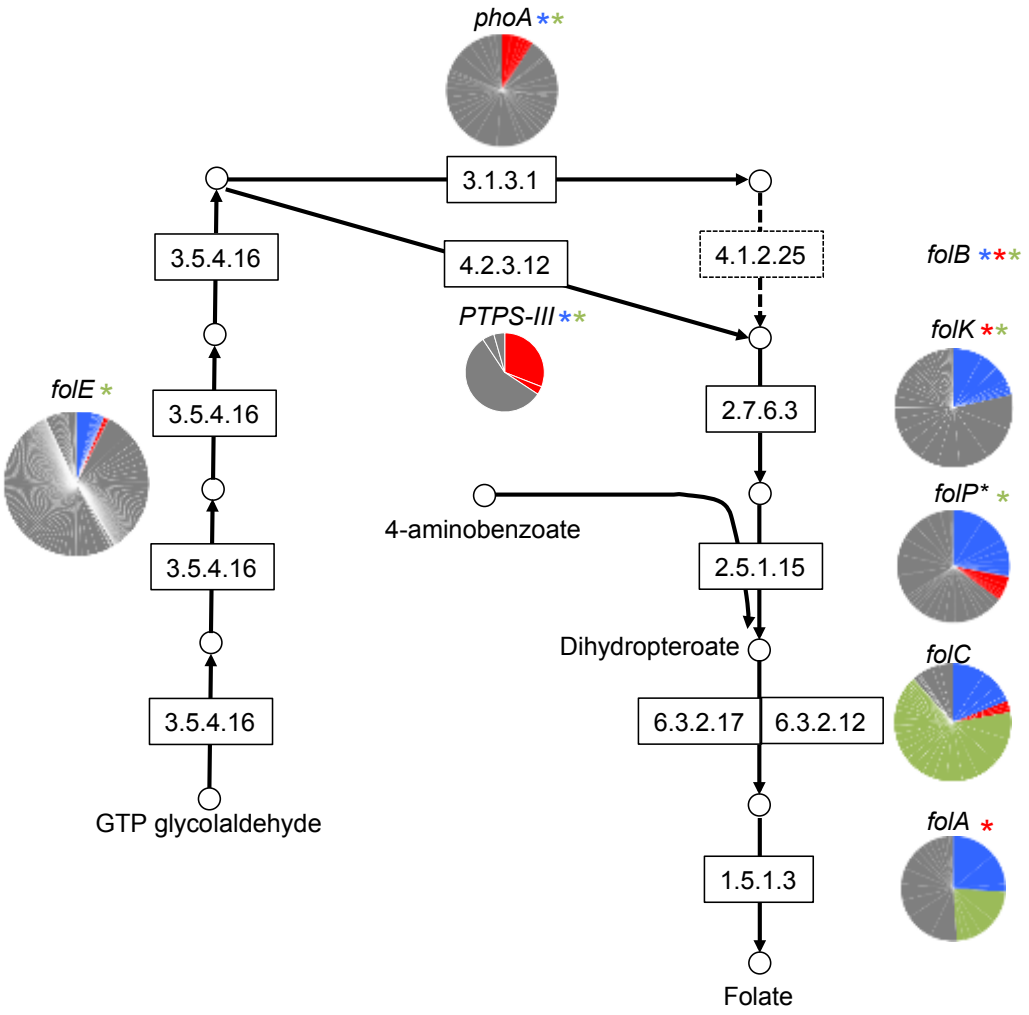

Figure 10. Expression levels of genes related to folate biosynthesis pathway. The gene expression level is shown as the number of reads in the RNA-seq data per gene length. *folE*, GTP cyclohydrolase I; *phoA*, alkaline phosphatase; *folP*, dihydropteroate synthase. The size of the pie charts and the colours are described in the legend of Figure 9. The asterisks in colours indicate that the gene was not found in the respective genomes examined in this study. The *folP* gene in another Fibrobacteres genome R8-0-F8 showing 96.3% sequence identity of the 16S rRNA gene with R8-3-H12 was used.

Nitrogen fixation

A molybdenum-dependent nitrogenase complex comprises essential catalytic components NifD and NifK and a nitrogenase reductase NifH. The TG3 R8-0-B4 and Fibrobacteres R8-3-H12 genomes possessed *nifH*, *nifD*, *nifK*, and the genes for a molybdenum transport system, although *nifH* in this TG3 genome was a partial sequence. In contrast, these genes were not found in the *Treponema* R8-4-B8 genome except for

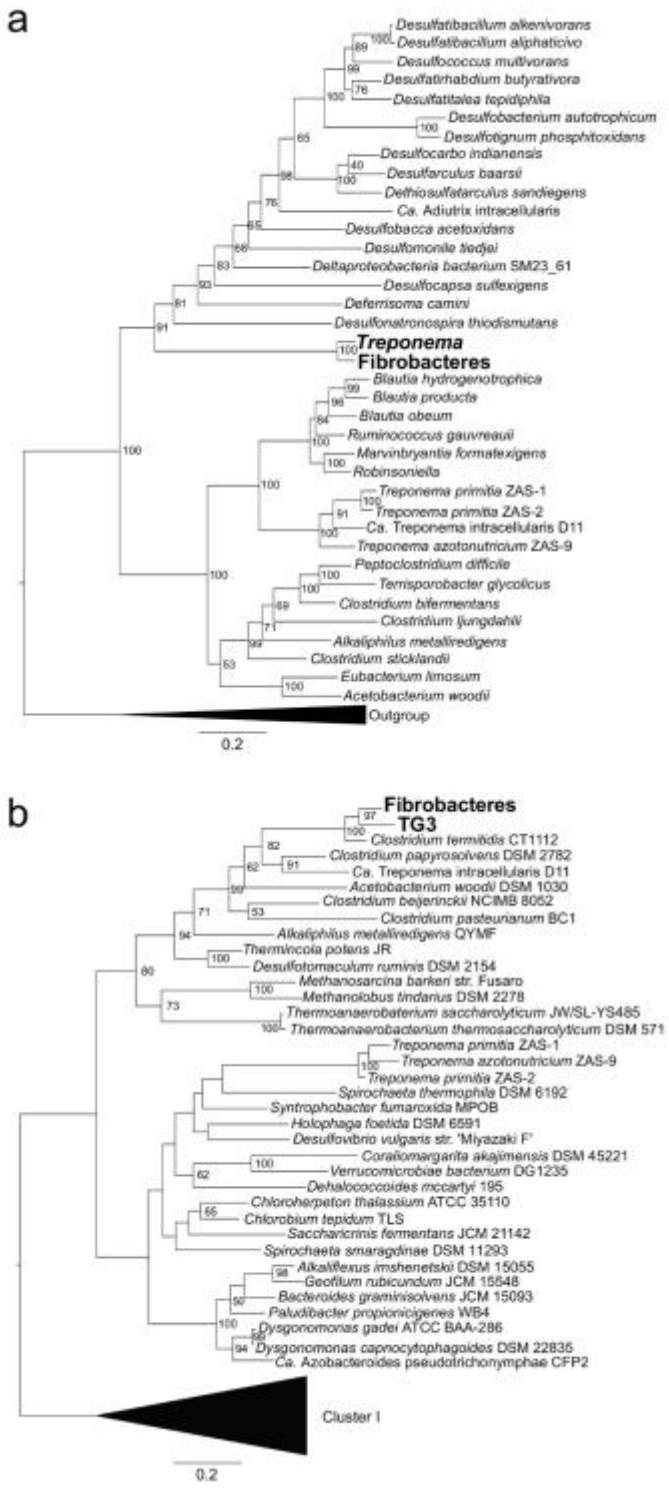

Figure 11. Phylogenetic trees based on the concatenated amino acid sequences of nitrogenase and CO dehydrogenase/acetyl-CoA synthase. (a) A total of 1,369 amino acid sites of AcsA, AcsB, AcsC, AcsD and AcsF were used to construct the maximum likelihood tree using the LG+G+I model. Bootstrap values as percentages are indicated at the nodes. The protein sequence encoded by *acsB* of another *Treponema* single-cell genome was used as mentioned in the Figure 6 legend. The *acsB* and *acsD* genes in the Fibrobacteres and *Treponema* genomes are fused and each corresponding domain sequence were used for the analysis. (b) A total of 508 amino acid sites of NifH, NifD and NifK were used to construct the maximum likelihood tree using the LG+G+I model. Sequences of Cluster I of Nif were used as outgroups. The NifH sequence encoded by another TG3 genome mentioned in the Figure 6 legend was used. Bootstrap values as percentages are indicated at the nodes when the values were over 50%. Scales indicate 0.2 substitutions per site.

nifH. A phylogenetic analysis indicated that *nifH* genes in these TG3 and Fibrobacteres genomes belong to Cluster III of *nifH*, but *nifH* gene in this *Treponema* genome belongs to Cluster IV of *nifH* (data not shown). The function of Cluster IV of *nifH* is not entirely understood, although Zheng *et al*., 2016 proposed the possible function of a Cluster IV member in nitrogen fixation. Phylogenetic analysis based on a concatenated NifH, NifD and NifK amino acid sequence revealed that the NifHDK of TG3 R8-0-B4 and Fibrobacteres R8-3-H12 were closely related to each other and branched out among Clostridia members (Figure 11b). Moreover, the genes for the nitrogenase cofactor biosynthesis proteins, *nifE* and *nifN*, were present in these TG3 and Fibrobacteres genomes, and their NifN protein contained a NifB domain as in the case of '*Ca*. Treponema intracellularis'. The presence of a minimal gene set (*nifHDKENB*) required for nitrogen fixation (Wang *et al*., 2013) in these Fibrobacteres and TG3 genomes suggests their potential for diazotrophy. The RNA-seq analysis indicated that the *nifH* genes closely related to these Fibrobacteres and TG3 genes accounted for approximately 90% of the expressed level of *nifH* genes in the termite gut (Figure 6).

Biosynthesis of amino acids, cofactors and nucleotides

The single-cell genomes of TG3, Fibrobacteres and *Treponema* encoded most of the genes involved in the biosynthesis of the proteinaceous 20 amino acids, but the Fibrobacteres R8-3-H12 genome typically lacked genes for the methionine biosynthetic pathway (Figure 12a). The presence of a gene for methionine ABC transporter in this Fibrobacteres genome suggests the uptake of methionine. Each of the three single-cell genomes possessed genes for an ammonium transporter and various amino acid ABC transporters, and many genes for the biosynthesis of cofactors (Figure 12b), purines, and pyrimidines.

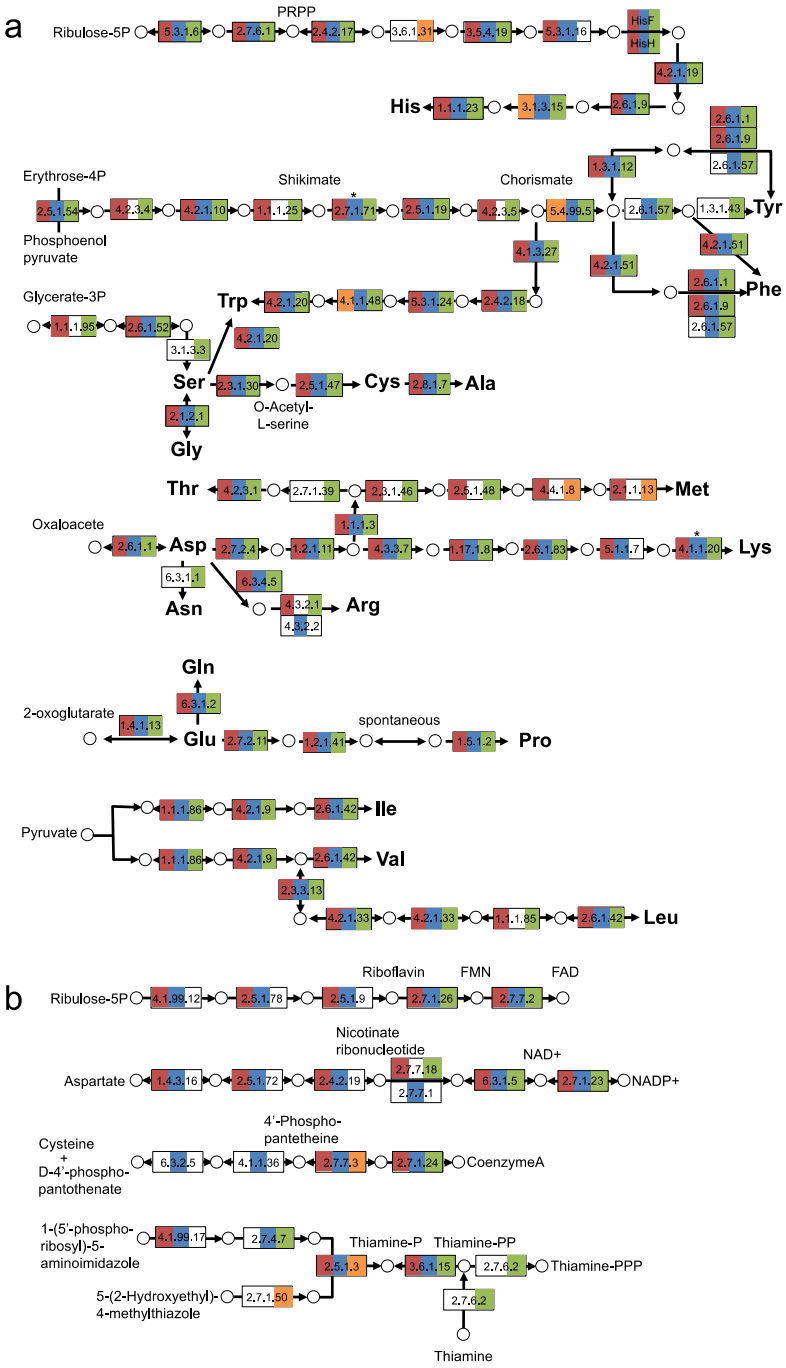

Figure 12. Predicted biosynthetic pathways for amino acids (a) and cofactors (b) in the TG3, Fibrobacteres and *Treponema* genomes. The colours indicate presence of the pathway in the TG3 (red), Fibrobacteres (blue), *Treponema* (green) and other single-cell (orange) genomes. Asterisks show that the genes were partial sequence. The white indicates the gene is absent. Several reaction steps absent in the genomes may be functional with yet-uncharacterized enzymes because these steps are not identified in the genomes of the cultured close relative *Treponema primitia* such as reactions 5.3.1.16 and 4.1.99.17, and of *Chitinivibrio alkaliphilus* such as the reactions 3.1.3.3, 6.3.1.1, 2.7.1.39, 2.7.4.7 and 2.7.1.50. The *Treponema* bacterium may require several vitamins for growth because its genome lacks genes for biosynthesis for riboflavin, nicotinate ribonucleotide, and 4'-phosphopantethein as in the genome of *T. primitia*

Discussion

The draft genomes of the three dominant bacterial species in the gut of *N. takasagoensis* were obtained by single-cell genome sequencing. The genome analyses reveal the

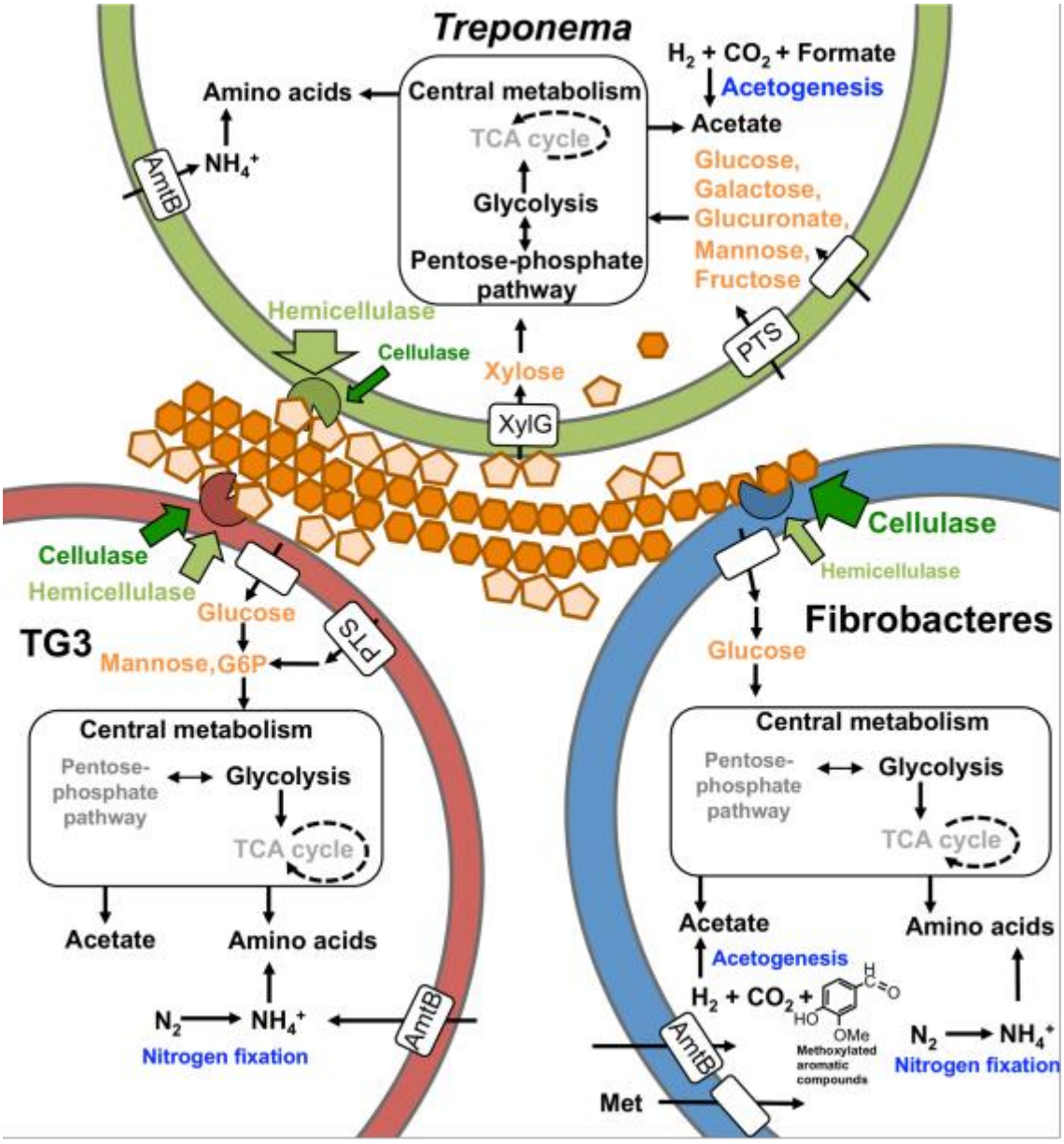

Figure 13. Hypothesised symbiotic roles among the three dominant bacteria of TG3, Fibrobacteres and *Treponema*. The TG3, Fibrobacteres and *Treponema* bacteria co-ordinately digest lignocellulose, utilise the different carbon sources, and then generate acetate. In addition, the *Treponema* and Fibrobacteres bacteria produce acetate by reductive acetogenesis using different methyl-group donors. The TG3 and Fibrobacteres bacteria fix nitrogen. Pentose phosphate pathway and TCA cycle described in grey indicate that genes for the pathway or the cycle were only partially found. AmtB, ammonium transporter; XylG, xylose transporter; PTS, phosphotransferase system IIA components; G6P, glucose-6-phosphate; Met, methionine.

intertwining functions of the three dominant bacteria in terms of lignocellulose digestion, reductive acetogenesis, and nitrogen fixation (Figure 13). Based on the analyses of the expression levels of the closely related genes found in the genomes, we successfully evaluated their contributions in the gut microbial community.

Lignocellulose is composed of cellulose, hemicellulose and lignin. Each component is combined to one another in a complicated manner, and thus many kinds of GHs are required for the efficient digestion. Indeed, the previous metagenomic analysis reported the presence of more than 700 genes of GHs classified into 45 GH families in the gut community of *Nasutitermes* sp. (Warnecke *et al*., 2007). In the present study, we found genes encoding a variety of cellulolytic and hemicellulolytic enzymes in the single-cell genomes of the three dominant species and genes related to these (hemi)cellulolytic enzymes were detected as mostly major degradation genes expressed in the gut community. Nevertheless, we could not detect any cellobiohydrolase-encoding genes in these genomes and also in the RNA-seq data of the entire gut microbial community. This enzyme digests the crystalline cellulose component of wood particles. The absence of cellobiohydrolase genes is in agreement with the previous metagenome analysis of gut bacteria of *Nasutitermes* sp. (Warnecke *et al*., 2007). In lower termites, cellulolytic gut protists possess cellobiohydrolases that play a crucial role in the digestion of crystalline cellulose. The gut symbionts in the higher termites seem to have evolved a unique lignocellulose degradation system, one that does not require cellobiohydrolases.

Interestingly, the ratio of gene abundance for cellulases to hemicellulases is different among the dominant TG3, Fibrobacteres and *Treponema* species, implying their division of roles in lignocellulose digestion in the gut community. The compositions of GH families in the TG3 and Fibrobacteres single-cell genomes were similar to those reported previously in the genomes reconstructed from metagenome data for species in these bacterial groups (Abdul Rahman *et al*., 2015); however, in the TG3 members, the xylanase genes of GH10 and GH11 were found only in the present study. Mikaelyan et al. 2014 reported that members of the phylogenetic clusters encompassing these bacteria were able to adhere to cellulose fibres. We suggest that the three dominant, fibre-adhering bacteria compose a coordinated system to accomplish effective digestion with their

division of roles. The Fibrobacteres species mainly digests cellulose components of fibres, whereas the *Treponema* species mainly targets the hemicellulose components. The TG3 species can work on both cellulose and hemicellulose. The hemicellulolytic abilities increase the exposure of the cellulose regions and the cellulolytic abilities also help to expose the hemicellulose regions, and thus their coordination works synergistically enhancing the digestion. Of course, other bacteria in the gut community should be also involved in lignocellulose digestion, e.g. xylan degradation as shown by the abundance of expressed GH11, GH43 and GH74 genes unrelated to those encoded by the three dominant bacterial genomes (see Figure 6). Furthermore, the differences in the compositions of GHs among the dominant bacteria suggest that they produce different mono-/oligosaccharides from lignocellulose, and that they likely utilise different sugars to alleviate competition for growth substrates in the gut. This possibility is supported by the fact that the pathways utilising xylose and glucuronate, products from hemicellulose, were found exclusively in the genome of the *Treponema* species.

Termites utilise acetate produced by their gut symbionts as a major carbon and energy source. The metabolic pathways predicted from the single-cell genomes indicate that all the three dominant bacterial species are able to ferment monosaccharaides to acetate. In wood-feeding lower and higher termites, reductive acetogenesis from H_2 and CO_2 contributes to the energy demand of the host termite (Breznak., 2000), and at least in lower termites, *Treponema* species including an endosymbiont of a cellulolytic protist are likely responsible for this activity (Ohkuma *et al*., 2015; Leadbetter *et al*., 1999). We consider that the Fibrobacteres and *Treponema* members play substantial roles for reductive acetogenesis in the gut, because both genomes possessed the genes of ACS/CODH and transcripts of their related genes were detected in considerable amount in the gut community (see Figure 6). However, several genes involved in the methyl branch of Wood-Ljungdahl pathway were not found in both genomes.

In the metagenome studies of gut community of higher termites, conflicting results were reported for formate dehydrogenase gene (*fdh*); in one case of *Nasutitermes* sp., *fdh* gene is rarely detected (Warnecke *et al*., 2007), while in the other termite species it is detected (He *et al*., 2013). An alternative way to produce formate with pyruvate-formate lyase

(encoded by *pflD* gene) is hypothesised in the former case. Here, both *fdh* and *pflD* genes were detected in the RNA-seq data of the gut community, but their organismal origins were largely unidentified and *Treponema*-like sequences comprised only a small fraction of the expressed genes. The absence/paucity of genes for formate production may indicate that *Treponema* species in the gut of higher termites preferentially utilise formate supplied by other members of the gut community. Indeed, formate accumulates in a significant level in the *Nasutitermes* gut (Köhler *et al*., 2012) and at least in a lower termite, labelled formate injected into the gut is incorporated to acetate at a considerable rate (Pester & Brune., 2007).

In the case of the dominant Fibrobacteres bacterium, the methyl-branch of the Wood-Ljungdahl pathway is incomplete because of the lack of key genes such as *fdh*, *fhs* and *metF* in the genome. Instead, it can probably utilise *O*-methyl group of methoxylated aromatic compounds as a methyl donor, which is common in many homoacetogens including isolates from termite guts (Graber & Breznak., 2004; Boga *et al*., 2003). Such compounds are rich in lignin, and *O*-demethylation of lignin during the gut passage is detected in a lower termite (Geib *et al*., 2008). The dominant *Treponema* and Fibrobacteres bacteria likely contribute to reductive acetogenesis in the gut but they show division of roles in terms of the use of methyl donors.

Nitrogen fixation by the termite-gut bacteria and upgrading of the fixed nitrogen to essential nitrogenous compounds (e.g. amino acids) are pivotal in the symbiosis because the host termite thrives on dead wood extremely poor in nitrogen (Breznak., 2000). The genomes of the dominant Fibrobacteres and TG3 species contained the minimal gene set needed for nitrogen fixation and we consider that they are responsible for nitrogen fixation in the gut. Indeed, their related *nifH* genes shared a great majority of the expressed genes in the gut community (see Figure 6). The ability of nitrogen fixation for TG3 and Fibrobacteres members are debatable in the recent metagenome and the following genome reconstruction study because of the presence of only a limited number of *nif* genes among the minimal gene set (Abdul Rahman *et al*., 2015). Highly similar sequences to the *nifH* genes in the genomes of the dominant TG3 and Fibrobacteres species have been detected in the gut community of *N. takasagoensis* and other wood-

feeding higher termites (Ohkuma *et al*., 1999; Yamada *et al*., 2007). These sequences are most abundant among the detected *nifH* sequences, suggesting that TG3 and Fibrobacteres members play crucial roles in nitrogen fixation in wood-feeding higher termites.

Possible horizontal gene transfers (HGTs) across distantly related taxa are depicted in this study in two important functions in the gut community. One occurs in genes for the ACS/CODH complex of the Fibrobacteres and *Treponema* members, and the other in *nifHDK* genes of the Fibrobacteres and TG3 members (see Figure 11). The former is sister to the sequences of Deltaproteobacteria members, but there is no known species identified as close relatives (Figure 11A). The latter is branched out among the sequences from Clostidia members and a close relative is *Clostridium termitidis*, an isolate from the gut of the higher termite *Nasutitermes lujae* (Hethener *et al*., 1992). The *nifHDK* gene sequence of the previously identified nitrogen-fixing endosymbiont ('*Ca*. Treponema intracellularis') of a cellulolytic protist in a lower termite is also branched out near *C. termitidis* and a horizontal gene transfer is suggested (Ohkuma *et al*., 2015). In both cases, the sequences of the gut symbionts of *N. takasagoensis* studied here are the closest neighbours to each other. In each case, a plausible scenario is that at first one of the gut symbiont species has acquired the gene laterally from a distantly related group of bacteria, e.g. Deltaproteobacteria or Clostridia, and then the acquired gene has been secondarily transferred to another gut symbiont species. The gut microbial community is densely populated and provides much opportunity for gene exchange. When the acquired gene is beneficial to the recipient as well as other community members or the host termite, the recipient has established a niche suitable for the symbiosis and possibly become a dominant population in the community.

The single-cell genome approach as described here is advantageous to infer the functions of individual members in a complex microbial community. HGTs seem to be hardly detected when binning of metagenome sequences and the following manual curation uses the phylogenetic context to identify the organismal origins. The binning of metagenomes, though powerful, is solely based on the bioinformatics and thus has limitation in itself. The single-cell genome is promising and complementary to metagenomics; however, the

completeness of the obtained genome sequences is often not satisfactory. Here, we examined the method using specific primers for WGA, and by applying this method, the recovery of the genome sequence was successfully improved as judged from the increase in the size of assembled contigs and the completeness of the genome (Table 2). Indeed, we tested the effects of the use of specific primers for *de novo* assembly of the obtained sequence data of the single-cell genome (Figure 1). Judging from the N50 length, the assembly of the short sequence reads selected from both WGA samples amplified with standard random hexamers and specific primers greatly improved when compared with that of only the random hexamers. This method, if applied to a single-cell genome study may improve the genome reconstruction where initial sequencing is unsatisfactory for the assembly and genome recovery.

In conclusion, this study discloses that the dominant bacterial species of TG3, Fibrobacteres and *Treponema* contribute to the major nutritional symbiotic roles of lignocellulose digestion, nitrogen fixation and reductive acetogenesis, and individually play intertwining roles. The divisions of roles found in these three dominant species are complementary to one another and can explain the maintenance of respective responsible species. As shown in biosyntheses of folate and other cofactors (Figure 10 and Figure 12b), these dominant bacterial species particularly the *Treponema* species also seem to depend on other community members for provision of several cofactors or their precursors. The complex gut community comprises diverse microbial species and their intertwining symbiotic functions likely attribute to the formation and maintenance of the complexity of the gut community. Furthermore, these bacteria have probably obtained and exchanged the important genes for the symbiotic functions through HGTs, which should be crucial in the evolution and adaptation of the gut bacteria to establish efficient and stable symbiotic relationships in the gut community. This study also indicates that the single-cell genome approach together with metatranscriptomics greatly improves our understanding of symbiotic relationships in a complex microbial community.

Data availability

The single-cell genome assemblies of the TG3, Fibrobacteres and *Treponema* isolates have been deposited at DNA Data Bank of Japan (DDBJ) under the accession numbers BDGZ01000001-BDGZ01000536 (TG3 R8-0-B4), BDHC01000001-BDHC01000780 (TG3 R8-4-B1), BDHA01000001-BDHA01000724 (Fibrobacteres R8-3-H12), BDHB01000001-BDHB01001042 (Fibrobacteres R8-0-F8), BDHE01000001-BDHE01000529 (*Treponema* R8-4-B8), BDHF01000001-BDHF01000611 (*Treponema* R80B11-R83G3) and BDHD01000001-BDHD01001541 (*Treponema* R6D11). The RNA-seq reads have been deposited at DDBJ Sequence Read Archive (DRA) under the accession number DRA005097.

Acknowledgements

This work was supported in part by Grants-in-Aid for Scientific Research from Japan Society for Promotion of Science (JSPS), Nos. 26660075 (to M.Y.), 23117003 (to M.O. and Y.H.) and 26292047 and 16K14797 (to M.O.), by a grant from the Institute of Fermentation, Osaka (to M.Y.), by the Special Postdoctoral Research Program of RIKEN (to M.S.), and by a RIKEN Competitive Program for Creative Science and Technology (to M.O).

Author contributions

M.Y., D.S., Y.H and M.O. designed the research. M.Y. and M.S. sorted single cells. M.Y. and H.K. sequenced amplified DNA, M.Y. and D.S. assembled and annotated single-cell genomes. M.Y. and D.S performed RNA-seq. M.Y., D.S., Y.H. and M.O. wrote the paper.

Competing financial interests

The authors declare no competing financial interests.

References

Abdul Rahman N, Parks DH, Vanwonterghem I, *et al.* (2015) A Phylogenomic Analysis of the Bacterial Phylum Fibrobacteres. *Front Microbiol* **6**, 1469.

Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169.

Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol Biol* **12**, 186.

Aziz RK, Bartels D, Best AA, *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75.

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

Boga H, Ludwig W, Brune A (2003) Sporomusa aerivorans sp. nov., an oxygen-reducing homoacetogenic bacterium from the gut of a soil-feeding termite. *Int J Syst Evol Microbiol* **53**, 1397-1404.

Brune A (2014) Symbiotic digestion of lignocellulose in termite guts. *Nat Rev Microbiol* **12**, 168-180.

Brune A, & Ohkuma M (2011) Role of the termite gut microbiota in symbiotic digestion. In: Bignell DE, Roisin Y, Lo N, (eds). Biology of Termite: A Modern Synthesis. Springer-Verlag: pp 439-475.

Buchfink B, Xie C, Huson DH, (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60.

de Crécy-Lagard V, EI Yacoubi B, de la Garza RD, *et al.* (2007) Comparative genomics of bacterial and plant folate synthesis and salvage: predictions and validations. *BMC Genomics* **8**, 245.

Desai MS, & Brune A. (2012) Bacteroidales ectosymbionts of gut flagellates shape the nitrogen-fixing community in dry-wood termites. *ISME J* **6**, 1302-1313.

Edgar RC. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. **5**, 113.

Geib SM, Filley TR, Hatcher PG, *et al.* (2008) Lignin degradation in wood-feeding insects. *Proc Natl Acad Sci U S A* **105**, 12932-12937.

Graber JR, & Breznak JA. (2005) Folate cross-feeding supports symbiotic homoacetogenic spirochetes. *Appl Environ Microbiol* **72**, 1883-1889.

Graber JR, & Breznak JA. (2004) Physiology and nutrition of *Treponema primitia*, an H2/CO2-acetogenic spirochete from termite hindguts. *Appl Environ Microbiol* **70**, 1307-1314.

Haas BJ, Papanicolaou A, Yassour M, *et al.* (2013) De novo transcript sequence

reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494-1512.

He SM, Ivanova N, Kirton E, *et al.* (2013) Comparative Metagenomic and Metatranscriptomic Analysis of Hindgut Paunch Microbiota in Wood- and Dung-Feeding Higher Termites. *Plos One* **8**, 14.

Hethener P, Brauman A, & Garcia J. (1992) *Clostridium termitidis* sp. nov., a cellulolytic bacterium from the gut of the woodfeeding termite, *Nasutitermes lujae*. *Syst Appl Microbiol* **15**, 52–58.

Hongoh Y, Deevong P, Hattori S, *et al.* (2006) Phylogenetic diversity, localization, and cell morphologies of members of the candidate phylum TG3 and a subphylum in the phylum Fibrobacteres, recently discovered bacterial groups dominant in termite guts. *Appl Environ Microbiol* **72**, 6780-6788.

Hongoh Y, Deevong P, Inoue T, *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, Sharma VK, Prakash T, *et al.* (2008a) Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Proc Natl Acad Sci U S A* **105**, 5555-5560.

Hongoh Y, Sharma VK, Prakash T, *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, Strassert JF, Köhler T, *et al.* (2016) 'Candidatus Adiutrix intracellularis', an endosymbiont of termite gut flagellates, is the first representative of a deep-branching clade of Deltaproteobacteria and a putative homoacetogen. *Environ Microbiol* **18**, 2548-2564.

Jewell KA, Scott JJ, Adams SM & Suen C. (2013) A phylogenetic analysis of the phylum Fibrobacteres. *Sys Appl Microbiol* **36**, 376-382.

Köhler T, Dietrich C, Scheffrahn, RH & Brune A. (2012) High-resolution analysis of gut environment and bacterial microbiota reveals functional compartmentation of the gut in wood-feeding higher termites (*Nasutitermes* spp.). *Appl Environ Microbiol*

78, 4691-4701.

Kumar S, Jones M, Koutsovoulos G, Clarke M & Blaxter M. (2013) Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front Genet* **4**, 237.

Langmead B & Saizberg SL. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359.

Leadbetter JR, Schmidt TM, Graber JR & Breznak JA. (1999) Acetogenesis from H2 plus CO2 by spirochetes from termite guts. *Science* **283**, 686-689.

Mikaelyan A, Köhler T, Lampert N, *et al.* (2015) Classifying the bacterial gut microbiota of termites and cockroaches: A curated phylogenetic reference database (DictDb). *Syst Appl Microbiol* **38**, 472-482.

Mikaelyan A, Dietrich C, Köhler T, *et al.* (2015) Diet is the primary determinant of bacterial community structure in the guts of higher termites. *Mol Ecol* **24**, 5284-5295.

Mikaelyan A, Strassert JF, Tokuda G & Brune A. (2014) The fibre-associated cellulolytic bacterial community in the hindgut of wood-feeding higher termites (*Nasutitermes spp.*). *Environ Microbiol* **16**, 2711-2722.

Naidu D & Ragsdale SW. (2001) Characterization of a three-component vanillate O-demethylase from *Moorella thermoacetica*. *J Bacteriol* **183**, 3276-3281.

Ohkuma M, Noda S, Hattori S, *et al.* (2015) Acetogenesis from H2 plus CO2 and nitrogen fixation by an endosymbiotic spirochete of a termite-gut cellulolytic protist. *Proc Natl Acad Sci USA* **112**, 10224-10230.

Ohkuma M, Noda S & Kudo T. (1999) Phylogenetic diversity of nitrogen fixation genes in the symbiotic microbial community in the gut of diverse termites. *Appl Environ Microbiol* **65**, 4926-4934.

Pester M & Brune A. (2007) Hydrogen is the central free intermediate during lignocellulose degradation by termite gut symbionts. *ISME J* **1**, 551-565.

Powell S, Forslund F, Szklarczyk D, *et al.* (2014) eggnog v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* **42**, D231-D239.

Pribat A, Jeanguenin L, Lara-Núñez A, *et al.* (2009) 6-pyruvoyltetrahydropterin synthase paralogs replace the folate synthesis enzyme dihydroneopterin aldolase in diverse bacteria. *J Bacteriol* **191**, 4158-4165.

Rinke C, Schwientek P, Sczyrba A*, et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431-437.

Rosenthal AZ, Matson EG, Eldar A & Leadbetter JR. (2011) RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. *ISME J* **5**, 1133-1142.

Seemann T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069.

Smith KS & Ferry JG. (2000) Prokaryotic carbonic anhydrases. *FEMS Microbiol Rev* **24**, 335-366.

Sorokin DY, Gumerov VM, Rakitin AL, *et al.* (2014) Genome analysis of Chitinivibrio alkaliphilus gen. nov., sp. nov., a novel extremely haloalkaliphilic anaerobic chitinolytic bacterium from the candidate phylum Termite Group 3. *Environ Microbiol* **16**, 1549-1565.

Sorokin DY, Rakitin AL, Gumerov VM, *et al.* (2016) Phenotypic and Genomic Properties of Chitinispirillum alkaliphilum gen. nov., sp. nov., A Haloalkaliphilic Anaerobic Chitinolytic Bacterium Representing a Novel Class in the Phylum Fibrobacteres. *Front Microbiol* **7**, 407.

Stamatakis A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690.

Suen G, Weimer PJ, Stevenson DM, *et al.* (2011) The complete genome sequence of Fibrobacter succinogenes S85 reveals a cellulolytic and metabolic specialist. *PLoS One* **6**, e18814.

Tamura K, Peterson D, Peterson N, *et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28, 2731-2739.

Tokuda G & Watanabe H. (2007) Hidden cellulases in termites: revision of an old hypothesis. *Bio Lett* **3**, 336-339.

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

Wang L, Zhang L, Liu Z, *et al.* (2013) A minimal nitrogen fixation gene cluster from *Paenibacillus sp*. WLY78 enables expression of active nitrogenase in Escherichia coli. *PLoS Genet* **9**, e1003865.

Warnecke F, Luginbuhl P, Ivanova N, *et al.* (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**, 560-565.

White BA, Lamed R, Bayer EA & Flint HJ. (2014) Biomass utilization by gut microbiomes. *Annu Rev Microbiol* **68**, 279-296.

Yamada A, Inoue T, Noda S, Hongoh Y & Ohkuma M. (2007) Evolutionary trend of phylogenetic diversity of nitrogen fixation genes in the gut community of wood-feeding termites. *Mol Ecol* **16**, 3768-3777.

Yuki M, Kuwahara H, Shintani M, *et al.* (2015) Dominant ectosymbiotic bacteria of cellulolytic protists in the termite gut also have the potential to digest lignocellulose. *Environ Microbiol* **17**, 4942-4953.

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.

Chapter IV

Acetogenesis from H_2 plus CO_2 and nitrogen fixation by an endosymbiotic spirochete of a termite-gut cellulolytic protist

Chapter IV

This Chapter has already been published in the Proceedings of the National Academy of Sciences USA, August 18th 2015, Volume 112, number 33, pages 10224-10230. And is submitted here in its original format. My roles in this study were the assembly, annotation, genome analysis and figure and table production of single cell isolated *Eucomonympha* endosymbionts. I also had a part in the writing and reviewing of the main text.

Classification: Biological Science, Evolution

Acetogenesis from H_2 plus CO_2 and nitrogen fixation by an endosymbiotic spirochete of a termite-gut cellulolytic protist

Moriya Ohkuma[a,b,1], Satoko Noda[a,c], Satoshi Hattori[d], Toshiya Iida[a], Masahiro Yuki[b], David Starns[a,e], Jun-ichi Inoue[a], Alistair C. Darby[e], and Yuichi Hongoh[a,f]

[a]Japan Collection of Microorganisms/Microbe Division, RIKEN BioResource Center, and [b]Biomass Research Platform Team, RIKEN Biomass Engineering Program Cooperation Division, RIKEN Center for Sustainable Resource Science, Ibaraki 305-0074, Japan; [c]Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Yamanashi 400-8511, Japan; [d]Department of Food, Life, and Environmental Sciences, Yamagata University, Yamagata 997-8555, Japan; [e]Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom; and [f]Department of Biological Sciences, Tokyo Institute of Technology, Tokyo 152-8550, Japan

[1]To whom correspondence may be addressed. Email: mohkuma@riken.jp. Address: Japan Collection of Microorganisms/Microbe Division, RIKEN BioResource Center, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan; Tel: +81-29-829-9101, Fax: +81-29-836-9561

Author contributions: M.O., S.N., S.H., and T.I. designed research; S.N., S.H., T.I., M.Y., and Y.H. performed research; M.O., S.N., S.H., T.I., M.Y., D.S., J.I., and A.C.D. analysed data; and M.O., A.C.D., and Y.H. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The sequences reported in this study have been deposited in the DNA Data Bank of Japan [accession nos. BBPV01000001–798, BBPW01000001–757, BBPX01000001–878, BBPY01000001–946, and BBPZ01000001–698 (genomes) and LC012866–LC012879 (others)].

Abstract

Symbiotic associations of cellulolytic eukaryotic protists and diverse bacteria are common in the gut microbial communities of termites. Besides cellulose degradation by the gut protists, reductive acetogenesis from H_2 plus CO_2 and nitrogen fixation by gut bacteria play crucial roles in the host termites' nutrition by supplying the major carbon and energy source and nitrogen poor in their diet, respectively. Fractionation of these activities and the identification of key genes from the gut community of the wood-feeding termite *Hodotermopsis sjoestedti* revealed that substantial activities in the gut, nearly 60% of reductive acetogenesis and almost exclusively for nitrogen fixation, were uniquely attributed to the endosymbiotic bacteria of the cellulolytic protist in the genus *Eucomonympha*. The rod-shaped endosymbionts were surprisingly identified as a spirochete species in the genus *Treponema*, which usually exhibits a characteristic spiral morphology. The endosymbionts likely utilise H_2 produced by the protist for these dual functions. Although H_2 is known to inhibit nitrogen fixation in some bacteria, it seemed to rather stimulate this important mutualistic process. In addition, the single-cell genome analyses revealed its potentials of the use of sugars advantageous for its energy requirement, and of the biosynthesis of valuable nutrients for the host such as amino acids from the fixed nitrogen. These metabolic interactions are suitable for the dual functions of the endosymbiont and reconcile its substantial contributions in the gut.

Keywords: endosymbiosis; spirochetes; single-cell genomics; adaptive evolution; metabolic interaction

Significance

Termites thrive on nutritionally unbalanced and recalcitrant woody cellulose owing to their complex microbial community in the gut. Although key functions of the entire gut community have been characterised, the responsible microbes and their metabolic interactions remain unclear. We present a novel endosymbiotic relationship between termite-gut cellulolytic eukaryotic protists and spirochete bacteria through a combined investigation of biochemistry, microbial ecology, and genomics of the endosymbionts that perform the dual functions of fixing CO_2 and N_2 using H_2 produced by the protists. The unveiled endosymbiotic relationship, though seemingly still in an initial developmental stage, is efficient for the abilities of both partners, and substantially benefits the carbon, nitrogen, and energy metabolism of the host termite.

Endosymbiotic associations between eukaryotic cells and bacteria, in which these partners share their unique abilities, have a profound impact on the ecological adaptation and niche expansion. In the gut of termites, there are various examples of species-specific symbiotic associations between protists (single-cell eukaryotes) and bacteria, and if not all, they have cospeciated (1–6). The social behavior of termites has promoted the stable and sustained vertical inheritance of gut microbes between termite generations, which may be advantageous for the emergence and development of these symbiotic associations in the gut community (7). The gut microbial community is responsible for the utilisation of nutritionally unbalanced and recalcitrant woody cellulose (6, 8). Owing to this ability, termites are a keystone of global carbon cycles and economically important as pests to timber constructions, and are expected for biotechnological application in producing biofuel from cellulosic biomass (5, 6, 8). However, the gut community comprises unique and diverse species, which are mostly yet-uncharacterised due to our historical inability to culture them in the laboratory.

In wood-feeding 'lower' termites that harbour unique flagellated protists in their guts, these protists play a central role in cellulose digestion. The protists phagocytose ingested wood particles and almost completely decompose and ferment the cellulose to produce acetate, H_2, and CO_2 (6, 8). The host termite utilises the produced acetate as a major carbon and energy source. The produced H_2 is a key metabolic intermediate that fuels many bacteria in the gut (9). The gut bacteria are also important for the nutrition of the host termite, carrying out both CO_2-reducing acetogenesis (hereafter, reductive acetogenesis) and nitrogen fixation (6, 10). The H_2 and CO_2 produced from the cellulose fermentation are converted by the gut bacteria via reductive acetogenesis to acetate, which accounts for up to one-third of the carbon and energy demand of the host termite (11). Termites thrive on dead wood extremely poor in nitrogen, so the fixation of atmospheric N_2 is crucial for acquisition of nitrogen for the termite (10). The responsible bacteria of these two functions are inferred based on the characterisations of cultured isolates of spirochetes in the genus *Treponema*, the most abundant constituent of the gut bacterial microbiota in many wood-feeding termites (10). One species *Treponema primitia* is known to carry out reductive acetogenesis and another *Treponema azotonutricium* is a nitrogen fixer (12–15). Indeed, gene-based analyses suggest that treponemes are often major contributors for reductive acetogenesis, and to a lesser extent

nitrogen fixation, in the gut microbial community (16–19).

The genome sequences of the two bacterial species that are endosymbionts of two different gut protists have been reported and their roles are inferred (20, 21). These endosymbionts have the genetic potential to utilise sugars produced during the cellulose degradation in the protist cells, and instead, to upgrade nitrogenous nutrients for the host protists. In one of these endosymbionts, an ability of nitrogen fixation is also predicted (20). Furthermore, gene inventory studies suggest the importance of protist-associated bacteria for H_2 metabolism or nitrogen fixation in the termite guts (19, 22). Therefore, the protist-associated bacteria are very likely to play important roles in the gut metabolism, but their actual activities and contributions are rarely evaluated and our knowledge of the symbiotic relationships is still fragmentary.

In this study, we localised activities of both reductive acetogenesis and nitrogen fixation in the gut of a wood-feeding termite *Hodotermopsis sjoestedti* (Termopsidae), and identified key genes in the responsible bacterial species. We demonstrate that both functions were attributed to a single *Treponema* species endosymbiotic for a cellulolytic protist in the genus *Eucomonympha* (phylum Parabasalia). To further understand this endosymbiotic relationship, we examined the single-cell genome sequences of the endosymbiont and reconstructed its metabolic ability.

Results

Fractionation of Reductive Acetogenesis. To localise reductive acetogenesis activity in the gut community, the reduction of $^{14}CO_2$ to acetate was measured. Using differential low-speed centrifugation technique, the gut contents were fractionated into the large protist-enriched fraction and the fraction containing small protists and freely swimming bacteria (small protist/bacteria fraction). The former fraction was extensively washed to remove free-swimming bacteria. Because reductive acetogenesis is sensitive to oxygen, all the sample preparations and assays were performed under anoxic conditions. The large-protist fraction constituted up to 63% of the whole-gut activity, while only less than 15% was detected in the small protist/bacteria fraction (Table 1). The supply of H_2 stimulated reductive acetogenesis activity in all the fractions. Enzymes in the acetyl-CoA (Wood-Ljungdahl) pathway for reductive acetogenesis (23) showed a higher level of

activity in the large protist fraction (Table S1).

The large-protist fraction almost exclusively consisted of the cellulolytic protists of the genera *Eucomonympha* and *Trichonympha*. The ^{14}C radio isotope assay under the presence of exogenously supplied H_2 was sensitive enough to allow the analysis of pools of manually isolated cells of each protist species independently. Strong activity was observed for *Eucomonympha* (9.5±2.6 pmol/hr/cell, n=3), whereas only weak activity was detected in *Trichonympha* (0.31±0.27 pmol/h/cell, n=3). The activity associated with *Eucomonympha* was estimated to be approximately 58% of the whole gut. Thus, the substantial activity of reductive acetogenesis in the gut was associated with the *Eucomonympha* protists.

Genes involved in the acetyl-CoA pathway, *acsA* (also known as *cooS*) and *acsB* encoding two subunits of the acetyl-CoA synthase /carbon monoxide dehydrogenase complex (ACS/CODH) respectively, a gene encoding formyltetrahydrofolate synthetase, and a gene *acsF* encoding an Ni-insertion protein to ACS/CODH were successfully PCR-amplified and identified from associated bacteria of manually isolated *Eucomonympha* cells (*SI Text*). Transcribed sequences corresponding to the identified *acsA* and *acsB* genes were detected in RNA extracted from the gut community (*SI Text*). An *in situ* hybridisation experiment against mRNA of the *acsB* gene confirmed expression and specific localisation to the *Eucomonympha* cells (Fig. S1).

Nitrogen Fixation Associated with *Eucomonympha* Protist. Nitrogen fixation activity was investigated after fractionation of the gut contents as in the reductive acetogenesis assay described above. The acetylene reduction assay for living termites showed an activity of 171.3±92.5 nmol/hr/g fresh-weight termite (n=4). Substantial activity was observed in the large protist fraction (105.5±27.4 nmol/hr/g fresh-weight termite; n=6), whereas no activity was detected at all in the small protist/bacteria fraction (n=4). The nitrogen fixation ability was then examined with $^{15}N_2$ stable isotope incorporation in the cell mass (Table 2), and again, the activity was detected with the large protist fraction but not with the small protist/bacteria fraction. When the termites were fed on starch, the condition known for cellulolytic protists to disappear, no acetylene reduction was

detected. The results indicate that bacteria associated with large cellulolytic protists are responsible for nitrogen fixation in the gut.

Nitrogenase, the enzyme catalysing nitrogen fixation, is generally inhibited by the presence of H_2 (24). Considering this inhibitory effect of H_2, nitrogen fixation in the termite gut is enigmatic because H_2 is known to accumulate at a high partial pressure in the gut lumen (6, 9). Therefore, we examined the effect of H_2 on ^{15}N-incorporating nitrogen fixation activity (Table 2; compare the activities of the large protist fraction under $^{15}N_2$ and $^{15}N_2$+H_2). The activity detected in the large protist fraction was not inhibited, but rather seemingly stimulated, by the presence of H_2. Acetylene reducing activity in the large protist fraction under the presence of 20% H_2 also significantly increased (133.6±40.2 nmol/hr/g fresh-weight termite; n=4) when compared with that under the presence of argon gas instead (62.5±13.3 nmol/hr/g fresh-weight termite; n=3).

Nitrogen fixation associated with large cellulolytic protist species was further investigated by the presence of a nitrogen fixation gene, *nifH*, which encodes nitrogenase reductase and is used as a molecular marker for nitrogen fixing bacteria in termite guts (18, 22, 25). The *nifH* gene sequence was successfully amplified by PCR with manually isolated cells of *Eucomonympha*, whereas no amplification was detected with isolated *Trichonympha* cells. The control amplifications of bacterial 16S rRNA gene sequence were detected in both protist species. The results strongly suggest that the *Eucomonympha*-associated bacteria are responsible for the nitrogen fixation activity in the gut.

The *nifH* sequences of the associated bacteria of *Eucomonympha* belonged to a so-called alternative nitrogenase group of *nifH* (hereafter designated as *anfH*) represented by the previously reported HSN10 sequence (AB011910) from the gut microbial community of *H. sjoestedti* (25). The repertoire of transcribed *nifH* genes were investigated by RT-PCR using RNA extracted from the gut community. Although PCR often introduces some biases depending on the primer applied, the majority of the detectable transcribed sequences corresponded to the *anfH* sequence of the *Eucomonympha*-associated bacteria (23 of 25 clones). The specific localisation of the *anfH* mRNA with *Eucomonympha* cells was confirmed by *in situ* hybridisation (Fig. S1). The other transcribed sequences belonged to an ordinary nitrogenase group of *nifH* and encoded identical protein to the

previously reported HSN20 sequence (AB011918) (25). No *nifH* or *anfH* sequence has been detected so far from the gut community of this termite except for those of the HSN10 and HSN20 groups, and their corresponding gene sequences were found in the same genome (see below). The results support the major contribution of the *Eucomonympha*-associated bacteria to nitrogen fixation in the gut.

Identification of *Eucomonympha* Endosymbiont. *Eucomonympha* cells were numerous in the gut, observed at 1,935±206 cells per termite, with each *Eucomonympha* cell harboring a dense population of rod-shaped endosymbiotic bacteria amounted to $>10^4$ bacterial cells per single protist cell (Fig. 1). Therefore, the protist and its bacterial endosymbionts represent abundant species in the gut community. The bacterial endosymbionts of *Eucomonympha* were observed to be separated from the host protist cytoplasm by an electron sparse inter-membrane space and often found in the proximity of hydrogenosomes-like organelles surrounded by a single membrane.

PCR-amplified bacterial 16S rRNA gene sequences from carefully isolated *Eucomonympha* cells were clonally analysed and a single clone group with minimal sequence variation (<1.3% nucleotide difference) was obtained (*SI Text*). The sequence information was used to design specific probes for fluorescence *in situ* hybridisation (FISH). The FISH gave a specific signal to the majority of rod-shaped endosymbiotic bacteria of *Eucomonympha* (Fig. 1b-e, and *SI Text*). An analysis of bacterial 16S rRNA gene sequences in the gut community showed that the endosymbionts were the most abundant bacteria (21 of 218 clones; Fig. S2), supporting its dominance in the gut. Phylogenetic analyses (Fig. 2) revealed that the sequences of the endosymbionts were derived from spirochetes of the genus *Treponema* and belonged to the previously defined termite *Treponema* cluster II (26), which consists of the ectosymbionts attached to the cell surface of gut protists as far as their localisations were examined (27, 28). This cluster formed a clade with saccharolytic treponemes that contained *Treponema brennaborense*, an isolate from a digital dermatitis ulcer of a cow (29), as a close cultured relative, but distantly related to the termite *Treponema* cluster I that comprises isolates and clone sequences from termite guts. We propose a novel species, '*Candidatus* Treponema intracellularis' for this endosymbiont (*SI Text*).

Genome of the Endosymbiont. Single cells of the endosymbiotic treponeme species of *Eucomonympha* were sorted, and after whole genome amplification, the genome sequences of five single cells were examined individually. These five single cells showed less than 0.8% nucleotide difference in their 16S rRNA gene sequences, well within a species-level difference. The completeness of the single-cell genomes, judged on the presence of conserved single copy genes (30), ranged from 40 to 85% (Table S2). The estimated genome sizes were 2.67–3.21 Mb, which was slightly smaller than the genome size of the isolated *Treponema* strains from the termite gut (3.79–4.06 Mb) and nearly equivalent to *T. brennaborense* (3.06 Mb). The GC content was 55.2±0.28.

Among the five single cell genomes, all the acetyl-CoA pathway genes were compositely detected, although no single genome harbored all these genes (Table. S3). All acetyl-CoA pathway genes encoded homologous proteins to those of *T. primitia* and those detected in the termite-gut microbial communities (Table S3 and Figures S3–S6). Two gene sets encoding alternative and ordinary nitrogenases and genes involved in metal cofactor assembly for nitrogenase were also found in the single-cell genomes (Fig. S7). The encoded alternative nitrogenase subunits were closely related to those of *T. primitia* strain ZAS-1. In contrast, the ordinary nitrogenase subunits and the metal cofactor assembly proteins were distantly related to those found in the genomes of *Treponema* species and rather they were similar to those of some firmicutes (Fig. S8 and Fig. S9), implying the acquisition of the genes by a horizontal gene transfer.

Analyses of the single-cell genomes suggested that the endosymbiotic treponeme of *Eucomonympha* is able to utilise glucose, maltose, mannose, fructose, and glucuronate, and ferment them through glycolysis to produce acetate, although no gene for enolase in the glycolytic pathway was found (Table. S4). The genomes have genes for synthesis and utilisation of glycogen for storing energy-rich carbohydrates. The presence of urease and ammonium transporter genes in the genomes suggested that the endosymbionts are enabled to utilise urea and ammonia. The fixed or utilised nitrogen is likely assimilated first by glutamine synthetases, and then utilised for biosynthesis of nitrogenous nutrients. The genomes were found to compositely harbour most of the genes involving biosynthesis of amino acids and many genes for biosynthesis of cofactors (Fig. S10).

Discussion

The dual functions of reductive acetogenesis and nitrogen fixation discovered in the bacterial endosymbiont specis of the *Eucomonympha* protist, together with the cellulolytic ability of the protist, clearly benefit the host termite in terms of the efficient carbon, nitrogen, and energy metabolism. The cells of the cellulolytic protist occupy a large volume of the gut and their cytoplasm provides a considerable space and a safe niche for the bacteria to proliferate. Indeed the endosymbiont species is predominant in the gut community, which reconciles its contribution to the substantial biochemical activities in the gut. Probably, the functional potentials for both reductive acetogenesis and nitrogen fixation optimally match the intracellular niche provided by the cellulolytic host. It is noted that the previously isolated termite-gut treponemes exhibit only either one or the other of the two phenotypes (12–14) though their genomes predict these potentials. Therefore, there should be important endosymbiotic interactions suitable for these dual functions (see Fig. 3).

The host *Eucomonympha* protist degrades cellulose, and as in other parabasalian protists, very likely ferments it to acetate, H_2 and CO_2. H_2 is particularly important as the sink of reducing equivalents produced during the fermentation and effective removal of H_2 theoretically enhances the fermentation of cellulose. Therefore, endosymbionts that are capable of utilising the produced H_2 and CO_2 for reductive acetogenesis would help enhance cellulose decomposition by consuming H_2 and gain its own energy source. However, availability of H_2 may not be a limiting factor for gut bacteria as H_2 partial pressures are high throughout the gut lumen in most termites investigated so far (6, 9) and indeed exogenous supply of H_2 does not stimulate reductive acetogenesis in intact guts (31). Nevertheless, H_2 consumption within the H_2-producing protist cells may have some impacts on the protist metabolism. As intracellular populaions of gut protists, methanogens that produce methane presumably from H_2 plus CO_2 are known well and distributed widely, but generally limited only to smaller protist species (32). Indeed in the termite *H. sjoestedti*, associated methanogens are observed with the gut small protist species (33). Possibly, the niche segregation of reductive acetogens and methanogens determines their relative activities in the gut.

Nitrogen fixation by the endosymbiont is clearly advantageous for the host protist not

only because the protist can utilise the fixed nitrogen but also because the protist can monopolise the supply of more valuable nitrogenous nutrients converted by the endosymbiont such as amino acids directly usable for the host protein synthesis. This is expected to give the protist a selective advantage and allow the protist to grow efficiently and stably, independent of other gut bacteria, which in turn, benefits the host termite for stable maintenance of the essential cellulolytic protists. As shown in this study, the nitrogen fixation activity of the endosymbiont is resistant to H_2. This nature, probably as the consequence of adaptive evolution in this niche where H_2 accumulates abundantly, is required for the endosymbiosis within the protist cell. The nitrogen fixation reaction itself is also known to produce H_2. Possibly, H_2 serves as a reductant of the nitrogenase of the endosymbiont through its hydrogenase function as in the case of CO_2-reduction in acetogenesis because the activity is seemingly stimulated by the presence of H_2, although the electron transfer mechanism needs further studies.

The endosymbiotic treponeme of *Eucomonympha* could utilise sugars and uronates, both abundant substrates available from the (hemi)cellulose degradation in the protist cell. This ability suggests the mixotrophic nature of simultaneous utilisations of H_2 plus CO_2 for reductive acetogenesis and sugars (or uronates) for fermentation because these substrates coexis and are available abundantly for the endosymbiont unless the host termite starves. Such mixotrophy is shown in *T. primitia* ZAS-2, which can use maltose and H_2 plus CO_2 simultaneously to increase cell yield and acetate production than growth with either substrate alone (34). The mixotrophic nature of the endosymbiont is especially advantageous for satisfying the high-energy demand of the nitrogen fixation reaction. Acetogenesis solely from H_2 plus CO_2 is energy conservative (23) but the net energy yield may not be enough for the high level of nitrogen fixation activity that supports the nitrogen demands of the gut entire community and the host termite. The almost exclusive contribution of the endosymbiont species to the gut nitrogen fixation activity indicates that the intracellular location is an ideal niche for this function, and once endosymbiotic nitrogen fixation established, other nitrogen fixers may have become obsolete with gradual loss of this ability.

Spirochetes are generally characterised by their conspicuous spiral morphology and vigorous motility. Flagella located in the periplasmic space are wrapped around the

cytoplasm to produce the unique helical locomotion. However, the rod-shaped morphology of the *Eucomonympha* endosymbiont is unique among spirochetes, although non-helical spirochete species of the genus *Sphaerochaeta* have been isolated from the termite gut and other environments (35, 36). By localising within the protist cell, motility has become unnecessary, which likely has caused the dispensing of its spiral morphology, since both the motility and the spiral morphology are tightly linked (37). Indeed in the TEM observations of the endosymbiont cells, no periplasmic or other flagellum was detected (see Fig. 1a). Furthermore, there are no genes for flagellum and its biosynthesis found in the genome sequences of the endosymbiotic treponeme (*SI text*).

In contrast to the endosymbionts of termite-gut cellulolytic protists in the genus *Pseudotrichonympha* and *Trichonympha* , both of which have smaller genomes (1.1 Mb in each case) rich in pseudogenes and thus are suggested their ongoing genome erosion (20, 21), the estimated genome size of the endosymbiont of *Eucomonympha* does not look severely reduced. This is probably because the dual endosymbiotic roles described in this study require a larger gene repertoire and metabolic networks. Nevertheless, analyses of cluster of orthologous groups (COGs) of protein show that the endosymbiont of *Eucomonympha* are more similar to the two endosymbionts of termite-gut protists in the gene content of some COGs than the cultured treponemes (Fig. S11). Commonly in all the endosymbionts, relative gene abundance in COGs of 'carbohydrate metabolism and transport' and 'signal transduction mechanisms' as well as 'cell motility' decreased, and that in 'translation' and 'coenzyme metabolism' increased, when compared with cultured treponemes. Given the genome size and gene content, it is possible that the acquisition of the endosymbiont by *Eucomonympha* was a more recent event and that the endosymbiont is in an initial developmental stage but in a dynamic process of adaptive evolution as an endosymbiont. To support this, there are transposable elements in high densities in the genomes (e.g. putative transposases amounted to 5.5–9.3% coding sequences, Table S3). Such sequences are often observed in symbionts and pathogens in initial stages of genome reduction (38).

The protist species *Teranympha mirabilis* in the gut of the termite *Reticulitermes speratus* is the closest known relative of *Eucomonympha* spp. in *H. sjoestedti* (39). From the isolated *T. mirabilis* cells, we successfully identified sequences of bacterial 16S rRNA

gene and genes involved in reductive acetogenesis and nitrogen fixation, all of which were closely related to those of the endosymbiont of *Eucomonympha* (*SI text*). The results suggest that a common ancestor of *Eucomonympha* and *Teranympha* acquired an endosymbiotic treponeme species that had the dual functions. The *Eucomonympha-Teranympha* lineage is phylogenetically sister to *Pseudotrichonympha* (39), but their endosymbionts are completely different. The *Pseudotrichonympha* endosymbiont belongs to the order Bacteroidales (40). It is possible that, after the two lineages separated, they have acquired the endosymbionts independently or that the endosymbiont of one lineage has been replaced by another. In either case, the endosymbionts of both lineages seem to share many functions such as nitrogen fixation, utilisation of urea and ammonia, glycogen metabolism, and H_2 consumption, though H_2 is not used for reductive acetogenesis in the endosymbiont of *Pseudotrichonympha*, in addition to upgrading nitrogenous nutrients and the utilisation of sugars which are also common to the genome-sequenced endosymbiont of *Trichonympha* that belongs to the phylum Elusimicrobia (21). These common features suggest that different bacterial species have convergently established the similar intracellular niches of the cellulolytic protists and that these functions are important for the establishment of the endosymbiotic relationships. In the case of the *Trichonympha* endosymbiont, the sequenced genome carries neither acetyl-CoA pathway genes nor nitrogen fixation genes. However, associations of the second bacterial symbionts which are different species depending on the *Trichonympha* lineages have been reported, and they likely involve H_2 metabolism (19, 41, 42). One of them belong to the genus *Desulfovibrio* and genes for sulfate reduction and hydrogenase are identified although the sulfate concetraion in the guts is usually low (41). Another is an unclassified bacterial species in Deltaproteobacteria and the investigations of its FDH gene implys its importance for gut H_2 economy, possibly as a reductive acetogen (19).

The endosymbiont of *Eucomonympha* is related to lineages of ectosymbiotic treponemes of termite-gut protists. This phylogenetic relationship implies an evolution from ecto- to endosymbioses, as is also suggested in Bacteroidales endo- and ectosymbionts of termite-gut protists (43). The ectosymbiotic attachments of treponemes onto the protist cells are often observed (27, 28) and they are hypothesised to catalyse reductive acetogenesis (12). Termite-gut microbial communities harbor a variety of endo- and ecto-symbiotic relationships, and thus provide attractive models for comparative studies in order to

understand how these symbiotic relationships have been established, adapted, and coevolved. Such studies using empirically testable models of endosymbioses can give us fruitful indications for contentious evolutionary trajectories of eukaryotic cells and their organelles from ancient endosymbioses (44).

Materials and Methods

Analytical methods. The termite specimens were collected and maintained as described previously (25, 27). Termites were introduced into an anaerobic chamber (Bactron model II, Sheldon Manufacturing Inc) with an O_2-free atmosphere of N_2/ H_2/CO_2 (80/10/10, v/v/v) and all the following sample preparations were performed under the anoxic condition at ambient temperature. Guts were dissected in 0.5–1.0ml of buffered saline solution (16.2mM K_2HPO_4, 10.35mM KH_2PO_4, 36.75mM NaCl, 32.25mM KCl, 0.795mM $CaCl_2$, 7.95mM $MgCl_2$, 1.0mM dithiothreitol). Gut debris was removed by filtration through nylon mesh. The gut content was centrifuged at 23 ×*g*, and after five rounds of the centrifugation, the resulting supernatant was used as the small flagellate/bacteria fraction. The cell pellet of the first centrifugation was washed 7 to 15 times with the same buffer and used as the large flagellate fraction. Using a micromanipulation system (Transferman NK2, Eppendorf) operated in the anaerobic chamber, two hundred cells of each *Eucomonympha* and *Trichonympha* protist were collected from a large protist fraction prepared as described above but with the buffer containing 3.25mM reduced glutathione instead of 1.0mM dithiothreitol. After sealing reaction vials with butyl rubber stoppers and removing from the anaerobic chamber, the gas phases were exchanged to H_2 or N_2 for the fractions and to H_2 for the isolated protist cells. The reductive acetogenic activity was measured as described previously (11) with slight modifications. Reaction vials (4.2 ml) had a final liquid volume of 0.25 to 0.50 ml and contained 1.95 μmol/ml concentration of $NaH^{14}CO_3$ (specific activity, 0.117MBq/μmol) and was incubated at 30°C. The incubation time was 2.5–5 hours for the fractions and 24–72 hours for the isolated protist cells. An aliquot of the reaction sample (10–50 μl) was analysed by HPLC equipped with an ion-exclusion reverse phase column (Shodex RSpak KC-811) and an on-line flow scintillation analyzer (Ramona 2000 Raytest). Elution was isocratic with 0.1% H_3PO_4 (flow rate 1.0ml/min) at 50°C. Under

this condition, acetate, formate, lactate, succinate and other acids were clearly separated. The buffer with glutathione was also used for the fractionation and the assay of the fractions and the activities showed the similar patterns and equivalent levels to those with the buffer containing dithiothreitol instead.

Acetylene reduction assay of a living termite specimen was performed as described previously (25). For the fractionated samples, the gas phase of the cell suspension was replaced first with 100% N_2 and then with N_2 gas containing 16% acetylene. After 30 min to 3 hours incubation, 0.1ml of gas was assayed as described previously (25). The termites were fed on soluble starch (Nacarai Tesuque) moistened with distilled water for 7–9 days and acetylene reduction was also assayed as described above. Microscopic oberservations of gut contents of the starch-fed termites confimed that large protists were almost completely disappeared but some small protists and spirochete-like bacteria were retained. For the ^{15}N incorporation assay, the gas phase of the cell fraction in 10 ml vial was composed of $^{15}N_2$ (99.7 atom%) or Ar or N_2. After 20 hours incubation at ambient temperature, the cells were harvested and dried at 70°C. ^{15}N concentrations of the dried samples were determined in duplicate measurements by an automatic gas chromatograph-mass spectrometer (EA1110-DELTA plus Advantage ConFloIII system, Thermo Finnigan). Atom % excess was the difference between sample and reference and $\delta\ ^{15}N$ (‰) was obtained with the notation $\{(R_s - R_r)/R_r\}\times 1{,}000$, where R_s and R_r are atomic ratio of sample and reference, respectively. As the reference, the atomospheric nitrogen was used (^{15}N atom % was 0.366).

Gene identification and analyses. The protist cells were manually isolated using a micromanipulator as described previously (27, 28), and used as templates for PCR of respective genes (*SI Text*). The PCR products were cloned and their DNA sequences were analysed. Analyses of *nifH* genes were conducted as described previously (45). Maximum likelihood phylogenetic analyses were conducted with the unambiguously aligned sequences using RAxML MPI version 8.1.2 (46). Confidence of tree topology was estimated with bootstrap analyses of 500 resamplings. FISH and transmission electron microscopy (TEM) were conducted as described previously (27, 40).

Single-cell genome sequencing and analyses. Manually isolated *Eucomonympha* cells were pooled, ruptured, and treated with DNase I (Promega). After staining with

CellTracker™ Green CMFDA (Life Technologies), single bacterial cells were sorted using a MoFlo XDP (Beckman Coulter) as described previously (47). The genome DNA of the sorted cells was amplified with the REPLI-g UltraFast Mini kit (Qiagen). After checking the identity and purity based on the PCR-amplified 16S rRNA gene sequence, the genome sequencing was conducted on the Illumina MiSeq platform. The generated quality-passed sequence reads were adapter and quality trimmed and assembled using SPAdes 3.1.1 (48). After checking the assembled sequences in the context of GC content and coverage using the blobology bash (49), the reads were filtered and reassembled using the same program. The resulting sequences were annotated with the PROKKA pipeline (50) and uploaded to the RAST server (51), and the annotations were checked manually with BLAST searches against public databases.

ACKNOWLEDGMENTS. We thank members of previous RIKEN Environmental Molecular Biology Laboratory, T. Kudo, H. Yuzawa, T. Sato, and other technical staff for supporting this work. D.S. was a RIKEN International Program Associate under a joint graduate school program with the University of Liverpool. This work was supported in part by a grant for Precursory Research for Embryonic Science and Technology from Japan Science and Technology Agency (to M.O.), Grants-in-Aid for Scientific Research from Japan Society for Promotion of Science (JSPS), Nos. 19380055, 23117003, and 26292047 (to M.O.), and a fund for Next Generation World-Leading Researchers from JSPS (to Y.H.).

References

1. Noda S, et al. (2007) Cospeciation in the triplex symbiosis of termite gut protists (*Pseudotrichonympha* spp.), their hosts, and their bacterial endosymbionts. *Mol Ecol* 16(6):1257–1266.
2. Ohkuma M (2008) Symbioses of flagellates and prokaryotes in the gut of lower termites. *Trends Microbiol* 16(7):345–352.
3. Ikeda-Ohtsubo W, Brune A (2009) Cospeciation of termite gut flagellates and their bacterial endosymbionts: *Trichonympha* species and '*Candidatus* Endomicrobium

trichonymphae'. *Mol Ecol* 18(2):332–342.

4. Desai MS, et al. (2010) Strict cospeciation of devescovinid flagellates and *Bacteroidales* ectosymbionts in the gut of dry-wood termites (*Kalotermitidae*). *Environ Microbiol* 12(8):2120–2132.
5. Hongoh Y (2011) Toward the functional analysis of uncultivable, symbiotic microorganisms in the termite gut. *Cell Mol Life Sci* 68(8):1311–1325.
6. Brune A (2014) Symbiotic digestion of lignocellulose in termite guts. *Nat Rev Microbiol* 12(3):168–180.
7. Lo N, Eggleton P (2000) Termite phylogenetics and co-cladogenesis with symbionts. *Biology of Termites: A Modern Synthesis*, eds Bignell DE, Roisin Y, Lo N (Springer, Dordrecht), pp 27–50.
8. Watanabe H, Tokuda G (2010) Cellulolytic systems in insects. *Annu Rev Entomol* 55:609–632.
9. Pester M, Brune A (2007) Hydrogen is the central free intermediate during lignocellulose degradation by termite gut symbionts. *ISME J* 1(6):551–565.
10. Breznak JA (2000) Ecology of prokaryotic microbes in the guts of wood- and litter-feeding termites. *Termites: Evolution, Sociality, Symbioses, Ecology*, eds Abe T, Bignell DE, Higashi M (Kluwer Academic Publishers, Dordrecht), pp 209–231.
11. Breznak JA, Switzer JM (1986) Acetate synthesis from H_2 plus CO_2 by termite gut microbes. *Appl Environ Microbiol* 52(4):623–630.
12. Leadbetter JR, Schmidt TM, Graber JR, Breznak JA (1999) Acetogenesis from H_2 plus CO_2 by spirochetes from termite guts. Science 283(5402):686–689.
13. Lilburn TG, et al. (2001) Nitrogen fixation by symbiotic and free-living spirochetes. *Science* 292(5526):2495–2498.
14. Graber JR, Leadbetter JR, Breznak JA (2004) Description of *Treponema azotonutricium* sp. nov. and *Treponema primitia* sp. nov., the first spirochetes isolated from termite guts. *Appl Environ Microbiol* 70(3):1315–1320.
15. Rosenthal AZ, Matson EG, Eldar A, Leadbetter JR (2011) RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. *ISME J* 5(7):1133–1142.
16. Salmassi TM, Leadbetter JR (2003) Analysis of genes of tetrahydrofolate-dependent metabolism from cultivated spirochaetes and the gut community of the termite

Zootermopsis angusticollis. *Microbiology* 149(Pt 9):2529–2537.

17. Pester M, Brune A (2006) Expression profiles of fhs (FTHFS) genes support the hypothesis that spirochaetes dominate reductive acetogenesis in the hindgut of lower termites. *Environ Microbiol* 8(7):1261–1270.
18. Yamada A, Inoue T, Noda S, Hongoh Y, Ohkuma M. (2007) Evolutionary trend of phylogenetic diversity of nitrogen fixation genes in the gut community of wood-feeding termites. *Mol Ecol* 16(18):3768–3777.
19. Rosenthal AZ, et al. (2013) Localizing transcripts to single cells suggests an important role of uncultured deltaproteobacteria in the termite gut hydrogen economy. *Proc Natl Acad Sci USA* 110(40):16163–16168.
20. Hongoh Y, et al. (2008) Genome of an endosymbiont coupling N_2 fixation to cellulolysis within protist cells in termite gut. *Science* 322(5904):1108–1109.
21. Hongoh Y, et al. (2008) Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Proc Natl Acad Sci USA* 105(14):5555–5560.
22. Desai MS, Brune A (2012) *Bacteroidales* ectosymbionts of gut flagellates shape the nitrogen-fixing community in dry-wood termites. *ISME J* 6(7):1302–1313.
23. Ragsdale SW, Pierce E (2008) Acetogenesis and the Wood–Ljungdahl pathway of CO_2 fixation. *Biochim Biophys Acta* 1784(12):1873–1898.
24. Burgess BK, Wherland S, Newton WE, Stiefel EI (1981) Nitrogenase reactivity: insight into the nitrogen-fixing process through hydrogen-inhibition and HD-forming reactions. *Biochemistry* 20(18):5140–5146.
25. Ohkuma M, Noda S, Kudo T (1999) Phylogenetic diversity of nitrogen fixation genes in the symbiotic microbial community in the gut of diverse termites. *Appl Environ Microbiol* 65(11):4926–4934.
26. Ohkuma M, Iida T, Kudo T (1999) Phylogenetic relationships of symbiotic spirochetes in the gut of diverse termites. *FEMS Microbiol Lett* 181(1):123–129.
27. Iida T, Ohkuma M, Ohtoko K, Kudo T (2000) Symbiotic spirochetes in the termite hindgut: phylogenetic identification of ectosymbiotic spirochetes of oxymonad protists. *FEMS Microbiol Ecol* 34(1):17–26.
28. Noda S, Ohkuma M, Yamada A, Hongoh Y, Kudo T (2003) Phylogenetic position and *in situ* identification of ectosymbiotic spirochetes on protists in the termite gut. *Appl Environ Microbiol* 69(1):625–633.

29. Schrank K et al. (1999) *Treponema brennaborense* sp. nov., a novel spirochaete isolated from a dairy cow suffering from digital dermatitis. *Int J Syst Bacteriol* 49(1):43–50.
30. Rinke C, et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499(7459):431–437.
31. Tholen A, Brune A (2000) Impact of oxygen on metabolic fluxes and i*n situ* rates of reductive acetogenesis in the hindgut of the wood-feeding termite *Reticulitermes flavipes*. *Environ Microbiol* 2(4):436–449.
32. Brune A (2010) Methanogens in the digestive tract of termites. *Microbiology Monographs vol 9: (Endo)symbiotic methanogenic archaea*, ed Hackstein JHP (Springer, Heidelberg) pp 81–100.
33. Tokura M, Ohkuma M, Kudo T (2000) Molecular phylogeny of methanogens associated with flagellated protists in the gut and with the gut epithelium of termites. *FEMS Microbiol Ecol* 33(3):233–240.
34. Graber JR, Breznak JA (2004) Physiology and nutrition of *Treponema primitia*, an H_2/CO_2-acetogenic spirochete from termite hindguts. *Appl Environ Microbiol* 70(3):1307–1314.
35. Dröge S, Fröhlich J, Radek R, König H (2006) *Spirochaeta coccoides* sp. nov., a novel coccoid spirochete from the hindgut of the termite *Neotermes castaneus*. *Appl Environ Microbiol* 72(1):392–397.
36. Ritalahti KM, et al. (2012) *Sphaerochaeta globosa* gen. nov., sp. nov. and *Sphaerochaeta pleomorpha* sp. nov., free-living, spherical spirochaetes. *Int J Syst Evol Microbiol* 62(Pt 1):210–216.
37. Motaleb MA, et al. (2000) *Borrelia burgdorferi* periplasmic flagella have both skeletal and motility functions. *Proc Natl Acad Sci USA* 97(20):10899–10904.
38. Moran NA, Plague GR (2004) Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev* 14(6):627–333.
39. Ohkuma M, et al. (2005) Molecular phylogeny of parabasalids inferred from small subunit rRNA sequences, with emphasis on the Hypermastigea. *Mol Phylogenet Evol* 35(3):646–655.
40. Noda S, et al. (2005) Endosymbiotic *Bacteroidales* bacteria of the flagellated protist *Pseudotrichonympha grassii* in the gut of the termite *Coptotermes formosanus*. Appl

Environ Microbiol 71(12):8811–8817.

41. Sato T, et al. (2009) *Candidatus* Desulfovibrio trichonymphae, a novel intracellular symbiont of the flagellate *Trichonympha agilis* in termite gut. *Environ Microbiol* 11(4):1007–1015.
42. Strassert JF, et al. (2012) '*Candidatus* Ancillula trichonymphae', a novel lineage of endosymbiotic *Actinobacteria* in termite gut flagellates of the genus *Trichonympha*. *Environ Microbiol* 14(12):3259–3270.
43. Noda S, Hongoh Y, Sato T, Ohkuma M (2009) Complex coevolutionary history of symbiotic Bacteroidales bacteria of various protists in the gut of termites. *BMC Evol Biol* 9:158.
44. McCutcheon JP, Keeling PJ (2014) Endosymbiosis: protein targeting further erodes the organelle/symbiont distinction. *Curr Biol* 24(14):R654–655.
45. Noda S, Ohkuma M, Usami R, Horikoshi K, Kudo T (1999) Culture-independent characterization of a gene responsible for nitrogen fixation in the symbiotic microbial community in the gut of the termite *Neotermes koshunensis*. *Appl Environ Microbiol* 65(11):4935–4942.
46. Stamatakis A 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
47. Shintani M, et al. (2014) Single-cell analyses revealed transfer ranges of IncP-1, IncP-7, and IncP-9 plasmids in a soil bacterial community. *Appl Environ Microbiol* 80(1):138–145.
48. 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.
49. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M (2013) Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front Genet* 4:237.
50. Seemann T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068–2069.
51. Aziz RK, et al. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75.

Tables and figures

Table 1. Fractionation of $^{14}CO_2$-reducing acetogenic activity of the *H. sjoestedti* gut contents

Fraction	Exogenous H_2		Endogenous H_2	
	Activity	Ratio (%)	Activity	Ratio (%)
Whole gut	0.52±0.23	100	0.23±0.12	100
Large protists	0.33±0.15	63	0.14±0.12	61
Small protists /bacteria	0.08±0.04	15	0.03±0.03	13

Activity (μmol/hr/g termite) is the mean ± standard deviation for n=7 experiments. The atmosphere in the reaction vials was replaced to 100% H_2 (exogenous H_2) or to 100% N_2 (endogenous H_2). In the latter condition, gut microorganisms endogenously supplied the reductant (e.g. H_2, formate, or other produced during their fermentation). The ratio of activity after fractionation was calculated with the activity in the whole gut as 100%. Total bacterial cells in the small protists/bacteria fraction were 3.7 times as many as protist-associated bacterial cells in the large protist fraction, whereas the amount of total protein in the former fraction was approximately one half of that in the latter fraction that contains large protists.

Table 2. Fractionation of $^{15}N_2$ incorporation and effect of H_2

Fraction	Gas phase	$^{15}N_2$ incorporation	
		δ ^{15}N (‰)	Atom % excess
Large protists	$^{15}N_2$	38.3	0.014
Large protists	$^{15}N_2$ + H_2	43.7	0.016
Large protists	$^{15}N_2$ + Ar	28.7	0.011
Large protists	Ar	-5.5	-0.002
Large protists	Ar + H_2	-5.5	-0.002
Large protists	–	-6.8	-0.003
Small protists/bacteria	$^{15}N_2$	-5.5	-0.002
Small protists/bacteria	Ar	-2.7	-0.001
Small protists/bacteria	–	-2.7	-0.001

^{15}N incorporation was measured in duplicate mass spectrometry measurements and the mean values were used for calclulations. The differences in the duplicate measures of the ^{15}N atom % was <0.3%. The reactions were prepared under N_2, and then the gas phase of the reaction vials was replaced with $^{15}N_2$ gas [$^{15}N_2$ (99.7 atom%):Ar = 4:6] or Ar gas, or none (indicated by –). In the experiments of $^{15}N_2$ + H_2, $^{15}N_2$ + Ar, and Ar + H_2 gas phases, 40% (vol/vol) of the gas phase was replaced with H_2, Ar, and H_2, respectively.

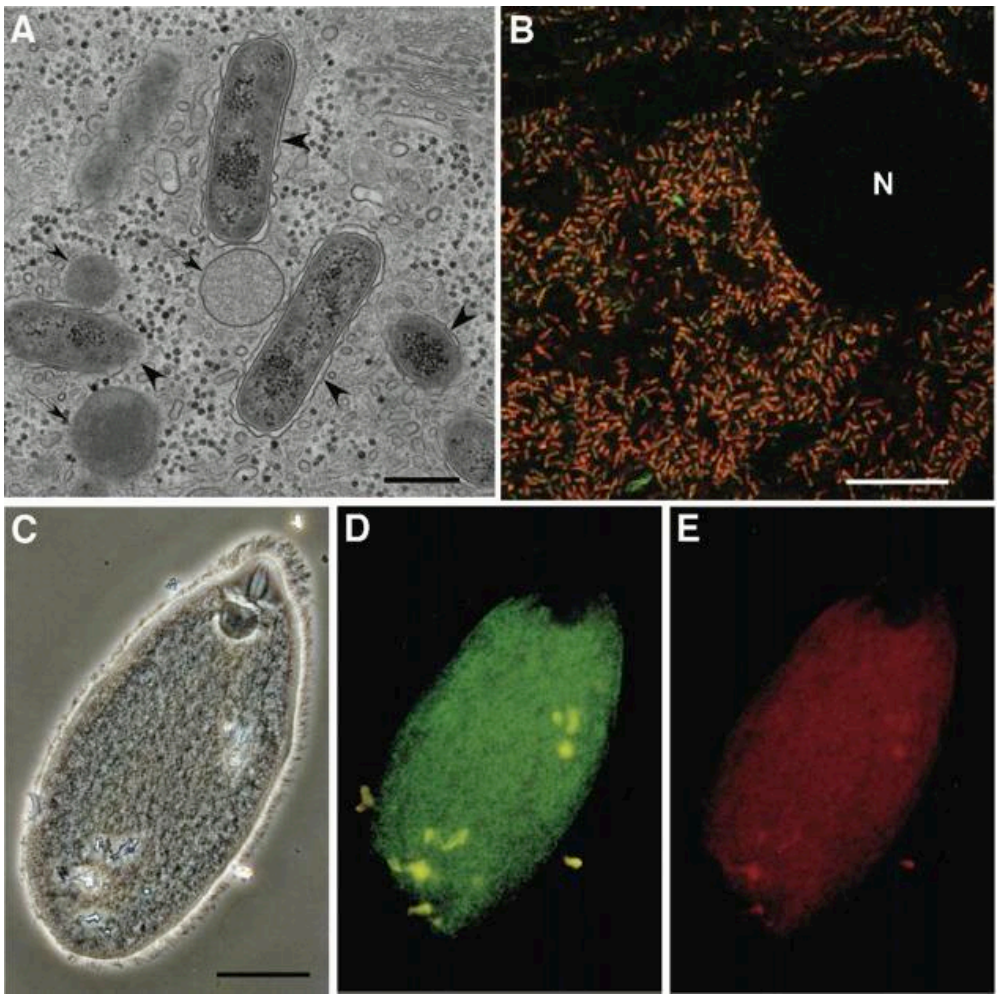

Fig. 1. Rod-shaped treponeme endosymbionts filling the *Eucomonympha* cell. (**A**), Transmission electron micrograph of the endosymbiont cells (arrowheads) within a *Eucomonympha* cell and its hydrogenosome-like organelles (asterisks). Scale bar, 0.5 μm. (**B**)-(**E**), FISH of the *Eucomonympha* endosymbionts (**C**), Phase contrast image of a *Eucomonympha* cell. Scale bar, 50 μm. (**B**), (**D**)-(**E**), Simultaneous hybridisations with two probes specific for the endosymbiont (**D**; in green) and with a general bacterial probe (**E**; in red). Amorphous yellow signals in (**D**) and corresponding red signals in (**E**) were derived from autofluorescence of ingested wood particles. (**B**), Merge image of laser-scanning confocal microscope of the endosymbiont cells. Almost all the intracellular bacteria were hybridised with both the specific and the general bacterial probes (in orange-yellow). N indicates the host nucleus where no bacterium is detected. Scale bar, 10 μm.

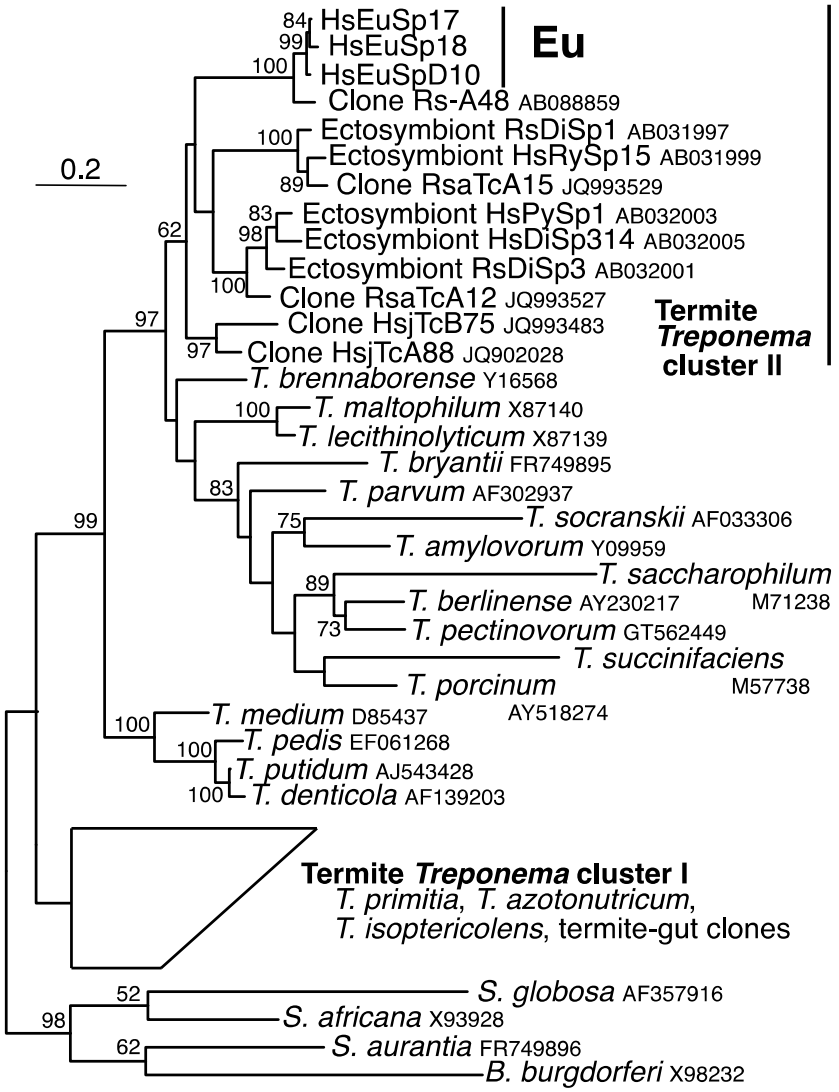

Fig. 2. Phylogenetic relationship of the *Eucomonympha* endosymbiont species and representatives of termite-gut and described treponemes, inferred based on the 16S rRNA gene sequences. The sequences identified from the endosymbionts of *Eucomonympha* are indicated by 'Eu'. The previously described termite *Treponema* clusters I and II (26) are indicated. Ectosymbiont sequences are identified from termite-gut oxymonad protists (27). Clones RsaTcA12, RsaTcA15, HsjTcA88, and HsjTcB75 are obtained from cell suspensions of *Trichonympha* protists (39). *T.*, *Treponema*; *S.*, *Spirochaeta* for *S. africana* and *S. aurantia*, and *Sphaerochaeta for S. globosa*; and *B.*, *Borrelia*. Bootstrap values over 50% are shown at nodes. Scale bar corresponds to 0.2 substitutions per nucleotide position.

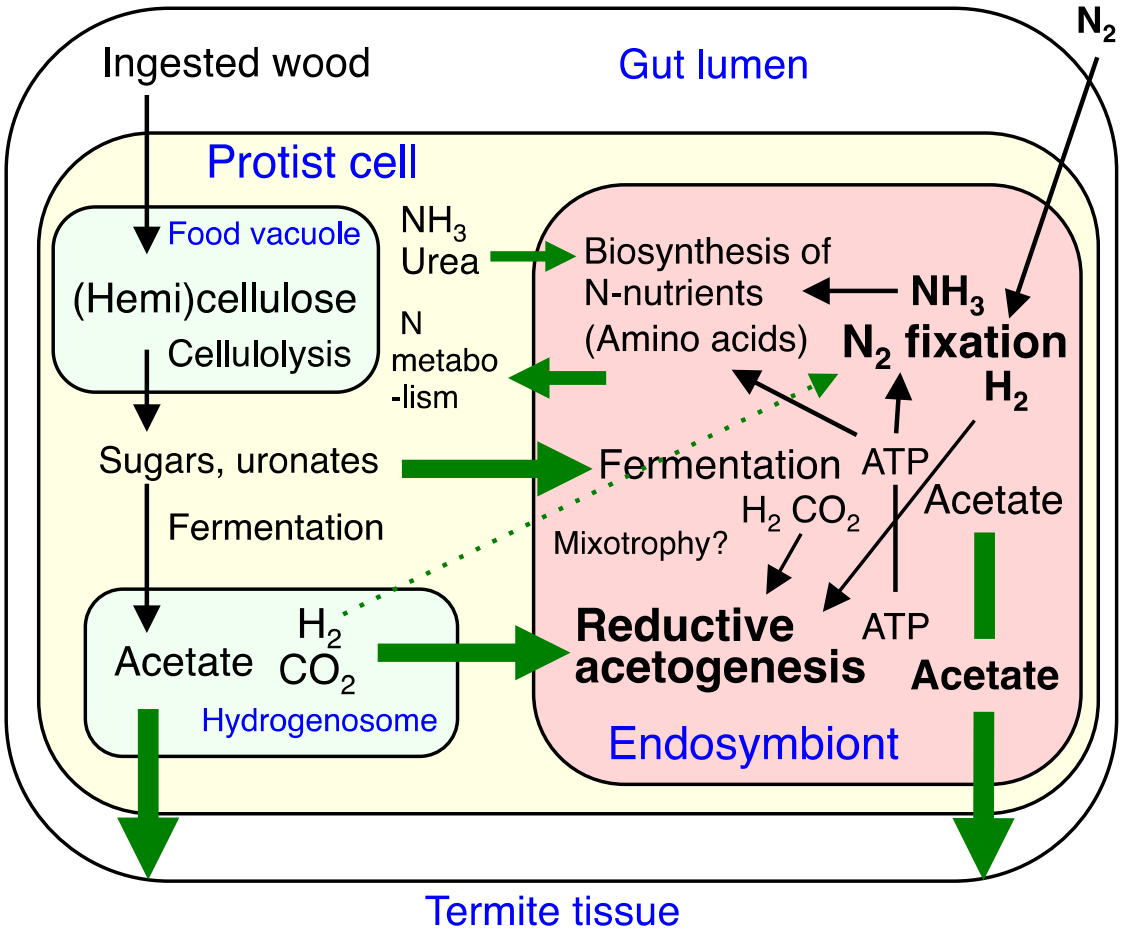

Fig. 3. A schematic view of the endosymbiotic relationships. The *Eucomonympha* protist endocytoses, degrades and ferments (hemi)cellulose to acetate, H_2 and CO_2 through metabolisms in food vacuole, cytoplasm, and possibly hydrogenosome. The produced H_2 and CO_2 are utilised by the endosymbiont for reductive acetogenesis. Probably, H_2 is also utilised by nitrogen fixation as the reductant (doted arrow). The degradation intermediates, sugars (such as glucose) and uronates are also utilised by the endosymbionts for fermentation. Termite absorbs acetate produced by both protist and endosymbiont as the major carbon and energy source. The endosymbiont fixes N_2 and utilises NH_3 and urea if available, and biosynthesises nitrogenous nutrients that are possibly supplied to the host protist. H_2 produced by nitrogen fixation and fermentation of the endosymbiont is probably re-utilised by reductive acetogenesis. The products of reductive acetogenesis and nitrogen fixation are shown in bold. The arrows in green indicate inter-species transfers of metabolic compounds.

Supporting Information

SI Text

Identification of the bacterial endosymbiont of *Eucomonympha*. To identify the endosymbiont phylogenetically, we first PCR-amplified and examined its 16S rRNA gene sequence using manually isolated *Eucomonympha* cells according to the method previously described (1). In two independent experiments, we obtained the sequences represented by HsEuSp17 and HsEuSp18. These sequences of the *Eucomonympha* endosymbiont were closely related to the previously reported partial sequence of the clone Hs32 (AB015826) (2), showing >98.1% identity. For the FISH identification, we designed two probes, IIC-404 (5'- TTTCGGCCTTCGTCATACAC) and IIC-637 (5'-ACTCTAGYCACCTAGTTCTC), specific for the 16S rRNA gene sequences for the *Eucomonympha* endosymbiont. When used individually, each of the two probes gave a weak but significant signal in almost all endosymbiont cells. When two probes were used simultaneously, stronger positive signals were obtained. With each probe no positive signal was given by free-swimming and ectosymbiotic spirochete-like bacteria in the gut. A previously designed probe for the termite *Treponema* cluster II (2) shares the identical target sequence with the HsEuSp17 and HsEuSp18 sequences, and indeed gave significant positive signals in the endosymbionts of many *Eucomonympha* cells. However, endosymbionts in some *Eucomonympha* cells gave no signal despite their abundance, therefore we isolated these *Eucomonympha* cells and identified the sequence HsEuSpD10, which was almost identical to the sequences of HsEuSp17 and HsEuSp18 (<1.3% nucleotide difference, respectively). We found a single nucleotide change in the probe-target region of the HsEuD10 sequence. In the termite *H. sjoestedti*, three similar but distinct lineages of *Eucomonympha*-like sequences of the eukaryotic small-subunit rRNA gene have been identified (3). The sequence divergence detected in the *Eucomonympha* endosymbiont species probably reflects their variation depending on the host protist lineages. The specific probes for termite *Treponema* cluster I (4) and any other examined here such as probes for the Elusimicrobia endosymbiont of *Trichonympha* protists (5) gave no hybridisation signal in the endosymbiont cells of *Eucomonympha*.

Description of '*Candidatus* Treponema intracellularis'. *Treponema intracellularis* (in.tra.cell.u.la'ris. L. adv. *intra*, within; M. L. n. *cellula* cell; L. adj. *intracellularis* within cells, referring to the presence of the organism within protist cells). Cells are short rods in 0.90-2.04 μm × 0.30-0.52 μm size (average 1.33±0.31 μm × 0.41±0.05 μm). Cells have no flagellum, and surrounded by two membranes. Probable for activities of reductive acetogenesis from H_2 plus CO_2 and nitrogen fixation. The bacterium is specifically found in the cytoplasm of the parabasalian protist *Eucomonympha* spp. in the hindgut of the termite *Hodotermopsis sjoestedti*. The assignment is based on the 16S rRNA gene sequences (accession nos. LC012866–LC012868) and hybridisation with the 16S rRNA-targeted oligonucleotide probes (5'- TTTCGGCCTTCGTCATACAC or 5'-ACTCTAGYCACCTAGTTCTC). The draft genome sequences of the five single cells are available (accession Nos. BBPV01000001–798, BBPW01000001–757, BBPX01000001–878, BBPY01000001–946, and BBPZ01000001–698).

Acetyl-CoA pathway genes of the endosymbiont of *Eucomonympha*. From manually isolated *Eucomonympha* cells, we successfully identified the gene fragments for the acetyl-CoA pathway using PCR primers designed in this study. The sequences of the primers were: 5'-RTIGGIATHTGYTGYACIGS (primer CbF3) and 5'-TTRTSIACRCAISWICCCATRTG (primer CbR2) for *acsA* (or *cooS*), 5'-HTITGYCARWSITTYGCICC (primer AaF1) and 5'-AYIRCYTCRAARCAICCRCA (primer AaR1) for *acsB*, and 5'-GGIGGIGCIRCIGGIGGNGG (primer FsF2) and 5'-ATRTTIGCRAAIGGICCICCRTG (primer FsR2) for formyltetrahydrofolate synthetase (FTHFS) gene. The PCR condition applied for the PCR was 35 cycles at 94ºC for 30s, 50ºC for 30s, and 72ºC for 2 min. The PCR produced a 701 bp, 281 bp, and 527 bp DNA fragments for *acsA*, *acsB* and FTHFS gene, respectively. After cloning and analyses, we identified the sequence highly similar to the corresponding gene of *T. primitia* in each gene case (see below). At the time of design of these degenerate primers, conserved amino acid sequences among available data for reductively acetogenic bacteria were carefully selected and the respective gene fragments were successfully amplified from *T. primitia* strain ZAS-2 (obtained as DSM 12427) and strains of other reductively acetogenic bacteria. However, after the gene sequences have appeared in *T. primitia* and *T. azotonutricium* strains (6), in *CooS* clones from termite guts (7) as well as in the

endosymbiont of *Eucomonympha* (this study), the primers for *acsA* and *acsB* genes were found to have mismatches with the target gene sequences. Nevertheless, using these primers, the sequences of the genes of the endosymbiont were indeed identified from the isolated protist cells and RNA extracted from the gut community (see below).

Because *acsA* and *acsB* are generally clustered in the genome, we amplified the DNA region between these two genes using specific primers (5'-ATTCCCGGMGCAATTCATGTT and 5'- GACGCCTGATTCACCACTTC), to obtain 3.4 kb DNA sequence. This sequence showed only less than 2.7% nucleotide differences in the overlapping regions with the *acsA* and *acsB* gene fragments. We found *acsF* between *acsA* and *acsB* in this 3.4 kb region. The encoding protein sequences of these genes were closely related to those of *T. primitia* strain ZAS-2 (6-8), showing amino acid identities of 89% for AcsA, 81% for AcsB, 75% for AcsF, and 88% for FTHFS. These PCR-amplified sequences were highly similar to the corresponding gene sequences in the single-cell genomes, showing less than 3.8% nucleotide difference. These PCR-amplified sequences have been deposited to DDBJ under accession nos. LC012869–LC012872.

The transcribed *acsA* and *acsB* genes were investigated by reverse transcription-PCR using RNA extracted from the gut community. After cloning and sequencing, highly similar sequences to those identified from the *Eucomonympha* cells (less than 3.0% and 2.9% nucleotide differences, respectively) were detected, although the abundance was not always high (4 of 30 clones and 8 of 21 clones, respectively) probably owing to the PCR biases. The localisation and expression of the identified *acsB* as well as *anfH* gene were further examined by *in situ* hybridisation (Fig. S1). Each of the *acsB* and the *anfH* fragments was cloned into a pGEM-T vector (Promega) for the probe preparation. After excision of the regions between the cloning site and T7 or SP6 promoters using its flanking restriction sites, each gene fragment was transcribed *in vitro* with T7 and SP6 RNA polymerases (Qiagen) in the presence of digoxigenin (DIG)-11-UTP with a DIG RNA labelling kit (Boehringer), to obtain sense and anti-sense RNA probes. The cells were fixed as described previously (9), and treated subsequently with 0.25N HCl, 1 mg/ml lysozyme, and 0.1 mg/ml Proteinase K for 30 min each. After dehydration in a series of ethanol, hybridisation was conducted as described previously (9, 10). Hybridisation

signals were detected with the alkaline phosphatase-conjugating anti-DIG antibody using a DIG nucleic acid detection kit (Boehringer).

Genes related to motility and outer membrane functions.

As shown in the COG profiles (Fig. S7), the genome sequences of the endosymbiont of *Eucomonympha* had no genes related to the "cell motility" function, which contains genes for flagellum and its biosynthesis. In the genomes of non-helical *Sphaerochaeta* species, *S. globosa* and *S. pleomorpha*, the absence of important genes is reported not only for flagellar proteins but also for methyl-accepting chemotaxis proteins, directly interacting with flagellar proteins and involving in the motility (11). In the annotations of the genomes of the endosymbiont of *Eucomonympha*, these proteins were not found. Whether the homologous sequences are present or absent in the genomes was further examined by BLAST searches (tBLASTn) using protein sequences of *Treponema denticola* as queries. There was no homologous sequence for flagellum, its biosynthesis, and chemotaxis in the genomes of the endosymbiont of *Eucomonympha* (no hit or insignificant hits only with e-values >-3 in the BLAST searches), while the orthologous genes were present in the genomes of *T. brennaborense* and *T. primitia* ZAS-2 (hits with e-values <-42). The genes absent in the genomes are listed as follows. Genes for flagellar and motor proteins: the filament FliC; the filament cap FliD; the hook FlgE; the hook capping protein FlgD; hook-filament junction FlgL and FlgK; the distal rod FlgG; the proximal rod FlgB and FlgC; the MS ring FliF; the motor switch/ C ring FliG, FliM, and FliN; and the motor MotA and MotB. Genes for flagellar biosynthesis proteins: FlhA, FlhB, FliH, FliP, and FliR. Genes for chemotaxis proteins: the sensor kinase CheA; the purine-binding chemotaxis protein CheW, methyltransferase CheR, and methyl-accepting chemotaxis sensory transducers (MCPs). Genes for flagellar proximal rod protein FliE and flagellar biosynthesis proteins FliO and FliQ were also absent in the genomes of the endosymbiont of *Eucomonympha*, although these proteins were not so homologous to the query sequences in *T. brennaborense* and *T. primitia* ZAS-2 (hits with e-values -23 to -16). These flagellar components and biosynthesis proteins are considered essential for the flagellar function and conserved among most bacteria as a core set (12).

In contrast, the COG profiles indicated the presence of genes involved in the "cell

wall/membrane/envelop biogenesis" function in the genomes of the endosymbiont of *Eucomonympha* (Fig. S7). Indeed, a set of genes for peptidoglycan biosynthesis is found in the genomes, including genes for the penicillin-binding proteins that are absent in the *Sphaerochaeta* genomes (11). In addition, some genes necessary for outer membrane proteins such as *lol* genes for lipoprotein transport, *yaeT* for outer membrane protein assembly complex, and *ompH* for outer membrane chaperone were present, although genes for lipopolysaccharide biosynthesis were incomplete as in most *Treponema* species. The cell having two membranes with electron sparse inter-membrane space is a common morphological character with the endosymbionts of other termite-gut cellulolytic protists in the genus *Pseudotrichonympha* and *Trichonympha*. In the genomes of these endosymbionts, the *lol* genes are absent or become pseudogenes, suggesting defects in their outer membrane function, although the genomes have peptidoglycan biosynthesis genes (13, 14).

Identification of the treponeme endosymbiont of *Teranympha*. *Teranympha mirabilis* in the gut of the termite *Reticulitermes speratus*, is the closest known relative of *Eucomonympha* spp. in the termite *H. sjoestedti* (3). Although these two termite species belong to distantly related families, the compositions of their gut protists are exceptionally quite similar, and a horizontal transfer of gut protists between these termites is suggested (15). The protist species in the genus *Eucomonympha* are also reported in wood-feeding cockroaches in the genus *Cryptocercus*, but are far distantly related to *Eucomonympha* spp. in *H. sjoestedti* than *T. mirabilis* and probably than termite-gut protists in the genus *Pseudotrichonympha* (16, 17). To know whether the related treponeme species is associated with *T. mirabilis*, we examined bacterial 16S rRNA gene sequences from manually isolated cells of *T. mirabilis*. In two independent experiments, we identified the sequences RsTeSp1 and RsTe409 that showed 3.05–3.65% nucleotide differences with the sequence identified from the endosymbionts of *Eucomonympha*. The sequences RsTeSp1 and RsTe409 showed less than 0.40% nucleotide difference to each other and to the sequence Rs-A48, the closest relative of the endosymbionts of *Eucomonympha* (see Fig. 1 in the main text), identified from the gut community of *R. speratus*. The target sequences of the probes for the endosymbionts of *Eucomonympha* (IIC-404 and IIC-637) were conserved in the sequences RsTeSp1 and RsTe409, so these

two probes were used for the FISH identification of associated bacteria with *T. mirabilis*. The cells of *T. mirabilis* harbour a dense population of endosymbiotic bacteria specifically stained with these FISH probes (Fig. S7). The cells of the endosymbiont were rod-shaped, 1.44±0.39 μm × 0.32±0.02 μm in size, not flagellated, and separated from the host protist cytoplasm by electron sparse inter-membrane space as in the case of the endosymbionts of *Eucomonympha*. The results indicate that *T. mirabilis* harbours *Treponema* species as endosymbionts that are very similar to that of *Eucomonympha* with respect to their phylogeny and morphology.

To investigate whether the endosymbiotic *Treponema* species of *T. mirabilis* is involved in reductive acetogenesis and nitrogen fixation, we identified the gene fragments of *acsA*, *acsB*, and FTHFS gene by the PCRs from manually isolated *T. mirabilis* cells using the primers described above. The DNA region between *acsA* and *acsB* was also amplified to find the similar gene structure to that of the endosymbiont of *Eucomonympha*. The coding protein sequences were closely related to those of the endosymbiont of *Eucomonympha*, showing 98.5%, 96.8%, 92.2%, and 98.1% identities for AcsA, AcsB, AcsF, and FTHFS, respectively. In addition, the *nifH* sequence identified from *T. mirabilis* corresponded to sequence RSN-TKY9 (D83124), an abundantly obtained sequence from the gut community of *R. speratus* (18). The RSN-TKY9 sequence was closely related to the *nifH* sequence identified from the endosymbiont of *Eucomonympha* (HSN20), showing 99.3% identity for the coding protein. Although mere the presence of these genes is not the direct evidence and further studies are necessary, it may be possible that the endosymbiotic treponeme of *T. mirabilis* can perform the same functions of reductive acetogenesis and nitrogen fixation as the endosymbiont of *Eucomonympha*.

We tentatively assigned the endosymbiont of *T. mirabilis* to '*Ca.* Treponema intracellularis', although a more comparative study such as the genome analysis is needed to justify this assignment. The sequence data from the endosymbionts of *T. mirabilis* have been deposited to DDBJ under the accession nos. LC012874– LC012879.

SI References

1. Iida T, Ohkuma M, Ohtoko K, Kudo T (2000) Symbiotic spirochetes in the termite

hindgut: phylogenetic identification of ectosymbiotic spirochetes of oxymonad protists. *FEMS Microbiol Ecol* 34(1):17–26.

2. Ohkuma M, Iida T, Kudo T (1999) Phylogenetic relationships of symbiotic spirochetes in the gut of diverse termites. *FEMS Microbiol Lett* 181(1):123–129.
3. Ohkuma M, et al. (2005) Molecular phylogeny of parabasalids inferred from small subunit rRNA sequences, with emphasis on the Hypermastigea. *Mol Phylogenet Evol* 35(3):646–655.
4. Noda S, Ohkuma M, Yamada A, Hongoh Y, Kudo T (2003) Phylogenetic position and in situ identification of ectosymbiotic spirochetes on protists in the termite gut. *Appl Environ Microbiol* 69(1):625–633.
5. 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(3):467–476.
6. Rosenthal AZ, Matson EG, Eldar A, Leadbetter JR (2011) RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. *ISME J* 5(7):1133–1142.
7. Matson EG, Gora KG, Leadbetter JR (2011) Anaerobic carbon monoxide dehydrogenase diversity in the homoacetogenic hindgut microbial communities of lower termites and the wood roach. *PLoS One* 6(4):e19316.
8. Matson EG, Zhang X, Leadbetter JR (2010) Selenium controls transcription of paralogous formate dehydrogenase genes in the termite gut acetogen, *Treponema primitia*. *Environ Microbiol* 12(8):2245–2258.
9. Inoue T, Moriya S, Ohkuma M, Kudo T (2005) Molecular cloning and characterization of a cellulase gene from a symbiotic protist of the lower termite, *Coptotermes formosanus*. *Gene* 349:67–75.
10. Inoue J, Saita K, Kudo T, Ui S, Ohkuma M (2007) Hydrogen production by termite gut protists: characterization of iron hydrogenases of Parabasalian symbionts of the termite *Coptotermes formosanus*. *Eukaryot Cell* 6(10):1925–1932.
11. Caro-Quintero A, Ritalahti KM, Cusick KD, Löffler FE, Konstantinidis KT (2012) The chimeric genome of *Sphaerochaeta*: nonspiral spirochetes that break with the prevalent dogma in spirochete biology. *MBio* 3(3): e00025-12.
12. Liu R, Ochman H (2007) Stepwise formation of the bacterial flagellar system. *Proc*

Natl Acad Sci USA 104(17):7116–7121.

13. Hongoh Y, et al. (2008) Genome of an endosymbiont coupling N_2 fixation to cellulolysis within protist cells in termite gut. *Science* 322(5904):1108–1109.
14. Hongoh Y, et al. (2008) Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Proc Natl Acad Sci USA* 105(14):5555–5560.
15. Kitade O (2004) Comparison of symbiotic flagellate faunae between termites and a wood-feeding cockroach of the genus *Cryptocercus*. *Microbes Environ* 19(3):215–220.
16. Carpenter KJ, Keeling PJ (2007) Morphology and phylogenetic position of *Eucomonympha imla* (Parabasalia: Hypermastigida). *J Eukaryot Microbiol* 54(4):325–332.
17. Ohkuma M, Noda S, Hongoh Y, Nalepa CA, Inoue T (2009) Inheritance and diversification of symbiotic trichonymphid flagellates from a common ancestor of termites and the cockroach *Cryptocercus*. *Proc Biol Sci* 276(1655):239–245.
18. Ohkuma M, Noda S, Usami R, Horikoshi K, Kudo T (1996) Diversity of nitrogen fixation genes in the symbiotic intestinal microflora of the termite *Reticulitermes speratus*. *Appl Environ Microbiol* 62(8):2747–2752.
19. Meßmer M, Wohlfarth G, Diekert G (1993) Methyl chloride metabolism of the strictly anaerobic, methyl chloride-utilizing homoacetogen strain MC. *Arch Microbiol* 160(5):383–387.
20. Hattori S, Galushko AS, Kamagata Y, Schink B (2005) Operation of the CO dehydrogenase/acetyl coenzyme A pathway in both acetate oxidation and acetate formation by the syntrophically acetate-oxidizing bacterium *Thermacetogenium phaeum*. J Bacteriol 187(10):3471–3476.
21. Seemann T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068–2069.
22. Rinke C, et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499(7459):431–437.
23. Axley MJ, Böck A, Stadtman TC (1991) Catalytic properties of an *Escherichia coli* formate dehydrogenase mutant in which sulfur replaces selenium. *Proc Natl Acad Sci USA* 88(19):8450–8454.
24. Ballor NR, Paulsen I, Leadbetter JR (2012) Genomic analysis reveals multiple [FeFe]

hydrogenases and hydrogen sensors encoded by treponemes from the H_2-rich termite gut. *Microb Ecol* 63(2):282–294.

25. 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(11):6590–6599.
26. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 9(4): 286–298.
27. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 56(4):564–77.
28. Stamatakis A, Hoover P, Rougemont J. (2008) A rapid bootstrap algorithm for the RAxML Web servers. 57(5):758–771.
29. Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21(9):2104–2105.
30. Rosenthal AZ, et al. (2013) Localizing transcripts to single cells suggests an important role of uncultured deltaproteobacteria in the termite gut hydrogen economy. *Proc Natl Acad Sci USA* 110(40):16163–16168.
31. Zhang X, Matson EG, Leadbetter JR (2011) Genes for selenium dependent and independent formate dehydrogenase in the gut microbial communities of three lower, wood-feeding termites and a wood-feeding roach. *Environ Microbiol* 13(2):307–323.
32. Salmassi TM, Leadbetter JR (2003) Analysis of genes of tetrahydrofolate-dependent metabolism from cultivated spirochaetes and the gut community of the termite *Zootermopsis angusticollis*. *Microbiology* 149(9):2529–2537.
33. Pester M, Brune A (2006) Expression profiles of fhs (FTHFS) genes support the hypothesis that spirochaetes dominate reductive acetogenesis in the hindgut of lower termites. *Environ Microbiol* 8(7):1261–1270.
34. Ottesen EA, Leadbetter JR (2010) Diversity of formyltetrahydrofolate synthetases in the guts of the wood-feeding cockroach *Cryptocercus punctulatus* and the omnivorous cockroach *Periplaneta americana*. *Appl Environ Microbiol* 76(14):4909–4913.
35. Ottesen EA, Leadbetter JR (2011) Formyltetrahydrofolate synthetase gene diversity in the guts of higher termites with different diets and lifestyles. *Appl Environ*

Microbiol 77(10):3461–3467.

36. Zheng H et al. Comprehensive phylogenetic diversity of [FeFe]-hydrogenase genes in termite gut microbiota. *Microbes Environ* 28(4):491–494.
37. Noda S, Ohkuma M, Usami R, Horikoshi K, Kudo T (1999) Culture-independent characterization of a gene responsible for nitrogen fixation in the symbiotic microbial community in the gut of the termite *Neotermes koshunensis*. *Appl Environ Microbiol* 65(11):4935–4942.
38. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
39. Inoue J, et al. (2014) Distribution and evolution of nitrogen fixation genes in the phylum *Bacteroidetes*. *Microbes Environ* 30(1):44–50.
40. Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R (2012) Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics* 13:162.
41. Powell S, et al. (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* 42(Database issue):D231–239.

Table S1. Fractionation of enzymatic activities for acetyl-CoA pathway

Enzyme	Whole gut content	Large protist	Small protist/bacteria
Formate dehydrogenase	2.36×10^3	0.54×10^3 (23%)	0.13×10^3 (5.5%)
Formyl tetrahydrofolate synthetase	69.2	53.7 (78%)	4.05 (5.9%)
Carbon monoxide dehydrogenase	805	192 (24%)	97.9 (12%)
Hydrogenase	2.60×10^4	1.44×10^4 (55%)	0.41×10^4 (16%)

Activity is an average of duplicate or triplicate measurements, expressed as nmol/min/g fresh-weight termite in each case. Whole gut content and the two fractions were prepared as described in the main text using the buffer containing dithiothreitol, but for the small protist/bacteria fraction, the cells were harvested by centrifugation. Crude extracts were obtained by sonication on ice in the anaerobic chamber and were used for the enzyme assays under anoxic conditions as described previously (19). Detailed assay conditions were followed by the methods described previously (20). Percentages in parentheses indicate the ratios of the activities after the fractionation calculated with the activities of whole gut as 100%. Note that the large proportion of the enzymatic activities was associated with the large protist fraction. One may consider that the endosymbionts could be protected by the host protist cell against exposures of O_2 or other inhibitory chemicals much more than free-swimming bacteria. However, the cell-free enzymes are equally susceptible to the exposure and the results of the enzyme assays supported that the majority of the potentials for reductive acetogenesis are indeed associated with the large protist fraction.

Table S2. Assembly and feature details of single-cell genomes of the endosymbiotic treponeme of *Eucomonympha* and comparison with related treponemes and other endosymbiont species of termite-gut protists.

	Single cell genomes of *Eucomonympha* endosymbiont					*T. brennabo -rense DSM 12668*	*T. primitia* ZAS-2	*T. azoto -nutricium* ZAS-9	*Azobacteroides pseudotricho -nymphae* CfPt1-2	*Endomicrobium trichonymphae* Rs-D17
	C1	D2	D11	E8	E12					
Assembled total bases	1,413,236	1,939,663	1,614,612	1,133,094	2,422,424	3,055,580	4,059,867	3,855,671	1,114,206	1,125,857
Number of contigs	802	891	758	719	948	1	1	1	1	1
Contig N50 bases	6,737	8,274	10,228	6,878	14,468	–	–	–	–	–
Longest contig bases	47,471	41,615	39,077	53,423	118,189	–	–	–	–	–
GC content, %	55.6	55.1	55.2	55.4	55.1	51.5	50.8	49.8	32.7	35.2
[a]CDS	1,448	1,989	1,660	1,156	2,397	2,585	3,499	3,403	932	1,119
Hypothetical CDS	774	999	842	592	1,122	720	1,077	1,079	273	372
Transposases	83	109	104	108	183	12	21	33	0	0
[b]Genome completeness, %	52.52	60.43	60.43	40.29	84.89	94.96	96.40	97.12	97.84	97.12
[c]Estimated genome size, Mb	2.69	3.21	2.67	2.81	2.85	–	–	–	–	–

The single-cell genomes of the endosymbiont of *Eucomonmpha* are compared with those of *T. brennaborense* DSM 12168 (CP002696), treponeme isolates from the termite gut, ZAS-2 (CP001843) and ZAS-9 (CP001841) (6), and endosymbionts of termite-gut protists CfPt1-2 (AP010656) (13) and Rs-D17 (AP009511) (14). [a]Protein coding sequences estimated here using the PROKKA program (21). [b]Genome completeness was estimated based on the presence of 139 single-copy genes common in bacteria (22). These 139 genes were searched for in the CDS of the single-cell genomes using the HMMER3 software (http://hmmer.org/). [c]The genome size was estimated based on the genome completeness. Despite complete genome sequences of the compared bacteria, their genome completeness is not 100% owing to the absence of some searched single-copy genes, indicating that the genome size of the endosymbiont of *Eucomonympha* was slightly overestimated.

Table S3. Acetyl-CoA pathway gene found in the single cell genomes and protein sequence identity to those of *Treponema primitia*.

Enzyme	gene	C1	D2	D11	E8	E12	Identity[a]
[b]Formate dehydrogenase I	*fdh1*	124	97	58	-	13	87%
[b]Formate dehydrogenase II	*fdh2*	282	59	753	20	62	63%
Formyl-[c]THF synthetase	*fts*	223	-	-	4	54	91%
Methenyl-THF cyclohydrolase/ methylene-THF dehydrogenase	*folD*	798	40	17	36	77	81%
Methylene-THF reductase	*metF*	-	-	-	114	81	59%
THF: Fe-S-Co Methyl transferase	*acsE*	-	402	2	-	215	73%
Carbon monoxide dehydrogenase subunit	*acsA* (*cooS*)	-	802	97	-	-	85%
Acetyl-CoA synthase subunit	*acsB*	171	278	-	-	-	83%
Ni-insertion protein	*acsF*	-	-	483	-	-	82%
Large Subunit Fe-S-Co protein	*acsC*	-	-	1	-	144	72%
Small Subunit Fe-S-Co protein	*acsD*	-	4	1	-	144	82%
[d]FeFe-hydrogenase	*hyd*	83	12	-	-	13	66%

C1 to E12 represent the five single-cell genomes of the endosymbiont of *Eucomonympha*. Contig numbers localising these genes are indicated. Hyphens indicate that the genes were not found. [a]Percent identity of protein sequence with *Treponema primitia* strain ZAS-2. We found two distinct genes for formate dehydrogenase (FDH), as designated I and II here, each of which has an inframe non-canonical TGA codon putatively encoded selenocysteine (23). The FDH I is more closely related to the selenocysteine-containing FDH, and the FDH II to cysteine-only FDH, of *T. primitia* ZAS-2 (8). [c]Tetrahydrofolate. [d]FeFe-hydrogenase homologous to the protein encoded by *hndA3* gene of *T. primitia* ZAS-2. In the single-cell genome E12, this hydrogenase gene is clustered together with not only *hndB*- and *hndC*-like genes as in *T. primitia* ZAS-2 (24) but also with *fdh1* encoding FDH I. See also figures S3 to S6 for the phylogenetic analyses of FDH, formyl-THF synthetase, carbon monoxide dehydrogenase subunit, and FeFe-hydrogenase, respectively.

Table S4. Presence of genes for key metabolic functions in single-cell genomes of endosymbiont of *Eucomonympha*.

Enzyme	C1	D2	D11	E8	E12
Glycolytic pathway					
Hexokinase (EC2.7.1.1)	-	+	+	+	+
Glucose-6-phosphate isomerase (EC5.3.1.9)	-	-	-	-	+
Pyrophosphate-dependent fructose 6-phosphate-1-kinase (EC2.7.1.90)	-	+	+	-	+
Fructose-bisphosphate aldolase class II (EC4.1.2.13)	-	+	+	-	+
Triosephosphate isomerase (EC5.3.1.1)	+	-	-	-	+
NAD-dependent glyceraldehyde-3-phosphate dehydrogenase (EC1.2.1.12)	+	+	-	+	+
Phosphoglycerate kinase (EC2.7.2.3)	-	-	+	-	+
2,3-bisphosphoglycerate-independent phosphoglycerate mutase (EC5.4.2.1)	+	+	+	+	+
Enolase (EC4.2.1.11)	-	-	-	-	-
Pyruvate, phosphate dikinase (EC2.7.9.1)	+	-	+	+	+
Pyruvate metabolism					
Pyruvate-flavodoxin oxidoreductase (EC1.2.7.-)	-	+	+	+	+
Acetyl-coenzyme A synthetase (EC6.2.1.1)	-	+	-	-	-
Phosphate acetyltransferase (EC2.3.1.8)	-	+	-	-	+
Acetate kinase (EC2.7.2.1)	-	+	+	-	-
Phosphoenolpyruvate carboxykinase [GTP] (EC4.1.1.32)	+	+	-	-	+
Malate dehydrogenase (EC 1.1.1.37), similar to archaeal MJ1425	-	+	+	-	+
NADP-dependent malic enzyme (EC 1.1.1.40)	+	+	-	+	+
Oxaloacetate decarboxylase (EC 4.1.1.3)	+	-	+	-	-
Non-oxidative pentose phosphate pathway					
Ribose 5-phosphate isomerase A (EC5.3.1.6)	+	+	-	+	+
Ribulose-phosphate 3-epimerase (EC5.1.3.1)	-	-	+	+	+
Transketolase (EC2.2.1.1)	+	+	+	+	+
Sugar and uronate utilisation					
Mannose-6-phosphate isomerase (EC5.3.1.8)	-	+	-	+	+
4-alpha-glucanotransferase (amylomaltase) (EC2.4.1.25)	-	+	+	+	+
Trehalase (EC3.2.1.28)	+	+	+	-	+
Uronate isomerase (EC5.3.1.12)	-	-	-	+	+
D-mannonate oxidoreductase (EC1.1.1.57)	+	+	-	+	+
Mannonate dehydratase (EC4.2.1.8)	-	+	+	+	+
2-dehydro-3-deoxygluconate kinase (EC2.7.1.45)	-	-	-	-	+
2-dehydro-3-deoxyphosphogluconate aldolase (EC4.1.2.14)	-	+	-	+	+
Glycogen metabolism					
Glucose-1-phosphate adenylyltransferase (EC2.7.7.27)	+	+	+	+	+
Glycogen synthase, ADP-glucose transglucosylase (EC2.4.1.21)	+	+	+	-	+
Glycogen phosphorylase (EC2.4.1.1)	-	+	-	-	+
Glycogen branching enzyme, GH-57-type, archaeal (EC2.4.1.18)	+	+	+	+	-
Phosphoglucomutase (EC5.4.2.2)	-	-	-	-	-
Urea and ammonia utilisation					
UreA Urease gamma subunit	-	-	+	-	+
UreB Urease beta subunit	+	-	+	-	+
UreC Urease alpha subunit	-	-	+	-	+
UreD Urease accessory protein	-	-	-	-	+
UreF Urease accessory protein	-	-	+	-	+
UreG Urease accessory protein	-	-	-	-	+
UreI Urea channel (AmiS/UreI transporter)	-	+	+	-	+
Amt ammonium transporter	-		+	-	+
Glutamine synthetase type I (EC6.3.1.2)	-	+	+	+	+
Glutamine synthetase type III, GlnN (EC6.3.1.2)	-	-	+	+	+

The gene for the glycolytic pathway enzyme enolase (EC 4.2.1.1) was absent in all the single-cell genomes and the gene for phosphoglucomutase (EC 5.4.2.2) involving glycogen metabolism was not found in the genomes, but the latter gene is also absent in *T. brennaborense* and *T. primitia* ZAS-2 though the other genes for glycogen metabolism are conserved. Interestingly in genomes of treponemes and spirochetes, we failed to detect urease proteins highly homologous to those of the endosymbiont of *Eucomonympha*, which are homologous to those of firmicutes (e.g. *Clostridium* sp. (WP_008423975) showed 83% identity with the gamma subunit protein).

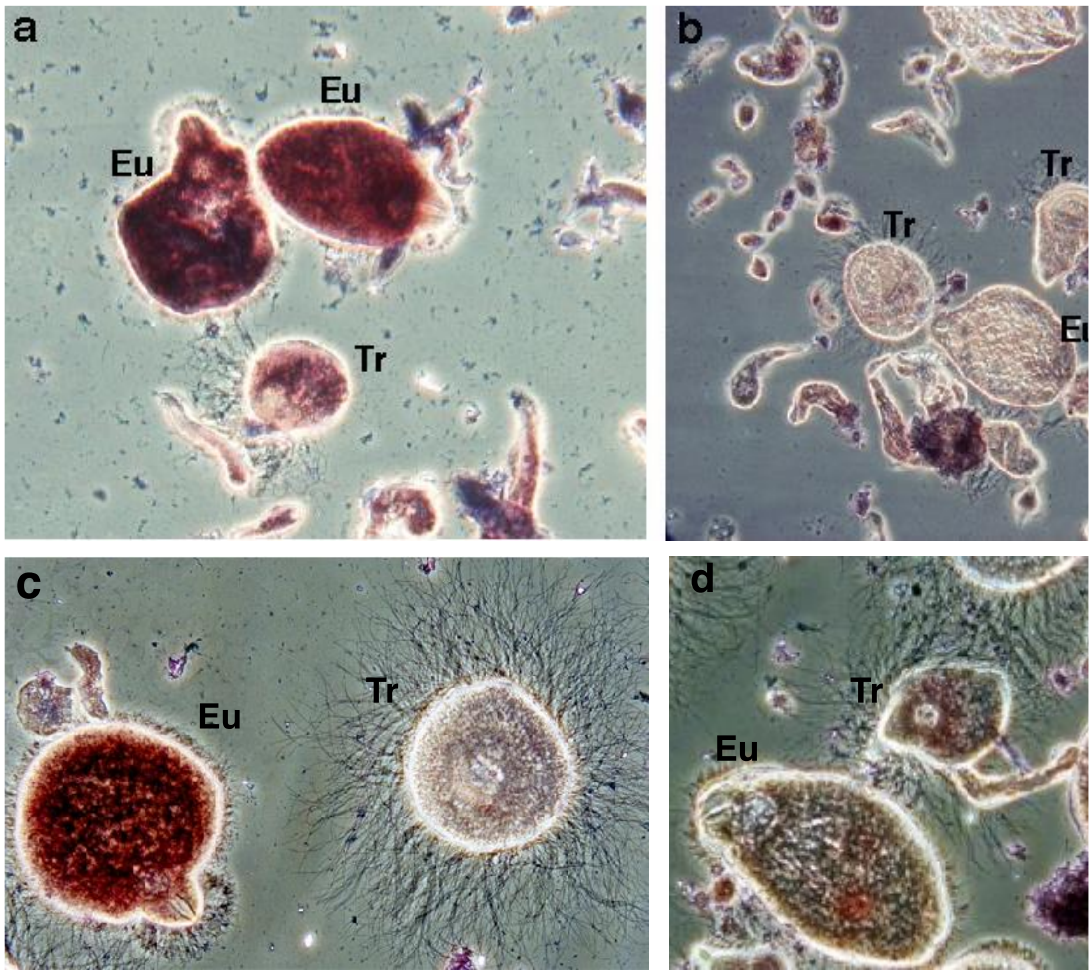

Fig. S1. *In situ* detection of the expression of *acsB* and *anfH* in the endosymbiont of *Eucomonympha*. (*a*) *In situ* hybridisation against *acsB* mRNA with an anti-sense probe highlighted strong signals in *Eucomonympha* cells (labelled with Eu) but only weakly in *Trichonympha* (labelled with Tr) and other protist cells. (*b*) Control hybridisation with a sense probe of *acsB* gave no significant hybridisation signal in any protist cell. Hybridisation with an anti-sense probe for *anfH* gave specific strong signals in *Eucomonympha* cells (*c*), but no significant signal with a control sense probe in any protist cell (*d*). Signal-like accumulation of dyes was the background reaction with amorphous particles in the gut. Another control experiment without any probe gave only the similar background reactions.

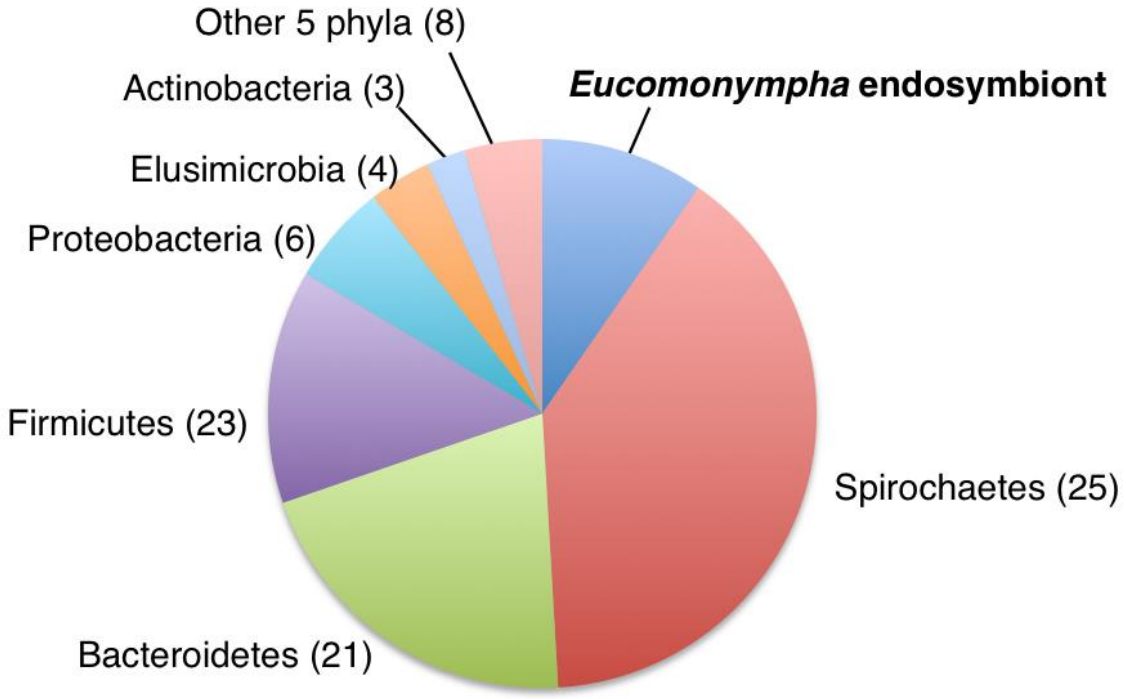

Fig. S2. Bacterial community structure in the gut of *H. sjoestedti* and predominance of the treponeme endosymbiont of *Eucomonympha*. A total of 218 clones of the bacterial 16S rRNA gene sequences amplified by PCR from the gut community were sorted into 91 phylotypes (defined with >97.0% sequence identity) in 11 bacterial phyla. The endosymbiont of *Eucomonympha* represented the most abundant phylotype and amounted to 21 clones (9.6% of total clones). The pie chart shows the relative abundance of the detected phyla and of the phylotype of the endosymbiont. Parentheses after the names of phyla indicate the number of phylotypes wherein. The clone analysis was conducted by the method described previously (25).

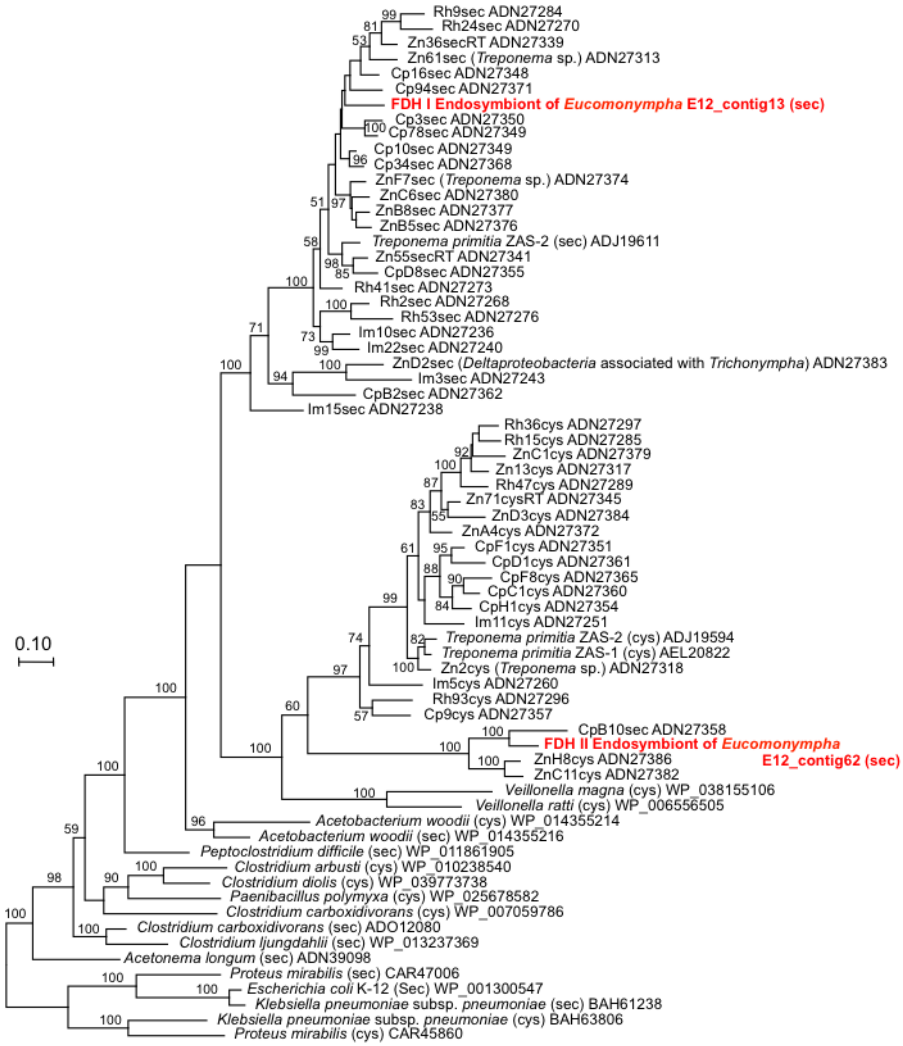

Fig. S3. Phylogenetic positions of two formate dehydrogenases (FDH I and FDH II) of the endosymbiont treponeme species of *Eucomonympha*. Protein sequences homologous to hydrogenase-linked FDHs and representative sequences from the gut of termites and the related *Cryptocercus* cockroach were aligned using the MAFFT program (26), and after removing ambiguously aligned sites with the Gblocks program (27), 586 sites were used for the analysis. The maximum likelihood tree was inferred using the RaxML program (28) with the LG+I+G model selected by the ProtTest (29). Numbers at nodes indicate percent bootstrap values when greater than 50%. The scale bar denotes 0.10 substitutions per site. FDH sequences containing putative selenocysteine in the active site are designated with "sec", while those having cysteine in the corresponding site are designated with "cys", in clone names or in parentheses, respectively. The organismal origins of clones are indicated in parentheses when experimentally identified (30). FDH I of the endosymbiont of *Eucomonympha* was grouped with selenocysteine-containing FDH sequences from treponemes and gut clones, and this group was a sister to the group containing deltaproteobacterium associated with the *Trichonympha* protist. FDH II of the endosymbiont of *Eucomonympha* and its closest relative CpB10sec, though they were selenocysteine-containing FDHs, were branched out within the cluster comprised of cysteine-only FDHs of *Treponema* and *Veillonella*. Clones tagged with Rh, Zn, Im, and Cp are recovered from *Reticulitermes hesperus*, *Zootermopsis nevadensis*, *Incisitermes minor*, and *Cryptocercus punctulatus*, respectively (31).

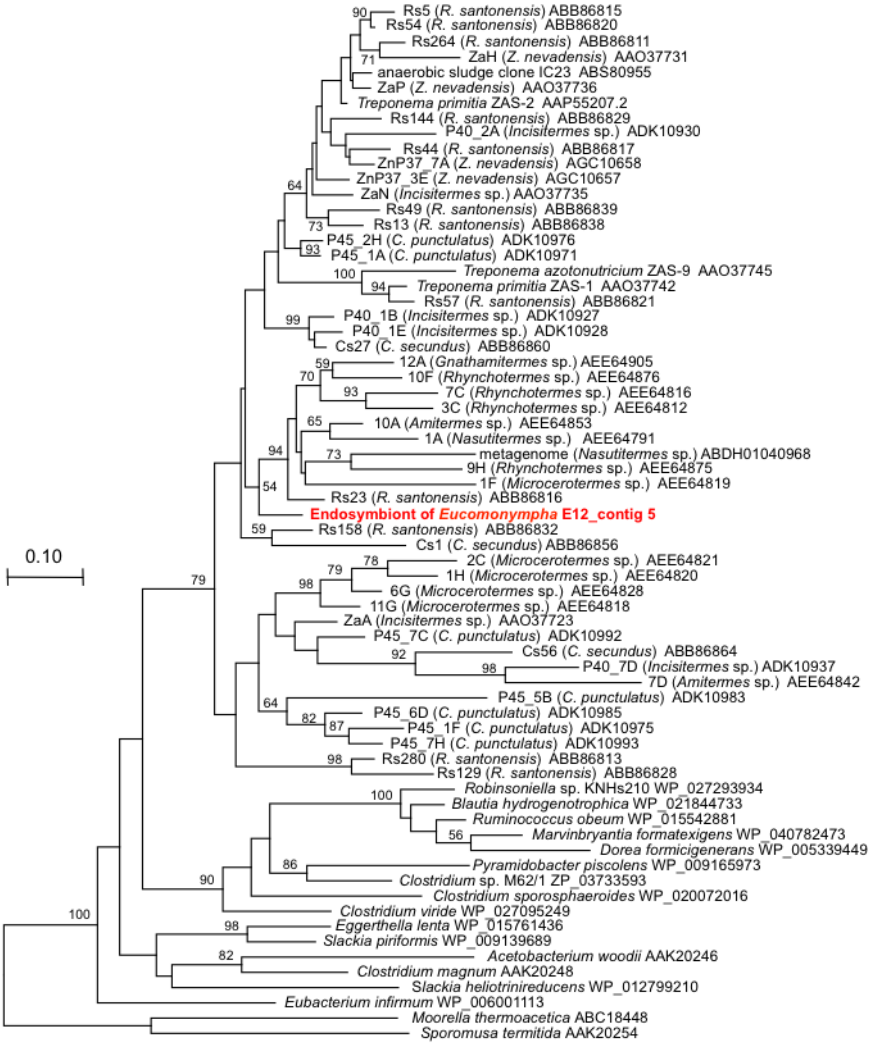

Fig. S4. Phylogenetic position of formyltetrahydrofolate synthetase of the endosymbiont of *Eucomonympha*. The maximum likelihood tree was inferred with the LG+I+G+F model based on unambiguously aligned 278 sites of protein sequences. Details are given as Fig. S3 in the text. Species of the termites and the cockroach from which the clones are recovered (32–35) are shown in parentheses after the clone names.

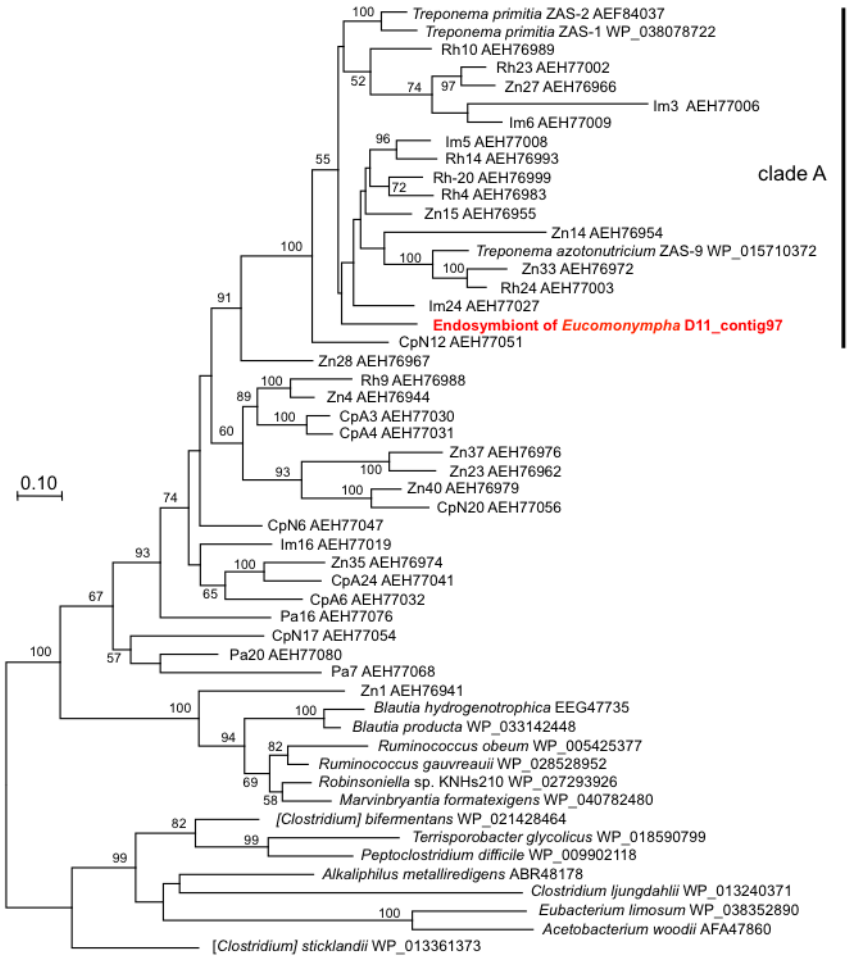

Fig. S5. Phylogenetic position of carbon monoxide dehydrogenase (AcsA or CooS) of the endosymbiont of *Eucomonympha*. The maximum likelihood tree was inferred with the LG+I+G model based on unambiguously aligned 446 sites of protein sequences. Details are given as Fig. S3 in the text. AcsA (CooS) sequence of the endosymbiont is included in the "clade A" (indicated by vertical bar) defined by Matson et al. (2011) (7). This clade also contains sequences from treponemes. Clones tagged with Rh, Zn, Im, Cp, and Pa are recovered from *Reticulitermes hesperus*, *Zootermopsis nevadensis*, *Incisitermes minor*, *Cryptocercus punctulatus*, and the cockroach *Periplaneta americana*, respectively (7).

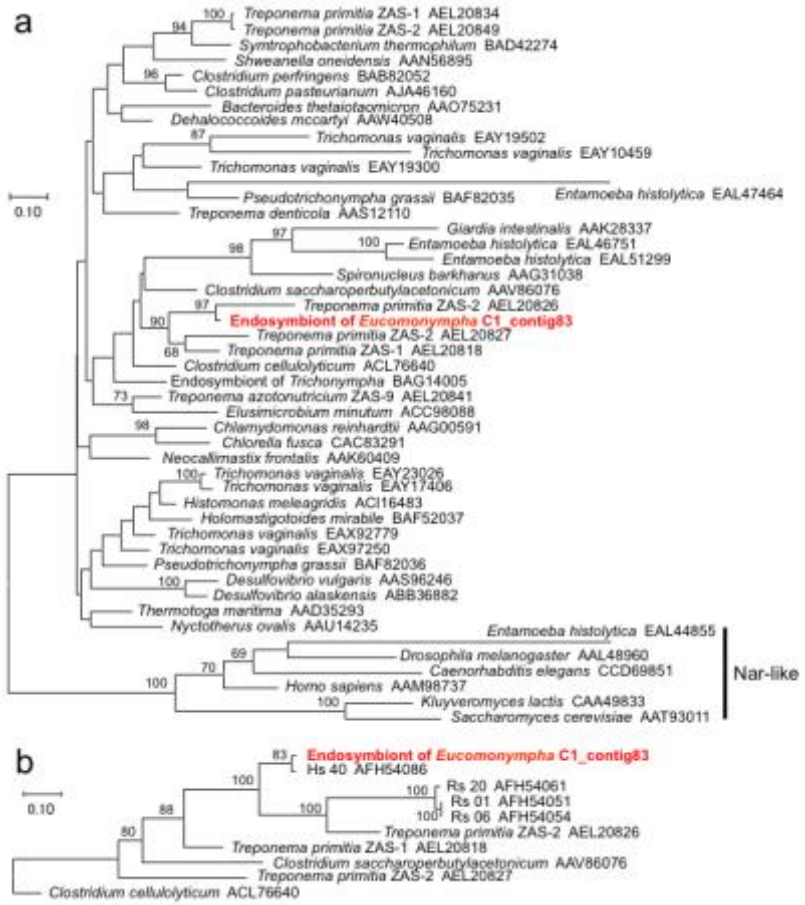

Fig. S6. Phylogenetic position of FeFe-hydrogenase of the endosymbiont of *Eucomonympha*. The phylogenetic relationship of diverse FeFe-hydrogenases is shown in (a) and that of close relatives to the endosymbiont of *Eucomonympha* is shown in (b). The maximum likelihood trees were inferred with the LG+I+G model based on unambiguously aligned 135 sites of protein sequences in (a) and with the LG+G model based on 131 sites of H-cluster domain of the proteins in (b), respectively. Nar-like proteins that are related to FeFe-hydrogenases but do not catalyse hydrogenase reactions were used as the outgroups in (a). Details are given as Fig. S3 in the text. Clones tagged with Hs and Rs in (b) are recovered from the termites *H. sjoestedti* and *R. speratus* (36).

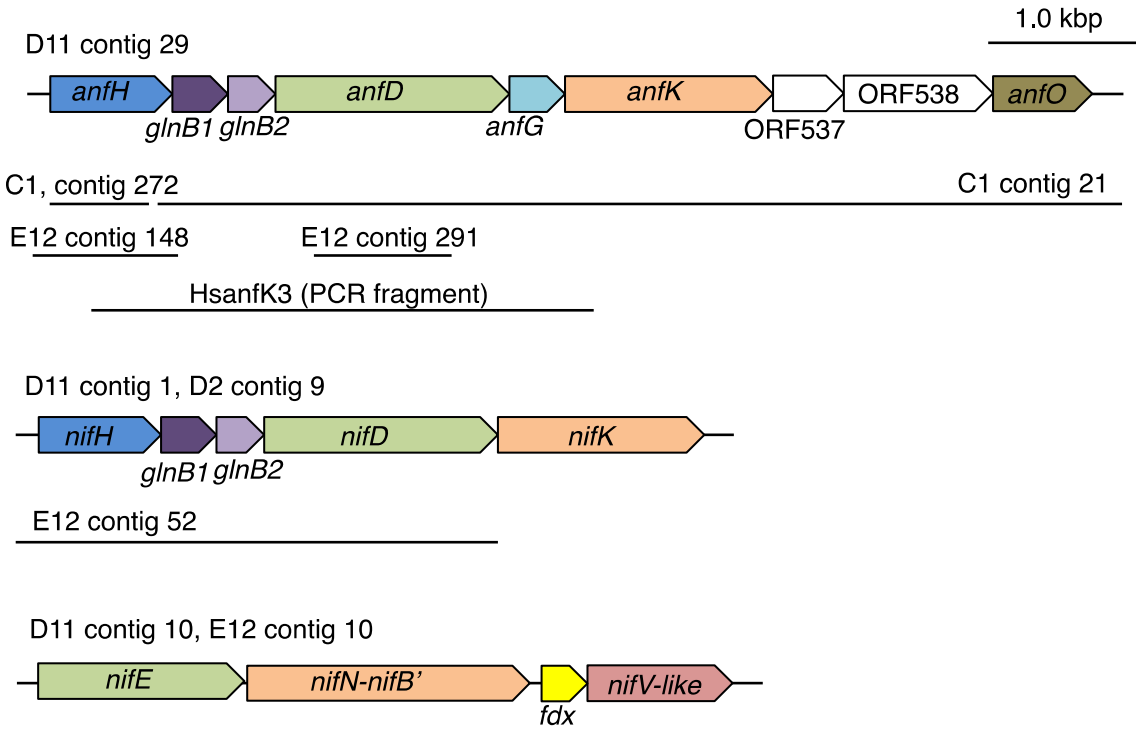

Fig. S7. Nitrogen fixation gene clusters found in the single-cell genomes of the endosymbiont of *Eucomonympha*. Bars below the gene structures shows the sequences found in other single cell genomes. The sequence HsanfK3 (LC012873) was obtained from the gut microbial community of *H. sjoestedti* by PCR using a primer specific for the Hs10 group of *anfH* (5'-CACTTCTACAACCAG) and the ANFK primer (37) for the conserved sequence in *anfK*. Genes *glnB1* and *glnB2* encode homologs of nitrogen regulatory protein PII. ORF537 is annotated as an FMN-binding protein related to pyridoxamine 5'-phosphate oxidase. ORF538 encodes a hypothetical protein. Gene *fdx* encodes ferredoxin. Gene *nifV*-like encodes homocitrate synthase. Gene designated as *nifN-nifB'* encodes a protein fused NifN and NifB but the NifB part lacks a radical S-adenosylmethionine domain.

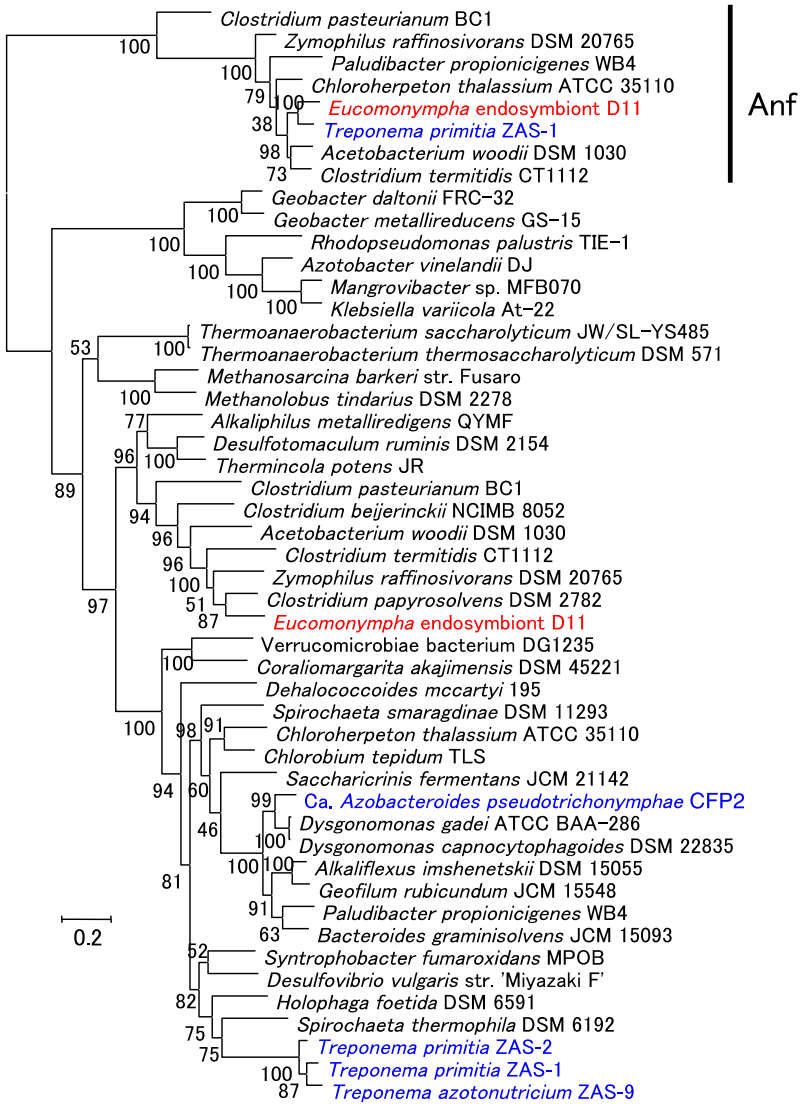

Fig. S8. Phylogenetic tree with concatenated protein sequences of NifH, NifD, and NifK including the corresponding AnfH, AnfD, and AnfK. The tree was inferred by maximum likelihood method using RaxML 8.1.2 MPI version (38) with the LG+F+G model for each protein as described previously (39). A vertical bar indicates the sequences of AnfH, AnfD, and AnfK. The sequences of the endosymbiont of *Eucononympha* found in the single-cell genome D11 were used for the inference and are shown in red. The sequences from termite-gut symbionts are shown in blue. Numbers at nodes indicate percentage bootstrap values. The scale bar denotes 0.2 substitutions per site. Note that the NifHDK of the endosymbiont of *Eucononympha* is distantly related to those of the termite-gut treponemes. The AnfHDK of the endosymbiont of *Eucononympha* is closely related to that of *T. primitia* stain ZAS-1, in which nitrogen fixation has not been reported.

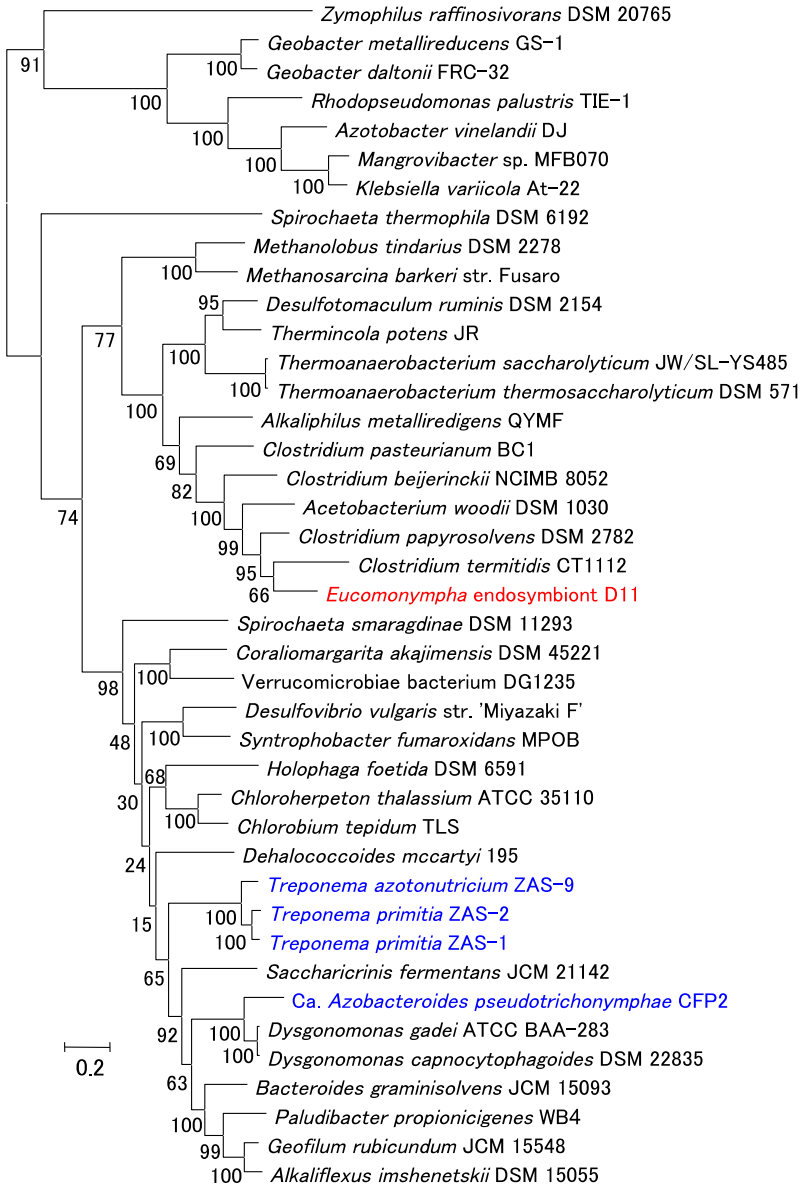

Fig. S9. Phylogenetic tree with concatenated sequences of NifE and NifN. The tree was inferred by maximum likelihood method with the LG+ G model for each protein as described previously (39). Details are given as Fig. S4 in the text. NifE and NifN are needed for metal cofactor assembly of nitrogenase. Again, the NifEN of the endosymbiont of *Eucomonympha* is distantly related to those of the termite-gut treponemes. In the single-cell genomes of the endosymbiont, there were gene clusters comprising eight pairs of *nifE* and *nifN* homologous genes. They are not included in this tree, because they were distantly related to *nifE* and *nifN* and phylogenetically classified to the *nifE*-like and *nifN*-like groups of sequences encoding unknown functions (40).

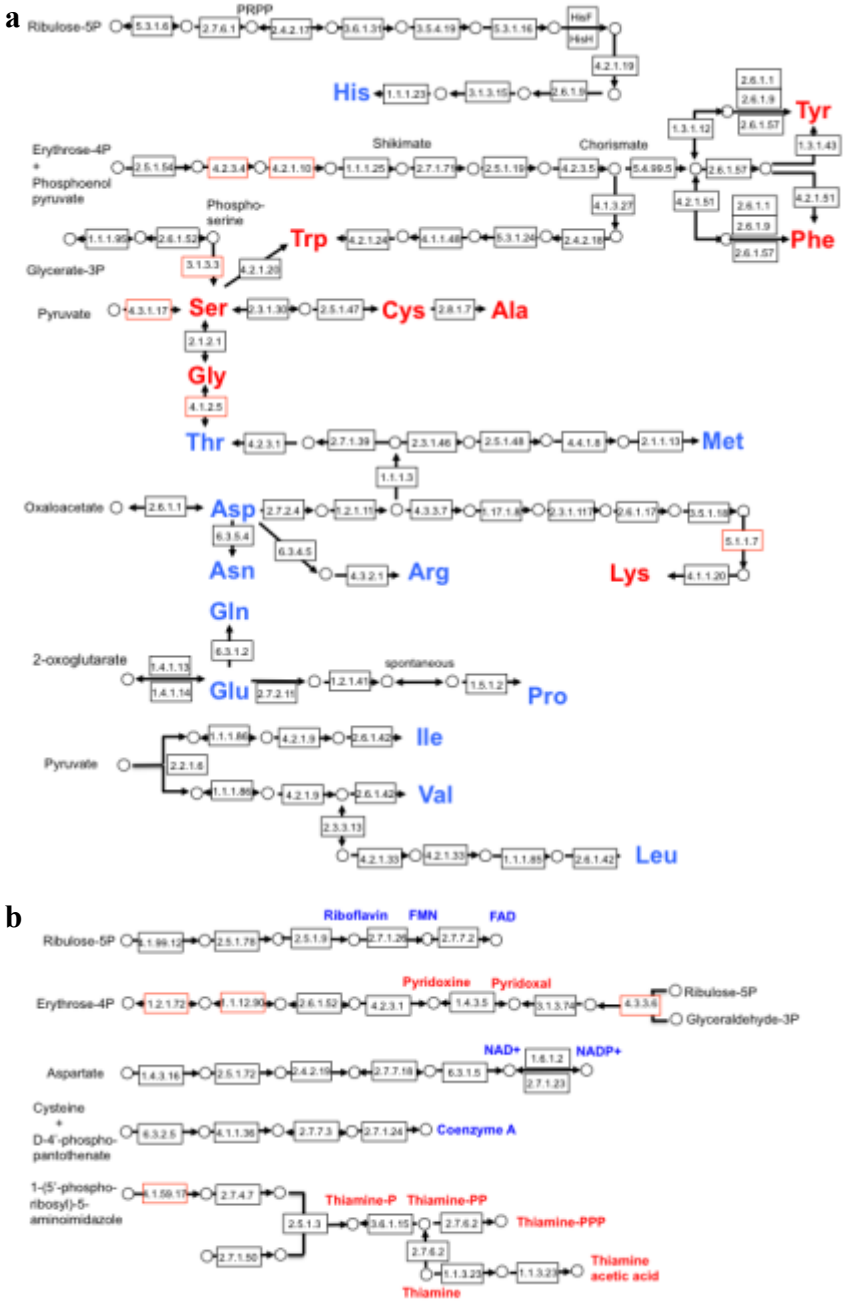

Fig. S10. Biosynthetic pathways of amino acids (*a*) and cofactors (*b*) reconstructed compositely from the single-cell genomes of the endosymbiont of *Eucomonympha*. The pathways were reconstructed according to the KEGG Pathway database (http://www.genome.jp/kegg/pathway.html). Amino acids and cofactors in blue indicate that their complete biosynthetic pathway was successfully reconstructed. Those in red indicate that the pathway was partial, but they can be synthesised if some other amino acids or intermediate compounds are available. Because genes for several reaction steps absent in the endosymbiont genomes have not been identified in the genomes of *T. primitia* and *T. brennaborense*, such as reactions of EC 4.2.1.10, EC 3.1.3.3, and EC 5.1.1.7, these pathways may be functional by yet-uncharacterised enzymes.

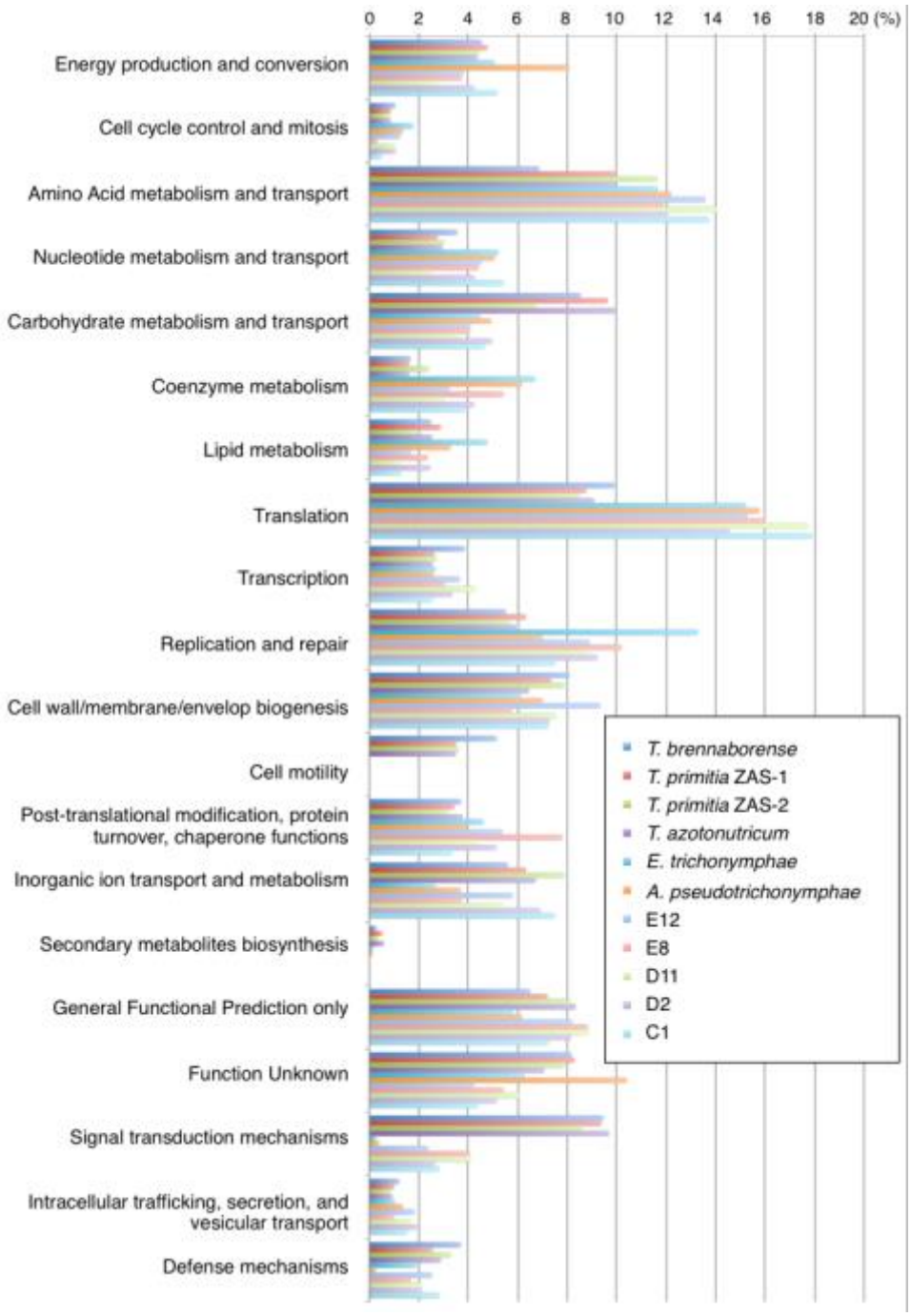

Fig. S11. Comparison of cluster of orthologous groups of proteins (COGs) profiles. The COGs profiles of the five single-cell genomes of the endosymbiont of *Eucomonympha* (C1, D2, D11, E8, and E12) were compared with four cultured treponemes and two endosymbiont species of termite-gut protists ('*Ca.* Endomicrobium trichonymphae' Rs-D17 and '*Ca.* Azobacteroides pseudotrichonymphae' CfPt1-2). The COGs of the treponemes were assigned using spiNOG, a non-supervised orthologous groups database for Spirochaetes in eggNOG (41), and those of Rs-D17 and CfPt1-2 were retrieved from the eggNOG database.

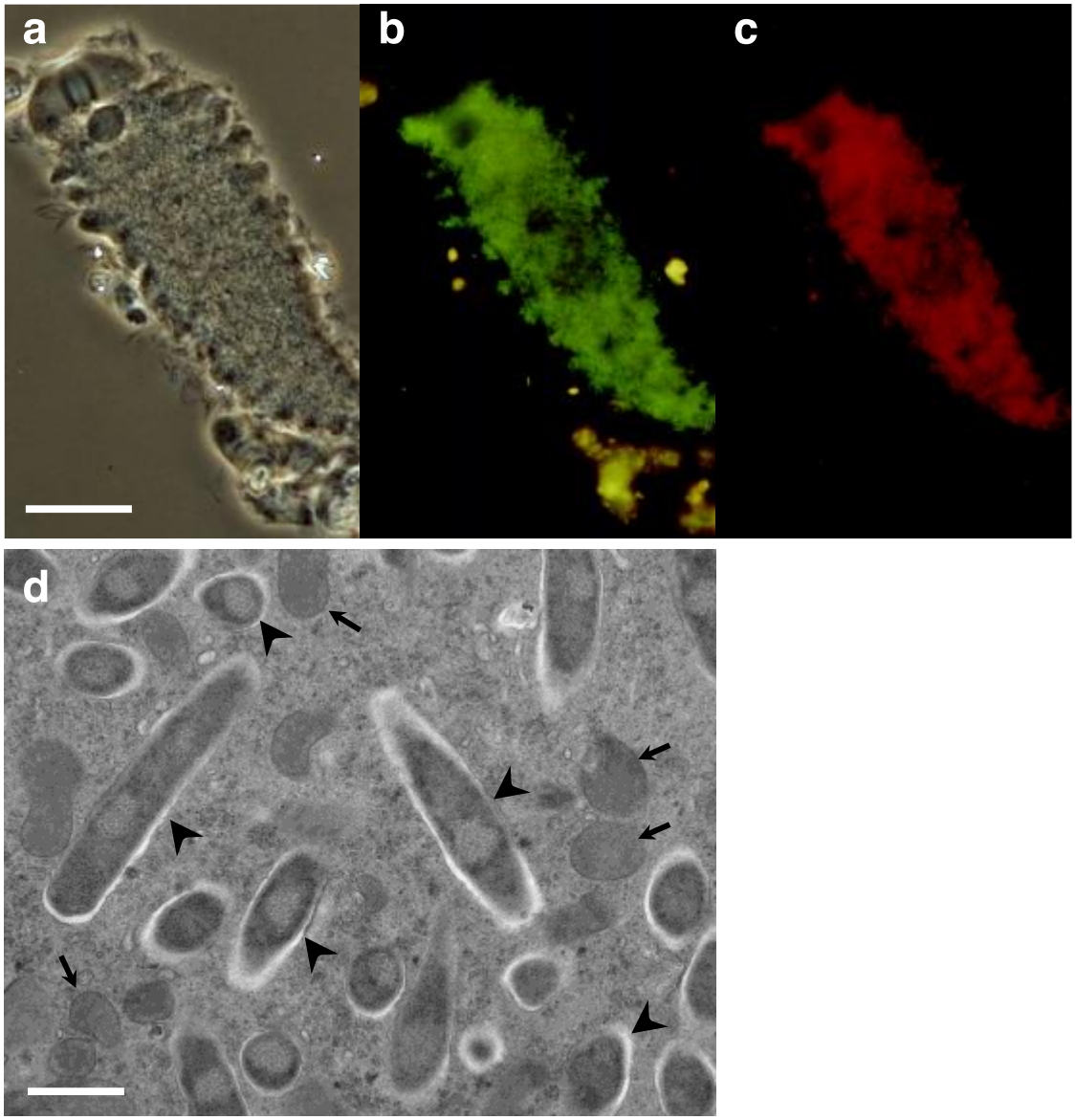

Fig. S12. FISH identification of endosymbiotic treponeme species of *Teranympha mirabilis* (*a*-*c*) and its morphology (*d*). (*a*) Phase contrast image of *T. mirabilis*. Scale bar, 50 µm. (*b*) Hybridisation with the specific probes (in green). (*c*). Hybridisation with the general bacterial probe (in red). Amorphous yellow signals in (*b*) and corresponding red signals in (*c*) were derived from autofluorescence of ingested wood particles. (*d*) TEM image of ultrathin section of a *T. mirabilis* cell, showing rod-shaped endosymbionts (arrowheads) and hydrogenosome-like organelles (asterisks). Scale bar, 0.5 µm.

Chapter V

Comparative genomic approaches to classify the taxonomy of the genus *Treponema*.

Introduction

Treponemes are highly motile, helical bacteria, from the phylum Spirochaetae. These bacteria inhabit various environments and have various lifestyles. Most culturable references have been host associated treponemes, and either pathogens or gut microbiota. Pathogenic treponemes affect multiple host species, in humans the sexually transmitted pathogen *Treponema pallidum*, causing syphilis and yaws and was the first treponeme genome sequenced (Fraser *et al*., 1998). There are treponeme species also responsible for dental pathogenesis in the development of periodontitis (*Treponema medium and Treponema phagedensis*). In bovines, treponeme species act as both pathogens and gut symbionts, species such as *Treponema brennaborense* cause lesion formation in digital dermatitis, whereas *Treponema bryantii* is a bovine ruminal treponeme that aids in biomass degradation in enhancing cellulose breakdown (Stanton and Canale-Parola, 1980). Treponemes have also been detected in the digestive tracts of indigenous human populations (Schnorr *et al*., 2014) and isolated from swine (*Treponema succinifaciens*), and ovine gut environments (*Treponema zioleckii*).

The largest reservoir of symbiotic treponeme species has emerged from wood eating termites, where acetogenic and diazotrophic species have been previously isolated (Graber *et al*., 2004). Metagenomic analysis described the number of phylotypes of treponemes within a single termite at over 100 (Warnecke *et al*., 2007). Termite associated *Treponema* have been classified into two phylogenetic clusters, one cluster associated with bovine rumen species (*Treponema* cluster II; Ohkuma *et al*., 1999) and the second cluster associated with the free living thermophilic mat species *Treponema caldarium* (*Treponema* cluster I). *Treponema* cluster I, contains both cultured strains *Treponema azotonutricium* (ZAS-9) and *Treponema primitia* (ZAS-1 and ZAS-2) isolated from the lower termite *Zootermopsis angusticolis* (Graber *et al*., 2004). Both of these clusters, certain members have been observed at forming both endosymbiotic and ectosymbiotic relationships (Iida et al., 2000; Ohkuma *et al*., 2015). The differences between the lifestyles of free living *Treponema* in these environments with and without protists have not been categorised. Here using sequenced higher termite phylotypes from sub cluster Ic (hybrid assembled samples R80H11, R81E12 and R83G3 from chapter II), the sequenced endosymbiont (‘Candidatus *Treponema intracellularis’* from chapter IV) and new treponeme sequences isolated from *Reticulitermes spearatus,* we aim to establish categorical differences between these species.

Treponema and Spirochaetes have an established spiral morphology whereby the flagella are located concurrently within the periplasm. The advantage of this is chemotactically, the

bacterium can manoeuver efficiently and within very viscous environments. Exceptions to this dogma have arisen in species from the genus *Sphaerochaeta* (Ritalahti *et al*., 2012) and in the endosymbiont 'Candidatus *Treponema intracellularis*' (Ohkuma *et al*., 2015) showing the distinctive absence of orthologs related to motility and chemotaxis.

Here the first treponeme genomes isolated from the lower termite *Reticulitermes speratus* are analysed in conjunction with the genomes of other *Treponema* species to detect environmentally specific orthologous genes and core genomes, using hierarchal clustering and tree classification statistical validation. The Random forest algorithm is a useful classification algorithm for inferring importance to certain variables in defining a class. Random Forest was utilised to classify the relative importance of *Lactobacillus lactus* strains orthologs based on phenotype derived traits (Bayjanov *et al*., 2013). Here the algorithm is used to analyse orthologs of statistical importance in classifying the *Treponema* species by environment, mode of life and in terms of the wood-eating termite differences between lower and higher associated orthologs.

Methods

Sampling, isolation and sequencing

Reticulitermes spearatus termites were sampled from mount Tsukuba in Ibaraki prefecture, Japan (36°13'47.3"N 140°07'27.1"E). A *R. spearatus* gut sample was prepared and dissected as previously described, due to the termite's size, 10 worker guts were used to generate the sample. The sample was dyed using celltracker CMFDA and fluorescence activated cell sorted on Beckman Coulters MoFloXDP as previously described (Yuki *et al*., 2015). Single cell isolates were lysed and whole genomes amplified using a previous protocol (Yuki *et al*., 2015), and screened using 16S rRNA gene sequence identity. Samples successfully identified as Treponema and that passed quality control thresholds were utilised in Nextera sample library preparation and sequenced on the Illumina MiSeq platform. Illumina 2x300bp V3 chemistry was used to generate clusters.

Strain Selection

All cultured and genome sequenced type strains of all available treponemes at the time (July 2016) were utilised in the analysis. Treponeme species isolated from human and rabbit blood, the sexually transmitted infections (STI), *Treponema pallidum* and *Treponema paraluiscuniculi*; species isolated from bovine rumen: *Treponema*. sp JC4, *Treponema*. sp C6A8, *Treponema bryantii*, *Treponema saccharophilum*; species isolated from bovine digital dermatitis lesions: *Treponema brennaborense*, *Treponema pedis*; from porcine gut: *Treponema succinifaciens*; species from human periodontal samples: *Treponema denticola*, *Treponema putidum*, *Treponema maltophilum*, *Treponema medium*, *Treponema lecithinolyticum*, *Treponema socranskii*, *Treponema phagedenis* and *Treponema vincentii*; from a thermal spring: *Treponema caldarium*; treponemes isolated from the higher termite (*Nasutitermes*) and species isolated from the lower termite gut: *Treponema azotonutricium*, *Treponema primitia*, 'Canditatus *Treponema intracellularis'* and treponemes isolated in the current study from *Reticulitermes speratus*.

Genome assembly and annotation

Sequenced genomic reads were quality and adapter trimmed, and assembled into contigs using the SPAdes (Bankevich *et al*., 2012) assembler in single cell mode and evaluated using QUAST (Gurevich *et al*., 2013) and a custom perl script. Assembled samples were annotated using the Prokka pipeline (Seemann, 2014) and single cell completeness evaluated using the single copy database (Rinke *et al*., 2013) and used to estimate overall genome size. Additional Treponema genomes used in the analyses were downloaded as contig nucleic acid fasta files from the NCBI database and JGI's IMG database, contig files were run through the same Prokka annotation steps to be used in the downstream analysis. Three sample Prokka generated genbank files were uploaded to RAST (Aziz *et al*., 2008) and compared manually for SEED and KEGG pathway information

Phylogenetic analysis

16S rRNA gene sequences were classified into phylogenetic *Treponema* sub clusters using Mothur (Schloss *et al*., 2009) and DictDB (Mikaelyan *et al*., 2015). 16S rRNA gene sequences were extracted from the annotation or downloaded from NCBI and aligned by MUSCLE (Edgar, 2004), the optimal nucleic acid substitution model was selected for in MEGA (GTR+I+G; Tamura *et al*., 2011) and then used to create a 16S rRNA gene sequence maximum likelihood phylogenetic tree based on those parameters.

Genome Clustering

To approximate complete gene sets, Prokka annotated amino acid files from *Reticulitermes* derived treponeme genomes were concatenated based on individual samples phylogenetic positions. CD-Hit was used to condense these redundant sequences appropriately using default parameters (Li and Godzik, 2006).

Prokka annotated amino acid files were used in the cluster analysis using OrthoMCL (Li *et al.,* 2003) in mode 1 with default parameters. The subsequent generated ortholog table was transformed using a custom perl script into presence and absence binary table to be used further downstream. The ortholog table was also used to generate overall core genome and habitat related ortholog summaries. The binary table was used to generate a presence absence plot using gplots in R and this was then transformed into a binary distance matrix.

The binary ortholog table was used as a matrix input to the randomForest package in R (Liaw and Wiener, 2002). The tuneRF function was used to calculate the optimal number of variables sampled at each split. The general variable importance was calculated for each of the environmental variables.

Results and Discussion

Reticulitermes single cell genomes

We sequenced 12 FACS isolated *Treponema* identified samples from the termite *R. speratus*. General genome assembly and annotation statistics were calculated (Table 1.; specific samples Table 2). The completeness of samples ranged from 28% to 68%, based on the presence of single copy genes (Rinke *et al*., 2013). All sample GC contents ranged between 47 and 57%, all except one genome was above 50% GC content. The full genomic 16S rRNA gene sequences were recovered from the assemblies and used to generate the phylogenetic tree (Figure 1) Samples were classified as belonging to sub cluster Ia and Ib, using the DictDB (Mikaelyan *et al*., 2015) and when shown in the tree are relatively clustered together. Sample 4H E3 was classified as sub cluster Ia but clusters with sub cluster Ib and sample R3-C5 was classified as Ib but clustered with Ic samples from the higher termite *Treponema*. The average evolutionary divergence was 0.068 base substitutions across the 16S rRNA gene sequence of all samples. 16S copy number was 1 per sample, Spirochetes tend to have low copy numbers and this probably reflects the ecological niche (Klappenbach *et al*., 2000), with termite's recalcitrant diet and lack of usable nitrogen (Graber *et al*., 2004).

Table 1. Properties of the *Reticulitermes* isolated treponeme genomes

Number of Samples	12
Genome Completeness %*	28-68
Estimated Genome Size	3.5Mb
Average Protein Coding Sequences	1,635
Average GC content %	51.8
16S rRNA Copy Number	1
Sub Cluster phylogeny	a and b

Table 2. Genome summaries of *Reticulitermes* isolated treponeme single cell assemblies.

	10H-A6	AH-A2	4H-E3	2nd-8H-B8	2nd-4H-A10	4H-D10	2nd-4H-E11	2nd-8H-C6	2nd-8H-B3	R3-C5	10H-E11	8H-G7
Total Gb	3.26	0.81	1.80	2.14	1.52	1.12	2.12	1.05	0.78	0.56	2.30	1.76
No. contigs	969	487	1151	1199	860	510	825	453	372	230	962	828
N50	16075	4566	3152	4419	4432	6899	13846	7316	7933	34377	7839	7796
Longest contig (bp)	88217	20312	18046	26915	29790	27266	82888	30854	40754	57689	73415	69375
GC content %	51.0	53.0	47.7	50.7	52.7	54.1	57.6	52.4	51.9	53.1	52.7	51.3
tRNA	33	6	18	17	14	11	28	11	8	8	22	23
5S rRNA	1	1	0	1	2	2	1	1	1	1	1	1
16S rRNA	1	1	1	1	1	1	1	1	1	1	1	1
23S rRNA	1	1	0	1	1	1	1	1	1	1	1	0
CDS	3154	827	2017	2297	1515	1218	2059	1135	822	527	2270	1780
SCG*	62	34	50	45	71	44	38	35	31	32	74	62
Genome complete %	56.9	31.2	45.9	41.3	65.1	40.4	34.9	32.1	28.4	29.4	67.9	56.9
Sub cluster	a	a	a	a	a	a	a	a	a	b	b	b

*SCG single cell genes based on the 139 single gene database (Rinke *et al*., 2013).

Treponeme Ortholog analysis

The whole *R. speratus* isolated treponeme genomes applied into the orthoMCL analysis were six samples from sub cluster Ia (samples 10H-A6, AH-A2 and 4H-E3; 2nd-8H-B8, 2nd-4H-A10 and 4H-D10) and three from sub cluster Ib (samples R3-C5, 10H-E11_5 and 8H-G7), with three samples per amino acid fasta file used. These samples were chosen based on 16S rRNA gene sequence similarity and their overall GC content similarity (Figure 1; Table 2). The samples amino acid fasta files were concatenated, to generate a more complete representation of a genome, due to their fragmentary recovery and amino acid sequence redundancy reduced. The generated ortholog table clustered 52,852 proteins into 7,053 orthologous groups, based on the 27 taxa listed on the treponeme 16S rRNA gene sequence phylogenetic tree (Figure 1), and was transformed into a presence absence plot (Figure 2).

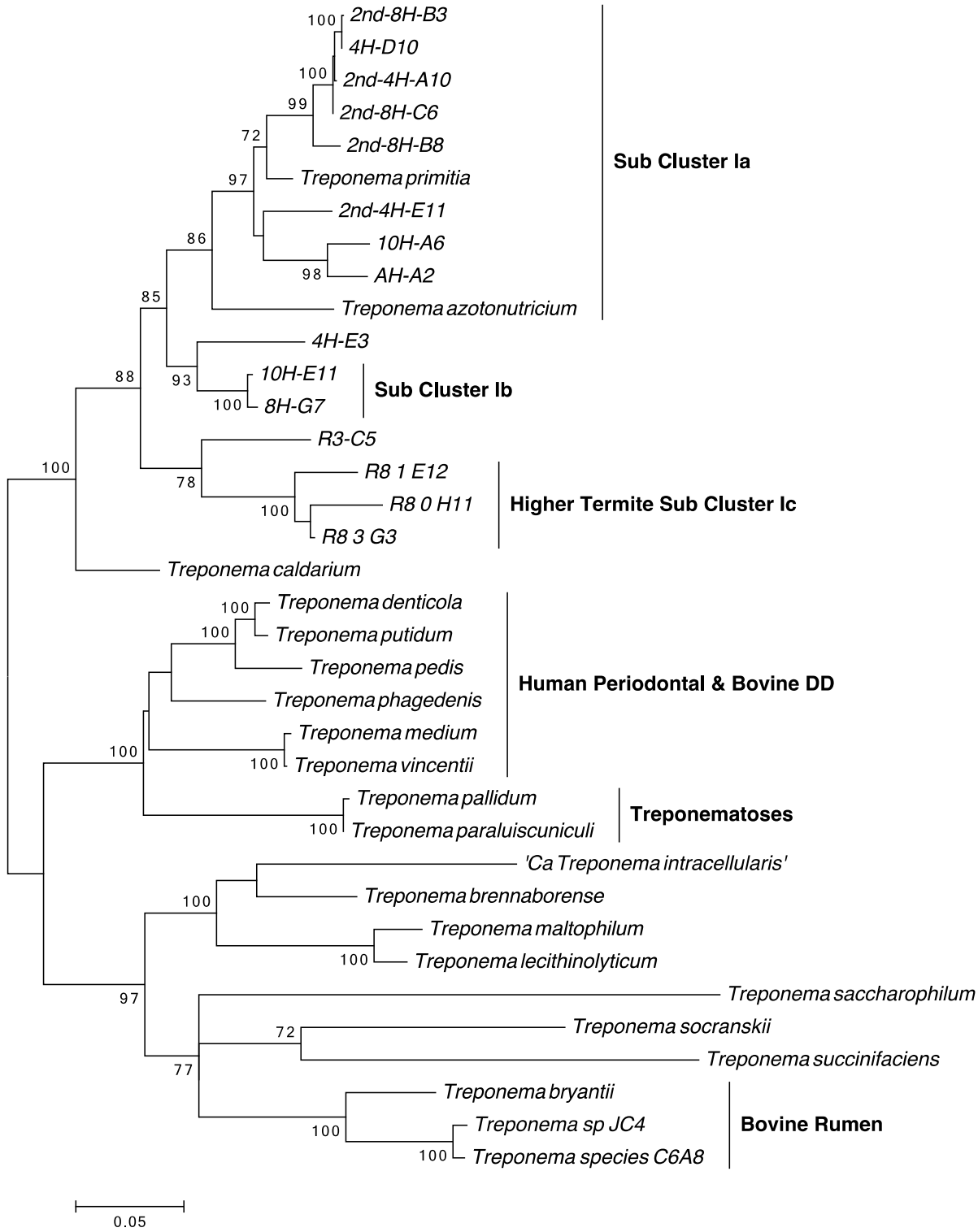

Figure 1. Maximum likelihood 16S rRNA gene phylogenetic tree. Samples sequenced from *R. speratus* gut *Treponema* cluster near the top of the tree and are listed in Table 2. The tree was created using the generalised time reversible model with I and G. The final alignment was made from 1287 nucleic acid positions, the analysis was subject to 1000 bootstrap resamplings, shown on branch bifurcations. Bovine DD, Bovine Digital Dermatitis.

.

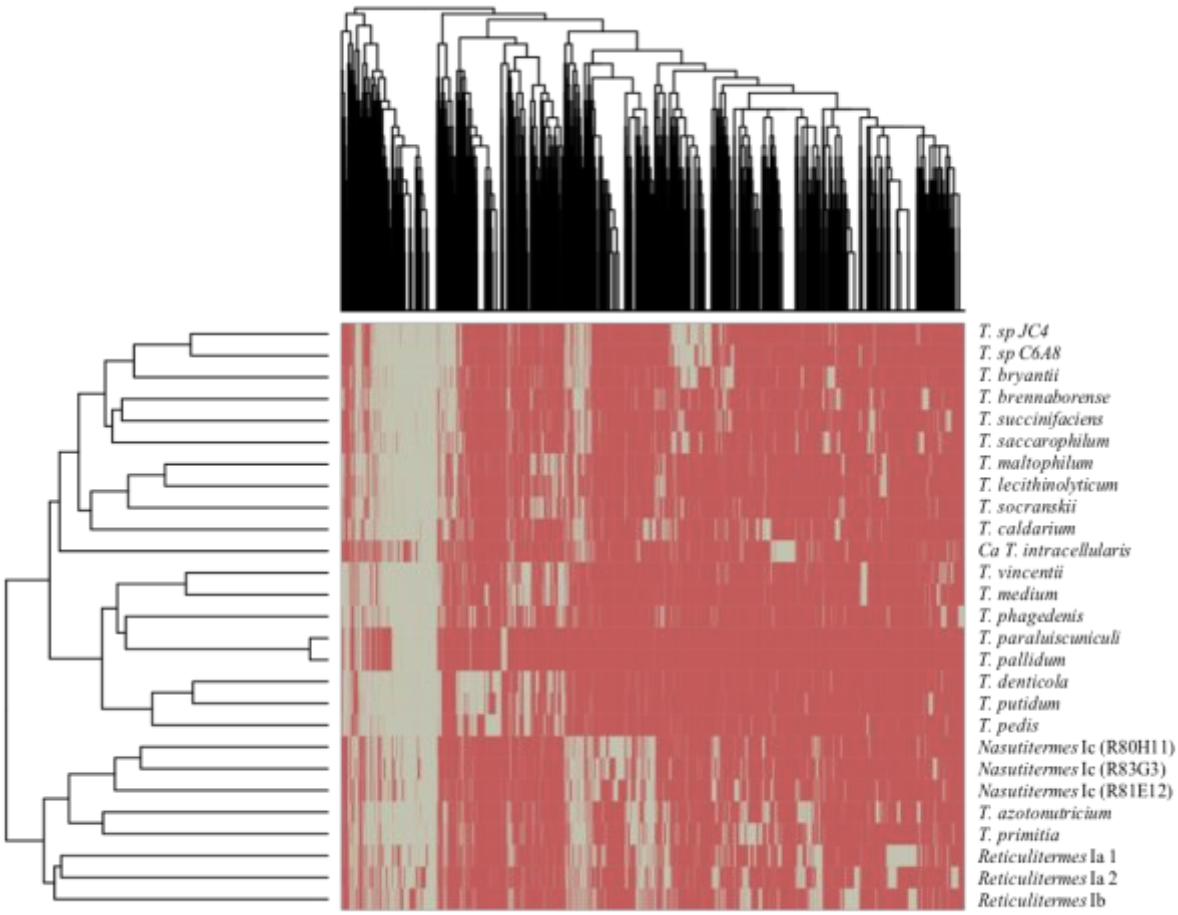

Figure 2. Presence and absence plot of orthologous groups throughout the whole *Treponema* genus, utilising whole sequenced genomes. The top tree represents the orthologous protein clusters based on sequence homology, the left tree represents the relatedness of samples based on the presence and absence of orthologous groups. White indicates presence and red represents absence of an ortholog.

There is a dense set of core orthologs represented to the left of the plot. Here the least orthologs are seen in three taxa, 'Candidatus *Treponema intracellularis*' and two species responsible for syphilis, *T. pallidum* and *T. paraluiscuniculi*. The termite gut derived '*Ca T. intracellularis*' is an endosymbiont of a protist in the lower termite and has lost orthologous groups essential for 'self-sufficiency' in the transition from free-living to endosymbiont (reductive evolution), these groups include motility, carbohydrate metabolism and transport. The two treponematosis causing species have evolved as pathogens, relying on the hosts blood for survival, this in term removes the need for certain metabolic and biosynthetic pathways, and subsequent removal and streamlining of the genome (reductive genome evolution (Moran, 2002)). There were several clusters specific for environment, non-endosymbiont, termite associated genomes clustered together, human periodontal and the treponematoses share the same cluster, 'Ca *T. intracellularis*' was solitary and the last cluster was split between additional periodontal and

bovine associated samples. Some taxa followed the 16S rRNA gene sequence phylogeny however parts did not correspond to their taxonomic positions (Figure 1), *T. caldarium* clustered away from termite associated termite treponeme cluster I sequences. The presence absence plot was transformed into a distance matrix which gave a more accurate representation of the *Treponema* clusters (Figure 3).

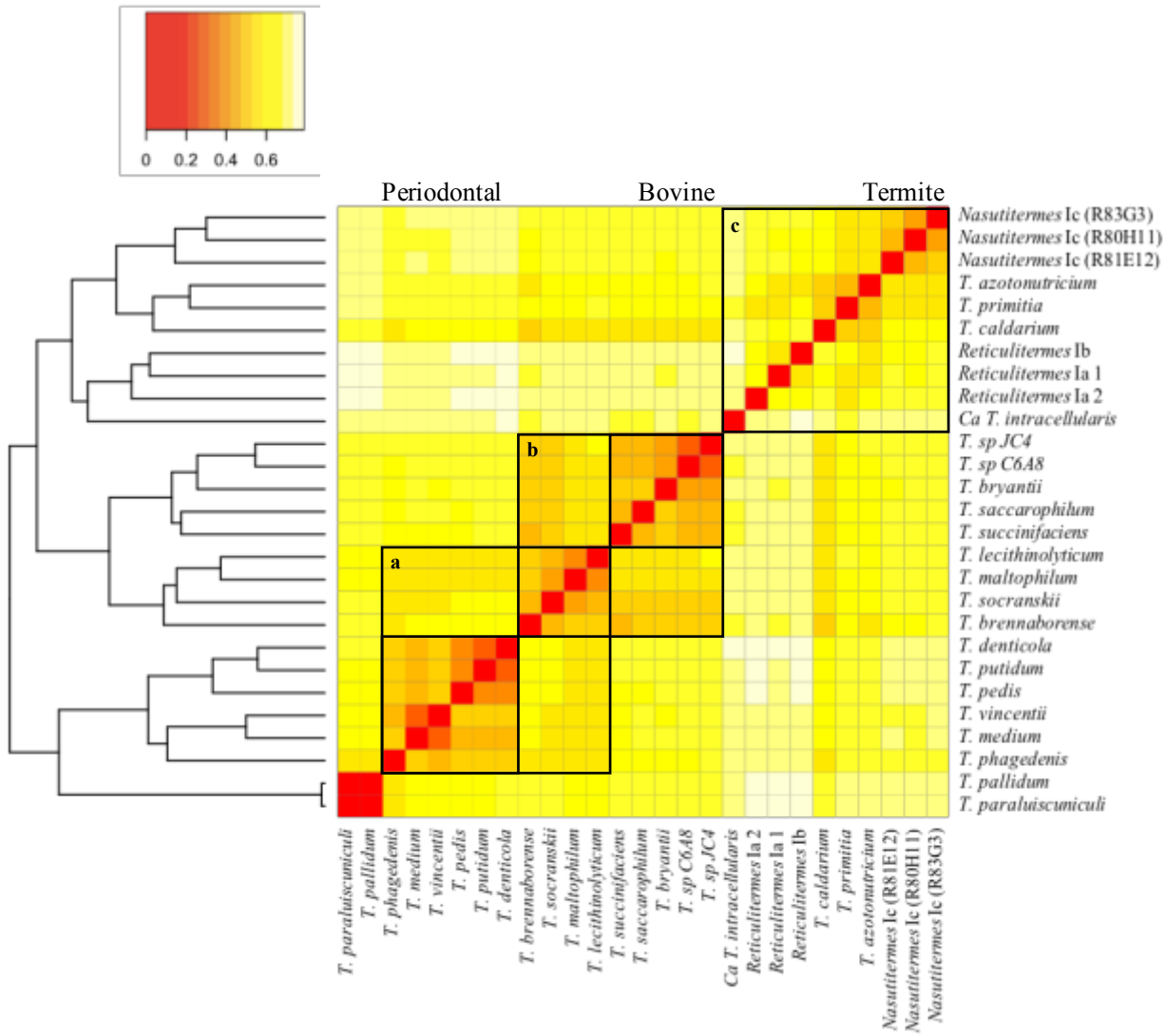

Figure 3. Distance matrix based on presence and absence of 7,053 orthologous groups of the Treponema genus. The key in the top right corner indicates the similarity between species, red being identical to white indicating the greatest distance between species. Boxes represent environment, box a) represents human periodontal species, box b) represents those associated with human periodontal, bovine digital dermatitis and rumen species and box c) represents termite associated species exception *T. caldarium* from a thermophilic microbial mat.

The distance matrix showed that the treponemes genomes associated into three main clusters that were differentiated by host human, bovine and termite. Human related periodontal strains shared the most similarity from *T. phagedenis* to *T. denticola* (Figure 3a). Overlap occurred with human periodontal and bovine digital dermatitis strains (Figure 3a, 3b), it is noted that species of oral treponemes are capable of simple sugar fermentation, having less of a pathogenic profile (Chan and McLaughlin, 2000) and this was evident with *T. socranskii* and *T. brennaborense*, whose profiles were similar to those of rumen species. The latter also shared an increased similarity with that of *T. caldarium*.

Termite associated samples clustered by host organism *Nasutitermes*, *Reticulitermes*, and *Zootermopsis* (Figure 3c). Interestingly *T. caldarium* shares similarity with the termite gut treponemes species but also with the digital dermatitis pathogen *T. brennaborense*. The endosymbiont '*Ca T. intracellularis*' clustered with the termite species but was relatively dissimilar to most species except *T. primitia* and the rumen species, this is most likely due to being an endosymbiont with a specialised lifestyle of reductive acetogenesis, shared with *T. primitia* and taxonomically derived from termite treponeme cluster II, respectively.

Random forests classification

The random forest algorithm classified importance to certain orthologs based on the defining variable. It uses a decision tree classification algorithm, based on the presence and absence of orthologs to score the importance, and with the generation of trees, the algorithm classifies the most popular class within the forest (Breiman, 2001). The random forests classification utilised two data sets, the entire *Treponema* genus and a subset of termite derived species only, here the most important orthologs associated with the treponemes environment (periodontal, lower and higher termite, rumen, digital dermatitis, STI). The most important orthologs shown in the random forest classification had the greatest decrease in accuracy and in the gini index (Figure 4).

The two top most important orthologs were classified from the human periodontal species and both assigned as potential virulence factors. OrthoMCL 693, an internalin-J precursor, is a virulence factor and potentially involved in the degradation and perturbation of host epithelia important in the pathogenesis of periodontal *T. denticola* (Seshadri *et al*., 2004). Here it was found as 25 paralogous genes in eight taxa all associated with periodontitis. The second most important ortholog was BrnA antitoxin, another putative virulence factor (Heaton *et al*., 2012), may be important in the pathogenesis by periodontal treponemes. In relation to the rumen

treponeme species, OrthoMCL 3809 was classified the most important ortholog. This codes for the fermentation-respiration switch protein FrsA and is involved in modulating the PTS system (Phosphotransferase system), which is potentially important to species involved in sugar fermentation and adaption to different dietary carbon substrates in the rumen (Koo *et al*., 2004).

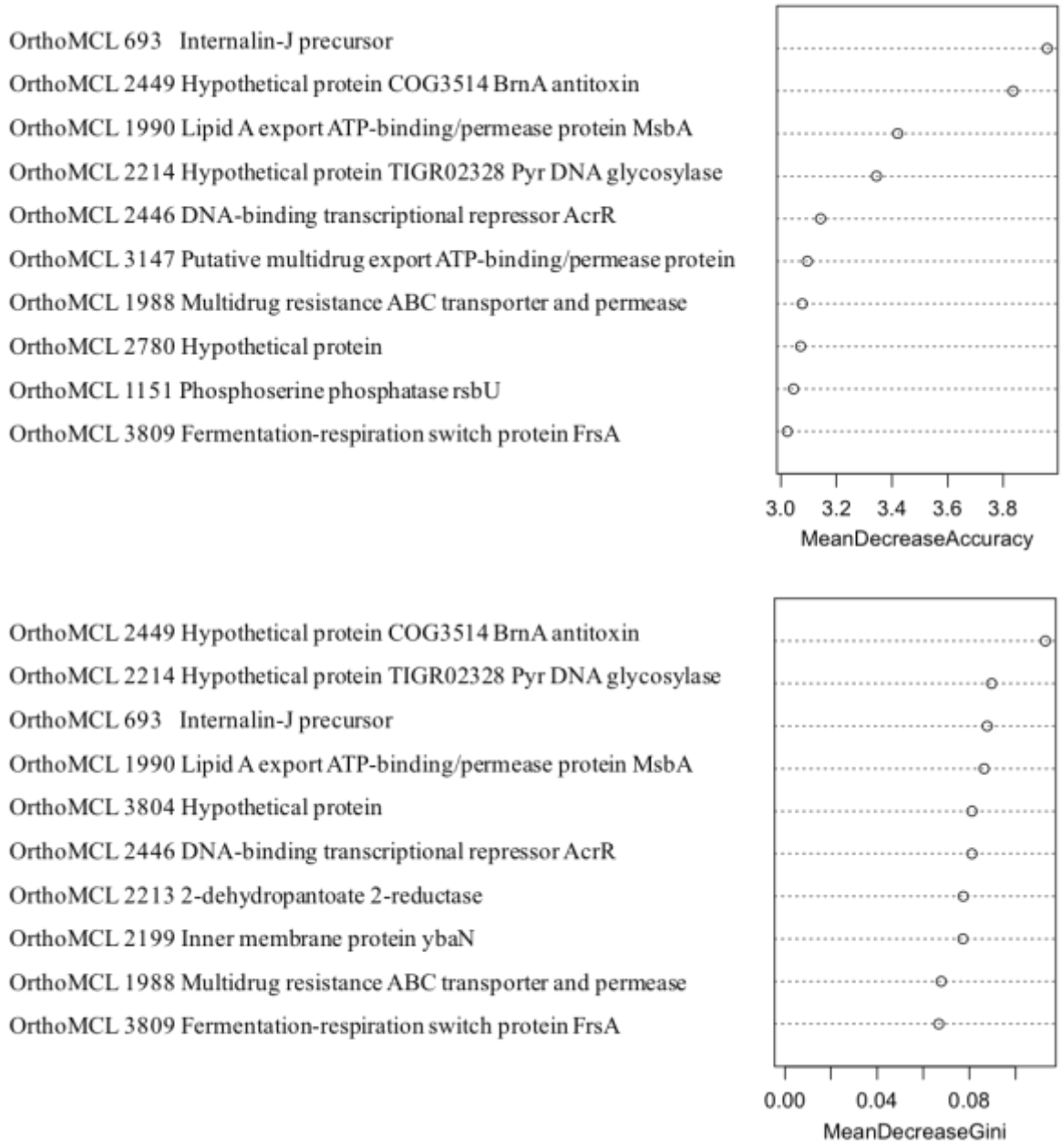

Figure 4. **Variable importance plots of *Treponema* orthologs in differentiating the environmental variable.** The greater the decrease in accuracy and gini the more important the ortholog in classifying the species by the environmental variable.

Table 3. The principal orthologs characterising higher and lower termite *Treponema* species using random forests classification.

Ortholog	Putative encoded protein	Mean Decrease Accuracy	Mean Decrease Gini	Termite Treponema
ORTHOMCL1015	Hypothetical protein	2.229300804	0.017333333	Lower
ORTHOMCL4985	Domain of unknown function (DUF3560)	2.141724088	0.017333333	Higher
ORTHOMCL1417	Phosphatase yihX	2.00401204	0.024	Lower
ORTHOMCL1016	Glucose-1-phosphate thymidylyltransferase 1	2.00401204	0.019111111	Lower
ORTHOMCL5049	GloB, putative metal-dependent RNase	1.986388742	0.018388889	Higher
ORTHOMCL3087	Anaerobic benzoate catabolism transcriptional regulator	1.967759644	0.024888889	Higher
ORTHOMCL4982	Hypothetical protein	1.940975704	0.017333333	Higher
ORTHOMCL5059	Hypothetical protein(COG3680)	1.734654744	0.036444444	Higher
ORTHOMCL3564	Hypothetical protein	1.734654744	0.010222222	Higher
ORTHOMCL4127	Phosphoenolpyruvate synthase	1.720451972	0.019555556	Higher

In higher termites, the highest importance scoring ortholog was OrthoMCL 4994, this was annotated as thymidylate synthase (ThyA), throughout the *Treponema* genus there is a conserved thymidylate synthase (ThyX) and this is evident in the lower termite cultured treponemes, however in these higher termites ThyA is present, coupled with dihydrofolate reductase. This ortholog was most likely assigned the high importance due to its uniqueness among these treponemes and its protein domain structure is different from that of ThyX.

Lower and higher termite ortholog importance classifications were classified separately based on the random forests classication (Table 3). The ortholog categorised by random forests as having high importance, OrthoMCL 4127 is shared by the higher termite treponeme species and also by *T. caldarium*. This ortholog was annotated as phosphoenolpyruvate synthase that irreversibly catalyses the last step in glycolysis, the enzyme shares the same function as the rate limiting reversible pyruvate phosphate dikinase. Higher termites that possess both enzymes are putatively more efficient at regulating glycolysis and gluconeogenesis, and has also been shown to exist in termite gut protists (Slamovits and Keeling, 2006). Most lower and higher termite orthologs that were highlighted were associated with central metabolic pathways, for example OrthoMCL 1417 that putatively encodes phosphatase yihX, this enzyme dephosphorylates alpha-D-glucose 1-phosphate which is another enzyme affiliated with carbohydrate metabolism and transport.

Conclusion

Treponemes are a highly diverse genus of bacteria, treponemes are significantly difficult to culture, they are unique, with distinctive morphologies and variability in physiology and lifestyle. As such the *Treponema* genus is still an underrepresented bacteria genus in current databases, the diversity between species is still so vast and represents gaps in our current understanding. The sequencing of additional strains both culturable and unculturable will contribute to understanding the mechanisms to which this genus has evolved both multiple lifestyles of pathogenicity and symbiosis.

The power and depth of the Random forests classifications of this genus will improve with an increase in the number of strains used and with the addition of phenotypic metadata than what was used in the current study. The algorithm is useful in determining genes responsible for, or associated with phenotypes when there are distinct binary variables, for example antimicrobial resistance, sulphate reduction. Using complete genomes should also increase the utilisation of this method, although this is not essential for the operation of this algorithm the accuracy of selected ortholog importance is decreased.

References

Aziz RK, Bartels D, Best AA*, et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75.

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

Bayjanov JR, Starrenburg MJ, van der Sijde MR, Siezen RJ, van Hijum SA (2013) Genotype-phenotype matching analysis of 38 Lactococcus lactis strains using random forest methods. *BMC Microbiol* **13**, 68.

Breiman L (2001) Random Forests. *Mach Learn* **45**, 5-32.

Chan ECS, McLaughlin R (2000) Taxonomy and virulence of oral spirochetes. *Oral Microbiol Immunol* **15**, 1-9.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797.

Fraser CM, Norris SJ, Weinstock GM*, et al.* (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375-388.

Graber JR, Leadbetter JR, Breznak JA (2004) Description of Treponema azotonutricium sp. nov. and Treponema primitia sp. nov., the first spirochetes isolated from termite guts. *Appl Environ Microbiol* **70**, 1315-1320.

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075.

Heaton B, Herrou J, Blackwell A, *et al.* (2012) Molecular Structure and Function of the Novel BrnT/BrnA Toxin-Antitoxin System of Brucella abortus. *J Biol Chem* **287**, 12098-12110.

Iida T, Ohkuma M, Ohtoko K, Kudo T (2000) Symbiotic spirochetes in the termite hindgut: phylogenetic identification of ectosymbiotic spirochetes of oxymonad protists. *FEMS Microbiol Ecol* **34**, 17-26.

Klappenbach JA, Dunbar JM, Schmidt TM (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* **66**, 1328-1333.

Koo BM, Yoon MJ, Lee CR, *et al*. (2004) A novel fermentation/respiration switch protein regulated by enzyme IIAGlc in *Escherichia coli*. *J Biol Chem* **279**, 31613-31621.

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659.

Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News* ***2***, 18-22.

Mikaelyan A, Köhler T, Lampert N*, et al.* (2015) Classifying the bacterial gut microbiota of termites and cockroaches: A curated phylogenetic reference database (DictDb). *Syst Appl Microbiol* **38**, 472-482.

Moran N (2002) Microbial Minimalism: Genome Reduction in Bacterial Pathogens. *Cell* **108**, 583-586

Ohkuma M, Noda S, Hattori S*, et al.* (2015) Acetogenesis from H2 plus CO2 and nitrogen fixation by an endosymbiotic spirochete of a termite-gut cellulolytic protist. *Proc Natl Acad Sci U S A* **112**, 10224-10230.

Rinke C, Schwientek P, Sczyrba A*, et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431-437.

Ritalahti KM, Justicia-Leon SD, Cusick KD*, et al.* (2012) Sphaerochaeta globosa gen. nov., sp. nov. and Sphaerochaeta pleomorpha sp. nov., free-living, spherical spirochaetes. *Int J Syst Evol Microbiol* **62**, 210-216.

Schloss PD, Westcott SL, Ryabin T*, et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537-7541.

Schnorr SL, Candela M, Rampelli S*, et al.* (2014) Gut microbiome of the Hadza hunter-gatherers. *Nat Commun* **5**, 3654.

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069.

Seshadri R, Myers GS, Tettelin H*, et al.* (2004) Comparison of the genome of the oral pathogen Treponema denticola with other spirochete genomes. *Proc Natl Acad Sci U S A* **101**, 5646-5651.

Slamovits CH, Keeling PJ (2006) Pyruvate-Phosphate Dikinase of Oxymonads and Parabasalia and the Evolution of Pyrophosphate-Dependent Glycolysis in Anaerobic Eukaryotes. *Eukaryot Cell* **5**, 148-154

Stanton TB, Canale-Parola E (1980) *Treponema bryantii* sp. nov., a Rumen Spirochete that Interacts with Cellulolytic Bacteria. *Arch Microbiol* **127**, 145-156.

Tamura K, Peterson D, Peterson N*, et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739.

Warnecke F, Luginbuhl P, Ivanova N*, et al.* (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**, 560-U517.

Yuki M, Kuwahara H, Shintani M*, et al.* (2015) Dominant ectosymbiotic bacteria of cellulolytic protists in the termite gut also have the potential to digest lignocellulose. *Environ Microbiol* **17**, 4942-4953.

Chapter VI

General Discussion

General Discussion

6.1 A brief history

Spirochetes were first observed and described in the guts of termites in the latter half of the nineteenth century by renowned parasitologist Joseph Leidy. It was not until the advent of molecular biology where 16S rRNA gene sequence analysis could identify these organisms. In terms of termite spirochete research, the cultivation of two species of treponemes was an important milestone in the physiology of these symbiotic bacteria. The pure culture of *Treponema primitia* (Leadbetter *et al*., 1999) allowed analysis of its physiology, and its reductive acetogenesis from H_2 and CO_2. This was accompanied by the pure culture of *Treponema azotonutricium* isolated from the same termite species (Graber *et al*., 2004) and the discovery of proficient nitrogen fixation and diazotrophic growth of the treponeme. These two organisms generated both carbon and nitrogen sources essential for the host termite's dietary needs. Then, subsequent studies showed that the co-culturing of these treponemes enhanced each other's growth reciprocally, due to the provisioning of metabolic intermediates evidenced by metatranscripomic analysis (Rosenthal *et al*., 2011).

Treponemes from higher termites were never cultured. Metagenomics explored the hindgut from a species of the wood feeding termite genus *Nasutitermes*, where the whole gut community was profiled in another milestone piece of research (Warnecke *et al*., 2007). This research found similar physiological mechanisms as those present in the lower termite gut. Individual treponeme genomes were not recovered but contigs were putatively classified and 16S rRNA gene sequences were used to taxonomically assign treponeme groups.

Phylogenetic research continued from the two treponeme clusters (Ohkuma *at al*., 1999, Iida *et al*., 2000) to full database assignments (Mikaeylan *et al*., 2015a). The *Treponema* cluster I to which both pure culture species *T. primitia* and *T. azotonutricium* were assigned, was further divided into sub clusters where abundances of sub clusters differed among different termite species, and these microbial community structures were determined by the host's diet (Mikaeylan *et al*., 2015b).

Treponemes inhabiting the gut of lower termites also associate as ecto and endo symbionts of the gut protists. The treponemes are not alone in associating with the protist gut population, as various phyla form these relationships. Another milestone was the elucidation of two

endosymbionts, an Elusimicrobia of *Trichonympha agilis* (Hongoh *et al*., 2008a) and a Bacteroidales endosymbiont of the protist *Pseudotrichonympha grassii* (Hongoh *et al*., 2008b). These two endosymbiont genomes were the first instance of applying single cell genomics to exploring termite gut symbioses.

6.2 The wood feeding higher termite *Nasutitermes takasagoensis*.

The tropical arboreal termite *Nasutitermes takasagoesnsis* is an apical wood feeding species. This termite's microbiota has appropriated the wood fibre niche, filled by protists in the basal termite lineages. Components of the wood fibre associated community are the sub cluster Ic and If treponemes of *Treponema* cluster I, TG3 and Fibrobacteres (Mikaelyan *et al*., 2014). The genomes for these community members have been explored in this thesis to understand the contribution these groups of bacteria make to the overall nutritional symbiosis. These were the first treponeme genomes isolated from a higher termite and the first of sub clusters Ic, If and Ih; sub cluster Ia genomes of *T. azotonutricium* and *T. primitia* are already published. Due to the nature of SCG, genome recovery was not 100% and even using hybrid assembly this was not attained. Hybrid assembly condenses multiple samples into one assembly; this may cause the heterogeneity between individual genomes to be hidden and differences in genome coverage and rare reads (Bankevich *et al*., 2012) from the sequencing can cause some genomic information during the genome assembly to be lost.

These are the first SCG assemblies of Fibrobacteres and TG3. These taxa had been previously analysed as composite genomes from metagenomics samples (Abdul Rahman *et al*., 2015), however the composite assembly yielded no ribosomal RNA from their *Nasutitermes* derived Fibrobacteres genome.

The wood fibre-associated bacteria all play roles in lignocellulose degradation. These are evidenced by multiple glycosyl hydrolase enzymes and transporters encoded in their genomes. Both cellulose and hemicellulose are putatively broken down synergistically, the need for cellobiohydrolases within this symbiosis is potentially alleviated by the suggested mechanisms of cellulose degradation akin to those in the rumen cellulolytic *Fibrobacter succinogenes* (Wilson, 2009) and most likely performed by the related Fibrobacteres and TG3. Hemicellulases are important to utilise the entirety of the wood fibre so that lignin remains and is excreted. The diversity of the treponeme sub clusters allows for niche partitioning of the lignocellulose resource, and the utilisation of xylose and derivations by the different sub clusters.

The products of the cellulose and hemicellulose fermentation are utilised in acetogenesis by the treponemes and to a lesser extent the Fibrobacteres, thus most hydrogen produced from these reactions is converted to acetate by reductive acetogenesis (Brune and Ohkuma, 2011). The fibre-associated treponeme sub clusters Ic and If would be in an advantageous position to utilise these by products produced by TG3 and Fibrobacteres. The possession of multiple methyl accepting chemotaxis proteins and motility related genes infers their importance for the treponemes.

All three groups of fibre-associated bacteria have the potential for nitrogen fixation, with the presence of the minimal genes required for nitrogen fixation (Wang *et al*., 2013). TG3 and Fibrobacter genomes share the same nitrogenase phylogenetic cluster, whereas the treponemes nitrogenase is of a separate cluster but still believed to be functional (Zheng *et al*., 2016). As nitrogen is deficient in the host termites natural diet, the ability of all members to carry out this function maybe crucial, their presence maybe important for sustaining the genes within this microbiome environment. The treponeme sub clusters also recycle urea, allowing the preservation of nitrogen in this environment and upscaling it to other nitrogenous compounds.

6.3 The *Eucomonympha* endosymbiont.

Wood feeding lower termites harbour cellulolytic protists that sequester wood particles and degrade the lignocellulose. The protist *Eucomonympha* processes the wood particles but harbours endosymbiotic treponemes. Endosymbionts of termite gut protists have been shown to fix nitrogen (Hongoh *et al*., 2008b) and their roles in hydrogen metabolism suggested (Desai and Brune, 2012). The genomes of 'Ca *T. intracellularis*' strains were the first treponeme endosymbiont sequenced. The study also revealed a similar endosymbiont in the protist *Teranympha* from *R. speratus*. Metabolic activity was elucidated by biochemical analysis and genome analysis to establish that the endosymbiont plays major roles in the fixing of nitrogen and in reductive acetogenesis within the protist utilising by-products of fermentation. The endosymbiont shows the beginnings of genome reduction with the loss of orthologs assigned to motility, and thus losing its canonical spiral morphology. The genome however is still large in comparison to other endosymbionts with multiple gene sets specialised for its symbiotic roles and large amounts of mobile genetic elements that usually precede genome erosion (Zheng *et al*., 2016).

6.4 Struggles and limitations of single cell genomics

Microbial single cell genomics relies on improvements in both molecular techniques and bioinformatics. However, the knowledge garnered is still limited by itself; multi-omic approaches are becoming more affordable and offer a greater more in-depth understanding of microbial communities in diverse environments. Here presenting single cell genomics with metatranscriptomics gave greater substance to the results (Chapters II & III). Biochemical analyses with single cell genomics can also evidence results on a more intrinsic level (Chapter IV). Although the SCG generated assemblies are many times incomplete, they can still be used in comparative genomic studies. Incomplete genomes offer insight based on the presence of functional genes. Dismissing a putative function of the organism is misleading, as absence could be due to the incompleteness of the genome. The random forest classification algorithm can overcome missing values in the data and thus have some use in analysing the genomes (Chapter V).

Metagenomics is the most effective partner with single cell genomics. Here shotgun sequences can be combined to fill gaps and even close genomes. Hybrid assemblies (Chapter II) can also be used of identical samples (clonal) to a lesser effect. This benefits from the high throughput nature of sorting and successful sample gating. However, once a cell is isolated and DNA is extracted, the cell is lost. Successive rounds of MDA can be used to generate larger quantities of DNA; however, the reaction increases the probability of contamination at each sample handling and increases the amplification bias in portions of the DNA previously amplified. Improvements in bioinformatics have helped with some of the inherent problems with SCG. SPAdes assembler (Bankevich *et al*., 2012) addressed some of these errors, formation of chimeras (incorrect orientation of genome sequence), produced during the MDA and in dealing with the differences in coverage. BAYESHAMMER error correction tool also improves single cell assembly (Nikolenko *et al*., 2013).

6.4 Future directions

The diversity of the termite gut microbiota provides an abundant source of research with the use of single cell genomics. There are still abundant individual symbiotic relationships to be uncovered in terms of roles in associations with protists to those in higher termites that are yet to be cultured. Even looking at within strains SCG can uncover hidden diversity (Engel *et al*.,

2014). The genomic sequences can even illuminate methods and conditions for culturing these fastidious individuals. Culturing the treponemes from sub clusters Ic and If can classify the roles of these individuals based on biochemical analysis and to see if co-culturing the individuals with TG3 and Fibrobacteres truly enhances their growth and wood particle degradation.

The *Reticulitermes speratus* samples from Chapter V, require further exploration. They are the first sequenced treponemes isolated from the higher termite *R. speratus* and the first genomes of treponemes sub cluster Ib. Further analysis would highlight potential metabolic functions or evolutionary ideas, and using more individual samples would improve the random forests classification of termite orthologs.

Using combined omics approaches can allow for deeper understanding of these symbiotic relationships and in addressing interactions between community members. In future studies when exploring a complex microbial community combining metagenomics, SCG and metatranscriptomics would be the optimal way of analysing a community. This multi-omic approach could categorise abundant taxa and their contributions to the community and look at other key players and the interactions between them.

Termite gut nutritional symbiosis research still has a long way to go in exploring the complete complexity of this efficient biomass converter. In applying the symbiotic model and utilising components of this symbiosis we may be able to generate the artificial and efficient recycling of waste organic material. Novel enzymes may be mined from these sequences and not only those for lignocellulosic digestion but those that may have antimicrobial or other medicinal value, as there are many examples of bioprospecting natural resources, e.g. leaf cutter ants (Currie *et al*., 1999; Knight *et al*., 2003). Further applied science can lead to the effective use of the gut as a means of pest control. This has been recently tested in the paratransgenesis based termite control of *Coptotermes formosanus* with *Trabulsiella odontotermitis* (Tikhe *et al*., 2016). This is the genetic engineering of native gut bacteria to induce the expression of lethal toxins against the host and is easily transmitted through termite colonies.

References

Abdul Rahman N, Parks DH, Vanwonterghem I *et al*. (2015) A phylogenomic analysis of the bacterial phylum Fibrobacteres. *Front Microbiol* **6**, 1469.

Bankevich A, Nurk S, Antipov D, Gurevich AA *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477.

Brune A, Ohkuma M (2011) Role of the termite gut microbiota in symbiotic digestion. In: Bignell DE, Roisin Y, Lo N, (eds). Biology of Termite: A Modern Synthesis. Springer-Verlag: pp 439-475.

Currie CR, Scott JA, Summerbell RC, Malloch D (1999) Fungus-growing ants use antibiotic-producing bacteria to control garden parasites. *Nature* **398**, 701-704

Desai MS, Brune A (2012) Bacteroidales ectosymbionts of gut flagellates shape the nitrogen-fixing community in dry-wood termites. *ISME J* **6**, 1302-1313.

Engel P, Stepanauskas R, Moran NA (2014) Hidden diversity in honey bee gut symbionts detected by single-cell genomics. *PLoS Genet* **10**, e1004596.

Graber JR, Leadbetter JR, Breznak JA (2004) Description of *Treponema azotonutricium* sp. nov. and *Treponema primitia* sp. nov., the first spirochetes isolated from termite guts. *Appl Environ Microbiol* **70**, 1315-1320.

Hongoh Y, Sharma VK, Prakash T, *et al.* (2008a) Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Proc Natl Acad Sci U S A* **105**, 5555-5560.

Hongoh Y, Sharma VK, Prakash T, *et al.* (2008b) Genome of an Endosymbiont Coupling N2 Fixation to Cellulolysis Within Protist Cells in Termite Gut. *Science* **322**, 1108-1109.

Iida T, Ohkuma M, Ohtoko K, Kudo T (2000) Symbiotic spirochetes in the termite hindgut: phylogenetic identification of ectosymbiotic spirochetes of oxymonad protists. *FEMS Microbiol Ecol* **34**, 17-26.

Knight V, Sanglier JJ, DiTullio D *et al*. (2003) Diversifying microbial natural products for drug discovery. *Appl Microbiol Biotechnol* **5**, 446-458.

Leadbetter JR, Schmidt TM, Graber JR, Breznak JA (1999) Acetogenesis from H2 plus CO2 by spirochetes from termite guts. *Science* **283**, 686-689.

Mikaelyan A, Strassert JF, Tokuda G, Brune A (2014) The fibre-associated cellulolytic bacterial community in the hindgut of wood-feeding higher termites (*Nasutitermes spp*.). *Environ Microbiol* **16**, 2711-2722.

Mikaelyan A, Köhler T, Lampert N, *et al.* (2015a) Classifying the bacterial gut microbiota of termites and cockroaches: A curated phylogenetic reference database (DictDb). *Syst Appl Microbiol* **38**, 472-482.

Mikaelyan A, Dietrich C, Köhler T, *et al.* (2015b) Diet is the primary determinant of bacterial community structure in the guts of higher termites. *Mol Ecol* **24**, 5284-5295.

Nikolenko S, Korobeynikov A, Alekseyev M (2013) BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* **14**, S1:S7.

Rosenthal AZ, Matson EG, Eldar A, Leadbetter JR. (2011) RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. *ISME J* **5**, 1133-1142.

Tikhe CV, Martin TM, Howells A, *et al*. (2016) Assessment of genetically engineered *Trabulsiella odontotermitis* as a 'Trojan Horse' for paratransgenesis in termites. *BMC Bioinform* **16**, 202.

Wang L, Zhang L, Liu Z, *et al.* (2013) A minimal nitrogen fixation gene cluster from *Paenibacillus sp*. WLY78 enables expression of active nitrogenase in Escherichia coli. *PLoS Genet* **9**, e1003865.

Warnecke F, Luginbuhl P, Ivanova N, *et al.* (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**, 560-565.

Wilson D (2009) Evidence for a novel mechanism of microbial cellulose degradation. *Cellulose* **16**, 723-727.

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.

Appendix

The following additional papers I have co-authored during my doctoral thesis research.

Starns D, Oshima K, Suda W, Iino T, Yuki M, Inoue J-I, Kitamura K, Iida T, Darby AC, Hattori M, and Ohkuma M. (2014) Draft Genome Sequence of *Cytophaga fermentans* JCM 21142T, a Facultative Anaerobe Isolated from Marine Mud. *Genome Announc* **2**.

My contribution to this study was the writing of the paper, the genome annotation and analysis.

Blow F, Gioti A, **Starns D**, Ben-Yosef M, Pasternak Z, Jurkevitch E, Vontas J, and Darby AC. (2016) Draft Genome Sequence of the Bactrocera oleae Symbiont "Candidatus *Erwinia dacicola*". *Genome Announc* **4**.

My contribution to this study was the isolation of single cells by FACS and MDA of single cells.

Yuki M, Kuwahara H, Shintani M, Izawa K, Sato T, **Starns D**, Hongoh Y, and Ohkuma M. (2015) Dominant ectosymbiotic bacteria of cellulolytic protists in the termite gut also have the potential to digest lignocellulose. *Environ Microbiol*. **17**, 4942-53.

My contribution to this study was the bioinformatic analysis.

genomeA
Journals.ASM.org

Draft Genome Sequence of *Cytophaga fermentans* JCM 21142^T, a Facultative Anaerobe Isolated from Marine Mud

David Starns,[a,b] Kenshiro Oshima,[c] Wataru Suda,[c] Takao Iino,[a] Masahiro Yuki,[d] Jun-Ichi Inoue,[a] Keiko Kitamura,[a] Toshiya Iida,[a] Alistair Darby,[b] Masahira Hattori,[c] Moriya Ohkuma[a,d]

Japan Collection of Microorganisms/Microbe Division, RIKEN BioResource Center, Tsukuba, Ibaraki, Japan[a]; Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom[b]; Center for Omics and Bioinformatics, Graduate School of Frontier Sciences, the University of Tokyo, Kashiwa, Chiba, Japan[c]; Biomass Research Platform Team, RIKEN Biomass Engineering Program Cooperation Division, RIKEN Center for Sustainable Resource Science, Tsukuba, Ibaraki, Japan[d]

***Cytophaga fermentans* strain JCM 21142^T is a marine-dwelling facultative anaerobe. The draft genome sequence of this strain revealed its diverse chemoorganotrophic potential, which makes it capable of metabolizing various polysaccharide substrates. The genome data will facilitate further studies on its taxonomic reclassification, its metabolism, and the mechanisms pertaining to bacterial gliding.**

Received 21 February 2014 **Accepted** 10 March 2014 **Published** 27 March 2014

Citation Starns D, Oshima K, Suda W, Iino T, Yuki M, Inoue J-I, Kitamura K, Iida T, Darby A, Hattori M, Ohkuma M. 2014. Draft genome sequence of *Cytophaga fermentans* JCM 21142^T, a facultative anaerobe isolated from marine mud. Genome Announc. 2(2):e00206-14. doi:10.1128/genomeA.00206-14.

Address correspondence to Moriya Ohkuma, mohkuma@riken.jp.

The bacterium *Cytophaga fermentans*, belonging to the phylum *Bacteriodetes* and first described in 1955 (1), is found commonly in marine mud, near shores, and on decaying marine organisms (2). Metabolic studies on the species in the order *Cytophagales* have shown that they are diverse chemoorganotrophs able to degrade many biomacromolecular compounds, including chitin and cellulose (2). *C. fermentans* is a rod-shaped, Gram-negative, facultative anaerobe, with anaerobic growth at the expense of organic compounds (1). Like other cytophagas, it is motile, using gliding to move across solid surfaces. The mechanism for gliding still remains to be clarified but recently a model was created for *Flavobacterium johnsoniae* in which helical surface proteins are thought to play a crucial role (3). Although gliding motility is a typical feature of cytophagas, phylogenetic analyses based on the 16S rRNA gene sequence revealed that *C. fermentans* is misclassified as a species of *Cytophaga* and should be classified to the order *Bacteroidales*, not to *Cytophagales* (4, 5).

The type strain *Cytophaga fermentans* JCM 21142 was sequenced *de novo* using the Ion Torrent PGM system, generating 1,115,426 quality-filtered reads. These reads were assembled using Newbler version 2.8 (Roche) into 299 contigs (longest, 218,358 bp) with an N_{50} length of 59,017 bp. The resulting draft genome sequence is 5,649,512 bp, with 38.6× redundancy and a G+C content of 37.4%. The draft genome sequence was annotated using the Victoria Bioinformatics Consortiums' pipeline Prokka (Prokaryotic Genome Annotation System) version 1.5.2 and the genome annotation was completed using Prodigal (6) version 2.6, to identify 4,781 protein coding sequences. Aragorn (7) version 1.2.34 predicted 46 tRNAs, Infernal (8) version 1.1 predicted 24 noncoding RNAs (ncRNAs), and RNAmmer (9) version 1.2. predicted 3 rRNAs.

From the annotation, 580 genes and gene fragments were assigned to 78 different Carbohydrate-Active Enzyme database (CAZy) families (10). Those identified included genes for galactosidases, mannosidases, xylanases, chitinases, glucosidases, glucanases, and agarases, suggesting the utilization of various compounds as carbon and energy sources. The annotation also uncovered genes for carbon fixation, supporting the CO_2-requiring nature of *C. fermentans* in a defined medium (1), as well as bacterial gliding membrane protein families Gld and Spr, which could allow further elucidation of the mechanism of bacterial gliding. The sequence data and annotated genome will allow improved comparative phylogenetic analyses to better establish the *C. fermentans* taxonomic position within the *Bacteriodetes* phylum.

Nucleotide sequence accession numbers. The genome sequence of *Cytophaga fermentans* strain JCM 21142^T has been deposited in DDBJ/EMBL/Genbank under the accession numbers BAMD01000001 through BAMD01000299.

ACKNOWLEDGMENTS

This work was supported by the Genome Information Upgrading Program of the National BioResource Project from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

We thank Hiromi Kuroyanagi (the University of Tokyo) for technical support.

REFERENCES

1. **Bachmann BJ.** 1955. Studies on *Cytophaga fermentans*, n.sp., a facultatively anaerobic lower myxobacterium. J. Gen. Microbiol. **13:**541–551. http://dx.doi.org/10.1099/00221287-13-3-541.
2. **Reichenbach H.** 2006. The order *Cytophagales*, p 549–590. *In* Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E (ed), The prokaryotes: a handbook on the biology of bacteria, vol 7. Springer-Verlag, New York, NY.
3. **Nakane D, Sato K, Wada H, McBride MJ, Nakayama K.** 2013. Helical flow of surface protein required for bacterial gliding motility. Proc. Natl. Acad. Sci. U. S. A. **110:**11145–11150. http://dx.doi.org/10.1073/pnas.1219753110.

4. **Gherna R, Woese CR.** 1992. A partial phylogenetic analysis of the "flavobacter-bacteroides" phylum: basis for taxonomic restructuring. Syst. Appl. Microbiol. **15:**513–521. http://dx.doi.org/10.1016/S0723-2020(11)80110-4.
5. **Paster BJ, Dewhirst FE, Olsen I, Fraser GJ.** 1994. Phylogeny of *Bacteroides*, *Prevotella*, and *Porphyromonas* spp. and related bacteria. J. Bacteriol. **176:**725–732.
6. **Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11:**119. http://dx.doi.org/10.1186/1471-2105-11-119.
7. **Laslett D, Canback B.** 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. **32:**11–16. http://dx.doi.org/10.1093/nar/gkh152.
8. **Nawrocki EP, Eddy SR.** 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics **29:**2933–2935. http://dx.doi.org/10.1093/bioinformatics/btt509.
9. **Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35:**3100–3108. http://dx.doi.org/10.1093/nar/gkm160.
10. **Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B.** 2009. The Carbohydrate-Active Enzymes Database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. **37:**D233–D238. http://dx.doi.org/10.1093/nar/gkn663.

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

Draft Genome Sequence of the *Bactrocera oleae* Symbiont "*Candidatus* Erwinia dacicola"

Frances Blow,[a] Anastasia Gioti,[b] David Starns,[a] Michael Ben-Yosef,[c] Zohar Pasternak,[d] Edouard Jurkevitch,[d] John Vontas,[b,e] Alistair C. Darby[a]

Institute of Integrative Biology, University of Liverpool, Liverpool, Merseyside, United Kingdom[a]; Institute of Molecular Biology & Biotechnology, Foundation for Research & Technology Hellas, Heraklion Crete, Greece[b]; Department of Entomology, The Hebrew University of Jerusalem, Rehovot, Israel[c]; Department of Plant Pathology and Microbiology, Robert H. Smith Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, Rehovot, Israel[d]; Department of Crop Science, Agricultural University of Athens, Athens, Greece[e]

"*Candidatus* Erwinia dacicola" is a *Gammaproteobacterium* that forms a symbiotic association with the agricultural pest *Bactrocera oleae*. Here, we present a 2.1-Mb draft hybrid genome assembly for "*Ca.* Erwinia dacicola" generated from single-cell and metagenomic data.

Received 20 July 2016 **Accepted** 26 July 2016 **Published** 15 September 2016

Citation Blow F, Gioti A, Starns D, Ben-Yosef M, Pasternak Z, Jurkevitch E, Vontas J, Darby AC. 2016. Draft genome sequence of the *Bactrocera oleae* symbiont "*Candidatus* Erwinia dacicola." Genome Announc 4(5):e00896-16. doi:10.1128/genomeA.00896-16.

Address correspondence to Alistair C. Darby, acdarby@liv.ac.uk.

The association between *Bactrocera oleae* (*Tephritidae*) and a bacterial symbiont was first discovered in 1909 (1). Several studies have since identified this organism as "*Candidatus* Erwinia dacicola" (*Enterobacteriaceae*) (2–4). "*Ca.* Erwinia dacicola" plays a role in nutrient provisioning, particularly during juvenile development in unripe olives, where it is essential for larval survival (5–8). However, due to a lack of genomic resources and the inability to culture "*Ca.* Erwinia dacicola," the metabolic basis of its association with *B. oleae* remains elusive. We present a draft of the "*Ca.* Erwinia dacicola" genome sequence that will inform future investigations into the functional and evolutionary foundations of the symbiosis.

Multiple single-cell (eight) and metagenomic (two) libraries were used to generate a draft hybrid genome assembly of "*Ca.* Erwinia dacicola." Single-cell libraries were prepared from the guts of adult female flies collected in Heraklion, Greece, stained with CellTracker deep red, and sorted on a Sony SH800. Genomic DNA was amplified using the REPLI-g kit (Qiagen) and was validated as "*Ca.* Erwinia dacicola" by amplification of the 16S rRNA gene, followed by digestion with the restriction enzyme PstI (9). Shotgun libraries were then prepared with the NEBNext Ultra DNA kit (New England Biosciences) and sequenced on an Illumina MiSeq sequencer at the Centre for Genomic Research, University of Liverpool, United Kingdom. Metagenomic shotgun and mate-pair (2- to 6-kb) libraries were prepared from gastric ceca dissected from third-instar larvae isolated in Israel from unripe olives and from a mixture of ripe and unripe olives, respectively. DNA for the shotgun library was extracted using an adapted cetyltrimethylammonium bromide (CTAB) method (10) with additional bead beating and lysozyme digestion, and the library was prepared with the Ovation rapid DR multiplex system (NuGen). The mate-pair library was prepared with the gel-free NexteraMate protocol from DNA extracted with the Chemagic DNA bacteria kit (Chemagen). Both libraries were prepared and sequenced on an Illumina MiSeq sequencer by LGC Genomics GmbH (Berlin, Germany).

Reads were assembled with SPAdes version 3.7.1 (11) in single-cell mode. Using Blobology (12), contigs identified as belonging to other organisms based on coverage and G+C content were excluded, and the 12,519,932 reads that mapped back to putative "*Ca.* Erwinia dacicola" contigs were extracted and reassembled with SPAdes. The result was a 2.1-Mb assembly comprising 333 scaffolds (>500 bp) at ~1,000× coverage, with an N_{50} of 9,998 bp. The assembly was assessed as in reference 13 and was found to be 92% complete in comparison to free-living bacteria and 100% complete in comparison to the endosymbiotic bacteria of aphids and tsetse flies, *Buchnera aphidicola* and *Wigglesworthia glossinidia*, respectively. Its G+C content (53.5%) is similar to that observed in other members of the *Erwinia* genus (14) and higher than that of other vertically transmitted endosymbiotic bacteria (15). The genome contains 2,407 protein-coding genes and 28 RNA-coding genes, based on annotation with PROKKA version 1.5.2 (16).

Accession number(s). This whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession no. MAZB00000000. The version described in this paper is version MAZB01000000 and BioProject no. PRJNA326914.

ACKNOWLEDGMENTS

We thank Anastasia Kampouraki for technical support and sample collection. The single-cell isolation and sequencing were conducted in the Single Cell Lab at the Centre for Genomic Research, University of Liverpool.

The funders had no role in the study design, data analysis, decision to publish, or preparation of the manuscript.

FUNDING INFORMATION

This work was funded by the Biotechnology and Biological Sciences Research Council (BBSRC), Oxitec Ltd. through an iCASE studentship (BB/K501773/1) awarded to A.C.D., and the ARISTEIA Action of the Opera-

tional Programme Education and Lifelong Learning and is cofunded by the European Social Fund (ESF) and National Resources (code number 4937) to J.V.

REFERENCES

1. **Petri L**. 1909. Ricerche sopra i batteri intestinali della mosca olearia. Memorie della Regia Stazione di Patologia Vegetale di Roma, Rome, Italy.
2. **Capuzzo C, Firrao G, Mazzon L, Squartini A, Girolami V**. 2005. "*Candidatus* Erwinia *dacicola*", a coevolved symbiotic bacterium of the olive fly *Bactrocera oleae* (Gmelin). Int J Syst Evol Microbiol **55:**1641–1647. http://dx.doi.org/10.1099/ijs.0.63653-0.
3. **Estes AM, Hearn DJ, Bronstein JL, Pierson EA**. 2009. The olive fly endosymbiont, "*Candidatus* Erwinia dacicola," switches from an intracellular existence to an extracellular existence during host insect development. Appl Environ Microbiol **75:**7097–7106. http://dx.doi.org/10.1128/AEM.00778-09.
4. **Sacchetti P, Granchietti A, Landini S, Viti C, Giovannetti L, Belcari A**. 2008. Relationships between the olive fly and bacteria. J Appl Entomol **132:**682–689. http://dx.doi.org/10.1111/j.1439-0418.2008.01334.x.
5. **Hagen KS**. 1966. Dependence of the olive fly, *Dacus oleae*, larvae on symbiosis with *Pseudomonas savastanoi* for the utilization of olive. Nature **209:**423–424. http://dx.doi.org/10.1038/209423a0.
6. **Ben-Yosef M, Aharon Y, Jurkevitch E, Yuval B**. 2010. Give us the tools and we will do the job: symbiotic bacteria affect olive fly fitness in a diet-dependent fashion. Proc Biol Sci **277:**1545–1552. http://dx.doi.org/10.1098/rspb.2009.2102.
7. **Ben-Yosef M, Pasternak Z, Jurkevitch E, Yuval B**. 2014. Symbiotic bacteria enable olive flies (*Bactrocera oleae*) to exploit intractable sources of nitrogen. J Evol Biol **27:**2695–2705. http://dx.doi.org/10.1111/jeb.12527.
8. **Ben-Yosef M, Pasternak Z, Jurkevitch E, Yuval B**. 2015. Symbiotic bacteria enable olive fly larvae to overcome host defences. R Soc Open Sci **2:**150170. http://dx.doi.org/10.1098/rsos.150170.
9. **Estes AM, Segura DF, Jessup A, Wornoayporn V, Pierson EA**. 2014. Effect of the symbiont *Candidatus* Erwinia dacicola on mating success of the olive fly *Bactrocera oleae* (Diptera: Tephritidae). Int J Trop Insect Sci **34:**S123–S131. http://dx.doi.org/10.1017/S1742758414000174.
10. **Xin Z, Chen J**. 2012. A high throughput DNA extraction method with high yield and quality. Plant Methods **8:**26. http://dx.doi.org/10.1186/1746-4811-8-26.
11. **Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelsky A, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R**. 2013. Assembling genomes and mini-metagenomes from highly chimeric reads. Res Comput Mol Biol **7821:**158–170.
12. **Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M**. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front Genet **4:**237. http://dx.doi.org/10.3389/fgene.2013.00237.
13. **Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM**. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature **499:**431–437. http://dx.doi.org/10.1038/nature12352.
14. **Starr MP, Chatterjee AK**. 1972. The genus *Erwinia*: enterobacteria pathogenic to plants and animals. Annu Rev Microbiol **26:**389–426. http://dx.doi.org/10.1146/annurev.mi.26.100172.002133.
15. **Moran NA**. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A **93:**2873–2878. http://dx.doi.org/10.1073/pnas.93.7.2873.
16. **Seemann T**. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics **30:**2068–2069. http://dx.doi.org/10.1093/bioinformatics/btu153.