Thesis submitted for the degree
Doctor of Philosophy

Functional Metagenomics and Evolution of Termite Gut Microbiome

by
Jigyasa Arora

Supervisor: Thomas Bourguignon
March 11, 2022

Declaration of Original Authorship

I, Jigyasa Arora, declare that the thesis entitled "Functional Metagenomics and Evolution of Termite Gut Microbiome" and the data presented in it are original and my own work.

I confirm that:

- This work was done solely while a candidate for the research degree at the Okinawa Institute of Science and Technology Graduate University, Japan.
- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly acknowledged. Quotations from the work of others have been clearly indicated and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.
- None of this work has been previously published elsewhere.

Name: Jigyasa Arora

Signature:  Date: 11th March 2022

Abstract

Termites are amongst the most abundant terrestrial animals on earth, primarily due to their ability to digest lignocellulose, the most abundant organic molecule. Lignocellulose is broken down in the termite gut with the help of symbiotic microbes, including protists, archaea, and bacteria. Studies using fragments of the 16S ribosomal RNA gene have shown that termites and their gut bacteria have had a complex coevolutionary history. In addition, the bacterial communities found in termite guts vary with termite diet. Up to now, studies have been focusing on termite species that are easy to sample or have a pest status. This sampling bias against early-evolving termite lineages and lineages feeding on substrates other than wood preclude a global understanding of the evolutionary history of termites and their gut microbes. To fill this gap, I sequenced whole gut metagenomes of 201 termite samples and one sample of a species of *Cryptocercus*, the cockroach genus sister to termites. The samples were selected from across the termite tree of life and represent termite phylogenetic and dietary diversity. My thesis showcases that (i) the gut microbiome of all termites possess similar genes for carbohydrate breakdown and other metabolic pathways involved in the digestion of carbohydrates. The proportion of these genes vary with termite phylogeny and diet. Still, the acquisition of a soil diet from a wood-feeding ancestor was accompanied by changes in gene abundance rather than by the acquisition of new genes and pathways. Using ten single-copy protein-coding marker gene sequences, (ii) I studied the pattern of coevolution between termites and their gut bacteria. Significant cophylogenetic signals with termites were found for tens of gut bacterial lineages that were either acquired by the common ancestor of all termites or by specific termite lineages. Finally, (iii) I investigated the role of horizontal gene transfer on the acquisition of carbohydrate metabolizing gene families by the termite gut microbes. I found gene family-specific transfers from the environment and from bacteria belonging to lineages present in the termite gut, suggesting that horizontal gene transfer events are common among bacteria of termite guts. Overall, my Ph.D. thesis sheds new light on how the gut microbiome has coevolved with its termite hosts since the inception of this nutritional symbiosis, some 150 million years ago.

Acknowledgments

Firstly, I would like to acknowledge the support of my thesis supervisor, Thomas Bourguignon, who guided me throughout the thesis process. Thomas patiently taught me different aspects of the paper writing process and instilled the importance of presenting one's work to the best of our abilities. I am also grateful to Yukihiro Kinjo, who taught me bioinformatics and mentored me to analyze and troubleshoot large sequencing data that I worked on during my thesis. This thesis could not have been possible without the support of Lucia Zifcakova, who continuously encouraged me to improve through our discussions on microbial genomics. I would also like to thank Ales Bucek, who examined the urate oxidase functions in termite transcriptomes in chapter one (Figure S1.4) and taught me how to examine the termite phylogenetic history in chapter two. I want to thank the lab members of the OIST Evolutionary Genomics Unit-especially Crystal-Leigh Clitheroe, Tracy Audisio, Menglin Wang, and Esra Kaymak-who each played an essential role during my thesis journey by teaching me lab protocols such as DNA isolation and library preparation, engaging in scientific discussions and giving feedback.

I would also like to thank Vincent Hervé for sharing the 16S rRNA data for chapter one, and Andreas Brune, Vincent Hervé, and Gaku Tokuda for giving constructive feedback on chapter one. I would also like to acknowledge the OIST sequencing section for performing the sequencing, especially Ryo Koyanagi, for providing sequencing advice. I would like to acknowledge further the OIST Scientific Computing and Data Analysis section for helping me troubleshoot problems in software installation and providing support for running jobs on OIST's High-Performance Computing Cluster.

Lastly, and most importantly, I am grateful to my family for always supporting my decisions and working together to make it a reality, finally, to Ankur Dhar, my partner in crime, for always having my back through the best and worst of times.

List of Abbreviations

AA- Auxiliary Activities
ack- acetate kinase
acs- acetyl-CoA synthase
ANI- Average Nucleotide Identity
apr- adenylylsulfate reductase
arcC- carbamate kinase
argF- ornithine carbamoyltransferase
argG- argininosuccinate synthase
argH- argininosuccinate lyase
CAZymes- carbohydrate-active enzymes
CBM- Carbohydrate Binding Modules
CDS- coding sequences
CE- Carbohydrate Esterases
dsr- dissimilatory sulfite reductase
fdhF- formate dehydrogenase H
FDR- False Discovery Rate
fhs- formate-tetrahydrofolate ligase
folD- methylenetetrahydrofolate dehydrogenase
GH- Glycoside Gydrolases
glnA- glutamine synthetase
gltBD- glutamate synthase
GT- Glycoside Transferases
hdr- heterodisulfide reductase
HGT- Horizontal Gene Transfers
HMM- Hidden Markov Model
MAGS- Metagenome Assembled Genomes
mcrABG- methyl-coenzyme M reductase
metF- 5,10-methylenetetrahydrofolate reductase
ML- Maximum likelihood
mtaABC- methanol:coenzyme M methyltransferase
ORF- Open Reading Frames
PACo- Procrustean Approach to Cophylogeny
PCA- Principal Component Analysis
pta- phosphotransacetylase
PUL- Polysaccharide Utilization Loci
RF- Robinson Foulds
rRNA- ribosomal RNA
sat- sulfate adenylyltransferase
TPM- Transcripts Per Million
tRNA- transfer RNA genes
WLP- Wood-Ljungdahl pathway
Xus- Xylan utilization system

Table of Content

List of figures

List of tables

Introduction section

Insects are the most diverse animal group on earth (Basset et al. 2012). They have adapted to most terrestrial ecosystems and have experienced exceptional ecological diversification since their crustacean ancestor colonized lands over 400 million years ago (Engel and Grimaldi 2004; Misof et al. 2014). Their ecological success is partly due to the many associations they have established with the microbes they host in their gut (Felton et al. 2018).

Gut bacteria have been shown to contribute to host immunity (Lemaitre and Hoffmann 2007), nutrition (Khan and Ahmad 2018), and development (Chouaia et al. 2012). Many insects acquire their gut symbionts from the environment through horizontal transfers. For example, the bean bug *Riptortus pedestris* acquires its symbiont, *Burkholderia*, from the soil every generation. However, *R. pedestris* is not dependent on *Burkholderia* to reach adulthood (Kikuchi et al. 2005; Kikuchi et al. 2007). Alternatively, some gut symbionts are heritable and vertically transmitted from parents to offspring. The most intricate symbiosis generally relies on vertical transfers, which allows the establishment of stable relationships over extended evolutionary time scales (Engel and Moran 2013; Vavre and Kremer 2014; Groussin et al. 2020). For example, plataspid stinkbugs inherit vertically their endosymbiont, a *γ-Proteobacteria*, via capsule transfer. The symbionts-rich capsule is deposited by the mother onto the egg mass and is orally ingested by the hatchling, helping with host development and contributing to reproductive fitness (Hosokawa et al. 2006). Multiple other mechanisms can also lead to the vertical transfer of gut bacteria. These mechanisms include, among others, the transfer of gut bacteria-rich tubules from the mother to the egg surface (Kikuchi et al. 2009), the transfer of midgut bacteria into the oocyte cytoplasm leading to trans-generational transfer during oogenesis (Kuechler et al. 2011), and the consumption of parent's excrement, a phenomenon referred to as proctophagy (Buchner 1965). In social insects, gut symbionts can be passed among individuals by social interactions. For example, the young honeybees *Apis mellifera* acquire their bacterial symbionts from the fecal fluid of adult workers after emergence (Martinson et al. 2012; Kwong and Moran 2016). Similarly, some species of ants, such as, the herbivorous ants of the Cephalotini tribe, practice proctodeal (anus-to-mouth) trophallaxis among nestmates (Anderson et al. 2012; Łukasik et al. 2017; Hu et al. 2018). Termites can acquire their gut microbes vertically through proctodeal trophallaxis (Nalepa 2017; Michaud et al. 2020) or horizontally from their diet (Mikaelyan et al. 2015a; Mikaelyan et al. 2017a). Many gut bacterial clades are specific to termites. These clades are shared among phylogenetically divergent termite lineages and are also present in *Cryptocercus*, the sub-social wood-feeding cockroach sister of termites (Dietrich et al. 2014; Bourguignon et al. 2018). These characteristics make termites a unique model to study the role of factors such as host phylogenetic distance and diet on gut microbial taxonomic composition and function. In addition, termites host specific microbial communities (Hongoh et al. 2005; Dietrich et al. 2014; Bourguignon et al. 2018) whose members might exchange some genes through horizontal gene transfers (Ottesen and Leadbetter 2011; Ikeda-Ohtsubo et al. 2016; Tokuda et al. 2018). This process might allow the acquisition of new gene families by various gut bacterial clade, and have an adaptative value for the termite host, shaping the evolution of the gut environment.

Termite as a model system to study the gut microbiome

Termites are the oldest social insect lineage (Legendre et al. 2015). They descend from a cockroach ancestor (Lo et al. 2000a) and diverged from their sister group, *Cryptocercus*, a genus

of subsocial wood-feeding cockroaches, some 170 million years ago (Ma) (Bourguignon et al. 2015). The most recent common ancestor of termites was estimated to have evolved 149 Ma (Bourguignon et al. 2015; Bucek et al. 2019). Traditionally, termites are classified into two major groups: the paraphyletic lower termites and the monophyletic higher termites. Lower termites comprise the basal lineages of termites, which are associated with flagellates living in their gut that assist their host in wood digestion (Brune 2014). Lower termites also host diverse communities of bacteria in their gut that also participate to wood digestion, complementing the gut flagellates (Ohkuma and Brune 2011; Brune 2014). Phylogenetic analyses of termites, using mitochondrial genomes and transcriptomes, retrieved Mastotermitidae as the first diverging family sister of all other termites, and Rhinotermitidae as the apical family of lower termites, within which the higher termites are nested (Bourguignon et al. 2015; Bucek et al. 2019). The higher termites consist of a single termite family, the Termitidae, which makes up about 70% of all described termite species (Bourguignon et al. 2017). The oldest Termitidae fossils are dated to early Eocene, 50 Ma (Engel et al. 2011), and modern Termitidae subfamilies diversified 35-23 Ma (Bourguignon et al. 2017). All Termitidae lost flagellate symbionts and established new associations to digest plant-derived organic matter (Brune and Dietrich 2015). The early diverging higher termite lineages developed fungiculture (subfamily Macrotermitinae) and bactericulture (subfamily Sphaerotermitinae) (Garnier-Sillam et al. 1989; Rouland-Lefèvre et al. 2006). These fungi and bacteria are not part of the gut microbiota but are cultivated within the nest, helping termites to digest plant matter. The other subfamilies of higher termites are not associated with fungal or bacterial gardens and depend solely on gut prokaryotes to process their food. They specialized to feed on new dietary resources, such as soil, grass, lichens, wood, and leaf litter. This dietary diversification contributed to the ecological and evolutionary success of Termitidae, which became one of the dominant organic matter decomposers in tropical and subtropical terrestrial ecosystems (Bourguignon et al. 2017).

The lignocellulose in the plant matter is made up of three major components, cellulose (38-50%), hemicellulose (23-32%), and lignin (15-25%) (Béguin 1990). The mastication of the plant matter by mandibles reduce wood into smaller fragments and initiates the plant matter digestion in termite guts (Muegge et al. 2011). These wood particles enter into the digestive tract where the cellulose is hydrolyzed by the host enzymes secreted by the salivary glands in lower termites or by the midgut epithelium in higher termites (Watanabe et al. 1998; Tokuda et al. 2004). Although digestion begins in the early sections of the gut, the major part of the digestion process occurs in the hindgut and is performed by gut microbes (Figure 1). The hindgut is also the location in the gut where nutrient absorption by the host takes place (Breznak and Brune 1994). The hindgut has been estimated to harbor ~10^8 viable cells per micro liter of hindgut fluid (Schultz and Breznak 1978). These microbial communities include more than 1,000 species-level operational taxonomic units (OTUs) of bacteria and archaea and, in all lower termites, an assemblage of flagellate protists (Hongoh 2011; Brune 2014; Dietrich et al. 2014). Consisting of multiple interacting partners, the termite gut microbiome is largely stable and is composed of many bacterial phylotypes that are shared by related termite lineages (Hongoh et al. 2005; Dietrich et al. 2014; Bourguignon et al. 2018). For example, the Spirochetes genus *Treponema* Ia forms a monophyletic termite-specific cluster (Bourguignon et al. 2018; Song et al. 2021) sister to sequences found in other environments, and the relative abundance of *Ruminococcaceae* and *Porphyromonadaceae* mirror the termite host tree (Dietrich et al. 2014; Abdul Rahman et al. 2015). These similarities among bacterial communities of related termite species are largely maintained by vertical transfers, which, in termites, are most likely

performed by proctodeal trophallaxis among nestmates, allowing inheritance of gut microbes across generations of the termite hosts (Nalepa 2017; Michaud et al. 2020). Despite the preponderance of vertical transfer of bacterial communities, termites also host bacteria acquired from their environment. For example, lab-reared colonies of *Hodotermes mossambicus* have a colony-specific gut microbial signature, possibly used for nestmate recognition (Minkley et al. 2006). Diet has also emerged as a major contributor to gut microbiome diversity. For example, the genera of Nasutitermitinae and Termitinae feeding on a cellulose-rich diet have a higher abundance of Spirochaetota and Fibrobacteriota phyla than the genera feeding on soil (Warnecke et al. 2007; Köhler et al. 2012; Mikaelyan et al. 2015a). In contrast, soil-feeding lineages host many Firmicutes (Thongaram et al. 2005; He et al. 2013), which have an OTU-level richness three to five times higher than wood-feeding lineages (He et al. 2013; Marynowska et al. 2020). The gut microbial communities of the fungus-cultivating Macrotermitinae, on the other hand, are dominated by Bacteroidetes and Firmicutes, two phyla that are generally dominant in the gut of omnivorous animals (Figure1; Dietrich et al. 2014; Bourguignon et al. 2018).

The successful symbiosis established with gut microbes has enabled termites to digest lignocellulose from different types of plant matter (Bignell and Eggleton 1995). The functional characterization of the digestion process by the gut microbiome has been performed using culture-based studies (Breznak and Switzer 1986; Leadbetter et al. 1998; Graber et al. 2004; Song et al. 2021), gene-expression analysis (Liu et al. 2018; Tokuda et al. 2018; Liu et al. 2019), and culture-independent sequencing such as metagenomics and metatranscriptomics surveys. These different methods showcase the functional potential of gut microbes in different aspects of lignocellulose digestion process. The gut microbes produce cellobiohydrolases, endoglucanases, β-glucosidases, and hemicellulases, a consortium of enzymes that act together to efficiently digest the plant matter (Warnecke et al. 2007; He et al. 2013; Poulsen et al. 2014; Marynowska et al. 2020). The cleavage of cellulose and hemicellulose into monomeric carbohydrates releases H_2, which is utilized by microbes performing various metabolic functions. These key functions include sulfate reduction, reductive acetogenesis, and methanogenesis, that generate short-chain fatty acids such as acetyl-CoA, acetate, and methane and provide energy to the host (Pester and Brune 2006; Brune and Ohkuma 2011; Brune 2014). The lignocellulose in the plant matter is poor in nitrogen and contains a low amount of amino acids and vitamins (Brune and Ohkuma 2011). Termite gut microbes compensate for the nitrogen deficiency by fixing atmospheric nitrogen, recycling nitrogen, and metabolically converting nitrogenous waste products into amino acids (Hongoh and Ohkuma 2010; Hongoh 2011; Ohkuma et al. 2015).

Functional analyses of the termite gut microbiome have been carried out for an increasingly large number of termite species (Warnecke et al. 2007; He et al. 2013; Liu et al. 2018; Tokuda et al. 2018; Hervé et al. 2020; Marynowska et al. 2020). But there has been a marked sampling bias towards wood-feeding termite species and species with pest status. The function and taxonomy of the gut microbial communities of termite lineages belonging to early diverging wood-feeding families and lineages with soil-feeding habits have not been carefully examined yet. There is insufficient data to determine the role of ecological factors such as diet and host evolutionary history on gut microbial functions. The mode of acquisition of gut microbes across different termite lineages and dietary habits, the duration of the coevolution with the host, and the frequency of horizontal transfers among the termite-specific microbial taxa remain largely unknown and should be examined in more details. This Ph.D. thesis aims to address these

questions through shotgun metagenome sequencing of the whole guts of 201 termite samples and one sample of the sister group of termites, the cockroach *Cryptocercus*. My thesis provides new data on the gut microbiome of previously understudied lower termite lineages, such as the families Stolotermitidae, Serritermitidae, and Rhinotermitidae, and higher termite sub-families, such as Termitinae, Foraminitermitinae, and Apicotermitinae. These newly sequenced termite samples are representative of the termite phylogenetic and dietary diversity that evolved since termites came to be some 150 million years ago. Our sequencing dataset provides an opportunity to generate a global synthesis of the dynamics of termite gut microbial communities at geological timescales. Our dataset also provides the opportunity to examine the role of ecological and evolutionary forces acting on termite gut microbial communities. In my chapter one, I will provide a global picture of gut microbial composition and their lignocellulose degradation ability. I annotated microbial genes from metagenome contigs and metagenome-assembled genomes (MAGs) to reconstruct the metabolic potential of microbial taxa present in the gut of termite species sampled across the termite tree of life. I analyzed the extent to which the termite phylogenetic history and dietary habits determine the relative abundance of the lignocellulose digesting genes and the microbes encoding these genes. The changes in the taxonomy and function of termite gut microbes during two significant events in termite history, the loss of flagellates in higher termites and the acquisition of a soil diet from wood-feeding ancestors, were also examined in detail.

In chapter two, to complement my efforts to characterize termite gut microbial composition across termite lineages and dietary habits, I investigated how long the gut microbes have been associated with the host. To do so, I extracted ten single-copy protein-coding marker genes from the 201 termite gut metagenome assemblies I generated. I use these marker gene sequences and homologous sequences obtained from non-termite-gut environment to produce individual marker gene phylogenetic trees. I found a series of termite-specific monophyletic clusters nested within lineages composed of sequences derived from environments other than termite guts. These phylogenetic clusters were examined for coevolution with the termite host. The phylogenies of these termite-specific microbial clusters were reviewed to determine whether they were acquired by the common ancestor of all modern termites, by termites sharing a specific diet, or by specific termite lineages. These analyses were performed on ten protein-coding marker genes occuring as a single copy in bacterial and archaeal genomes, providing a better phylogenetic resolution than the 16S ribosomal RNA gene (Sunagawa et al. 2013; Lan et al. 2016).

In chapter three, I inspected horizontal transfers among bacteria of gene families involved in cellulose, hemicellulose and chitin digestion using MAGs. MAG annotation was based on concatenated single-copy protein sequences (Chen et al. 2020) known not to be subject to horizontal gene transfers. Specifically, seven glycosyl hydrolase gene families and four sub-families that were shown in chapter one to significantly coevolve with the termite host were chosen for phylogenetic inferences. The phylogenetic trees of these gene families were compared with the phylogenetic tree of MAGs in order to identify putative horizotal transfers. These analyses sheds light on the evolutionary history of microbial gene families present in the guts of termites.

Finally, in the concluding chapter, I summarize the findings of this thesis and discuss its contribution to our understanding of the evolution of gut microbial composition and function during the 150 million years of termite evolution.

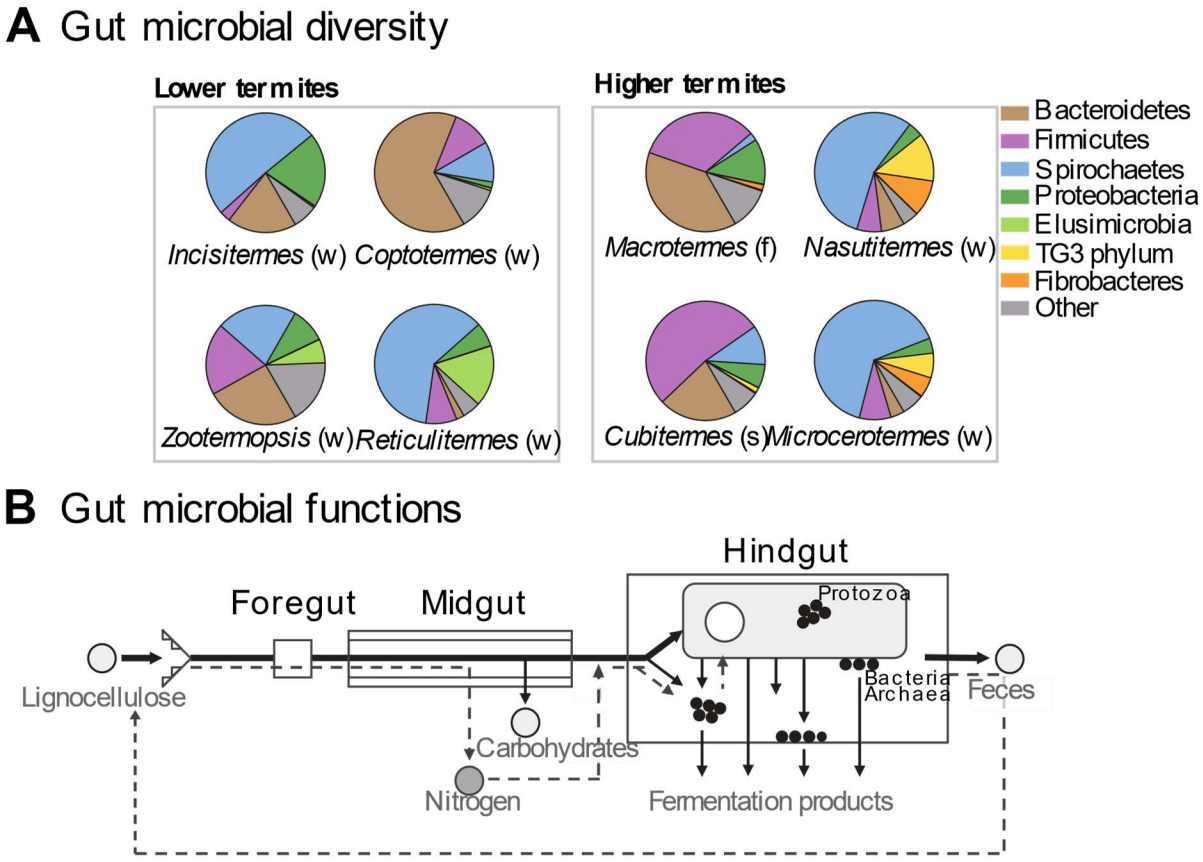

Figure A. 1- Termite gut microbiota composition and functions. (a) Phylum-level distribution of gut microbes in termite species representing major host groups (w, wood-feeding; f, fungus-cultivating; s, soil-feeding; modified from Brune, 2015). (b) Schematic of symbiotic digestion in termites. The *bold* lines represent the path of lignocellulose digestion. The *thinner* lines show the formation of soluble degradation products reabsorbed by the host and the *dashed* lines indicate nitrogen recycling by termites (modified from Brune and Ohkuma, 2011).

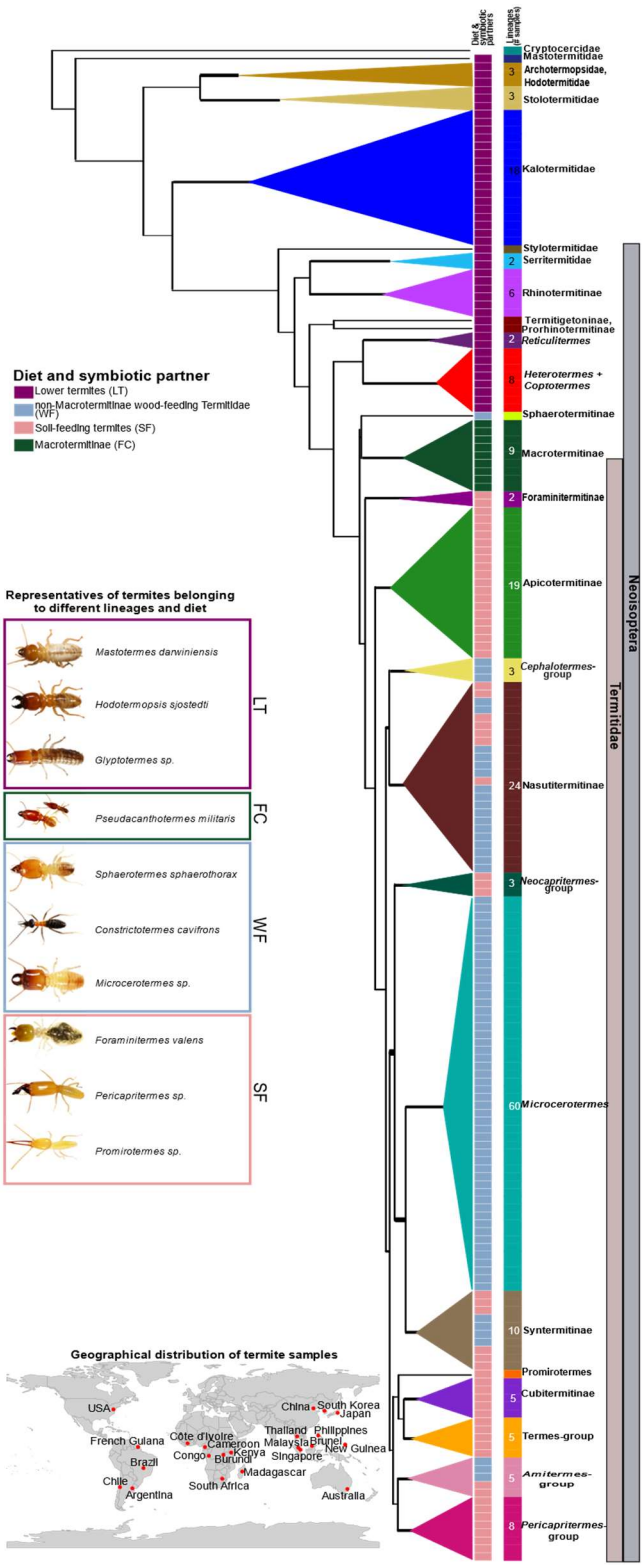

Figure A. 2- Phylogenetic and ecological diversity of termite samples examined in this study. Geographical distribution of collected samples and images of termite soldiers' representative of termite dietary groups are shown on the left (Picture credit- Ales Bucek). Number of samples collected per termite lineage is showcased on top of the lineage bar.

Chapter one- The functional evolution of termite gut microbes

Introduction

The gut of most animals is colonized by diverse communities of microbes. The human gut microbiome has received considerable attention because of its link to diseases and aging (Sommer and Bäckhed 2013; Kundu et al. 2017). As is the case for humans, the gut microbiome of many animals has positive effects on host fitness (Moran et al. 2019) , and in some extreme cases, like that for termites, gut microbes have a dominant role in host nutrition (Brune 2014; Brune and Dietrich 2015).

Termites are one of the few animal lineages feeding on substrates distributed along the wood-soil decomposition gradient (Donovan et al. 2001; Bourguignon et al. 2011). Although termites produce their own endogenous cellulases (Watanabe et al. 1998; Tokuda et al. 2004), their ability to decompose wood or soil organic matter largely depends on symbiosis with mutualistic gut microbes (Watanabe and Tokuda 2010; Brune and Ohkuma 2011), including bacteria, archaea and, in the case of lower termites, cellulolytic flagellates. The cellulolytic flagellates of termites are typically found nowhere else other than in termite guts and are efficiently transmitted across host generations (Nalepa 2017; Michaud et al. 2020). Similarly, many of the prokaryotes present in termite guts are found nowhere else in nature (Bourguignon et al. 2018; Hervé et al. 2020). Their vertical mode of inheritance is supported by the observations that differences among termite gut prokaryotic and protist communities tend to increase as phylogenetic distances among termite hosts increase (Abdul Rahman et al. 2015; Tai et al. 2015). In addition, the diet of the termite host, which largely correlates with the termite phylogeny (Bourguignon et al. 2011), also shapes the termite gut microbial communities (Dietrich et al. 2014; Mikaelyan et al. 2015a). This phylosymbiotic pattern observed between gut microbial communities and their hosts is not unique to termites, and is shared with many other groups of animals (Brooks et al. 2016; Lim and Bordenstein 2020). Whether the termite phylogeny is recapitulated by gut microbial functions, as it is recapitulated by the taxonomic composition of microbial communities, remains unknown.

Investigations of termite gut microbe genomes has revealed that, in addition to the production of enzymes involved in lignocellulose digestion, gut microbes have numerous nutritional functions, including nitrogen fixation and recycling abilities that supplement the nitrogen-poor diet of their host (Lilburn et al. 2001; Yamada et al. 2007; Hongoh et al. 2008; Ohkuma and Brune 2011). While metagenomics and metatranscriptomics surveys of termite guts have been carried out for an increasingly large number of termite species (Warnecke et al. 2007; He et al. 2013; Liu et al. 2018; Tokuda et al. 2018; Marynowska et al. 2020), often with the prospect of harvesting cellulolytic enzymes able to convert plant biomass into biofuel (*e.g.* Tartar *et al.*, 2009; Calusinska *et al.*, 2020), there has been a marked sampling bias towards easy-to-sample wood-feeding termite species, and species with pest status. Far less is known about the function and taxonomy of the gut prokaryotic communities of other termite lineages, such as basal wood-feeding lineages, or lineages with soil-feeding habits (Hervé et al. 2020). Because of this gap in our knowledge, it remains largely unclear how the taxonomy and function of gut microbiome has been evolving since termites and *Cryptocercus* diverged >150 Million years ago (Bourguignon et al. 2015; Bucek et al. 2019). Similarly, how the acquisition of a diet based on soil has affected the taxonomy and function of gut microbial communities remains an open question. A metagenomics survey based on a comprehensive sampling of termites is required to answer these questions.

In this study, we sequenced whole gut metagenomes of 146 species representatives of the phylogenetic and ecological diversity of termites, including many lineages that have remained undocumented. We used the assembled prokaryotic contigs of this dataset, the largest of its kind, to determine (1) when important gut prokaryotic pathways involved in nutritional functions were acquired by termites and *Cryptocercus*; (2) to which extent termite phylogeny is predictive of gut prokaryote taxonomic and functional composition; and (3) the taxonomic and functional changes experienced by gut prokaryote communities following the acquisition of a diet of soil.

Material and methods

Sample collection

We collected a total 145 termite samples and one sample of the cockroach *Cryptocercus kyebangensis* (Table S1.1, Figure S1.1). These samples were representative of the global termite diversity. All samples were preserved in RNA-later®, shipped to Okinawa, and stored at -80°C until DNA extraction.

DNA extraction and sequencing

Genomic DNA extraction was performed on the whole guts of five workers using the NucleoSpin Soil kit (Macherey-Nagel) according to manufacturer's protocol. Library preparation was performed using the KAPA Hyperplus Kit, which is based on a unique dual tag indexing approach that minimizes the effects of index hopping. Libraries were either PE250-sequenced on the Illumina HiSeq2500 platform or PE150-sequenced on the Illumina HiSeq4000 platform (Table S1.1).

Data filtering and assembly of metagenomic reads

Raw reads were filtered based on their quality. Reads with average Phred quality score below 30 were removed using Trimmomatic v 0.33 (Bolger et al. 2014). The "SLIDINGWINDOW" was set to "4:30" to trim low quality bases (Phred quality score below 30) from the 3' end of the reads. We removed the 16 base pairs at the 5' end of each read using the "HEADCROP" option because we observed over-represented k-mers in this region of the reads. Reads shorter than 50bps were removed.

The quality-controlled reads were assembled into contigs using SPAdes v 3.11.1 (Nurk et al. 2017) with the "meta" option and k-mer sizes of 21, 31, 41, 51, 71. The assembly quality was checked using the "metaquast" option of QUAST v 3.1 (Quality Assessment for Genome Assemblies) based on weighted median contig size (N50) (Gurevich et al. 2013) and percent of reads mapped to the contigs (Langmead and Salzberg 2012; Papudeshi et al. 2017). Only the reads mapped to prokaryotic contigs were examined in this study (see the *taxonomic annotation* and *functional annotation* sections below).

Termite phylogenetic tree reconstruction

We build a phylogenetic tree of termites using mitochondrial genomes retrieved from metagenome assemblies. Mitochondrial contigs derived from termites were identified using BLAST search (sequence length >5000 and percent identity >90) (Altschul et al. 1990) against previously published whole mitochondrial genomes of termites (Bourguignon et al. 2015; Bourguignon et al. 2016; Bourguignon et al. 2017; Wang et al. 2019). Mitochondrial genomes were complete, or near-complete, in most cases. Each contig derived from mitochondrial genomes was annotated using the MITOS webserver (Bernt et al. 2013). The 13 protein-coding genes, two rRNA genes, and 22 tRNA genes were aligned with MAFFT v 7.305 (Katoh et al. 2002) using default settings. The alignments were concatenated and the third codon position of protein-coding genes was removed. The dataset was partitioned into four subsets: one for the

first codon position of protein-coding genes, one for the second codon position of protein-coding genes, one for the two rRNA genes, and one for the 22 tRNA genes. A Bayesian phylogenetic tree was generated using BEAST v 2.4.8 (Suchard et al. 2018). We used an uncorrelated relaxed lognormal clock model (Drummond et al. 2006), and a Birth Death speciation process as tree prior (Gernhard 2008). The molecular clock was calibrated using nine fossil calibrations used by Bucek *et al.*, (2019) (Table S1.19). The fossil calibrations were implemented as exponential priors on node times. Because transcriptome-based phylogenies unambiguously support the monophyly of Sphaerotermitinae and Macrotermitinae (Bucek et al. 2019) (unlike mitochondrial genome-based phylogenies; Bourguignon *et al.*, 2017), we constrained Sphaerotermitinae + Macrotermitinae to be monophyletic. Similarly, we constrained non-Stylotermitidae Neoisoptera to form a monophyletic group. The MCMC chain was sampled every 1000 steps over a total of 0.4 billion generations. The convergence of the chain was assessed using Tracer v 1.7.1 (Rambaut et al. 2018), and the initial 10 percent was discarded. We carried out two replicate MCMC runs to ensure convergence of the chain.

Reconstruction of Metagenome Assembled Genomes

We reconstructed Metagenome Assembled Genomes (MAGs) from metagenomes contigs using CONCOCT v 0.4.0 (Alneberg et al. 2014) implemented in the metawrap software v 0.9 (Uritskiy et al. 2018) with default parameters. MAG quality checking, based on 43 single-copy marker genes (Table S1.9), was performed with CheckM v 1.0.11 (Parks et al. 2015). High-quality MAGs, medium-quality MAGs, and low-quality MAGs with upward of 30% completeness and downward of 10% contamination were retained (Table S1.9) (Bowers et al. 2017). We retained low-quality MAGs that were at least 30% complete because, in some cases, they were endowed with complete pathways. Despite having fewer single-copy marker genes, 65.35% of these MAGs possessed more than 10 transfer RNA genes (tRNA) and 17.54% had at least one of the three ribosomal RNA genes (rRNA). All MAGs that did not meet these criteria were discarded. In addition, we discarded MAGs with obvious mismatches among marker genes. To identify these MAGs, we built Maximum Likelihood phylogenetic trees for all 43 single-copy marker genes with FastTree v 2.1.11 (Price et al. 2010). MAGs that fall in different phyla for different marker genes were considered as having obvious mismatches and were discarded. The rRNA genes were extracted using METAXA2 software (Bengtsson-Palme et al. 2015), tRNA genes were predicted via tRNAscan-SE tool (Chan and Lowe 2019), and MAG coverage was calculated with the "metawrap quant_bins" command of the metawrap software (Uritskiy et al. 2018).

Taxonomic annotation

The annotation of genomic features of bacterial and archaeal contigs and MAGs was carried out with Prokka v 1.14 (Seemann 2014). This step allowed the identification of coding sequences (CDS), ribosomal RNAs (rRNA), and transfer RNAs (tRNA), which were used in downstream analyses. To identity the taxonomy of the metagenome contigs, we taxonomically annotated single-copy marker genes and other protein-coding genes in contigs longer than 1000bps. 40 single-copy marker genes were extracted using mOTU software ver1 (Sunagawa et al. 2013; Wu et al. 2013). Single-copy marker genes were taxonomically annotated using DIAMOND BLASTp (Buchfink et al. 2015) with e-value $\leq$ 1e-24 and output format 102, which uses the lowest common ancestor algorithm for annotation. Other protein-coding genes were annotated using the same settings as marker genes but with DIAMOND BLASTx algorithm. Both annotations were performed using the GTDB ver 95 database as reference (Parks et al. 2020).

Taxonomic annotation of MAGs was based on bacterial and archaeal reference trees using GTDB-Tk v1.3.0 based on GTDB ver 95 (Chaumeil et al. 2019).

We used the genomic DNA extracted from whole termite guts to produce 16S rRNA gene PCR amplicon sequences. PCR reactions were carried out using the primer pairs 515F (XXXXXGTGTGYCAGCMGCCGCGGTAA, Parada *et al.*, 2016) and 806R (XXXXXXXXCCGGACTACNVGGGTWTCTAAT, (Apprill et al. 2015)). All pairs of primers were endowed with unique dual tag indexes (8X overhang on the forward primer and 5X overhang on the reverse primer) to minimize the effects of index hopping between libraries. We conducted PCR amplifications using Takara Tks Gflex DNA Polymerase with the following conditions: initial denaturation (3 min at 94°C), 30 cycles of amplification (45 s at 94°C, 60 s at 50°C, and 90 s at 72°C), and a terminal extension (10 min at 72°C). All PCR reactions were scaled down using one half of the reagents recommended in the manufacturers protocol. Prepared libraries were mixed in equimolar concentration and paired-end-sequenced on the Illumina MiSeq platform. The analysis of the 16S rRNA gene amplicon sequences was performed with mothur v1.44.1 (Schloss et al. 2009) following the standard procedure for Illumina data analysis described by Kozich et al. (2013). After removing low-quality reads and chimera, sequences were clustered into operational taxonomic units (OTUs) at a sequence similarity level of 97% using VSEARCH (Rognes et al. 2016). Sequences were classified using the naïve Bayesian classifier (Wang et al. 2007) implemented in mothur and the SILVA reference database release 138 (Quast et al. 2013). The abundance of every family inferred from both 16S rRNA gene amplicon data and metagenomic data was then compared. In total 143 prokaryote lineages received identical family-level annotation in both datasets.

Functional annotation

We carried out functional annotation of the CDSs identified with Prokka v.1.14.5 (Seemann 2014) for all contigs and MAGs that were taxonomically annotated as bacteria or archaea using the "metagenome" option. We used the CAZy database (Lombard et al. 2014) as a reference to identify CDSs with carbohydrate metabolizing properties. Protein sequences were searched against a set of profile Hidden Markov models (HMM) representing CAZy domains deposited in the dbCAN database release 7 (Yin et al. 2012). We used an e-value lower than e-30 and coverage greater than 0.35 as thresholds to extract best domain matches.

Hydrogenases were annotated by means of HMM searches against the Pfam database version 32.0 (El-Gebali et al. 2019) using an e-value cut-off of e-30. Catalytic subunits of hydrogenases were classified into different classes using the k-nearest neighbor algorithm implemented in the HydDB webtool (Søndergaard et al. 2016). For the [FeFe] hydrogenase Group A4, we carried out manual inspection of the conserved motifs in the protein sequence (Schuchmann et al. 2018).

We reconstructed prokaryotic metabolic pathways from our metagenomes with KOFam scan v.1.1.0 (Kanehisa et al. 2016; Graham et al. 2018). We used the KEGG database as a reference and e-value cut-off of e-30. Each protein sequence was annotated to gene family level with the KEGG-Decode python module (Graham et al. 2018). The MAG metabolic pathways were annotated with KOFam scan v.1.1.0 using default settings. As some MAG gene families appeared to be absent after annotation against the KEGG database, to confirm, or reject, the absence of these gene families, we carried out BLAST searches (Amino acid identity >60% and alignment length > 100 amino acids) against the Annotree protein sequence database (Mendler et al. 2019).

Relative abundance of gene families

The relative abundance of CDSs was calculated by mapping the raw reads on the sequences. Briefly, the reads were mapped to the assembled contigs annotated as bacteria or archaea. Relative abundance was calculated for each CDS using Salmon v.1.4.0 with the "meta" option. Salmon corrects for GC-content bias, gene-length differences, and sampling effort (Srivastava et al. 2020). Relative abundance of CDSs obtained as Transcripts per Million (TPM) values were retained for downstream analysis if they were embedded into contigs longer than 1000 bps and had more than 1 TPM value. Individual TPM counts were normalized using centered log(2)-ratio (clr) transformation to account for the compositional structure and unequal numbers of reads in our metagenome data. Clr transformation enhances sub-compositional comparisons (gene vs gene, bacteria vs bacteria) and reduces spurious correlations. Positive and negative TPM values indicate positive and negative departure from the overall compositional mean, which is zero (Gloor et al. 2017). Clr transformation of marker genes and functional genes was performed using the R package *propr* using 0.65 as a pseudo count to account for zero values (Quinn et al. 2017). We did not calculate TPM for MAGs, but instead used presence/absence to investigate pathway completeness.

Statistical Analysis

We investigated whether the abundance of the genes and pathways of interest was phylogenetically autocorrelated to the time-calibrated tree of termites. To do so, we calculated the Moran's I phylogenetic autocorrelation index using the R package *phylosignal* (Keck et al. 2016) on CDSs embedded in contigs longer than 1000 bps and with TPM value higher than 1. This analysis was carried out for each bacterial and archaeal phylum present in at least 5% of the metagenomes, using the combined 40 single-copy marker genes (see Table S1.3). A 5% false discovery rate (FDR) correction was calculated using the p.adjust function implemented in the R package *stats* (R Core Team, 2014). Similarly, we calculated the Moran's I phylogenetic autocorrelation index for each 211 CAZymes present in more than 10% of gut metagenomes and carried out a 5% false discovery rate FDR correction. Finally, the analysis was performed for each gene involved in the reductive acetogenesis, sulfate reducing, nitrogen recycling and nitrogen fixating pathways, and the *mcrABG* gene of the methanogenesis pathway combined. We applied a 5% FDR correction.

To examine whether the abundance of the genes and pathways of interest differed with termite diet and presence of non-prokaryotic co-symbionts, we performed phylogenetic ANOVA using the procD.pgls function implemented in the R package *geomorph* (Adams and Otárola-Castillo 2013). A 5% FDR correction was calculated using the p.adjust function implemented in the R package *stats* (R Core Team, 2014). Termite diet was determined based on literature data (Donovan et al. 2001; Bourguignon et al. 2011), and was considered as made of wood or soil. Wood-feeding termite species included feeding groups I and II (including grass and leaf litter), while soil-feeding termites included feeding groups III and IV (*sensu* Donovan et al. 2001). Non-prokaryotic co-symbionts are found in two groups of wood-feeding termites: the lower termites, which include all termites with the exclusion of Termitidae and host cellulolytic flagellates in their gut, and the Macrotermitinae, a subfamily of Termitidae that cultivates cellulolytic *Termitomyces* in fungal combs. Therefore, we recognized four groups of termites: the lower termites (LT), the soil-feeding termites (all Termitidae, SF), the Macrotermitinae (FC), and the non-Macrotermitinae wood-feeding Termitidae (WF). Similar analysis was performed on prokaryotic lineages encoding CAZyme families present in more than 10% of termite gut

metagenomes in contigs longer than 5000 bps, to ensure correct taxonomic annotation. All metagenome contigs longer than 5000 bps with dinitrogen-fixing genes were also examined.
We visualized termite samples according to the abundance of CAZyme families present in their gut metagenomes using Principal Component Analysis (PCA). The PCA was performed using the prcomp function implemented in the R package *stats* (R Core Team 2014) and visualized using the R package *ggbiplot* (Vu 2011). Similar analyses were performed on the genes involved in reductive acetogenesis, sulfate reduction, dissimilatory nitrate reduction, urea degradation, glutamate biosynthesis, arginine biosynthesis, ammonia transport, nitrogen fixation, and *mcrABG* genes of the methanogenesis pathway.

Uricase genes encoded by termites

We searched the 53 termite transcriptomes previously published by Buček et al. (2019) for the presence of uricases. These transcriptomes were either derived from whole worker bodies or from worker heads, and included species of all termite families. Protein sequences of predicted uricases from termites (XP_023702357, GFG34960), cockroaches (PSN45555, CDO39394), fireflies (KAF529609, XP_031344605), sawflies (XP_015591878, XP_015521616), ant (XP_011159093), fruit fly (NP_476779), and rat (NP_446220) were used as a query in TBLASTn searches. The longest open reading frames for all significant TBLASTn search hits (E-value < 10^{-30}) were identified and translated using hmmer2go obtained from https://github.com/sestaton/HMMER2GO. The nonsense proteins that did not provide any significant BLASTx hit against NCBI RefSeq database (E-value < 10^{-10}) were discarded. The remaining predicted protein sequences, derived from 23 transcripts, were assigned KEGG annotations using eggNOG-Mapper version 4.5 (Huerta-Cepas et al. 2019). The protein sequences were aligned using CLUSTAL W (Larkin et al. 2007) and the alignment was visually inspected.

Results

The taxonomic composition of termite gut prokaryotes

We sequenced whole gut metagenomes, including the hindgut containing the bulk of the gut microbiota, of 145 termite species and one cockroach species (Table S1.1, Figure S1.1). This included species from the nine termite families, species from the eight subfamilies of Termitidae, and one species of the wood-feeding cockroach *Cryptocercus*, the sister group of termites (Lo et al. 2000b). Our shotgun sequencing approach generated an average of 72.5 million reads per sample that were assembled into an average of 92,237 scaffolds >1000 bps, constituting 63.3% of mapped reads. The proportions of prokaryotic reads were on average 18.4% in lower termites and 20.5% in higher termites.

We used 40 marker genes (Sunagawa et al. 2013; Wu et al. 2013) to determine the taxonomy and estimate the abundance of each major bacterial lineage present in the 129 termite gut metagenome assemblies including upward of 10,000 contigs longer than 1000 bps. Shorter contigs were removed from the analyses. The bacterial community composition and abundance inferred from marker gene data showed similarities at the phylum level to that inferred from 16S rRNA gene amplicon sequences (Figure S1.2). However, the abundance distribution estimated by both approaches showed some disagreements for several families (Dietrich et al. 2014; Mikaelyan et al. 2015a; Bourguignon et al. 2018). Notably, *Dysgomonadaceae*, *Ruminococcaceae*, *Synergistaceae*, and *Oscillospiraceae* occurred at low abundances among the marker genes but were represented by many 16S rRNA gene sequences in most termite species (Dietrich et al. 2014; Mikaelyan et al. 2015a; Bourguignon et al. 2018) (Table S1.2).

These discrepancies are likely the result of variation in 16S rRNA gene copy number (Větrovský and Baldrian 2013; Edgar 2018), which are higher in these lineages, or possibly because of artifacts generated during 16S rRNA gene amplicon PCR cycles. They might also reflect the incomplete coverage of our metagenomes or, to a certain extent, the differences in the databases used for classification.

In total, we identified 114 family-level bacterial lineages, belonging to 19 phyla and represented in the gut of more than 5% of termite species (Table S1.3). Many other bacterial family-level lineages were recorded from the gut of no more than a few termite species, and were possibly transient, and not strictly associated with termite guts. We calculated the Moran I index on the abundance of these 114 family-level bacterial lineages to test whether bacterial abundance is correlated with termite phylogeny. We found a phylogenetic autocorrelation signal for 59 of the 114 bacterial lineages, and this signal remained significant at a 5% false discovery rate (FDR) correction for 27 bacterial lineages, including some of the most abundant bacterial lineages (Figure 1.1, Table S1.4). For example, the wood-fiber-associated *Fibrobacteraceae* (Mikaelyan et al. 2014; Tokuda et al. 2018) are dominant in the gut of *Microcerotermes*, Nasutitermitinae, and related termite lineages, and are either undetectable or occur at low abundance in the assemblies of other termite lineages. Another example is the *Endomicrobiaceae* that comprise flagellate-associated (Stingl *et al.*, 2005; Zheng *et al.*, 2015) and free-living Endomicrobia (Ikeda-Ohtsubo et al. 2016; Mikaelyan et al. 2017b), which were abundant in lower termites and almost entirely absent in higher termites.

Our dense taxonomic sampling of diverse termite hosts also allowed us to identify bacterial lineages whose association with termites has remained largely unreported. For example, we found that the *Holophagaceae*, a bacterial family of Acidobacteriota previously reported from the gut of three humus-feeding termite species (Mikaelyan et al. 2015a) and two species of Nasutitermitinae (Dietrich et al. 2014), is widely distributed in Nasutitermitinae, Foraminitermitinae, the *Cephalotermes*-group, and the *Pericapritermes*-group (Figure 1.1). Altogether, our results demonstrate that termite phylogeny is remarkably predictive of the gut bacterial community composition, and therefore that a strong phylosymbiotic signal is present for termite gut bacteria, as has been demonstrated for termite gut protists (Tai et al. 2015) and for several other groups of insects (Lim and Bordenstein 2020).

Using the same 40 marker genes and 129 metagenome assemblies used for bacteria, we investigated the diversity of gut-associated Archaea across the termite phylogenetic tree. In total, we identified 16 family-level archaeal lineages, including *Methanoculleaceae* and *Methanocorpusculaceae* (order *Methanomicrobiales*), *Methanosarcinaceae* (order *Methanosarcinales*), *Methanobacteriaceae* (order *Methanobacteriales*), *Methanomethylophilaceae* (order *Methanomassiliicoccales*), and *UBA233* (class *Bathyarchaeia*). All but nine family-level lineages were present in the gut of more than 5% of termite species. The abundance of *Methanosarcinaceae*, *UBA233*, and an unclassified family-level lineage of *Bathyarchaeia* showed significant autocorrelation signals with the termite phylogenetic tree when no FDR correction was applied (Figure 1.1, Table S1.4). *Bathyarchaeia* occurred in the clade of Termitidae excluding Macrotermitinae, Sphaerotermitinae, and Foraminitermitinae confirming previous reports (Loh et al. 2021), and *Methanosarcinaceae* was found in Macrotermitinae, Nasutitermitinae, and in Cubitermitinae and related termite lineages (Figure 1.1). Archaea represented in average less than 1% of the gut prokaryotes in wood-feeding termite species, while their proportion reached 4.6% in Macrotermitinae and 10.6% in soil-feeding termite species, and was exceptionally high in the soil-feeding

Mimeutermes in which 59.8% of the marker genes were assigned to *Bathyarchaeia*. Our results are in line with the higher archaeal-to-bacterial ratios reported in soil-feeding termites as compared to their wood-feeding counterpart, reflecting the higher methane emission rates of soil-feeding termites (Brune 2018; Brune 2019).

The carbohydrate-active enzymes of termite gut prokaryotes

We investigated the evolution of prokaryotic carbohydrate-active enzymes (hereafter: CAZymes) using the same 129 gut metagenome assemblies used to investigate gut prokaryotic composition. The *de novo* assemblies of these 129 gut metagenomes contained an average of 127,159 prokaryotic open reading frames (ORF). We identified ORFs coding for CAZymes using Hidden Markov model searches against the dbCAN2 database (Zhang et al. 2018). As a first step, we investigated the evolution of enzymes derived from prokaryotes with no consideration of their taxonomic origin. In total, we found 346 CAZyme categories in 129 gut metagenomes that consisted of 205 glycoside hydrolases (GHs), 57 glycoside transferases (GTs), 18 enzymes with carbohydrate-binding modules (CBMs), 16 carbohydrate esterases (CEs), 41 polysaccharide lyases (PLs), and 9 redox enzymes with auxiliary activities (AAs) (Table S1.5). We did not find any CAZymes in only one gut metagenome (that of *Araujotermes parvellus*, at e-value cut-off below e-30). For the other 128 gut metagenomes, the number of CAZyme categories varied between 5 and 139 per gut metagenome. Five GH families, GH2, GH3, GH10, GH31, and GH77, were found in more than 85% of the termite species. 14 GHs, seven of which had putative lignocellulolytic activity, were found in 75 to 85% of the termite species. Therefore, glycoside hydrolases previously found to be abundant in the gut of particular termite species (*e.g.* Warnecke *et al.*, 2007, Calusinska *et al.*, 2020) are universally part of the gut enzymatic repertoire of termites and *Cryptocercus*.

We calculated the Moran I index on the abundance of 211 CAZymes, including 146 CAZyme families and 65 sub-families, present in more than 10% of termite species, and found an autocorrelation signal with the termite phylogenetic tree for 107 CAZymes. The autocorrelation signal remained significant after FDR correction for 77 CAZymes (Figure 1.2, Table S1.6). Therefore, as for gut prokaryotic composition, which present a phylosymbiotic signal, termite phylogeny is predictive of the CAZyme repertoire present in termite guts.

Two factors that potentially affect the prokaryotic CAZyme repertoire of termite gut prokaryotes are diet and co-occurring non-prokaryotic cellulolytic symbiotic partners. We distinguished four termite groups: soil-feeding Termitidae (SF) and wood-feeding Termitidae excluding Macrotermitinae (WF), which host no other symbionts than gut prokaryotes (Brune, 2014), the fungus-cultivating Macrotermitinae (FC), which feed on wood or plant litter and cultivate cellulolytic fungi of the genus *Termitomyces* (Rouland-Lefèvre et al. 2006), and lower termites (LT), which feed on wood and host cellulolytic flagellates in their gut (Inoue et al. 2000). Overall, the abundance of prokaryotic CAZymes was the highest in WF and the lowest in SF, while LT and FC fell between these two extremes (Table S1.7). This is consistent with the scarcity of lignocellulose in the diet of SF, which predominantly feed on the nitrogen-rich fraction of the soil, including microbial biomass and organic residues associated with clay particles (Ji and Brune 2001; Ji and Brune 2005; Ngugi et al. 2011; Ngugi and Brune 2012). The intermediate abundance of prokaryotic CAZymes in FC and LT reflects their dependence on *Termitomyces* fungi for lignocellulose digestion (Poulsen et al. 2014) and on gut flagellates that encode for diverse cellulolytic enzymes (Yamin 1981; Nishimura et al. 2020), respectively. Task partitioning between gut prokaryotes and other symbionts –in which both partners participate in different steps of wood digestion and provide different sets of CAZymes– could

be revealed from the gut metagenomes of LT and FC. Principal component analysis revealed that the prokaryotic CAZyme repertoire differs considerably among SF, LT, FC, and WF (Figure 1.3A). To characterize more accurately the contribution of termite gut prokaryotes to wood digestion, whenever possible, we identified the substrate of each 211 CAZymes (including 146 families and 65 subfamilies) present in more than 10% of termite species. We individually compared the abundance of these 211 CAZymes using phylogenetic ANOVA. We found that 178 comparisons were significantly different, and 177 comparisons remained significant after FDR corrections (Figure 1.3A, Table S1.7). Notably, we found that the combined seven GHs exclusively identified as cellulases were significantly depleted in LT as compared to other termite groups and were significantly depleted in FC and SF as compared to WF (Figure 1.2, Table S1.7). A similar pattern was found for the combined 29 GHs exclusively identified as hemicellulases, which were significantly more abundant in WF than in other termite groups (Figure 1.3A, Table S1.7). Therefore, the gut metagenomes of LT and FC appear to be depleted in prokaryotic GHs targeting cellulose as compared to WF, possibly reflecting task partitioning between termite gut prokaryotes and other eukaryotic symbionts such as cellulolytic flagellates in LT and *Termitomyces* in FC. Task partitioning between gut prokaryotes and *Termitomyces* in FC was previously suggested for *Macrotermes natalensis* (Poulsen et al. 2014), with gut symbionts primarily participating to the final digestion of oligosaccharides and *Termitomyces* performing the breakdown of complex carbohydrates. In support of this hypothesis, several GHs, such as GH8, GH26, GH45, GH5_2, and GH53, largely depleted from the gut metagenomes of LT were highly expressed by the gut cellulolytic flagellates of *C. formosanus* (Nishimura et al. 2020), and were abundant in the gut metagenomes of WF. However, several GHs encoded by gut prokaryotes are also highly expressed by the gut cellulolytic flagellates of *C. formosanus* (*e.g.* GH13_8, GH36, GH3, GH92, GH133) (Nishimura et al. 2020). The extant of the complementarity between the CAZyme repertoires of gut flagellates and prokaryotes is therefore unclear and requires further investigation.

We next investigated the taxonomic origin of the prokaryotic CAZymes found in the same 129 whole gut metagenomes. We focused on the 19 GHs found in more than 10% of termite species and embedded in contigs longer than 5000 bps, allowing taxonomic annotation based on several genes. Contigs including genes with discordant taxonomic annotations potentially indicate horizontal gene transfers, as is common among bacteria (Ochman et al. 2000), and were removed. We found that Bacteroidota were a significant source of GH10, GH130, GH2, GH20, GH28, GH29 ,GH30, GH31, and GH9 in FC and LT, while, as previously described (Marynowska et al. 2020), they rarely encoded these GHs in non-Macrotermitinae Termitidae (WF and SF) (Figure 1.4, Table S1.8). In contrast, Fibrobacteres, which were very rare in LT, were a significant source of GH10, GH11, GH130, GH18, GH2, GH26, GH3, GH30, GH43, GH8, GH9, and GH94 in WF. Two other bacterial phyla, Spirochaetota and Firmicutes A, encoded most of the investigated GHs and were important contributors of GHs in WF (Figure 1.4, Table S1.8). Therefore, the primary contributors of GHs are distinct between lower and higher termites. These results are consistent with previous reports, which indicate a possible involvement of the ectosymbiotic Bacteroidota of some oxymonadid flagellates in cellulose and hemicellulose hydrolysis (Yuki *et al.*, 2015; Treitli *et al.*, 2019) in lower termites, while Fibrobacteres, Spirochaeota, and/or Firmicutes are major agents in cellulose and hemicellulose degradation in higher termites (Warnecke et al. 2007; Calusinska et al. 2010; He et al. 2013; Tokuda et al. 2018; Marynowska et al. 2020). Our comprehensive analyses strongly indicate that the loss of cellulolytic flagellates in the ancestor of higher termites was accompanied by a

major reworking of the cellulolytic bacterial communities, from Bacteroidota in LT to Fibrobacterota and Spirochaeota in WF and to Firmicutes in SF.

CAZymes are often organized as polysaccharide utilization loci (PULs) that target complex polysaccharides (Terrapon et al. 2015). To search for PULs in our metagenomes, we reconstructed metagenome-assembled genomes (MAGs) by grouping contigs with similarities in sequence composition and depth of coverage. In total, we obtained 654 prokaryotic MAGs that ranged in completeness from 30% to 100% with <10% contamination for lineage-specific marker genes. We included MAGs with completeness between 30% to 50% as several such MAGs possessed complete pathways of interest (Figure S1.3, Table S1.9). The 654 MAGs included members of 16 phyla of bacteria and four phyla of archaea and included representatives of all major prokaryote phyla known to be present in termite gut. We found 128 PULs distributed across 130 MAGs, including 31 MAGs of Bacteroidota, 71 MAGs of Firmicutes, 13 MAGs of Proteobacteria, 12 MAGs of Spirochaetota, two MAGs of Actinobacteria, and one MAG of Verrucomicrobiota (Table S1.10). Sixteen PULs, found in 10 MAGs, had all the PUL components and mainly targeted lignocellulose components such as cellulose and xylan, and saccharides such as melibiose, alignate, and lactose. 107 PULs found in 74 MAGs encoded for more than one substrate but did not have all the PUL components, possibly reflecting the incompleteness of our MAGs or missing components nonessential for their activity, as experimentally demonstrated in the xylan utilization system (*Xus*) of a Bacteroidota associated with *Pseudacanthotermes* (Wu 2018). Altogether, our data provide an overview of the PUL distribution in termite gut microbes.

Reductive acetogenesis in termite gut

The fermentation of wood fibers by the termite gut microbiota produces mostly acetate, which is used by the termite host, but also H_2 and CO_2 (Hungate 1939; Brune 2014). Most of the H_2 is used to produce additional acetate by the reduction of CO_2 (Breznak and Switzer 1986; Brauman et al. 1992; Pester and Brune 2007). We focused on the genes of seven enzymes of the Wood-Ljungdahl pathway (WLP) of reductive acetogenesis that are present in all acetogens from termite guts identified to date, namely formate dehydrogenase H (*fdhF*), formate-tetrahydrofolate ligase *(fhs),* methylenetetrahydrofolate dehydrogenase *(folD),* 5,10-methylenetetrahydrofolate reductase *(metF)*, acetyl-CoA synthase (*acsABCDE)*, phosphotransacetylase (*pta*) and acetate kinase (*ack*), which are essential to operate the bacterial WLP (Schuchmann and Müller 2014). We compared the relative abundance of these markers across the 129 whole gut metagenomes used for previous analyses and found a significant phylogenetic autocorrelation signal with the termite phylogenetic tree for five of the seven enzymes, two of which remain significant after FDR correction (*fdhF* and *acsABCDE*) (Figure 1.5, Table S1.11). Together with the five other enzymes, which also occur in many other bacteria, the simultaneous presence of *fdhF* and *acsABCDE* is a strong predictor for the distribution of reductive acetogenesis across the termite phylogenetic tree.

The seven enzymes encoded by all acetogens significantly differed in relative abundance among the four termite groups. They were generally more abundant in LT and WF than in FC and SF (Figure 1.3B, Table S1.11). These analyses are in agreement with previous studies that measured the potential rates of acetogenesis in a smaller set of termite species, and corroborate the hypothesis that reductive acetogenesis is mostly associated with a diet of wood and is less important in fungus-cultivating Macrotermitinae and in soil-feeding lineages (Brauman et al. 1992; Tholen and Brune 1999).

To determine the identity of the acetogens, we searched each MAG for the genes of the seven enzymes associated with reductive acetogenesis. We found 44 MAGs associated with six termite families and *Cryptocercus* that encoded at least five of the seven enzymes, but none of these MAGs contained the complete set of genes (Table S1.12, Figure 1.6A). In addition to formate dehydrogenase H (*fdhF*), we also searched for the genes encoding [FeFe] hydrogenase Group A4 (*HydA*) and the iron-sulfur cluster proteins (*HycB3*, *HycB4*), the other subunits of the hydrogen-dependent CO_2 reductase (HDCR) complex catalyzing the first step of CO_2 reduction to formate (Schuchmann and Müller 2012; Ikeda-Ohtsubo et al. 2016). Two MAGs lacked *fdhF* but contained all other genes of the WLP and the HDCR complex (Table S1.12, Figure 1.6A). These MAGs belonged to the Desulfobacterota family *Adiutricaceae*, which comprises the putatively acetogenic *Candidatus* Adiutrix intracellularis, a flagellate endosymbiont from the archotermopsid *Zootermopsis*, and numerous uncharacterized representatives from other lower and higher termites (Ikeda-Ohtsubo et al. 2016). Like *Ca.* Adiutrix intracellularis, none of the four MAGs encoded a sulfate reduction pathway. They were found in the rhinotermitid *Dolichorhinotermes* and in the higher termite *Microcerotermes*, indicating that the putatively free-living members of *Adiutricaceae* from higher termites (which lack gut flagellates) are also acetogenic.

Because none of the other MAGs encoded a complete WLP, we could not unambiguously attribute acetogenic status to any other prokaryote lineage. Considering the high rates of reductive acetogenesis in many lower and higher termites, particularly the wood-feeding species (Brauman et al. 1992), this may be explained either by the incompleteness of our MAGs or the failure to assemble any genomes of the populations responsible for the acetogenic activity. Based on the low free energy yields of both reductive acetogenesis and methanogenesis, it has been speculated that the proportion of (hydrogenotrophic) acetogens among the prokaryotic community in termite hindguts may be as low as that of (hydrogenotrophic) methanogens (Loh et al. 2021). The problem of genome assembly from low abundance populations would be exacerbated by a high species diversity among members of a particular metabolic guild. Alternatively, the absence of a complete reductive acetogenesis pathway among our MAGs may be genuine. This could be the case among the MAGs assigned to the family Treponemataceae B. Although the first isolate of this lineage is a homoacetogen with a complete WLP (Leadbetter et al. 1999), none of the other species isolated to date are acetogenic (Song et al. 2021). With the exception of *Treponema primitia* (Graber *et al.*, 2004), *Candidatus* Treponema intracellularis (Ohkuma et al. 2015), and *Candidatus* Adiutrix (Ikeda-Ohtsubo et al. 2016), the identity of the populations responsible for reductive acetogenesis in termite guts, including the putatively acetogenic *Candidatus* Termitimicrobium (Bathyarcheia; Loh et al., 2021) remains open to speculation.

Methanogenesis in termite gut

The methanogenic archaea present in the gut of termites consume a large fraction of H_2 and are responsible for 3% of global methane emissions (Brune 2018; Brune 2019). We searched the 129 gut metagenomes used in earlier analyses for genes that are part of methanogenesis pathways. Because of the low abundance of Archaea in termite guts (Brune 2019; Loh et al. 2021), the abundance of genes involved in methanogenesis was often near, or below, our detection threshold. As a consequence, we were unable to analyze each gene independently, but instead calculated the Moran's I index using the abundance of genes encoding the methyl-coenzyme M reductase complex (*mcrABG*), which catalyzes the final step of methanogenesis

(Evans et al. 2019), and found no autocorrelation signal with the termite phylogenetic tree (Figure 1.5, Table S1.11).
We compared the abundance of *mcrABG* among the four termite groups and found no significant differences (Figure 1.3B, Table S1.11). However, this lack of significance probably reflects the low abundance of archaeal reads in our assemblies, rather than an actual uniformity of methanogenesis pathways across termites, as methane emission rates are known to be diet-related and particularly high in species feeding on soil (*e.g.*, Brauman *et al.*, 1992; Bignell and Eggleton, 1995; Bignell *et al.*, 1997; Sugimoto *et al.*, 1998).

We searched our gut metagenomes for operons encoding *mcrABG*, and found 14 operons, belonging to four methanogenic archaeal orders, *Methanomassiliicoccales*, *Methanobacteriales*, *Methanomicrobiales*, and *Methanosarcinales*, derived from the gut metagenomes of 14 termite species, including four of the eight families of LT, and five of the nine subfamilies of Termitidae (Table S1.13). All *mcrABG* operons of LT were classified to *Methanobacteriales*, which is in agreement with previous reports on the prevalence of *Methanobacteriales* in LT (Brune 2019). An exception was found in the gut metagenome of *Porotermes quadricollis*, which yielded an *mcrABG* operon from *Methanomethylophilaceae* (order *Methanomassiliicoccales*). This is unusual, because members of this order are frequently encountered in higher termites and millipedes (Paul et al. 2012) but have been detected only once in the lower termite *Reticulitermes speratus* (Shinzato et al. 2001).
Next, we analyzed the methanogenic capacities of 26 MAGs of Archaea reconstructed from the gut metagenomes of 23 termite species from four termite families and the cockroach *Cryptocercus*. Only 13 MAGs belonging to *Methanomicrobiales*, *Methanobacteriales*, *Methanosarcinales*, and *Methanomassiliicoccales* encoded the *mcrABG* complex, indicating that the assemblies are incomplete (Figure 1.6B, Table S1.14). Five of these 13 MAGs possessed complete pathways for methylotrophic methanogenesis and one MAG possessed complete pathways for hydrogenotrophic methanogenesis (Figure 1.6B). The five MAGs showing genomic evidence of methylotrophic methanogenesis included one MAG of *Methanosarcinales* (genus *Methanimicrococcus*) and four MAGs of *Methanomassiliicoccales*, including three MAGs classified to genus *Methanoplasma* and one MAG classified to family *Methanomethylophilaceae*. Only two MAGs of *Methanoplasma* encoded a methanol:coenzyme M methyltransferase (*mtaABC*) complex, which is required for growth on methanol and typical for all members of this lineage (Lang et al. 2015), and only one of the MAG of *Methanosarcinales* and one MAG of *Methanoplasma* encoded a complete heterodisulfide reductase complex (*HdrA2B2C2/mvhADG*) present in most methanogens (Thauer et al. 2008; Buckel and Thauer 2013), underscoring the incompleteness of the MAGs. The same was true for hydrogenotrophic methanogenesis, for which only one MAG belonging to *Methanobacteriaceae* (genus *Methanobrevibacter C*) possessed most of the genes required for the reduction of CO_2 to methane, including a heterodisulfide reductase (*HdrABC/mvhADG*) complex, an iron-sulfur flavoprotein along with a F420-independent hydrogenase (*Fdh*), and a F420 reducing hydrogenase (*FrhABC*) (Figure 1.6B, Table S1.14). The absence of aceticlastic methanogens is in agreement with previous reports (Brune 2018; Brune 2019). Overall, our results highlight the diversity of methanogens found in termite guts, and the diversity of the pathways they use.

Sulfate-reducing prokaryotes

Sulfate-reducing bacteria are potential H_2-consumers in the gut of termites and *Cryptocercus* (Brauman et al. 1992; Kuhnigk et al. 1996; Dröge et al. 2005) (Figure 1.5). However, sulfate concentration is low in termite gut, as is H_2 consumption by sulfate-reducing bacteria (Dröge et al. 2005; Brune and Ohkuma 2011). We found all the genes of the dissimilatory sulfate reduction pathway, namely, the two subunits of adenylylsulfate reductase (*aprA* and *aprB*), sulfate adenylyltransferase (*sat*), and dissimilatory sulfite reductase (*dsrAB*), in six out of eight lower termite families, all the higher termite subfamilies, and *Cryptocercus*. The abundance of *aprAB* and *sat* were significantly correlated with the termite phylogenetic tree, and the correlation remained significant after FDR correction for *sat* (Figure 1.5, Table S1.11).

Comparisons of the four termite groups showed that the abundance of *aprAB* was significantly higher in WF than in SF and the abundance of *sat* was significantly higher in LT than WF and SF (Figure 1.3B, Table S1.11). While sulfate reducers have been isolated from the guts of LT, FC and SF (Brauman et al. 1992; Kuhnigk et al. 1996), we found metagenomic evidence that sulfate reduction is also prevalent in WF.

Next, we analyzed the sulfate-reducing capabilities of our 654 MAGs and found a complete pathway for dissimilatory sulfate reduction in four MAGs (Figure 1.6C, Table S1.15). Three of these MAGs, found in the termites *Parrhinotermes*, *Reticulitermes*, and *Tumulitermes*, were assigned to *Desulfovibrionaceae* (Desulfobacterota), which are common in the termite gut and generate energy via sulfate respiration (Sato et al. 2009; Kuwahara et al. 2017). Of note, the fourth MAG, retrieved from the gut metagenome of the apicotermitine *Heimitermes laticeps*, belonged to the Proteobacteria family *Burkholderiaceae*, a bacterial family that was, prior to this study, largely unreported from termite guts, and that is abundant in Apicotermitinae and in the termite clade that includes the Cubitermitinae, the *Pericapritermes*-group, and the *Termes*-group. The evidence for dissimilatory sulfate reduction in *Burkholderiaceae* termite guts suggest that the capacity for sulfate respiration is more widely distributed than expected.

Nitrogen recycling by termite gut prokaryotes

Because the content of nitrogen in wood is low, termites have evolved mechanisms of nitrogen conservation. The termite gut microbiota contributes to the nitrogen metabolism of its host by recycling nitrogen (Breznak 2000; Hongoh 2011). Like most insects, termites convert waste products from nitrogen metabolism into uric acid, but, unlike other insects, the gut prokaryotes of termites degrade uric acid into ammonia, which is subsequently assimilated by the gut microbiota (Brune, 2014). We searched the 129 metagenomes used for previous analyses and found only few genes possibly involved in uric acid degradation, including 11 *aegA* (a putative oxidoreductase suspected to be involved in uric acid degradation by Enterobacteriaceae (Iwadate and Kato 2019)) in six termite species. Since the uricolytic prokaryotes isolated from termite guts are strict anaerobes (Potrikus and Breznak 1980; Potrikus and Breznak 1981; Thong-On et al. 2012), it is likely that they use alternative, so far unknown, pathways. Termite tissues reportedly lack uricase activity (Potrikus and Breznak 1981), but when we examined the transcriptomes of 53 termite species generated by Buček *et al.* (2019), we found evidence for the expression of a gene encoding urate oxidase in 20 termite species belonging to four termite families (Figure S1.4). This indicates that termites should be able to carry out the first step of uric acid degradation. However, the extent of the contribution of the termite host to uricolysis and the identity of the uricolytic prokaryotes and their catabolic pathways remain unknown.

The metagenomes of all termite families and *Cryptocercus* included numerous prokaryotic genes from other pathways involved in the production of ammonia (Figure 1.5, Table S1.11),

including ureases (*ureABC*), which degrade urea into ammonia (Hongoh and Ohkuma 2010; Ohkuma et al. 2015), and some of the genes of the dissimilatory nitrate reduction pathway (*narGHI*, *napAB*, *nrfAB*), which convert nitrate into ammonia. Among those, the abundance of *ureABC* genes significantly correlated with the termite phylogenetic tree after FDR correction (Figure 1.5, Table S1.11). We also found in the metagenomes of all termite families and *Cryptocercus* genes from pathways involved in amino acid biosynthesis from ammonia, including glutamine synthetase (*glnA*) and glutamate synthase (*gltBD*), the genes involved in the synthesis of glutamate from ammonia, and carbamate kinase (*arcC*), ornithine carbamoyltransferase (*argF*), argininosuccinate synthase (*argG*) and argininosuccinate lyase (*argH*), the genes involved in arginine biosynthesis from ammonia (Yan 2007). The abundance of *gltBD* correlated with the termite phylogenetic tree after FDR correction (Figure 1.5, Table S1.11). Therefore, the termite phylogeny is a good predictor of the enteric abundance of some of the prokaryotic genes involved in ammonia metabolism in termites.

We compared the four termite groups using the relative abundance of the nitrogen-recycling genes and found that the abundance of *ureABC* differed among termite groups, with the gut metagenomes of LT and WF significantly enriched in *ureABC* as compared to those of SF and FC (Figure 1.3B, Table S1.11). In contrast, the abundance of some of the genes of the dissimilatory nitrate reduction pathway, such as *napAB* and *narGHI*, was significantly reduced in the gut metagenomes of WF compared to SF and FC (Figure 1.3B, Table S1.11). This suggests that the high rates of nitrate ammonification previously found in two soil-feeding species (Ngugi and Brune 2012) is a characteristic that all soil-feeding termites share with fungus-cultivating termites. We also found that *gltBD* was significantly enriched in LT as compared to other termite groups, while *argFGH* was significantly enriched in LT and WF as compared to SF (Figure 1.3B, Table S1.11). The low abundance of genes involved in ammonia assimilation in soil-feeding termites is likely linked to their diet, which includes soil peptidic residues (Ji and Brune 2001; Ji and Brune 2005).

Next, we searched our 654 MAGs to determine the taxonomic identity of the prokaryotes involved in nitrogen recycling. Six MAGs possessed the three ureases *ureABC*, thence encoded enzymes to convert urea into ammonia, and 15 MAGs included a complete dissimilatory nitrate reduction pathway that convert nitrate into ammonia. All these MAGs belonged to diverse lineages of Proteobacteria and Campylobacterota (order *Campylobacterales*), except for one MAG of Firmicutes (genus *Bacillus*) found in *Foraminitermes rhinoceros* and endowed with *ureABC*, *narGHI* and *nirBD* (Figure 1.7A, Table S1.16). We also found numerous MAGs capable of ammonia assimilation into glutamate and arginine, indicating that ammonia is an important source of nitrogen for many termite gut prokaryotes. 91 MAGs possessed *glnA* and *gltBD* for glutamate biosynthesis from ammonia, while 26 MAGs possessed the four genes *arcC*, *argF*, *argG*, and *argH* for arginine biosynthesis from ammonia via the urea cycle, including 12 MAGs that also contained the glutamate biosynthesis pathway (Figure 1.7A, Table S1.16). 66 MAGs encoding glutamate biosynthesis genes, and 15 MAGs with arginine biosynthesis genes, also possessed the ammonium transporter *Amt*. These MAGs belonged to ten phyla, including 19 MAGs of Proteobacteria from six families,18 MAGs of Bacteroidota, of which eight belonged to the family *Azobacteroidaceae*, ten MAGs of Actinobacteria of three families, eight MAGs of the Spirochaetes family *Treponemataceae* B, eight MAGs of Firmicutes, six MAGs of *Campylobacterota*, five MAGs of Firmicutes A, three MAGs of Planctomycetota, three MAGs of Desulfobacterota, and one MAG of Verrucomicrobia (Figure 1.7A, Table S1.16).

Therefore, a great many bacterial lineages contribute to the nitrogen metabolism of their termite hosts.

Nitrogen fixation by termite gut prokaryotes

Many species of wood-feeding termites host dinitrogen-fixing prokaryotes in their gut, which compensate for the low nitrogen content of wood (Breznak 2000). They fix nitrogen with either the molybdenum-dependent (Nif), vanadium-dependent (Vnf), or iron-only alternative nitrogenases (Anf) (Ohkuma et al. 1999; Yamada et al. 2007; Inoue et al. 2015). We found gene homologs for the structural subunits of these nitrogenases (collectively referred to as *nifDHK*) in all metagenomes of termite families and in *Cryptocercus* (Figure 1.5). Their abundance significantly correlated with the termite phylogeny after FDR correction (Figure 1.5, Table S1.11), as was the case for several other pathways involved in nitrogen economy. There were significant differences among termite groups, with the nitrogenase reads in the gut metagenomes of non-FC wood-feeders (LT and WF) being 24.4-fold more abundant than in SF and 20.2-fold more abundant than in FC (Figures 1.3B, 1.5, Table S1.11). This is in line with the higher rate of N_2 fixation measured in LT and WF than in SF and FC (Yamada et al. 2007), and reflects the high amount of nitrogen present in soil and fungi, making the energy-demanding process of N_2 fixation unnecessary (Brune and Ohkuma 2011; Hongoh 2011).

To identify the diazotrophs present in the gut of termites, we taxonomically classified contigs longer than 5000 bps that contained the six genes present in all diazotrophs, *nifDHK* (which encode nitrogenase), and *nifB*, *nifE*, *nifN*, which encode proteins involved in nitrogenase biosynthesis (Dos Santos et al. 2012). We identified 15 contigs matching these criteria in the gut metagenomes of 12 termite species, representing five of the nine termite families (Table S1.17). These contigs were assigned to diverse prokaryote lineages, including nine contigs of diverse Bacteroidota, three contigs of the Spirochaetota order *Treponematales*, two contigs of Proteobacteria family *Enterobacteriaceae*, and one contig of the archaeal genus *Methanobrevibacter*. We carried out the same analyses on our MAGs and found 18 MAGs that contained a *nifHDKBEN* cluster, including seven MAGs that belonged to phyla not represented in the contigs >5000 bps. Among these seven MAGs, there were four MAGs of the Actinobacteriota family *UBA8131*, one MAG of the Planctomycetota family *Thermoguttaceae*, one MAG of the Verrucomicrobiota family *Chthoniobacteraceae*, and one MAG of Firmicutes C order *Acidaminococcales* (Figure 1.7A, Table S1.16). Therefore, the taxonomy of diazotrophs found in our termite species set corroborates previous evidence that termites host diverse communities of diazotrophs in their guts (Ohkuma et al. 2001; Yamada et al. 2007; Desai and Brune 2012).

We next investigated the taxonomic distribution of diazotrophs across termites. We focused on contigs longer than 5000 bps that included genes with concordant taxonomic annotation and that contained a *nifHDK* operon (Figure 1.7B, Table S1.18). In lower termites, the dominant diazotroph was an undescribed Bacteroidota allied to an ectosymbiont of the *Cryptocercus* gut flagellate *Barbulanympha* (Tai et al. 2016). This undescribed Bacteroidota was found in three of the eight families of LT. It was also largely absent from the gut metagenomes of *Coptotermes* and *Heterotermes*, which harbor the flagellate endosymbiont *Candidatus* Azobacteroides as the main diazotroph (Hongoh et al. 2008). The diazotrophs of Termitidae belonged to various phyla. Notably, we found the N_2-fixing *Candidatus* Azobacteroides in the nasutitermitine *Coatitermes* (which lacks gut flagellates), and a N_2-fixing *Treponematales* in *Mastotermes*, highlighting that the dominant lineages of diazotrophs in particular termite lineages are also harbored at a low abundance by unrelated species of termites and *Cryptocercus* (Figure 1.7B, Table S1.18).

Therefore, our results indicate that the phylogenetic position of termite species determined, at least partly, the taxonomy of their dominant diazotrophs.

Discussion

The metagenomics and metatranscriptomics surveys of termite guts carried out so far targeted a limited number of termite species (*e.g.* Warnecke *et al.*, 2007; He *et al.*, 2013; Liu *et al.*, 2018; Tokuda *et al.*, 2018; Marynowska *et al.*, 2020), and thus did not permit an investigation of how the gut microbiome of these social roaches has been evolving in term of function and composition since termites and *Cryptocercus* diverged >150 Million years ago. To address this issue, and to provide a global picture of the taxonomic and functional composition of the termite gut microbiome, we generated gut metagenomes for a comprehensive set of 145 termite species and one species of *Cryptocercus*. The analyses of this dataset revealed that: (1) gut prokaryotic genes involved in the main nutritional functions are generally present across termites and *Cryptocercus*, suggesting these genes were already harbored by the common ancestor of termites and *Cryptocercus*; (2) the termite phylogenetic tree is largely predictive of the gut bacterial community composition and the nutritional function they exert; and (3) the acquisition of a diet of soil was accompanied by a change in the stoichiometry of genes and metabolic pathways involved in important nutritional functions rather than by the acquisition of new genes and pathways.

The analyses of our 146 gut metagenomes indicated that prokaryotic CAZymes, genes of the reductive acetogenesis, sulfur reduction, and methanogenesis pathways, and genes involved in nitrogen fixation and recycling, are present across the nine termite families and *Cryptocercus*. Therefore, the nutritional functions previously known to be performed by the gut prokaryotic symbionts of particular termite species (*e.g.* Warnecke *et al.*, 2007; Calusinska *et al.*, 2020) are likely performed in the gut of all termites and *Cryptocercus* spp. These results strongly suggest that the gut prokaryotes performing important nutritional functions were already harbored by the common ancestor of termites and *Cryptocercus*. Following this scenario, the ancestor of termites and *Cryptocercus* did not only acquire their charismatic gut cellulolytic flagellates (Nalepa 1991), but also acquired several bacterial and archaeal lineages that make up a sizable fraction of the gut microbiota of modern termite species. In support of this hypothesis, many termite gut bacteria phylotypes form monophyletic groups present in the gut of various termite families and distantly related to bacteria found in other environments, such as in the guts of other animals, including cockroaches (Bourguignon et al. 2018). Therefore, as the cockroach-like ancestor of termites and *Cryptocercus* evolved wood-feeding, it is likely that it recruited facultative gut microbes able to degrade wood and participate in the nitrogen economy as essential gut symbionts.

Our analyses indicate that the phylogenetic position of termite species is partly predictive of the functions of gut bacterial communities. This is best illustrated by CAZymes whose abundance often correlated with the termite phylogenetic tree. Correlation with the termite phylogenetic tree, however, was not found for some genes, such as the *mcrABG* genes of the methanogenesis pathway, the genes of sulfate reduction pathway, and the genes of the dissimilatory nitrate reduction pathway. Whether this lack of correlation is genuine, or whether it reflects insufficient depth of sequencing, is unclear and requires further study. In any case, our results indicate that the phylosymbiotic patterns observed between termites and their gut bacterial and protist communities (Tai et al. 2015; Abdul Rahman et al. 2016) are also found for some gut microbial functions, which, at least partly, recapitulates the termite phylogeny.

The comparison of four termite groups, soil-feeding Termitidae (SF), fungus-cultivating Macrotermitinae (FC), non-Macrotermitinae wood-feeding Termitidae (WF), and lower termites (LT), reveals that genes and metabolic pathways important to termites are present in all termite species, but their abundances vary among groups. Notably, the gut metagenomes of SF possessed on average fewer CAZymes, nitrogenases, reductive acetogenesis, and sulfate-reducing genes than the gut metagenomes of other termite groups. Therefore, as pointed out by Marynowska *et al.* (2020), the gut prokaryote communities of SF retain important carbohydrate metabolism capabilities. Nevertheless, our dataset clearly indicate that these abilities are much reduced in soil-feeders compared to wood-feeders. Overall, our results support the idea that the acquisition of soil-feeding was accompanied by changes in the abundance of the gut prokaryote metabolic pathways important to termite nutrition.

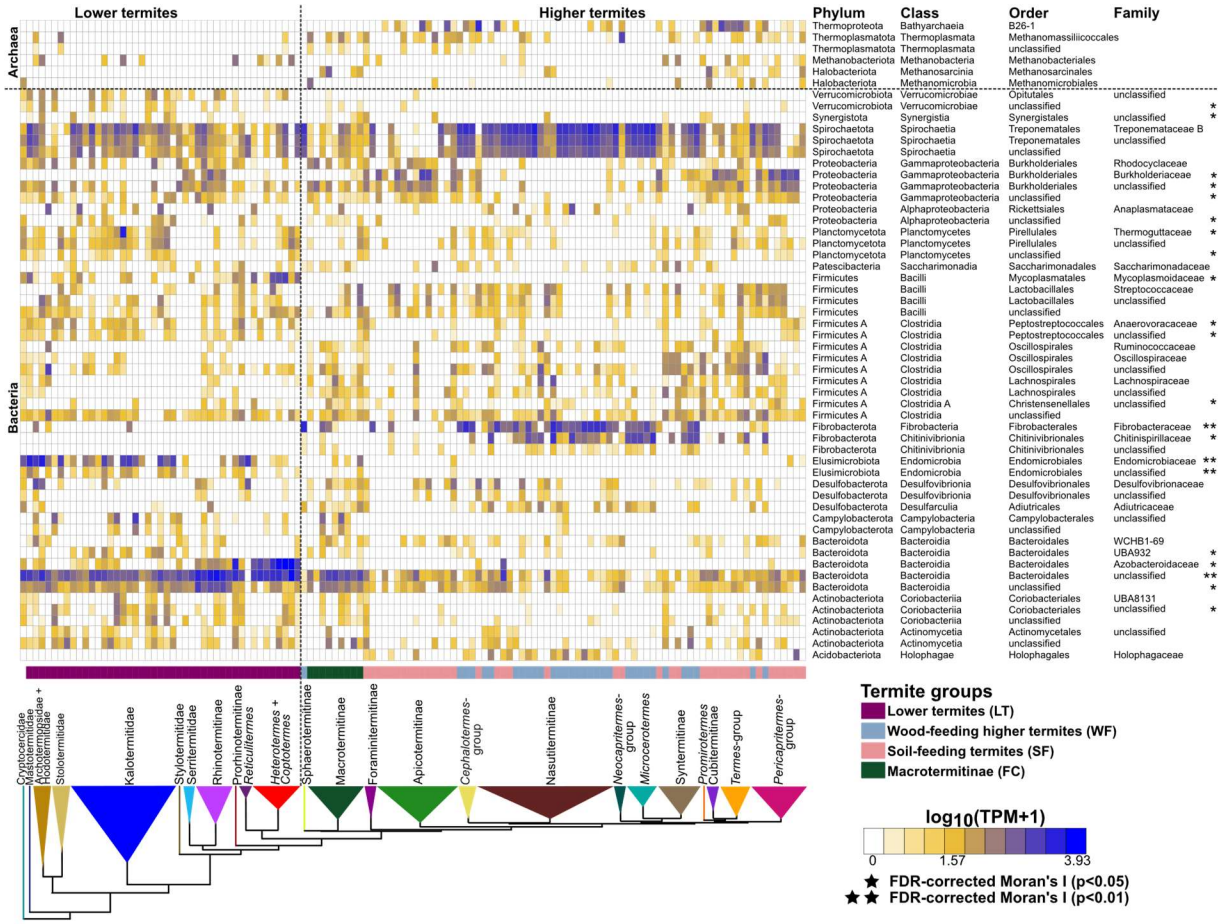

Figure 1. 1. Relative abundance of the top 50 bacterial lineages and the major archaeal orders found in the gut metagenomes of termites and *Cryptocercus*. The relative abundance of prokaryotic taxa was inferred from 40 single-copy marker genes. The color scale represents the logarithm of transcripts per million (TPM). The tree represents a simplified time-calibrated phylogenetic tree reconstructed using host termite and *Cryptocercus* mitochondrial genome sequences. Prokaryotic taxa presenting significant phylogenetic autocorrelation with the host phylogeny at a 5% false discovery rate (FDR) are indicated with an asterisk (*p < 0.05; **p < 0.01).

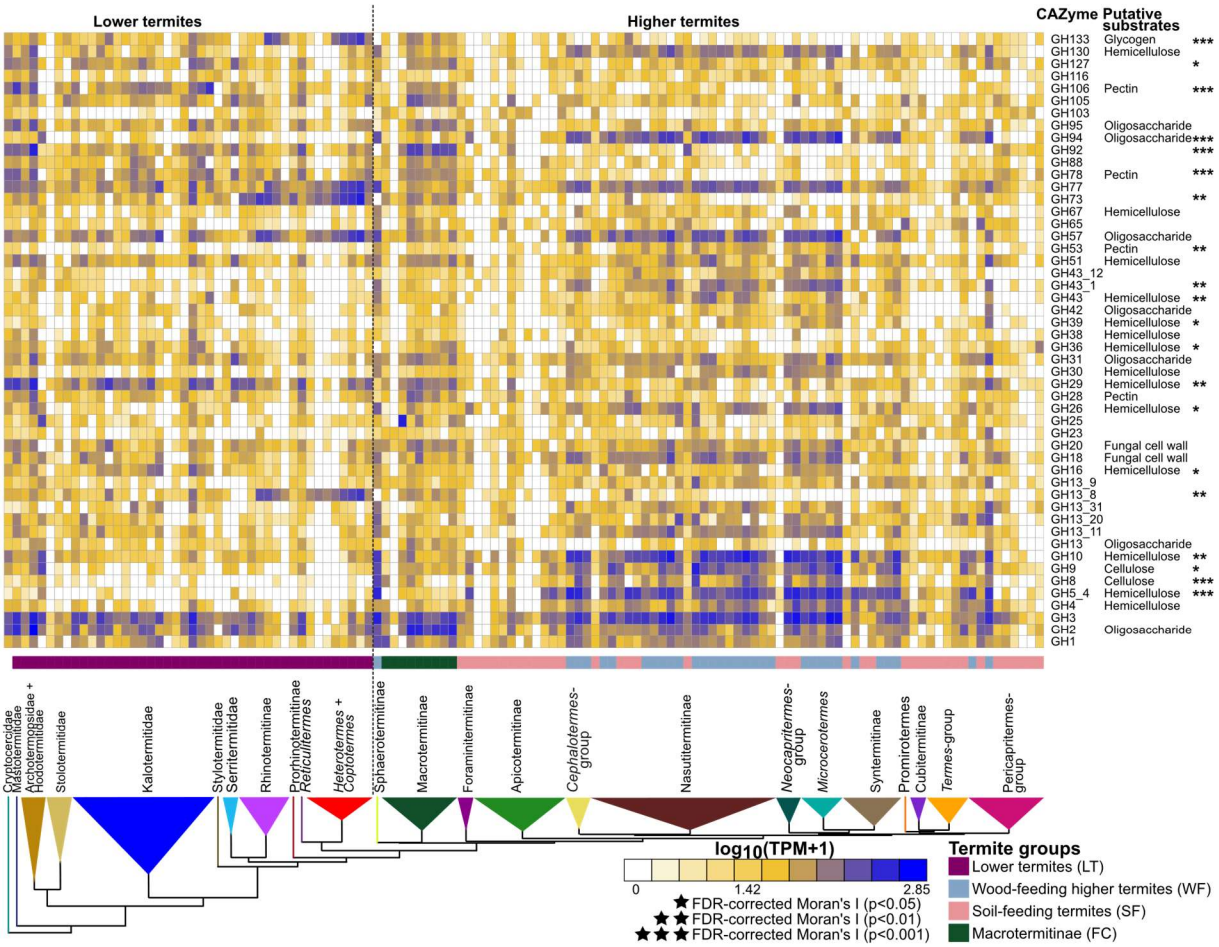

Figure 1. 2. Relative abundance of CAZymes found in gut metagenomes of termites and *Cryptocercus*. The heatmap shows the 50 most abundant CAZymes. The color scale represents the logarithm of transcripts per million (TPM). The tree represents a simplified time-calibrated phylogenetic tree reconstructed using host termite and *Cryptocercus* mitochondrial genomes. Genes showing significant phylogenetic autocorrelation with the host phylogeny at a 5% false discovery rate (FDR) are indicated with asterisks (*p < 0.05; **p < 0.01).

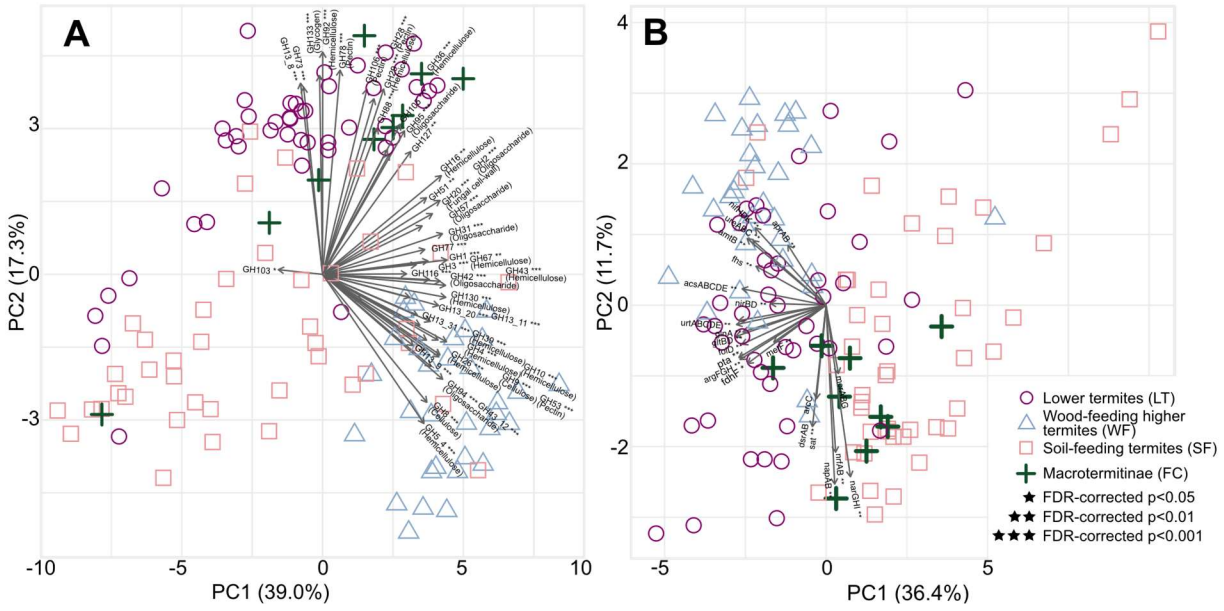

Figure 1. 3. Principle component analysis (PCA) bi-plots showing the distribution of prokaryotic genes involved in lignocellulose digestion in the gut of termites and *Cryptocercus*. (A) PCA performed on the relative abundance of CAZymes found in 129 gut metagenome assemblies. The 50 glycoside hydrolases (GHs) that contributed the most to separation of termite diets are plotted (see Table S1.7). (B) PCA inferred from relative abundance of metabolic genes involved in lignocellulose digestion after carbohydrate degradation. The symbols indicate host feeding habits. The species identity of each data point is available in Table S1.1. Asterisks indicate significant differences among the four termite groups at 5% false discovery rate (FDR, *p < 0.05; **p < 0.01; ***p < 0.001).

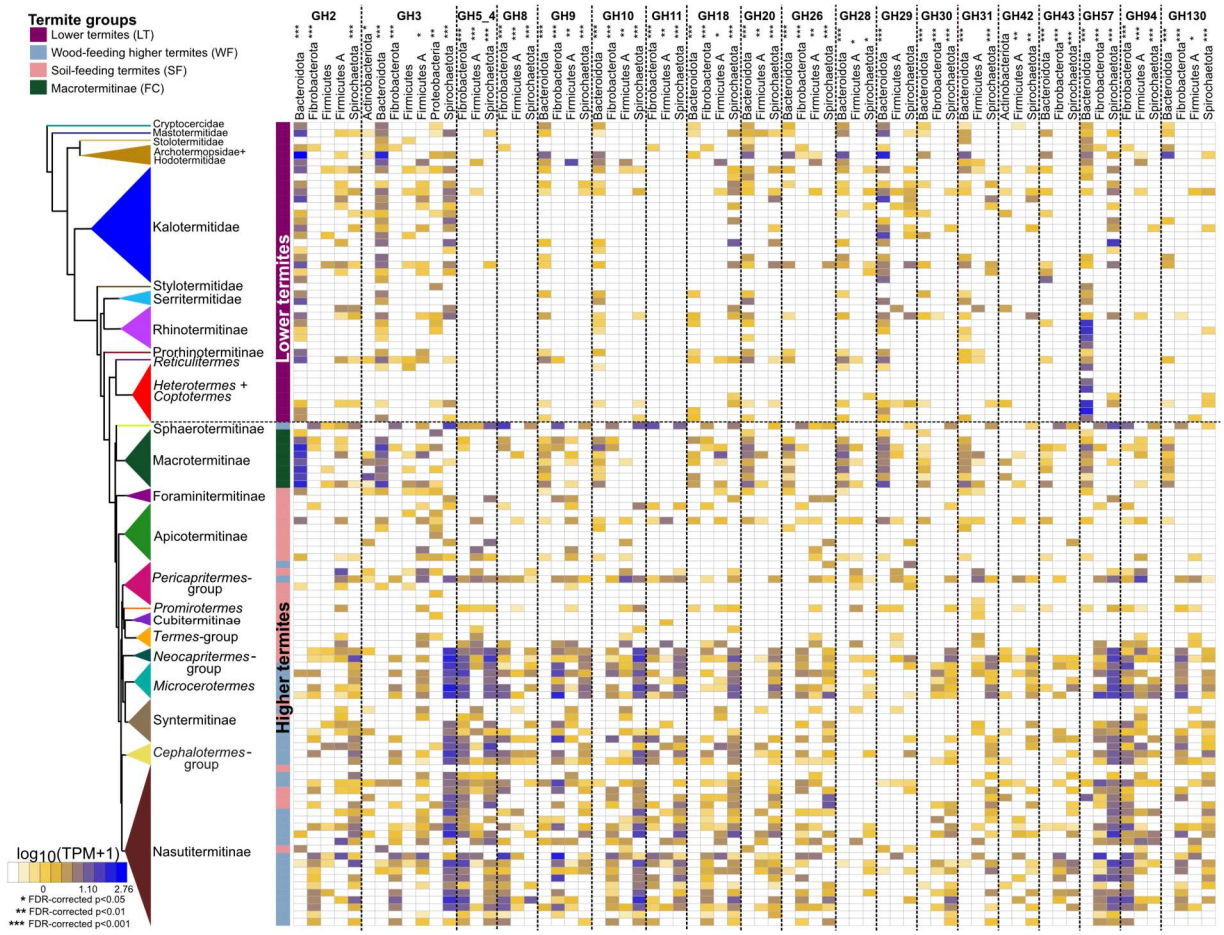

Figure 1. 4. CAZyme families, and their taxonomic origin, for enzymes derived from contigs longer than 5000 bps and present in 10% of gut metagenomes. The color scale represents the log-transformed transcripts per million (TPM). The tree represents a simplified time-calibrated phylogenetic tree reconstructed using host termite and *Cryptocercus* mitochondrial genomes. Asterisks indicate significant differences among the four termite groups at 5% false discovery rate (FDR, *p < 0.05; **p < 0.01; ***p < 0.001).

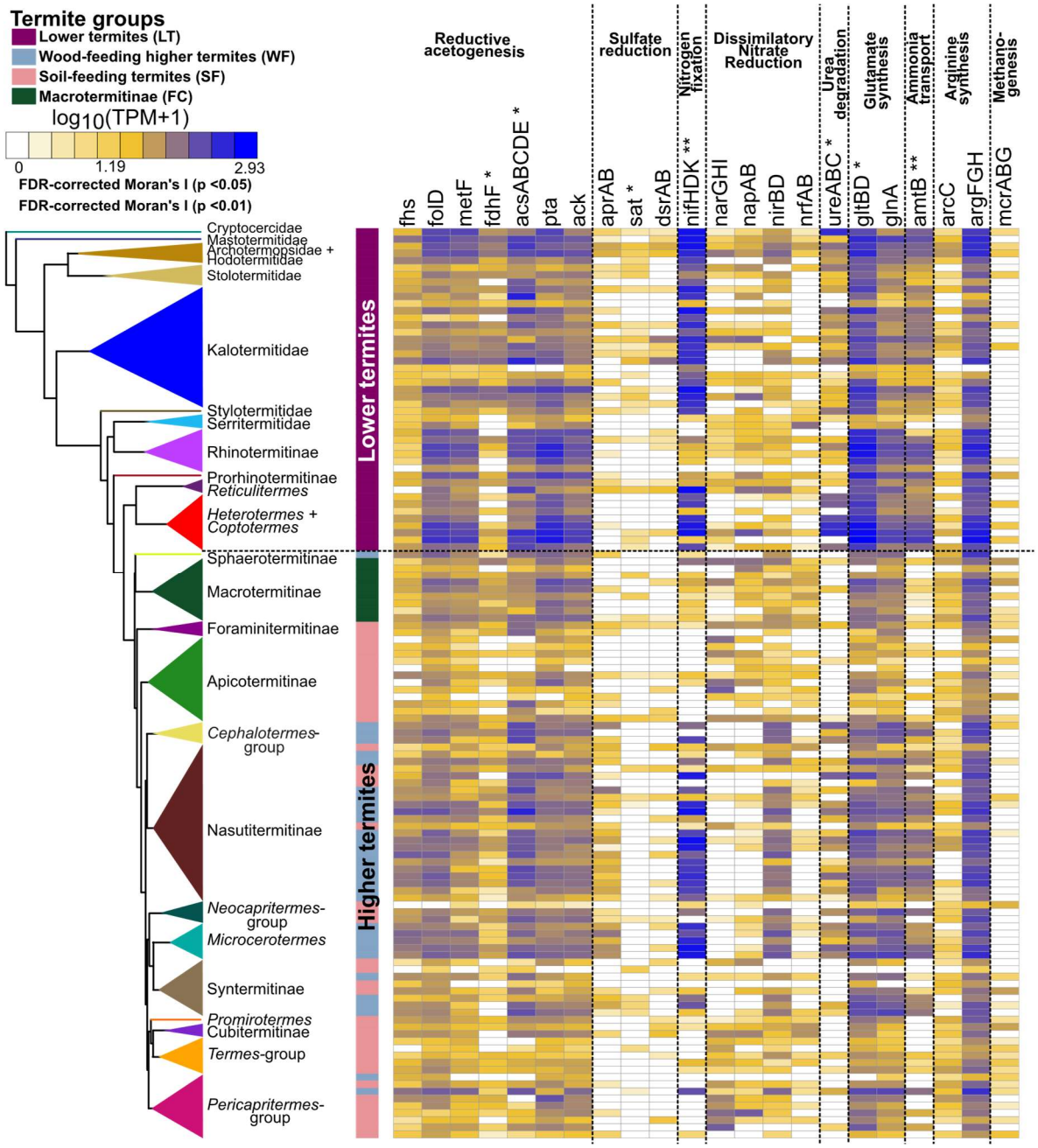

Figure 1. 5. Relative abundance of prokaryotic genes belonging to metabolic pathways involved in the final steps of the lignocellulose digestion in the gut of termites and *Cryptocercus*. The color scale represents the logarithm of transcripts per million (TPM). The tree represents a simplified time-calibrated phylogenetic tree reconstructed using host termite and *Cryptocercus* mitochondrial genomes. Full names of the gene families and their corresponding KEGG IDs are available in Table S1.11.

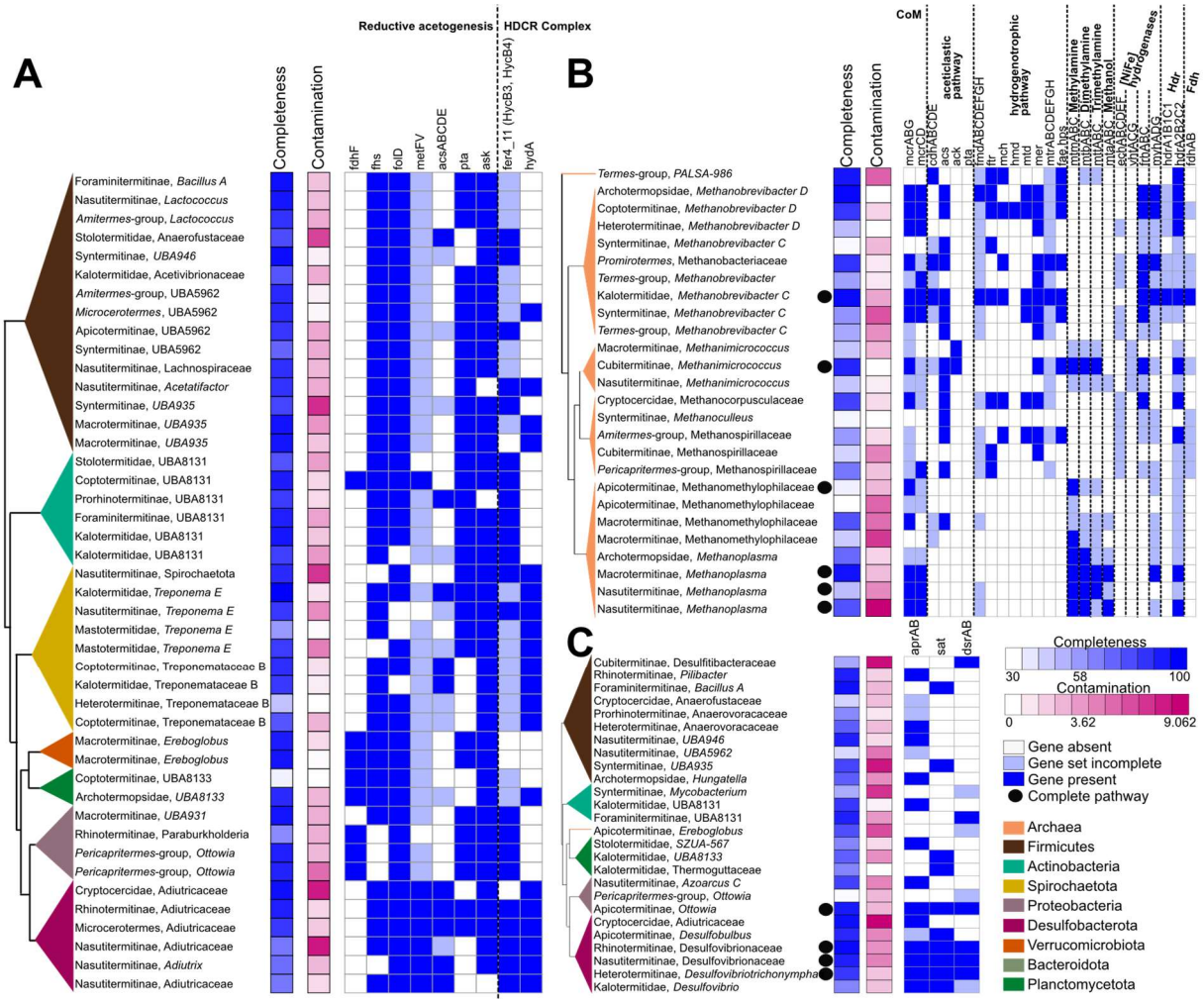

Figure 1. 6. Metabolic pathways involved in the final steps of lignocellulose digestion found in gut metagenome assembled genomes (MAGs) reconstructed in this study. (A) Genes involved in reductive acetogenesis, (B) methanogenesis, and (C) sulfate reduction found in MAGs. The trees represent simplified maximum likelihood phylogenetic trees of the MAGs reconstructed using 43 single-copy marker genes. MAG completeness and contamination, based on CheckM analyses, is shown beside the tree. Dark blue squares indicate gene presence, light blue squares indicate that incomplete gene sets, and open squares indicate gene absence. Detailed information on the gene families and their KEGG IDs are available in Tables S1.12, S1.14, and S1.15.

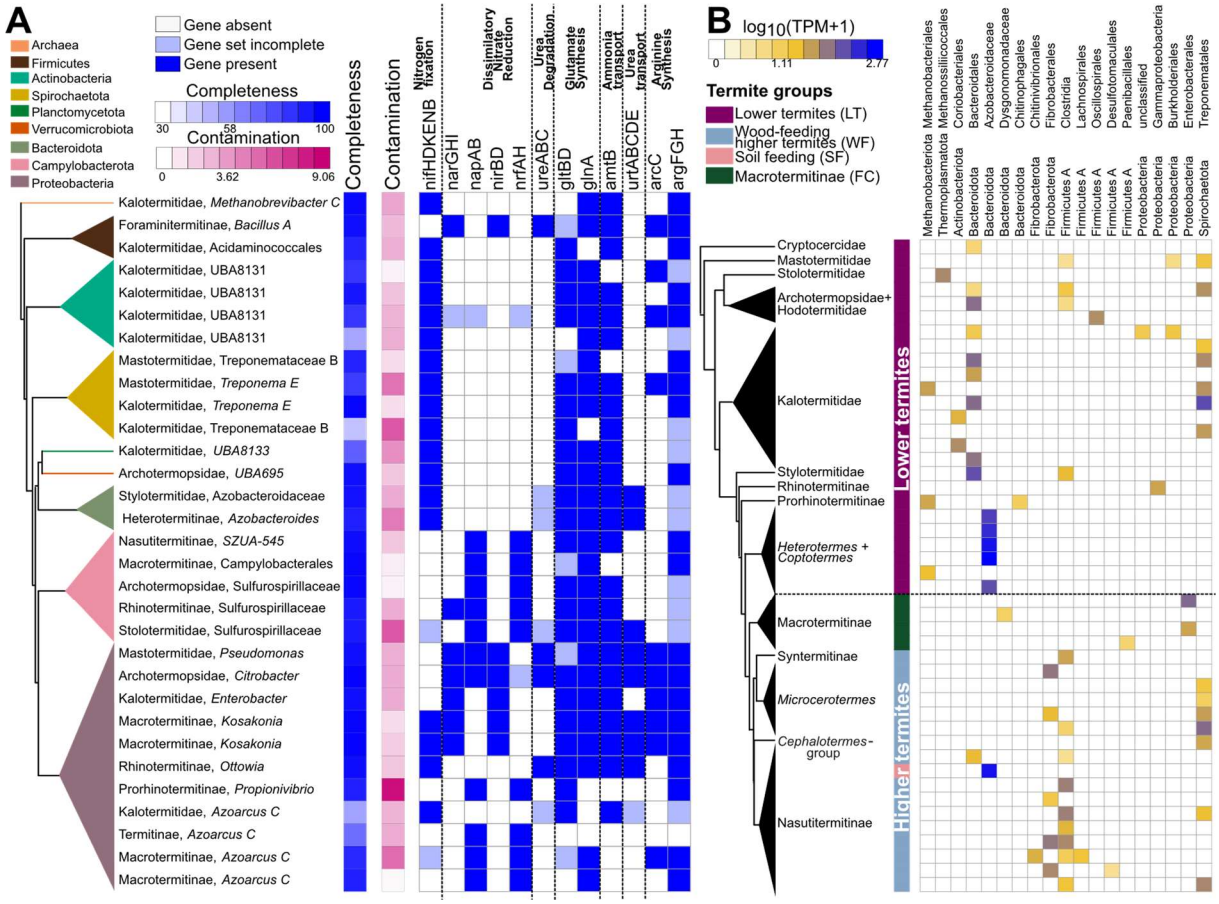

Figure 1. 7. Nitrogen metabolism in the gut of termites and *Cryptocercus*. (A) Metagenome-assembled genomes (MAGs) with complete nitrogen fixation or dissimilatory nitrate reduction pathways. All pathways potentially involved in the nitrogen metabolism, namely nitrogen fixation, dissimilatory nitrate reduction, ureases, glutamate metabolism, ammonia transport, urea transport, and arginine metabolism are represented. The tree represents a simplified maximum likelihood phylogenetic tree of the MAGs inferred from 43 marker genes. Completeness and contamination of MAGs, based on CheckM analysis, are shown beside the tree. Dark blue squares indicate gene presence, light blue squares indicate that incomplete gene sets, and open squares indicate gene absence. (B) Abundance of NifHDK operons (*nifHDK*, *vnfHDK*, or *anfHDK*) present in contigs longer than 5000bps across gut metagenomes. The color scale represents the log-transformed transcripts per million (TPM). The tree represents a simplified time-calibrated phylogenetic tree reconstructed using host termite and *Cryptocercus* mitochondrial genomes.

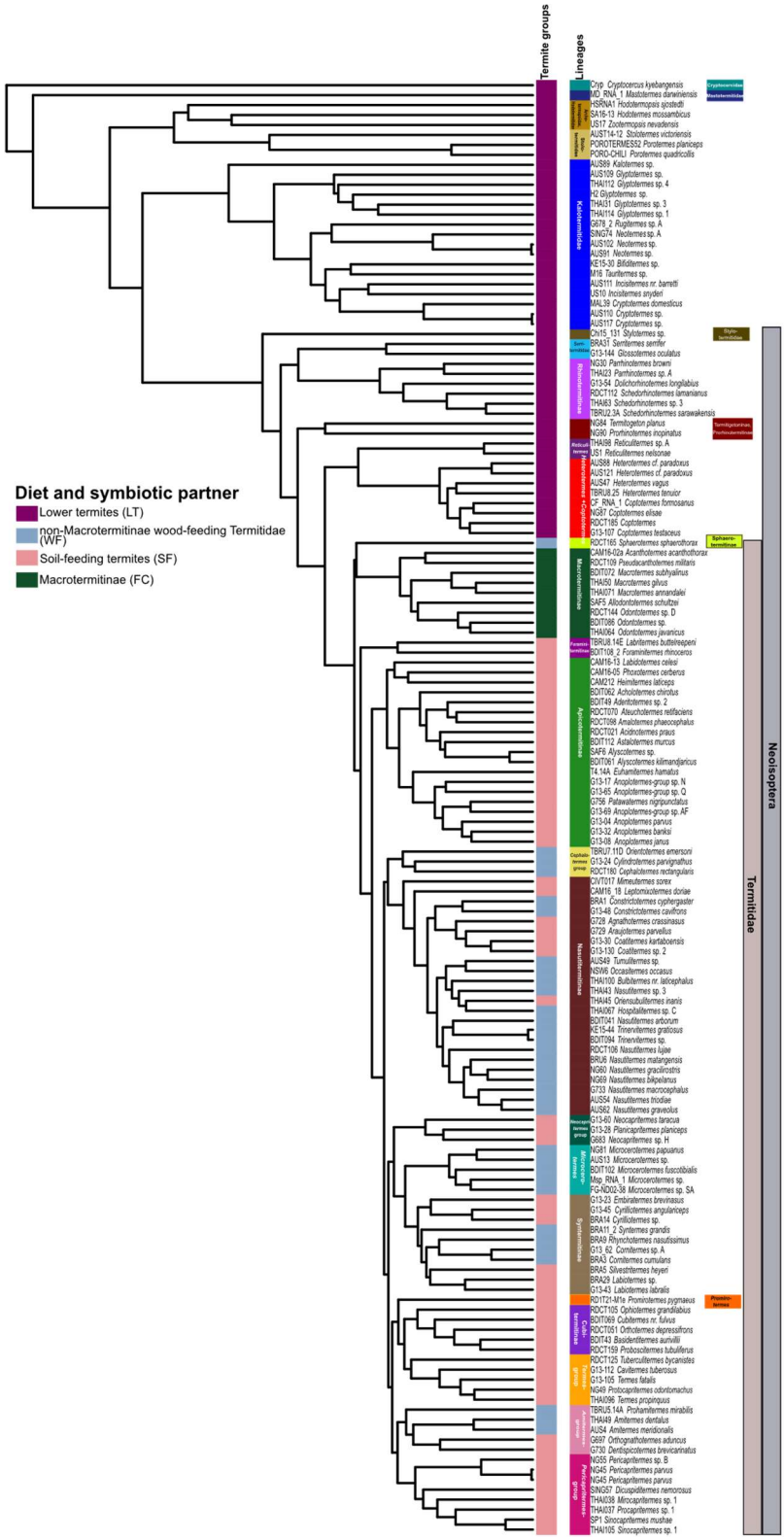

Figure S1. 1. Time-calibrated phylogenetic tree of termites and *Cryptocercus* inferred from mitochondrial genome sequences.

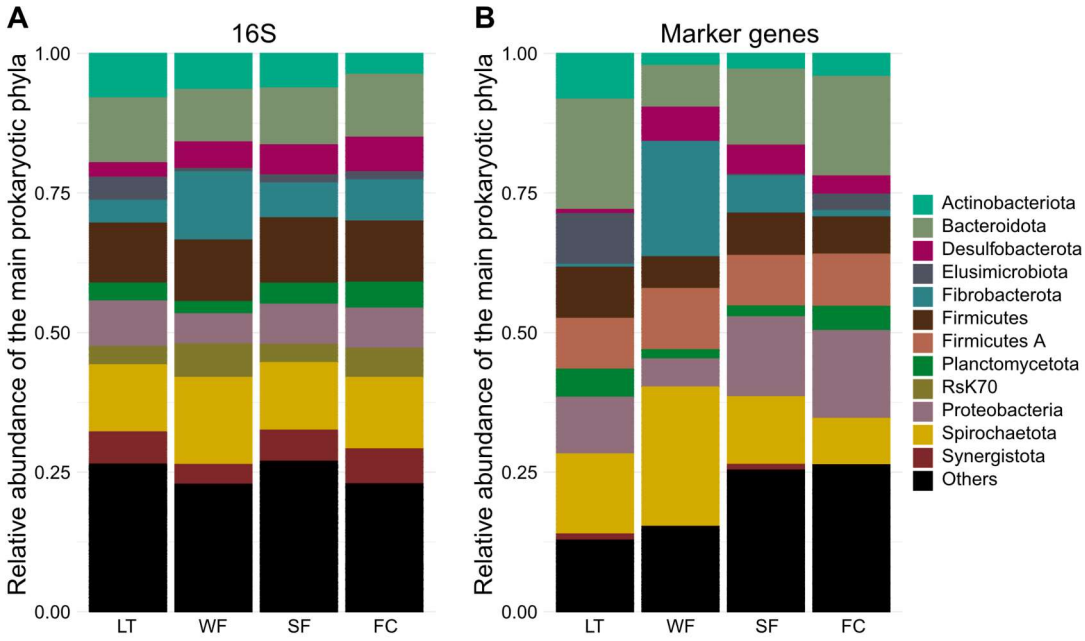

Figure S1. 2. Relative abundance of archaeal and bacterial phyla inferred from the termite gut metagenomes and the 16S rRNA amplicon data of 74 termite samples.

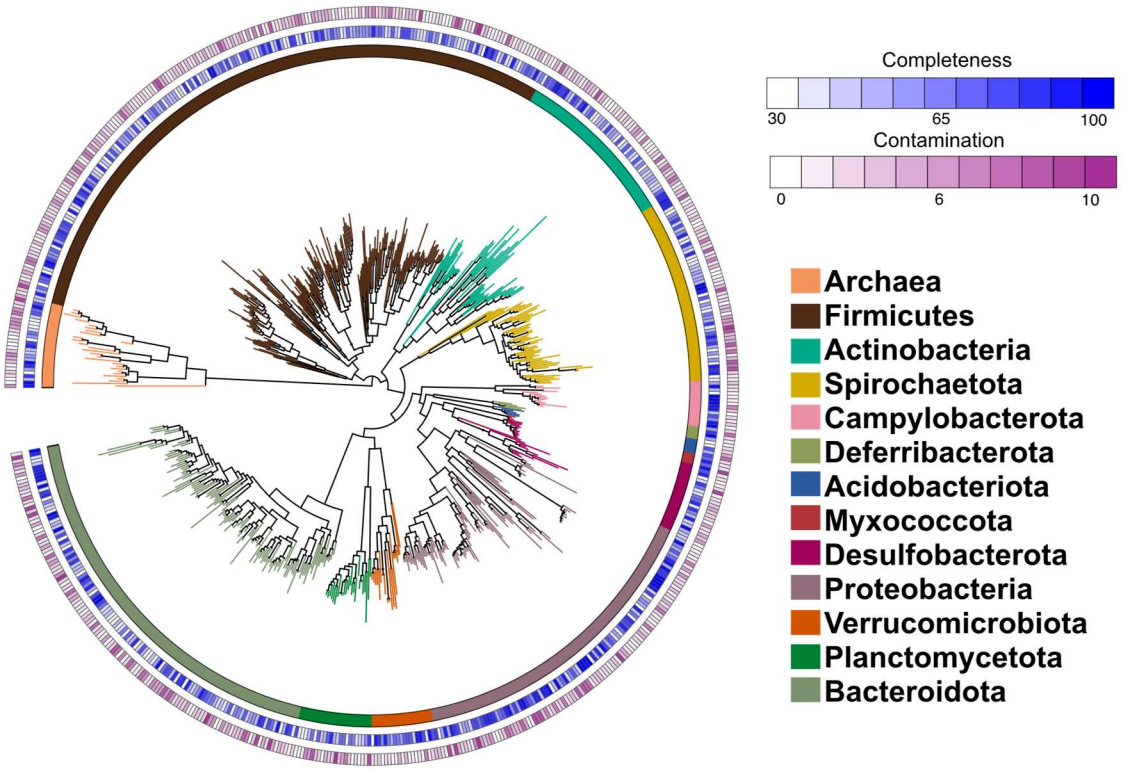

Figure S1. 3. Maximum likelihood phylogenetic tree inferred from 43 single-copy marker genes of 654 metagenome-assembled genomes (MAGs). The completeness and contamination of MAGs was inferred with CheckM (Park *et al.*, 2015). Detailed information about each MAG is available in Table S1.9.

Figure S1. 4. Protein sequence alignment of predicted uricases from 53 termite transcriptomes previously published in Buček et al. (2019).

Table S1. 1. Termite samples sequenced in the study (Link to complete sheet).

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
Cryp	Cryptocercus kyebangensis	Cryptocercidae		cockroach	Cryptocercidae	31.97526	25.58971
MD_RNA_1	Mastotermes darwiniensis	Mastotermitidae		lower termites (LT)	Mastotermitidae	41.87783	27.96191
HSRNA1	Hodotermopsis sjostedti	Archotermopsidae		lower termites (LT)	Archotermopsidae + Hodotermitidae	35.18442	20.56014
SA16-13	Hodotermes mossambicus	Hodotermitidae		lower termites (LT)	Archotermopsidae + Hodotermitidae	85.40099	9.85826
US17	Zootermopsis nevadensis	Archotermopsidae		lower termites (LT)	Archotermopsidae + Hodotermitidae	88.71237	62.3424
AUST14-12	Stolotermes victoriensis	Stolotermitidae		lower termites (LT)	Stolotermitidae	66.18394	12.20878
POROTERMES52	Porotermes planiceps	Stolotermitidae		lower termites (LT)	Stolotermitidae	64.873	8.979102
PORO-CHILI	Porotermes quadricollis	Stolotermitidae		lower termites (LT)	Stolotermitidae	81.91489	14.85188
AUS89	Kalotermes sp.	Kalotermitidae		lower termites (LT)	Kalotermitidae	78.7156	22.53374
AUS109	Glyptotermes sp.	Kalotermitidae		lower termites (LT)	Kalotermitidae	70.68229	14.17981
THAI112	Glyptotermes sp. 4	Kalotermitidae		lower termites (LT)	Kalotermitidae	86.74643	11.69556
H2	Glyptotermes sp.	Kalotermitidae		lower termites (LT)	Kalotermitidae	71.47136	42.25133
THAI31	Glyptotermes sp. 3	Kalotermitidae		lower termites (LT)	Kalotermitidae	60.13413	25.90219
THAI114	Glyptotermes sp. 1	Kalotermitidae		lower termites (LT)	Kalotermitidae	71.59612	9.308026

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
G678_2	Rugitermes sp. A	Kalotermitidae		lower termites (LT)	Kalotermitidae	80.43422	16.15866
SING74	Neotermes sp. A	Kalotermitidae		lower termites (LT)	Kalotermitidae	60.72135	18.0381
AUS102	Neotermes sp.	Kalotermitidae		lower termites (LT)	Kalotermitidae	69.66017	16.0229
AUS91	Neotermes sp.	Kalotermitidae		lower termites (LT)	Kalotermitidae	77.29557	21.18377
KE15-30	Bifiditermes sp.	Kalotermitidae		lower termites (LT)	Kalotermitidae	75.75747	5.475805
M16	Tauritermes sp.	Kalotermitidae		lower termites (LT)	Kalotermitidae	64.2507	20.64455
AUS111	Incisitermes nr. barretti	Kalotermitidae		lower termites (LT)	Kalotermitidae	70.97918	11.96546
US10	Incisitermes snyderi	Kalotermitidae		lower termites (LT)	Kalotermitidae	91.2567	9.651034
MAL39	Cryptotermes domesticus	Kalotermitidae		lower termites (LT)	Kalotermitidae	84.42584	49.52997
AUS110	Cryptotermes sp.	Kalotermitidae		lower termites (LT)	Kalotermitidae	67.37727	14.15388
AUS117	Cryptotermes sp.	Kalotermitidae		lower termites (LT)	Kalotermitidae	84.46147	17.64912
Chi15_131	Styloterms sp.	Stylotermitidae		lower termites (LT)	Stylotermitidae	80.53662	9.075423
BRA31	Serritermes serrifer	Serritermitidae		lower termites (LT)	Serritermitidae	58.0141	8.317455
G13-144	Glossotermes oculatus	Serritermitidae		lower termites (LT)	Serritermitidae	76.42737	12.96352
NG30	Parrhinotermes browni	Rhinotermitidae	Rhinotermitinae	lower termites (LT)	Rhinotermitinae	87.22521	30.29251
THAI23	Parrhinotermes sp. A	Rhinotermitidae	Rhinotermitinae	lower termites (LT)	Rhinotermitinae	67.77182	11.03226
G13-54	Dolichorhinotermes longilabius	Rhinotermitidae	Rhinotermitinae	lower termites (LT)	Rhinotermitinae	88.31052	19.84806
RDCT112	Schedorhinotermes lamanianus	Rhinotermitidae	Rhinotermitinae	lower termites (LT)	Rhinotermitinae	87.96551	17.53398

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
THAI63	Schedorhinotermes sp. 3	Rhinotermitidae	Rhinotermitinae	lower termites (LT)	Rhinotermitinae	96.66679	6.981693
TBRU2.3A	Schedorhinotermes sarawakensis	Rhinotermitidae	Rhinotermitinae	lower termites (LT)	Rhinotermitinae	71.58745	12.02827
NG84	Termitogeton planus	Rhinotermitidae	Termitigetoninae	lower termites (LT)	Termitigetoninae + Prorhinotermitinae	48.37778	6.832767
NG90	Prorhinotermes inopinatus	Rhinotermitidae	Prorhinotermitinae	lower termites (LT)	Termitigetoninae + Prorhinotermitinae	90.96845	21.2302
THAI98	Reticulitermes sp. A	Rhinotermitidae	Heterotermitinae	lower termites (LT)	Reticulitermes	73.94166	5.586651
US1	Reticulitermes nelsonae	Rhinotermitidae	Heterotermitinae	lower termites (LT)	Reticulitermes	84.2058	24.54687
AUS88	Heterotermes cf. paradoxus	Rhinotermitidae	Heterotermitinae	lower termites (LT)	Heterotermitinae + Coptotermitinae	87.94349	6.370943
AUS121	Heterotermes cf. paradoxus	Rhinotermitidae	Heterotermitinae	lower termites (LT)	Heterotermitinae + Coptotermitinae	82.19711	13.92354
AUS47	Heterotermes vagus	Rhinotermitidae	Heterotermitinae	lower termites (LT)	Heterotermitinae + Coptotermitinae	80.05473	6.092687
TBRU8.25	Heterotermes tenuior	Rhinotermitidae	Heterotermitinae	lower termites (LT)	Heterotermitinae + Coptotermitinae	80.95694	8.955118
CF_RNA_1	Coptotermes formosanus	Rhinotermitidae	Coptotermitinae	lower termites (LT)	Heterotermitinae + Coptotermitinae	47.10889	23.15731
NG87	Coptotermes elisae	Rhinotermitidae	Coptotermitinae	lower termites (LT)	Heterotermitinae + Coptotermitinae	77.04872	15.14363
RDCT185	Coptotermes	Rhinotermitidae	Coptotermitinae	lower termites (LT)	Heterotermitinae +	63.65561	32.46552

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
					Coptotermitinae		
G13-107	Coptotermes testaceus	Rhinotermitidae	Coptotermitinae	lower termites (LT)	Heterotermitinae + Coptotermitinae	66.78495	20.209
RDCT165	Sphaerotermes sphaerothorax	Termitidae	Sphaerotermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Sphaerotermitinae	75.47496	64.87452
CAM16-02a	Acanthotermes acanthothorax	Termitidae	Macrotermitinae	fungal-cultivating termites (FC)	Macrotermitinae	67.81468	4.835535
RDCT109	Pseudacanthotermes militaris	Termitidae	Macrotermitinae	fungal-cultivating termites (FC)	Macrotermitinae	67.97339	16.11661
BDIT072	Macrotermes subhyalinus	Termitidae	Macrotermitinae	fungal-cultivating termites (FC)	Macrotermitinae	50.39735	24.74305
THAI50	Macrotermes gilvus	Termitidae	Macrotermitinae	fungal-cultivating termites (FC)	Macrotermitinae	65.89293	28.97269
THAI071	Macrotermes annandalei	Termitidae	Macrotermitinae	fungal-cultivating termites (FC)	Macrotermitinae	59.56507	36.1678
SAF5	Allodontotermes schultzei	Termitidae	Macrotermitinae	fungal-cultivating termites (FC)	Macrotermitinae	82.54487	23.84826
RDCT144	Odontotermes sp. D	Termitidae	Macrotermitinae	fungal-cultivating termites (FC)	Macrotermitinae	70.71704	22.01581
BDIT086	Odontotermes sp.	Termitidae	Macrotermitinae	fungal-cultivating termites (FC)	Macrotermitinae	83.22163	54.61164
THAI064	Odontotermes javanicus	Termitidae	Macrotermitinae	fungal-cultivating termites (FC)	Macrotermitinae	72.91799	22.15938
TBRU8.14E	Labritermes buttelreepeni	Termitidae	Foraminitermitinae	soil-feeding termites (SF)	Foraminitermitinae	53.98647	10.09766

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
BDIT108_2	Foraminitermes rhinoceros	Termitidae	Foraminitermitinae	soil-feeding termites (SF)	Foraminitermitinae	57.81105	36.34262
CAM16-13	Labidotermes celesi	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	41.8243	4.850426
CAM16-05	Phoxotermes cerberus	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	45.66922	5.044952
CAM212	Heimitermes laticeps	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	29.92338	10.41394
BDIT062	Acholotermes chirotus	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	37.03967	5.893563
BDIT49	Aderitotermes sp. 2	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	46.10138	4.117369
RDCT070	Ateuchotermes retifaciens	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	58.57889	6.103523
RDCT098	Amalotermes phaeocephalus	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	77.63623	8.990482
RDCT021	Acidnotermes praus	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	43.59681	6.358763
BDIT112	Astalotermes murcus	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	53.67653	8.207886
SAF6	Alyscotermes sp.	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	46.87505	2.931437
BDIT061	Alyscotermes kilimandjaricus	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	55.24161	4.646086
T4.14A	Euhamitermes hamatus	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	71.00791	16.00352
G13-17	Anoplotermes-group sp. N	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	50.98634	7.130324
G13-65	Anoplotermes-group sp. Q	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	42.32126	12.12778
G756	Patawatermes	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	40.78011	6.183912

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
	nigripunctatus						
G13-69	Anoplotermes-group sp. AF	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	39.34427	4.111358
G13-04	Anoplotermes parvus	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	52.16008	9.374422
G13-32	Anoplotermes banksi	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	61.05825	17.73347
G13-08	Anoplotermes janus	Termitidae	Apicotermitinae	soil-feeding termites (SF)	Apicotermitinae	47.8667	6.925519
TBRU7.11D	Orientotermes emersoni	Termitidae	Termitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Cephalotermes-group	59.6076	19.13339
G13-24	Cylindrotermes parvignathus	Termitidae	Termitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Cephalotermes-group	72.60047	33.8204
RDCT180	Cephalotermes rectangularis	Termitidae	Termitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Cephalotermes-group	62.3229	28.24273
CIVT017	Mimeutermes sorex	Termitidae	Nasutitermitinae	soil-feeding termites (SF)	Nasutitermitinae	56.88422	12.22632
CAM16_18	Leptomixotermes doriae	Termitidae	Nasutitermitinae	soil-feeding termites (SF)	Nasutitermitinae	36.48502	10.39486
BRA1	Constrictotermes cyphergaster	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	63.3244	34.73879
G13-48	Constrictotermes cavifrons	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding	Nasutitermitinae	52.26081	19.08209

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
				Termitidae (WF)			
G728	Agnathotermes crassinasus	Termitidae	Nasutitermitinae	soil-feeding termites (SF)	Nasutitermitinae	50.87874	21.34311
G729	Araujotermes parvellus	Termitidae	Nasutitermitinae	soil-feeding termites (SF)	Nasutitermitinae	71.46997	4.842118
G13-30	Coatitermes kartaboensis	Termitidae	Nasutitermitinae	soil-feeding termites (SF)	Nasutitermitinae	53.29051	21.84086
G13-130	Coatitermes sp. 2	Termitidae	Nasutitermitinae	soil-feeding termites (SF)	Nasutitermitinae	56.18332	13.77836
AUS49	Tumulitermes sp.	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding (WF)	Nasutitermitinae	54.59143	20.89804
NSW6	Occasitermes occasus	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	57.94807	36.21402
THAI100	Bulbitermes nr. laticephalus	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	67.45778	48.04222
THAI43	Nasutitermes sp. 3	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	69.49039	51.67027
THAI45	Oriensubulitermes inanis	Termitidae	Nasutitermitinae	soil-feeding termites (SF)	Nasutitermitinae	54.6895	11.51772
THAI067	Hospitalitermes sp. C	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	61.03721	16.7846

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
BDIT041	Nasutitermes arborum	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	73.41439	55.59922
KE15-44	Trinervitermes gratiosus	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	81.1271	44.00075
BDIT094	Trinervitermes sp.	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	83.52211	38.71172
RDCT106	Nasutitermes lujae	Termitidae	Nasutitermitinae	non-Macrotermitinae (WF)	Nasutitermitinae	60.63951	18.02744
BRU6	Nasutitermes matangensis	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	62.08778	39.18875
NG60	Nasutitermes gracilirostris	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	67.25998	47.02199
NG69	Nasutitermes bikpelanus	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	68.59457	39.54295
G733	Nasutitermes macrocephalus	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	68.00779	44.4868

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
AUS54	Nasutitermes triodiae	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	64.14361	37.9091
AUS62	Nasutitermes graveolus	Termitidae	Nasutitermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Nasutitermitinae	65.09912	37.34616
G13-60	Neocapritermes taracua	Termitidae	Termitinae	soil-feeding termites (SF)	Neocapritermes-group	48.38753	7.374561
G13-28	Planicapritermes planiceps	Termitidae	Termitinae	soil-feeding termites (SF)	Neocapritermes-group	72.11616	42.85146
G683	Neocapritermes sp. H	Termitidae	Termitinae	soil-feeding termites (SF)	Neocapritermes-group	72.2035	13.92595
NG81	Microcerotermes papuanus	Termitidae	Termitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Microcerotermes	67.05145	48.77541
AUS13	Microcerotermes sp.	Termitidae	Termitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Microcerotermes	70.20581	24.08388
BDIT102	Microcerotermes fuscotibialis	Termitidae	Termitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Microcerotermes	62.3123	34.04349
Msp_RNA_1	Microcerotermes sp.	Termitidae	Termitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Microcerotermes	77.66529	55.08129

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
FG-ND02-38	Microcerotermes sp. SA	Termitidae	Termitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Microcerotermes	80.91325	62.13664
G13-23	Embiratermes brevinasus	Termitidae	Syntermitinae	soil-feeding termites (SF)	Syntermitinae	42.88905	5.606929
G13-45	Cyrilliotermes angulariceps	Termitidae	Syntermitinae	soil-feeding termites (SF)	Syntermitinae	26.17747	5.38118
BRA14	Cyrilliotermes sp.	Termitidae	Syntermitinae	soil-feeding termites (SF)	Syntermitinae	55.34307	7.815018
BRA11_2	Syntermes grandis	Termitidae	Syntermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Syntermitinae	58.33776	26.67263
BRA9	Rhynchotermes nasutissimus	Termitidae	Syntermitinae	non-Macrotermitinae (WF)	Syntermitinae	53.76813	13.98033
G13_62	Cornitermes sp. A	Termitidae	Syntermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Syntermitinae	47.55362	28.52899
BRA3	Cornitermes cumulans	Termitidae	Syntermitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Syntermitinae	57.35727	20.46551
BRA5	Silvestritermes heyeri	Termitidae	Syntermitinae	soil-feeding termites (SF)	Syntermitinae	49.05699	7.04267
BRA29	Labiotermes sp.	Termitidae	Syntermitinae	soil-feeding termites (SF)	Syntermitinae	40.54052	10.27183
G13-43	Labiotermes labralis	Termitidae	Syntermitinae	soil-feeding termites (SF)	Syntermitinae	29.90363	8.713193
RD1T21-M1e	Promirotermes pygmaeus	Termitidae	Termitinae	soil-feeding termites (SF)	Promiroterm es	69.41176	21.61321

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
RDCT105	Ophiotermes grandilabius	Termitidae	Cubitermitinae	soil-feeding termites (SF)	Cubitermitinae	39.59984	8.179341
BDIT069	Cubitermes nr. fulvus	Termitidae	Cubitermitinae	soil-feeding termites (SF)	Cubitermitinae	30.1377	11.43037
RDCT051	Orthotermes depressifrons	Termitidae	Cubitermitinae	soil-feeding termites (SF)	Cubitermitinae	36.37838	4.885623
BDIT43	Basidentitermes aurivillii	Termitidae	Cubitermitinae	soil-feeding termites (SF)	Cubitermitinae	18.54868	3.735083
RDCT159	Proboscitermes tubuliferus	Termitidae	Cubitermitinae	soil-feeding termites (SF)	Cubitermitinae	36.81179	14.6745
RDCT125	Tuberculitermes bycanistes	Termitidae	Termitinae	soil-feeding termites (SF)	Termes-group	53.27242	10.21767
G13-112	Cavitermes tuberosus	Termitidae	Termitinae	soil-feeding termites (SF)	Termes-group	53.06879	8.631003
G13-105	Termes fatalis	Termitidae	Termitinae	soil-feeding termites (SF)	Termes-group	50.75059	10.58474
NG49	Protocapritermes odontomachus	Termitidae	Termitinae	soil-feeding termites (SF)	Termes-group	47.03072	24.54328
THAI096	Termes propinquus	Termitidae	Termitinae	soil-feeding termites (SF)	Termes-group	45.89494	16.63874
TBRU5.14A	Prohamitermes mirabilis	Termitidae	Termitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Amitermes-group	34.06541	9.208603
THAI49	Amitermes dentalus	Termitidae	Termitinae	non-Macrotermitinae wood-feeding Termitidae (WF)	Amitermes-group	43.49859	14.44067
AUS4	Amitermes	Termitidae	Termitinae	non-Macrotermit	Amitermes-group	73.60822	39.05673

Sampleids	Species name	Family	Subfamily	Termite group	Termite lineage	Percent all mapped reads	Percent of microbial mapped reads
	meridionalis			inae wood-feeding Termitidae (WF)			
G697	Orthognathotermes aduncus	Termitidae	Termitinae	soil-feeding termites (SF)	Amitermes-group	49.29855	29.98896
G730	Dentispicotermes brevicarinatus	Termitidae	Termitinae	soil-feeding termites (SF)	Amitermes-group	45.85359	14.32617
NG55	Pericapritermes sp. B	Termitidae	Termitinae	soil-feeding termites (SF)	Pericapritermes-group	39.17136	6.041115
NG45	Pericapritermes parvus	Termitidae	Termitinae	soil-feeding termites (SF)	Pericapritermes-group	39.28964	8.597467
NG45_10	Pericapritermes parvus	Termitidae	Termitinae	soil-feeding termites (SF)	Pericapritermes-group	57.44647	12.67198
SING57	Dicuspiditermes nemorosus	Termitidae	Termitinae	soil-feeding termites (SF)	Pericapritermes-group	36.25868	18.94728
THAI038	Mirocapritermes sp. 1	Termitidae	Termitinae	soil-feeding termites (SF)	Pericapritermes-group	39.9187	6.190228
THAI037	Procapritermes sp. 1	Termitidae	Termitinae	soil-feeding termites (SF)	Pericapritermes-group	52.62302	9.785872
SP1	Sinocapritermes mushae	Termitidae	Termitinae	soil-feeding termites (SF)	Pericapritermes-group	53.69253	10.88233
THAI105	Sinocapritermes sp. 1	Termitidae	Termitinae	soil-feeding termites (SF)	Pericapritermes-group	40.91531	6.529427

Table S1. 2. Relative abundance of family-level prokaryotic taxa inferred from gut metagenome and 16S rRNA amplicon data of 74 termite samples. The prokaryotic taxonomy was determined with GTDB for marker genes and with SILVA for 16S rRNA data. The relative abundance was clr-transformed to account for differences in sequencing method and sequencing depth among metagenome samples. (link)

Table S1. 3. Taxonomic distribution of major bacterial and archaeal groups based on relative abundance of 40 single-copy marker genes. We analyzed the marker genes present

in contigs longer than 1000 bps in >5% of gut metagenomes. The relative abundance is represented as transcripts per million (TPM). **(**link**)**

Table S1. 4. Moran's I phylogenetic autocorrelation index calculated for 123 prokaryote families. Significance was assessed with 9999 random permutations. P-values <0.05 are indicated by asterisks. (link)

Table S1. 5. Relative abundance of microbial CAZymes in gut metagenomes with upward of 10000 contigs longer than 1000 bps. Relative abundance is given as transcripts per million (TPM). (link)

Table S1. 6. Moran's I phylogenetic autocorrelation index calculated for 211 prokaryotic CAZymes present in more than 10% of gut metagenomes. Significance was assessed with 9999 random permutations. P-values <0.05 are indicated by asterisks. (link)

Table S1. 7. Phylogenetic ANOVA calculated for 211 prokaryotic CAZymes present in more than 10% of gut metagenomes. Significance was assessed with 9999 random permutations. P-values of phylogenetic ANOVA and pairwise comparisons were adjusted at 5% false discovery rate (FDR). The relative abundance of each CAZyme for the four termite groups are indicated by mean TPM values. Significance of pairwise comparisons between termite groups are indicated by asterisks (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$). (link)

Table S1. 8. Phylogenetic ANOVA comparing the taxonomic origin of the 19 prokaryotic CAZymes found in 10% of gut metagenomes and embedded in contigs longer than 5000 bps. Significance was assessed with 9999 random permutations. The relative abundance of each CAZyme for the four termite groups are indicated by mean TPM values. Significance of pairwise comparisons between termite groups are indicated by asterisks (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$). (link)

Table S1. 9. Information about the 654 MAGs reconstructed in this study. (link)

Table S1. 10. Distribution of polysaccharide utilization loci (PULs) across the MAGs. PULs with at least one GH and Bacteroidota PULs with at least one susCD complex are shown. MAGs containing PULs with all the components are highlighted in grey. (link)

Table S1. 11. Moran's I phylogenetic autocorrelation index and phylogenetic ANOVA performed on the genes involved in the final steps of the lignocellulose digestion in the gut of termites and *Cryptocercus*. For genes composed of multiple subunits, all subunits were summed together. Significance was assessed with 9999 random permutations. P-values were adjusted at 5% false discovery rate (FDR). The relative abundance of each gene for the four termite groups are indicated by mean TPM values. Significance of pairwise comparisons between termite groups are indicated by asterisks (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$). (link)

Table S1. 12. Distribution of genes involved in reductive acetogenesis among MAGs. Distribution is shown as presence (1) and absence (0). Asterisks indicate genes that were annotated using BLASTx search against the AnnoTree database (perc. identity >60%, align. length >100 aa). Other genes were annotated using HMM search against the KEGG or Pfam databases. [FeFe] hydrogenase GroupA4 were annotated using the Hyddb webtool followed by manual inspection of the conserved motifs. The total number of HycB3 (PF13247) found in

each MAG is shown. MAGs with almost complete reductive acetogenesis pathway (>5 genes) and HDCR complex are highlighted in grey. (link)

Table S1. 13. Relative abundance of methyl-coenzyme M reductase (mcrABG) gene complex present in metagenome contigs longer than 5000 bps. Contigs were annotated using BLASTx search against the GTDB database. Relative abundance of the gene family is shown as raw TPM. (link)

Table S1. 14. Distribution of genes involved in methanogenesis among MAGs. Distribution is shown as presence (1) and absence (0). Asterisks indicate genes that were annotated using BLASTx search against the AnnoTree database (perc. identity >60%, align. length >100 aa). Other genes were annotated using HMM search against the KEGG or Pfam databases. Highlighted MAGs have a complete Methanogenesis pathway. (link)

Table S1. 15. Distribution of genes involved in sulfate reducing among MAGs. Distribution is shown as presence (1) and absence (0). Asterisks indicate genes that were annotated using BLASTx search against the AnnoTree database (perc. identity >60%, align. length >100 aa). MAGs with complete sulfate reducing pathway are highlighted. (link)

Table S1. 16. Genes involved in nitrogen metabolism and fixation found in our MAGs. Distribution is shown as presence (1) and absence (0). Asterisks indicate genes that were annotated using BLASTx search against the AnnoTree database (perc. identity >60%, align. length >100 aa). MAGs with complete nitrogen fixation or dissimilatory nitrate reduction pathways are highlighted. (link)

Table S1. 17. Contigs endowed with a NifHDKENB (nifHDKENB, vnfHDKENB, or anfHDKENB) gene complex found in gut metagenomes. The relative abundance is given as raw TPM. (link)

Table S1. 18. Contigs endowed with a NifHDK (nifHDK, vnfHDK, or anfHDK) gene complex found in termite gut metagenomes. The relative abundance is given as raw TPM. (link)

Table S1. 19. Fossil calibrations used to calibrate the time-calibrated tree of termites and Cryptocercus. (link)

Chapter two-Evidence of coevolution between termites and their gut bacteria at geological time scale

Main text

Symbiotic associations with microbes are pervasive across the animal tree of life (McFall-Ngai et al. 2013). Some of these associations, between coevolving mutually dependent partners, have lasted over extended evolutionary timescales. Animals host most of their symbiotic microbes in their gut, and some lineages of mammalian gut microbes may have coevolved with their hosts on time scale of several million years (Moeller et al. 2016). Although the timescale of these symbiotic associations allows for co-speciation to occur, they are short lived in comparison to the partnerships some insects have established with maternally transmitted intracellular bacterial endosymbionts that last over tens, or even hundreds, of millions of years (Moran et al. 2008). There are no clear examples of animals coevolving with their gut microbes through vertical transmission across host generations on such timescales.

Gut bacterial communities are often positively correlated with the phylogenetic tree of their hosts (Lim and Bordenstein 2020). However, this pattern is not necessarily indicative of coevolution and can be generated by ecological filtering imposed by host traits such as diet and gut pH (Mazel et al. 2018; Lim and Bordenstein 2020). Cophylogenetic patterns, whereby the trees of two interacting partners show congruence in term of topology and timing, provide stronger support for coevolution (de Vienne et al. 2013; Groussin et al. 2020). For example, maternally transmitted intracellular bacterial endosymbionts and their insect hosts have closely matching phylogenetic trees, in line with their long-term coevolution at geological timescales (e.g. (Jousselin et al. 2008; Kinjo et al. 2021). Cophylogenetic analyses have rarely been performed for animals and their gut microbes, perhaps because many studies have relied upon 16S rRNA sequences, a marker that diverges at a rate of about 1% per 50 million years (Ochman et al. 1999) and does not provide the taxonomic resolution required to resolve coevolutionary patterns between animals and their gut microbes. The limitations of the 16S rRNA gene can be overcome by using protein-coding marker genes obtained from gut metagenomes.

We analyzed the coevolutionary patterns between termites and their gut microbes. Termites host unique gut microbial communities composed of archaea, bacteria, and flagellates, the latter of which was lost in the Termitidae (Brune 2014). Some lineages of termite gut bacteria are ubiquitous in all termite species but are not found outside termite guts (Bourguignon et al. 2018). This raises the possibility that some lineages of termite gut bacteria were already associated with the ancestor of modern termites ~150 million years ago (Bourguignon et al. 2015; Bucek et al. 2019) and were since vertically transmitted across generations of termite hosts through exchange of fecal fluid, a phenomenon called proctodeal trophallaxis (Nalepa et al. 2001). To identify cases of long-term coevolution with vertical transmission, we searched for cophylogenetic patterns between termites and their gut bacteria. We compared a mitochondrial genome phylogenetic tree of termites with phylogenetic trees of gut bacteria reconstructed using ten independent universally occurring protein-coding marker genes (Sunagawa et al. 2013). The sequences were derived from 202 termite gut metagenomes that we combined with sequences from the GTDB database (Parks et al. 2020). Our dataset included representatives of the nine termite families, the eight subfamilies of Termitidae, and 60 samples of 27 species of *Microcerotermes*, a pantropical termitid genus that appeared ~20 million years ago (Bourguignon et al. 2017). Therefore, our dataset captured both intraspecific variations and ancient divergences of the termite hosts.

We built one Maximum Likelihood phylogenetic tree for every prokaryotic phylum and every marker gene and searched for termite-specific clusters that comprised sequences derived from upward of ten termite species and included no sequences from non-termite environments. The highest number of termite-specific clusters was found for the signal recognition protein FtsY (COG0552), the marker gene we used as a reference. In total, we identified 30 termite-specific clusters belonging to 13 prokaryotic phyla from the phylogenetic trees reconstructed with COG0552. We examined the cophylogenetic signal between each termite-specific cluster and their termite hosts using three different methods: PACo (Balbuena et al. 2013), generalized Robinson Foulds metric (Smith 2020), and the algorithm described by Nye et al. (2005). 20 out of 30 clusters showed strong cophylogenetic signal with all three methods (Table 2.1). Cophylogenetic analyses performed on the other nine marker genes representing the same termite-specific clusters of prokaryotes were highly congruent (Table 2.1), indicating that the choice of marker gene did not influence the outcome of our analyses. Prokaryotic lineages showing strong cophylogenetic signals included key components of the gut microbiota of termites such as the Spirochaetota family *Treponemataceae B*, the Fibrobacterota genus *Fibromonas*, and the Firmicutes family Ruminococcaceae (Dietrich et al. 2014; Mikaelyan et al. 2015a; Bourguignon et al. 2018). These results suggest that lineages of prokaryotes exclusively found in termite guts coevolve with their termite hosts.

Because significant cophylogenetic signals do not necessarily imply coevolution (de Vienne et al. 2013; Groussin et al. 2020), we searched our phylogenetic trees for additional evidence of coevolution. The existence of termite-specific clusters within prokaryote phylogenetic trees can be explained by horizontal transfers of gut prokaryotes among termite species and/or by vertical transfers of gut prokaryotes over millions of generations of termite hosts. We reasoned that the dominance of vertical transfers over horizontal transfers should lead to the emergence of lineages of prokaryotes only found in the guts of specific termite clades. We indeed found such termite-clade-specific lineages within our prokaryote phylogenetic trees (Figure 2.1). For example, we found prokaryote clades specific to *Microcerotermes*, the genus we sampled the most intensively, belonging to the Spirochaetota family *Treponemataceae B*, the Fibrobacterota genus *Fibromonas*, and the Desulfobacterota genus *Adiutrix* (Figure 2.1). The members of these prokaryotic clades were only found in *Microcerotermes* species including those collected in the same localities. They were found in the guts of *Microcerotermes* species collected across four continents and six biogeographic realms, indicating that species of *Microcerotermes* spread across the tropical regions of the world over the last 20 million years (Bourguignon et al. 2017) together with their specific gut prokaryotes. This close association between termites and their gut prokaryotes was not unique to *Microcerotermes* and was also found in other termite clades sampled less intensively. For example, we found that a group of Nasutitermitinae, including species that diverged ~25 million years ago (Bourguignon et al. 2017), hosted several lineages of the Spirochaetota family *Treponemataceae B* and Desulfobacterota genus *Adiutrix* specific to these termite species. Similarly, several lineages of the Spirochaetota family *Treponemataceae B* were unique to a group of Kalotermitidae that included species sharing a common ancestor >50 million years ago (Buček et al. 2021). These examples of extreme specificity between termite clades and some of their gut prokaryotes highlight the absence of horizontal transfers between termite species belonging to different clades. Therefore, gut prokaryotes specific to a termite clade are transmitted vertically from parents to offspring and/or horizontally among species of the termite clade.

A close examination of the prokaryote clades specific to *Microcerotermes* revealed patterns of cophylogeny, often repeated several times over the clade phylogenetic trees, albeit with missing termite representatives (Figure 2.2). These patterns were most evident for the phylogenetic trees of Spirochaetota family *Treponemataceae B* and the Fibrobacterota group annotated as *Chitinivibrionales*, two phyla of bacteria making up over half of the bacterial abundance in the gut of *Microcerotermes* (Bourguignon et al. 2018). They are indicative of symbiont extinction (or insufficient sequencing depth) and speciation taking place within the termite hosts, two processes expected to blur the cophylogenetic patterns (Groussin et al. 2020). Several termite-specific bacterial lineages, less speciose than *Treponemataceae B* and Fibrobacterota, depicted much clearer patterns of cophylogeny. For example, the phylogenetic tree of the Desulfobaterota genus *Adiutrix* found in the sister group of termites *Cryptocercus*, in three families of termites, and in six subfamilies of Termitidae present only a few mismatches with the phylogenetic tree of termites (Figure 2.1C). Similarly, the phylogenetic trees of Proteobacteria family *Rhodocyclaceae* (tree 6) and Acidobacteria family *Holophagaceae* (tree 7) showed similar patterns of cospeciation and mirrored remarkably well the phylogenetic trees of termites and Termitidae, respectively. Our results therefore reveals cophylogenetic pattern between termites and some of their gut bacterial symbionts arising from vertical inheritance and cospeciation over several tens of millions of years.

The majority of microbial phyla examined in this study show a significant coevolution pattern with the termite host. Some microbial groups were congruent with specific termite lineages such as specificity of Spirochaetota family *Treponemataceae B* and the Fibrobacterota genus *Fibromonas* for *Microcerotermes* genus. In contrast, other microbial groups like Desulfobaterota genus *Adiutrix* and Proteobacteria family *Rhodocyclaceae* were present across multiple termite lineages. These results support the hypothesis of partner fidelity feedback (Foster and Wenseleers 2006) in which the stable associations are maintained due to co-operative feedback between interacting partners. Gut microbes assist in the host's nutritional needs reciprocating the host that provide a stable gut environment.

Materials and Methods

Sample collection and metagenome analyses

We collected a total of 201 termite samples and one sample of the cockroach *Cryptocercus kyebangensis* (Table S2.1). All samples were preserved in RNA-later® and stored at -80°C until DNA extraction. We performed the DNA extraction and sequencing procedure and assembled the metagenomes as described previously (*Data filtering and assembly of metagenomic reads* sub-section of chapter one)

Sequence data extraction

Ten single-copy protein-coding marker genes (Sorek et al. 2007; Milanese et al. 2019) were extracted from the assemblies using profile hidden markov model (HMM) as implemented in the mOTU software (Sunagawa et al. 2013). Genomes and metagenome-assembled genomes available in GTDB v.95 (Parks et al. 2020) were downloaded, and the same ten single-copy marker genes were extracted as described above.

Taxonomic annotation of the marker genes

The taxonomic annotation of the ten marker genes extracted from termite gut metagenome assemblies was performed with DIAMOND BLASTP (Buchfink et al. 2015) using evalue ≤ 1e-24 and output format 102, which uses the lowest common ancestor algorithm for annotation. The blast search was performed against the GTDB database v.95 (Parks et al. 2020). For

downloaded genomes, we used the taxonomic annotation available from GTDB v.95. The marker gene sequences from termite gut metagenomes and from the GTDB database were analyzed separately for every phylum. We reconstructed the phylogenetic tree of every phylum including more than ten sequences.

Reconstruction of marker gene phylogenetic trees

Sequences shorter than half the mean length of the marker gene were removed to improve the accuracy of phylogenetic reconstructions (Wiens et al. 2003; von Mering et al. 2007). Nucleotide sequences were translated into protein sequences using the command *transeq* of the Emboss v.6.6.0 (Madeira et al. 2019) and the bacterial genetic code table (code 11). Protein sequences were aligned using MAFFT v.7.305 with the *–auto* option (Katoh and Standley 2013). Protein alignments were back-translated into their corresponding nucleotide alignments using PAL2NAL (Suyama et al. 2006). Nucleotide sequences were converted into purines (R) and pyrimidines (Y) using BMGE v.1.12 (Criscuolo and Gribaldo 2010) to account for the variability of GC content observed across bacterial sequences. Maximum likelihood (ML) phylogenetic trees were generated using IQ-TREE v.1.6.12 (Nguyen et al. 2014) with the GTR+I+G model of base substitution. Node supports were assessed using the ultrafast bootstrap method (Minh et al. 2013) with the command *-bb 2000* for 2000 bootstrap replicates. The phylogenetic trees of every phylum were rooted using outgroup taxa determined from the prokaryotic tree of life (Parks et al. 2017; Parks et al. 2020). The phylogenetic trees of prokaryotic clades composed of sequences found exclusively in termite guts and represented by more than ten termite species were extracted from the phylogenetic trees of each phylum. This procedure was followed for each marker gene.

Phylogenetic reconstruction of termites

We reconstructed the phylogenetic tree of termites using mitochondrial genome sequences. Termite mitochondrial contigs were identified using BLAST searches (Altschul et al. 1990). Sequences longer than 5000 bp and more than 90% percent identical to the previously published whole mitochondrial genomes of termites (Bourguignon et al. 2015; Bourguignon et al. 2016; Bourguignon et al. 2017; Wang et al. 2019) were identified. Complete or near-complete mitochondrial genomes were annotated using the MITOS web server (Bernt et al. 2013). We reconstructed a Bayesian phylogenetic tree using BEAST v.2.4.8 (Suchard et al. 2018) following the approach described previously (*Termite phylogenetic tree reconstruction* subsection of chapter one)

Matching termite-specific prokaryote clades across marker gene trees

We found between 8 to 32 termite-specific prokaryote clades per marker gene. The phylogenetic trees reconstructed with the marker gene coding for COG0552 yielded one of the largest number of termite-specific prokaryote clades and the most number of sequences per clade and were used as references. We attempted to link every termite-specific prokaryote clade found in the phylogenetic trees reconstructed with COG0552 with their counterparts found in the phylogenetic trees reconstructed with the other nine other marker genes. To do so, we searched our 201 gut metagenomes for contigs including at least two of the ten marker genes. The position of each marker gene sequence in their respective phylogetic trees was used to match termite-specific prokaryote clades across marker gene trees. We also used the ten marker genes of the termite gut bacterial genomes found in the GTDB database. Out of 194425 genomes downloaded from GTDB database, 37 were associated with termite guts.

Cophylogenetic analyses

We carried out cophylogenetic analyses between the phylogenetic trees of termites and termite-specific prokaryote clades using three approaches. For the first approach, we used the R package PACo (Procrustean Approach to Cophylogeny) (Balbuena et al. 2013) that uses Procrustean superimposition to estimate the cophylogenetic signal between two phylogenies. The host and symbiont phylogenetic trees were converted into distance matrices using the *copheretic()* function of the vegan R package (Oksanen 2014). The software was run using the *backtrack* method of randomization that conserves the overall degree of interactions between the two trees (Hutchinson et al. 2017). The second approach was the generalized Robinson Foulds (RF) metric (Smith 2020). This method was implemented using the *ClusteringInfoDistance()* function of the TreeDist R package (Smith 2020). For the third approach, the host and symbiont phylogenetic trees were matched to find an optimal 1-to-1 map between branches using the method explained by Nye et al. (2006) and implemented in *NyeSimilarity()* function of the TreeDist R package (Smith 2020). Because the two methods implemented in the TreeDist R package do not allow for multiple symbiont tips in one host, each host tip was split into a number of tips of zero branch length equal to the number of prokaryote symbionts present in the metagenome corresponding to that given tip (Perez-Lamarque and Morlon 2019; Satler et al. 2020). Congruence between the host and symbiont trees was determined by significance testing using 10,000 random permutations.

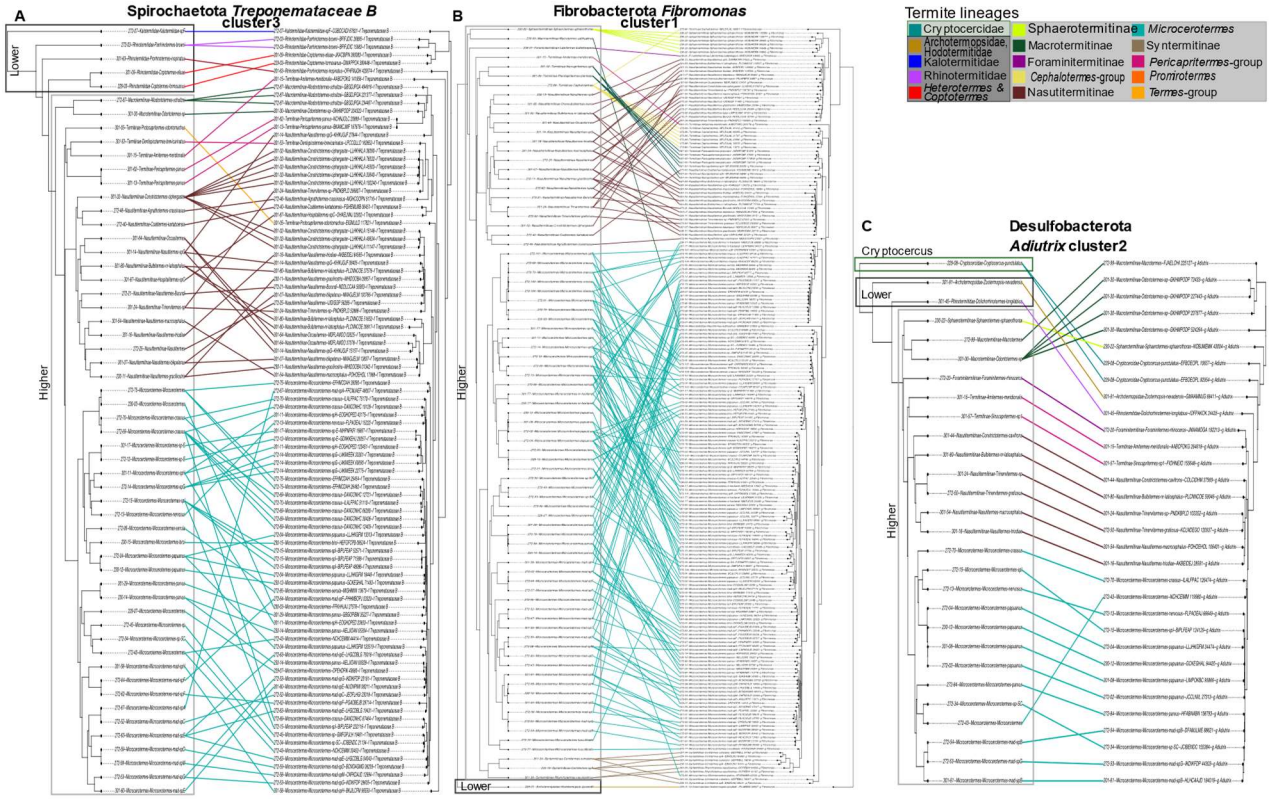

Figure 2. 1. Selected coevolution plots between termite tree and microbial gene trees from COG0552 single copy marker gene. (A) *Treponemataceae B* (see tree 1 in Figure S2.1 for additional details). (B) *Fibromonas* (see tree 2 in Figure S2.1 for additional details) and (C) *Adiutrix* (see tree 3 in Figure S2.1 for additional details).

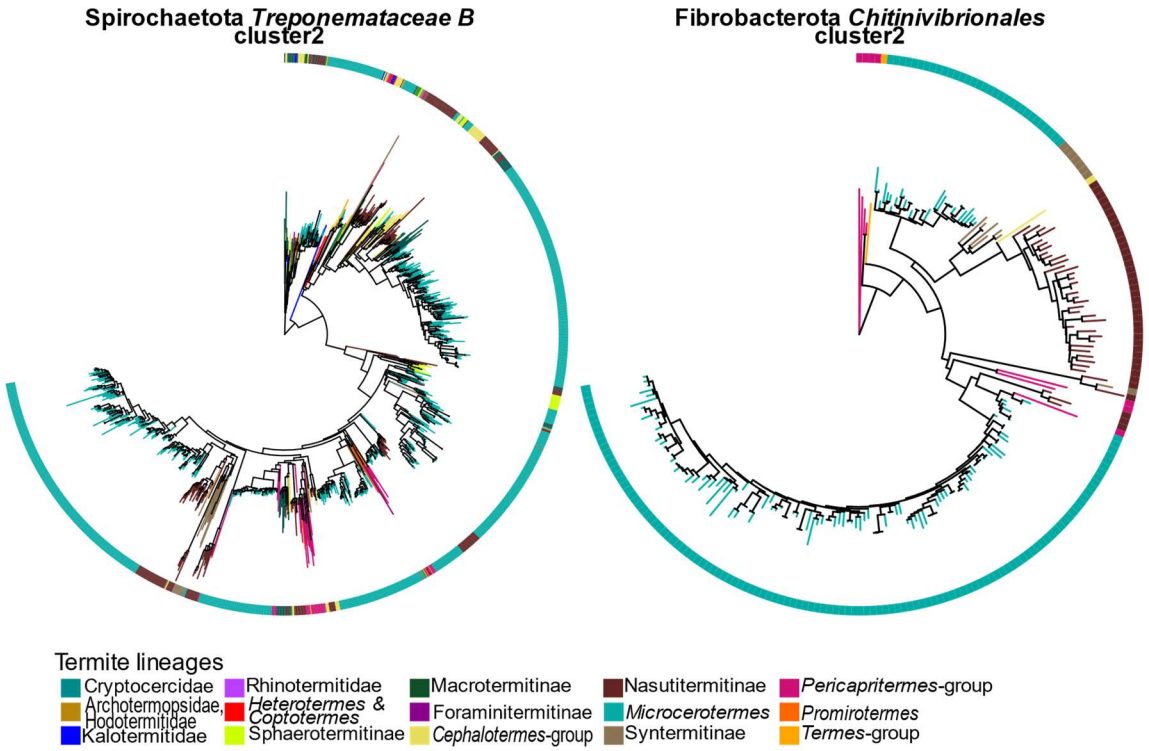

Figure 2. 2. Phylogenetic trees of microbial COG0552 marker gene with multiple clusters of *Microcerotermes*-specific sequences. (A) *Treponemataceae B* (see tree 4 in Figure S2.1 for additional details) and (B) *Chitinivibrionales* (see tree 5 in Figure S2.1 for additional details) showed coevolution with the termite host. The coevolution pattern was repeated multiple times across the tree, specifically for the *Microcerotermes* genus indicative of symbiont extinction and speciation.

Figure S2. 1. Phylogenetic trees of bacterial lineages belonging to COG0552 single-copy marker gene detected across 201 termite samples and one *Cryptocercus* cockroach. The best-hit taxonomy of the gene trees are as follows: *Treponemataceae B* cluster three (tree 1), *Fibromonas* cluster one (tree 2), *Adiutrix* (tree 3), *Treponemataceae B* cluster two (tree 4), *Chitinivibrionales* cluster two (tree 5), *Rhodocyclaceae* cluster one (tree 6) and *Holophagaceae* cluster one (tree 7). (link)

Table 2. 1. Coevolution statistics of termite-specific microbial clusters based on COG0552 as representative marker gene. The tree info column is based on presence of marker genes on the same contig or presence of marker gene belonging to single cell isolates or cultured isolates from publicly available termites (common) or only present in COG0552 trees. P-values of overall coevolution signal of microbial clusters was examined using three algorithms. P-value symbols shown here are based on all three coevolution algorithms having the same symbol. If one algorithm is not significant, then P-value symbol is showed as not-significant. If the symbols are different across the three algorithms, then the least significant P value symbol (eg- * or **) is represented. A putative taxonomic rank based on presence of five or more sequences annotated at that rank in the representative cluster (best hit) is shown.

Tree info	Phyla	Classification	COG0012	COG0016	COG0018	COG0172	COG0215	COG0495	COG0525	COG0533	COG0541	COG0552
comm	Spirochaetota	f__Treponemataceae_B	***	***	***	***	***	**	***	***	*	***
common	Fibrobacterota	g__Fibromonas	***				***		***	***	***	***
common	Firmicutes_A	f__Ruminococcaceae	***						***	***		***
common	Firmicutes_A	f__Oscillospiraceae						***				***
common	Bacteroidota	c__Bacteroidia	***	***	***	***	***	***	***	***	***	***
common	Bacteroidota	o__Bacteroidales								***		*
common	Proteobacteria	f__Rhodocyclaceae	***	***							***	***
common	Fibrobacterota	o__Chitinivibrionales		***						***		***
common	Thermoplasmatota	f__Methanomethylophilaceae		*		***						**
common	Proteobacteria	f__Burkholderiaceae								***		NS
COG0552 tree	Acidobacteriota	f__Holophagaceae										***
COG0552 tree	Desulfobacterota	g__Adiutrix										***

Table S2. 1. Information about the 202 termite gut metagenomes sequenced in this study. (link)

Chapter three-Horizontal transfers and multiple acquisitions drive the gut microbial functional evolution

Introduction

Lignocellulose is the most abundant biomolecule on earth (Swift et al. 1981). It is mainly composed of lignin, cellulose, and hemicellulose, which are essential food sources for some organisms (Baldrian et al. 2012; Allison et al. 2013). Glycosyl hydrolases (GHs) are groups of enzymes that help deconstruct lignocellulose (Berlemont et al. 2009; Nyyssönen et al. 2013) and breakdown chitin (Lindahl and Finlay 2006), thence supporting essential ecosystem processes (Allison et al. 2013; Tauzin et al. 2016). These enzymes are endogenously released in the gut of termites (Tokuda 2019) and are produced by a rich community of gut microbes (Hess et al. 2011; Brune and Dietrich 2015; Tamura et al. 2017). The gut microbes target multiple polysaccharides, digesting and breaking them down into short-chain fatty acids that are used by the host as a source of energy (Graber et al. 2004; Rosenthal et al. 2011). The microbial GH families include novel source of enzymes whose characterization may contribute to the biofuel industry to generate renewable energy (Dodd and Cann 2009; Himmel and Bayer 2009).

Many GH families are secreted inside termite guts. Termites feed on different types of plant matter such as wood, soil, grass, and litter (Donovan et al. 2001; Eggleton and Tayasu 2001). Altogether, termites have been estimated to consume ~3-7 billion tons of lignocellulose annually (Prins and Kreulen 1991; Breznak and Brune 1994) using a combination of endogenous enzymes, secreted in the salivary glands or midgut (Slaytor 2000; Tokuda et al. 2004), and enzymes produce by the microbes they host in their hindgut (Brune and Ohkuma 2011; Brune and Dietrich 2015). Termites have developed essential mutualistic associations with gut microbes since their inception, 150 Mya (Bourguignon et al. 2015; Bucek et al. 2019). The early evolving lineages of termites, called lower termites, harbor gut flagellates that have cellulolytic and hemicellulolytic activities (Tokuda 2019; Nishimura et al. 2020). Termites also established numerous symbiotic associations with gut bacteria and archaea that, alike their gut flagellates, produce enzymes breaking down lignocellulose into simple sugars (Wenzel et al. 2002; Treitli et al. 2019).

Unlike lower termites, the Termitidae (higher termites), lost their gut flagellates and established new ways to digest lignocellulose (Brune 2014). The early diverging higher termite subfamilies Macrotermitinae and Sphaerotermitinae externally cultivate fungi and bacteria in their nests, respectively (Garnier-Sillam et al. 1989; Rouland-Lefèvre et al. 2006). Other higher termite subfamilies developed complete dependence on gut bacteria and archaea for lignocellulose digestion (Brune and Ohkuma 2011; Brune 2014).

Numerous gut microbial GH families have been found in the termite gut via metagenomics and meta-transcriptomics surveys. The expression profile of carbohydrate-active enzymes (CAZymes) suggests that their distribution varies with termite diet (He et al. 2013; Marynowska et al. 2020). These enzymes are often assembled into putative operons, improving carbohydrate-degrading efficiency (Liu et al. 2019). They are distributed across numerous gut microbes and are able to act on multiple substrates simultaneously (Liu et al. 2018; Liu et al. 2019). Metagenomics-based taxonomic analyses have shown that CAZymes are present in the genomes of multiple termite gut microbial phyla such as Spirochaetota, Firmicutes, Fibrobacteriota, and Bacteroidota (Warnecke et al. 2007; He et al. 2013; Tokuda et al. 2018). These analyses have generated a catalog of CAZymes encoded in the termite gut, revealing

functional redundancy (Marynowska et al. 2020) and specialization (Warnecke et al. 2007) of termite gut microbes for carbohydrate metabolism.

The functional redundancy observed in termite gut microbiota has many ecological and evolutionary implications. It provide robustness to the system and prevents the loss of functions essential to the termite hosts (Xu et al. 2007). It also selects for unrelated microbes possessing functionally similar sets of genes (Ley et al. 2006), some of which were acquired via Horizontal Gene Transfers (HGTs). Although HGTs are rampant in animal guts, due to shared resources and close proximity (Shterzer and Mizrahi 2015; Lerner et al. 2017), only a few examples of HGT are available for microbes living inside the gut of termites. HGT may have occurred for genes involved in CO2-reductive acetogenesis, a metabolic process that occurs after the fermentation of cellulose (Breznak and Switzer 1986). One of the key genes, formyl tetrahydrofolate synthase (FTHFS), has possibly been acquired via inter-phyla transfers between Spirochaetota and Firmicutes (Ottesen and Leadbetter 2011). Possible HGT of the gene coding for Glucose-6-phosphate transporter, *uhpC*, has also been suggested via phylogenetic analysis of termite gut microbes. The flagellate endosymbiont *Candidatus* Endomicrobium trichonymphae strain Rs-D17 possibly acquired this gene by HGT from its free-living relative *Endomicrobium proavitum* (Zheng et al. 2017).

The composition and function of termite gut microbiome may also be affected by competition among microbes. Competition may select for microbial phyla specialized for specific functions important for the termite hosts. These phyla could then be maintained across ecological and evolutionary timescales through selection on the host (Xu et al. 2007; Shterzer and Mizrahi 2015). One of the most abundant termite gut phyla, Spirochaetota, was found to be specialized for Methyl Accepting Chemotaxis protein (MCP) family in termites of different diets (He et al. 2013). Spirochaetota might be advantaged in highly viscous hindgut environment by its ability to position itself along the physicochemical gradient to find new substrates (Bellgard et al. 2009).

It is still unclear how CAZymes present in termite guts have been acquired and evolved. To explore this question, we examined, in a phylogenetic framework, (i) the role of independent acquisitions of CAZymes from the environment, (ii) the functional redundancy among termite gut microbes, and (iii) the specialization of gut phyla to specific substrates as suggested by their presence in specific CAZymes. We extracted 10 CAZymes corresponding to seven most abundant GH families and four GH subfamilies from 924 metagenome-assembled genomes (MAGs) reconstructed from the gut metagenomes of 201 termite samples and one *Cryptocercus* sample. We then reconstructed the phylogenetic tree of each GH family including sequences derived from our MAGs and from 2393 publicly available microbial genomes. Using these phylogenetic trees, we characterized the evolution of GH genes across termite and non-termite microbial genomes. Our results suggest that each GH family has an independent evolutionary history. Some are specialized for the termite gut environment, encoded by a single termite gut microbial phylum, while others have been independently acquired from the environment by multiple phyla. Our analyses shed light on the evolutionary history of CAZymes present in termite guts and highlight the role of HGTs in driving termite gut adaptation.

Materials and Method

Metagenome sequencing and analysis

We collected a total of 201 termite samples and one sample of the cockroach *Cryptocercus kyebangensis*. Genomic DNA was extracted from whole guts of five workers using the

NucleoSpin Soil kit (Macherey-Nagel) according to manufacturer's protocol. Two library preparation methods were performed on 173 samples sequenced with Illumina HiSeq4000 using the KAPA Hyperplus kit. In the first method, a unique pair of dual indexes (non-redundant) were used, while in the second method, a unique combination of dual indexes were used. The remaining 29 samples were sequenced once on Illumina HiSeq 2500 with unique pair of dual indexes. The unique dual indexing method uses distinct index sequences for each sample, whereas a combinatorial design of dual indexes was used in the second method. The combinatorial design allows each index to be shared across the lane, but each sample has a unique combination (Sinha et al. 2017; Costello et al. 2018). This method is affected by tag switching. The reads from all the samples and library preparation methods were quality checked, filtered, and assembled as described previously (*Data filtering and assembly of metagenomic reads* sub-section of chapter one).

Reconstruction of Metagenome Assembled Genomes (MAGs)

Metagenome Assembled Genomes (MAGs) were reconstructed from metagenome contigs using CONCOCT v 0.4.0 (Alneberg *et al.*, 2014) as implemented in the metawrap software v 0.9 (Uritskiy et al. 2018) with default parameters. MAG quality check was performed with CheckM v 1.0.11 (Parks *et al.*, 2015) based on 43 single-copy marker genes (Table S3.1). The refineM tool was used to remove contigs with potential contamination based on genome properties such as GC content, tetranucleotide signatures, and coverage using the *outliers* and *filter_bins* commands. The overall MAG taxonomy was compared with the taxonomy of scaffolds containing the 16S rRNA gene using the *taxon_profile* command in refineM. The contaminated contigs were removed using the *taxon_filter* command (Parks et al 2017). A second round of quality check was performed with CheckM software, and MAGs with upward of 30% completeness and less than 10% contamination were retained (Bowers *et al.*, 2017).

Forty single-copy protein-coding marker genes were extracted using profile Hidden Markov Model (HMM) based approach as implemented in the mOTU software (Sunagawa et al. 2013). Maximum Likelihood (ML) trees of the MAGs were built for each marker gene family using FastTree v 2.1.11 (Price et al. 2010), and those that recovered the monophyly of each microbial phylum after manual inspection were retained (Coleman et al. 2021). This resulted in a final set of 924 MAGs belonging to 13 phyla.

Functional annotation

Protein sequences from the MAGs were annotated using Prokka v.1.14.5 (Seemann, 2014) with the *rfam, addgenes,* and *metagenome* parameters. Carbohydrate metabolizing genes were annotated using the CAZy database as a reference (Lombard *et al.*, 2014). The protein sequences were searched against a set of profile HMMs representing CAZy domains deposited in the dbCAN database release 9 (Zhang et al. 2018). We used an e-value below e-30 and coverage greater than or equal to 0.35 as thresholds to extract the best matches.

Ten of the most abundant CAZymes, including 7 GH families and 4 subfamilies, present across multiple host lineages and dietary habits (Table S1.5, S1.6) were selected. Intact carbohydrate degrading CDSs as determined with Pseudofinder v.0.10 (Syberg-Olsen et al. 2021) using standard parameters and GTDB ver95 database as reference (Parks et al 2020) were used for downstream phylogenetic analysis.

Publicly available MAGs

To examine the relationship between the selected CAZymes from the termite gut and those from other environments, we performed a BLASTP search of 10 selected CAZymes against the GTDB ver 95 (Parks et al. 2020) database. Protein sequences belonging to 13 phyla and

corresponding to the top 100 samples based on percent identity, e-value, and sequence length were extracted from each BLAST hit. Using this approach, we reduced the original 19,000 MAGs from the database to 2393. From the 2393 genomes, we only retained those that had all 40 single-copy protein-coding marker gene trees assign to the same phylum. MAGs presenting discrepancies among marker genes were removed. Additionally, we clustered non-termite MAGs at a level of 95% Average Nucleotide Identity (ANI) using the drep package (Olm et al. 2020) with the *ANIm* parameter to remove related taxa from the same environment. The final dataset of non-termite genomes contained 1779 representatives that we used for phylogenetic analyses.

MAG based species tree

Marker-genes were extracted from the final set of 1779 genomes obtained from the GTDB database and from the 924 MAGs reconstructed from our metagenomes. Sequences longer than 50% of the mean marker-gene length were retained. Maximum likelihood (ML) tree was generated for the concatenated 40 marker genes using FASTREE (Price et al. 2010). Termite gut MAGs that formed a cluster of four or more representatives were kept.

Gene trees of selected carbohydrate degrading enzymes

Sequences belonging to the most abundant 10 CAZymes (belonging to 7 families and four sub-families) of the termite gut (Table S1.5, S1.6) were extracted from all MAGs and genomes. Sequences longer than 50% of the mean length of each CAZyme were retained. Maximum Likelihood (ML) trees were built using IQTREE v.1.6.12 (Minh et al. 2020). We reconstructed four trees for each CAZyme family using four different models, namely, (i) CAT model on protein-coding nucleotide sequences without partitions (Price et al. 2010), (ii) GTR+I+G model on protein-coding nucleotide sequences with all codons partitioned (Chernomor et al. 2016; Minh et al. 2020), (iii) GTR+I+G model on protein-coding sequences with two partitions, one for the combined first and second codon positions and one for the third codon positions (Chernomor et al. 2016; Minh et al. 2020), to account for third codon saturation over long co-evolution with the host (Kikuchi et al. 2009) and (iv) GTR+I+G model on protein-coding sequences encoded as purines Rs (A or G) and pyrimidines Ys (T or C) to account for GC bias in endosymbiont genomes (Husník et al. 2011; Lo et al. 2016). RY coding of protein-coding nucleotide sequences was performed using BMGE v.1.12 with default parameters (Criscuolo and Gribaldo 2010). Each gene tree was run with at least 1000 bootstrap replicates and was rooted with randomly selected sequences from one of the selected CAZyme other than the gene of interest, for example, GH10 was rooted with sequences from GH45. The best-supported of the trees was determined with the approximately unbiased test (AU) (Shimodaira 2002) using 100,000 replicates. The phylogenetic trees were visualized using ITOL software (Letunic and Bork 2021) and ggtree package in R (Yu et al. 2017).

Results and Discussion

MAGs from termite whole gut metagenomes

We sequenced 202 whole gut metagenomes from 201 termite samples and one *Cryptocercus* species. A total of 173 metagenomes were sequenced on Illumina HiSeq 4000 twice, once with unique pairs of dual indexes (hereafter called unique dual indexing) and once with a unique combination of dual indexes (hereafter called combinational indexing). The remaining 29 samples were sequenced once on Illumina HiSeq 2500 using unique dual indexing approach. The unique dual indexing approach use distinct, unrelated, dual index sequences. In contrast, the combinational indexing approach uses unique combination of dual indexes but repeat

indexes multiple times (Illumina 2017; Sinha et al. 2017). The combinatorial distribution of indexes makes samples prone to index switching due to the presence of residual excess free primers or adapters that can lead to the sequencing of wrong sample index (Sinha et al. 2017; Costello et al. 2018). To account for sequencing-mediated errors, we added multiple refinement steps to metagenome analysis.

Metagenome Assembled Genomes (MAGs) were extracted from metagenome assemblies using the metaWRAP pipeline (Uritskiy et al. 2018). The MAGs were extracted separately from each sequencing library. The libraries obtained with combinational indexing were also assembled individually, but due to potential index cross-contamination, we cannot exclude the presence of sequences originating from a different termite host in our MAGs. This loss of information is analogous to the co-assembly method commonly used to generate MAGs from many samples (Uritskiy and DiRuggiero 2019; Quince et al. 2021). The co-assembled MAGs are generated by combining contigs from all the samples, generating a collapsed average of multiple strains, capturing the core microbial diversity (Stewart et al. 2018; Uritskiy and DiRuggiero 2019). The MAGs from combinational indexing samples may, in some cases, represent the core microbes across multiple termite samples rather than species-specific diversity.

We generated 1015 MAGs using libraries prepared with the unique dual indexing approach and 656 MAGs from libraries prepared with the combinational indexing approach. To improve the MAG quality and remove spurious contigs, we used a combination of GC content, tetranucleotide signature, and contig coverage information for MAG refinement (Parks et al. 2017). Additionally, chimeric contigs were removed if their taxonomic annotation were incongruent with 16S rRNA gene (Parks et al. 2017). MAGs were also filtered based on 40 single-copy protein-coding gene trees (Coleman et al. 2021). Specifically, we inspected every individual marker gene tree and remove MAGs assigned to different phyla by different marker genes. Finally, MAGs derived from termite guts forming termite-specific monophyletic clusters were identified from a phylogenetic tree reconstructed using 40 single-copy marker genes.

The refinement steps described above, relying upon the analyses of genome properties and marker gene trees, resulted in a final dataset of 924 MAGs (Figure 3.1). These 924 MAGs included 521 MAGs reconstructed from the libraries sequenced with unique dual indexing approach, and 403 MAGs from the libraries sequenced with combinational indexing approach. Each MAG had >30% completeness and <10% contamination (Bowers et al. 2017). 175 of the 924 MAGs were high quality MAGs (>90% completeness and <5% contamination), 529 MAGs were of medium quality (>50% completeness and <10% contamination) and 220 MAGs were of low quality (<50% but >30% completeness and <10% contamination) (Table S3.1, Figure 3.2). Taxonomic analyses of our MAGs indicated that the major microbial phyla characteristic of the termite gut environment were represented (Dietrich et al. 2014; Bourguignon et al. 2018; Hervé et al. 2020). In addition, we found 22 MAGs of the phylum Verrucomicrobiota and four MAGs annotated to the order *Methanomicrobiales (Halobacteriota phylum)*. This is noteworthy as these MAGs were not represented in the previous termite MAG catalog (Hervé et al. 2020), adding new sequence data that could help improving our understanding of termite gut microbial functional diversity.

Protein sequences annotated to carbohydrate metabolism

The protein sequences from the termite gut MAGs were annotated against the CAZy database (Cantarel et al. 2009) using dbCAN2 (Zhang et al. 2018). Intact full length protein sequences without signs of pseudogenization detected via Pseudofinder (Syberg-Olsen et al. 2021) were analyzed. In total, we obtained 38,222 CAZyme proteins belonging to 336 CAZyme gene

families, including 130 GH families. From an ongoing analysis on 130 GHs, we present the results of 10 selected CAZymes from 7 GH families (including 4 subfamilies) whose abundance in termite guts significantly correlated with the phylogenetic tree of termite host (Table S1.5, S1.6). These GH families had a putative substrate specificity against cellulose (GH5_1, GH5_2, GH9, GH45) (Zhang et al. 2010; Aspeborg et al. 2012; Lombard et al. 2014), hemicellulose (GH5_4, GH5_10, GH10, GH11) (Aspeborg et al. 2012; Paës et al. 2012; Han et al. 2013; Tokuda et al. 2018), or chitin (GH18, GH20) (Poulsen et al. 2014; Hu et al. 2019).

To reconstruct the relationship between selected CAZymes and sequences that were not associated with termites, we compiled a dataset comprising MAGs derived from termite gut and publicly available genomes from the Genome Taxonomy Database (GTDB ver95) (Parks et al. 2020). Genomes from the GTDB database were selected using BLASTP comparisons with protein sequences from the selected CAZymes from termite MAGs. The best 100 GTDB genomes were selected based on percentage identity, e-value, and sequence length criteria from the BLAST output of each gene family, which resulted in a final set of 2393 GTDB genomes. This subset was dereplicated at 95% ANI (species level), reducing the dataset to 1779 genomes. We reconstructed a phylogenetic tree of these 1779 genomes obtained from GTDB and the 924 MAGs reconstructed from our termite gut metagenomes.

BLAST similarity searches of the 10 selected CAZymes from termite guts showed that average protein identity was below 55% (over >100 amino acid length; Figure 3.3), indicating that termite gut CAZymes are novel and poorly represented in the publicly available database. Only 2595 of 143,906 protein sequences from the GTDB database genomes had sequence identity >=90% with the termite gut CAZymes. These protein sequences belonged to environments such as soil, root, or human-related microbes.

Species tree and Gene trees

We used the mOTU software (Sunagawa et al. 2013) to identify 40 single-copy protein-coding marker genes from the final dataset of 2703 MAGs consisting of MAGs derived from termite gut samples and genomes from non-termite environments. We reconstructed a phylogenetic tree and found multiple termite-specific monophyletic clusters of MAGs reconstructed from termite gut metagenomes (Figure 3.4). Multiple termite-specific clusters were found for each phylum at different microbial taxonomic levels. For example, at class level, termite gut MAGs of *Desulfarculia* and *Desulfovibrionia* (Desulfobacterota phylum) formed to clade sister to sequences from the human gut, wastewater, and lake sediments. The phylum Firmicutes also contained multiple termite-specific clusters belonging to the family *Ruminococcaeae*. These clusters were interspersed with sequences from mammalian guts (Figure S3.1), suggesting that, as shown by previous 16S rRNA-based analyses, many termite gut microbes belong to lineages adapted to the intestinal environment that were presumably exchanged among unrelated host species (Dietrich et al. 2014; Bourguignon et al. 2018). The termite-specific clusters consisting of four or more MAGs were kept for downstream phylogenetic analyses. This ensured sufficient representation within termite-specific cluster, preventing spurious inference of HGT from the comparison of species trees and gene trees.

We generated Maximum likelihood (ML) phylogenetic trees for 10 selected CAZymes found in 2703 MAGs. We reconstructed four trees for each CAZyme tree using four models: (i) CAT model on protein-coding nucleotide sequences without partitions (Price et al. 2010), (ii) GTR+G+I model on protein-coding nucleotide sequences with all codon positions given separate partitions (Chernomor et al. 2016; Minh et al. 2020), (iii) GTR+G+I model on protein-coding sequences with one partition for the first and second codon positions and one partition

for the third codon position (Chernomor et al. 2016; Minh et al. 2020) to account for third codon saturation for ancient divergences (Kikuchi et al. 2009) and (iv) GTR+G+I model on protein-coding sequences encoded as purines Rs (A or G) and pyrimidines Ys (T or C) to account for GC bias in endosymbiont genomes (Husník et al. 2011; Lo et al. 2016).

We compared the phylogenetic trees reconstructed with the four models using the approximately unbiased (AU) test (Shimodaira 2002). The trees reconstructed with a GTR+G+I model and with protein-coding sequences separated into two partitions, one partition for the first and second codon positions and one partition for the third codon position, were selected (Table 3.1) for all the CAZymes. For eight GH families, the trees reconstructed with a GTR+G+I model and three partitions (third model) or each codon position (second model), were equally likely. Two GH families showed that codon partitioning models and RY coding model were equally likely. None of the GH families supported the use of a CAT model (first model). As corroborated by previous analyses, we found that models giving the third codon positions a separate partition and models based on a RY coding are more suitable for resolving the evolution of gut symbionts (Kikuchi et al. 2009; Husník et al. 2011).

Cellulases

Cellulose is the major component of plant matter (20-40%) (Tomme et al. 1995) and can be hydrolyzed by the action of enzymes belonging to multiple GH families (Aspeborg et al. 2012; Lombard et al. 2014). We examined the evolutionary history of several GHs that cleave cellulose chains via endo-β-1,4-glucanases activity, such as GH5_2, GH9, GH45, and GH5_1. The subfamily GH5_1 can also depolymerize cellulose chains by cellobiohydrolase activity (Aspeborg et al. 2012).

The GH families 9 (n=327) and 5_2 (n=165) included the highest number of sequences derived from the termite gut environment, while GH5_1 (n=47) and GH45 (n=64) were comparatively less abundant. The phylogenetic relationship of the termite gut GHs with other environments indicated a complex evolutionary history. GH9 and GH5_2 were found in MAGs belonging to multiple microbial phyla, such as Spirochaetota, Bacteroidota, Firmicutes A, or Fibrobacteriota. Comparison with the species tree topology indicated that they were acquired multiple times by termites of different diets and lineages. This was evidenced by sequences from other gut environments, such as marine sediments, soil, and anaerobic sludge interspersed with those of termite gut (Figure 3.5A, Figure S3.2). Represented by fewer sequences, both GH5_1 and GH45 were encoded by the genomes of MAGs belonging to a single microbial phylum. The family *Ruminococcaceae* (Firmicutes A phylum) was specialized for GH5_1, whereby MAGs belonging to unique dual indexed higher termites of different diets such as *Microcerotermes* sp., *Amitermes* sp., and *Constrictotermes* sp. were distributed with sequences found in human and other mammal guts (Figure 3.5B, Figure S3.2). These phylogenetic analyses suggests that the ability to metabolize cellulose is not unique to termite gut as the CAZymes involved in the process are also found in other anaerobic environments.

GH45 formed a single termite-specific cluster including sequences of MAGs belonging to Fibrobacteriota (Figure 3.5C, Figure S3.2). The sequences belonged to two of the most abundant families, *Chitinispirillaceae* (previously TG3 subphylum 2) (Hongoh et al. 2006; Abdul Rahman et al. 2016) and *Fibrobacteraceae* (Abdul Rahman et al. 2016). They formed a clade sister to sequences from soil, anaerobic sludge, and guts of ruminating animals. The specificity of Fibrobacteriota to termite guts has been observed in previous 16S rRNA analyses (Hongoh et al. 2006; Mikaelyan et al. 2017a; Bourguignon et al. 2018). GH45 sequences from our dataset were mainly found in libraries prepared with the combinational indexing approach. Deeper

sequencing of the gut metagenomes is required to corroborate microbial diversity and termite species specificity for the GH45 family.

Hemicellulases

Hemicellulose is a complex polysaccharide family that modulates the interactions with cellulose chains in the plant matter (Busse-Wicher et al. 2014; Grantham et al. 2017). Representing one-third of the lignocellulose biomass (Oinonen et al. 2013), it consists of mannans and xylans with distinct biomechanical properties (Berglund et al. 2020).

We reconstructed phylogenetic trees to examine the mode of inheritance of GH5_4, GH10, and GH11, three enzymes with endo-β-1,4-xylanase activity (Aspeborg et al. 2012; Paës et al. 2012; Tokuda et al. 2018), and GH5_10, an enzyme that has endo-β-1,4-mannanase activity (Aspeborg et al. 2012). GH5_4 (n=267), GH10 (n=232), and GH11 (n=111) were the most commonly found hemicellulases in our termite gut MAGs, which is corroborated by previous meta-transcriptomics analysis (Tokuda et al. 2018; Marynowska et al. 2020). GH5_10 (n=50) was comparatively less common in our MAGs.

Comparison with the species tree showed the presence of multiple termite-specific clusters belonging to several microbial phyla for all hemicellulase GH trees. Specifically, the termite clusters from the GH5_4, GH11, and GH5_10 trees consisted mainly of MAGs reconstructed from higher termite metagenomes. The GH5_4 tree included three termite-specific clusters belonging to the family *Ruminococcaeae* (Firmicutes A phylum) that were associated with higher termite species of different diets. These clusters were sister to sequences found in the human gut, other mammalian guts, and anaerobic environments such as soil, anaerobic sludge, and hydrothermal sediments. One of the *Ruminococcaeae* cluster also contained sequences from Spirochaetota and Actinobacteriota phyla suggesting that past horizontal transfers took place in the termite guts (Figure 3.5E, Figure S3.3).

Five MAG clusters associated with Nasutitermitinae were found in the phylogenetic tree of GH11. Like GH5_4, these clusters consisted of sequences mostly associated with higher termites such as Apicotermitinae, *Microcerotermes* sp., and Syntermitinae (Figure S3.3). We found similar trends for the GH5_10 phylogenetic tree, suggesting that the hemicellulolytic gene families were acquired by the common ancestor of all higher termites after the loss of gut flagellates.

The phylogenetic tree of GH10, on the other hand, included MAG clusters associated with the fungus-cultivating Macrotermitinae and the lower termite lineages Kalotermitidae and Archotermopsidae that possess gut flagellates. These termite-specific clusters mostly belonged to Bacteroidota and Actinobacteriota. Non-Macrotermitinae higher termites of various diets were associated with separate clusters belonging to Firmicutes A and Spirochaetota. These clusters were closely related to sequences found in the guts of humans and ruminants (Figure 3.5D, Figure S3.3). Our phylogenetic tree indicates that Bacteroidota and Actinobacteriota mediating hemicellulolytic activity were already present in the common ancestor of termites. But non-Macrotermitinae higher termites acquired bacteria endowed with GH10 independently. Although a thorough examination of MAGs from multiple termite species is required to corroborate this hypothesis, it appears that hemicellulose degrading genes were affected by a combination of vertical and horizontal transfers in termites.

Chitinases

Although chitin is not a component of plant polysaccharides (Heyn 1936), chitin-degrading enzymes, such as β-glycan and α-mannans, are abundant in the termite gut microbiome (Brune 2014; Poulsen et al. 2014; Liu et al. 2018). Chitin can make up 10-20% of the cell wall of

filamentous fungi (Brown et al. 2020; Garcia-Rubio et al. 2020), and the action of chitinases help termite gut microbes to digest the fungal biomass growing on degraded plant matter or the fungal material cultivated in the nests of Macrotermitinae (Hu et al. 2019). We examined the phylogenetic history of two chitinase GH families, GH18 (n=237) and GH20 (n=183). Both GH families were encoded by several microbial phyla forming multiple termite-specific clusters. For example, we found four termite-specific groups belonging to *Lachnospiraceae* (Firmicutes A phylum) in the phylogenetic tree of GH18. These clusters were composed exclusively of sequences from fungus-cultivating Macrotermitinae (Figure 3.5F). Similarly, GH20 had three clusters of the order Bacteroidales (Bacteroidota phylum) associated with Macrotermitinae (Figure 3.5G, Figure S3.4). Both these phyla are highly abundant in fungus-cultivating termites (Poulsen et al. 2014; Hu et al. 2019). The multiple termite-specific clusters suggest chitinolytic activity was acquired multiple times by Macrotermitinae.

Other microbial phyla also encode these GH families. Spirochaetota and Firmicutes A generated many termite-specific groups sister to sequences found in the soil, marine sediments, human gut, *etc.* (Figure 3.5F, Figure 3.5G). These groups support the notion that these microbes are specialized for fungal rich environments (Kielak et al. 2013), whereby multiple transfers have taken place to the termite gut. Interestingly, these two phyla were represented by clusters of sequences from several higher termite species of various diets, which sometimes include sequences from Macrotermitinae. This supports previous reports of chitinase gene families in other higher termite guts (Warnecke et al. 2007; He et al. 2013; Hu et al. 2019), and indicates that microbes encoding these GHs were probably already present in the common ancestor of higher termites.

Conclusion

In this study, we analyzed the evolutionary history of gut microbial genes with carbohydrate degrading functions known to be important in the termite gut (Warnecke et al. 2007; Hu et al. 2019; Marynowska et al. 2020). Phylogenetic comparison of termite gut CAZymes with that of other environments revealed that multiple gut microbial phyla have converged on similar functions. Lateral transfers from other environments such as guts of ruminants and humans, soil, anaerobic digestors *etc*. were found (Figure 3.5). This clustering supports the notion that microbes essential for the carbohydrate metabolism adapted to other environments were acquired by termites (Xu et al. 2007; Brulc et al. 2009; Pope et al. 2010). Specifically, a close phylogenetic clustering of gene families from different gut environments suggests that inter-microbial interactions shape the gut functional potential (Shterzer and Mizrahi 2015). To thrive in the host, the gut microbes share resources (Graber et al. 2004), metabolites (Rosenthal et al. 2011), and potentially acquire genes by HGT (Tokuda et al. 2018). The gut environment is characterized by a high degree of functional redundancy (Xu et al. 2007) between phylogenetically distantly related taxa, allowing the maintenance of a stable gut ecosystem. We observed similar trends in our gene trees. Sequences belonging to several termite gut phyla such as Spirochaetota and Firmicutes A were interspersed in termite-specific clusters probably driven by lateral transfers in the gut.

These microbial interactions have a profound effect on the host. We found some of the most abundant gut microbes, such as Spirochaetota, Firmicutes A, Fibrobacteriota, encoding multiple gene families with activity on different substrates (Figure 3.5). The ability to utilize multiple substrates such as cellulose and hemicellulose, two major components of plant matter, provides the termite host with the capacity to feed on different types of plant matter (Tokuda et al. 2018).

The flexibility to exploit new feeding habits might have contributed to termites ecological and evolutionary success.

Along with the dominant gut microbes, we successfully assembled MAGs from 11 other microbial phyla, constituting a major component of the termite gut environment (Dietrich et al. 2014; Bourguignon et al. 2018; Hervé et al. 2020). The relative abundance of these microbes has been found to vary with termite diet (Mikaelyan et al. 2015a; Mikaelyan et al. 2017a). We observe a similar diet-specific clustering of carbohydrate metabolizing functions. We found that chitinolytic activity was mainly encoded by Firmicutes A and Bacteroidota, two of the most abundant microbial phyla found in fungus cultivating Macrotermitinae (Poulsen et al. 2014; Hu et al. 2019). *Lachnospiraceae* (Firmicutes A phylum) formed multiple Macrotermitinae-specific clusters suggesting that fungus-cultivating termites acquired chitinase activity multiple times independently to feed on fungal gardens.

Host evolutionary history is another determining factor of the termite gut microbiome (Dietrich et al. 2014; Abdul Rahman et al. 2015; Bourguignon et al. 2018). Phylogenetic analyses of putative hemicellulase gene families clustered higher termite species of different diets and phylogenetic history together. This suggests that the loss of flagellates ~50 Mya led to the acquisition of microbes with hemicellulolytic activity by the common ancestor of higher termites. This host-mediated selection of microbes in the gut environment can also generate specialization of function by an individual phylum (Xu et al. 2007). Fibrobacteriota phylum showcased a distinct termite specificity with respect to other environments in GH45 phylogenetic analysis (Figure 3.5). But all the sequences belonged to MAGs obtained from libraries prepared with the combinational indexing approach, requiring a deeper sequencing to confirm the host identity of GH45 sequences from our Fibrobacteriota MAGs.

Future comparative analyses of microbial gene families, as we did in this chapter, will help to understand the evolution of other enzymes essential to termites (Brune 2014; Brune and Dietrich 2015). We reconstructed MAGs from the metagenomes of 201 termite and one cockroach species, increasing the number of MAGs from termite guts (Abdul Rahman et al. 2016; Hervé et al. 2020). However, 81% of our MAGs (749 out of 924) were of medium or low quality (i.e., <50% completeness). Future work involving a deeper sequencing and generation of circular MAGs (Chen et al. 2020) is needed to resolve the complete picture of lignocellulose digestion by termite gut microbes.

Table 3. 1. GH families analyzed in this study, their potential substrate specificity and AU model test to select the best tree topology.

Gene family	sub family	Potential activity	Potential substrate	AU test	Reference for substrate specificity
GH5	GH5_2	endo-β-1,4-glucanase	cellulase	Separate codon or third codon	Aspeborg et al 2012
GH5	GH5_1	endo-β-1,4-glucanase or cellobiohydrolase	cellulose	Separate codon or third codon or RY	Aspeborg et al 2012
GH9	GH9	endo-β-1,4-glucanase	cellulose	Separate codon or third codon	Zhang et al 2010; Nguyen et al 2019
GH45	GH45	β-1,4-endoglucanases	cellulose	Separate codon or third codon	Lombard et al 2013
GH18	GH18	chitinase	chitin	Separate codon or third codon	Hu et al 2019
GH20	GH20	β-N-acetylglucosaminidase	chitin	Third codon	Meekrathok et al 2020; Hu et al 2019
GH5	GH5_4	endo-β-1,4-glucanases or endo-β-1,4-xylanase	hemicellulose	Separate codon or third codon	Aspeborg et al 2012
GH5	GH5_10	endo-β-1,4-mannanase	hemicellulose	Separate codon or third codon	Aspeborg et al 2012
GH10	GH10	endo-1, 4-beta-xylases	hemicellulose	Separate codon or third codon	Verma et al 2012; Han et al 2013
GH11	GH11	endo-1, 4-beta-xylases	hemicellulose	Separate codon or third codon	Tokuda et al 2018; Paes et al 2012

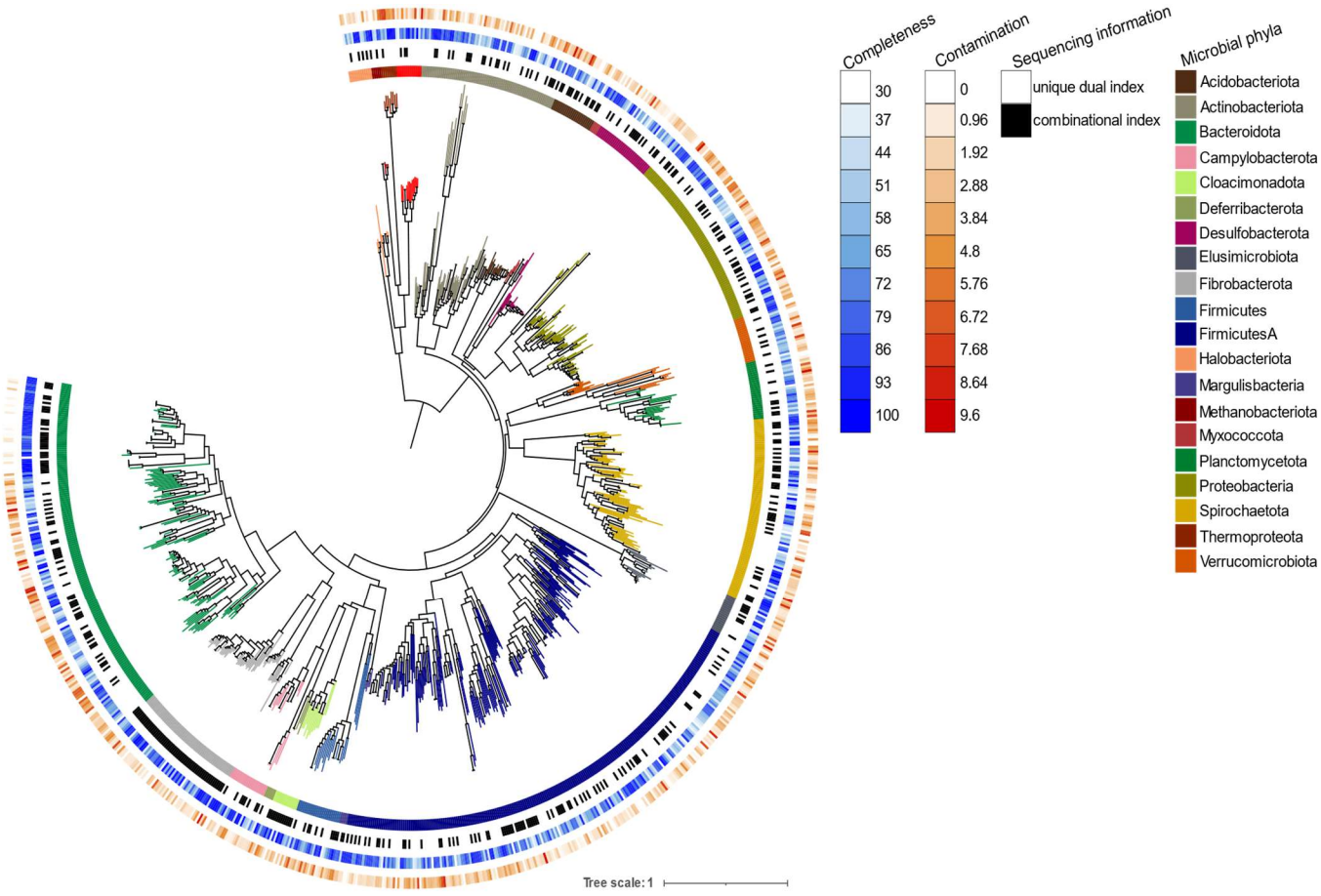

Figure 3. 1. Phylogenetic tree of 924 MAGs from the termite gut. The tree was generated from concatenated universally occurring single-copy protein-coding sequences.

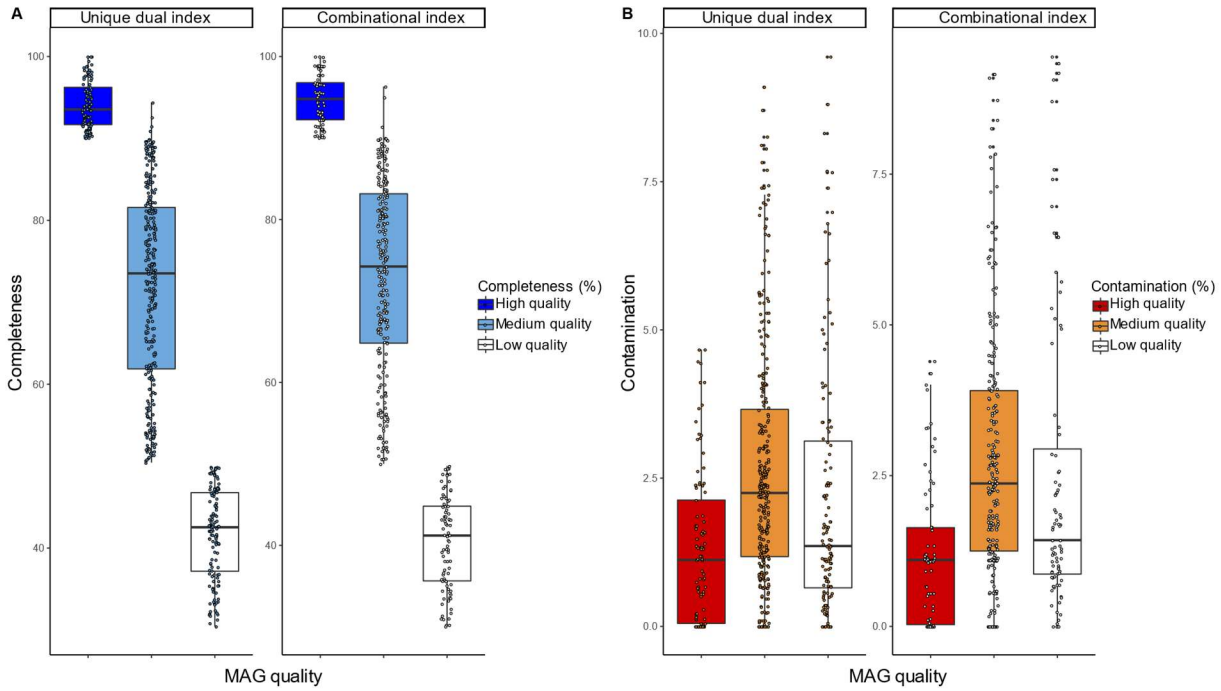

Figure 3. 2. MAG recovery information. Percent completeness (a) and contamination (b) metrics of high-, medium- and low-quality MAGs from samples sequenced using a unique dual index or combinational index. (c) MAGs are classified at different taxonomic ranks by GTDB-Tk (Chaumeil et al. 2019).

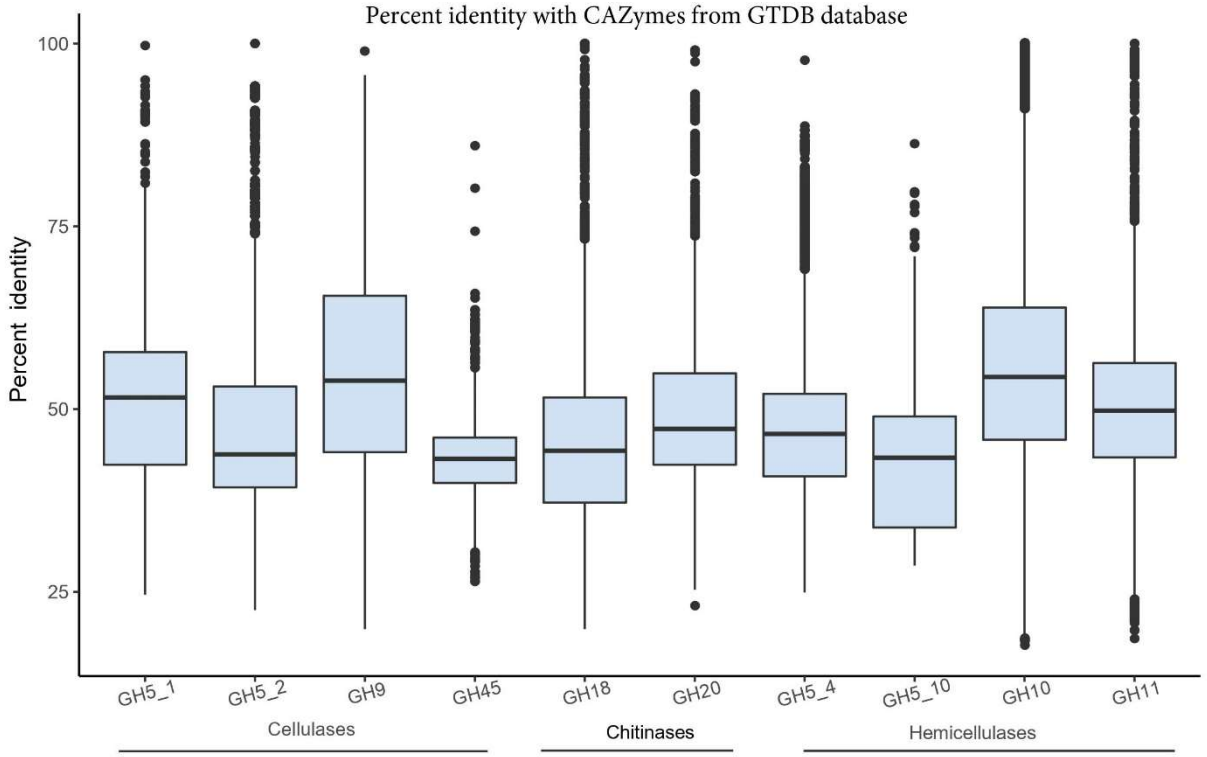

Figure 3. 3. Percent identity between CAZymes from termite gut MAGs and GTDB database. Boxes represent the interquartile range, center lines indicate median values, and whiskers and dots present the most extreme datapoints.

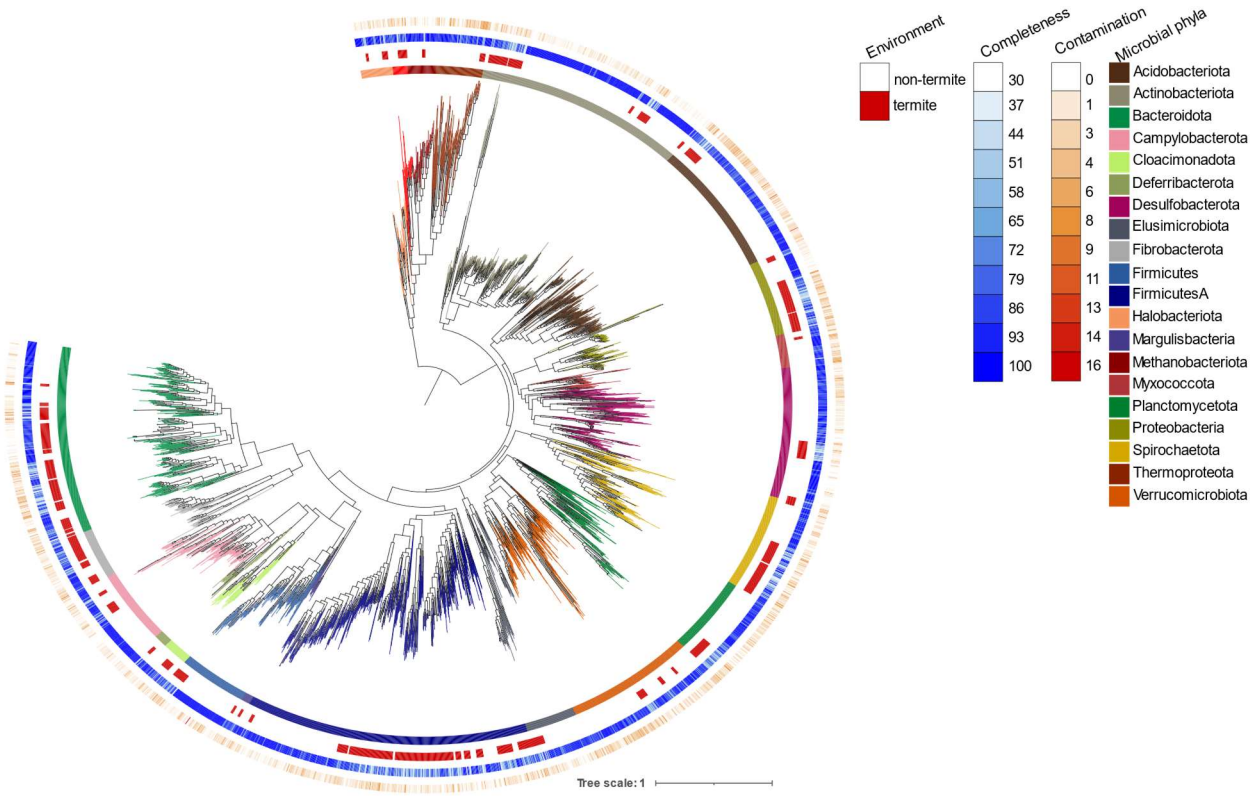

Figure 3. 4. Phylogenetic tree of 2703 MAGs from termite gut and non-termite environments analyzed in this study. The tree was generated from concatenated universally occurring single-copy protein-coding sequences. The non-termite samples correspond to publicly available genomes from the GTDB database (Parks, et al. 2020). The taxon names and support values are shown in Figure S3.1.

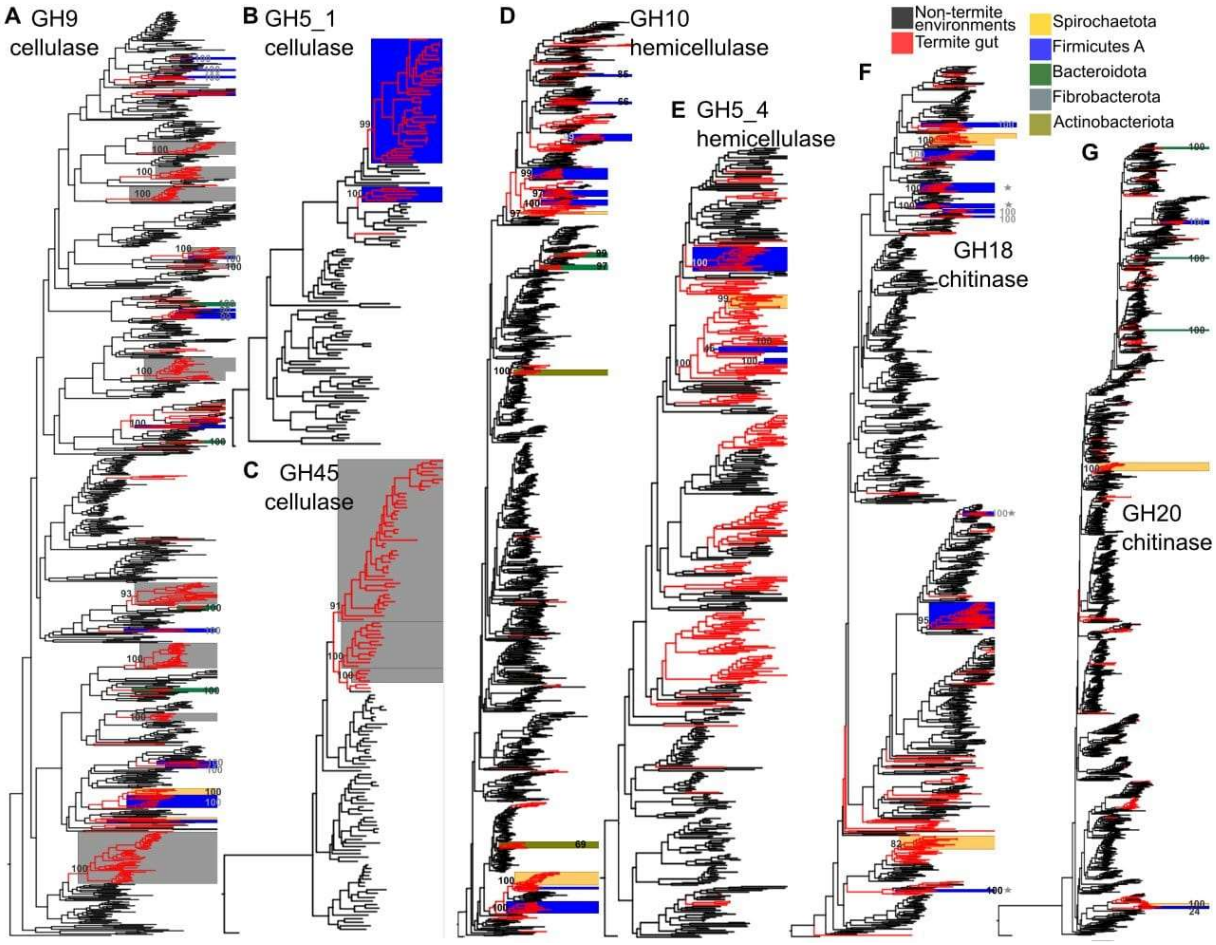

Figure 3. 5. Selected phylogenetic trees showing relationship of microbial CAZymes from termite gut and non-termite environments. Trees were inferred using GTR+I+G model with two partitions: combined first and second codons and third codon using IQTREE. Close inspection of the trees shows multiple microbial phyla performing the same function (highlighted clades) and close relationship of termite gut sequences (in red) with other environments (in black). Bootstrap support values as inferred via the ultrafast bootstrap approximation (UFBoot) are indicated for highlighted nodes. Asterisk next to blue highlighted clades in (F) distinguishes *Lachnospiraceae* clades from other Firmicutes A clades. Taxon names and bootstrap support values for each tree and the remaining trees examined in this chapter are shown in Supplementary figures 2-4.

Figure S3. 1. The phylogenetic tree of 2703 MAGs from the termite gut and close relatives from other environments. The tree was generated from concatenated universally occurring single-copy protein-coding marker gene sequences in FastTree. Branch support values were inferred using the Shimodaira-Hasegawa (SH) like approximate likelihood ratio test. The taxon names in the trees end with information of their source of isolation. MAGs generated in this study, corresponding to unique dual index samples, end with `id_` followed by the termite species name. Whereas the MAGs from combinational index samples end with `id_tagswitched`. Sequences from the GTDB database end with `id_` followed by their source of isolation based on their NCBI biosample ids. (link)

Figure S3. 2. Phylogenetic trees of Glycosyl hydrolases (GHs) with putative substrate specificity against cellulose in the plant matter. The trees were inferred using the GTR+I+G model with two partitions: combined first and second codons and third codon using IQTREE. The branch support values were inferred via the ultrafast bootstrap approximation (UFBoot). The taxon names in the trees end with information of their source of isolation. MAGs generated in this study, corresponding to unique dual index samples, end with `id_` followed by the termite species name. Whereas the MAGs from combinational index end with `id_tagswitched`. Sequences from the GTDB database end with `id_` followed by their source of isolation based on their NCBI biosample ids. (link)

Figure S3. 3. Phylogenetic trees of Glycosyl hydrolases (GHs) with putative substrate specificity against hemicellulose in the plant matter. The trees were inferred using the GTR+I+G model with two partitions: combined first and second codons and third codon using IQTREE. The branch support values were inferred via the ultrafast bootstrap approximation (UFBoot). The taxon names in the trees end with information of their source of isolation. MAGs generated in this study, corresponding to unique dual index samples, end with `id_` followed by the termite species name. Whereas the MAGs from combinational index end with `id_tagswitched`. Sequences from the GTDB database end with `id_` followed by their source of isolation based on their NCBI biosample ids. (link)

Figure S3. 4. Phylogenetic trees of Glycosyl hydrolases (GHs) with putative substrate specificity against chitin. The trees were inferred using the GTR+I+G model with two partitions: combined first and second codons and third codon using IQTREE. The branch support values were inferred via the ultrafast bootstrap approximation (UFBoot). The taxon names in the trees end with information of their source of isolation. MAGs generated in this study, corresponding to unique dual index samples, end with `id_` followed by the termite species name. Whereas the MAGs from combinational index end with `id_tagswitched`. Sequences from the GTDB database end with `id_` followed by their source of isolation based on their NCBI biosample ids. (link)

Table S3. 1. MAGs generated in this study. Genomic characteristics, taxonomic affiliation, and host information is provided. (link)

Conclusion section

The evolutionary history of termites over 150 million years, their ecological success in the tropics, and their gut microbiome have been extensively studied. Many studies have helped improve our understanding of termite phylogeny (Bourguignon et al. 2015; Bucek et al. 2019), physiology (Watanabe et al. 2014; Miura and Maekawa 2020), and gut microbiome (Ohkuma and Brune 2011; Brune and Dietrich 2015). Termites thrive on wood, soil, grass, or leaf litter, which is rare among animals (Donovan et al. 2001; Eggleton and Tayasu 2001). The early-evolving termite species, called lower termites, developed a nutritional symbiosis with cellulolytic flagellates, bacteria, and archaea (Brune 2006; Ohkuma and Brune 2011). The evolution of higher termites was accompanied by the loss of gut flagellates and the development of fungiculture (Garnier-Sillam et al. 1989) and bactericulture (Rouland-Lefèvre et al. 2006) in the nests of Macrotermitinae and Sphaerotermitinae, respectively. All higher termites also retained gut prokaryotic symbionts that helped them to process their food (Brune and Dietrich 2015). The higher termites also diversified their diet as many species evolved to feed on soil (Eggleton et al. 1998). Termites have also been studied from an applied perspective. Termites are the cause of major economic losses as they are pests of crops and building structures (Su and Scheffrahn 2000). Understanding the gut microbial functions has potential applications in agriculture and pest management (Scharf 2015). In addition, the gut microbes have evolved innovative ways to help termites digest wood. By using a combination of gene families (Liu et al. 2018), acting in concert on multiple substrates (Liu et al. 2019), termite gut microbes efficiently metabolize lignocellulose. These enzymes are of potential use for biotechnological applications in the biofuel industry (Dodd and Cann 2009; Himmel and Bayer 2009). However, many unanswered questions remain about the evolutionary history of termites. For example, the role of host phylogeny and diet on gut microbial functions, the co-evolution dynamics, and the frequency of horizontal transfers among the microbial taxa in the gut are still unclear.

This thesis aimed to address these gaps by analyzing 202 whole gut metagenomes belonging to 201 termite samples and one sample of the sister group of termites, the *Cryptocercus* cockroach. These metagenomes include previously understudied termite species, belonging to early-evolving termite lineages and higher termites feeding on soil. Our dataset is representative of the termite phylogenetic and dietary diversity across 150 million years of evolution.

Our taxonomic and functional profiling of termite gut microbiomes indicated that genes and metabolic pathways involved in carbohydrate degradation were present across distinct termite lineages and *Cryptocercus* cockroach. Phylogenetic comparative methods suggested that the termite phylogenetic history is largely predictive of the gut prokaryotic functions. For example, the relative abundance of carbohydrate degrading enzymes was correlated with the loss of gut flagellates in higher termites. We found that Glycosyl hydrolases (GHs) potentially targeting cellulose were depleted in lower termite gut metagenomes as compared to wood-feeding higher termites. Additionally, several GH families found in our lower termite gut metagenomes were absent in gut flagellate genomes (Nishimura et al. 2020).

We also examined the changes in the gut microbial functions of higher termite species with different diets. The acquisition of a soil diet was mainly accompanied by changes in gene abundance rather than by the acquisition of new genes and pathways. For example, genes involved in metabolic pathways such as reductive acetogenesis, sulfate reduction, and nitrogen metabolism were present across termite species with various dietary habits. But their relative abundance varied with diet. The genes involved in reductive acetogenesis, a pathway that produces acetate after fermentation of wood particles (Hungate 1939; Brune 2014), were more

abundant in wood-feeding termites than soil-feeding ones. An opposite trend was observed for genes of the dissimilatory nitrate reduction pathway involved in nitrogen-recycling (Ngugi et al. 2011), which were more abundant in termites feeding on soil, a substrate richer in nitrogen than wood (Ji and Brune 2001; Ji and Brune 2005). The presence of key genes and pathways in all termites suggest that microbial gene families having the major metabolic functions were already present in the common ancestor of higher termites.

Overall, the results from chapter one represent significant advances in our understanding of the role of host evolutionary history and dietary diversity on gut microbial functions. They provide a global picture of gut microbial carbohydrate metabolism and metabolic pathways nutritionally important for the termite host.

Previous research studies on termite gut microbial composition were based on 16S rRNA analysis (Dietrich et al. 2014; Mikaelyan et al. 2015a; Bourguignon et al. 2018). But using 16S rRNA gene has several limitations. For example, many bacterial genomes encode for multiple copies of the 16S rRNA gene (Coenye and Vandamme 2003), affecting estimation of bacterial abundance from 16S data. The 16S rRNA gene also has a low phylogenetic resolution (Lan et al. 2016), thence is inappropriate for phylogenetic reconstructions. To overcome the limitations of analyses based on the 16S rRNA gene, chapter two is based on single-copy protein-coding marker genes (Sunagawa et al. 2013) obtained from metagenome assemblies. We performed phylogenetic analyses on these marker genes found in termite gut metagenomes together with non-termite sequences mined from publicly available genomes. We extracted several monophyletic groups including exclusively of microbes commonly found in the termite gut (Brune 2014). We carried out co-evolution analyses using three algorithms (Balbuena et al. 2013; Smith 2020) and found multiple examples of significant coevolution between termites and some lineages of termite gut microbes. For example, the microbial groups Spirochaetota family *Treponemataceae B* and the Fibrobacterota genus *Fibromonas* were mainly found in higher termite guts (Hongoh et al. 2006) with a high proportion in *Microcerotermes* sp. and Nasutitermitinae sub-family (Mikaelyan et al. 2015b) forming termite lineage-specific clusters. On the other hand, *Adiutrix* (Desulfobaterota) and *Rhodocyclaceae* (Proteobacteria) consisted of sequences from *Cryptocercus* cockroach, lower termites, and higher termites of different diets (Dietrich et al. 2014; Pramono et al. 2017). The presence of these microbial group across multiple termite lineages suggests a vertical mode of inheritance for the last 150 million years, since modern termite cladogenesis was initiated. Overall, our marker-gene-based phylogenetic analyses indicated that some termite gut microbes were already present in the common ancestor of all termites. In contrast, other gut microbes were acquired individually by each termite lineage or multiple lineages with a similar diet. Our results showcase that lineage specificity occurs at various scales of host evolutionary history.

In chapter three, we reconstructed the phylogenetic trees of seven GH families and four subfamilies known to be abundant in the gut of termites. We attempted to determine the role of horizontal gene transfers on the evolution of microbial functions. To do so, we generated 924 metagenome-assembled genomes (MAGs) from termite species representing the host evolutionary history and dietary diversity. These MAGs represent the major microbial phyla present in the termite gut environment (Dietrich et al. 2014; Bourguignon et al. 2018; Hervé et al. 2020). We also found novel MAGs belonging to microbes abundant in the gut of termites but absent from previously generated MAG catalogs (Abdul Rahman et al. 2016; Hervé et al. 2020), improving the representation of termite gut microbes. We performed phylogenetic analyses on GHs with putative substrate activity against the major components of plant matter

termites feed on, such as cellulose and hemicellulose. The GH families targeting chitin, a component of the fungal cell wall (Heyn 1936), involved in digesting fungal biomass growing on decomposing wood and fungal material on the nest (Hu et al. 2019) were also examined.

Comparison of GHs from termite gut with those from other environments generated clusters interspersed with sequences found in similar anaerobic environments such as the guts of ruminats and humans, soil, *etc*. Along with transfers from the environment, intermicrobial interactions within the termite gut were also found to shape the functional potential of the microbes (Shterzer and Mizrahi 2015). A high degree of functional redundancy between phylogenetically distantly related taxa was observed. Whereby multiple phyla were interspersed with each other in termite-specific clusters. For example, two of the most abundant termite gut phyla, Firmicutes A and Spirochaetota, were interspersed with each other suggesting horizontal transfers within the termite gut.

The role of host evolutionary history and dietary diversity was also found to affect the evolutionary history of GH families. The chitinolytic GHs were found in MAGs of Firmicutes A and Bacteroidota associated with fungus-cultivating Macrotermitinae species. These two phyla are the most abundant microbes in the gut of Macrotermitinae (Poulsen et al. 2014; Hu et al. 2019), suggesting a termite lineage-specific trend. On the other hand, the hemicellulolytic GHs formed phylogenetic clusters represented by MAGs reconstructed from higher termite gut metagenomes. This indicated that hemicellulose degrading ability was probably acquired by the common ancestor of higher termites. These results corroborate the hypothesis that loss of flagellates in higher termites was accompanied by dramatic changes in gut microbial composition and functions (Dietrich et al. 2014). We also observed specialization of carbohydrate degradation in the termite gut by a specific microbial phylum. Most Fibrobacteriota MAGs encoded a GH45 belonging to a phylogenetic cluster sister to sequences from other environments. Therefore, the detailed analysis of seven GH families and four subfamilies suggests that the ability to utilize multiple substrates such as cellulose, hemicellulose, and chitin by the termite gut microbes might have provided the host with the ability to feed on different types of plant matter (Tokuda et al. 2018).

Overall, host evolutionary history and dietary habits shape the taxonomic and functional abundance of gut microbes. The results of my thesis help us understand the evolution of termite gut microbes during key events in termite history. Looking forward, future work on comparative genomics using circularized MAGs with strain level information (Chen et al. 2020) will help us understand population-level diversity in the gut. Long read sequencing could allow the assembling of complete microbial genome (Xie et al. 2020), shed light on the genomic architecture of the endosymbionts, and generate a complete picture of functional processes taking place in the gut (Stewart et al. 2018; Singleton et al. 2021). The genomic approaches can be complemented by meta-transcriptomics (Marynowska et al. 2020) and visualization methods (Hongoh 2011) to examine the expression profile and spatial distribution of the microbes in the gut compartments with different physiochemical conditions (Brune 2014).

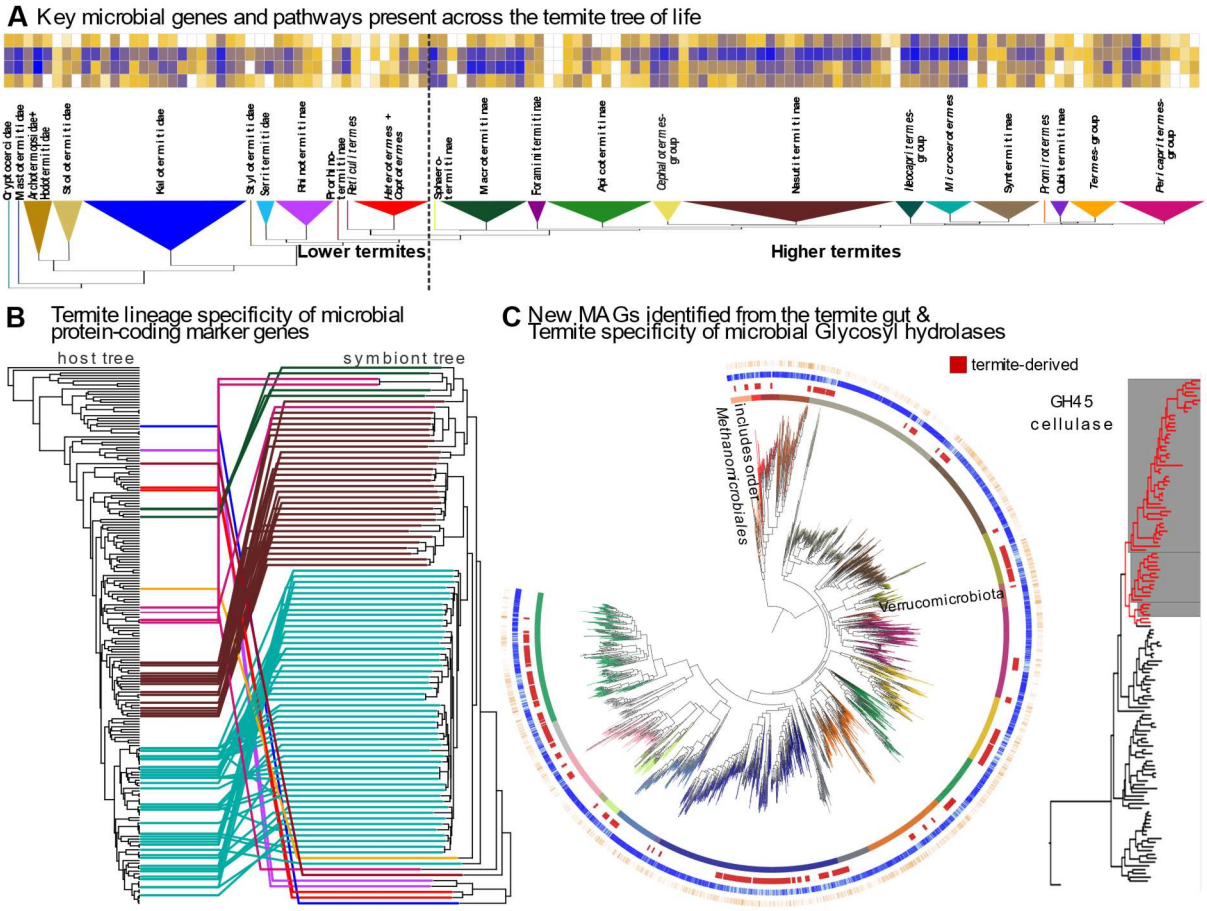

Figure A. 3- Summary of the thesis. (A) The microbial gene families were found to be abundant across the termite tree showcasing that the genes were acquired early in termite evolution. (B) The phylogenetic analysis of universally occurring single-copy protein-coding marker genes of microbial phyla indicated termite lineage specificity. (C) Metagenome assembled genomes (MAGs) undescribed in previous MAG catalogs were found. The two microbial groups are mentioned on the MAG tree. Termite gut specificity of Glycosyl hydrolases (GHs) gene trees from the MAGs with respect to sequences from other environments were determined. The cophylo lines in (B) are colored according to termite lineages mentioned in (A).

References

Abdul Rahman N, Parks DH, Vanwonterghem I, Morrison M, Tyson GW, Hugenholtz P. 2016. A Phylogenomic Analysis of the Bacterial Phylum Fibrobacteres. *Front Microbiol* **6**: 1469-1469.

Abdul Rahman N, Parks DH, Willner DL, Engelbrektson AL, Goffredi SK, Warnecke F, Scheffrahn RH, Hugenholtz P. 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.

Adams DC, Otárola-Castillo E. 2013. geomorph: an r package for the collection and analysis of geometric morphometric shape data. *Methods Ecol Evol* **4**: 393-399.

Allison SD, Lu Y, Weihe C, Goulden ML, Martiny AC, Treseder KK, Martiny JB. 2013. Microbial abundance and composition influence litter decomposition response to environmental change. *Ecology* **94**: 714-725.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**: 1144-1146.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J Mol Biol* **215**: 403-410.

Anderson KE, Russell JA, Moreau CS, Kautz S, Sullam KE, Hu Y, Basinger U, Mott BM, Buck N, Wheeler DE. 2012. Highly similar microbial communities are shared among related and trophically similar ant species. *Mol Ecol* **21**: 2282-2296.

Apprill A, McNally S, Parsons R, Weber L. 2015. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat Microb Ecol* **75**.

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

Balbuena JA, Míguez-Lozano R, Blasco-Costa I. 2013. PACo: A Novel Procrustes Application to Cophylogenetic Analysis. *PLOS ONE* **8**: e61048.

Baldrian P, Kolařík M, Stursová M, Kopecký J, Valášková V, Větrovský T, Zifčáková L, Snajdr J, Rídl J, Vlček C et al. 2012. Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *ISME J* **6**: 248-258.

Basset Y, Cizek L, Cuénoud P, Didham RK, Guilhaumon F, Missa O, Novotny V, Ødegaard F, Roslin T, Schmidl J et al. 2012. Arthropod Diversity in a Tropical Forest. *Science* **338**: 1481-1484.

Béguin P. 1990. Molecular Biology Of Cellulose Degradation. *Annu Rev Microbiol* **44**: 219-248.

Bellgard MI, Wanchanthuek P, La T, Ryan K, Moolhuijzen P, Albertyn Z, Shaban B, Motro Y, Dunn DS, Schibeci D et al. 2009. Genome Sequence of the Pathogenic Intestinal Spirochete Brachyspira hyodysenteriae Reveals Adaptations to Its Lifestyle in the Porcine Large Intestine. *PLOS ONE* **4**: e4641.

Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DG, Nilsson RH. 2015. METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol Ecol Resour* **15**: 1403-1414.

Berglund J, Mikkelsen D, Flanagan BM, Dhital S, Gaunitz S, Henriksson G, Lindström ME, Yakubov GE, Gidley MJ, Vilaplana F. 2020. Wood hemicelluloses exert distinct biomechanical contributions to cellulose fibrillar networks. *Nat Commun* **11**: 4692.

Berlemont R, Delsaute M, Pipers D, D'Amico S, Feller G, Galleni M, Power P. 2009. Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. *ISME J* **3**: 1070-1081.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M, Stadler PF. 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol* **69**: 313-319.

Bignell DE, Eggleton P. 1995. On the elevated intestinal pH of higher termites (Isoptera: Termitidae). *Insectes Sociaux* **42**: 57-69.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Bourguignon T, Lo N, Cameron SL, Šobotník J, Hayashi Y, Shigenobu S, Watanabe D, Roisin Y, Miura T, Evans TA. 2015. The evolutionary history of termites as inferred from 66 mitochondrial genomes. *Mol Biol Evol* **32**: 406-421.

Bourguignon T, Lo N, Dietrich C, Šobotník J, Sidek S, Roisin Y, Brune A, Evans TA. 2018. Rampant Host Switching Shaped the Termite Gut Microbiome. *Curr Biol* **28**: 649-654.e642.

Bourguignon T, Lo N, Šobotník J, Ho SYW, Iqbal N, Coissac E, Lee M, Jendryka MM, Sillam-Dussès D, Křížková B et al. 2017. Mitochondrial Phylogenomics Resolves the Global Spread of Higher Termites, Ecosystem Engineers of the Tropics. *Mol Biol Evol* **34**: 589-597.

Bourguignon T, Lo N, Šobotník J, Sillam-Dussès D, Roisin Y, Evans TA. 2016. Oceanic dispersal, vicariance and human introduction shaped the modern distribution of the termites *Reticulitermes*, *Heterotermes* and *Coptotermes*. *Proc Royal Soc B* **283**: 20160179.

Bourguignon T, Šobotník J, Lepoint G, Martin J-M, Hardy OJ, Dejean A, Roisin Y. 2011. Feeding ecology and phylogenetic structure of a complex neotropical termite assemblage, revealed by nitrogen stable isotope ratios. *Ecol Entomol* **36**: 261-269.

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**: 725-731.

Brauman A, Kane MD, Labat M, Breznak JA. 1992. Genesis of Acetate and Methane by Gut Bacteria of Nutritionally Diverse Termites. *Science* **257**: 1384-1387.

Breznak JA. 2000. Ecology of prokaryotic microbes in the guts of wood- and litter-feeding termites. In *Termites: evolution, sociality, symbioses, ecology*, (ed. T Abe, et al.), pp. 209-231. Kluwer Academic Publishers, Boston.

Breznak JA, Brune A. 1994. Role of microorganisms in the digestion of lignocellulose by termites. *Annu Rev Entymology* **39**: 453-487.

Breznak JA, Switzer JM. 1986. Acetate Synthesis from H2 plus CO2 by Termite Gut Microbes. *Appl Environ Microbiol* **52**: 623-630.

Brooks AW, Kohl KD, Brucker RM, van Opstal EJ, Bordenstein SR. 2016. Phylosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History. *PLoS Biol* **14**: e2000225.

Brown HE, Esher SK, Alspaugh JA. 2020. Chitin: A "Hidden Figure" in the Fungal Cell Wall. *Curr Top Microbiol Immunol* **425**: 83-111.

Brulc JM, Antonopoulos DA, Miller ME, Wilson MK, Yannarell AC, Dinsdale EA, Edwards RE, Frank ED, Emerson JB, Wacklin P et al. 2009. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci USA* **106**: 1948-1953.

Brune A. 2006. Symbiotic Associations Between Termites and Prokaryotes. *Prokaryotes* **1**: 439-474.

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

Brune A. 2018. Methanogens in the digestive tract of termites. In (Endo)symbiotic methanogenic archaea. In *Microbiology monographs*, Vol 19 (ed. JHP Hackstein), pp. 81–101. Springer.

Brune A. 2019. Methanogenesis in the digestive tracts of insects and other arthropods. In Biogenesis of hydrocarbons. In *Handbook of hydrocarbon and lipid microbiology*, (ed. AJM Stams, DE Sousa), pp. 229–260. Springer.

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, Ohkuma M. 2011. Role of the Termite Gut Microbiota in Symbiotic Digestion. In *Biology of Termites: a Modern Synthesis*, doi:10.1007/978-90-481-3977-4_16 (ed. DE Bignell, et al.), pp. 439-475. Springer Netherlands, Dordrecht.

Bucek A, Šobotník J, He S, Shi M, McMahon DP, Holmes EC, Roisin Y, Lo N, Bourguignon T. 2019. Evolution of Termite Symbiosis Informed by Transcriptome-Based Phylogenies. *Curr Biol* **29**: 3728-3734.e3724.

Buček A, Wang M, Šobotník J, Sillam-Dussès D, Mizumoto N, Stiblík P, Clitheroe C, Lu T, González Plaza JJ, Mohagan A et al. 2021. Transoceanic voyages of drywood termites (Isoptera: Kalotermitidae) inferred from extant and extinct species. *bioRxiv* doi:10.1101/2021.09.24.461667.

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

Buchner P. 1965. *Endosymbiosis of animals with plant microorganisms*. Interscience Publishers, New York, NY.

Buckel W, Thauer RK. 2013. Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1827**: 94-113.

Busse-Wicher M, Gomes TC, Tryfona T, Nikolovski N, Stott K, Grantham NJ, Bolam DN, Skaf MS, Dupree P. 2014. The pattern of xylan acetylation suggests xylan may interact with cellulose microfibrils as a twofold helical screw in the secondary plant cell wall of Arabidopsis thaliana. *Plant J* **79**: 492-506.

Calusinska M, Happe T, Joris B, Wilmotte A. 2010. The surprising diversity of clostridial hydrogenases: a comparative genomic perspective. *Microbiology* **156**: 157-1588.

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.

Chan PP, Lowe TM. 2019. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol Biol* **1962**: 1-14.

Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**: 1925-1927.

Chen L, Anantharaman K, Shaiber A, Eren AM, Banfield JF. 2020. Accurate and complete genomes from metagenomes. *Genome Res* doi:10.1101/gr.258640.119.

Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology* **65**: 997-1008.

Chouaia B, Rossi P, Epis S, Mosca M, Ricci I, Damiani C, Ulissi U, Crotti E, Daffonchio D, Bandi C et al. 2012. Delayed larval development in Anopheles mosquitoes deprived of Asaiabacterial symbionts. *BMC Microbiol* **12**: S2.

Coenye T, Vandamme P. 2003. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol Lett* **228**: 45-49.

Coleman GA, Davín AA, Mahendrarajah TA, Szánthó LL, Spang A, Hugenholtz P, Szöllősi GJ, Williams TA. 2021. A rooted phylogeny resolves early bacterial evolution. *Science* **372**: eabe0511.

Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, Granger B, Green L, Howd T, Mason T et al. 2018. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genom* **19**: 332.

Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology* **10**: 210.

de Vienne DM, Refrégier G, López-Villavicencio M, Tellier A, Hood ME, Giraud T. 2013. Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *New Phytol* **198**: 347-385.

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

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.

Dodd D, Cann IKO. 2009. Enzymatic deconstruction of xylan for biofuel production. *Glob Change Biol Bioenergy* **1**: 2-17.

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

Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R. 2012. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genom* **13**: 162.

Dröge S, Limper U, Emtiazi F, Schönig I, Pavlus N, Drzyzga O, Fischer U, König H. 2005. In vitro and in vivo sulfate reduction in the gut contents of the termite Mastotermes darwiniensis and the rose-chafer Pachnoda marginata. *J Gen Appl Microbiol* **51**: 57-64.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol* **4**: e88.

Edgar RC. 2018. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* **6**: e4652-e4652.

Eggleton P, Davies RG, Bignell DE. 1998. Body Size and Energy Use in Termites (Isoptera): The Responses of Soil Feeders and Wood Feeders Differ in a Tropical Forest Assemblage. *Oikos* **81**: 525-530.

Eggleton P, Tayasu I. 2001. Feeding groups, lifetypes and the global ecology of termites. *Ecol Res* **16**: 941-960.
El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A et al. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res* **47**: D427-d432.
Engel MS, Grimaldi DA. 2004. New light shed on the oldest insect. *Nature* **427**: 627-630.
Engel MS, Grimaldi DA, Nascimbene PC, Singh H. 2011. The termites of Early Eocene Cambay amber, with the earliest record of the Termitidae (Isoptera). *Zookeys* doi:10.3897/zookeys.148.1797: 105-123.
Engel P, Moran NA. 2013. The gut microbiota of insects – diversity in structure and function. *FEMS Microbiology Reviews* **37**: 699-735.
Evans PN, Boyd JA, Leu AO, Woodcroft BJ, Parks DH, Hugenholtz P, Tyson GW. 2019. An evolving view of methane metabolism in the Archaea. *Nat Rev Microbiol* **17**: 219-232.
Felton GW, Chung SH, Hernandez MGE, Louis J, Peiffer M, Tian D. 2018. Herbivore Oral Secretions are the First Line of Protection Against Plant-induced Defences. In *Annual Plant Reviews online*, doi:https://doi.org/10.1002/9781119312994.apr0506, pp. 37-76.
Foster KR, Wenseleers T. 2006. A general model for the evolution of mutualisms. *J Evol Biol* **19**: 1283-1293.
Garcia-Rubio R, de Oliveira HC, Rivera J, Trevijano-Contador N. 2020. The Fungal Cell Wall: Candida, Cryptococcus, and Aspergillus Species. *Front Microbiol* **10**.
Garnier-Sillam E, Toutain F, Villemin G, Renoux J. 1989. Études Préliminaires des Meules Originales du Termite XylophageSphaerotermes sphaerothorax (Sjostedt). *Insectes Sociaux* **36**: 293-312.
Gernhard T. 2008. The conditioned reconstructed process. *J Theor Biol* **253**: 769-778.
Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* **8**.
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.
Graham ED, Heidelberg JF, Tully BJ. 2018. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J* **12**: 1861-1866.
Grantham NJ, Wurman-Rodrich J, Terrett OM, Lyczakowski JJ, Stott K, Iuga D, Simmons TJ, Durand-Tardif M, Brown SP, Dupree R et al. 2017. An even pattern of xylan substitution is critical for interaction with cellulose in plant cell walls. *Nat Plants* **3**: 859-865.
Groussin M, Mazel F, Alm EJ. 2020. Co-evolution and Co-speciation of Host-Gut Bacteria Systems. *Cell Host Microbe* **28**: 12-22.
Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072-1075.
Han Q, Liu N, Robinson H, Cao L, Qian C, Wang Q, Xie L, Ding H, Wang Q, Huang Y et al. 2013. Biochemical characterization and crystal structure of a GH10 xylanase from termite gut bacteria reveal a novel structural feature and significance of its bacterial Ig-like domain. *Biotechnology and Bioengineering* **110**: 3093-3103.
He S, Ivanova N, Kirton E, Allgaier M, Bergin C, Scheffrahn RH, Kyrpides NC, Warnecke F, Tringe SG, Hugenholtz P. 2013. Comparative Metagenomic and Metatranscriptomic Analysis of Hindgut Paunch Microbiota in Wood- and Dung-Feeding Higher Termites. *PLOS ONE* **8**: e61126.

Hervé V, Liu P, Dietrich C, Sillam-Dussès D, Stiblik P, Šobotník J, Brune A. 2020. Phylogenomic analysis of 589 metagenome-assembled genomes encompassing all major prokaryotic lineages from the gut of higher termites. *PeerJ* **8**: e8614.

Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T et al. 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**: 463-467.

Heyn ANJ. 1936. Molecular Structure of Chitin in Plant Cell-Walls. *Nature* **137**: 277-278.

Himmel ME, Bayer EA. 2009. Lignocellulose conversion to biofuels: current challenges, global perspectives. *Curr Opin Biotechnol* **20**: 316-317.

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

Hongoh Y, Deevong P, Hattori S, Inoue T, Noda S, Noparatnaraporn N, Kudo T, Ohkuma M. 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, Moriya S, Trakulnaleamsai S, Ohkuma M, Vongkaluang C, Noparatnaraporn N, Kudo T. 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, Ohkuma M. 2010. Termite Gut Flagellates and Their Methanogenic and Eubacterial Symbionts. In *(Endo)symbiotic Methanogenic Archaea*, doi:10.1007/978-3-642-13615-3_5 (ed. JHP Hackstein), pp. 55-79. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hongoh Y, Sharma VK, Prakash T, Noda S, Toh H, Taylor TD, Kudo T, Sakaki Y, Toyoda A, Hattori M et al. 2008. Genome of an Endosymbiont Coupling N2 Fixation to Cellulolysis Within Protist Cells in Termite Gut. *Science* **322**: 1108-1109.

Hosokawa T, Kikuchi Y, Nikoh N, Shimada M, Fukatsu T. 2006. Strict Host-Symbiont Cospeciation and Reductive Genome Evolution in Insect Gut Bacteria. *PLoS Biol* **4**: e337.

Hu H, da Costa RR, Pilgaard B, Schiøtt M, Lange L, Poulsen M, Tringe SG. 2019. Fungiculture in Termites Is Associated with a Mycolytic Gut Bacterial Community. *mSphere* **4**: e00165-00119.

Hu Y, Sanders JG, Łukasik P, D'Amelio CL, Millar JS, Vann DR, Lan Y, Newton JA, Schotanus M, Kronauer DJC et al. 2018. Herbivorous turtle ants obtain essential nutrients from a conserved nitrogen-recycling gut microbiome. *Nat Commun* **9**: 964-964.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**: D309-D314.

Hungate RE. 1939. Experiments on the Nutrition of Zootermopsis. III. The anaerobic carbohydrate dissimilation by the intestinal protozoa. *Ecology* **20**: 230-245.

Husník F, Chrudimský T, Hypša V. 2011. Multiple origins of endosymbiosis within the Enterobacteriaceae (γ-Proteobacteria): convergence of complex phylogenetic approaches. *BMC Biology* **9**: 87.

Hutchinson MC, Cagua EF, Balbuena JA, Stouffer DB, Poisot T. 2017. paco: implementing Procrustean Approach to Cophylogeny in R. *Methods Ecol Evol* **8**: 932-940.

Ikeda-Ohtsubo W, Strassert JF, Köhler T, Mikaelyan A, Gregor I, McHardy AC, Tringe SG, Hugenholtz P, Radek R, Brune A. 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.

Illumina I. 2017. Effects of index Misassignment on multiplexing and downstream analysis., Vol 2021.

Inoue J-i, Oshima K, Suda W, Sakamoto M, Iino T, Noda S, Hongoh Y, Hattori M, Ohkuma M. 2015. Distribution and evolution of nitrogen fixation genes in the phylum Bacteroidetes. *Microbes Environ* **30**: 44-50.

Inoue T, Kitade O, Yoshimura T, Yamaoka I. 2000. Symbiotic Associations with Protists. In *Termites: Evolution, Sociality, Symbioses, Ecology*, doi:10.1007/978-94-017-3223-9_13 (ed. T Abe, et al.), pp. 275-288. Springer Netherlands, Dordrecht.

Iwadate Y, Kato JI. 2019. Identification of a Formate-Dependent Uric Acid Degradation Pathway in Escherichia coli. *J Bacteriol* **201**.

Ji R, Brune A. 2001. Transformation and mineralization of 14C-labeled cellulose, peptidoglycan, and protein by the soil-feeding termite Cubitermes orthognathus. *Biol Fertil Soils* **33**: 166-174.

Ji R, Brune A. 2005. Digestion of peptidic residues in humic substances by an alkali-stable and humic-acid-tolerant proteolytic activity in the gut of soil-feeding termites. *Soil Biol Biochem* **37**: 1648-1655.

Jousselin E, Desdevises Y, Coeur d'acier A. 2008. Fine-scale cospeciation between Brachycaudus and Buchnera aphidicola: Bacterial genome helps define species and evolutionary relationships in aphids. *Proceedings Biological sciences / The Royal Society* **276**: 187-196.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**: D457-462.

Katoh K, Misawa K, Kuma Ki, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059-3066.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**: 772-780.

Keck F, Rimet F, Bouchez A, Franc A. 2016. phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecol Evol* **6**: 2774-2780.

Khan A, Ahmad W. 2018. *Termites and Sustainable Management*. Springer International Publishing.

Kielak AM, Cretoiu MS, Semenov AV, Sørensen SJ, van Elsas JD. 2013. Bacterial chitinolytic communities respond to chitin and pH alteration in soil. *Appl Environ Microbiol* **79**: 263-272.

Kikuchi Y, Hosokawa T, Fukatsu T. 2007. Insect-Microbe Mutualism without Vertical Transmission: a Stinkbug Acquires a Beneficial Gut Symbiont from the Environment Every Generation. *Appl Environ Microbiol* **73**: 4308-4316.

Kikuchi Y, Hosokawa T, Nikoh N, Meng X-Y, Kamagata Y, Fukatsu T. 2009. Host-symbiont co-speciation and reductive genome evolution in gut symbiotic bacteria of acanthosomatid stinkbugs. *BMC Biology* **7**: 2.

Kikuchi Y, Meng XY, Fukatsu T. 2005. Gut symbiotic bacteria of the genus Burkholderia in the broad-headed bugs Riptortus clavatus and Leptocorisa chinensis (Heteroptera: Alydidae). *Appl Environ Microbiol* **71**: 4035-4043.

Kinjo Y, Lo N, Martín PV, Tokuda G, Pigolotti S, Bourguignon T. 2021. Enhanced Mutation Rate, Relaxed Selection, and the "Domino Effect" are associated with Gene Loss in Blattabacterium, A Cockroach Endosymbiont. *Molecular Biology and Evolution* **38**: 3820-3831.

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.

Kuechler SM, Dettner K, Kehl S. 2011. Characterization of an obligate intracellular bacterium in the midgut epithelium of the bulrush bug *Chilacis typhae* (Heteroptera, Lygaeidae, Artheneinae). *Appl Environ Microbiol* **77**: 2869-2876.

Kuhnigk T, Branke J, Krekeler D, Cypionka H, König H. 1996. A Feasible Role of Sulfate-Reducing Bacteria in the Termite Gut. *Syst Appl Microbiol* **19**: 139-149.

Kundu P, Blacher E, Elinav E, Pettersson S. 2017. Our Gut Microbiome: The Evolving Inner Self. *Cell* **171**: 1481-1493.

Kuwahara H, Yuki M, Izawa K, Ohkuma M, Hongoh Y. 2017. Genome of 'Ca. Desulfovibrio trichonymphae', an H2-oxidizing bacterium in a tripartite symbiotic system within a protist cell in the termite gut. *ISME J* **11**: 766-776.

Kwong WK, Moran NA. 2016. Gut microbial communities of social bees. *Nat Rev Microbiol* **14**: 374-384.

Lan Y, Rosen G, Hershberg R. 2016. Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. *Microbiome* **4**: 18.

Lang K, Schuldes J, Klingl A, Poehlein A, Daniel R, Brunea A. 2015. New mode of energy metabolism in the seventh order of methanogens as revealed by comparative genome analysis of "Candidatus methanoplasma termitum". *Appl Environ Microbiol* **81**: 1338-1352.

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

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948.

Leadbetter JR, Crosby LD, Breznak JA. 1998. *Methanobrevibacter filiformis* sp. nov., a filamentous methanogen from termite hindguts. *Arch Microbiol* **169**: 287-292.

Leadbetter JR, Schmidt TM, Graber JR, Breznak JA. 1999. Acetogenesis from H2 Plus CO2 by Spirochetes from Termite Guts. *Science* **283**: 686-689.

Legendre F, Nel A, Svenson GJ, Robillard T, Pellens R, Grandcolas P. 2015. Phylogeny of Dictyoptera: Dating the Origin of Cockroaches, Praying Mantises and Termites with Molecular Data and Controlled Fossil Evidence. *PLOS ONE* **10**: e0130127.

Lemaitre B, Hoffmann J. 2007. The Host Defense of Drosophila melanogaster. *Annu Rev Immunol* **25**: 697-743.

Lerner A, Matthias T, Aminov R. 2017. Potential Effects of Horizontal Gene Exchange in the Human Gut. *Frontiers in Immunology* **8**.

Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**: W293-W296.
Ley RE, Peterson DA, Gordon JI. 2006. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**: 837-848.
Lilburn TG, Kim KS, Ostrom NE, Byzek KR, Leadbetter JR, Breznak JA. 2001. Nitrogen Fixation by Symbiotic and Free-Living Spirochetes. *Science* **292**: 2495-2498.
Lim SJ, Bordenstein SR. 2020. An introduction to phylosymbiosis. *Proc Royal Soc B* **287**: 20192900.
Lindahl BD, Finlay RD. 2006. Activities of chitinolytic enzymes during primary and secondary colonization of wood by basidiomycetous fungi. *New Phytol* **169**: 389-397.
Liu C, Zou G, Yan X, Zhou X. 2018. Screening of multimeric β-xylosidases from the gut microbiome of a higher termite, Globitermes brachycerastes. *Int J Biol Sci* **14**: 608-615.
Liu N, Li H, Chevrette MG, Zhang L, Cao L, Zhou H, Zhou X, Zhou Z, Pope PB, Currie CR et al. 2019. Functional metagenomics reveals abundant polysaccharide-degrading gene clusters and cellobiose utilization pathways within gut microbiota of a wood-feeding higher termite. *ISME J* **13**: 104-117.
Lo N, Tokuda G, Watanabe H, Rose H, Slaytor M, Maekawa K, Bandi C, Noda H. 2000a. Evidence from multiple gene sequences indicates that termites evolved from wood-feeding cockroaches. *Curr Biol* **23**: 515-538.
Lo N, Tokuda G, Watanabe H, Rose H, Slaytor M, Maekawa K, Bandi C, Noda H. 2000b. Evidence from multiple sequences indicates that termites evolved from wood-feeding cockroaches. *Curr Biol* **10**: 801-804.
Lo W-S, Huang Y-Y, Kuo C-H. 2016. Winding paths to simplicity: genome evolution in facultative insect symbionts. *FEMS Microbiology Reviews* **40**: 855-874.
Loh HQ, Hervé V, Brune A. 2021. Metabolic Potential for Reductive Acetogenesis and a Novel Energy-Converting [NiFe] Hydrogenase in Bathyarchaeia From Termite Guts - A Genome-Centric Analysis. *Front Microbiol* **11**: 635786.
Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**: D490-495.
Łukasik P, Newton JA, Sanders JG, Hu Y, Moreau CS, Kronauer DJC, O'Donnell S, Koga R, Russell JA. 2017. The structured diversity of specialized gut symbionts of the New World army ants. *Mol Ecol* **26**: 3808-3825.
Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD et al. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research* **47**: W636-W641.
Martinson VG, Moy J, Moran NA. 2012. Establishment of characteristic gut bacteria during development of the honeybee worker. *Appl Environ Microbiol* **78**: 2830-2840.
Marynowska M, Goux X, Sillam-Dussès D, Rouland-Lefèvre C, Halder R, Wilmes P, Gawron P, Roisin Y, Delfosse P, Calusinska M. 2020. Compositional and functional characterisation of biomass-degrading microbial communities in guts of plant fibre- and soil-feeding higher termites. *Microbiome* **8**: 96.
Mazel F, Davis KM, Loudon A, Kwong WK, Groussin M, Parfrey LW. 2018. Is Host Filtering the Main Driver of Phylosymbiosis across the Tree of Life? *mSystems* **3**: e00097-00018.
McFall-Ngai M, Hadfield MG, Bosch TCG, Carey HV, Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF et al. 2013. Animals in a bacterial world, a new

imperative for the life sciences. *Proceedings of the National Academy of Sciences* **110**: 3229-3236.
Mendler K, Chen H, Parks DH, Lobb B, Hug LA, Doxey AC. 2019. AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res* **47**: 4442-4448.
Michaud C, Hervé V, Dupont S, Dubreuil G, Bézier AM, Meunier J, Brune A, Dedeine F. 2020. Efficient but occasionally imperfect vertical transmission of gut mutualistic protists in a wood-feeding termite. *Mol Ecol* **29**: 308-324.
Mikaelyan A, Dietrich C, Köhler T, Poulsen M, Sillam-Dussès D, Brune A. 2015a. Diet is the primary determinant of bacterial community structure in the guts of higher termites. *Mol Ecol* **24**: 5284-5295.
Mikaelyan A, Köhler T, Lampert N, Rohland J, Boga H, Meuser K, Brune A. 2015b. Classifying the bacterial gut microbiota of termites and cockroaches: A curated phylogenetic reference database (DictDb). *Syst Appl Microbiol* **38**: 472-482.
Mikaelyan A, Meuser K, Brune A. 2017a. Microenvironmental heterogeneity of gut compartments drives bacterial community structure in wood- and humus-feeding higher termites. *FEMS Microbiology Ecology* **93**.
Mikaelyan A, Strassert JFH, Tokuda G, Brune A. 2014. The fibr… associated cellulolytic bacterial community in the hindgut of woo,, feeding higher termites (Nasutitermes spp.). *Environ Microbiol* **16**: 2711-2722.
Mikaelyan A, Thompson CL, Meuser K, Zheng H, Rani P, Plarre R, Brune A. 2017b. High-resolution phylogenetic analysis of Endomicrobia reveals multiple acquisitions of endosymbiotic lineages by termite gut flagellates. *Environ Microbiol Rep* **9**: 477-483.
Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho LP et al. 2019. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* **10**: 1014.
Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast Approximation for Phylogenetic Bootstrap. *Mol Biol Evol* **30**: 1188-1195.
Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**: 1530-1534.
Minkley N, Fujita A, Brune A, Kirchner WH. 2006. Nest specificity of the bacterial community in termite guts (Hodotermes mossambicus). *Insectes Sociaux* **53**: 339-344.
Misof B Liu S Meusemann K Peters RS Donath A Mayer C Frandsen PB Ware J Flouri T Beutel RG et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**: 763-767.
Miura T, Maekawa K. 2020. The making of the defensive caste: Physiology, development, and evolution of the soldier differentiation in termites. *Evol Dev* **22**: e12335.
Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV, Muller MN, Pusey AE, Peeters M, Hahn BH, Ochman H. 2016. Cospeciation of gut microbiota with hominids. *Science* **353**: 380-382.
Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet* **42**: 165-190.
Moran NA, Ochman H, Hammer TJ. 2019. Evolutionary and Ecological Consequences of Gut Microbial Communities. *Annu Rev Ecol Evol Syst* **50**: 451-475.

Muegge BD, Kuczynski J, Knights D, Clemente JC, González A, Fontana L, Henrissat B, Knight R, Gordon JI. 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332**: 970-974.

Nalepa CA. 1991. Ancestral transfer of symbionts between cockroaches and termites: an unlikely scenario. *Proc Biol Sci* **246**: 185-189.

Nalepa CA. 2017. What Kills the Hindgut Flagellates of Lower Termites during the Host Molting Cycle? *Microorganisms* **5**.

Nalepa CA, Bignell DE, Bandi C. 2001. Detritivory, coprophagy, and the evolution of digestive mutualisms in Dictyoptera. *Insectes Sociaux* **48**: 194-201.

Ngugi DK, Brune A. 2012. Nitrate reduction, nitrous oxide formation, and anaerobic ammonia oxidation to nitrite in the gut of soil-feeding termites (Cubitermes and Ophiotermes spp.). *Environ Microbiol* **14**: 860-871.

Ngugi DK, Ji R, Brune A. 2011. Nitrogen mineralization, denitrification, and nitrate ammonification by soil-feeding termites: a ^{15}N-based approach. *Biogeochemistry* **103**: 355-369.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**: 268-274.

Nishimura Y, Otagiri M, Yuki M, Shimizu M, Inoue JI, Moriya S, Ohkuma M. 2020. Division of functional roles for termite gut protists revealed by single-cell transcriptomes. *ISME J* **14**: 2449-2460.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**: 824-834.

Nyyssönen M, Tran H, Karaoz U, Weihe C, Hadi M, Martiny J, Martiny A, Brodie E. 2013. Coupled high-throughput functional screening and next generation sequencing for identification of plant polymer decomposing enzymes in metagenomic libraries. *Front Microbiol* **4**.

Ochman H, Elwyn S, Moran NA. 1999. Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences* **96**: 12638-12643.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299-304.

Ohkuma M, Brune A. 2011. Diversity,Structure,and Evolution of the Termite Gut Microbial Community. In *Biology of Termites: a Modern Synthesis*, (ed. DE Bignell, et al.). Springer Netherlands.

Ohkuma M, Noda S, Hattori S, Iida T, Yuki M, Starns D, Inoue J, Darby AC, Hongoh Y. 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, Hongoh Y, Kudo T. 2001. Coevolution of symbiotic systems of termites and their gut microorganisms.

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.

Oinonen P, Areskogh D, Henriksson G. 2013. Enzyme catalyzed cross-linking of spruce galactoglucomannan improves its applicability in barrier films. *Carbohydr Polym* **95**: 690-696.

Oksanen J, Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. and Wagner, H. 2014. Vegan: Community Ecology Package. R Package Version 2.2-0.

Olm MR, Crits-Christoph A, Diamond S, Lavy A, Carnevali PBM, Banfield JF, Woyke T. 2020. Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5**: e00731-00719.

Ottesen EA, Leadbetter JR. 2011. Formyltetrahydrofolate synthetase gene diversity in the guts of higher termites with different diets and lifestyles. *Appl Environ Microbiol* **77**: 3461-3467.

Paës G, Berrin J-G, Beaugrand J. 2012. GH11 xylanases: Structure/function/properties relationships and applications. *Biotechnology Advances* **30**: 564-592.

Papudeshi B, Haggerty JM, Doane M, Morris MM, Walsh K, Beattie DT, Pande D, Zaeri P, Silva GGZ, Thompson F et al. 2017. Optimizing and evaluating the reconstruction of Metagenome-assembled microbial genomes. *BMC Genom* **18**: 915.

Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* **38**: 1079-1086.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043-1055.

Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**: 1533-1542.

Paul K, Nonoh JO, Mikulski L, Brune A. 2012. "Methanoplasmatales", Thermoplasmatales-Related Archaea in Termite Guts and Other Environments, Are the Seventh Order of Methanogens. *Appl Environ Microbiol* **78**: 8245-8253.

Perez-Lamarque B, Morlon H. 2019. Characterizing symbiont inheritance during host–microbiota evolution: Application to the great apes gut microbiota. *Mol Ecol Resour* **19**: 1659-1671.

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**: 1261-1270.

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

Pope PB, Denman SE, Jones M, Tringe SG, Barry K, Malfatti SA, McHardy AC, Cheng J-F, Hugenholtz P, McSweeney CS et al. 2010. Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proceedings of the National Academy of Sciences* **107**: 14793-14798.

Potrikus CJ, Breznak JA. 1980. Uric Acid-Degrading Bacteria in Guts of Termites [Reticulitermes flavipes (Kollar)]. *Appl Environ Microbiol* **40**: 117-124.

Potrikus CJ, Breznak JA. 1981. Gut bacteria recycle uric acid nitrogen in termites: A strategy for nutrient conservation. *Proc Natl Acad Sci U S A* **78**: 4601-4605.

Poulsen M, Hu H, Li C, Chen Z, Xu L, Otani S, Nygaard S, Nobre T, Klaubauf S, Schindler PM et al. 2014. Complementary symbiont contributions to plant decomposition in a fungus-farming termite. *Proc Natl Acad Sci USA* **111**: 14500-14505.

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

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**: e9490.

Prins RA, Kreulen DA. 1991. Comparative aspects of plant cell wall digestion in insects. *Animal Feed Science and Technology* **32**: 101-118.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590-596.

Quince C, Nurk S, Raguideau S, James R, Soyer OS, Summers JK, Limasset A, Eren AM, Chikhi R, Darling AE. 2021. STRONG: metagenomics strain resolution on assembly graphs. *Genome Biology* **22**: 214.

Quinn TP, Richardson MF, Lovell D, Crowley TM. 2017. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Sci Rep* **7**: 16252.

R Core Team. 2014. R: A language and environment for statistical computing.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* **67**: 901-904.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

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.

Rouland-Lefèvre C, Inoue T, Johjima T. 2006. Termitomyces/Termite Interactions. In *Intestinal Microorganisms of Termites and Other Invertebrates*, doi:10.1007/3-540-28185-1_14 (ed. H König, A Varma), pp. 335-350. Springer Berlin Heidelberg, Berlin, Heidelberg.

Satler JD, Bernhard KK, Stireman JO, III, Machado CA, Houston DD, Nason JD. 2020. Community Structure and Undescribed Species Diversity in Non-Pollinating Fig Wasps Associated with the Strangler Fig Ficus petiolaris. *Insect Systematics and Diversity* **4**.

Sato T, Hongoh Y, Noda S, Hattori S, Ui S, Ohkuma M. 2009. Candidatus Desulfovibrio trichonymphae, a novel intracellular symbiont of the flagellate Trichonympha agilis in termite gut. *Environ Microbiol* **11**: 1007-1015.

Scharf ME. 2015. Termites as Targets and Models for Biotechnology. *Annu Rev Entomol* **60**: 77-102.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ 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, Chowdhury NP, Müller V. 2018. Complex Multimeric [FeFe] Hydrogenases: Biochemistry, Physiology and New Opportunities for the Hydrogen Economy. *Front Microbiol* **9**.

Schuchmann K, Müller V. 2012. A bacterial electron-bifurcating hydrogenase. *J Biol Chem* **287**: 31165-31171.

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.

Schultz JE, Breznak JA. 1978. Heterotrophic bacteria present in hindguts of wood-eating termites [Reticulitermes flavipes (Kollar)]. *Appl Environ Microbiol* **35**: 930-936.
Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068-2069.
Shimodaira H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. *Systematic Biology* **51**: 492-508.
Shinzato N, Matsumoto T, Yamaoka I, Oshima T, Yamagishi A. 2001. Methanogenic Symbionts and the Locality of their Host Lower Termites. *Microbes Environ* **16**: 43-47.
Shterzer N, Mizrahi I. 2015. The animal gut as a melting pot for horizontal gene transfer. *Can J Microbiol* **61**: 603-605.
Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, Kondrotaite Z, Karst SM, Dueholm MS, Nielsen PH et al. 2021. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat Commun* **12**: 2009.
Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, Chan CKF, Nabhan AN, Su T, Morganti RM et al. 2017. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *bioRxiv* doi:10.1101/125724: 125724.
Slaytor M. 2000. Energy Metabolism in the Termite and Its Gut Microbiota. In *Termites: Evolution, Sociality, Symbioses, Ecology*, doi:10.1007/978-94-017-3223-9_15 (ed. T Abe, et al.), pp. 307-332. Springer Netherlands, Dordrecht.
Smith MR. 2020. Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics* **36**: 5007-5013.
Sommer F, Bäckhed F. 2013. The gut microbiota — masters of host development and physiology. *Nat Rev Microbiol* **11**: 227-238.
Søndergaard D, Pedersen CN, Greening C. 2016. HydDB: A web tool for hydrogenase classification and analysis. *Sci Rep* **6**: 34212.
Song Y, Hervé V, Radek R, Pfeiffer F, Zheng H, Brune A. 2021. Characterization and phylogenomic analysis of Breznakiella homolactica gen. nov. sp. nov. indicate that termite gut treponemes evolved from non-acetogenic spirochetes in cockroaches. *Environ Microbiol* **23**: 4228-4245.
Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**: 1449-1452.
Srivastava A, Malik L, Sarkar H, Zakeri M, Almodaresi F, Soneson C, Love MI, Kingsford C, Patro R. 2020. Alignment and mapping methodology influence transcript abundance estimation. *Genome Biology* **21**: 239.
Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, Liachko I, Snelling TJ, Dewhurst RJ, Walker AW et al. 2018. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun* **9**: 870.
Su N-Y, Scheffrahn RH. 2000. Termites as Pests of Buildings. In *Termites: Evolution, Sociality, Symbioses, Ecology*, doi:10.1007/978-94-017-3223-9_20 (ed. T Abe, et al.), pp. 437-453. Springer Netherlands, Dordrecht.
Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**: vey016.

Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB et al. 2013. Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods* **10**: 1196-1199.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609-W612.

Swift MJ, Heal OW, Anderson JM. 1981. Decomposition in Terrestrial Ecosystems. *The Quarterly Review of Biology* **56**: 96-96.

Syberg-Olsen MJ, Garber AI, Keeling PJ, McCutcheon JP, Husnik F. 2021. Pseudofinder: detection of pseudogenes in prokaryotic genomes. *bioRxiv* doi:10.1101/2021.10.07.463580: 2021.2010.2007.463580.

Tai V, Carpenter KJ, Weber PK, Nalepa CA, Perlman SJ, Keeling PJ. 2016. Genome Evolution and Nitrogen Fixation in Bacterial Ectosymbionts of a Protist Inhabiting Wood-Feeding Cockroaches. *Appl Environ Microbiol* **82**: 4682-4695.

Tai V, James ER, Nalepa CA, Scheffrahn RH, Perlman SJ, Keeling PJ. 2015. The role of host phylogeny varies in shaping microbial diversity in the hindguts of lower termites. *Appl Environ Microbiol* **81**: 1059-1070.

Tamura K, Hemsworth GR, Déjean G, Rogers TE, Pudlo NA, Urs K, Jain N, Davies GJ, Martens EC, Brumer H. 2017. Molecular Mechanism by which Prominent Human Gut Bacteroidetes Utilize Mixed-Linkage Beta-Glucans, Major Health-Promoting Cereal Polysaccharides. *Cell Rep* **21**: 417-430.

Tauzin AS, Kwiatkowski KJ, Orlovsky NI, Smith CJ, Creagh AL, Haynes CA, Wawrzak Z, Brumer H, Koropatkin NM. 2016. Molecular Dissection of Xyloglucan Recognition in a Prominent Human Gut Symbiont. *mBio* **7**: e02134-02115.

Terrapon N, Lombard V, Gilbert HJ, Henrissat B. 2015. Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics* **31**: 647-655.

Thauer RK, Kaster AK, Seedorf H, Buckel W, Hedderich R. 2008. Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat Rev Microbiol* **6**: 579-591.

Tholen A, Brune A. 1999. Localization and in situ activities of homoacetogenic bacteria in the highly compartmentalized hindgut of soil-feeding higher termites (Cubitermes spp.). *Appl Environ Microbiol* **65**: 4497-4505.

Thong-On A, Suzuki K, Noda S, Inoue J, Kajiwara S, Ohkuma M. 2012. Isolation and characterization of anaerobic bacteria for symbiotic recycling of uric acid nitrogen in the gut of various termites. *Microbes Environ* **27**: 186-192.

Thongaram T, Hongoh Y, Kosono S, Ohkuma M, Trakulnaleamsai S, Noparatnaraporn N, Kudo T. 2005. Comparison of bacterial communities in the alkaline gut segment among various species of higher termites. *Extremophiles* **9**: 229-238.

Tokuda G. 2019. Chapter Three - Plant cell wall degradation in insects: Recent progress on endogenous enzymes revealed by multi-omics technologies. In *Advances in Insect Physiology*, Vol 57 (ed. R Jurenka), pp. 97-136. Academic Press.

Tokuda G, Lo N, Watanabe H, Arakawa G, Matsumoto T, Noda H. 2004. Major alteration of the expression site of endogenous cellulases in members of an apical termite lineage. *Mol Ecol* **13**: 3219-3228.

Tokuda G, Mikaelyan A, Fukui C, Matsuura Y, Watanabe H, Fujishima M, Brune A. 2018. Fiber-associated spirochetes are major agents of hemicellulose degradation in the hindgut of wood-feeding higher termites. *Proc Natl Acad Sci USA* **115**: E11996-E12004.

Tomme P, Driver D, Amandoron E, Miller R, Antony R, Warren J, Kilburn D. 1995. Comparison of a fungal (family I) and bacterial (family II) cellulose- binding domain. *Journal of bacteriology* **177**: 4356-4363.

Treitli SC, Kolisko M, Husník F, Keeling PJ, Hampl V. 2019. Revealing the metabolic capacity of Streblomastix strix and its bacterial symbionts using single-cell metagenomics. *Proceedings of the National Academy of Sciences* **116**: 19675-19684.

Uritskiy G, DiRuggiero J. 2019. Applying Genome-Resolved Metagenomics to Deconvolute the Halophilic Microbiome. *Genes* **10**: 220.

Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**: 158.

Vavre F, Kremer N. 2014. Microbial impacts on insect evolutionary diversification: from patterns to mechanisms. *Curr Opin Insect Sci* **4**: 29-34.

Větrovský T, Baldrian P. 2013. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLOS ONE* **8**: e57923.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P. 2007. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**: 1126-1130.

Vu VQ. 2011. ggbiplot: A ggplot2 based biplot. R package version 0.55.

Wang M, Buček A, Šobotník J, Sillam-Dussès D, Evans TA, Roisin Y, Lo N, Bourguignon T. 2019. Historical biogeography of the termite clade Rhinotermitinae (Blattodea: Isoptera). *Mol Phylogenet Evol* **132**: 100-104.

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261-5267.

Warnecke F, Luginbühl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N et al. 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**: 560-569.

Watanabe D, Gotoh H, Miura T, Maekawa K. 2014. Social interactions affecting caste development through physiological actions in termites. *Front physiol* **5**.

Watanabe H, Noda H, Tokuda G, Lo N. 1998. A cellulase gene of termite origin. *Nature* **394**: 330-331.

Watanabe H, Tokuda G. 2010. Cellulolytic systems in insects. *Annu Rev Entomol* **55**: 609-632.

Wenzel M, Schönig I, Berchtold M, Kämpfer P, König H. 2002. Aerobic and facultatively anaerobic cellulolytic bacteria from the gut of the termite Zootermopsis angusticollis. *J Appl Microbiol* **92**: 32-40.

Wiens JJ, Chippindale PT, Hillis DM. 2003. When Are Phylogenetic Analyses Misled by Convergence? A Case Study in Texas Cave Salamanders. *Systematic Biology* **52**: 501-514.

Wu D, Jospin G, Eisen JA. 2013. Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PLOS ONE* **8**: e77033.

Wu H. 2018. Characterizing xylan-degrading enzymes from a putative xylan utilization system derived from termite gut metagenome. Vol PhD Thesis. INSA de Toulouse, France.

Xie H, Yang C, Sun Y, Igarashi Y, Jin T, Luo F. 2020. PacBio Long Reads Improve Metagenomic Assemblies, Gene Catalogs, and Genome Binning. *Front genet* **11**.

Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, Henrissat B, Coutinho PM, Minx P, Latreille P et al. 2007. Evolution of Symbiotic Bacteria in the Distal Human Intestine. *PLoS Biol* **5**: e156.

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

Yamin MA. 1981. Cellulose Metabolism by the Flagellate *Trichonympha* from a Termite Is Independent of Endosymbiotic Bacteria. *Science* **211**: 58-59.

Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* **40**: W445-451.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* **8**: 28-36.

Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* **46**: W95-W101.

Zhang XZ, Sathitsuksanoh N, Zhang YH. 2010. Glycoside hydrolase family 9 processive endoglucanase from Clostridium phytofermentans: heterologous expression, characterization, and synergy with family 48 cellobiohydrolase. *Bioresour Technol* **101**: 5534-5538.

Zheng H, Dietrich C, Brune A. 2017. Genome Analysis of Endomicrobium proavitum Suggests Loss and Gain of Relevant Functions during the Evolution of Intracellular Symbionts. *Appl Environ Microbiol* **83**: e00656-00617.