Czech University of Life Sciences Prague
Faculty of Tropical AgriSciences

Genomics of termites and their symbiotic microorganisms

DISSERTATION THESIS

MSc. Tereza Beránková

Professor Jan Šobotník, PhD

Associated professor Thomas Bourguignon, PhD

2025

I hereby declare that I have completed this thesis entitled "**Genomics of termites and their symbiotic microorganisms"** independently, all texts in this thesis are original, and that all information sources have been quoted and acknowledged by means of complete references. I also confirm that this work has not been previously submitted, nor is it currently submitted, for any other degree, to this or any other university.

In Prague date

.....

Name of the student

Acknowledgments

I would like to express my deepest gratitude to my supervisors, **Professor Jan Šobotník** and **Associate Professor Thomas Bourguignon**, for their invaluable guidance, support, and encouragement throughout my doctoral studies. Their expertise and insight have been instrumental in guiding this research and have provided me with the foundation to grow as a researcher. I am also profoundly thankful to **Jigyasa Arora** for her exceptional mentorship and for her expertise particularly in biostatistics, which significantly contributed to the analytical aspects of this work.

To my family and friends, I owe immense gratitude for their unwavering support, patience, and understanding throughout this journey. Their encouragement has been my foundation during challenging times, and their belief in me has been a constant source of strength.

Abstract and keywords

Termites are eusocial cockroaches with a substantial ecological impact in warm terrestrial habitats. Termites play a central role in lignocellulose decomposition and nutrient cycling, shaping ecosystems in tropical and subtropical regions. Their ability to digest dead vegetal matter depends on carbohydrate-active enzymes (CAZymes) produced by symbiotic gut bacteria with which they have co-evolved over long geological timescales. However, the evolutionary origins of these CAZymes, whether rooted in ancestral gut bacterial genomes and vertically transmitted, or acquired more recently from environmental bacteria, remain uncertain.

This research examines the evolutionary interplay between termites and their gut prokaryota, focusing on the CAZymes critical for lignocellulose digestion. We analyzed the metagenomes of 195 termite species and a wood-feeding cockroach of the genus *Cryptocercus* and identified 420 termite-specific clusters within 81 bacterial CAZyme gene trees. Of these, 404 clusters showed strong cophylogenetic patterns with termites, while 131 clusters contained sequences associated with *Cryptocercus* or *Mastotermes*, which represent the sister lineage to all other termites.

The study shows that numerous bacterial CAZymes have been conserved in the termite gut microbiota since the earliest stages of termite evolution, underscoring a deep-rooted symbiotic relationship. This co-evolutionary association enables termites to effectively digest lignocellulose, highlighting their critical ecological role in nutrient cycling.

Objective of the thesis

This research investigates the cophylogeny and acquisition of bacteria and carbohydrate-active enzymes (CAZymes) within termite gut microbiota. It focuses on analysing the gut metagenomes from 196 termite samples representing the termite phylogeny to uncover patterns of microbial co-evolution and functional adaptation. By examining the evolutionary dynamics of CAZymes, which are essential for lignocellulose digestion, this study also explores the processes that have shaped the acquisition and diversification of bacteria within termite gut ecosystems, providing insights into their ecological and evolutionary significance.

The specific objectives of this thesis are:

- To analyse how the composition of CAZymes and prokaryote varies across termite species with different dietary specializations (e.g., wood-eating vs. soil-eating termites).
- To investigate whether some CAZymes, like certain bacterial taxa, are specific to the termite environment or shared with non-termite environments.
- To analyse CAZyme clusters from termite and non-termite environments to determine whether vertical or horizontal transfer is the dominant mechanism:
    - If CAZymes are predominantly inherited vertically, their sequences will cluster distinctly within termite lineages and dietary groups, separate from non-termite environments.
    - If horizontal transfer is a dominant mechanism, CAZyme sequences from termites will exhibit mixed clustering with those from non-termite environments, indicating frequent external acquisition.
- To examine the distribution of CAZyme families across bacterial taxa within termite gut microbiota, identifying whether CAZyme abundance is concentrated in specific bacterial groups or shared across multiple taxa.

This research employs comparative metagenomic analyses of termite and non-termite environments to elucidate the role of prokaryote and CAZymes in termite evolution and their contribution to the ecological success of termites and their symbiotic gut microbiota.

List of Abbreviations

AA - Auxiliary Activities

CAZymes - carbohydrate-active enzymes

CBM - Carbohydrate Binding Modules

CE - Carbohydrate Esterases

GH - Glycoside Hydrolases

GT - Glycoside Transferases

HGT- Horizontal Gene Transfers

HMM - Hidden Markov Model

MAGS - Metagenome Assembled Genomes

ML - Maximum likelihood

UCE - Ultra Conserved Elements

TSC – Termite-Specific Cluster

FC – Fungus-Cultivating (termite feeding group)

WF – Wood-Feeders

SF – Soil-Feeders

MAG – Metagenome-Assembled Genome *(singular form of MAGS)*

COG – Cluster of Orthologous Groups

VGT - Vertical Gene Transfers

List of figure

List of tables

Table of Contents

1. Literature survey

1.1 Termite introduction and eusociality

Eusociality is defined by cooperative brood care, overlapping generations within a colony, and a division of labour into reproductive and non-reproductive groups (Wilson, 1971). Termites are eusocial insects known for their complex colonies, remarkable ability to decompose wood and vegetal matters. The eusociality has evolved independently in several insect groups, including termites, ants, bees, and wasps. Termites evolved in the Cretaceous period, approximately 170 million years ago, when they diverged from ancestors of wood-eating *Cryptocercus* (Blattodea). The group includes about 3,000 described species (Lo et al., 2000; Krishna et al., 2013; Bourguignon et al., 2015.). Based on the fossil evidence, termites underwent significant diversification approximately 130 million years ago, and additional radiation occurred 54 million years ago when the Termitidae family, known as "higher" termites, diverged from other termites and lost protists from their guts (Bourguignon et al., 2015; Engel et al., 2016; Inward et al., 2007).

Termites play a significant ecological role, primarily in tropical and subtropical environments, where they are important in nutrient cycling, soil watering, aerating, and drainage (Ashton et al., 2019). Their ability to decompose dead plant matter makes them ecological engineers as well as serious pests (Bignell et al., 2011). Termites and their symbiotic organisms living in their gut are essential for breaking down lignocellulose, the most abundant biopolymer on the Earth, and thus converting the dead phytomass into simpler compounds that enrich the soil. Furthermore, termites significantly influence soil properties by enhancing soil porosity, water infiltration, and organic matter content. These activities improve soil fertility and structure, benefiting plant growth and agricultural productivity (Jouquet et al., 2011).

Termites have revealed a limited resilience to environmental stress, especially temperature extremes, and are in their distribution restricted by isotherm +10°C (Emerson et al., 1955) Termites, have perfected social structures through advanced communication mechanisms, including pheromone use, which facilitates colony coordination and defence, highlighting their evolutionary ingenuity (Bordereau and Pasteels, 2011). While termites contribute positively to natural ecosystems, they also challenge human structures and agriculture,

leading to significant economic losses. Balancing their ecological roles and their impact on human activities requires the development of sustainable management strategies (Rust and Su, 2012; Su and Scheffrahn, 1990).

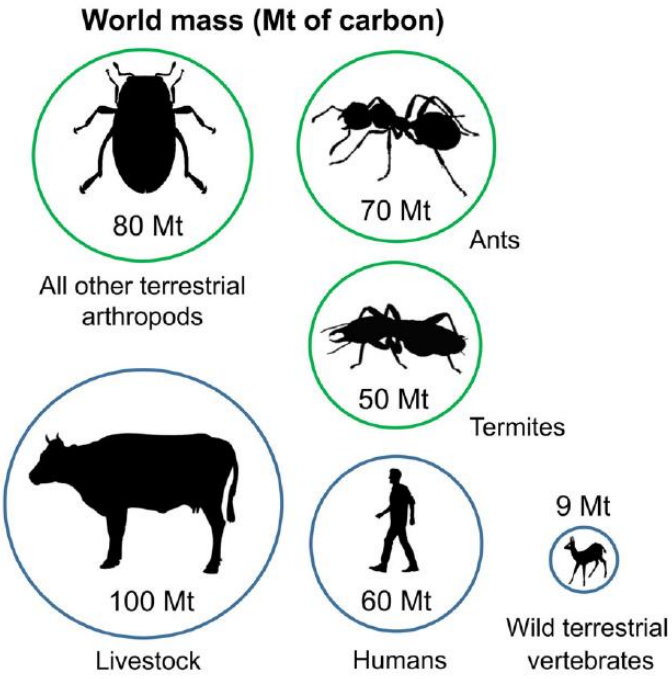


Figure 1: Comparison of global dry biomass (in megatons of carbon).

Ants and termites, distinct insect groups with eusocial organization, frequently interact in ecosystems, primarily as predator and prey. While phylogenetically distant, their high abundance and biomass make their interactions crucial for ecosystem processes, particularly for organic matter decomposition, nutrient cycling, and the redistribution of soil nutrients (Tuma et al., 2020). In some regions, termites exhibit remarkable population densities, with colony numbers reaching up to 70 million individuals per hectare in specific biotopes. Despite their ecological importance, ants and termites mostly interact through predation, with ants regulating termite populations and indirectly influencing ecosystem services such as litter decomposition and nutrient availability (Tuma et al., 2020).

Termites are dominant insects in tropical forests, where their abundance can lead to up to 60% of total macrofauna (Dahlsjö et al., 2014). Due to the high abundance of individuals, termites can process up to 60% of litter and dead wood annually (Ashton et al., 2019). Termites, despite their concealed lifestyle, frequently interact with ants, which are often their predators. Ant lineages such as the African *Dorylus* (army ants) are known for conducting

regular raids on termite colonies, as highlighted by Brady et al. (2014). These raids showcase the predatory dominance of army ants in tropical ecosystems, where they can significantly impact termite populations and behaviors. Another example is *Neoponera marginata*, which paralyzes termites and stores them as a living food supply, reflecting a unique adaptation for survival (Leal and Oliveira, 1995). Within the ant subfamilies *Ponerinae*, *Dorylinae*, and *Myrmicinae*, several species exhibit specialized predatory behaviours targeting termites. The interactions between these ants and termites have ecological and evolutionary significance, as they influence termite colony structure, defence mechanisms, and overall survival strategies. This predatory-prey relationship underscores the complex interdependencies within tropical ecosystems and highlights the role of ants as both predators and regulators of termite populations (Brady et al., 2014).

The origin of eusociality in termites is reflected in several different theories. The first theory, the “symbiont transfer hypothesis” (Cleveland, 1925; Nalepa, 1994), emphasizes the dependence of termites on symbiosis with microorganisms, as termites must be “inoculated” by the symbiotic culture after each molt (cuticle shedding event) by proctodaeal trophallaxis (liquid food exchange through the anus) (Bignell et al., 2011). The second theory emphasizes another specific means of termite reproduction, which is called "cyclic inbreeding" theory (Bartz, 1979). It is based on high inbreeding that increases individuals' relatedness and their mutual altruism (Thorne, 1997). Another theory, "chromosomal linkage", points to the presence of chromosomal chains that connect to the Y chromosome and can span up to half of the termite genome, causing sex-specific variation in the relatedness within a colony (Lacy, 1980). The ‘intraspecific conflict theory’ proposes that parental or sibling manipulation often leads to wing pad damage in nymphs (brachypterous individuals developing into winged imagoes), thereby altering their developmental trajectory. This alteration delays the emergence of imaginal characteristics, rendering the individual incapable of leaving the colony. As a result, these nymphs transition into the worker caste, contributing to the division of labour within the colony (Roisin 1994). However, none of these theories sufficiently explain the evolution of eusociality in termites, likely a combination of various factors from the above-mentioned theories played a role.

1.2 Termite evolution and phylogeny

The evolution of termites is characterized by a complex path from solitary to sophisticated eusocial insects, accompanied by significant morphological and behavioural adaptations to the new functions in food acquisition, defence of societies and protective structures construction. The ecological success of termites in tropical and subtropical environments is largely attributed to their unique symbiotic relationship with gut microorganisms, which facilitate the use of cellulose as a primary food source (Brune, 2014). The phylogenetic analyses conducted by Lo et al. (2000) and further supported by Inward, Beccaloni, and Eggleton (2007), provide compelling molecular evidence for the lineage comprising termites and wood-feeding cockroaches, elucidating the evolutionary origins of eusociality within the Isoptera.

Termites are split into two ecological (but not phylogenetic) groups, “lower” termites and “higher” termites, based on their symbiotic relationships and gut microbiota composition. “Lower” termites feed only on wood or dry grass, and rely predominantly on symbiotic protozoa living within their hindguts to aid in the digestion of cellulose from wood and other plant materials (Brune, 2014; Krishna & Weesner, 1970). The “lower” termites include families Mastotermitidae, Archotermopsidae, Hodotermopsidae, Hodotermitidae, Stolotermitidae, Kalotermitidae, Stylotermitidae, Serritermitidae, Rhinotermitidae, Termitogetonidae, Psammotermitidae and Heterotermitidae (Hellemans et al., 2024). In contrast, “higher” termites or Termitidae rely exclusively upon Prokaryotes in lignocellulose digestion. “Higher” termites consume a wide range of plant materials irrespective of their degree of decomposition, including wood, leaf litter, grass, soil, and herbivore dung, with some species cultivating fungi (Termitidae: Macrotermitinae) in their nests (Brune, 2014; Korb & Heinze, 2016).

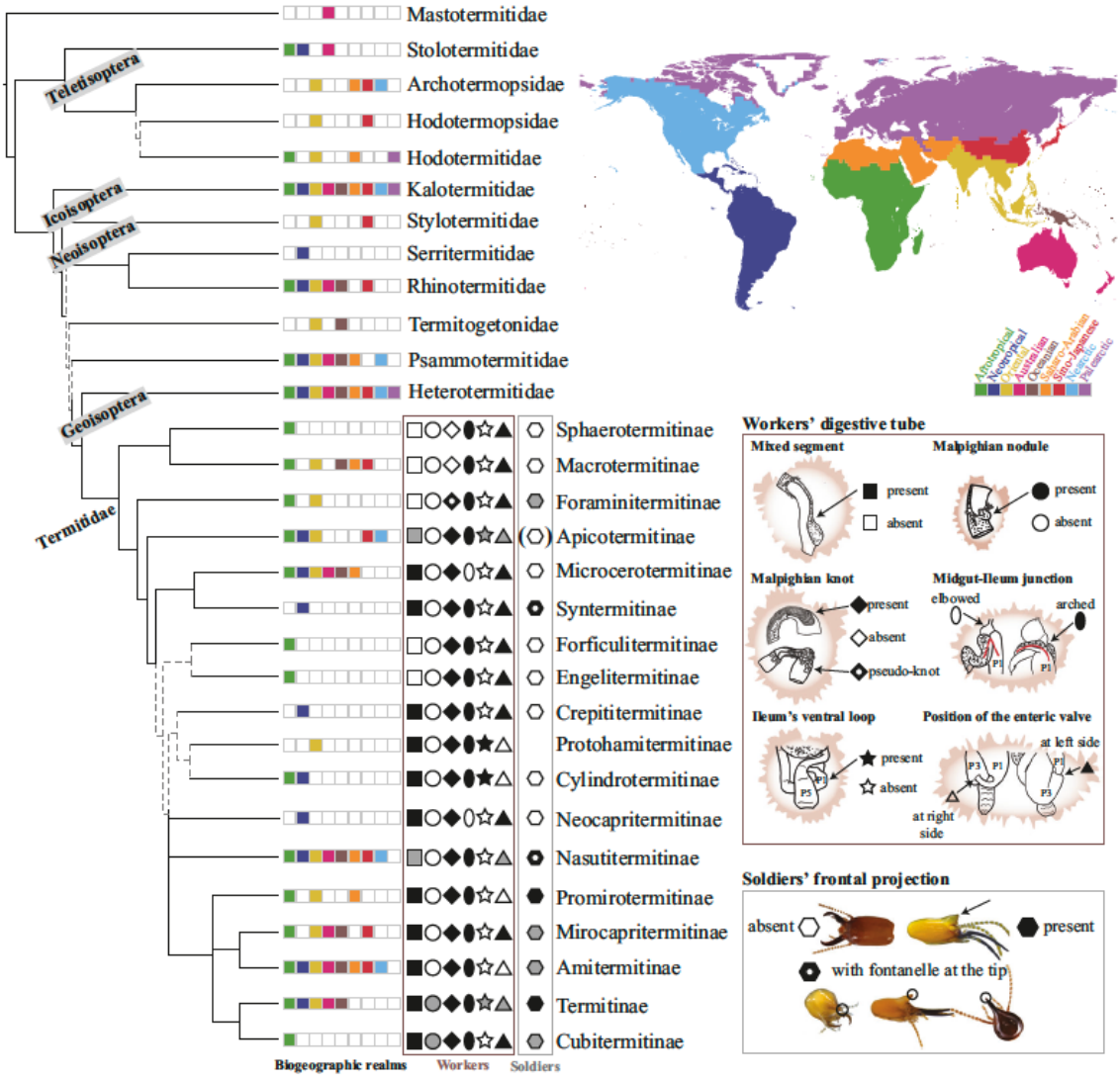


Figure 2: cladogram of Isoptera (Hellemans et al., 2024)

1.2.1 Mastotermitidae

The family Mastotermitidae is the sister group to all other termites with its divergence from the rest of the termites estimated to have occurred about 130 million years ago (Bourguignon et al. 2015; Buček et al. 2019). This family exhibits symplesiomorphic traits shared with cockroaches, such as the anal lobe of the hind wing, egg-laying in oothecae, and the presence of symbiotic bacteria (genus *Blattabacterium*) in the fat body. It also exhibits derived traits such as bifurcated ontogeny, extensive colony formation, advanced alarm communication, and a unique feature—multi-flagellate sperm cells (Krishna & Weesner 1969; Roisin 2000; Delattre et al. 2015). Fossil evidence shows that this family occurred across all continents, but the last living species of this family, *Mastotermes darwiniensis*, is found only in tropical

Australia (Krishna et al. 2013). This termite species exhibits a subterranean lifestyle, primarily feeds on dead wood, and can cause significant damage to wooden structures and buildings (Gay & Watson 1982).

1.2.2 Stolotermitidae

The family includes genera *Stolotermes* and *Porotermes* with a total of ten species (Krishna et al. 2013). They inhabit southern South America and Africa, East Australia, Tasmania, and New Zealand. They feed on damp wood and, although they are categorized among one-piece termites, they are capable of colonizing other food sources (Bignell & Eggleton 2000). Stolotermitidae, like Archotermopsidae, exhibit symplesiomorphic features, such as wings with rich venation, genital structure, and long cerci (appendages at the end of the abdomen) (Krishna & Weesner 1969).

1.2.3 Archotermopsidae

Representatives of this family belong to the genera *Archotermopsis* and *Zootermopsis* comprising a total of four species (Wang 2022). Along with the Stolotermitidae, this family also shares symplesiomorphic traits such as the presence of cerci and richly veined wings. (Wischnitzer 2013). Colonies of this family are formed by a smaller number of individuals, all of whom try to reproduce either as a neotenics (immature reproductive individuals), alates (winged adults), or fertile soldiers. In the species *Archotermopsis wroughtoni*, an adult colony may only consist of 40 individuals (Shellman-Reeve 1997). The family Archotermopsidae has a disjoint distribution, with representatives found in the Himalayas (*Archotermopsis wroughtoni*) or from Mexico to southern Canada (*Zootermopsis* spp.) (Krishna et al. 2013).

1.2.4 Hodotermopsidae

Representatives of this family belong to the genus *Hodotermopsis*, which includes a single species, *H. sjostedti*, distributed across East Asia (Wang 2022). The family Hodotermopsidae, newly elevated to familial rank, represents a distinct lineage within the early diverging termite clades. This family, previously considered a subfamily of Archotermopsidae, includes the genus Hodotermopsis. Hodotermopsidae is closely related to the families Hodotermitidae and Archotermopsidae, collectively forming part of the broader clade Teletisoptera. The divergence of Hodotermopsidae from its closest relatives occurred approximately 90 million

years ago, highlighting its ancient origins and evolutionary significance. Members of Hodotermopsidae are characterized by their adaptation to dampwood habitats, typically coniferous wood, and share ecological niches distinct from those of their relatives in arid and temperate regions. The genus Hodotermopsis, now placed within Hodotermopsidae, showcases these ecological specializations and emphasizes the evolutionary trajectory of this group. Phylogenetic analyses confirm the monophyly of Hodotermopsidae, differentiating it from the paraphyletic Archotermopsidae and supporting its classification as a separate family (Wang et al., 2022).

1.2.5 Hodotermitidae

The Hodotermitidae, also known as “harvester termites”, primarily inhabit arid and semi-arid regions, including deserts of Africa, the Middle East, and South Asia. This family comprises genera such as *Hodotermes*, *Microhodotermes*, and *Anacanthotermes*, which are specialized for feeding predominantly on dry grasses. These termites play a crucial ecological role in nutrient cycling within these regions, particularly through the decomposition of plant material in water-limited environments. Phylogenetic studies, including time-calibrated analyses, suggest that Hodotermitidae diverged from their closest relatives approximately 90 million years ago during the late Cretaceous period. The divergence within the family, such as the split between *Hodotermes* and *Microhodotermes*, occurred during the Oligocene, around 31 million years ago, potentially influenced by the expansion of arid biomes. The biogeographic disjunction between African and Asian lineages aligns with historical land bridges, such as the Gomphotherium land bridge connecting Africa and Eurasia approximately 18–20 million years ago (Wang et al., 2022).

1.2.6 Kalotermitidae

Kalotermitidae, commonly referred to as drywood termites, are the second largest family of termites, comprising approximately 450 described species across 23 genera (Buček et al., 2022; Krishna et al., 2013). Fossil evidence suggests their lineage dates back to the Late Cretaceous, around 99 million years ago, while molecular phylogenies place their divergence from ancestral termites at approximately 115 million years ago. The crown group Kalotermitidae likely emerged around 84 million years ago during the final breakup of Gondwana, highlighting their evolutionary adaptation to distinct ecological niches (Engel et

al., 2009; Buček et al., 2022). Kalotermitidae exhibit unique nesting and foraging behaviors, establishing small colonies within single pieces of wood, such as dead branches or logs (Eggleton, 2000). This restricted nesting behaviour limits their foraging range but facilitates long-distance dispersal, particularly via transoceanic rafting. Such dispersal methods have enabled Kalotermitidae to colonize isolated ecosystems, including remote islands like the Krakatau Islands after volcanic eruptions (Miura et al., 1998). Combined with human-mediated dispersal through the timber trade, species such as *Cryptotermes brevis* have established global populations far beyond its native ranges (Scheffrahn & Su, 2000; Buček et al., 2022). The global distribution of Kalotermitidae is primarily concentrated in tropical and subtropical regions between latitudes 45°N and 45°S (Thorne et al., 2000). Phylogenetic studies confirm the monophyly of this family, revealing that early lineages retained ancestral traits such as foraging across multiple pieces of wood, while modern species evolved to nest exclusively within single pieces of wood (Buček et al., 2022). Their diet and nesting strategy limit colony size, with most individuals retaining reproductive potential, while sterile soldiers play a defensive role consistent with their restricted habitat (Nalepa, 2017).

1.2.7 Stylotermitidae

The family Stylotermitidae is a member of Neoisoptera (characterized by the presence of a frontal gland), which also includes Serritermitidae and Rhinotermitidae, with Stylotermitidae occupying a basal position in the phylogenetic tree (Buček et al., 2019). It is among the least explored families, as it feeds exclusively on hardwood at the boundary between dead and living tissues, which is rather unusual for termites. This family contains a single extant genus, *Stylotermes*, with 45 described species living in Southeast Asia only (Krishna et al. 2013).

1.2.8 Serritermitidae

This Neotropical family contains two genera (*Serritermes, Glossotermes*) and three described species (Krishna et al. 2013). This family exhibits gender specialization - all workers and soldiers are males (Barbosa & Constantino 2017; Bourguignon et al. 2009). *Serritermes serrifer* is the only obligatory inquiline among “lower” termites, while the genus *Glossotermes* feeds on dry red rot wood in the Amazon basin (Emerson et al., 1955).

1.2.9 Rhinotermitidae

The family Rhinotermitidae has undergone significant revisions based on molecular phylogenetic studies and unlike other families, this family was previously considered polyphyletic (Bourguignon et al., 2015; Donovan et al., 2000; Inward et al., 2007). Under the revised classification, Rhinotermitidae sensu novo includes genera such as *Acorhinotermes*, *Dolichorhinotermes*, *Parrhinotermes*, *Rhinotermes*, and *Schedorhinotermes*. Soldiers in this family exhibit diverse morphologies, including monomorphic, dimorphic, and trimorphic forms. Their geographic distribution spans the Australian, Afrotropical, Neotropical, Oriental, Palaearctic, and Oceanian realms, highlighting their ecological adaptability (Hellemans et al., 2024; Wang et al., 2022).

1.2.10 Termitogetonidae

The family Termitogetonidae, elevated from subfamily rank (formerly Termitogetoninae), represents a monophyletic group with unique morphological and ecological characteristics. This family, which is monogeneric, contains the genus *Termitogeton* and is distributed primarily in the Oriental and Oceanian realms. Members of this family are characterized by their wood-feeding behaviour and distinct morphological traits, such as soldiers with heart-shaped heads and elongated, marginally toothless mandibles. The pronotum in imagoes is small, and both imagoes and soldiers are densely hairy, with dorso-ventrally flattened bodies (Hellemans et al., 2024).

1.2.11 Psammotermitidae

The family Psammotermitidae, previously part of Rhinotermitidae, was elevated to family status based on molecular and morphological evidence, confirming its monophyly and distinct evolutionary lineage. Psammotermitidae includes genera such as *Psammotermes* and *Prorhinotermes*. *Psammotermes* is primarily found in arid and semi-arid regions, particularly deserts, where it is well-adapted to extreme environmental conditions. Its ability to survive in such habitats reflects significant ecological specialization. In contrast, *Prorhinotermes* is commonly associated with coastal and insular habitats, where it has been observed using driftwood for dispersal, enabling long-distance colonization of islands. Phylogenetic studies suggest that Psammotermitidae diverged from other termite lineages approximately 80–90 million years ago during the Late Cretaceous. Their unique adaptations, such as their foraging

and nesting behaviours, highlight their ecological significance in nutrient cycling and decomposition within dry and resource-scarce environments.

1.2.12 Heterotermitidae

The family Heterotermitidae, elevated from its previous status as a subfamily within Rhinotermitidae, represents a monophyletic group of subterranean termites widely distributed in tropical, subtropical, and temperate regions. This family includes economically significant genera such as *Coptotermes*, *Heterotermes*, and *Reticulitermes*, known for their wood-feeding behaviour and substantial role in nutrient cycling. However, many species, like *Coptotermes formosanus* (Formosan subterranean termite), are among the most destructive structural pests worldwide. Heterotermitidae termites are characterized by narrow-mandible soldiers and alates with simple wing venation, adaptations that align with their subterranean lifestyle. Their evolutionary divergence, estimated to have occurred during the Late Cretaceous to early Paleogene, highlights their adaptability to diverse habitats and ecological niches. The family's ecological importance as decomposers is offset by their significant economic impact, making them a crucial focus for both ecological studies and pest management strategies (Hellemans et al., 2024; Wang et al., 2022).

1.2.13 Termitidae

The family Termitidae, known as "higher" termites, represents the most diverse and evolutionarily advanced group within the termite order. This family includes more than 2,000 described species across 18 subfamilies, making it the largest family within Isoptera. Termitidae are globally distributed, with their greatest diversity found in tropical and subtropical regions, where they play critical roles in various ecosystems. One of the defining characteristics of Termitidae is the absence of protozoan symbionts in their guts, a trait that distinguishes them from "lower" termites. Instead, they rely on bacterial and, in some cases, fungal symbionts for digestion.

Sphaerotermitinae are small, compact termites found in tropical regions. They are notable for their unique nesting behaviour, constructing spherical nests that are distinct from those of other subfamilies. Within these nests, they create bacterial combs, sponge-like structures that support the cultivation of bacteria. These bacterial gardens facilitate the decomposition of organic matter, enabling the termites to efficiently process a wider range of food resources.

Figure 3: nest and queen with soldiers and workers of Sphaerotermitinae

Macrotermitinae, known as fungus-growing termites, are widely distributed in Africa and Asia. This subfamily is characterized by its mutualistic relationship with *Termitomyces* fungi, cultivated within termite nests. Prominent genera include *Macrotermes*, *Odontotermes*, and *Microtermes*.

Figure 4: Termitomyces and soldier of Macrotermitinae

Foraminitermitinae is a relatively understudied subfamily and is notable for its members' ability to create intricate nesting structures, often reflecting adaptations to their environment.

Apicotermitinae rely on highly specialized gut symbionts for digestion. **Microcerotermitinae**, which includes the genus *Microcerotermes*, consists of small-sized termites distributed throughout tropical regions.

Syntermitinae, primarily distributed in South America, include genera such as *Syntermes* and *Cornitermes*. Known for their specialized nesting and foraging behaviours, some species build prominent mounds.

Forficulitermitinae is characterized by its distinct soldier mandibles, which are elongated. Individuals feed on the soil and some of them can be found in bare soil.

Engelitermitinae was recently discovered with species, *Engelitermes zambo*.

Figure 5: Engelitermes zambo

Neocapritermitinae includes termites with specialized adaptations, particularly in their soldier caste. A notable species within this subfamily is *Neocapritermes taracua*, which exhibits a remarkable defence mechanism. Soldiers of this species possess a unique defensive adaptation involving specialized pouches on their bodies filled with a blue liquid containing toxic compounds. When threatened, they rupture these pouches, releasing the toxic substance to deter predators.

Figure 6: Neocapritermes Taracua

Crepititermitinae, Protohamitermitinae, and Cylindrotermitinae are less commonly discussed subfamilies, with limited ecological and morphological data. Promirotermitinae and Mirocapritermitinae.

Amitermitinae, represented by the genus *Amitermes*, is widely distributed in tropical and subtropical regions. These termites exhibit significant ecological versatility and play important roles in their ecosystems. Termitinae, the largest subfamily within Termitidae, includes diverse genera such as *Microcerotermes* and *Quasitermes*. This subfamily is found across tropical and subtropical zones, displaying a wide range of ecological behaviours and

adaptations. Cubitermitinae are prominent in African savannahs, where they are recognized for their mound-building behaviour.

Figure 7: Cubitermitinae

1.3 Termite castes and life cycle

All termite species are eusocial and usually live in colonies with only one pair of reproductive, king and queen. A colony splits into several castes, each defined as a group of individuals that is specified by certain morphology, function, and behaviour. The castes of termites are as follows (according to Šobotník & Dahlsjö 2017):

Queen and King - Wingless female and male who mate, are the founders of the colony, and often the only fertile individuals.

Alate - The adult winged stage.

Larva - A young, nutritionally dependent individual that does not reveal wing pads or soldier traits.

Neotenic - A secondary sexual individual with juvenile characteristics, this individual originates from larvae, nymphs, pseudergates, or workers through a single or more molts.

Worker - A wingless stage arising by irreversible deviation from development into an imago, ensuring the functioning of the colony.

Pseudergate (false worker) - An individual functioning as a worker in the colony, but unlike a true worker, maintaining possibility of developing wings into adult stage.

Soldier - An individual defending the colony with a heavily sclerotized head and defensive adaptations such as large and strong mandibles or defensive secretions of exocrine glands.

Presoldier - A non-sclerotized individual with soldier traits, a transitional stage between a larva, pseudergate, or worker and a soldier.

Inter-caste - An individual having traits of two or more castes. Usually, it is a developmental abnormality.

Fertile Soldier - A soldier-like individual that retains reproductive capacity as a male or female.

Termite colonies are defended by soldiers, and food acquisition and maintenance performed by true workers or pseudergates. Workers search for food, feed dependent castes such as soldiers, larvae, and the reproductive pair (sometimes multiple reproductives), and build and repair the nest. Due to these aspects, a termite colony is often referred to as a 'superorganism' (Wilson ,1971). Overall, the colony is usually protected by a complex structure of tunnels (galleries) or nests, which keep the population safe and also ensure control of the microclimate (Bignell et al., 2010). Termites are an important food source for many predators (Redford & Dorea, 1984) and must also compete for food sources such as wood, leaf litter, or humus with other taxa (Deligne et al., 1981; Šobotník et al. 2010).

Termites have developed a specific defence system, best represented by the soldier caste present in the colonies of all termites, except for a few derived groups where this caste has secondarily disappeared (Noirot & Pasteels, 1987; Šobotník et al., 2015). Termites protect themselves with both active and passive defence mechanisms. The active defence primarily includes the strong mandibles of soldiers, and some use a phragmotic method of defence, which involves the soldier defending the attacked nest by sealing the access tunnel with its body and its strongly sclerotized and specifically shaped head facing the predators. Subsequently, the tunnel towards the nest is sealed by the workers with a mixture of wood

fragments, faeces, and secretions of labial glands (Deligne et al., 1981; Šobotník et al., 2010). Additionally, soldiers use defensive substances produced by frontal, labral, or labial glands to protect the nest (Prestwich, 1984; Šobotník et al. 2010b; Palma-Onetto et al. 2019). The passive way of protection is a cryptic, hidden way of life and construction of well-fortified nests (Korb, 2011; Šobotník et al., 2010).

1.4 Nesting strategies

Termites exhibit diverse nesting strategies that can be broadly categorized into single-site and multiple-site nesting. Single-site nesters confine their colonies to a single, centralized nest, where all colony members reside. These nests are typically well-fortified and serve as hubs for brood care, resource processing, and defence. Single-site nesting is commonly observed in species that rely on localized resources, such as wood-feeding termites for example family Archotermopsidae, which consume the material in which they nest (Shellman-Reeve, 1997). In contrast, multiple-site nesters distribute their colonies across several interconnected nesting sites. These nests are often linked via subterranean tunnels or above-ground foraging trails, enabling the colony to exploit resources over larger areas. This strategy allows for increased flexibility and resilience to environmental pressures, as multiple nests can reduce the risk of colony failure due to localized disturbances (Thorne & Traniello, 2003). Species such as *Nasutitermes* and *Reticulitermes* exemplify this nesting strategy, using satellite nests to support foraging activities far from the primary nest (Jones & Eggleton, 2011).

1.5 Feeding Strategies and Diet

The main component of termite diet is lignocellulose, a complex of cellulose, lignin, and hemicelluloses, and its digestion requires a broad array of enzymes produced by termites and symbiotic microorganisms in termites gut (Ohkuma & Brune 2011). Termites primarily feed on dead wood or grass, but consume also rotten wood, leaf litter, humus, or soil, and only occasionally live tissues, whether grass (Hodotermitidae), wood (Stylotermitidae), symbiotic fungi (Macrotermitinae) or microepiphytes (part of Nasutitermitinae) (Bignell et al., 2011).

1.6 Termite gut structure and endogenous enzymes

Termites have developed a highly specialized digestive system that allow them to break down lignocellulose efficiently. The termite gut is divided into three main regions: the foregut,

midgut, and hindgut. Each region plays a distinct role in digestion, with significant differences between “lower” and “higher” termites (Brune, 2014).

“Lower” termites, (e.g. Mastotermitidae, Kalotermitidae, Rhinotermitidae) harbour a complex community of flagellate protozoa in their hindguts. These protozoa possess their own cellulolytic enzymes, contributing significantly to the breakdown of lignocellulose. The combined action of termite-produced enzymes and protozoan enzymes ensures efficient digestion of cellulose and hemicellulose (Nakashima et al., 2002). “Higher” termites (family Termitidae) have lost the flagellate protozoa and rely entirely on their own enzymes and bacterial symbionts for lignocellulose digestion. The bacterial community in the hindgut of “higher” termites has adapted to take over the cellulolytic functions previously performed by protozoa, producing a range of enzymes that break down plant material (Brune, 2014).

1.6.1 Foregut and Midgut:

The foregut, comprising the pharynx, esophagus and crop, is responsible for the initial ingestion and temporary storage of food. Following this, the midgut serves as the primary site for enzymatic digestion, where endogenous enzymes such as cellulases and hemicellulases, produced by the termite, initiate the breakdown of lignocellulosic plant material (Tokuda et al., 1998).

1.6.2 Hindgut

The hindgut is divided into the paunch, colon, and rectum, and is significantly enlarged compared to the foregut and midgut. The hindgut of termites is a highly specialized and compartmentalized structure, functioning as a series of interconnected microbial bioreactors that facilitate the efficient digestion of lignocellulosic material (Köhler et al., 2012). This structure is divided into distinct compartments (P1 to P5), each characterized by unique physicochemical conditions, such as pH gradients, oxygen pressure, and metabolite concentrations, which support specific microbial communities adapted to localized environments. The P1 compartment, with its extremely alkaline conditions (pH 9.3–10.9), hosts microbes such as *Turicibacter* and Lactobacillales, initiating the breakdown of ingested material. The enteric valve (P2) regulates the flow of digesta into the hindgut but is less colonized due to its functional role (Köhler et al., 2012). The P3 compartment, or hindgut paunch, represents the primary site for anaerobic digestion, with anoxic conditions and

hydrogen accumulation (up to 12 kPa), fostering diverse bacterial communities, including *Spirochaetes*, *Fibrobacteres*, and TG3, which play central roles in lignocellulose fermentation. Transitioning through the tubular P4 compartment, where taxa like *Acidobacteriaceae* dominate, the digesta reaches the P5 or rectum, characterized by slightly acidic conditions and microbial activity focused on processing remaining fermentation products such as lactate. Together, these compartments create an optimized system for extracting energy and nutrients from plant material, reflecting the evolutionary adaptations of termites and their symbiotic microbiota. This functional integration not only supports termite survival but also highlights their ecological role in nutrient cycling and decomposition (Köhler et al., 2012).

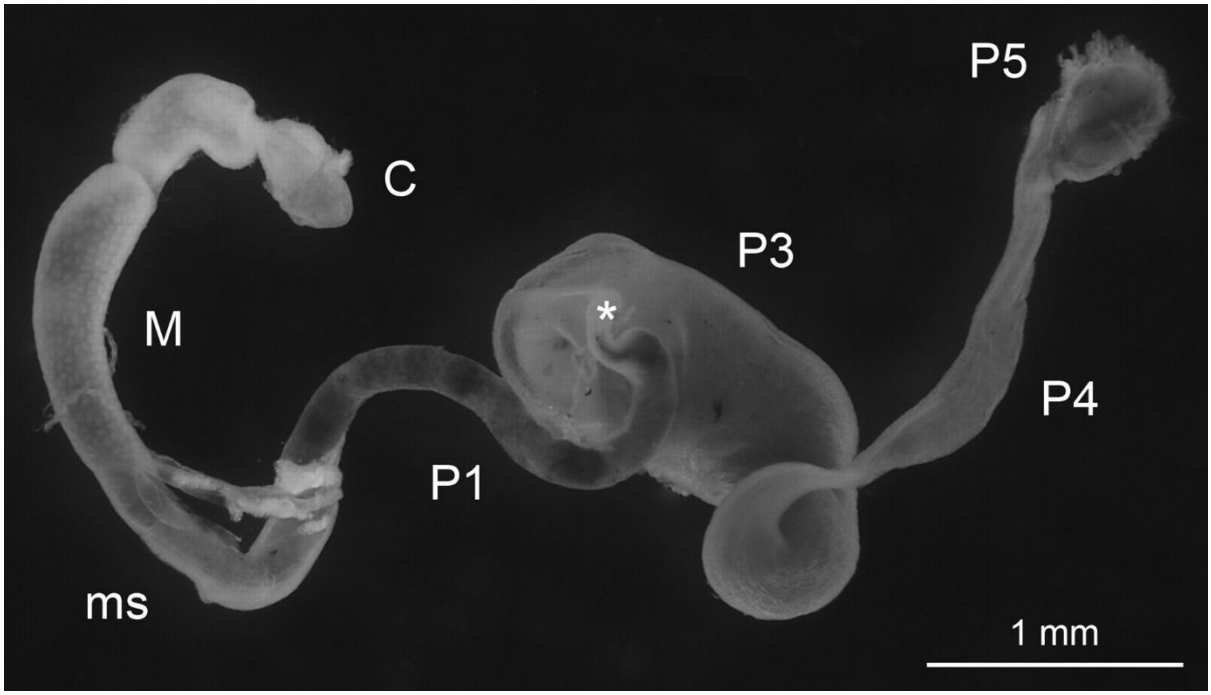


Figure 8: Intestinal tract of Nasutitermes corniger. The gut includes the crop (C), midgut (M), mixed segment (ms), and several hindgut segments (P1 to P5); the asterisk marks the position of the P2 (enteric valve) (Köhler et al., 2012)

1.6.3 Endogenous Enzymes

Termites produce their own cellulases, which are crucial for the initial stages of cellulose digestion. These enzymes include endoglucanases, cellobiohydrolases and β-glucosidases. Endoglucanases break down the internal bonds of cellulose chains to form shorter polysaccharide fragments. Cellobiohydrolases act on the ends of cellulose chains and release cellobiose units. β-Glucosidases hydrolyse cellobiose to glucose, which can be absorbed and utilized by the termite (Tokuda et al., 1998; Tokuda & Watanabe, 2007). In addition, termites produce hemicellulases, such as xylanases, which break down hemicellulose into xylose and

other sugars. These enzymes facilitate the degradation of more complex plant cell wall components and thus complement the action of cellulases (Nakashima et al., 2002).

1.7 Symbiotic organism in termite gut

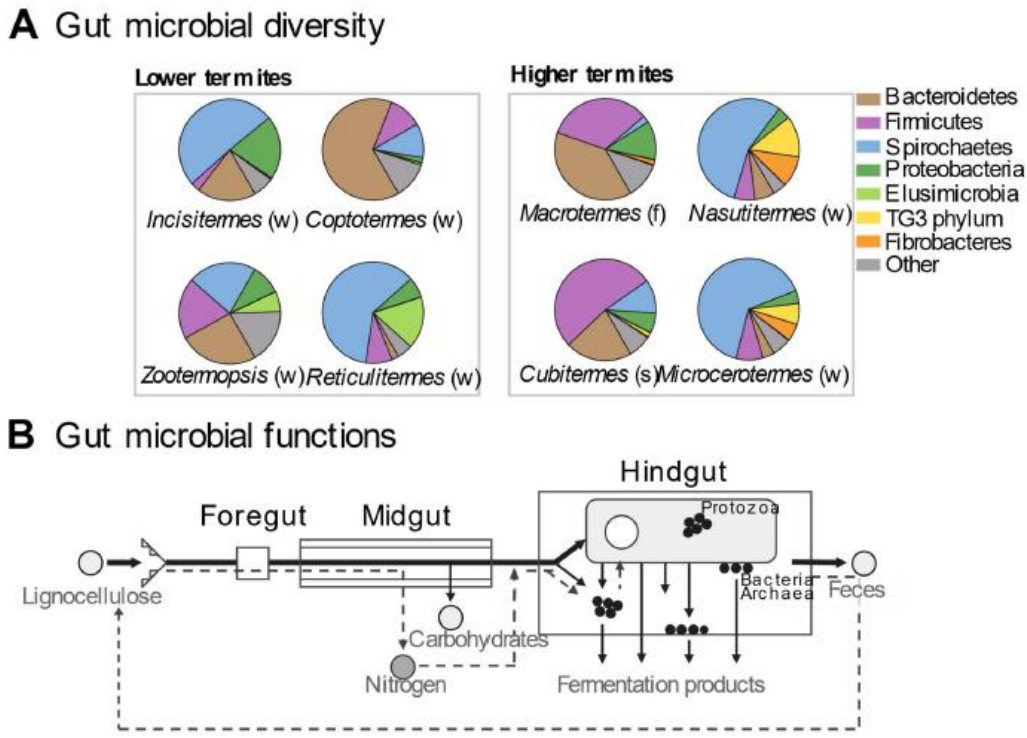


Figure 9: Termite gut microbiota composition and functions

(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).

Termite ability to digest lignocelluloses mostly rely on symbiotic organisms as bacteria, archaea, and protozoa. “Lower” termites typically harbour a rich community of protists in their hindguts. These protists (protozoa) are flagellates engaging in symbiotic relationships with the termite host. They are capable of cellulose degradation, breaking down the complex cellulose molecules into simpler sugars. The protists devour wood particles ingested by the termite, secreting cellulolytic enzymes that catalyse the breakdown of cellulose into glucose. This mutualistic interaction is crucial for the survival of “lower” termites, as the glucose

produced by the protists provides a vital energy source. The intricate balance within this symbiotic relationship is maintained through the process of proctodeal trophallaxis, whereby termites exchange hindgut contents, ensuring the transfer of protists to the next generation (Cleveland, 1925; Inoue, Moriya & Ohkuma, 2000). The “higher” termites do not rely on protists for lignocellulose digestion. Instead, their guts are populated by a diverse consortium of bacteria and archaea that produce cellulases and other glycoside hydrolases. The gut microbiome has been adapted to function efficiently in highly alkaline conditions, favouring the breakdown of cellulose and hemicellulose into fermentable sugars. The bacterial communities are adept at fermenting the sugars into acetate, which serves as the primary energy source for the termite. The absence of protists in “higher” termites also points to an advanced and more versatile digestive system, as these termites can consume a wider variety of lignocellulosic materials, including soil organic matter and humus, in addition to wood (Brune & Ohkuma, 2011; Warnecke et al., 2007).

1.7.1 Taxonomy of Bacteria in Termite gut

1.7.1.1 Firmicutes

Firmicutes of the genera *Clostridium*, *Ruminococcus* and *Desulfovibrio*, including the classification of *Desulfovibrio* under Proteobacteria, represent abundant taxa in microbial communities. *Clostridium* and *Ruminococcus* are known for their role in fermentation processes and gut health and contribute to the breakdown of complex carbohydrates in the digestive systems of animals, including termites. *Desulfovibrio*, although classified as Proteobacteria, shares functional features with Firmicutes in terms of energy metabolism, particularly in sulfate reduction processes. These genera are essential for understanding microbial ecology, energy cycling and symbiotic relationships in termite gut and other environments (Taib et al., 2020).

1.7.1.2 Bacteroidetes

The genera *Bacteroides* and *Prevotella* of the phylum Bacteroidetes are important for their role in the human gut microbiome, where they affect digestion and overall health. *Prevotella* is often associated with the high-fibre diets that are prevalent in rural communities, while Bacteroides is associated with Western diets rich in protein and fat. These genera are central

to discussions about the diversity of the gut microbiome, which reflects the host's diet, lifestyle and environmental factors. Their abundance and interactions within the gut ecosystem offer insights into the complex relationship between diet, microbiome composition and health outcomes (Gorvitovskaia et al., 2016).

1.7.1.3 Spirochaetes

The gut microbiome of termites, particularly rich in bacteria of the genus *Treponema* (phylum Spirochaetes), is crucial for the digestion of lignocellulosic material such as wood. Species of *Treponema* play an essential role in the termite gut ecosystem, as they are involved in fundamental processes including fibre hydrolysis, fermentation, homoacetogenesis, and nitrogen fixation. These processes enable termites to efficiently degrade and utilize wood as their primary food source, highlighting the symbiotic relationship between termites and their gut microbiota. The presence of *Treponema* in various termite species indicates a co-evolutionary history that has significantly contributed to the ecological success of termites (Abdul Rahman et al., 2015).

1.7.1.4 Proteobacteria

Species of the genus *Desulfovibrio* in the termite gut play a key role in nitrogen fixation and oxygen removal, facilitating the anaerobic environment necessary for the termite digestive process. This involvement underscores the complex symbiotic relationships in the termite gut microbiome. It highlights the importance of *Desulfovibrio* in maintaining the ecological balance required for termite survival and wood decomposition (Abdul Rahman et al., 2015).

1.7.1.5 Actinobacteria

Actinobacteria in termites are essential for breaking down the tough components of wood and plants, such as lignin, making other nutrients more accessible. These bacteria produce substances that help control harmful microbes in the termite gut and contribute to the balance of the microbial community. Their relationship with termites is mutually beneficial: termites provide a home and food, while actinobacteria aid the digestion. The diversity of actinobacteria varies depending on the termite species and the environment, indicating their

adaptability. Some of them can even convert nitrogen from the air into a form that termites can use, helping in their nutrition (Korsa et al., 2023).

1.7.2 Taxonomy of Archaea in Termite Guts

The most abundant archaea in termite gut are *Methanobrevibacter* species, which are critical for methane production through methanogenesis. This process is necessary for the digestion of lignocellulosic material, allowing termites to use wood as a primary food source. In "lower" wood-feeding termites such as *Reticulitermes spp*. and *H. sjostedti*, Methanobrevibacter species are the only methanogens. However, in "higher" termites, including *Microcerotermes sp.* and *Nasutitermes sp., Methanobrevibacter* coexists with methanogens from other orders such as Methanoplasmatales and Methanomicrobiales, indicating a more diverse community of archaea. Methanoplasmatales are particularly dominant in the gut of "higher" termites, where they represent 37.5-60.3% of the archaeal population. This group, previously classified as Thermoplasmatales, has been found in various termite species and is capable of producing methane from methanol, suggesting a new lineage of methanogens (Parks et al., 2020a; Shi et al., 2015). In addition, the presence of Thaumarchaeota in termite gut, which is involved in ammonia metabolism, points to a complex ecological function beyond methanogenesis. These archaea, distinct from archaea in other environments, suggest a unique evolutionary origin and specialized role in the termite gut ecosystem. The diversity of archaeal communities in termites is influenced by the diversity of methanogenic substrates provided by the complex bacterial communities of the gut. "Higher" termites with more complex gut compartments and physiological conditions exhibit greater archaeal diversity. This diversity is further enriched by different metabolites, such as formate and methanol, which are utilized by different methanogens, indicating a subtle interplay of metabolic processes in the termite gut. Despite the dominance of methanogens, the role of non-methanogenic archaea, especially in "higher" termites presents interesting questions for further research on their ecological functions and contribution to termite digestion and methane production (Shi et al., 2015).

1.7.3 Co-evolution of gut bacteria and termite

The diversity of termite species and their diet is closely linked to the evolution of their gut microbiota, highlighting the complex web of interactions that are essential for termite survival and ecosystem functioning (Engel and Moran, 2013).

The evolutionary history of termites, spanning over 150 million years, is tightly interwoven with the development of their gut microbial communities. Phylogenetic studies reveal strong cophylogenetic signals between termite lineages and their microbiota, demonstrating a long-term association. For example, bacterial taxa such as *Treponema* within the phylum Spirochaetes form monophyletic clusters specific to termites, while lineages within Firmicutes and Bacteroidetes also show termite-specific evolutionary patterns (Arora 2023, Bourguignon et al. 2018, Mikaelyan et al., 2015). These findings underscore vertical transmission as a critical mechanism for preserving these symbiotic relationships across generations (Engel and Moran 2013; Vavre and Kremer 2014; Groussin et al. 2020).

Proctodeal trophallaxis, the exchange of hindgut fluids among colony members, plays a pivotal role in this vertical inheritance, ensuring the continuity of coevolved microbial communities over evolutionary timescales (Nalepa 2017; Michaud et al. 2020). Furthermore, termite-specific microbial communities adapt to changes in host diets and phylogeny. For instance, wood-feeding termites ("Lower" termites and non-Macrotermitinae Termitidae subfamilies), harbour microbial communities enriched with Fibrobacteraceae, a bacterial family specialized in lignocellulose degradation, while soil-feeding termites (family Termitidae, particularly subfamilies like Cubitermitinae, Syntermitinae, Nasutitermitinae, and Apicotermitinae, among others) have gut microbiota dominated by nitrogen-metabolizing taxa such as Firmicutes (Mikaelyan et al. 2014; Tokuda et al. 2018). The coevolutionary relationship between termites and their gut microbiota is not static but highly dynamic. Vertical inheritance preserves essential microbial lineages, while horizontal gene transfer and occasional environmental acquisitions enrich the microbiota's genetic and functional diversity (Nalepa, 2017; Bourguignon et al., 2018). Recent studies further support vertical inheritance, demonstrating that some microbial lineages were conserved from a common ancestor of termites, while others were acquired through dietary and ecological transitions (Arora, 2023; Brune & Ohkuma, 2011).

Functional studies provide additional evidence of this coevolution also for nitrogen-fixing microbes, such as those from Bacteroidota, *Treponematales*, and *Methanobrevibacter*, compensate for termites' nitrogen-deficient diets, enhancing their survival and ecological fitness (*more in chapter 1.7.1. Nitrogen fixation*) (Yamada et al., 2007; Ohkuma et al., 1999). Horizontal gene transfer among gut microbes has further expanded the functional repertoire of these microbial communities, enabling them to produce a diverse array of carbohydrate-active enzymes (CAZymes) essential for lignocellulose degradation (*more in chapter 1.1. CAZymes*)(Tokuda et al., 2018).

1.8 Cazyme

The termite gut microbiome is characterized by a diverse array of polysaccharide-degrading enzymes, including xylanases for Xylan degradation and various pathways for the metabolism of cellobiose, a common intermediate in cellulose degradation. These metabolic pathways are essential to minimize inhibition of cellulases, thereby allowing more efficient wood decomposition. This complex microbial ecosystem in termite gut with Treponema as a key player is an example of evolutionary adaptation to a lignocellulosic diet and highlights the importance of the termite gut as a natural biomass conversion system (Liu et al., 2019).

Carbohydrate-active enzymes (CAZymes) are a broad class of enzymes that are critical to the digestion, modification, and synthesis of carbohydrates. CAZymes are essential for various processes, including the breakdown of complex sugars into simple sugars for digestion, constructing and remodelling cell walls, and storing and utilizing energy (Wardman et al., 2022).

CAZymes are categorized into different families and subfamilies, each characterized by the specific reactions they catalyse and the structures of the substrates they act upon. Glycoside hydrolases (GHs) with 189 families and 241 subfamilies, Glycosyltransferases (GTs) with 135 families. Polysaccharide lyases with 43 families and 63 subfamilies, Carbohydrate-esterase with 20 families, Carbohydrate-binding module with 101 families and Auxiliary Activities with 17 families (Cantarel et al., 2009; Wardman et al., 2022). These enzymes are not confined to a single group of organisms but are widespread across the kingdoms of life. They are found in microorganisms like bacteria and fungi, which play an essential role in the decomposition of organic matter and nutrient cycling in ecosystems. Plants also produce CAZymes, which are

involved in the development of plant structures (Pinard et al., 2015). Although these enzymes are widespread across various microbial species, research has shown that certain bacterial genera exhibit particularly high abundances of CAZyme-encoding genes. Bacteroides produce more likely glycoside hydrolases (GH2, GH3, GH5, GH43) and carbohydrate esterases (CE1, CE6) that break down dietary fibre (Koropatkin et al., 2012). Species of the genus *Clostridium* (for example in ruminant gut) use GH9 and GH48 to degrade cellulose and PL1 and PL9 to degrade pectin (Flint et al., 2008). Soil-dwelling Streptomyces species produce CAzymes such as GH12, GH74 (lichenases and xyloglucanases) and CE2, CE4 (deacetylases) to degrade plant biomass (Book et al., 2014). Bacillus species contribute to the decomposition of soil organic matter through GH13 and GH32 (amylases and inulinases) and GT2 and GT4 (glycosyltransferases) (Lombard et al., 2014). In the gut of herbivores, Ruminococcus species play an important role in cellulose degradation through GH5 and GH26 (mannanases and endoglucanases) and CBM32, which enhances enzyme binding to polysaccharides (Crouch et al, 2016). “Lower” termites have high levels of glycoside hydrolases (GH1, GH9) and carbohydrate esterases (CE1) for cellulose and hemicellulose degradation. “Higher” termites rely on bacterial symbionts in the hindgut to produce CAZymes, including glycoside hydrolases (GH5, GH11) and auxiliary activities (AA1, AA3) that aid in lignin degradation (Brune, 2014).

The study of CAZymes involves various methods to uncover their functions and mechanisms. Biochemical assays can reveal the activity of these enzymes, while gene expression analysis can show when and where these enzymes are produced. Genomic and metagenomic approaches have expanded our understanding of the diversity and evolution of CAZymes, shedding light on their roles in environmental ecosystems and potential applications in industries such as biofuel production. The Carbohydrate-Active CAZyme Database (CAZy) is an invaluable resource for researchers in the field, offering a comprehensive repository of information on the known CAZymes, including their structure, function, and classification (Cantarel et al., 2009). Studies like those of Henrissat and Davies (2013) provide structural and mechanistic insights, enhancing our comprehension of how CAZymes operate at a molecular level. Moreover, Pallen and Wren (2006) discuss the emergence and significance of CAZymes in prokaryotes, emphasizing their importance in microbial communities.

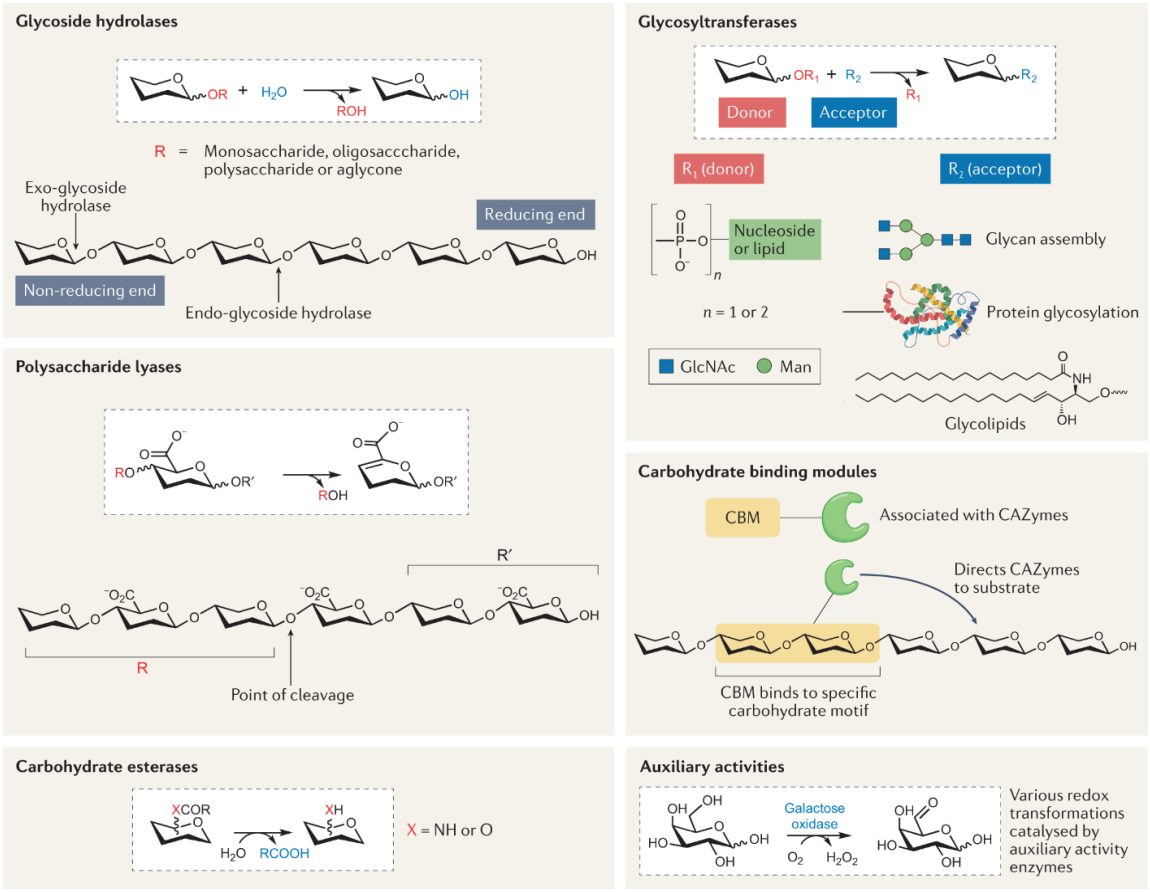


Figure 10: visualization of different CAZyme families and scheme of function (Wardman et al., 2022).

1.8.1 Glycoside hydrolases

Glycoside hydrolases (GHs) are enzymes that break down glycosidic bonds, and they are categorized into various families based on their structure, mechanism, and function. Each family of GHs has distinct characteristics and plays specific roles in carbohydrate metabolism (Lombard et al., 2014a).

For instance, enzymes in the GH3 family have been studied for their ability to hydrolyse and sometimes transglycosylate substrates, which means they can transfer glycosyl groups to other sugars or molecules. This family includes enzymes like β-glucosidases and N-acetylglucosaminidases. The catalytic mechanism of these enzymes often involves a conserved aspartate residue acting as a nucleophile, which is critical for catalysis. Notably, the positioning of the general acid/base residue, which is essential for the enzyme's function, can vary across the family, influencing the enzyme's activity and specificity(Lombard et al., 2010a, 2014b).

GH18, GH20, and GH85 are examples of families that utilize a mechanism where an acetamido group at the 2-position of the substrate assists in the cleavage of the glycosidic bond. This neighboring group participation leads to the formation of intermediates like oxazoline or oxazolinium ions (Davies and Henrissat, 1995). The GH99 family has enzymes that hydrolyze α-mannoside substrates without a typical catalytic nucleophile. Instead, these enzymes use the 2-hydroxyl group as an intramolecular nucleophile, leading to the formation of an epoxide intermediate. Myrosinases from the GH1 family are unique because they lack a general acid and utilize an external base, like ascorbate, to assist in catalysis. These enzymes hydrolyse thio glycosides found in plants (Henrissat and Davies, 1997).

Some retaining glycosidases, such as those in the GH33 and GH34 families, employ alternative nucleophiles like tyrosine instead of the conventional carboxylate residues. The use of tyrosine as a nucleophile is suggested to be advantageous when the anomeric centre of the substrate is negatively charged, as it avoids charge repulsion. Lastly, GH4, GH109, GH177, and GH179 families use a different mechanism involving NAD cofactors. They proceed via an elimination and redox mechanism rather than through the classical double-displacement reaction typically seen in other GH families (Davies and Henrissat, 1995; Lombard et al., 2014c).

1.8.2 Glycosyltransferases

Glycosyltransferases (GTs) are key enzymes in glycosylation processes, transferring sugar moieties from activated sugar donors to acceptor molecules. They are essential for various biological functions, including cell wall synthesis, signalling, and immune response modulation. The classification of GTs into families is based on amino acid sequence similarities, reflecting their structural and mechanistic features. This system highlights the diversity within GT families, grouping enzymes with different substrate specificities and catalytic mechanisms (Hartman et al., 2007). Modular GTs, often containing non-catalytic domains, play significant roles in substrate binding and enzyme localization, indicating a complex evolution to acquire new functions. The continuous update of the CAZy database reflects the growing understanding of GTs' roles in biology and their potential in biotechnological applications (Lairson et al., 2008).

1.8.3 Polysacchaaride Lyases

Polysaccharide lyases catalyse the non-hydrolytic cleavage of glycosidic bonds in polysaccharides, leading to the formation of a new double bond at the reducing end of the cleaved molecule. These enzymes are essential in the degradation of complex polysaccharides like pectins, heparin, and hyaluronan, which are found in the cell walls of plants and in extracellular matrices of animals. Their action is crucial in processes such as plant biomass degradation, pathogenesis, and various biotechnological applications (Lombard et al., 2010b).

1.8.4 Carbohydrate Esterases

Carbohydrate esterases are enzymes that de-esterify the acetyl or other ester groups from carbohydrate molecules. These enzymes are pivotal in the modification and degradation of plant biomass, especially in the breakdown of hemicelluloses and pectins, which often contain acetyl, methyl, and other ester-linked decorations. This de-esterification is often a prerequisite for further degradation by other carbohydrate-active enzymes, making CEs essential in biomass conversion, paper manufacturing, and the food industry (Armendáriz-Ruiz et al., 2018; Lombard et al., 2014a).

1.8.5 Auxiliary Activities

The Auxiliary Activities family encompasses a diverse group of enzymes that support the breakdown of complex carbohydrates by acting in conjunction with other CAZymes. They include various lytic polysaccharide monooxygenases (LPMOs), which catalyse the oxidative cleavage of polysaccharides, and other redox enzymes that target lignin, chitin, and cellulose. These enzymes are critical for lignocellulose deconstruction in biofuel production, enhancing the efficiency of other carbohydrate-active enzymes by providing access to otherwise resistant structures (Levasseur et al., 2013).

1.8.6 Carbohydrate-Binding Modules

Carbohydrate-binding modules are non-catalytic protein domains that bind to carbohydrates. They play a crucial role in targeting and increasing the efficiency of catalytic domains towards their substrate by binding to specific polysaccharides. CBMs are found in various architectures, from single domains to complex multimodular structures, and are involved in

numerous biological processes, including cellulose degradation, pathogenesis, and symbiosis. Their specificity and binding properties make them valuable tools in biotechnology and biofuel research (Armenta et al., 2017).

These enzymatic families are essential for the complete utilization and modification of carbohydrates, with applications in numerous fields such as bioenergy, bioremediation, food processing, and pharmaceuticals. Their study and manipulation continue to be an active area of research, driving innovations in sustainable technologies and bioproducts (Boraston et al., 2004; McLean et al., 2002).

1.9 Process of digestion in termites

The digestion process of lignocellulose begins with the mastication of wood, initiating the mechanical breakdown of lignocellulosic material. Termites use their mandibles to chew wood into smaller particles (primarily “lower” termites) or soil (Termitidae), which increases the surface area for enzymatic action. The chewed material is then mixed with saliva, which contains a suite of enzymes that begin the initial hydrolysis of cellulose. The food then passes into the midgut, which is the primary site for enzymatic digestion by endogenous enzymes produced by the termite itself. These include endoglucanases (GH9) that break internal bonds within cellulose chains, producing shorter polysaccharide fragments, cellobiohydrolases (GH1, GH6) that cleave cellobiose units from the ends of cellulose chains, β-glucosidases (GH1) that hydrolyse cellobiose and other oligosaccharides into glucose, xylanases (GH10, GH11) that break down hemicellulose, particularly Xylan, into xylose and other monosaccharides, and mannanases (GH5) that degrade hemicellulose components such as mannan into mannose. These enzymes facilitate the initial stages of lignocellulose degradation, making it more accessible for further breakdown by microbial symbionts in the hindgut (Tokuda et al., 1998; Tokuda & Watanabe, 2007).

Most of the cellulose digestion occurs within the hindgut, which acts as an anaerobic fermentation chamber. The hindgut environment is neutral to slightly alkaline, which is conducive to the activity of microbial enzymes. The primary difference between a hindgut of “lower” and “higher” termites lies in their symbiotic relationships and the composition of their gut microbiota. “Lower” termites depend on a combination of protozoa, which harbour

their own exo- and endosymbiotic bacteria. Protozoa devour wood particles, produce digesting enzymes and convert the particles into simpler sugars and acetate, which the termite can absorb and utilize (Cleveland, 1923). “Higher” termites rely solely on bacterial symbionts, which evolved a more diverse and specialized bacterial community capable of producing a wide range of CAZymes to compensate for the absence of protozoa.

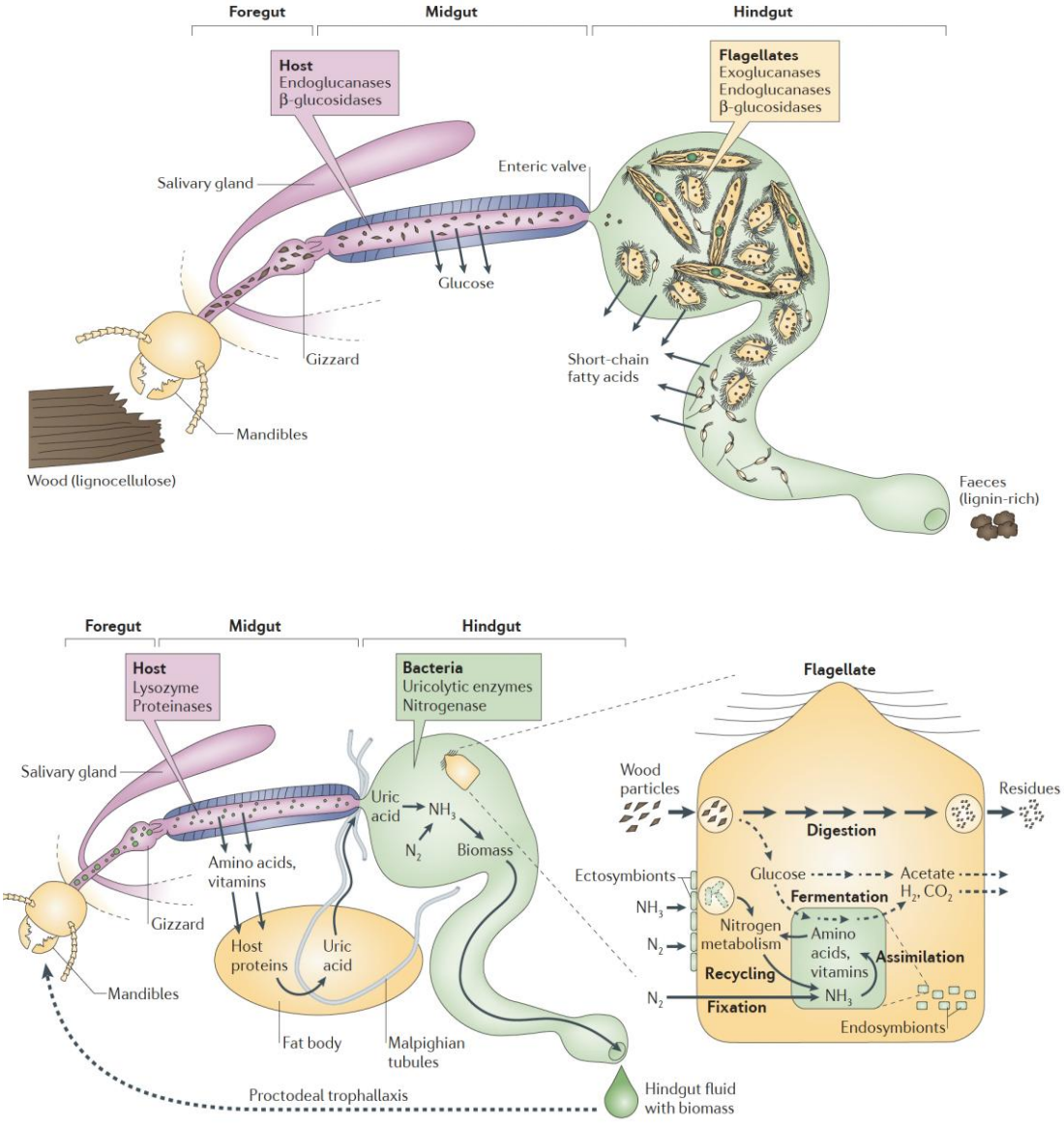


Figure 11: Schema of termite gut and digestion process (Brune, 2014)

1.9.1 Nitrogen fixation

Another critical aspect of termite digestion is nitrogen fixation. Many termite species consume a diet rich in cellulose but poor in nitrogen. To compensate, diazotrophic microorganisms in the termite gut fix atmospheric nitrogen (N_2) into biologically usable forms,

enriching the termite's nitrogen intake essential for synthesizing amino acids and nucleotides (Breznak, 1982).

Termite gut microbiota 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). Gene homologs of these nitronenases (nifDNK) is significantly corelating with termite phylogeny. Significant differences were observed among termite groups, with nitrogenase reads in the gut metagenomes of non-fungus-cultivating (non-FC) wood-feeders (LT and WF) being 24.4 times more abundant compared to soil-feeders (SF) and 20.2 times more abundant than in fungus-cultivating (FC) termites (Arora 2023). This aligns with the higher rates of N_2 fixation reported in LT and WF compared to SF and FC (Yamada et al., 2007) and highlights the high nitrogen content in soil and fungi, which reduces the necessity for the energy-intensive process of N_2 fixation (Brune & Ohkuma, 2011; Hongoh, 2011).

Genetic analysis confirmed the presence of genes responsible for nitrogen fixation in the gut microbiomes of termites. Specifically, members of Bacteroidota, Spirochaetota (order Treponematales), Proteobacteria (family Enterobacteriaceae), and the archaeal genus Methanobrevibacter. Additional analyses of metagenome-assembled genomes (MAGs) further revealed nitrogenase genes in lineages such as Actinobacteriota, Planctomycetota, Verrucomicrobiota, and Firmicutes (Arora, 2023). These findings corroborate previous evidence that termites harbor a diverse community of nitrogen-fixing microbes in their guts (Ohkuma et al., 2001; Yamada et al., 2007; Desai & Brune, 2012). Subsequently phylogenetic position of termite species determined, in some measure, the taxonomy of their dominant diazotrophs (Arora 2023).

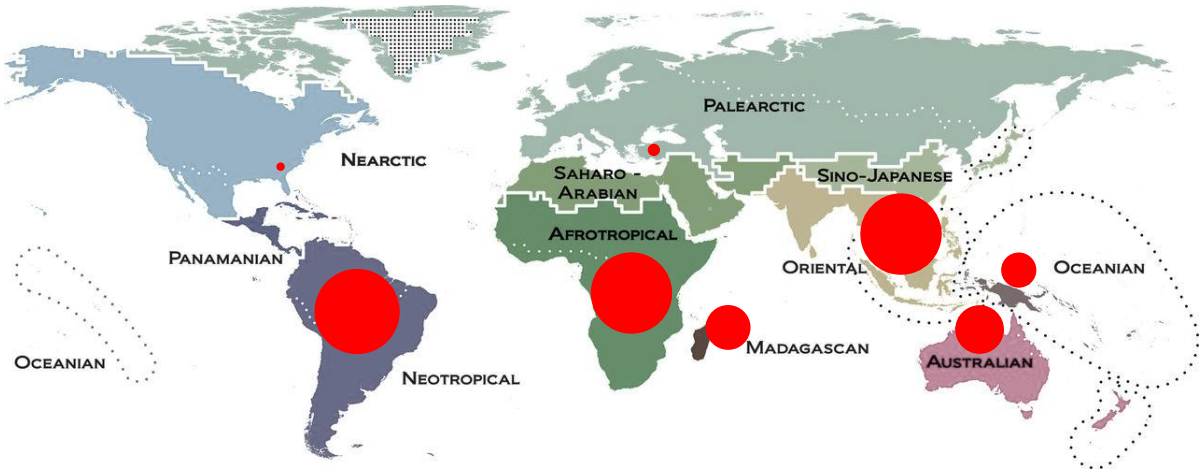
In addition to nitrogen fixation, termites enhance soil nitrogen content through organic matter decomposition and nutrient recycling. As termites consume plant material, the metabolic activity of their gut microbiota releases nitrogen compounds into the soil, making them available for plant uptake. Termite mounds and surrounding soil often have elevated nitrogen levels compared to areas without termite activity, emphasizing their role in nutrient cycling (Jones et al., 2005; Holt & Lepage, 2000). The mineralization of organic matter processed by termites further enhances soil fertility, converting nitrogen into forms accessible to plants and contributing to ecosystem productivity (Schmidt et al., 2019).

2. Methodology

2.1 Sample collection and preparation

Microbial data were obtained from gut metagenome analysis from 195 termite samples and 145 termite species and one *Cryptocercus* with detailed information in (supplementary table 1).

Termite guts were dissected from at least 10 workers for each sample and preserved in RNA-later® and stored at – 80 °C until DNA extraction.



(Holt et al., 2013)

Afro-tropical	**41**
Australia	**24**
Madagascar	**22**
Neo-arctic	**3**
Neo-tropical	**43**
Oceania	**17**
Oriental	**41**
Paleo-arctic	**5**

2.2 Genomic DNA extraction

Genomic DNA extraction was performed on the whole guts of five workers using the NucleoSpin Soil kit (Macherey-Nagel) according to the protocol. For the analysis of metagenome were use gut from workers:

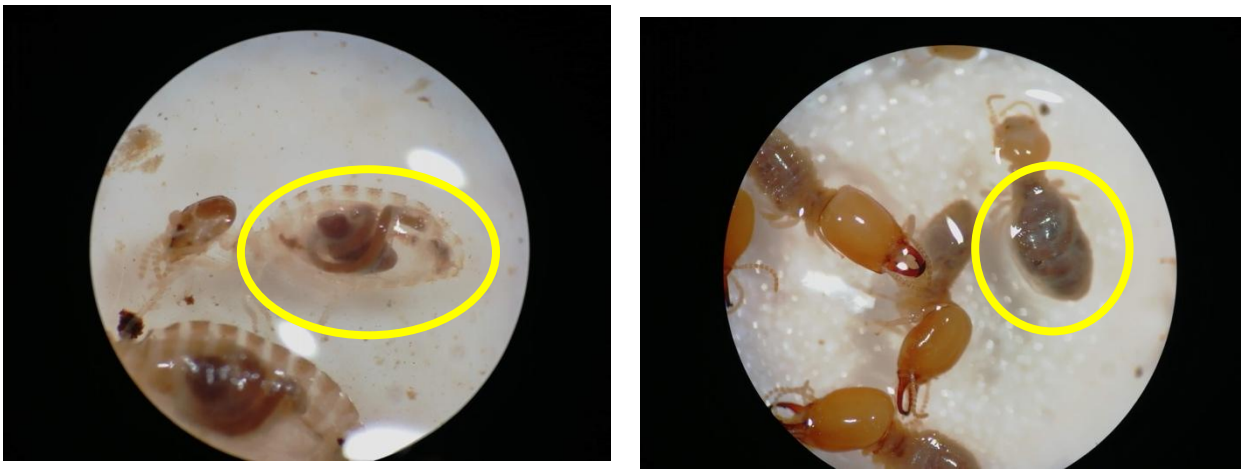


Figure 12: Termite workers under binocular

2.2.1 Protocol – purification of DNA from soil and sediment

Sample Preparation

Fresh sample material, ranging between 250 and 500 mg, is transferred into an MN Bead Tube Type A, which contains ceramic beads. It is essential to ensure that the tube is not filled beyond the 1 mL mark to maintain proper mixing and efficient lysis. Following this, 700 µL of Buffer SL1 or Buffer SL2 is added to the tube. Adjustments to the lysis buffer volume may be necessary depending on the nature of the sample. For very dry material, additional lysis buffer can be added until the tube is filled up to the 1.5 mL mark to ensure sufficient hydration and efficient lysis. Conversely, for very wet material, any excess liquid should be removed prior to the addition of the lysis buffer, which may involve spinning down the sample to separate excess moisture. These steps are critical for preparing the sample for downstream processes while maintaining the integrity of the extraction procedure.

Adjust Lysis Conditions

To optimize the lysis process, 150 µL of Enhancer SX is added to the sample, and the cap is securely closed. Enhancer SX is designed to maximize DNA yield, but it may also facilitate the release of humic acids. For samples with a high content of contaminants, adjustments to the

volume or complete omission of the enhancer can improve DNA purity. Refer to Section 2.5 for detailed recommendations on managing this balance.

Sample Lysis

The MN Bead Tubes are horizontally secured to a vortexer, either using tape or a specialized adapter. The samples are then vortexed at maximum speed at room temperature (18–25 °C) for 5 minutes, ensuring thorough disruption of the sample material.

Precipitate Contaminants

To reduce foam caused by the detergent, the samples are centrifuged for 2 minutes at 11,000 × g. At this stage, it is recommended to transfer the clear supernatant to a new collection tube, particularly for carbonate-rich samples, as this ensures consistency in yield. Next, 150 µL of Buffer SL3 is added, followed by vortexing for 5 seconds. The samples are incubated on ice (0–4 °C) for 5 minutes and centrifuged again for 1 minute at 11,000 × g to precipitate contaminants effectively.

Filter Lysate

A NucleoSpin® Inhibitor Removal Column (red ring) is placed into a 2 mL Collection Tube with a lid. Up to 700 µL of the clear supernatant from the previous step is loaded onto the column and centrifuged for 1 minute at 11,000 × g. For wet samples, such as sediments, where the clear supernatant volume exceeds 700 µL, the process is repeated with a fresh collection tube. Flow-throughs are combined after each spin. After filtration, the NucleoSpin® Inhibitor Removal Column is discarded, and any visible pellet in the flow-through is avoided by transferring the clear supernatant to a new collection tube.

Adjust Binding Conditions

To facilitate DNA binding, 250 µL of Buffer SB is added to the sample, and the tube is vortexed for 5 seconds. For samples preserved in Zymo DNA/RNA Shield, the total sample volume is quantified after adding Buffer SB, and 0.2 volumes of isopropanol are incorporated to enhance DNA recovery.

Bind DNA

The DNA-binding step is performed using a NucleoSpin® Soil Column (green ring) placed in a 2 mL Collection Tube. An initial volume of 550 µL of the prepared sample is loaded onto the column and centrifuged for 1 minute at 11,000 × g. The flow-through is discarded, and the column is reinserted into the collection tube. The remaining sample is then loaded, and the process is repeated to ensure all DNA binds to the column. After the final centrifugation step, the flow-through is discarded, and the column is retained for subsequent purification steps.

2.2.2 Library preparation using the KAPA HyperPlus Kit:

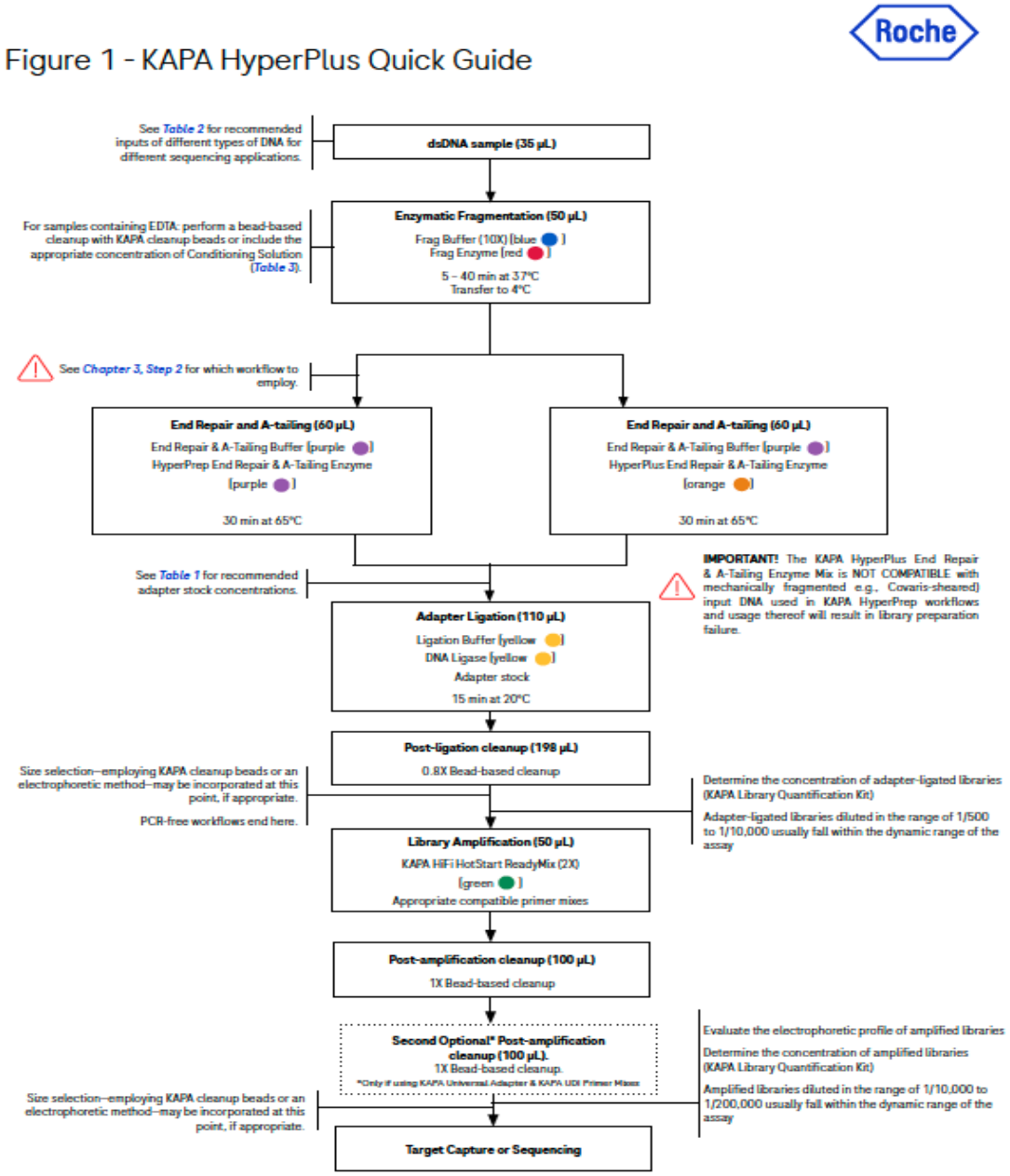


Figure 13: Quick guide of Library preparation protocol

Prepare the Sample Library

Step 1. Enzymatic Fragmentation

1. Dilute 1 ng – 1000 ng of DNA with 10 mM Tris-HCl, pH 8.0 – 8.5 (recommended) to a total volume of 35 µL in a 0.2 mL tube or well of a PCR plate.
    - *If the DNA preparation does not contain EDTA, dilute in 10 mM Tris-HCl (pH 8.0 – 8.5) in a total of 35 µL.*
    - *If the DNA preparation does contain EDTA, dilute in the EDTA-containing buffer in which samples are currently suspended, in a total of 30 µL. To each reaction with 30 µL of EDTA-containing DNA, add 5 µL of diluted Conditioning Solution.*
2. Assemble each Fragmentation reaction on ice as per the table below:

Component	Volume Per Individual Sample
1 ng – 1000 ng DNA (with Conditioning Solution, if needed)	35 µL
KAPA Frag Buffer (10X)*	5 µL
KAPA Frag Enzyme*	10 µL
Total volume	**50 µL**

Table 1: volumes for Fragmentation reaction

3. Mix the Fragmentation reaction thoroughly and centrifuge briefly. Return the plate/tube(s) on ice and proceed immediately to the next step.
    - If the Fragmentation reaction is not mixed properly, it can result in increased fragment size.
4. Incubate in a thermocycler, pre-cooled to +4°C and programmed as outlined below. Set the lid temperature to ~ +65°C (if possible):

    a. Pre-cool block: +4°C

    b. Fragmentation: +37°C - See table below

    c. Hold: +4°C

Mode fragment length	Incubation time at +37°C*	Optimization range
600 bp	5 min	3 – 10 min
350 bp	10 min	5 – 20 min
200 bp	20 min	10 – 25 min
150 bp	30 min	20 – 40 min

Table 2: Incubation in thermocycler

Step 2. End Repair and A-tailing

1. In the same plate/tube(s) in which enzymatic fragmentation was performed, assemble each End Repair and A-Tailing reaction as per table below:

Component	Volume Per Individual Sample
Fragmented, double-stranded DNA	50 µL
End Repair & A-Tailing Buffer*	7 µL
End Repair & A-Tailing Enzyme Mix**	3 µL
Total volume	**60 µL**

Table 3: End Repair and A-Tailing reaction mix

2. Mix the End repair and A-tailing reaction thoroughly and centrifuge briefly. Return the reaction plate/tube(s) to ice. Proceed immediately to the next step.
3. Incubate in a thermocycler programmed as outlined below. A heated lid is required for this step. If possible, set the temperature of the heated lid to ~ +85°C (instead of the usual +105°C).

Step	Temperature	Time
End repair and A-tailing	+65°C*	30 min
Hold	+4°C**	∞

Table 4: Incubation in thermocycler

Step 3. Adapter Ligation

1. Transfer the reaction from the thermocycler to ice. 2.
2. In the same plate/tube(s) in which End repair and A-tailing was performed, assemble each Adapter Ligation reaction on ice as per the table below:

Component	Volume Per Individual Sample
End repair and A-tailing reaction product	60 µL
KAPA Adapters (Chapter 2)	5 µL
PCR-grade water*	5 µL
Ligation Buffer*	30 µL
DNA Ligase*	10 µL
Total volume	**110 µL**

Table 5: Adapter Ligation reaction

3. Mix the Adapter Ligation reaction thoroughly and centrifuge.
4. Incubate the Adapter Ligation reaction at +20°C on a thermocycler for 15 minutes.

5. Following the incubation, proceed immediately to the next step.

Step 4. Purify the Sample Library using KAPA HyperPure Beads

1. To each Adapter Ligation reaction, add 88 µL of room temperature KAPA HyperPure Beads that have been thoroughly resuspended.

Component	Volume
Ligation reaction product	110 µL
KAPA HyperPure Beads	88 µL
Total volume	**198 µL**

Table 6: Adapter Ligation reaction for purifying of the samples

2. Once added, mix thoroughly and centrifuge briefly to collect all droplets. Do NOT allow beads to pellet.
3. Incubate the sample at room temperature for 5 minutes to allow the sample library to bind to the beads.
4. Place the sample on a magnet to capture the beads. Incubate until the liquid is clear.
5. Carefully remove and discard the supernatant.
6. Keeping the sample on the magnet, add 200 µL of freshly-prepared 80% ethanol.
7. Incubate the sample at room temperature for ≥30 seconds.
8. Carefully remove and discard the ethanol.
9. Keeping the sample on the magnet, add 200 µL of freshly-prepared 80% ethanol.
10. Incubate the sample at room temperature for ≥30 seconds.
11. Carefully remove and discard the ethanol. Remove residual ethanol without disturbing the beads.
12. Allow the beads to dry at room temperature, sufficiently for all the ethanol to evaporate.

- *Over-drying the beads may result in dramatic yield loss. Over-drying is indicated by a dry, cracked appearance of the bead pellet. The bead pellet should have a matte appearance when sufficiently dried.*

13. Remove the sample from the magnet.
14. Thoroughly resuspend the beads:

- 14.1 in 25 µL of elution buffer (10 mM Tris-HCl, pH 8.0 – 8.5) to proceed with Library Amplification (Chapter 4), or
- 14.2 in 55 µL of elution buffer (10 mM Tris-HCl, pH 8.0 – 8.5) to proceed with Double-sided Size Selection (Appendix B).

15. Incubate the sample at room temperature for 2 minutes to allow the sample library to elute off the beads.

16. 16. Place the sample on the magnet to collect the beads. Incubate until the liquid is clear.

17. 17. Transfer an appropriate volume of the clear supernatant/eluate to a fresh tube/well:

I. to proceed with Library Amplification (Chapter 4), transfer 20 µL of supernatant, or

II. to proceed with Double-sided Size Selection (Appendix B), transfer 50 µL of supernatant.

The remaining 5 µL can be used for quality control purposes e.g., quantification using the KAPA Library Quantification Kit.

18. Proceed to Chapter - Amplify The Sample Library (optional for sample inputs of ≥50 ng but mandatory if using KAPA Universal Adapter) or Chapter - Quality Control, if performing a PCR-free workflow (not applicable if using KAPA Universal Adapter).

Safe stopping point – If necessary, this is a safe stopping point. Purified, adapter-ligated library may be stored at +2°C to +8°C for 1 – 2 weeks or at -15°C to – 25°C for ≤1 month before amplification and/or sequencing. To avoid degradation, always store DNA in a buffered solution (10 mM Tris-HCl, pH 8.0 – 8.5) when possible, and minimize the number of freeze-thaw cycles.

Chapter 4. Amplify the Sample Library

Step 1. Prepare the Library Amplification Reaction

1. Assemble each Library Amplification reaction as per table below:

Component	Volume per Individual Library
KAPA HiFi HotStart ReadyMix (2X)	25 µL
KAPA Library Amplification Primer Mix* OR KAPA UDI Primer Mix**	5 µL
Adapter-ligated library	20 µL
Total volume	**50 µL**

Table 7: Library Amplification reaction

2. Mix thoroughly and centrifuge briefly. Immediately proceed to the next step.

Step 2. Perform the Library Amplification

1. Place the sample in the thermocycler and amplify the adapter-ligated library using the following Library Amplification program with the lid temperature set to +105°C:

Step	Temperature	Duration	Cycles
Initial denaturation	+98°C	45 sec	1
Denaturation	+98°C	15 sec	Variable, see Table 4 or Table 5 below for cycle numbers tailored to the KAPA Adapter that was used during Adapter Ligation
Annealing	+60°C	30 sec	
Extension	+72°C	30 sec	
Final extension	+72°C	1 min	1
Hold	+4°C	∞	1

Table 8: thermocycler and amplify

Input into library construction	Number of cycles required to generate	
	100 ng library	1 µg library
1 µg	0*	0 – 1*
500 ng	0*	2 – 3
250 ng	0 – 1*	3 – 5
100 ng	0 – 2*	5 – 6
50 ng	3 – 5	7 – 8
25 ng	5 – 6	8 – 10
10 ng	7 – 9	11 – 13
5 ng	9 – 11	13 – 14
2.5 ng	11 – 13	14 – 16
1 ng	13 – 15	17 – 19

Table 9: Recommended cycle numbers to generate 100 ng or 1 µg of amplified DNA when using KAPA UDI Adapters

Input amount	Amplification cycle number
50 – 500 ng*	3 – 4
10 ng	3 – 5
1 ng	6 – 8

*Table 10: Recommended number of amplification cycles to generate 4 nM** of amplified DNA when using*

Step 3. Purify the Amplified Sample Library using KAPA HyperPure Beads

Step 3a. Purify the Amplified Sample Library constructed using KAPA UDI Adapter & KAPA Library Amplification Primer Mix

1. Add 50 µL of room temperature, thoroughly resuspended, KAPA HyperPure Beads to each amplified sample library.
2. Mix the amplified sample library and KAPA HyperPure Beads thoroughly and centrifuge briefly to collect all droplets. Do NOT allow beads to pellet.
3. Incubate the sample at room temperature for 5 minutes to allow the amplified sample library to bind to the beads.
4. Place the sample on a magnet to capture the beads. Incubate until the liquid is clear.
5. Carefully remove and discard the supernatant.
6. Keeping the sample on the magnet, add 200 µL of freshly-prepared 80% ethanol.
7. Incubate the sample at room temperature for ≥30 seconds.
8. Carefully remove and discard the ethanol.
9. Keeping the sample on the magnet, add 200 µL of freshly-prepared 80% ethanol.
10. Incubate the sample at room temperature for ≥30 seconds.
11. Carefully remove and discard the ethanol. Remove residual ethanol without disturbing the beads.
12. Allow the beads to dry at room temperature, sufficiently for all of the ethanol to evaporate.
13. Remove the sample from the magnet
14. Thoroughly resuspend the beads in 25 µL (or appropriate volume) of 10 mM Tris-HCl, pH 8.0 – 8.5. Centrifuge briefly to collect all droplets. Do NOT allow beads to pellet.
15. . Incubate the sample at room temperature for 2 minutes to allow the amplified sample library to elute off the beads.

16. . Place the sample on the magnet to capture the beads. Incubate until the liquid is clear.
17. Transfer an appropriate volume of the clear supernatant to a fresh tube(s)/well and proceed with double-sided size selection (refer to Appendix B), library QC, target capture or sequencing (KAPA HyperPlus Kit, March 2024, Version 10.0).
18. Purified, amplified sample libraries can be stored at +2°C to +8°C for 1 – 2 weeks or at -15°C to -25°C for up to 3 months.

Step 3b. Purify the Amplified Sample Library constructed using KAPA Universal Adapter & KAPA UDI Primer Mixes

1. Add 50 µL of room temperature, thoroughly resuspended, KAPA HyperPure Beads to each amplified sample library.
2. Mix the amplified sample library and KAPA HyperPure Beads thoroughly and centrifuge briefly to collect all droplets.
3. Incubate the sample at room temperature for 5 minutes to allow the amplified sample library to bind to the beads.
4. Place the sample on a magnet to capture the beads. Incubate until the liquid is clear.
5. Carefully remove and discard the supernatant.
6. Remove the tubes from the magnet, and resuspend the beads in 50 µL of Nuclease-free, PCR-grade water or 10 mM Tris-HCl, pH 8.0 – 8.5.
7. Add 50 µL of room temperature, thoroughly resuspended, of KAPA HyperPure Beads to each sample.
8. Mix thoroughly by pipetting or vortexing, and centrifuge briefly to collect all droplets.
9. . Incubate the sample at room temperature for 5 minutes to allow the amplified sample library to bind to the beads.
10. Place the sample on a magnet to capture the beads. Incubate until the liquid is clear
11. Carefully remove and discard the supernatant.
12. Keeping the sample on the magnet, add 200 µL of freshly-prepared 80% ethanol.
13. Incubate the sample at room temperature for ≥30 seconds.
14. Carefully remove and discard the ethanol.
15. Keeping the sample on the magnet, add 200 µL of freshly-prepared 80% ethanol.
16. Incubate the sample at room temperature for ≥30 seconds.

17. Carefully remove and discard the ethanol. Remove residual ethanol without disturbing the beads.
18. Allow the beads to dry at room temperature, sufficiently for all of the ethanol to evaporate.
19. Remove the sample from the magnet.
20. Thoroughly resuspend the beads in 25 µL (or appropriate volume) of 10 mM Tris-HCl, pH 8.0 – 8.5. Centrifuge briefly to collect all droplets. Do NOT allow beads to pellet.
21. Incubate the sample at room temperature for 2 minutes to allow the sample library to elute off the beads.
22. Place the sample on the magnet to capture the beads. Incubate until the liquid is clear.
23. 3. Transfer an appropriate volume of the clear supernatant to a fresh tube(s)/well and proceed with double-sided size selection (refer to Appendix B), library QC, target capture or sequencing (KAPA HyperPlus Kit, March 2024, Version 10.0).
24. Purified, amplified sample libraries can be stored at +2°C to +8°C for 1 - 2 weeks or at -15°C to -25°C for up to 3 months.

Chapter 5. Quality Control

DNA Input	Expected Conversion Rate
1 – 10 ng	5 – 20%
11 – 100 ng	10 – 50%
>100 ng	50 – 100%

Table 11: Expected conversion rates for DNA input ranges.

Typical electrophoretic profiles for libraries prepared with the KAPA HyperPlus Kit:

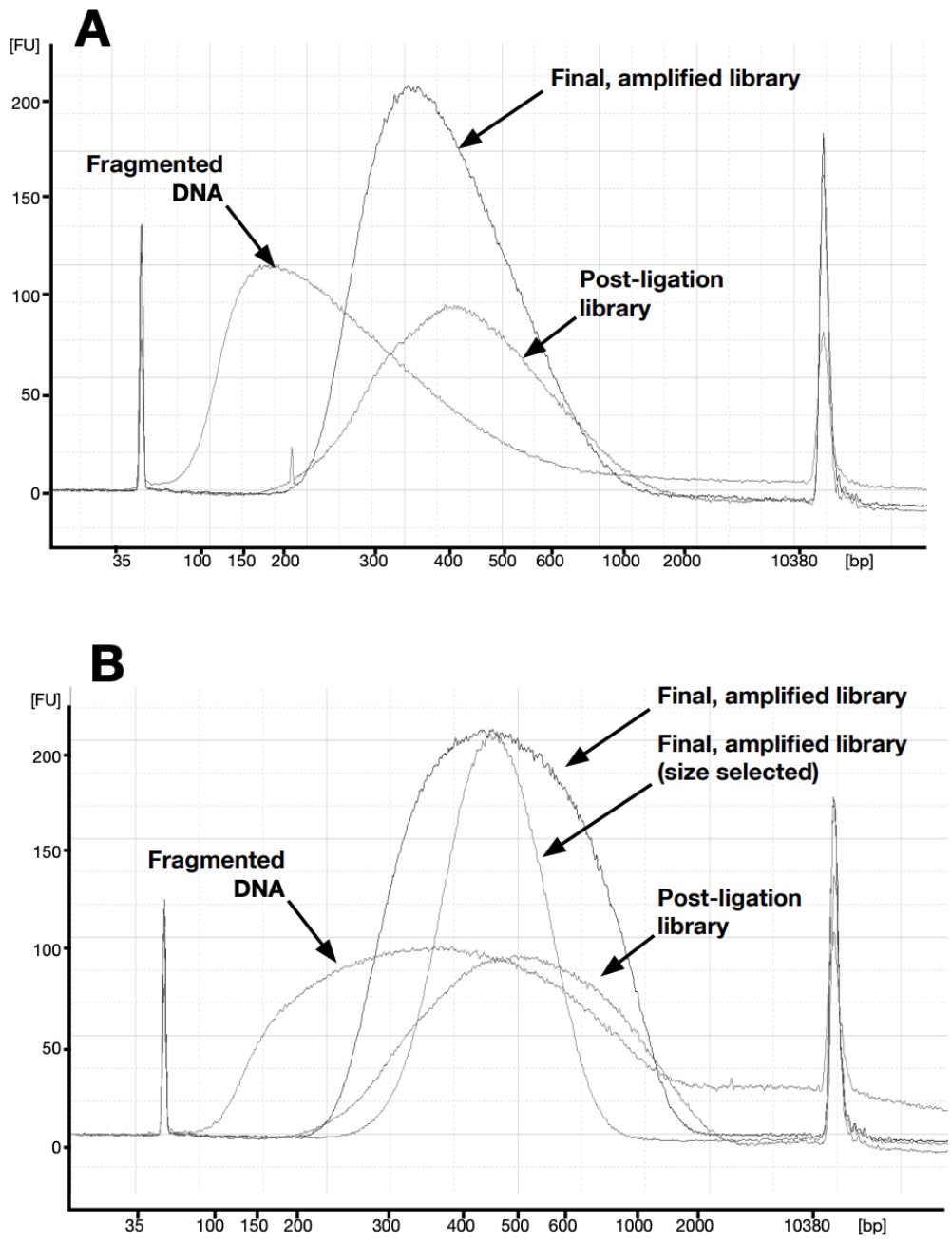


Figure 14: Examples of libraries prepared with the KAPA HyperPlus Kit

Libraries were send to Okinawa institute of science and technologies to perform PE250-sequenced on the Illumina HiSeq2500 platform or PE150-sequenced on the Illumina HiSeq4000 platform.

Figure 15: The HiSeq 2500 and HiSeq 4000 system Illumina

HiSeq 2500		HiSeq 3000	HiSeq 4000
Power and efficiency for large-scale genomics		Maximum throughput and lowest cost for production-scale genomics	
Production-scale genome, exome, transcriptome sequencing, and more			
Rapid run	High-output	—	—
1 or 2	1 or 2	1	1 or 2
10–300 Gb	50–1000 Gb	125–750 Gb	125–1500 Gb
7–60 hours	< 1–6 days	< 1–3.5 days	< 1–3.5 days
300 million	2 billion	2.5 billion	2.5 billion
2 × 250 bp	2 × 125 bp	2 × 150 bp	2 × 150 bp

Figure 16: Performanc of sequencing devices from Illumina

2.3 Preparing sequencing data for microbial annotation

Raw Illumina sequencing reads were processed to generate high-quality contigs using a standardized workflow. Quality control was performed using FastQC (Andrews, 2010) to assess base quality scores, GC content, and adapter contamination. Low-quality bases and adapter sequences were removed with Trimmomatic (Bolger et al., 2014), retaining reads longer than 50 bp. De novo assembly was conducted using SPAdes in “meta” mode (Bankevich et al., 2012), employing multiple k-mer sizes to reconstruct contigs and optimize assembly resolution. The assembled contigs were validated with QUAST (Gurevich et al., 2013), focusing on metrics such as N50, total length, and read mapping rates. Contigs shorter than 1,000 bp were excluded to enhance data reliability. Annotation was performed using the GTDB database (Parks et al., 2018) for taxonomic classification and DIAMOND BLAST (Buchfink et al., 2015).

2.4 Reconstruction of marker gene phylogenetic trees

Sequences from termite gut metagenome shorter than half the mean length of the marker gene were removed to improve the accuracy of phylogenetic reconstructions (Mering et al., 2007). 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). Aligned 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 these RY-recoded sequence alignments with IQ-TREE v. 1.6.12 (Nguyen et al., 2015). Substitution model for the tree reconstruction was used the GTR2 + G + I model of binary state. 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 selected from the bacterial tree of life (Parks et al., 2020a). The phylogenetic trees of archaeal and bacterial clades composed of sequences found exclusively in termite guts and represented by more than 10 termite species were extracted from the phylogenetic trees of each phylum. We refer to these trees, including sequences of termite gut bacteria exclusively, as termite-specific clades (TSCs). This procedure was followed for each marker gene. The phylogenetic trees reconstructed with the marker gene coding for COG0552 (ftsY)

were used as references. We attempted to link every TSC found in the phylogenetic trees reconstructed with COG0552 with their counterparts found in the phylogenetic trees reconstructed with the other nine marker genes. To do so, we searched the 198 gut metagenomes for contigs encompassing at least two of the 10 marker genes. The position of each marker gene sequence in their respective phylogenetic trees was used to match TSCs across marker gene trees. We also used the 10 marker genes of the termite gut bacterial genomes found in the GTDB database. Of the194425 genomes downloaded from the GTDB database, 37were associated with termite guts.

2.5 Preparing microbial contigs for CAZyme analysis

The open reading frames coding for CAZymes were identified among these metagenome contigs using Hidden Markov model searches against the dbCAN2 database (Zhang et al., 2018). Fragments of CAZyme sequences shorter than 50% of the expected CAZyme length were excluded from all analyses. Only hits with e-value lower than e-30 and coverage upward of 0.35 were considered for further analyses. For CAZymes composed of several modules, and it was necessary to separate the domains corresponding to specific CAZyme families (Beránková, 2024). CAZyme sequences from termite gut metagenomes were also searched against the GTDB database Release 207 (Parks et al., 2022) using Nucleotide-Nucleotide BLAST v2.10.0+ (Altschul et al., 1990) with default settings to obtain sequences not associated with termites. CAZyme sequences from non-termite environment from the GTDB database were filter from BLAST result using seqkit tool v2.0.0 (Shen et al., 2016).

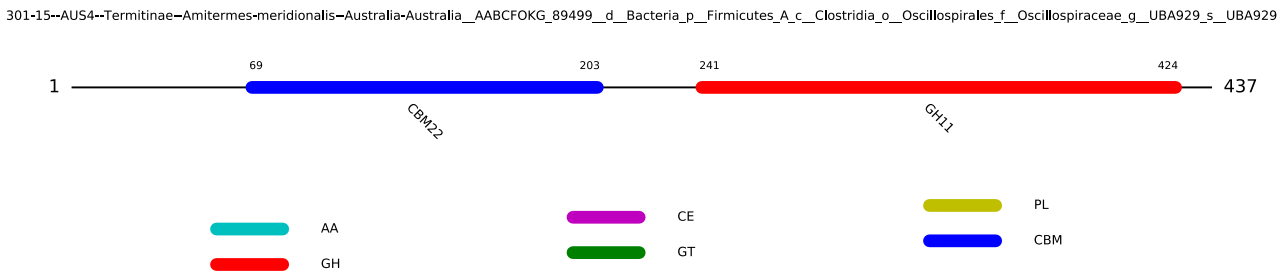


Figure 17: Scheme of hidden Markov model search

Visualization of identification of gene from CBM22 and GH11 in one termite gut metagenome contig (total number of contigs identified from microbial data was 101,941)

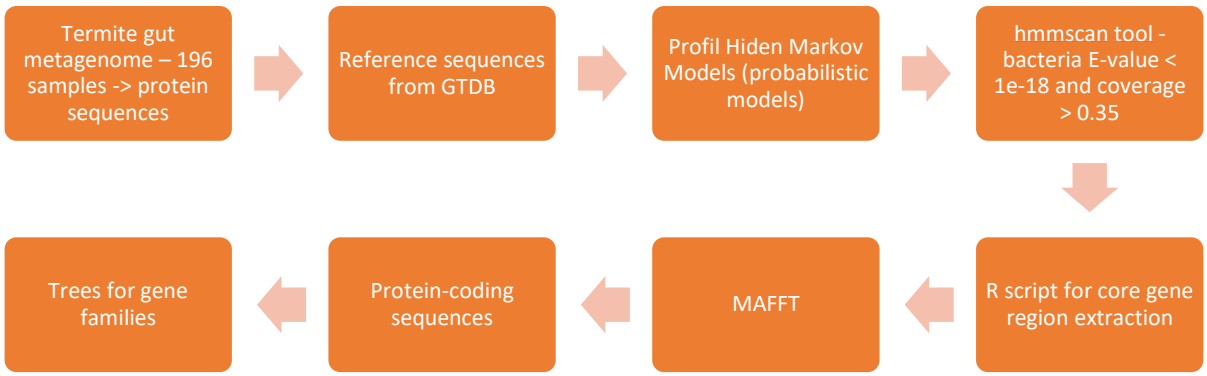


Figure 18: Simplified scheme for data analysis from annotated microbial contigs to individual CAZyme gene trees

2.6 Reconstruction of CAZyme phylogenetic trees

Phylogenetic trees were reconstructed for each CAZyme family comprising more than 20 sequences derived from termite gut bacteria. For large families divided into subfamilies, one phylogenetic tree was reconstructed for each CAZyme subfamily comprising more than 20 sequences derived from termite gut bacteria. Nucleotide sequences were converted to amino acid sequences using codon table 11 (bacterial and archaeal code) using Geneious Prime v2022.2.0. Protein sequences of each CAZyme gene family were aligned using MAFFT v7.490 with the"--auto setting" parameters recommended for multiple sequence alignments (Katoh et al., 2002; Katoh and Standley, 2013). Protein alignments were converted to nucleotide alignments using pal2nal v14.1-3 (Suyama et al, 2006) with codon table 11 (bacteria and archaea code). Maximum likelihood phylogenetic tree reconstructions were performed on nucleotide alignments using FastTree v2.1.11-2 (Price et al., 2009) with the "-gtr -gamma" setting. Phylogenetic trees of each CAZyme family were rooted using 20 sequences of related CAZyme families included in the analyses as outgroups, selected based on information available at www.cazy.org (Drula et al., 2022).

2.7 Identification of termite-specific CAZyme clusters

Phylogenetic trees of all CAZyme families were examined for clusters containing exclusively sequences derived from the gut metagenomes of termites and Cryptocercus. Clusters containing sequences from more than 20 termite and Cryptocercus samples were considered, and these clusters were designated as termite-specific clusters (TSCs). Clusters with

sequences from less than 20 samples were excluded from subsequent analyses. To assess the relative contribution of TSCs to termite wood digestion, the relative abundance of each TSC was calculated by mapping trimmed sequencing reads to CAZyme sequences. This procedure was performed separately for sequences from TSCs and sequences not belonging to any TSC. Reads were aligned using BWA mem v0.7.10 (Li and Durbin, 2009) and the resulting alignments were sorted ("sort") and fixed ("fixmate") using SAMtools v1.9 (Li et al., 2009). The number of reads mapping to each set of CAZymes was extracted using the SAMtools "flagstat" command. These reads were then used to estimate the proportion of CAZyms belonging to TSCs for each intestinal metagenome analysed in this study (Beránková et al 2024).

2.8 Termite tree reconstruction

Termite (host) tree was reconstructed with UCEs.(Ultra Conserved Elements) The phylogenetic tree was reconstructed with 322 of the 50,616 termite-specific UCE loci (Hellemans *et al.*, 2022). These 322 UCE loci were found in more than 57% of termite gut metagenomes and matched, at least partly, singly-annotated exons from the draft genome of *Zootermopsis nevadensis* (Terrapon et al., 2014). The maximum-likelihood phylogenetic tree was reconstructed using IQ-TREE v1.6.12 with a GTR+G+I model of nucleotide substitution and 1,000 ultrafast bootstrap replicates (UFB) to assess branch supports (Arora et al., 2023; Beránková et al., 2024; Hoang et al., 2018).

2.9 Cophylogenetic analysis of Prokaryota and Termites

Three approaches were applied to test for cophylogeny between termites and TSCs. The first approach employed the R package PACo (Procrustean Approach to Cophylogeny) (Balbuena et al., 2013), which implements Procrustean superimposition to estimate the cophylogenetic signal between two phylogenies. Host and symbiont phylogenetic trees (UCE tree) were converted into distance matrices using the cophenetic() function from the vegan R package (Oksanen et al., 2014). The analysis was performed using the backtracking method of randomization, which conserves the overall degree of interactions between the two trees (Oksanen et al., 2014).

The second approach relied on the generalized Robinson–Foulds (RF) metric, implemented via the ClusteringInfoDistance() function of the TreeDist R package (Smith, 2020). In the third approach, host and symbiont phylogenetic trees were matched to identify an optimal one-to-

one mapping between branches, following the method described by Nye et al. (Nye et al., 2006) and implemented in the NyeSimilarity() function of the TreeDist R package (Smith, 2020). Since these methods do not permit multiple symbiont tips per host, each host tip was split into a number of tips of zero branch length corresponding to the number of archaeal and bacterial symbionts present in the metagenome of that host (Perez-Lamarque and Morlon, 2019).

The strength of the cophylogenetic signal was quantified using each algorithm, with congruence between host and symbiont trees assessed through 10,000 random permutations. Analyses were conducted on phylogenetic trees reconstructed from both mitochondrial genomes and UCEs.

The number of host transfer events for each TSC was estimated using GeneRax software (Morel et al., 2020), a maximum-likelihood–based method that reconciles microbial gene trees with host trees and estimates horizontal transfer rates. This includes the probability of microbial transfer from one host to a non-ancestral random host. Each cophylogenetic analysis was performed twice, once using the termite tree derived from mitochondrial genomes and once with the tree based on UCEs.

Transfer rates obtained for TSC trees were compared with those calculated for 13 mitochondrial protein-coding genes (excluding third codon positions) and two rRNA genes. Since mitochondrial genomes do not undergo recombination, these genes share an identical evolutionary history and are not subject to horizontal transfer. Consequently, positive transfer rates in mitochondrial gene trees reflect phylogenetic reconstruction uncertainty and provide a baseline for interpreting horizontal transfer estimates. The evolutionary history of TSCs is considered predominantly shaped by vertical transfer when estimated rates of horizontal transfer fall within the baseline range established by mitochondrial genes.

2.9.1 Cophylogenetic analysis of CAZyme and Termites

Cophylogenetic analyses between termites and all TSCs were performed using three different approaches. The first approach used the Prokrust approach to cophylogeny, implemented in the R package PACo (Balbuena et al., 2013). For this approach, termite and TSC trees were converted to distance matrices using the cophenetic() function of the vegan R package

(Oksanen et al., 2014). The software was run using a backward randomization method to maintain the overall level of interactions between termite and TSC trees (Hutchinson et al., 2017). The second approach used a generalized Robinson Foulds (RF) metric (Smith, 2020), implemented in the ClusteringInfoDistance() function of the TreeDist R package (Smith 2020). The third approach used the Nye et al. (2006) method, implemented in the NyeSimilarity() function of the TreeDist R package (Smith 2020). In this approach, termite and TSC trees were compared to produce an optimal 1:1 map between branches. In the last two methods implemented in the TreeDist R package, each termite tip was divided into x tips with zero branch length, where x represents the number of CAZyme sequences associated with the metagenome corresponding to that termite tip (Perez-Lamarque and Morlon, 2019; Satler et al., 2019). The correspondence between termite and TSC trees was assessed using 10,000 random permutations (Beránková et al 2024).

2.10 Statistic analysis and scripts

Getting sequences from GTDB database were done on on cluster computation capabilities where was uploaded fasta file with GDTB database. To run this large amount of dataset were used usually those settings for capacity of computation resources:

```
#!/bin/bash
#SBATCH -p compute
#SBATCH -t 1-10:00:00
#SBATCH --mem 300G
#SBATCH -c 128
```

All files with metagenome information for each metagenome sample (supplementary table 1) were looped for the blast with command:

```
for f in *.fasta;
do
  blastp -query $f -db /bucket/BourguignonU/Terka/GTDBv207/GTDBp_new -outfmt '6 qseqid sseqid qcovhsp pident evalue bitscore length gapopen mismatch' -num_threads 120 -out $f*b.fasta;
done
```

Requirements for the information for the BLAST results were choose in this order:

- 6: Tabular format with user-specified fields.
- qseqid: Query sequence ID. This is the identifier of the sequence you provided as the query in the BLAST search.

- sseqid: Subject sequence ID. This is the identifier of the sequence from the database that matches your query sequence.
- qcovhsp: Query coverage per HSP (high-scoring segment pair). This shows the percentage of the query sequence covered by the aligned region(s) in the subject.
- pident: Percentage of identical matches. This indicates the percentage of identical amino acid residues in the aligned region.
- evalue: Expect value. This measures the statistical significance of the match; lower values indicate more significant matches.
- bitscore: Bit score. This provides a normalized score for the alignment, which takes into account the raw score and the scoring matrix.
- length: Alignment length. This is the total length of the aligned region between the query and subject sequences.
- gapopen: Number of gap openings. This indicates how many gaps were introduced in the alignment.
- mismatch: Number of mismatches. This counts the amino acids that do not match between the query and subject sequences in the alignment.

After running BLAST, the IDs of all sequences with at least 50% similarity to each metagenome from the termite gut were extracted. A command was then used to filter these IDs from the results and remove any duplicates:

```
for f in *blast.txt;
do
awk '{print $2}' $f > $f*id.txt;
done

# deleting duplicate lines
for f in *.hmm._id.txt; do sort -u $f > $f*nd.txt; done

renaming

for f in *.hmm._id.txt*nd.txt*; do mv -i -- "$f" "${f//.hmm._id.txt*nd.txt/_id_nd.txt}"; done
```

After obtaining unique IDs for each metagenome dataset, the corresponding sequences were extracted from GTDB and saved into FASTA files:

```
module load ncbi-blast/2.10.0+
for f in *_id_nd.txt;
```

```
do
blastdbcmd -entry_batch $f -db /GTDBv207/GTDBp -out $f*seq.fasta;
done
```

Sequences from GTDB and termite metagenome were combined together in one fasta file for each sample:

```
for f in *_termite_gutmetagenome_oist.fasta; do
t=${f/_termite_gutmetagenome_oist./_gtdb_hmm_nd.}
out=${f/_termite_gutmetagenome_oist./_prot_all.}
cat "$f" "$t" > "$out"
done
```

To every gene family was added outgroup:

```
for f in GH*prot_all.fasta; do cat $f o-GH11_outgroup.fasta > $f*protein_all.fasta; done
for f in GT*prot_all.fasta; do cat $f o-GT1_outgroup.fasta > $f*protein_all.fasta; done
for f in PL*prot_all.fasta; do cat $f o-PL4_outgroup.fasta > $f*protein_all.fasta; done
for f in CE*prot_all.fasta; do cat $f o-CE11_outgroup.fasta > $f*protein_all.fasta; done
for f in CBM*prot_all.fasta; do cat $f o-CBM67_outgroup.fasta > $f*protein_all.fasta; done
for f in AA*prot_all.fasta; do cat $f o-AA10_outgroup.fasta > $f*protein_all.fasta; done
```

Multiple alignment was done for every sample separately:

```
module load mafft/7.475-1
mafft --maxiterate 1000 --localpair *_prot_all.fasta > *_MA.fasta
```

Multiple alignment was then translated from the protein to the nucleic acid:

```
module load bioinfo-ugrp-modules  DebianMed/11.0
module load emboss/6.6.0
backtranseq -sequence protein_multiplealignment.fasta -cfile Eecoli.cut -outfile DNA_from_prot_MA.fasta

# translate to protein-coding

module load pal2nal/14.1-3

pal2nal.pl protein_multiplealignment.fasta DNA_from_protMA.fasta -output fasta > protein_coding.fasta -
codontable 11
```

Phylogenetic analysis were performed by fasttree and IQtree:

```
for f in *protcoding.fasta; do /FastTree $f -nt -gtr -gamma -out $f*tree; done

ruse /Tool/iqtree-2.1.2-Linux/bin/iqtree2 -s GH11_protcoding.fasta -nt AUTO -bb 1000 -rcluster 10 -m TEST -
pre gene1 -st CODON11
```

Cophylogeny analysis were performed by several steps in R. Scripts were first tested on one sample file and then run on all dataset. The script is designed to evaluate and quantify the coevolutionary dynamics between termites and their symbiotic microorganisms. By analyzing congruence between phylogenetic trees, it provides insights into shared evolutionary histories and potential host-symbiont co-divergence.

Key Packages Used:

• ape: Handles phylogenetic tree reading, manipulation, and distance matrix generation.

• paco: Performs phylogenetic congruence tests and identifies coevolutionary patterns.

• ggtree **and** treeio: Visualizes and modifies tree structures.

• TreeDist **and** TreeTools: Quantifies tree similarity using advanced metrics.

• HOME: Adds tips and enforces ultrametricity in host trees for comparative analyses.

To facilitate analysis, distance matrices are generated for both the host (termite) and symbiont trees. Using the cophenetic() function, pairwise evolutionary distances are computed for each tree. These matrices are further combined into an HE matrix, which links host and symbiont labels by marking matching tips in both trees. This matrix forms the foundation for assessing shared evolutionary histories.

A key part of the script involves PACo analysis, conducted with the paco package to evaluate phylogenetic congruence between hosts and symbionts. This process includes preparing matrices of host and symbiont distances and their associations, followed by a Principal Coordinate Transformation to correct for phylogenetic signal distortions. Permutation tests are employed to assess coevolutionary signals using randomization methods. Additionally, residual and link strength analyses are performed to measure the degree of individual host-symbiont interactions.

The script also manipulates trees to ensure compatibility between host and symbiont structures. This involves renaming taxa and adding tips with small branch lengths to symbiont and host trees. These steps, facilitated by packages like ggtree, treeio, and HOME, ensure that both trees are appropriately matched for analysis.

Tree distance comparisons are another critical component, utilizing methods such as the Generalized Robinson-Foulds (RF) and Nye Similarity metrics. These methods, implemented through the TreeDist and TreeTools packages, quantify tree topological similarities. By generating randomized trees, the script calculates p-values for the observed tree congruence, providing statistical rigor to the analysis.

The outputs generated by the script include matrices and result files that summarize coevolutionary metrics such as PACo p-values, adjusted phylogenetic trees with renamed taxa, and statistical results from tree similarity analyses. These outputs offer insights into the phylogenetic relationships between termites and their symbionts (whole script Supplementary file 1).

.

3. Results

3.1 Cophylogenetic Analysis of prokaryote and host

The sequences were derived from 196 termite gut metagenomes and one Cryptocercus metagenome combined with sequences from the GTDB database (Parks et al., 2020b). Separate ML phylogenies were reconstructed for each marker gene and for each bacterial and archaeal phylum. Each tree was then searched for TSCs composed exclusively of sequences associated with termites, represented in at least 10 termite species. This analysis identified between 8 and 34 TSCs per marker gene. As a reference marker gene, *ftsY* (COG0552) was selected, containing 2299 sequences forming 27 TSCs. These 27 TSCs of COG0552 were distributed across nine bacterial and two archaeal phyla. The cophylogenetic signal between each TSC (COG0552) and its termite host was examined using termite phylogenetic tree reconstructed with UCE data and three different methods: PACo (Balbuena et al., 2013), the generalized RF metric (Smith, 2020), and the tree alignment algorithm described by Nye et al. (Nye et al., 2006). Significant cophylogenetic signals with termites were observed in 18 of the 27 TSCs across all three methods.

The TSCs with the strongest cophylogenetic signals included key components of the gut microbiota of termites. For example, the families Ruminococcaceae (phylum Bacillota, formerly Firmicutes) and Breznakiellaceae (phylum Spirochaetota), respectively, made up 16.5% and 20.0% of the 16S rRNA gene sequences found in a survey of 94 termite species (Bourguignon et al., 2018). Brezna-kiellaceae generally have a fermentative metabolism and include strains capable of reductive acetogenesis (Leadbetter et al., 1999; Song et al., 2021). They have been isolated from the guts of cockroaches, suggesting that they were already present in the ancestor of termites and their cockroach sister group, Cryptocercidae (Brune et al., 2022; Song et al., 2021).

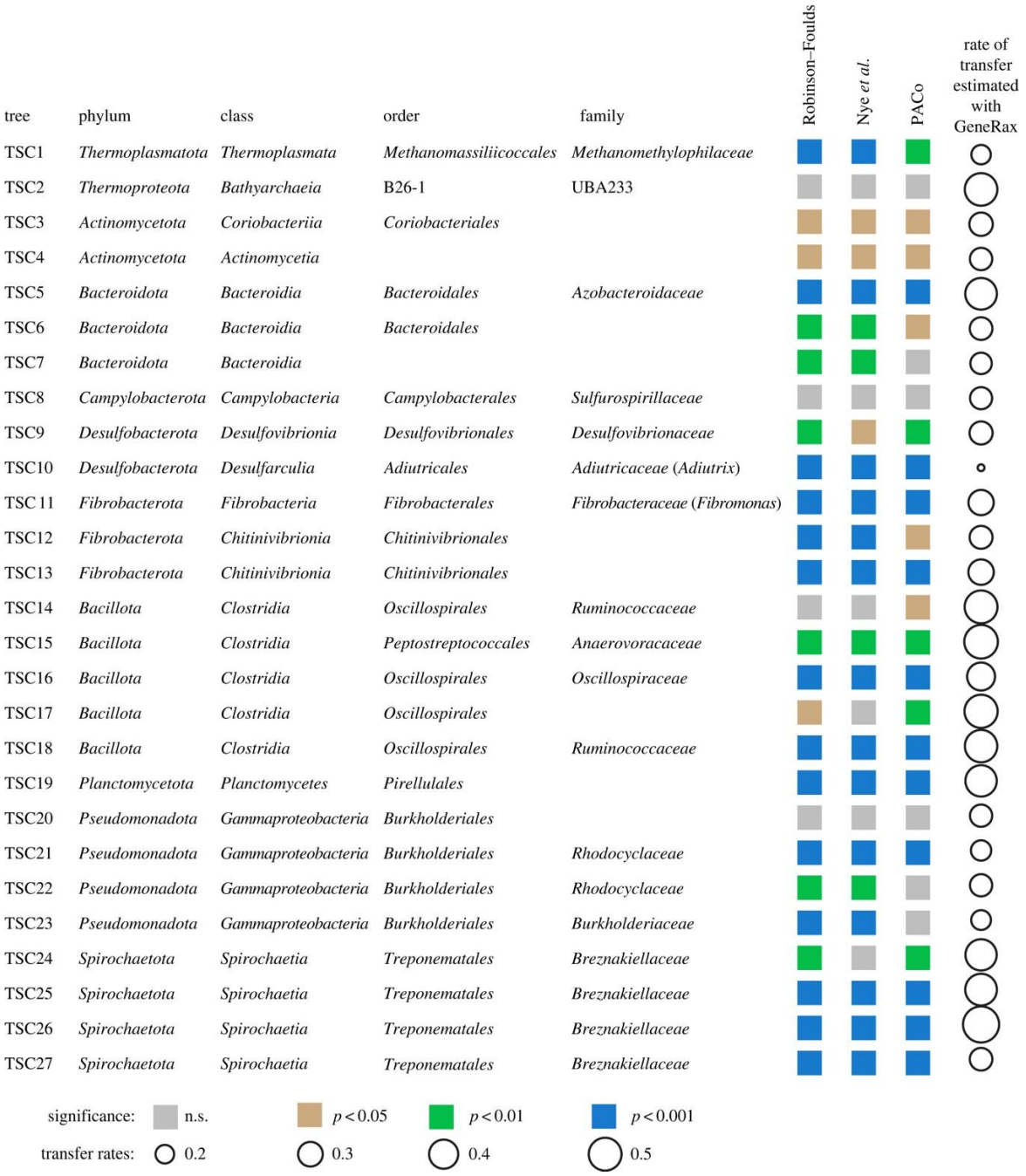


Figure 19: Results of the cophylogenetic analyses performed on the marker gene COG0552

Results of the cophylogenetic analyses performed on the marker gene COG0552 of 27 termite-specific archaeal and bacterial clades (TSCs). The cophy-logenetic analyses were performed with three different methods: PACo, the generalized RF metric, and the tree alignment algorithm described by Nye et al. (Nye et al., 2006).The transfer rates were estimated using the ML method implemented in the GeneRax software.

Therefore, TSCs with essential functions and a long history of association with termites show cophylogenetic signals. In principle, the observed cophylogenetic signals between TSCs and their termite hosts could be caused by two different mechanisms: (i) vertical transmission of gut bacteria from parent colonies to daughter colonies, which is caused by the transmission of gut bacteria among family members and results in the coevolution of symbionts and hosts; (ii) limited horizontal transfers of gut bacteria among the diverging termite species due to geographical barriers, which would not require vertical transfers and results in allopatric speciation (Groussin et al., 2020; Vienne et al., 2013). If vertical transfer were responsible for the cophylogenetic signals, it should give rise to bacterial lineages associated exclusively with specific termite clades and not shared with other sympatric termites.

The analysis identified termite clade-specific lineages (TCSLs) within many TSCs. For example, several TCSLs were found belonging to the family Breznakiellaceae, the genus *Fibromonas* (phylum Fibrobacterota), and the genus *Adiutrix* (phylum Desulfobacterota), which were exclusively associated with the densely sampled genus *Microcerotermes* (figure 13a–d). These TCSLs were absent from the guts of other termites, including many species that are sympatric with *Microcerotermes*, demonstrating that some TCSLs are endemic to the gut of specific termite genera, as previously hypothesized based on smaller datasets (Hongoh et al., 2005). They were found in the guts of *Microcerotermes* species collected across four continents and six biogeographic realms, indicating that *Microcerotermes* dispersed worldwide with their specific gut bacteria. The research also identified TCSLs associated with termite clades that were sampled less intensively. For instance, a group of Nasutitermitinae, sharing a common ancestor approximately 25 million years ago and sampled across multiple continents, hosted several TCSLs belonging to the family Breznakiellaceae and the genus *Adiutrix* (figure 13a,b,d).

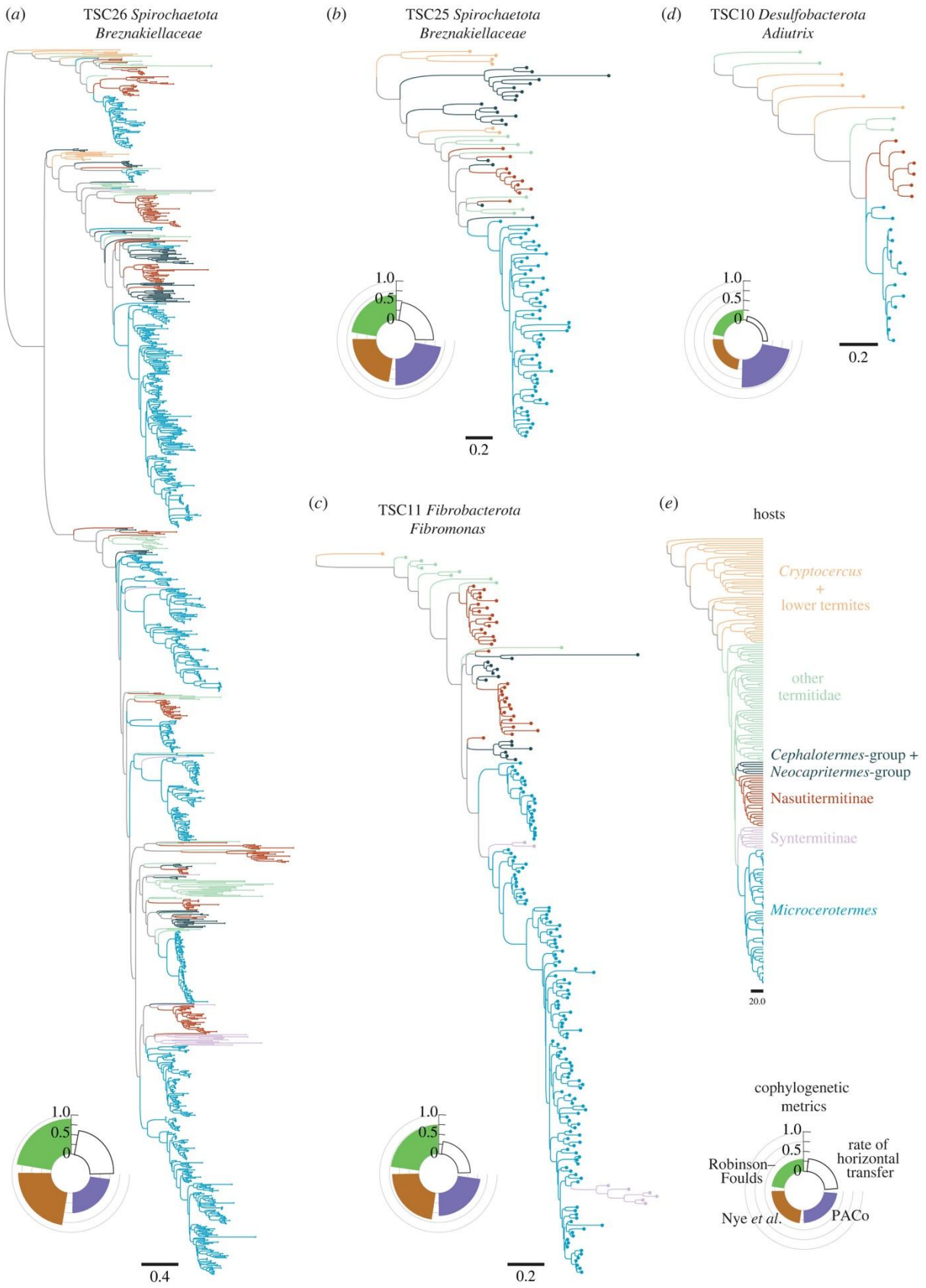


Figure 20: Selected phylogenetic trees of termite-specific bacterial clades (TSCs)

These examples of the absence of horizontal transfer of bacteria between sympatric termites belonging to different clades indicate that allopatry is not required to maintain the association between termite clades and their symbiotic bacteria. Therefore, even if allopatric speciation of termites and TCSLs likely occurred, TCSLs are transmitted vertically from parent colonies to daughter colonies and possibly horizontally among related termite species forming a clade.

The number of host transfer events for each TSC was estimated using the ML method implemented in the GeneRax software (Morel et al., 2020). The estimated transfer rates with the UCE-based termite phylogenetic tree varying between 0.13 and 0.61. Notably, 16 TSCs had rates of transfer falling between 0.11 and 0.32, the range of rates of transfer estimated for each of the 13 protein-coding and two rRNA mitochondrial genes used in this study to build the phylogenetic tree of termites (figure 14a).

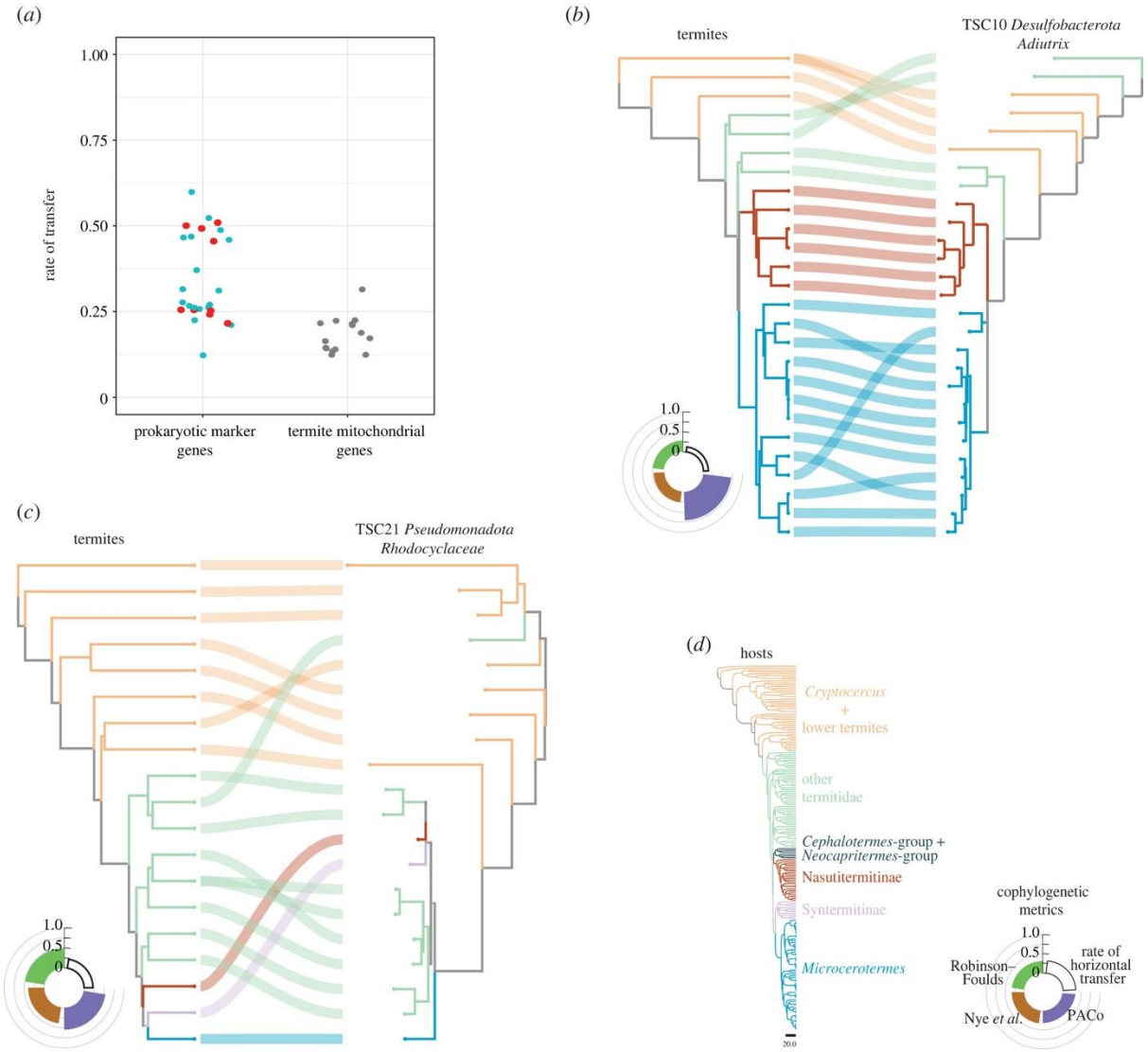


Figure 21: Rate of transfer and phylogenetic trees of some termite-specific bacterial clades (TSCs)

Mitochondrial genes are expected to experience no transfer and have an identical evolutionary history, providing a baseline for estimated rates of transfer values obtained for genes expected to experience no horizontal transfer. While these results do not prove the absence of horizontal transfers, they suggest that the cophylogenetic patterns observed between some TSCs and termites may not involve any horizontal transfers.

Cophylogenetic patterns would be obfuscated by bacterial extinction (or insufficient sequencing depth, from which it cannot be distinguished) and speciation taking place within non-speciating termite hosts (Groussin et al., 2020). Several TSCs, less speciose than Breznakiellaceae and *Fibromonas*, depicted patterns of cophylogeny across large parts of the termite phylogenetic tree (figure 14). For example, the phylogenetic tree of the genus *Adiutrix* found in the termite sister group Cryptocercidae, three families of termites, and across Termitidae, was highly congruent with the phylogenetic tree of termites (figure 14b). The phylogenetic tree of the family *Rhodocyclaceae* (phylum Pseudomonadota, formerly Proteobacteria) (figure 14c) is another example of a clade showing significant cophylogenetic signal with termites. These cophylogenetic patterns between termites and certain gut bacterial symbionts are interpreted as evidence of coevolution, with vertical transmission occurring over several tens of millions of years.

3.2 Taxonomy annotation of CAZyme sequences

The analysis of bacterial genomes from termite guts revealed 101,941 CAZyme sequences within the metagenome assemblies from 196 samples of termites and Cryptocercus. Detailed examination of termite gut metagenomes uncovered 5 to 168 CAZyme sequences in individual metagenome. These sequences are distributed among 259 distinct families and subfamilies, including 96 glycoside hydrolases (GHs), 42 glycosyltransferases (GTs), 11 Polysaccharide Lyases (PLs), 14 Carbohydrate Esterases (CEs), 5 auxiliary activities (AAs), and 12 Carbohydrate-Binding Modules (CBMs). Further, 11 of these CAZyme families are present in more than 55% of the analysed termite gut metagenomes. Following, 34 CAZymes were found in upward of 70% of gut metagenomes, nine of which, GH3, GH5, GH13, GH43, GH77, GT4, GT5, GT28, and GT51, were found in more than 90% of gut metagenomes.

Of the 259 reconstructed CAZyme trees, 116 contained at least one cluster of CAZyme from termite environment and from at least 20 termite and *Cryptocercus* samples (supplementary table 1) CAZyme GT51 contains 23 TSC, which is the largest amount of sequences in one cluster. Across all CAZyme trees were identified 420 TSC with average number of 120 sequences in one cluster. The largest TSC is in the tree for GH77, with 1080 sequences primarily belonging to *Breznakiellaceae* (phylum *Spirochaetota*, previously family *Treponemataceae*).

Four of the 420 maximum-likelihood phylogenetic trees of termite-specific bacterial clusters (TSCs).

All four trees showed strong cophylogenetic signals with termites. The trees included several termite clade-specific CAZyme clusters only found in Nasutitermitinae and *Microcerotermes*. Phylogenetic trees of (**A**) GH2 Cluster 10 composed of 97.4% of *Spirochaetota*, (**B**) GH77 Cluster 6 composed of 98.1% of *Spirochaetota*, (**C**) GH9 Cluster 7 composed of 100% of *Fibrobacterota*, and (**D**) GH8 Cluster 4 composed of 100% of *Fibrobacterota*. **E** Maximum-likelihood phylogenetic tree of termites inferred from UCEs. Black dots indicate CAZyme sequences assigned to a different bacterial phylum.

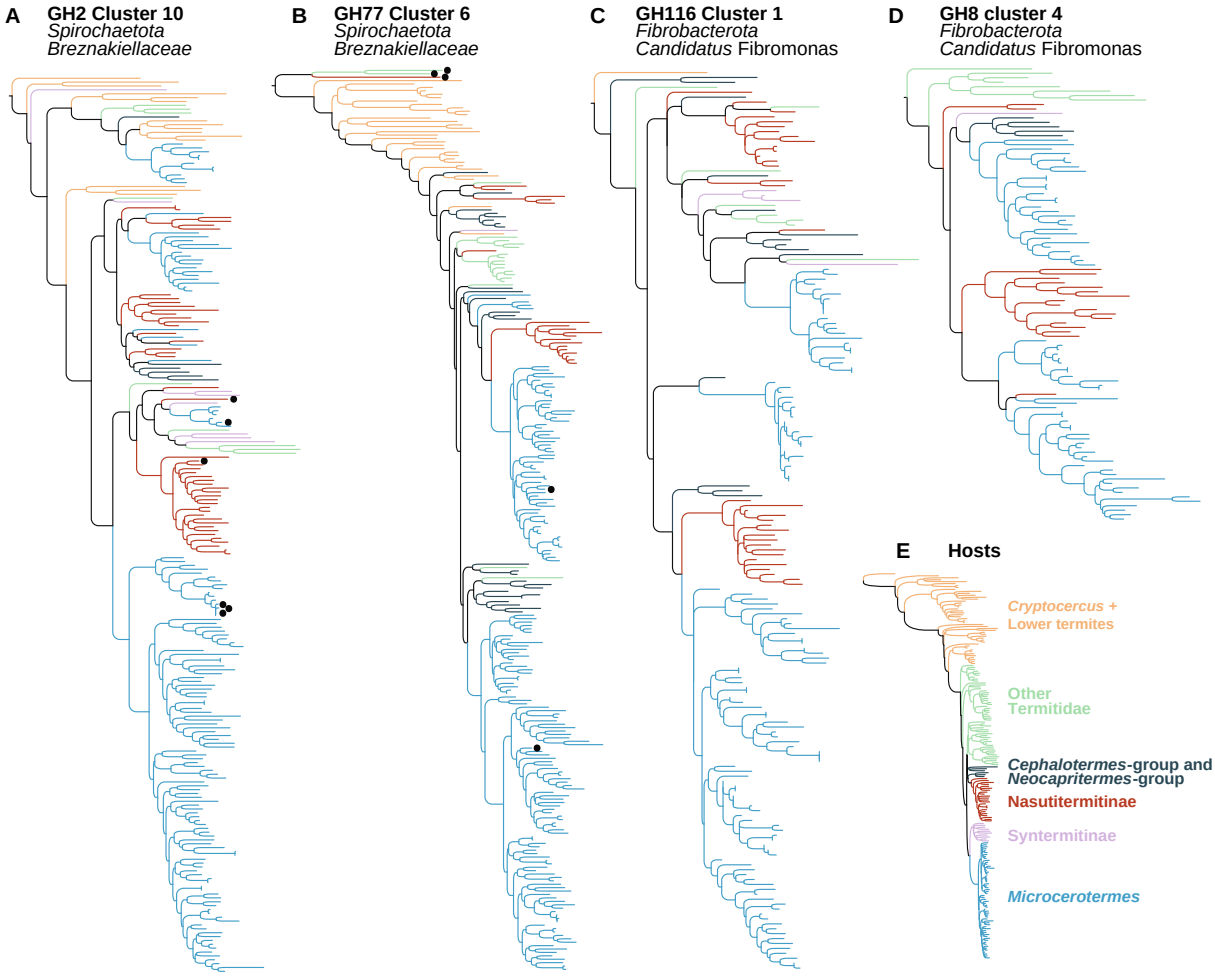


Figure 22: Four of the 420 maximum-likelihood phylogenetic trees of termite-specific bacterial clusters (TSCs).

Three of the 420 maximum-likelihood phylogenetic trees of termite-specific bacterial clusters (TSCs).

All three trees showed strong cophylogenetic signals with termites and included termite clade-specific CAZyme clusters associated with Kalotermitidae or non-Termitidae Neoisoptera. Phylogenetic trees of (**A**) GH57 Cluster 7 composed of *Bacteroidota* only and including the genus *Candidatus* Azobacteroides, (**B**) GH10 Cluster 13 composed of 98.5% of *Bacteroidota*, and (**C**) GH73 Cluster 3 composed of 97.6% of *Bacteroidota*. **D** Maximum-likelihood phylogenetic tree of termites inferred from UCEs. *CAZyme sequences annotated as *Candidatus* Azobacteroides; **CAZyme sequences assigned to *Candidatus* Azobacteroides with BLAST search against the GenBank database; X CAZyme sequences originally annotated as *Candidatus* Azobacteroides but with conflicting BLAST search against the GenBank database. Black dots indicate CAZyme sequences assigned to a different bacterial phylum.

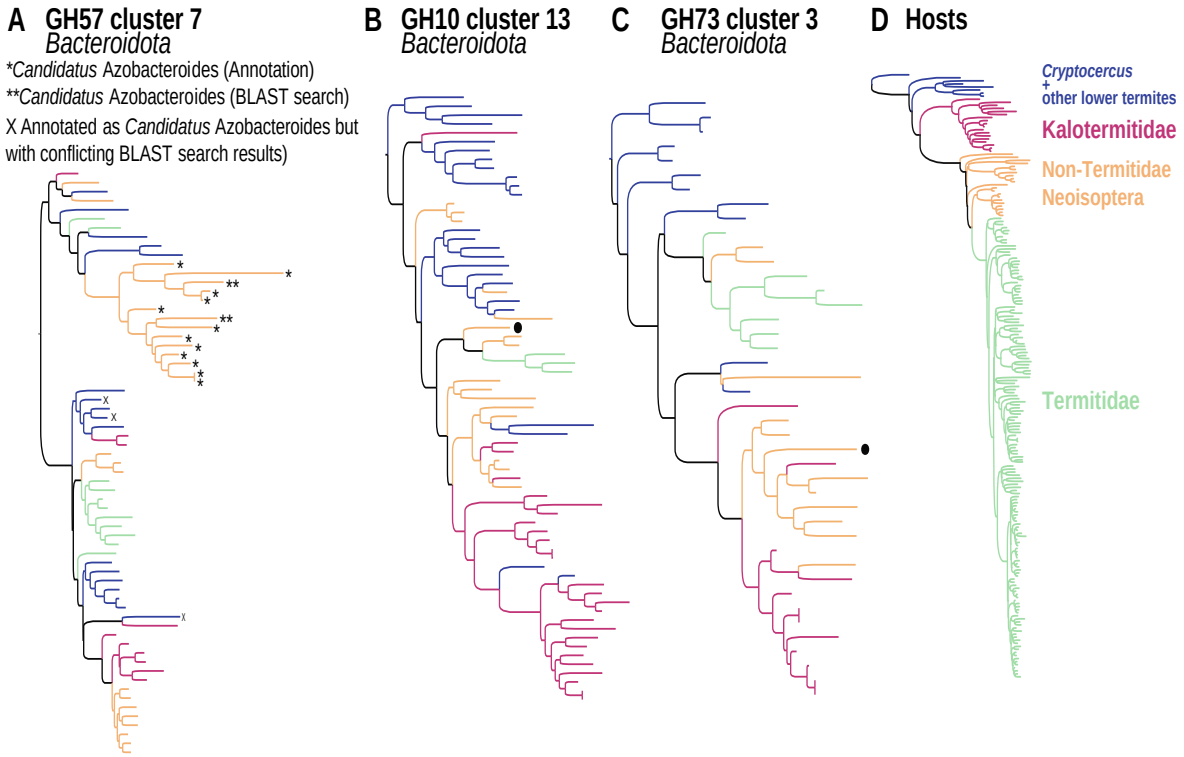


Figure 23: Three of the 420 maximum-likelihood phylogenetic trees of termite-specific bacterial clusters (TSCs).

Maximum-likelihood phylogenetic trees of three of the 131 termite-specific bacterial clusters (TSCs) containing at least one sequence of *Cryptocercus* and/or *Mastotermes.*

The three trees showed strong cophylogenetic signals with termites. Phylogenetic trees of (**A**) GH3 Cluster 4 composed of 97.3% of *Spirochaetota*, (**B**) CE1 Cluster 2 composed of 86.2% of *Spirochaetota*, and (**C**) GH29 Cluster 5 composed of 97.7% of *Bacteroidota*. **D** Maximum-likelihood phylogenetic tree of termites inferred from UCEs. **Cryptocercus kyebangensis*; ***Mastotermes darwiniensis*. Black dots indicate CAZyme sequences assigned to a different bacterial phylum.

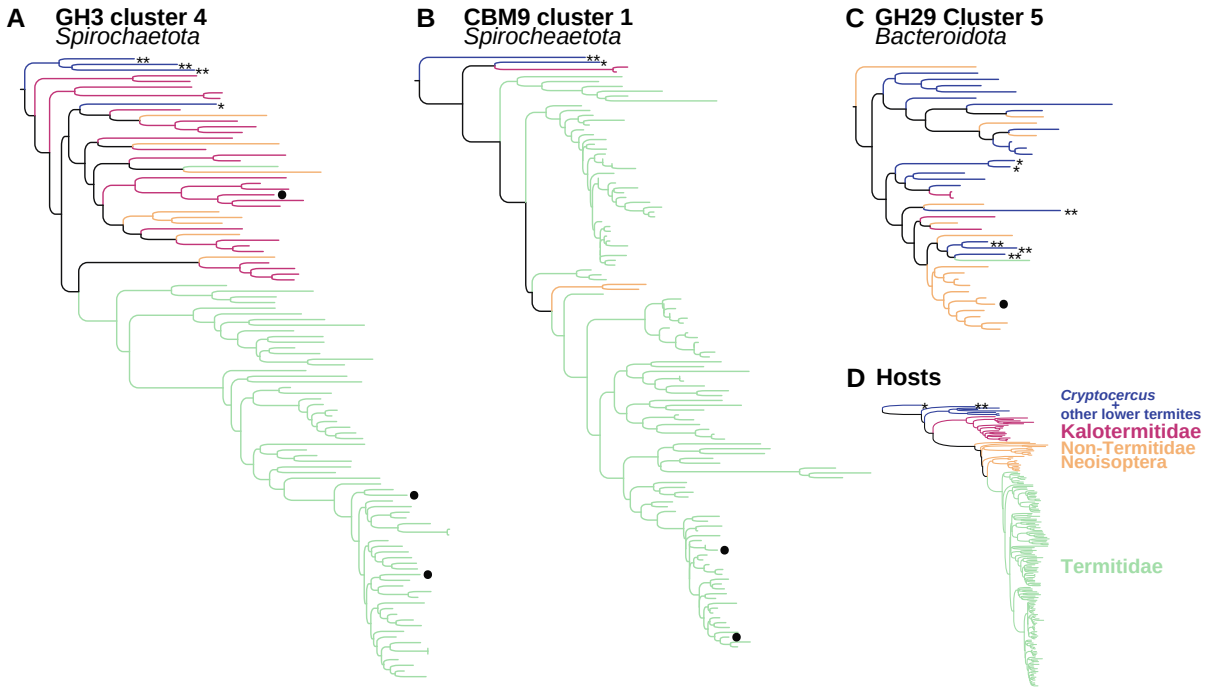


Figure 24: Maximum-likelihood phylogenetic trees of three of the 131 termite-specific bacterial clusters (TSCs) containing at least one sequence of Cryptocercus and/or Mastotermes.

Maximum-likelihood phylogenetic trees of six of the 175 termite-specific bacterial clusters (TSCs) strictly associated with Termitidae.

The six trees showed strong cophylogenetic signals with termites. Phylogenetic trees of (**A**) GH77 Cluster 10 composed of 99.4% of *Fibrobacterota*, primarily of the family Chitinispirillaceae, (**B**) GH57 Cluster 10 composed only of *Fibrobacterota*, primarily of the family Chitinispirillaceae, (**C**) GH5_2 Cluster 9 composed only of *Fibrobacterota* of the genus *Candidatus* Fibromonas, (**D**) GH26 Cluster 7 composed only of *Fibrobacterota*, primarily of the genus *Candidatus* Fibromonas, (**E**) GH4 Cluster 4 composed of 94.9% of *Spirochaetota*, and (**F**) GH57 Cluster 2 only composed of *Spirochaetota*. **G** Maximum-likelihood phylogenetic tree of Termitidae inferred from UCEs. Black dots indicate CAZyme sequences assigned to a different bacterial phylum.

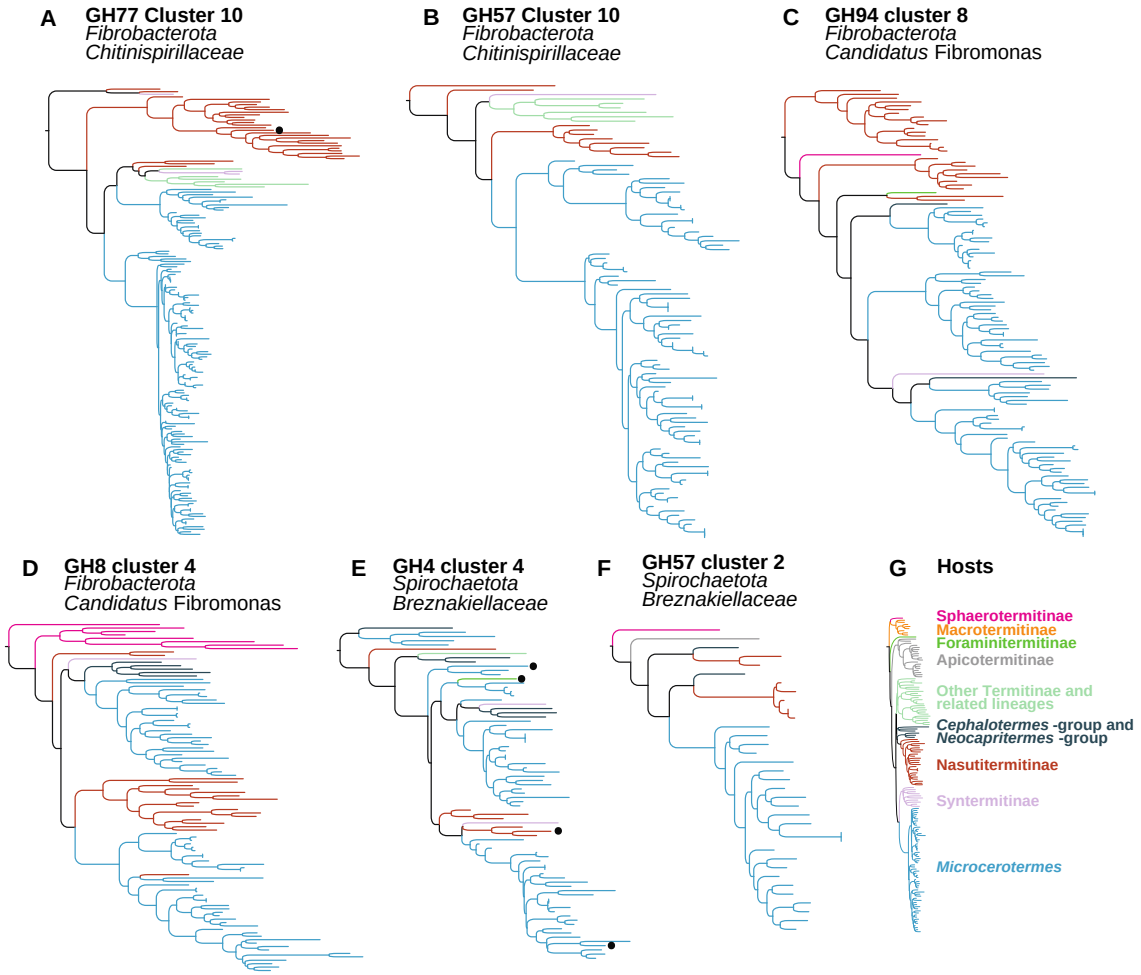


Figure 25: Maximum-likelihood phylogenetic trees of six of the 175 termite-specific bacterial clusters (TSCs) strictly associated with Termitidae.

3.3 Cophylogenetic Analysis of CAZymes and host

Research also focused on cophylogenetic analyses between all TSC and termite phylogenetic tree reconstructed using UCEs (Hellemans et al., 2022). Three cophylogenetic methods were used: PACo (Balbuena et al., 2013), the generalized Robison-Foulds metric (Smith 2020), and the method of Nye et al. (Nye et al., 2006). 392 of the 420 TSCs showed a significant cophylogenetic signal with the three methods, 315 of which were highly significant ($p < 0.001$) for all three methods (Supplementary table 2). TSCs with significant cophylogenetic signals made up an average of 42.3% of the CAZyme sequences to which 44.5% of the trimmed CAZyme reads mapped (Supplementary table 2).

ID cluster	Proportion of *Fibrobacterd*	Proportion of *Spirochaetot a*	Proportion of *Bacteroidd*	Proportion of *Firmicutes*	Proportion of other prokaryot	Total number of sequenc	PACo p-value	Nye et al. p-value	Generalized Robinson-Foulds p-value
GH77_cluster_2	0.00462963	0.99166667	0.00277778	0.000925926	0	1080	0,***	0,***	0,***
GT26_cluster_1	0.018009479	0.88909953	0.00189574	0.090047393	0.00094787	1055	0,***	0,***	0,***
GH5_4_cluster_10	0.031536114	0.37436419	0.00406918	0.578840285	0.01119023	983	0,***	0,***	0,***
GH18_cluster_3	0.009009009	0.98536036	0	0.005630631	0	888	0,***	0,***	0,***
GH13_11_cluster_3	0.019450801	0.97597254	0	0	0.00457666	874	0,***	0,***	0,***
GH5_4_cluster_4	0.974595843	0.02309469	0.00115473	0.001154734	0	866	0,***	0,***	0,***
GH130_cluster_4	0.372389791	0.59976798	0	0.027842227	0	862	0,***	0,***	0,***
GH10_cluster_6	0.090686275	0.82720588	0.00490196	0.036764706	0.04044118	816	0,***	0,***	0,***
GH3_cluster_6	0.004944376	0.98393078	0	0.008652658	0.00247219	809	0,***	0,***	0,***
GH5_12_cluster_1	0.024937656	0.96758105	0	0.003740648	0.00374065	802	0,***	0,***	0,***
GH57_cluster_13	0.020512821	0.97435897	0	0.003846154	0.00128205	780	0,***	0,***	0,***
GT51_cluster_19	0.010159652	0.98403483	0	0.004354136	0.00145138	689	0,***	0,***	0,***
GH3_cluster_2	0.004451039	0.98516321	0.00296736	0.007418398	0	674	0,***	0,***	0,***
GH5_39_cluster_1	0.604754829	0.35215453	0.02971768	0.011887073	0.00148588	673	0,***	0,***	0,***
GT35_cluster_1	0.011904762	0.98809524	0	0	0	672	0,***	0,***	0,***
GT28_cluster_20	0.020030817	0.97534669	0	0.004622496	0	649	0,***	0,***	0,***
GH20_cluster_3	0.012987013	0.96266234	0.00162338	0.021103896	0.00162338	616	0,***	0,***	0,***
GH57_cluster_3	0.030456853	0.95939086	0.00507614	0.005076142	0	591	0,***	0,***	0,***
GH5_52_cluster_2	0.840659341	0.15750916	0	0.001831502	0	546	0,***	0,***	0,***
GH94_cluster_11	0.954183267	0.0438247	0	0.001992032	0	502	0,***	0,***	0,***
GH3_cluster_12	0.034693878	0.24285714	0.00204082	0.716326531	0.00408163	490	0,***	0,***	0,***
GH43_1_cluster_3	0.006160164	0.46406571	0.44147844	0.0862423	0.00205339	487	0,***	0,***	0,***
GH30_8_cluster_4	0.676348548	0.29460581	0.00622407	0.01659751	0.00622407	482	0,***	0,***	0,***
GH9_cluster_4	0.074766355	0.28738318	0	0.63317757	0.0046729	428	0,***	0,***	0,***
GH3_cluster_18	0.018867925	0.98113208	0	0	0	424	0,***	0,***	0,***
GH13_20_cluster_4	0.824940048	0.14148681	0	0.026378897	0.00719425	417	0,***	0,***	0,***
GH39_cluster_1	0.029411765	0.15441177	0	0.81372549	0.00245098	408	0,***	0,***	0,***
CE9_cluster_3	0.017766497	0.85025381	0	0.010152284	0.12182741	394	0,***	0,***	0,***

Table 12: Cophylogenetic analysis done by methods Paco, Robison-Foulds metric and method by Nye et al.

Table showing 30 from 420 most abundant clusters investigated for cophylogeny.

Cophylogeny	*Fibrobacterota*	*Spirochaetota*	*Bacteroidota*	*Firmicutes A*	Others
PACo *p*-value < 0.001	82	55	49	26	142
0.05 > PACo *p*-value ≥ 0.001	5	3	27	5	19
PACo non-significant *p*-value	1	1	2	0	3
Nye et al. *p*-value < 0.001	85	59	56	25	153
0.05 > Nye et al. *p*-value ≥ 0.001	2	0	14	6	8
Nye et al. non-significant *p*-value	1	0	8	0	3
Robinson–Foulds *p*-value < 0.001	87	59	58	27	156
0.05 > Robinson–Foulds *p*-value ≥ 0.001	1	0	16	4	6
Robinson–Foulds non-significant *p*-value	0	0	4	0	2
Total	88	59	78	31	164

Table 13: P-values were estimated using three cophylogenetic analyses.

P-values were estimated using three cophylogenetic analyses (PACo, generalized Robinson Foulds (RF) metric, and Nye et al.'s method). TSCs were assigned to a bacterial phylum when more than 95% of sequences were assigned to this phylum. The phylum Firmicutes is split into multiple categories in the GTDB database, including *Firmicutes*_A, one category abundant in termite guts.

Cophylogeny	Fibrobacterota	Spirochaetota	Bacteroidota	Firmicutes A	Others
p < 0.001	78	51	40	21	125
0.05 > p > 0.001	5	6	30	9	27
non-significant	5	2	8	1	12
Total	88	59	78	31	164

Table 14: Simplified distribution of cophylogeny in different microbial groups

3.3.1 Vertical gene transfer (VGT)

Evidence of VGT is most clearly observed in the high degree of cophylogeny between termite hosts and their symbiotic gut microbiota, highlighting their long-term evolutionary association. For instance, phylogenetic trees of termite-specific microbial lineages, particularly those encoding CAZymes, often mirror the evolutionary history of their host termites. This alignment is indicative of coevolution and vertical inheritance, as seen in genes critical for lignocellulose degradation, such as those within the GH5 and GH45 families.

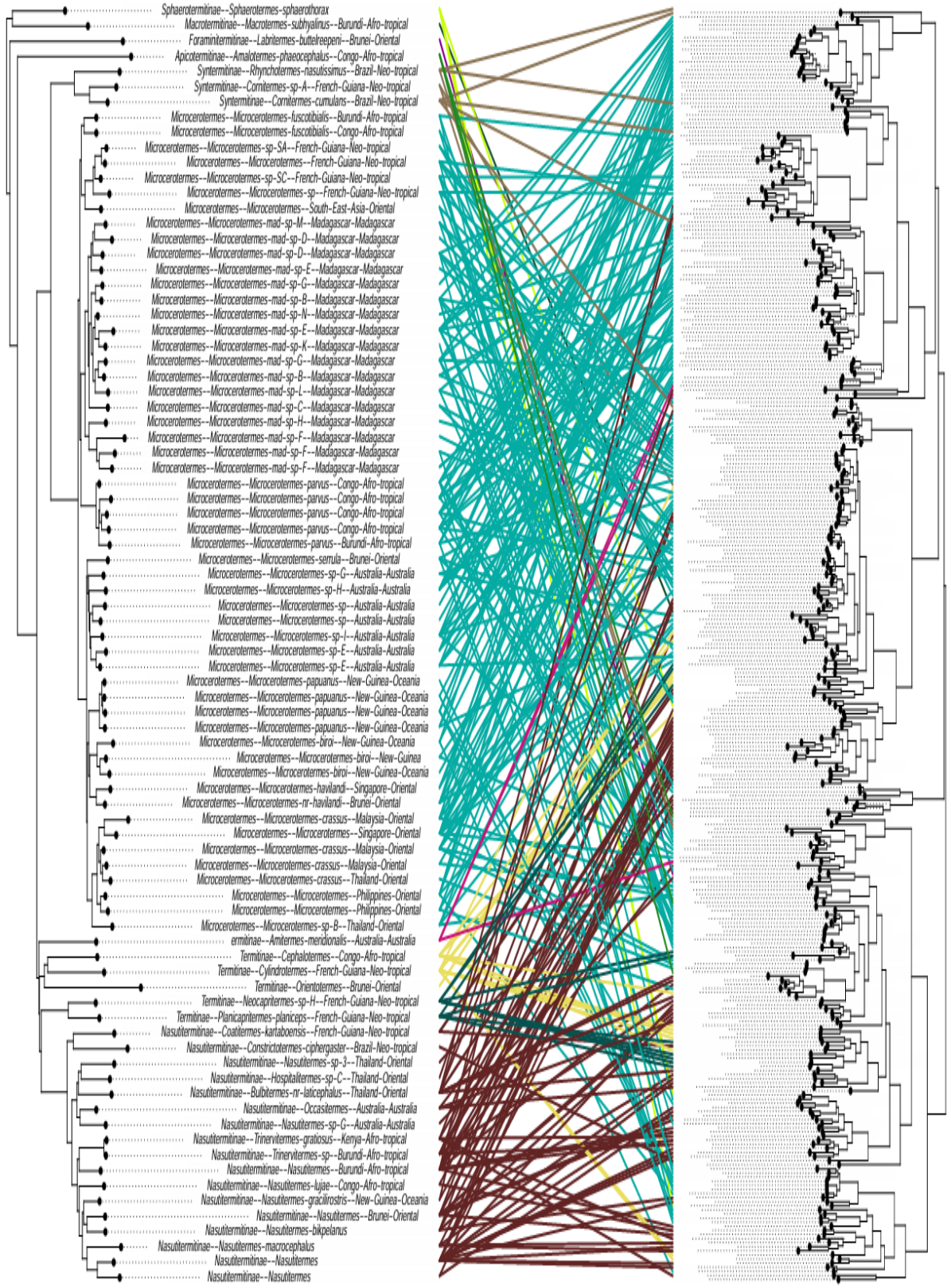


Figure 26: GH5 cluster 1 - cophylogeny analysis between CAZyme family and termite host tree

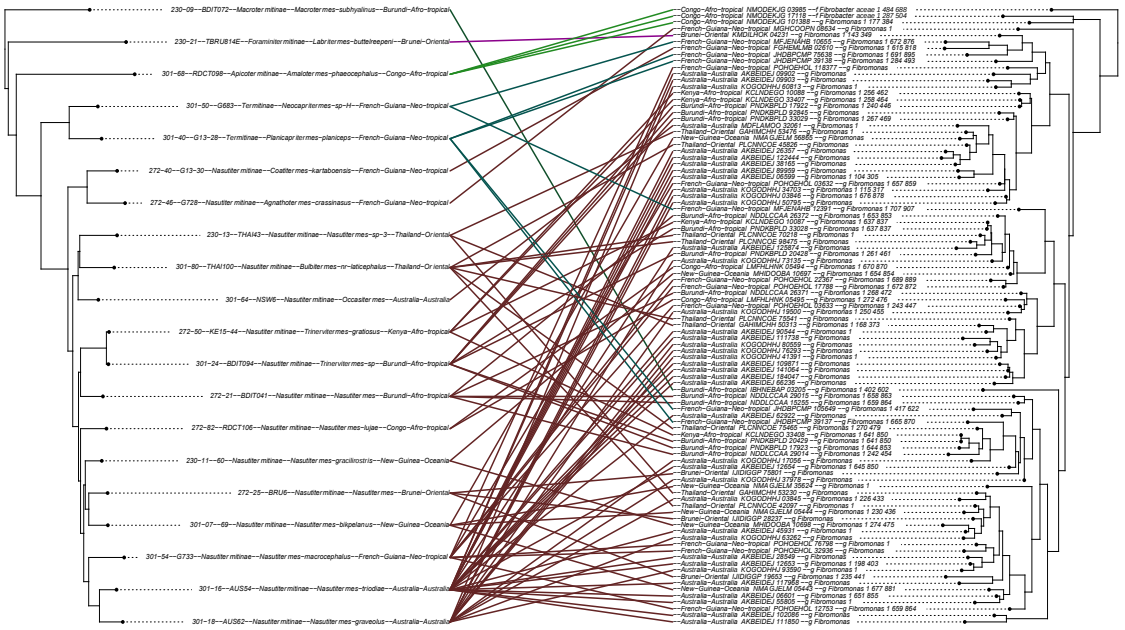


Figure 27: GH45 cluster 1 - cophylogeny analysis between CAZyme family and termite host tree

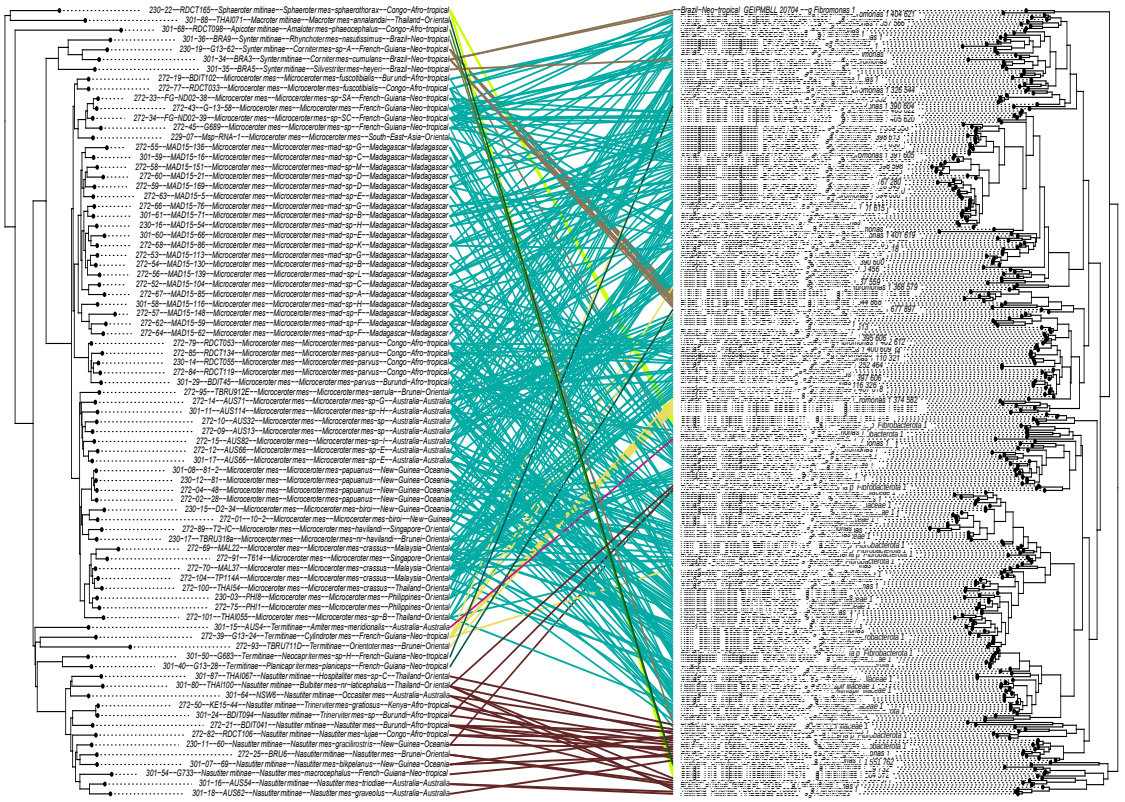


Figure 28: GH45 cluster 2 - cophylogeny analysis between CAZyme family and termite host tree

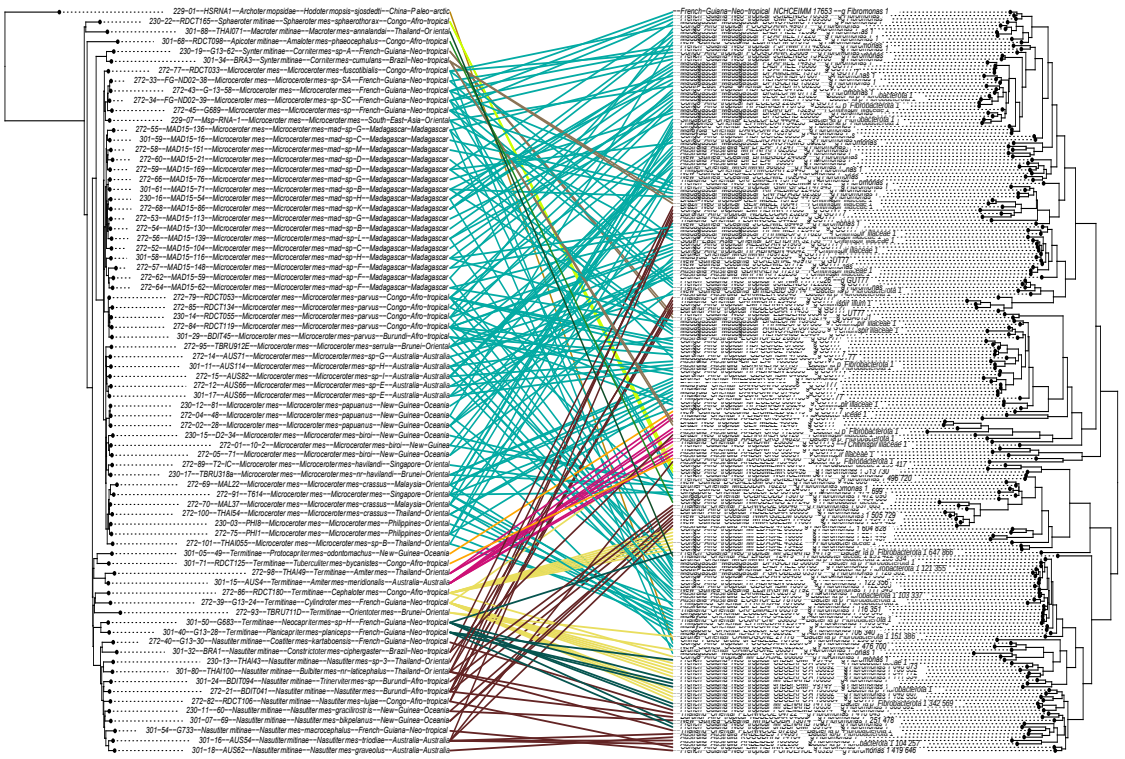


Figure 29: GH45 cluster 3 - cophylogeny analysis between CAZyme family and termite host tree

Figure 30: GH45 cluster 4 - cophylogeny analysis between CAZyme family and termite host tree

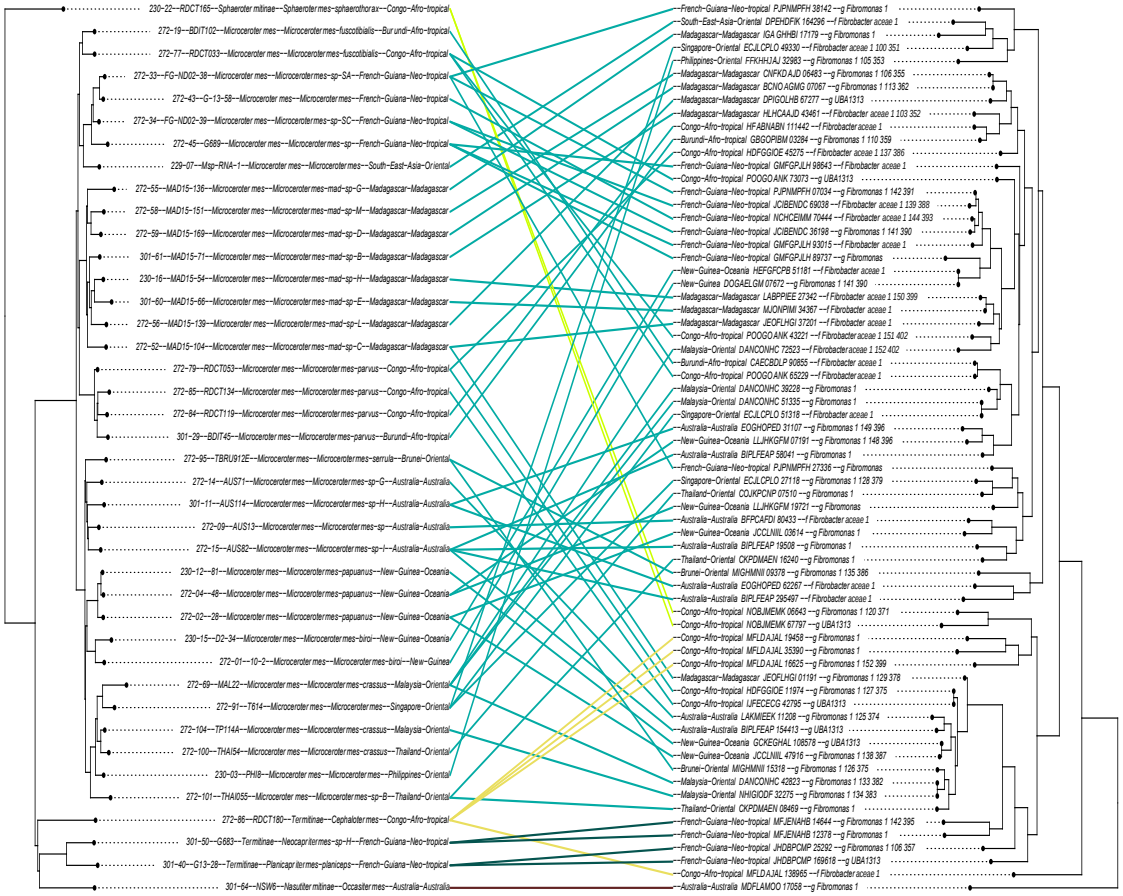


Figure 31: GH45 cluster 5 - cophylogeny analysis between CAZyme family and termite host tree

Termite-specific clusters (TSCs) of CAZymes provide additional evidence of VGT. These clusters include genes that are consistently found within termite-associated microbial communities but are absent or rare in environmental microbes (Bourguignon 2019). For example, termite-associated lineages within the bacterial phyla Spirochaetota and Firmicutes demonstrate phylogenetic patterns consistent with coevolution (Fig 7.).

In the case of basal termite lineages, such as those within the family Mastotermitidae, VGT plays a critical role in preserving ancestral microbial associations. Comparative analyses reveal that CAZyme genes in Mastotermitidae and their closest relatives, the wood-feeding cockroaches (*Cryptocercus*), share significant phylogenetic similarities (Fig 9.).

3.3.2 Horizontal gene transfer (HGT)

Evidence for HGT is derived from phylogenetic and functional analyses that demonstrate gene acquisition from external microbial sources rather than vertical inheritance alone.

Phylogenetic trees constructed for CAZyme genes, such as those encoding GH53, PLs, and CBMs, reveal incongruences when compared to the phylogenies of their bacterial hosts or termite lineages. For example, GH53 genes in termite-associated bacteria are more closely related to genes from soil-dwelling microbial lineages, indicating horizontal acquisition from environmental microbes.

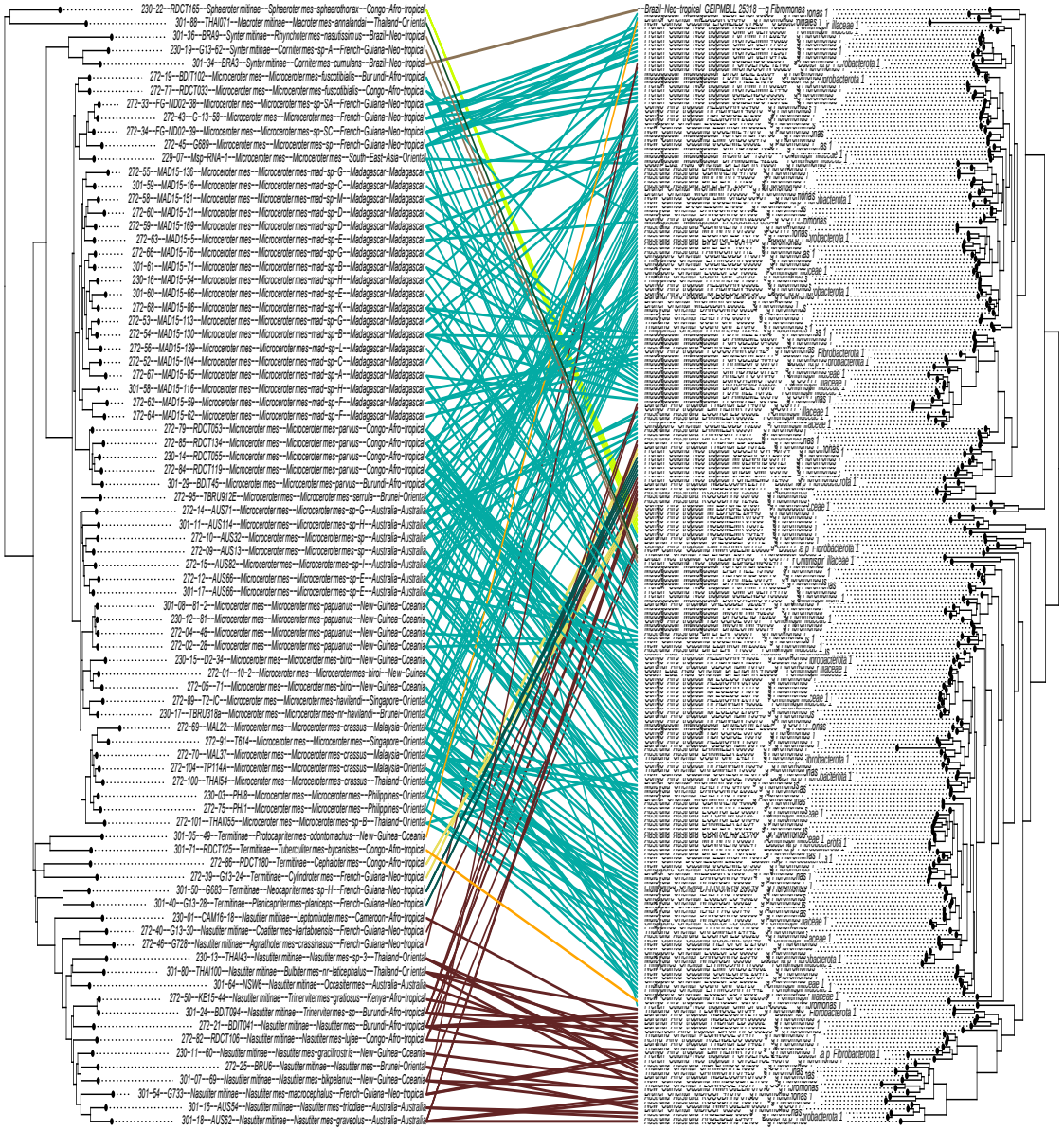


Figure 32: GH53 cluster 1 - cophylogeny analysis between CAZyme family and termite host tree

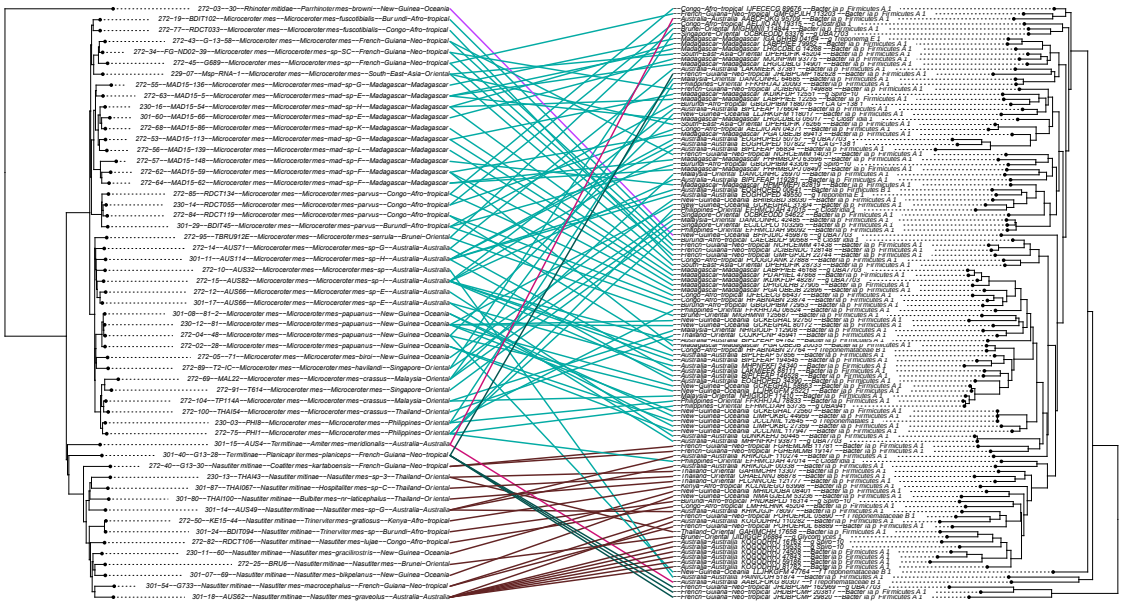


Figure 33: GH53 cluster 2 - cophylogeny analysis between CAZyme family and termite host tree

The broad taxonomic distribution of these CAZyme families provides additional support for HGT. Genes encoding GH53, PLs, and CBMs are widely distributed among bacteria found in soil and decaying plant material, suggesting that termite gut bacteria acquired these genes through interactions with external microbial communities.

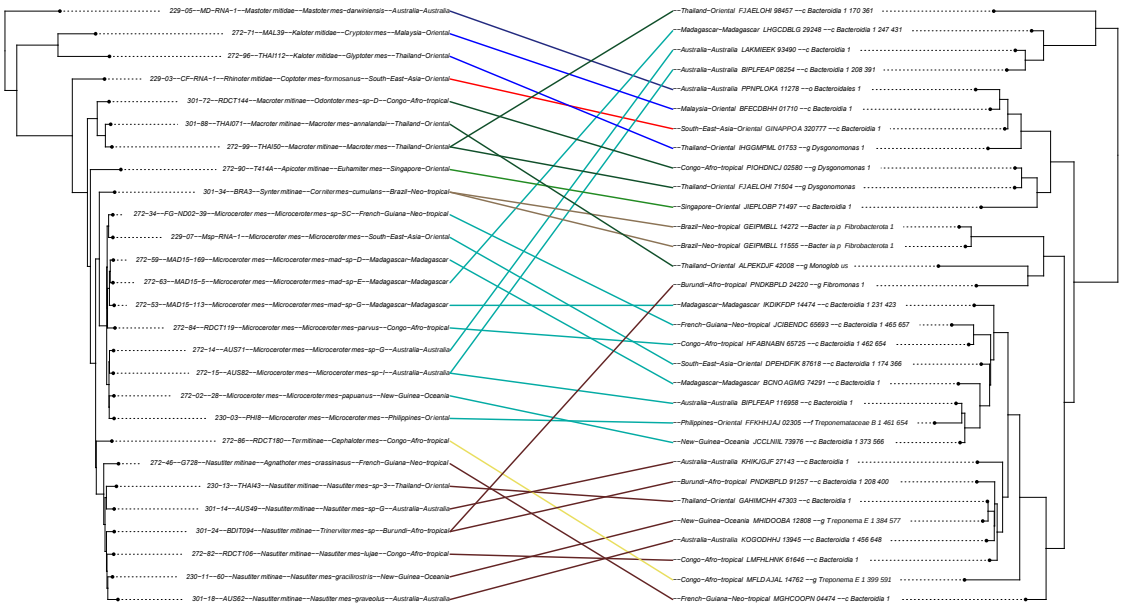


Figure 34: PL1_2 cluster 1 - cophylogeny analysis between CAZyme family and termite host tree

Termite-specific CAZyme clusters (TSCs) further illustrate the influence of HGT. Many of these clusters include sequences that display phylogenetic incongruences, indicating a mix of

termite-associated and environmentally derived microbial genes. For example, horizontally transferred genes such as PLs and CBMs, which enhance the breakdown of complex polysaccharides, have integrated into termite-specific clusters, enriching the functional diversity of the gut microbiota.

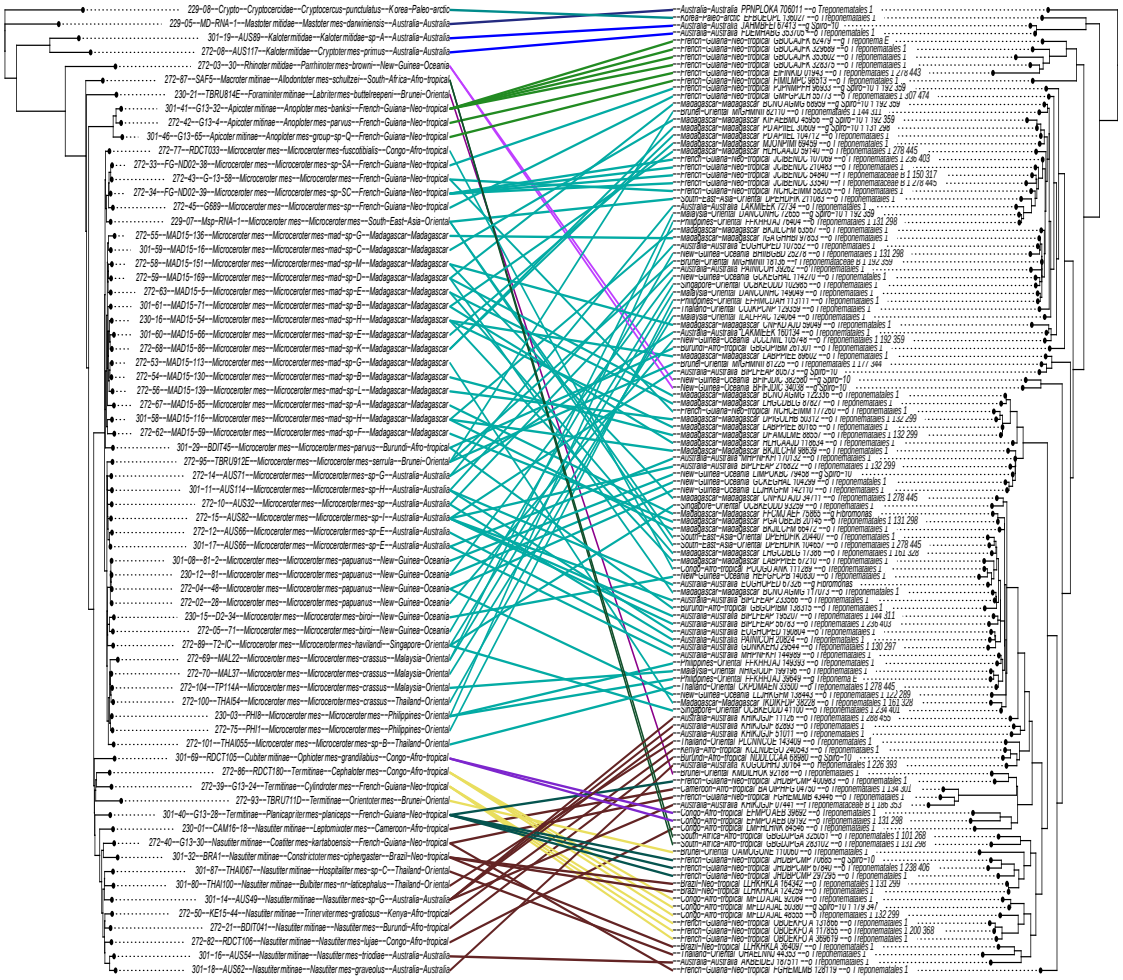


Figure 35: CBM9 cluster 1 - cophylogeny analysis between CAZyme family and termite host tree

The integration of these genes through HGT has significantly expanded the enzymatic capabilities of termite gut bacteria, enabling them to exploit diverse and nutrient-poor substrates effectively.

4. Discussion

Termites host unique gut microbial communities composed of cellulolytic bacteria, archaea and flagellates (Brune, 2014). It is well established that the gut flagellates have cospeciated with their hosts since their acquisition by the common ancestor of termites and wood-feeding cockroaches and were eventually lost in the most apical termite family, Termitidae (Ohkuma et al., 2009; Ohkuma and Brune 2011). However, there are also numerous bacterial lineages that occur ubiquitously in all termite species investigated but have never been found outside of termite guts (Mikaelyan et al. 2015, Bourguignon et al. 2018), raising the possibility that also gut bacteria have been vertically transmitted over the past 150 million years of termite evolution (Bourguignon et al. 2015; Bucek et al. 2019).

This thesis investigates evidence of cophylogeny between termites and their gut bacteria, as well as the CAZymes produced by this specific microbiota. Evidence was obtained by comparing the phylogenetic tree of termites with phylogenetic trees of gut bacteria reconstructed using ten independents, universally occurring protein-coding marker genes (Sunagawa et al., 2013). The sequences were derived from 196 termite gut metagenomes combined with sequences from the GTDB database (Parks et al., 2020). The dataset for this research comprises representatives from all termite families and all subfamilies of Termitidae. Special attention was given to the genus *Microcerotermes*, a pantropical termitid genus that emerged approximately 20 million years ago (Bourguignon et al., 2017). This genus is represented in the dataset with 30 species, some sampled multiple times, enabling an examination of both intraspecific variations and ancient divergences of the termite hosts.

The strongest cophylogenetic signals were identified within key components of the termite gut microbiota, including families such as *Ruminococcaceae* (phylum Bacillota, formerly Firmicutes) and *Breznakiellaceae* (phylum Spirochaetota). These families accounted for 16.5% and 20.0%, respectively, of the 16S rRNA gene sequences in a previous survey of 94 termite species (Dietrich et al., 2014; Mikaelyan et al., 2015; Bourguignon et al., 2018). Members of *Breznakiellaceae* are known for their fermentative metabolism and include strains capable of reductive acetogenesis (Leadbetter et al., 1999; Song et al., 2021). Their presence in the guts of cockroaches suggests an ancestral origin predating the divergence of termites and their sister group, *Cryptocercus* (Song et al., 2021; Brune et al., 2022). These findings underscore

the significance of TSCs with essential functions and a long-standing association with termites, as evidenced by their cophylogenetic signals.

Cophylogenetic signals between TSCs and their termite hosts could arise from two primary mechanisms: (i) vertical transmission of gut bacteria, leading to coevolution between symbionts and hosts; or (ii) limited horizontal transfers of gut bacteria among diverging termite species, which would result in allopatric speciation without requiring vertical transmission (de Vienne et al., 2013; Groussin et al., 2020). In cases where vertical transmission drives cophylogenetic signals, phylogenetic trees of TSCs are expected to align with termite phylogenies, resulting in bacterial lineages that are specific to particular termite clades and absent from sympatric termites outside those clades. Such termite-clade-specific lineages (TCSLs) were identified in multiple TSCs. Examples include TCSLs within the family *Breznakiellaceae*, the genus *Fibromonas* (phylum Fibrobacterota), and the genus *Adiutrix* (phylum Desulfobacterota), all of which were exclusively associated with the genus *Microcerotermes*. These TCSLs were found in *Microcerotermes* species across four continents and six biogeographic realms, indicating worldwide dispersal alongside specific gut bacteria. Notably, these lineages were absent from the guts of sympatric termites belonging to other clades.

Further examples of TCSLs were identified in less intensively sampled termite clades, such as a group of Nasutitermitinae sharing a common ancestor approximately 25 million years ago. This group, sampled across multiple continents, hosted several TCSLs from the families *Breznakiellaceae* and *Adiutrix*. These findings demonstrate a remarkable specificity between termite clades and their gut bacteria, with no evidence of horizontal bacterial transfer between sympatric termite clades. Therefore, although allopatric speciation of termites and TCSLs likely occurred, the primary mechanism maintaining these associations appears to be vertical transmission from parents to offspring, with occasional horizontal transfers among closely related species within a clade.

Host transfer events were estimated for each TSC using the maximum likelihood method in GeneRax (Morel et al., 2020). Transfer rates varied from 0.16 to 0.61 for TSCs exhibiting cophylogenetic signals. Notably, 18 TSCs showed transfer rates between 0.10 and 0.33, consistent with the rates estimated for the mitochondrial genes used to construct the termite phylogenetic tree. Since mitochondrial genes are not subject to horizontal transfer and share

an identical evolutionary history with their hosts, these transfer rates provide a baseline. The observed cophylogenetic patterns suggest minimal or no horizontal transfers in certain TSCs. Factors such as bacterial extinction, insufficient sequencing depth, or bacterial speciation within non-speciating termite hosts may obscure cophylogenetic patterns (Groussin et al., 2020). Several less speciose TSCs also displayed strong cophylogenetic patterns across extensive sections of the termite phylogenetic tree. For example, the genus *Adiutrix* was found in *Cryptocercus*, three termite families, and six subfamilies of Termitidae, with its phylogeny showing high congruence with that of termites. Similarly, the phylogenetic trees of *Rhodocyclaceae* (phylum Pseudomonadota, formerly Proteobacteria) and *Holophagaceae* (phylum Acidobacteriota) mirrored the phylogenetic trees of termites and Termitidae, respectively. These patterns provide compelling evidence of coevolution between termites and their gut bacterial symbionts, maintained through vertical transmission over tens of millions of years. These results substantiate the oldest known cophylogenetic patterns between animals and their gut bacteria, involving multiple bacterial lineages and their termite hosts over geological timescales. Some associations may even trace back to the origin of termites around 150 million years ago. This study supports previous assertions of termite-gut microbiota coevolution (Brune & Dietrich, 2015) and demonstrates the stability of vertical transmission mechanisms such as proctodeal trophallaxis, whereby termites exchange hindgut contents among nestmates (Nalepa et al., 2001).

The analysis of 101,941 CAZyme sequences in the gut metagenomes of termites and one Cryptocercus shed the light on enzymatic diversity present in termite gut and underscored the evolutionary relationship with symbiotic prokaryotes. The identification of CAZymes across a representative sampling of the termite phylogenetic tree highlights the extensive enzymatic toolkit that termites possess, facilitating their adaptation to diverse ecological niches by enabling efficient lignocellulose degradation. Analyzing prokaryotes and enzymatic families across termite species, supporting the hypothesis that the evolution of termite gut microbiota is closely linked with the dietary shift to lignocellulose (Bourguignon 2019, Arora 2022). The phylogenetic and cophylogenetic analyses conducted in this study reveal the presence of termite-specific clusters (TSCs), which are indicative of a co-evolutionary pattern between termites and their symbiotic gut microbiota. This observation is supported by the significant cophylogenetic signals detected in a majority of the TSCs, with a high level of

statistical significance ($p < 0.001$) across three different analytical methods. These results not only confirm but also extend the findings of previous studies such as those by Dietrich et al. (2014) and Bourguignon et al. (2018), who demonstrated coevolutionary relationships between termites and specific bacterial lineages within their gut microbiota. Moreover, the dominance of bacterial taxa such as *Breznakiellaceae* and *Candidatus* Fibromonas in the CAZyme sequences composing TSCs suggest the stable association of these bacterial lineages with termite guts.

Through the detailed examination of CAZyme trees, my research bring insights into the specific bacterial lineages and CAZyme clusters that have co-evolved with termites, highlighting the evolutionary background of this complex symbiotic network. The presence of termite-specific clusters, particularly those associated with Bacteroidota, elucidates the important role of microbial symbionts in shaping the dietary and ecological adaptability of termites. The example of GH57 Cluster 7 being composed of Bacteroidota is a significant finding that underscores the role of this phylum in the termite gut ecosystem. Within this cluster, sequences correspond to the genus *Candidatus* Azobacteroides, a lineage known for its symbiotic relationship with termites (ref.). This highlights the specialized nature of microbial communities within the termite gut, tailored to support the digestive needs of their hosts through specific enzymatic functions. Further investigation of CAZyme clusters also showed sequences exclusive to certain termite lineages such as *Nasutitermitinae* and *Microcerotermes*, emphasizing the specialized nature of these symbiotic relationships. This exploration of CAZyme diversity and its evolutionary implications in termite gut microbiomes underscores the complex symbiosis between termites and their microbiota. It reveals how specific CAZyme families have become integral to the survival and ecological success of termites, reflecting a shared evolutionary history of adaptation and specialization. Further, 11 of these CAZyme families are present in more than 55% of the analyzed termite gut metagenomes, pointing to a set of core enzymatic functions that are essential across a wide range of termite species. This suggests a conserved evolutionary strategy among termites to maintain a specific enzymatic toolkit for lignocellulose digestion and nutrient assimilation (ref.). 34 CAZymes were found in upward of 70% of gut metagenomes, nine of which, GH3, GH5, GH13, GH43, GH77, GT4, GT5, GT28, and GT51, were found in more than 90% of gut metagenomes, confirming that the primary CAZyme families are ubiquitous across all termite

species. Clusters GH2 Cluster 10 and GH77 Cluster 6, predominantly comprising *Spirochaetota*, and GH116 with GH8 Clusters, are exclusively formed by *Fibrobacterota*, exemplify the diverse microbial origins of CAZymes in termite guts. These clusters not only illustrate the variety of microbial life contributing to the termite's digestive process but also highlight the evolutionary depth of the termite-microbiota relationship (Bourguignon 2019). The presence of such distinct microbial communities within the termite gut points to a complex evolutionary history marked by mutualistic adaptation and specialization.

The analysis of the 420 maximum-likelihood phylogenetic trees of termite-specific bacterial clusters (TSCs), uncovering cophylogenetic signals of deep evolutionary links between termites and their gut microbiota. Notably, these trees showcased clusters of CAZymes that are specifically associated with distinct termite clades, such as Kalotermitidae and non-Termitida, highlighting the evolutionary intricacies of these symbiotic relationships. In the detailed exploration of termite-specific bacterial clusters (TSCs), my investigation focused on the unique subset of clusters that incorporate sequences from *Cryptocercus* and/or *Mastotermes*, two pivotal taxa in understanding termite evolutionary biology. This subset comprised three of the 131 identified TSCs, each exhibiting pronounced cophylogenetic signals that underscore the deep evolutionary connections between termites and their gut microbiota. The specificity of these clusters to ancient termite lineages provides a unique lens through which to view the evolutionary history of termite-microbiota symbiosis.

Another example is GH3 Cluster 4, predominantly composed of 97.3% *Spirochaetoda*, this cluster illustrates the significant role of *Spirochaetoda* in the termite gut environment, particularly in the digestion processes. The high percentage of *Spirochaetoda* within this cluster points to the evolutionary adaptation of these bacteria to the termite gut environment, reflecting a long history of co-evolution with their termite hosts. This cluster's association with sequences from both *Cryptocercus kyebangensis* and *Mastotermes darwiniensis*, representing early branching points in the termite evolutionary tree, highlights the ancient origins of this symbiotic relationship.

This research focused on important VGT which ensures the transmission of critical genes and microbial lineages across generations, preserving the metabolic capabilities necessary for termite survival in diverse ecological niches (Bourguignon et al., 2015; Buček et al., 2019). The

phylogenetic correspondence between termite lineages and their gut microbiota provides strong evidence for VGT. For example, phylogenetic trees of termite-specific microbial lineages, particularly those encoding CAZymes, often reflect the evolutionary history of their termite hosts. This correspondence is indicative of co-evolution and vertical inheritance, as seen in genes critical for lignocellulose degradation, such as those in the GH5 and GH45 families. These genes are conserved in all wood-feeding termite species and closely match the termite phylogeny, reflecting their origin in the ancestral termite and microbiota. Termite-specific clusters (TSCs) of CAZymes provide additional evidence of VGT. These clusters include genes that are consistently found within termite-associated microbial communities but are absent or rare in environmental microbes. For example, termite-associated lineages within the bacterial phyla Spirochaetota and Firmicutes demonstrate phylogenetic patterns consistent with coevolution. The deep integration of these microbial lineages into termite metabolism further supports their vertical transmission.

Also in the case of basal termite lineages, such as those within the family Mastotermitidae, VGT plays a critical role in preserving ancestral microbial associations. Comparative analyses reveal that CAZyme genes (GH5, GH45, GH9, in Mastotermitidae and their closest relative *Cryptocercus*, share significant phylogenetic similarities. This shared ancestry suggests that these genes were inherited from a common ancestor of termites and cockroaches, providing further evidence for the stability of vertically transmitted microbial lineages. Behavioral mechanisms also reinforce VGT in termites. Social interactions, such as proctodeal trophallaxis, facilitate the transfer of gut microbiota and their associated genes between colony members and across generations. This ensures the continuity of symbiotic relationships and the preservation of essential microbial lineages (Nalepa, 2011).

Horizontal gene transfer (HGT) has played a pivotal role in shaping the functional diversity of CAZyme families within termite gut microbiota and my research also investigated evidence for this transfer. Phylogenetic trees constructed for CAZyme genes, such as those encoding GH11, GH53, PLs, and CBMs, reveal incongruences when compared to the phylogenies of their bacterial hosts or termite lineages. For example, GH11 and GH53 genes in termite-associated bacteria are more closely related to genes from soil-dwelling microbial lineages, indicating horizontal acquisition from environmental microbes.

The broad taxonomic distribution of these CAZyme families provides additional support for HGT. Genes encoding GH11, GH53, PLs, and CBMs are widely distributed among bacteria found in soil and decaying plant material, suggesting that termite gut bacteria acquired these genes through interactions with external microbial communities. For instance, GH11 genes, critical for xylan degradation, are prominent in soil-feeding termites but trace their evolutionary origins to bacterial lineages outside the termite gut (Marynowska et al., 2020). Termite-specific CAZyme clusters (TSCs) further illustrate the influence of HGT. Many of these clusters include sequences that display phylogenetic incongruences, indicating a mix of termite-associated and environmentally derived microbial genes. For example, horizontally transferred genes such as PLs and CBMs, which enhance the breakdown of complex polysaccharides, have integrated into termite-specific clusters, enriching the functional diversity of the gut microbiota.

HGT-derived CAZyme families are also strongly associated with dietary transitions in termites. Soil-feeding termites exhibit a significant enrichment of CAZymes such as GH53 and PLs, which enable the digestion of humic acids and complex soil-derived carbohydrates. This contrasts with wood-feeding termites, which rely primarily on vertically inherited CAZyme clusters, such as GH5 and GH45, specialized for lignocellulose degradation. These dietary correlations highlight the adaptive role of HGT in facilitating termite evolution and ecological diversification. The microbial donors of HGT-derived genes include Spirochaetota, Firmicutes, and Bacteroidota, bacterial lineages commonly associated with soil and decaying organic matter are sources provided essential genes for cellulose, hemicellulose, and complex carbohydrate metabolism, which were subsequently incorporated into the termite gut microbiota. The integration of these genes through HGT has significantly expanded the enzymatic capabilities of termite gut bacteria, enabling them to exploit diverse and nutrient-poor substrates effectively.

In contrast, five TSCs restricted to Termitidae are suggestive of the latter mechanism, as they included more than 90% of CAZymes annotated as Spirochaetota, most of which from the *Breznakiellaceae*, a bacterial family present across the gut of most termites. Future studies are needed to determine whether the *Breznakiellaceae* populating the gut of the ancestor of Termitidaeacquired these CAZymes by horizontal transfer from bacteria not associated with termite guts (Beránková et al., 2024).

The acquisition of CAZyme genes through HGT underscores the dynamic interplay between genetic inheritance and ecological adaptation in termite gut microbiota. While vertical inheritance preserves core enzymatic functions, HGT introduces novel capabilities, enhancing the microbiota's functional repertoire and facilitating termite survival in varying ecological niches. These findings highlight the importance of HGT in the evolutionary success of termites and their symbiotic microbial communities. Overall, this study highlights how termite evolution and ecological adaptation have been deeply influenced by the long-term and dynamic symbiosis with their gut microbiota.

5. Conclusion

My research has uncovered the oldest known cophylogenetic patterns between animals and their symbiotic bacteria, encompassing multiple bacterial lineages and their corresponding termite hosts. These patterns span tens of millions of years and may date back to the emergence of termites approximately 150 million years ago. The findings reinforce earlier hypotheses of coevolution between termites and their gut microbiota and provide direct evidence that proctodeal trophallaxis—a social behaviour involving the exchange of hindgut contents among nestmates—serves as a stable mechanism for symbiont transmission across geological timescales.

This symbiotic relationship extends beyond mere coexistence, reflecting a dynamic interplay in which termite gut bacteria have not only adapted to their hosts but have also played a significant role in shaping termite dietary specialization and evolutionary success. The identification of termite clade-specific CAZyme clusters restricted to certain lineages suggests a complex co-evolutionary mechanism that contributes to the metabolic diversity and ecological adaptability of modern termites. These unique microbial assemblages further support the hypothesis that termite evolution has been tightly linked with microbial diversification, enabling expansion into new ecological niches through the acquisition of novel metabolic functions.

The broader implications of this study extend beyond termites, offering insights into the role of microbial symbioses in the adaptive radiation of host organisms. By elucidating the evolutionary dynamics of termite gut microbiomes, this research highlights how microbial partnerships have shaped the evolutionary trajectories of multicellular life. Such findings have relevance for evolutionary biology, microbiology, and ecology, presenting a compelling narrative of symbiotic evolution that invites further exploration of host-microbe interactions.

Moreover, the presence of CAZymes in the gut metagenomes of phylogenetically distant termite species points to the ancient origin of these symbiotic relationships. Combined with evidence for both vertical and horizontal transfer mechanisms, the findings underscore the dynamic nature of termite gut microbiomes. The acquisition and diversification of CAZymes

appear to have been pivotal in enabling termites to process lignocellulosic diets, contributing to their ecological success and diversification.

In conclusion, this study not only substantiates existing theories on termite–microbiota coevolution but also emphasizes the central role of CAZymes in the evolution of termite dietary specialization. By clarifying the diversity, distribution, and evolutionary history of CAZyme sequences in termite guts, this work advances understanding of the symbiotic interactions that underpin termite ecology. Future research—particularly through broader taxonomic sampling and functional characterization of CAZymes—will be essential to fully unravel the complexity and evolutionary importance of termite gut microbiomes.

7. References

Abdul Rahman, N., Parks, D.H., Willner, D.L., Engelbrektson, A.L., Goffredi, S.K., Warnecke, F., Scheffrahn, R.H., 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. https://doi.org/10.1186/s40168-015-0067-8

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Armendáriz-Ruiz, M., Rodríguez-González, J.A., Camacho-Ruíz, R.M., Mateos-Díaz, J.C., 2018. Carbohydrate Esterases: An Overview. Methods Mol. Biol. Clifton NJ 1835, 39–68. https://doi.org/10.1007/978-1-4939-8672-9_2

Armenta, S., Moreno-Mendieta, S., Sánchez-Cuapio, Z., Sánchez, S., Rodríguez-Sanoja, R., 2017. Advances in molecular engineering of carbohydrate-binding modules. Proteins Struct. Funct. Bioinforma. 85, 1602–1617. https://doi.org/10.1002/prot.25327

Arora, J., Buček, A., Hellemans, S., Beránková, T., Romero Arias, J., Fisher, B., Clitheroe, C.-L., Brune, A., Kinjo, Y., Šobotník, J., Bourguignon, T., 2023. Evidence of cospeciation between termites and their gut bacteria on a geological time scale. Proc. R. Soc. B 290. https://doi.org/10.1098/rspb.2023.0619

Ashton, L.A., Griffiths, H.M., Parr, C.L., Evans, T.A., Didham, R.K., Hasan, F., Teh, Y.A., Tin, H.S., Vairappan, C.S., Eggleton, P., 2019. Termites mitigate the effects of drought in tropical rainforest. Science 363, 174–177. https://doi.org/10.1126/science.aau9565

Balbuena, J.A., Míguez-Lozano, R., Blasco-Costa, I., 2013. PACo: A Novel Procrustes Application to Cophylogenetic Analysis. PLOS ONE 8, e61048. https://doi.org/10.1371/journal.pone.0061048

Bartz, S.H., 1979. Evolution of eusociality in termites. Proc. Natl. Acad. Sci. U. S. A. 76, 5764–5768. https://doi.org/10.1073/pnas.76.11.5764

Beránková, T., Arora, J., Romero Arias, J., Buček, A., Tokuda, G., Šobotník, J., Hellemans, S., Bourguignon, T., 2024. Termites and subsocial roaches inherited many bacterial-borne carbohydrate-active enzymes (CAZymes) from their common ancestor. Commun. Biol. 7, 1–9. https://doi.org/10.1038/s42003-024-07146-w

Bignell, D.E., Roisin, Y., Lo, N. (Eds.), 2011. Biology of Termites: a Modern Synthesis. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-90-481-3977-4

Boraston, A.B., Bolam, D.N., Gilbert, H.J., Davies, G.J., 2004. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem. J. 382, 769–781. https://doi.org/10.1042/BJ20040892

Bordereau, C., Pasteels, J.M., 2011. Pheromones and Chemical Ecology of Dispersal and Foraging in Termites, in: Bignell, D.E., Roisin, Y., Lo, N. (Eds.), Biology of Termites: A Modern Synthesis. Springer Netherlands, Dordrecht, pp. 279–320. https://doi.org/10.1007/978-90-481-3977-4_11

Bourguignon, T., Lo, N., Cameron, S.L., Šobotník, J., Hayashi, Y., Shigenobu, S., Watanabe, D., Roisin, Y., Miura, T., Evans, T.A., 2015. The Evolutionary History of Termites as Inferred from 66 Mitochondrial Genomes. Mol. Biol. Evol. 32, 406–421. https://doi.org/10.1093/molbev/msu308

Bourguignon, T., Lo, N., Dietrich, C., Šobotník, J., Sidek, S., Roisin, Y., Brune, A., Evans, T.A., 2018. Rampant Host Switching Shaped the Termite Gut Microbiome. Curr. Biol. CB 28, 649-654.e2. https://doi.org/10.1016/j.cub.2018.01.035

Brady, S.G., Fisher, B.L., Schultz, T.R., Ward, P.S., 2014. The rise of army ants and their relatives: diversification of specialized predatory doryline ants. BMC Evol. Biol. 14, 93. https://doi.org/10.1186/1471-2148-14-93

Brune, A., 2014. Symbiotic digestion of lignocellulose in termite guts. Nat. Rev. Microbiol. 12, 168–180. https://doi.org/10.1038/nrmicro3182

Brune, A., Song, Y., Oren, A., Paster, B.J., 2022. A new family for ‘termite gut treponemes’: description of Breznakiellaceae fam. nov., Gracilinema caldarium gen. nov., comb.

nov., Leadbettera azotonutricia gen. nov., comb. nov., Helmutkoenigia isoptericolens gen. nov., comb. nov., and Zuelzera stenostrepta gen. nov., comb. nov., and proposal of Rectinemataceae fam. nov. Int. J. Syst. Evol. Microbiol. 72, 005439. https://doi.org/10.1099/ijsem.0.005439

Buček, A., Menglin, W., Šobotník, J., Hellemans, S., Sillam-Dussès, D., Mizumoto, N., Stiblík, P., Clitheroe, C.-L., Lu, T., Gonzalez Plaza, J.J., Mohagan, A., Rafanomezantsoa, J.-J., Fisher, B., Engel, M., Roisin, Y., Evans, T., Scheffrahn, R., Bourguignon, T., 2022. Molecular Phylogeny Reveals the Past Transoceanic Voyages of Drywood Termites (Isoptera, Kalotermitidae). Mol. Biol. Evol. 39. https://doi.org/10.1093/molbev/msac093

Buček, A., Šobotník, J., He, S., Shi, M., Mcmahon, D., Holmes, E., Roisin, Y., Lo, N., Bourguignon, T., 2019. Evolution of Termite Symbiosis Informed by Transcriptome-Based Phylogenies. Curr. Biol. 29. https://doi.org/10.1016/j.cub.2019.08.076

Cantarel, B.L., Coutinho, P.M., 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. https://doi.org/10.1093/nar/gkn663

Cleveland, L.R., 1925. The Effects of Oxygenation and Starvation on the Symbiosis between the Termite, Termopsis, and Its Intestinal Flagellates. Biol. Bull. 48, 309–326. https://doi.org/10.2307/1536599

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 Evol. Biol. 10, 210. https://doi.org/10.1186/1471-2148-10-210

Dahlsjö, C.A.L., Parr, C.L., Malhi, Y., Meir, P., Chevarria, O.V.C., Eggleton, P., 2014. Termites promote soil carbon and nitrogen depletion: Results from an in situ macrofauna exclusion experiment, Peru. Soil Biol. Biochem. 77, 109–111. https://doi.org/10.1016/j.soilbio.2014.05.033

Davies, G., Henrissat, B., 1995. Structures and mechanisms of glycosyl hydrolases. Structure 3, 853–859. https://doi.org/10.1016/S0969-2126(01)00220-9

DONOVAN, S.E., JONES, D.T., SANDS, W.A., EGGLETON, P., 2000. Morphological phylogenetics of termites (Isoptera). Biol. J. Linn. Soc. 70, 467–513. https://doi.org/10.1111/j.1095-8312.2000.tb01235.x

Emerson, A.E., Emerson, A.E., Schmidt, K.P., 1955. Geographical origins and dispersions of termite genera. Chicago Natural History Museum, [Chicago]. https://doi.org/10.5962/bhl.title.2783

Engel, M.S., Barden, P., Riccio, M.L., Grimaldi, D.A., 2016. Morphologically Specialized Termite Castes and Advanced Sociality in the Early Cretaceous. Curr. Biol. 26, 522–530. https://doi.org/10.1016/j.cub.2015.12.061

Engel, P., Moran, N.A., 2013. The gut microbiota of insects - diversity in structure and function. FEMS Microbiol. Rev. 37, 699–735. https://doi.org/10.1111/1574-6976.12025

Gorvitovskaia, A., Holmes, S.P., Huse, S.M., 2016. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome 4, 15. https://doi.org/10.1186/s40168-016-0160-7

Groussin, M., Mazel, F., Alm, E.J., 2020. Co-evolution and Co-speciation of Host-Gut Bacteria Systems. Cell Host Microbe 28, 12–22. https://doi.org/10.1016/j.chom.2020.06.013

Hartman, M.C.T., Jiang, S., Rush, J.S., Waechter, C.J., Coward, J.K., 2007. Glycosyltransferase Mechanisms: Impact of a 5-Fluoro Substituent in Acceptor and Donor Substrates on Catalysis. Biochemistry 46, 11630–11638. https://doi.org/10.1021/bi700863s

Hellemans, S., Rocha, M.M., Wang, M., Romero Arias, J., Aanen, D.K., Bagnères, A.-G., Buček, A., Carrijo, T.F., Chouvenc, T., Cuezzo, C., Constantini, J.P., Constantino, R., Dedeine, F., Deligne, J., Eggleton, P., Evans, T.A., Hanus, R., Harrison, M.C., Harry, M., Josens, G., Jouault, C., Kalleshwaraswamy, C.M., Kaymak, E., Korb, J., Lee, C.-Y., Legendre, F., Li, H.-F., Lo, N., Lu, T., Matsuura, K., Maekawa, K., McMahon, D.P., Mizumoto, N., Oliveira, D.E., Poulsen, M., Sillam-Dussès, D., Su, N.-Y., Tokuda, G., Vargo, E.L., Ware,

J.L., Šobotník, J., Scheffrahn, R.H., Cancello, E., Roisin, Y., Engel, M.S., Bourguignon, T., 2024. Genomic data provide insights into the classification of extant termites. Nat. Commun. 15, 6724. https://doi.org/10.1038/s41467-024-51028-y

Hellemans, S., Wang, M., Hasegawa, N., Šobotník, J., Scheffrahn, R.H., Bourguignon, T., 2022. Using ultraconserved elements to reconstruct the termite tree of life. Mol. Phylogenet. Evol. 173. https://doi.org/10.1016/j.ympev.2022.107520

Henrissat, B., Davies, G., 1997. Structural and sequence-based classification of glycoside hydrolases. Curr. Opin. Struct. Biol. 7, 637–644. https://doi.org/10.1016/S0959-440X(97)80072-3

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S., 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35, 518–522. https://doi.org/10.1093/molbev/msx281

Holt, B.G., Lessard, J.-P., Borregaard, M.K., Fritz, S.A., Araújo, M.B., Dimitrov, D., Fabre, P.-H., Graham, C.H., Graves, G.R., Jønsson, K.A., Nogués-Bravo, D., Wang, Z., Whittaker, R.J., Fjeldså, J., Rahbek, C., 2013. An Update of Wallace's Zoogeographic Regions of the World. Science 339, 74–78. https://doi.org/10.1126/science.1228282

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. https://doi.org/10.1128/AEM.71.11.6590-6599.2005

Inward, D., Beccaloni, G., Eggleton, P., 2007. Death of an order: a comprehensive molecular phylogenetic study confirms that termites are eusocial cockroaches. Biol. Lett. 3, 331–335. https://doi.org/10.1098/rsbl.2007.0102

Jouquet, P., Traoré, S., Choosai, C., Hartmann, C., Bignell, D., 2011. Influence of termites on ecosystem functioning. Ecosystem services provided by termites. Eur. J. Soil Biol. 47, 215–222. https://doi.org/10.1016/j.ejsobi.2011.05.005

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010

Köhler, T., Dietrich, C., Scheffrahn, R.H., 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. https://doi.org/10.1128/AEM.00683-12

Korsa, G., Beyene, A., Ayele, A., 2023. Bacterial diversity from soil-feeding termite gut and their potential application. Ann. Microbiol. 73, 38. https://doi.org/10.1186/s13213-023-01741-8

Krishna, K., Grimaldi, D.A., Krishna, V., Engel, M.S., 2013. Treatise on the Isoptera of the World: Introduction. Bull. Am. Mus. Nat. Hist. 2013, 1–200. https://doi.org/10.1206/377.1

Lacy, R.C., 1980. The Evolution of Eusociality in Termites: A Haplodiploid Analogy? Am. Nat. https://doi.org/10.1086/283638

Lairson, L.L., Henrissat, B., Davies, G.J., Withers, S.G., 2008. Glycosyltransferases: Structures, Functions, and Mechanisms. Annu. Rev. Biochem. 77, 521–555. https://doi.org/10.1146/annurev.biochem.76.061005.092322

Leadbetter, J.R., Schmidt, T.M., Graber, J.R., Breznak, J.A., 1999. Acetogenesis from H2 Plus CO2 by Spirochetes from Termite Guts. Science 283, 686–689. https://doi.org/10.1126/science.283.5402.686

Leal, I.R., Oliveira, P.S., 1995. Behavioral ecology of the neotropical termite-hunting ant Pachycondyla (= Termitopone) marginata: colony founding, group-raiding and migratory patterns. Behav. Ecol. Sociobiol. 37, 373–383. https://doi.org/10.1007/BF00170584

Levasseur, A., Drula, E., Lombard, V., Coutinho, P.M., Henrissat, B., 2013. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol. Biofuels 6, 41. https://doi.org/10.1186/1754-6834-6-41

Liu, N., Li, H., Chevrette, M.G., Zhang, L., Cao, L., Zhou, H., Zhou, X., Zhou, Z., Pope, P.B., Currie, C.R., Huang, Y., Wang, Q., 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. https://doi.org/10.1038/s41396-018-0255-1

Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P.M., Henrissat, B., 2010a. A hierarchical classification of polysaccharide lyases for glycogenomics. Biochem. J. 432, 437–444. https://doi.org/10.1042/BJ20101185

Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P.M., Henrissat, B., 2010b. A hierarchical classification of polysaccharide lyases for glycogenomics. Biochem. J. 432, 437–444. https://doi.org/10.1042/BJ20101185

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B., 2014a. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 42, D490-495. https://doi.org/10.1093/nar/gkt1178

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B., 2014b. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 42, D490–D495. https://doi.org/10.1093/nar/gkt1178

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B., 2014c. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 42, D490–D495. https://doi.org/10.1093/nar/gkt1178

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. https://doi.org/10.1186/s40168-020-00872-3

McLean, B.W., Boraston, A.B., Brouwer, D., Sanaie, N., Fyfe, C.A., Warren, R.A.J., Kilburn, D.G., Haynes, C.A., 2002. Carbohydrate-binding Modules Recognize Fine Substructures of Cellulose *. J. Biol. Chem. 277, 50245–50254. https://doi.org/10.1074/jbc.M204433200

Mering, C. von, Hugenholtz, P., Raes, J., Tringe, S.G., Doerks, T., Jensen, L.J., Ward, N., Bork, P., 2007. Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments. Science. https://doi.org/10.1126/science.1133420

Minh, B.Q., Nguyen, M.A.T., von Haeseler, A., 2013. Ultrafast Approximation for Phylogenetic Bootstrap. Mol. Biol. Evol. 30, 1188–1195. https://doi.org/10.1093/molbev/mst024

Miura, T., Maekawa, K., Kitade, O., Abe, T., Matsumoto, T., 1998. Phylogenetic Relationships among Subfamilies in Higher Termites (Isoptera: Termitidae) Based on Mitochondrial Coii Gene Sequences. Ann. Entomol. Soc. Am. 91, 515–523. https://doi.org/10.1093/aesa/91.5.515

Morel, B., Kozlov, A.M., Stamatakis, A., Szöllősi, G.J., 2020. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. Mol. Biol. Evol. 37, 2763–2774. https://doi.org/10.1093/molbev/msaa141

Nalepa, C., 1994. Nourishment and the Origin of Termite Eusociality, in: Nourishment and Evolution in Insect Societies. pp. 57–104.

Nalepa, C.A., 2017. What Kills the Hindgut Flagellates of Lower Termites during the Host Molting Cycle? Microorganisms 5, 82. https://doi.org/10.3390/microorganisms5040082

Nalepa, C.A., 2011. Altricial Development in Wood-Feeding Cockroaches: The Key Antecedent of Termite Eusociality, in: Bignell, D.E., Roisin, Y., Lo, N. (Eds.), Biology of Termites: A Modern Synthesis. Springer Netherlands, Dordrecht, pp. 69–95. https://doi.org/10.1007/978-90-481-3977-4_4

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300

Nye, T.M.W., Liò, P., Gilks, W.R., 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics 22, 117–119. https://doi.org/10.1093/bioinformatics/bti720

Oksanen et al., 2014. 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. http://CRAN.Rproject.org/package=vegan.

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., Hugenholtz, P., 2020a. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat. Biotechnol. 38, 1079–1086. https://doi.org/10.1038/s41587-020-0501-8

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., Hugenholtz, P., 2020b. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat. Biotechnol. 38, 1079–1086. https://doi.org/10.1038/s41587-020-0501-8

Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., Hugenholtz, P., 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 50, D785–D794. https://doi.org/10.1093/nar/gkab776

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. https://doi.org/10.1111/1755-0998.13063

Pinard, D., Mizrachi, E., Hefer, C.A., Kersting, A.R., Joubert, F., Douglas, C.J., Mansfield, S.D., Myburg, A.A., 2015. Comparative analysis of plant carbohydrate active enZymes and their role in xylogenesis. BMC Genomics 16, 402. https://doi.org/10.1186/s12864-015-1571-8

Rust, M.K., Su, N.-Y., 2012. Managing social insects of urban importance. Annu. Rev. Entomol. 57, 355–375. https://doi.org/10.1146/annurev-ento-120710-100634

Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE 11, e0163962. https://doi.org/10.1371/journal.pone.0163962

Shi, Y., Huang, Z., Han, S., Fan, S., Yang, H., 2015. Phylogenetic diversity of Archaea in the intestinal tract of termites from different lineages. J. Basic Microbiol. 55, 1021–1028. https://doi.org/10.1002/jobm.201400678

Smith, M.R., 2020. Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. Bioinformatics 36, 5007–5013. https://doi.org/10.1093/bioinformatics/btaa614

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. https://doi.org/10.1111/1462-2920.15600

Su, N.-Y., Scheffrahn, R.H., 1990. Comparison of Eleven Soil Termiticides Against the Formosan Subterranean Termite and Eastern Subterranean Termite (Isoptera: Rhinotermitidae). J. Econ. Entomol. 83, 1918–1924. https://doi.org/10.1093/jee/83.5.1918

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. https://doi.org/10.1093/nar/gkl315

Taib, N., Megrian, D., Witwinowski, J., Adam, P., Poppleton, D., Borrel, G., Beloin, C., Gribaldo, S., 2020. Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. Nat. Ecol. Evol. 4, 1661–1672. https://doi.org/10.1038/s41559-020-01299-7

Terrapon, N., Li, C., Robertson, H.M., Ji, L., Meng, X., Booth, W., Chen, Z., Childers, C.P., Glastad, K.M., Gokhale, K., Gowin, J., Gronenberg, W., Hermansen, R.A., Hu, H., Hunt,

B.G., Huylmans, A.K., Khalil, S.M.S., Mitchell, R.D., Munoz-Torres, M.C., Mustard, J.A., Pan, H., Reese, J.T., Scharf, M.E., Sun, F., Vogel, H., Xiao, J., Yang, W., Yang, Zhikai, Yang, Zuoquan, Zhou, J., Zhu, J., Brent, C.S., Elsik, C.G., Goodisman, M.A.D., Liberles, D.A., Roe, R.M., Vargo, E.L., Vilcinskas, A., Wang, J., Bornberg-Bauer, E., Korb, J., Zhang, G., Liebig, J., 2014. Molecular traces of alternative social organization in a termite genome. Nat. Commun. 5, 3636. https://doi.org/10.1038/ncomms4636

Thorne, B.L., 1997. Evolution of Eusociality in Termites. Annu. Rev. Ecol. Syst. 28, 27–54. https://doi.org/10.1146/annurev.ecolsys.28.1.27

Thorne, B.L., Grimaldi, D.A., Krishna, K., 2000. Early Fossil History of the Termites, in: Abe, T., Bignell, D.E., Higashi, M. (Eds.), Termites: Evolution, Sociality, Symbioses, Ecology. Springer Netherlands, Dordrecht, pp. 77–93. https://doi.org/10.1007/978-94-017-3223-9_4

Tuma, J., Eggleton, P., Fayle, T.M., 2020. Ant-termite interactions: an important but under-explored ecological linkage. Biol. Rev. 95, 555–572. https://doi.org/10.1111/brv.12577

Vienne, D.M. de, Refrégier, G., López-Villavicencio, M., Tellier, A., Hood, M.E., Giraud, T., 2013. Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. New Phytol. 198, 347–385. https://doi.org/10.1111/nph.12150

Wang, M., Hellemans, S., Šobotník, J., Arora, J., Buček, A., Sillam-Dussès, D., Clitheroe, C., Lu, T., Lo, N., Engel, M.S., Roisin, Y., Evans, T.A., Bourguignon, T., 2022. Phylogeny, biogeography and classification of Teletisoptera (Blattaria: Isoptera). Syst. Entomol. 47, 581–590. https://doi.org/10.1111/syen.12548

Wardman, J.F., Bains, R.K., Rahfeld, P., Withers, S.G., 2022. Carbohydrate-active enzymes (CAZymes) in the gut microbiome. Nat. Rev. Microbiol. 1–15. https://doi.org/10.1038/s41579-022-00712-1

Wilson, E.O., 1971. The Insect Societies, Illustrated edition. ed. Belknap Press: An Imprint of Harvard University Press, Cambridge, Mass.

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P.K., Xu, Y., Yin, Y., 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 46, W95–W101. https://doi.org/10.1093/nar/gky418

8. Supplementary

8.1 Supplementary tables

Supplementary table 1 list of used termite samples

Sample ID	Run Number	Family	Subfamily	Species	biogeographic origin	diet	Metagenome accession number on MGRAST
Cryp	**229-08**	Cryptocercidae		*Cryptocercus kyebangensis*	Paleo-arctic	wood	mgm4955158.3
MD_RNA_1	**229-05**	Mastotermitidae		*Mastotermes darwiniensis*	Australia	wood	mgm4955159.3
HSRNA1	**229-01**	Hodotermopsidae		*Hodotermopsis sjostedti*	Paleo-arctic	wood	mgm4953835.3
US17	**301-91**	Archotermopsidae		*Zootermopsis nevadensis*	Neo-arctic	wood	mgm4815525.3
SA16-13	**301-75**	Hodotermitidae		*Hodotermes mossambicus*	Neo-tropical	grass	mgm4815519.3
POROTERMES52	**230-10**	Stolotermitidae		*Porotermes planiceps*	Neo-tropical	wood	mgm4782062.3
PORO-CHILI	**272-76**	Stolotermitidae		*Porotermes quadricollis*	Neo-tropical	wood	mgm4813750.3
AUST14-12	**272-17**	Stolotermitidae		*Stolotermes victoriensis*	Australia	wood	mgm4812123.3
KE15-30	**272-49**	Kalotermitidae		*Bifiditermes* sp.	Afro-tropical	wood	mgm4813747.3
MAD15-2	**272-61**	Kalotermitidae		*Bifiditermes* sp.nr.*madagascariensis*	Madagascar	wood	mgm4814055.3
MAL39	**272-71**	Kalotermitidae		*Cryptotermes domesticus*	Oriental	wood	mgm4813744.3
AUS110	**272-06**	Kalotermitidae		*Cryptotermes* sp. 1	Australia	wood	mgm4812116.3
AUS117	**272-08**	Kalotermitidae		*Cryptotermes* sp. 1	Australia	wood	mgm4812128.3
H2	**301-55**	Kalotermitidae		*Glyptotermes* sp.	Paleo-arctic	wood	mgm4815527.3
AUS109	**230-02**	Kalotermitidae		*Glyptotermes* sp. 1	Australia	wood	mgm4782046.3
THAI114	**272-97**	Kalotermitidae		*Glyptotermes* sp. 22	Oriental	wood	mgm4814057.3
THAI31	**230-24**	Kalotermitidae		*Glyptotermes* sp. 3	Oriental	wood	mgm4782053.3
THAI112	**272-96**	Kalotermitidae		*Glyptotermes* sp. 6	Oriental	wood	mgm4814077.3
AUS111	**272-07**	Kalotermitidae		*Incisitermes* nr. *barretti*	Australia	wood	mgm4812129.3
US10	**272-105**	Kalotermitidae		*Incisitermes snyderi*	Neo-arctic	wood	mgm4839824.3
AUS89	**301-19**	Kalotermitidae		*Kalotermes* sp.	Australia	wood	mgm4821337.3
AUS102	**301-10**	Kalotermitidae		*Neotermes insularis cf. malandensis*	Australia	wood	mgm4821351.3
AUS91	**301-20**	Kalotermitidae		*Neotermes insularis cf. malandensis*	Australia	wood	mgm4821366.3
SING74	**230-18**	Kalotermitidae		*Neotermes* sp. 8	Oriental	wood	mgm4782057.3
G678_2	**301-49**	Kalotermitidae		*Rugitermes* sp. A	Neo-tropical	wood	mgm4821371.3
M16	**272-51**	Kalotermitidae		*Tauritermes* sp.	Neo-tropical	wood	mgm4814081.3
Chi15_131	**272-31**	Stylotermitidae		*Stylotermes* sp.	Paleo-arctic	wood	mgm4814076.3
CF_RNA_1	**229-03**	Heterotermitidae		*Coptotermes formosanus*	Oriental	wood	mgm4953834.3
RDCT185	**230-04**	Heterotermitidae		*Coptotermes* sp.	Afro-tropical	wood	mgm4782058.3
G13-107	**230-26**	Heterotermitidae		*Coptotermes testaceus*	Neo-tropical	wood	mgm4782063.3
NG87	**301-09**	Heterotermitidae		*Coptotermes elisae*	Oceania	wood	mgm4821355.3
AUS47	**272-11**	Heterotermitidae		*Heterotermes vagus*	Australia	wood	mgm4812112.3
AUS88	**272-16**	Heterotermitidae		*Heterotermes cf. paradoxus*	Australia	wood	mgm4812131.3
TBRU8.25	**272-94**	Heterotermitidae		*Heterotermes tenuior*	Oriental	wood	mgm4814070.3
AUS121	**301-12**	Heterotermitidae		*Heterotermes cf. paradoxus*	Australia	wood	mgm4821359.3
THAI98	**301-90**	Heterotermitidae		*Reticulitermes* sp. A	Oriental	wood	mgm4815493.3
US1	**301-92**	Heterotermitidae		*Reticulitermes nelsonae*	Neo-arctic	wood	mgm4815520.3
NG90	**301-63**	Psammotermitidae		*Prorhinotermes inopinatus*	Oceania	wood	mgm4815523.3
G13-54	**301-45**	Rhinotermitidae		*Dolichorhinotermes longilabius*	Neo-tropical	wood	mgm4821365.3
NG30	**272-03**	Rhinotermitidae		*Parrhinotermes browni*	Oceania	wood	mgm4812121.3
THAI23	**301-82**	Rhinotermitidae		*Parrhinotermes* sp. A	Oriental	wood	mgm4815491.3
RDCT112	**301-70**	Rhinotermitidae		*Schedorhinotermes lamanianus*	Afro-tropical	wood	mgm4815487.3
TBRU2.3A	**272-92**	Rhinotermitidae		*Schedorhinotermes sarawakensis*	Oriental	wood	mgm4814060.3
THAI63	**272-102**	Rhinotermitidae		*Schedorhinotermes* sp. 3	Oriental	wood	mgm4839823.3
NG84	**272-72**	Termitigetonidae		*Termitogeton planus*	Oceania	wood	mgm4813751.3
G13-144	**272-37**	Serritermitidae		*Glossotermes oculatus*	Neo-tropical	wood	mgm4814059.3
BRA31	**230-25**	Serritermitidae		*Serritermes serrifer*	Neo-tropical	wood	mgm4782052.3
RDCT109	**272-83**	Termitidae	Macrotermitinae	*Pseudacanthotermes militaris*	Afro-tropical	fungus-growing	mgm4813754.3
CAM16-02a	**272-26**	Termitidae	Macrotermitinae	*Acanthotermes acanthothorax*	Afro-tropical	fungus-growing	mgm4814044.3
SAF5	**272-87**	Termitidae	Macrotermitinae	*Allodontotermes schultzei*	Afro-tropical	fungus-growing	mgm4813753.3
THAI071	**301-88**	Termitidae	Macrotermitinae	*Macrotermes annandalei*	Oriental	fungus-growing	mgm4815517.3
THAI50	**272-99**	Termitidae	Macrotermitinae	*Macrotermes gilvus*	Oriental	fungus-growing	mgm4839825.3
BDIT072	**230-09**	Termitidae	Macrotermitinae	*Macrotermes subhyalinus*	Afro-tropical	fungus-growing	mgm4782051.3
THAI064	**301-86**	Termitidae	Macrotermitinae	*Odontotermes javanicus*	Oriental	fungus-growing	mgm4815500.3

RDCT144	**301-72**	Termitidae	Macrotermitinae	*Odontotermes* sp. D	Afro-tropical	fungus-growing	mgm4815489.3
RDCT165	**230-22**	Termitidae	Sphaerotermitinae	*Sphaerotermes sphaerothorax*	Afro-tropical	wood-bacterial-comb	mgm4782048.3
TBRU8.14E	**230-21**	Termitidae	Foraminitermitinae	*Labritermes buttelreepeni*	Oriental	soil	mgm4782061.3
BDIT062	**230-23**	Termitidae	Apicotermitinae	*Acholotermes chirotus*	Afro-tropical	soil	mgm4782065.3
RDCT021	**301-65**	Termitidae	Apicotermitinae	*Acidnotermes praus*	Afro-tropical	soil	mgm4815515.3
BDIT049	**272-22**	Termitidae	Apicotermitinae	*Aderitotermes* sp. 2	Afro-tropical	soil	mgm4812117.3
BDIT061	**301-23**	Termitidae	Apicotermitinae	*Alyscotermes kilimandjaricus*	Afro-tropical	soil	mgm4821358.3
SAF6	**272-88**	Termitidae	Apicotermitinae	*Alyscotermes* sp.	Afro-tropical	soil	mgm4813746.3
RDCT098	**301-68**	Termitidae	Apicotermitinae	*Amalotermes phaeocephalus*	Afro-tropical	soil	mgm4815508.3
G13-32	**301-41**	Termitidae	Apicotermitinae	*Anoplotermes banksi*	Neo-tropical	soil	mgm4821356.3
G13-08	**301-48**	Termitidae	Apicotermitinae	*Anoplotermes janus*	Neo-tropical	soil	mgm4821354.3
G13-04	**272-42**	Termitidae	Apicotermitinae	*Anoplotermes parvus*	Neo-tropical	soil	mgm4814062.3
G13-69	**301-47**	Termitidae	Apicotermitinae	*Anoplotermes*-group sp. AF	Neo-tropical	soil	mgm4821350.3
G13-17	**272-38**	Termitidae	Apicotermitinae	*Anoplotermes*-group sp. N	Neo-tropical	soil	mgm4814058.3
G13-65	**301-46**	Termitidae	Apicotermitinae	*Anoplotermes*-group sp. Q	Neo-tropical	soil	mgm4821357.3
BDIT112	**301-27**	Termitidae	Apicotermitinae	*Astalotermes murcus*	Afro-tropical	soil	mgm4821369.3
RDCT070	**301-67**	Termitidae	Apicotermitinae	*Ateuchotermes retifaciens*	Afro-tropical	soil	mgm4815506.3
T4.14A	**272-90**	Termitidae	Apicotermitinae	*Euhamitermes hamatus*	Oriental	soil	mgm4813756.3
CAM212	**272-30**	Termitidae	Apicotermitinae	*Heimitermes laticeps*	Afro-tropical	soil	mgm4814072.3
CAM16-13	**272-28**	Termitidae	Apicotermitinae	*Labidotermes celesi*	Afro-tropical	soil	mgm4814042.3
G756	**272-47**	Termitidae	Apicotermitinae	*Patawatermes nigripunctatus*	Neo-tropical	soil	mgm4813749.3
CAM16-05	**272-27**	Termitidae	Apicotermitinae	*Phoxotermes cerberus*	Afro-tropical	soil	mgm4814065.3
THAI49	**272-98**	Termitidae	Amitermitinae	*Amitermes dentalus*	Oriental	wood	mgm4814053.3
AUS4	**301-15**	Termitidae	Amitermitinae	*Amitermes meridionalis*	Australia	grass	mgm4821348.3
G730	**301-53**	Termitidae	Amitermitinae	*Dentispicotermes brevicarinatus*	Neo-tropical	soil	mgm4815497.3
G697	**301-51**	Termitidae	Amitermitinae	*Orthognathotermes aduncus*	Neo-tropical	soil	mgm4815501.3
TBRU5.14A	**230-05**	Termitidae	Amitermitinae	*Prohamitermes mirabilis*	Oriental	wood	mgm4782055.3
BDIT043	**301-28**	Termitidae	Cubitermitinae	*Basidentitermes aurivillii*	Afro-tropical	soil	mgm4821342.3
BDIT069	**230-20**	Termitidae	Cubitermitinae	*Nititidermes fulvus*	Afro-tropical	soil	mgm4782049.3
RDCT105	**301-69**	Termitidae	Cubitermitinae	*Ophiotermes mirandus*	Afro-tropical	soil	mgm4815503.3
RDCT051	**301-66**	Termitidae	Cubitermitinae	*Orthotermes depressifrons*	Afro-tropical	soil	mgm4815518.3
RDCT159	**301-73**	Termitidae	Cubitermitinae	*Proboscitermes tubuliferus*	Afro-tropical	soil	mgm4815521.3
RDCT180	**272-86**	Termitidae	Cylindrotermitinae	*Cephalotermes rectangularis*	Afro-tropical	wood	mgm4813748.3
G13-24	**272-39**	Termitidae	Cylindrotermitinae	*Cylindrotermes parvignathus*	Neo-tropical	wood	mgm4814069.3
MAD15-5	**272-63**	Termitidae	Microcerotermitinae	*Microcerotermes aff. pauliani*	Madagascar	wood	mgm4814047.3
MAD15-169	**272-59**	Termitidae	Microcerotermitinae	*Microcerotermes aff. sikorae*	Madagascar	wood	mgm4814073.3
NG10	**272-01**	Termitidae	Microcerotermitinae	*Microcerotermes biroi*	Oceania	wood	mgm4812113.3
NG71	**272-05**	Termitidae	Microcerotermitinae	*Microcerotermes biroi*	Oceania	wood	mgm4812109.3
THAI54	**272-100**	Termitidae	Microcerotermitinae	*Microcerotermes crassus*	Oriental	wood	mgm4839822.3
TP1.14A	**272-104**	Termitidae	Microcerotermitinae	*Microcerotermes crassus*	Oriental	wood	mgm4839820.3
MAL22	**272-69**	Termitidae	Microcerotermitinae	*Microcerotermes crassus*	Oriental	wood	mgm4814071.3
MAL37	**272-70**	Termitidae	Microcerotermitinae	*Microcerotermes crassus*	Oriental	wood	mgm4814080.3
PHI1	**272-75**	Termitidae	Microcerotermitinae	*Microcerotermes crassus*	Oriental	wood	mgm4813739.3
RDCT033	**272-77**	Termitidae	Microcerotermitinae	*Microcerotermes fuscotibialis*	Afro-tropical	wood	mgm4813742.3
BDIT102	**272-19**	Termitidae	Microcerotermitinae	*Microcerotermes fuscotibialis*	Afro-tropical	wood	mgm4812115.3
T2-IC	**272-89**	Termitidae	Microcerotermitinae	*Microcerotermes havilandi*	Oriental	wood	mgm4813752.3
MAD15-85	**272-67**	Termitidae	Microcerotermitinae	*Microcerotermes* mad sp.A	Madagascar	wood	mgm4814068.3
MAD15-71	**301-61**	Termitidae	Microcerotermitinae	*Microcerotermes* mad sp.B	Madagascar	wood	mgm4815505.3
MAD15-16	**301-59**	Termitidae	Microcerotermitinae	*Microcerotermes* mad sp.C	Madagascar	wood	mgm4815492.3
MAD15-21	**272-60**	Termitidae	Microcerotermitinae	*Microcerotermes* mad sp.D	Madagascar	wood	mgm4814039.3
MAD15-59	**272-62**	Termitidae	Microcerotermitinae	*Microcerotermes* mad sp.F	Madagascar	wood	mgm4814051.3
MAD15-113	**272-53**	Termitidae	Microcerotermitinae	*Microcerotermes* mad sp.G	Madagascar	wood	mgm4814048.3
MAD15-116	**301-58**	Termitidae	Microcerotermitinae	*Microcerotermes* mad sp.H	Madagascar	wood	mgm4815507.3
MAD15-86	**272-68**	Termitidae	Microcerotermitinae	*Microcerotermes* mad sp.K	Madagascar	wood	mgm4814067.3
MAD15-151	**272-58**	Termitidae	Microcerotermitinae	*Microcerotermes* mad sp.M	Madagascar	wood	mgm4814054.3
MAD15-63	**272-65**	Termitidae	Microcerotermitinae	*Microcerotermes* mad sp.N	Madagascar	wood	mgm4814046.3
TBRU3.18a	**230-17**	Termitidae	Microcerotermitinae	*Microcerotermes* nr. *havilandi*	Oriental	wood	mgm4782066.3

NG28	**272-02**	Termitidae	Microcerotermitinae	*Microcerotermes papuanus*	Oceania	wood	mgm4812111.3
NG48	**272-04**	Termitidae	Microcerotermitinae	*Microcerotermes papuanus*	Oceania	wood	mgm4812108.3
NG81	**301-08**	Termitidae	Microcerotermitinae	*Microcerotermes papuanus*	Oceania	wood	mgm4782060.3
NG81	**230-12**	Termitidae	Microcerotermitinae	*Microcerotermes papuanus*	Oceania	wood	mgm4782060.3
RDCT055	**230-14**	Termitidae	Microcerotermitinae	*Microcerotermes parvus*	Afro-tropical	wood	mgm4782068.3
RDCT053	**272-79**	Termitidae	Microcerotermitinae	*Microcerotermes parvus*	Afro-tropical	wood	mgm4813740.3
RDCT119	**272-84**	Termitidae	Microcerotermitinae	*Microcerotermes parvus*	Afro-tropical	wood	mgm4813743.3
RDCT134	**272-85**	Termitidae	Microcerotermitinae	*Microcerotermes parvus*	Afro-tropical	wood	mgm4813755.3
BDIT045	**301-29**	Termitidae	Microcerotermitinae	*Microcerotermes parvus*	Afro-tropical	wood	mgm4821353.3
MAD15-136	**272-55**	Termitidae	Microcerotermitinae	*Microcerotermes pauliani*	Madagascar	wood	mgm4814066.3
TBRU9.12E	**272-95**	Termitidae	Microcerotermitinae	*Microcerotermes serrula*	Oriental	wood	mgm4814037.3
Msp_RNA_1	**229-07**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.	Oriental	wood	mgm4775431.3
PHI8	**230-03**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.	Oriental	wood	mgm4782067.3
D2-34	**230-15**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.	Oceania	wood	mgm4782056.3
AUS32	**272-10**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.	Australia	wood	mgm4812114.3
G13-58	**272-43**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.	Neo-tropical	wood	mgm4814082.3
G689	**272-45**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.	Neo-tropical	wood	mgm4842692.3
T6.14	**272-91**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.	Oriental	wood	mgm4814064.3
AUS13	**272-09**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.	Australia	wood	mgm4812125.3
MAD15-130	**272-54**	Termitidae	Microcerotermitinae	*Microcerotermes* sp. 1	Madagascar	wood	mgm4814040.3
MAD15-139	**272-56**	Termitidae	Microcerotermitinae	*Microcerotermes* sp. 1	Madagascar	wood	mgm4814036.3
MAD15-54	**230-16**	Termitidae	Microcerotermitinae	*Microcerotermes* sp. 2	Madagascar	wood	mgm4782059.3
MAD15-76	**272-66**	Termitidae	Microcerotermitinae	*Microcerotermes* sp. 2	Madagascar	wood	mgm4814075.3
MAD15-66	**301-60**	Termitidae	Microcerotermitinae	*Microcerotermes* sp. 2	Madagascar	wood	mgm4815499.3
THAI055	**272-101**	Termitidae	Microcerotermitinae	*Microcerotermes* sp. A	Oriental	wood	mgm4839821.3
AUS66	**272-12**	Termitidae	Microcerotermitinae	*Microcerotermes* sp. E	Australia	wood	mgm4812127.3
AUS66_2	**301-17**	Termitidae	Microcerotermitinae	*Microcerotermes* sp. E	Australia	wood	mgm4821331.3
FG-ND02-38	**272-33**	Termitidae	Microcerotermitinae	*Microcerotermes* sp. SA	Neo-tropical	wood	mgm4814074.3
AUS71	**272-14**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.G	Australia	wood	mgm4812110.3
AUS114	**301-11**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.H	Australia	wood	mgm4821372.3
AUS82	**272-15**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.I	Australia	wood	mgm4812130.3
FG-ND02-39	**272-34**	Termitidae	Microcerotermitinae	*Microcerotermes* sp.SC	Neo-tropical	wood	mgm4814049.3
MAD15-148	**272-57**	Termitidae	Microcerotermitinae	*Microcerotermes subtilis*	Madagascar	wood	mgm4814078.3
MAD15-62	**272-64**	Termitidae	Microcerotermitinae	*Microcerotermes subtilis*	Madagascar	wood	mgm4814043.3
MAD15-104	**272-52**	Termitidae	Microcerotermitinae	*Microcerotermes unidentatus*	Madagascar	wood	mgm4814052.3
SING57	**301-76**	Termitidae	Mirocapritermitinae	*Dicuspiditermes nemerosus*	Oriental	soil	mgm4815512.3
THAI038	**301-84**	Termitidae	Mirocapritermitinae	*Mirocapritermes* sp. 1	Oriental	soil	mgm4815488.3
NG45	**301-13**	Termitidae	Mirocapritermitinae	*Pericapritermes parvus*	Oceania	soil	mgm4821360.3
NG45_2	**301-62**	Termitidae	Mirocapritermitinae	*Pericapritermes parvus*	Oceania	soil	mgm4815524.3
NG55	**301-06**	Termitidae	Mirocapritermitinae	*Pericapritermes* sp. B	Oceania	soil	mgm4821338.3
THAI037	**301-83**	Termitidae	Mirocapritermitinae	*Procapritermes* sp. 1	Oriental	soil	mgm4815516.3
SP1	**301-57**	Termitidae	Mirocapritermitinae	*Sinocapritermes mushae*	Paleo-arctic	soil	mgm4815511.3
THAI105	**301-81**	Termitidae	Mirocapritermitinae	*Sinocapritermes* sp. 1	Oriental	soil	mgm4815504.3
G728	**272-46**	Termitidae	Nasutitermitinae	*Agnathotermes crassinasus*	Neo-tropical	soil	mgm4813735.3
THAI100	**301-80**	Termitidae	Nasutitermitinae	*Bulbitermes nr. laticephalus*	Oriental	wood	mgm4815513.3
G13-30	**272-40**	Termitidae	Nasutitermitinae	*Coatitermes kartaboensis*	Neo-tropical	soil	mgm4814041.3
G13-48	**301-44**	Termitidae	Nasutitermitinae	*Constrictotermes cavifrons*	Neo-tropical	lychen	mgm4821340.3
BRA1	**301-32**	Termitidae	Nasutitermitinae	*Constrictotermes cyphergaster*	Neo-tropical	lychen	mgm4821347.3
THAI067	**301-87**	Termitidae	Nasutitermitinae	*Hospitalitermes* sp. C	Oriental	lychen	mgm4815502.3
CAM16_18	**230-01**	Termitidae	Nasutitermitinae	*Leptomixotermes doriae*	Afro-tropical	soil	mgm4782050.3
CIVT017	**272-32**	Termitidae	Nasutitermitinae	*Mimeutermes sorex*	Afro-tropical	soil	mgm4814050.3
BDIT041	**272-21**	Termitidae	Nasutitermitinae	*Nasutitermes arborum*	Afro-tropical	wood	mgm4812124.3

NG69	**301-07**	Termitidae	Nasutitermitinae	*Nasutitermes bikpelanus*	Oceania	wood	mgm4821364.3
NG60	**230-11**	Termitidae	Nasutitermitinae	*Nasutitermes gracilirostris*	Oceania	wood	mgm4782045.3
AUS62	**301-18**	Termitidae	Nasutitermitinae	*Nasutitermes graveolus*	Australia	wood	mgm4821362.3
RDCT106	**272-82**	Termitidae	Nasutitermitinae	*Nasutitermes lujae*	Afro-tropical	wood	mgm4813734.3
G733	**301-54**	Termitidae	Nasutitermitinae	*Nasutitermes macrocephalus*	Neo-tropical	wood	mgm4815496.3
BRU6	**272-25**	Termitidae	Nasutitermitinae	*Nasutitermes matangensis*	Oriental	wood	mgm4814079.3
THAI43	**230-13**	Termitidae	Nasutitermitinae	*Nasutitermes* sp. 3	Oriental	wood	mgm4782054.3
AUS54	**301-16**	Termitidae	Nasutitermitinae	*Nasutitermes triodiae*	Australia	grass	mgm4821341.3
NSW6	**301-64**	Termitidae	Nasutitermitinae	*Occasitermes occasus*	Australia	grass	mgm4815526.3
THAI45	**301-85**	Termitidae	Nasutitermitinae	*Oriensubulitermes inanis*	Oriental	soil	mgm4815510.3
KE15-44	**272-50**	Termitidae	Nasutitermitinae	*Trinervitermes gratiosus*	Afro-tropical	grass	mgm4813736.3
BDIT094	**301-24**	Termitidae	Nasutitermitinae	*Trinervitermes* sp.	Afro-tropical	grass	mgm4821370.3
AUS49	**301-14**	Termitidae	Nasutitermitinae	*Tumulitermes sp.*	Australia	wood	mgm4821328.3
G13-60	**272-44**	Termitidae	Neocapritermitinae	*Neocapritermes taracua*	Neo-tropical	soil	mgm4814045.3
G13-28	**301-40**	Termitidae	Neocapritermitinae	*Planicapritermes planiceps*	Neo-tropical	soil	mgm4821336.3
G683	**301-50**	Termitidae	Neocapritermitinae	*Schievitermes globicornis*	Neo-tropical	soil	mgm4821346.3
RD1T21-M1e	**301-74**	Termitidae	Promirotermitinae	*Promirotermes pygmaeus*	Afro-tropical	soil	mgm4815522.3
TBRU7.11D	**272-93**	Termitidae	Protohamitermitinae	*Orientotermes emersoni*	Oriental	wood	mgm4814035.3
BRA3	**301-34**	Termitidae	Syntermitinae	*Cornitermes cumulans*	Neo-tropical	wood	mgm4821352.3
G13_62	**230-19**	Termitidae	Syntermitinae	*Cornitermes* sp. A	Neo-tropical	wood	mgm4782047.3
G13-45	**272-41**	Termitidae	Syntermitinae	*Cyrilliotermes angulariceps*	Neo-tropical	soil	mgm4814038.3
BRA14	**301-22**	Termitidae	Syntermitinae	*Cyrilliotermes* sp.	Neo-tropical	soil	mgm4821327.3
G13-23	**230-06**	Termitidae	Syntermitinae	*Embiratermes brevinasus*	Neo-tropical	soil	mgm4782069.3
G13-43	**301-43**	Termitidae	Syntermitinae	*Labiotermes labralis*	Neo-tropical	soil	mgm4821333.3
BRA29	**301-33**	Termitidae	Syntermitinae	*Labiotermes* sp.	Neo-tropical	soil	mgm4821344.3
BRA9	**301-36**	Termitidae	Syntermitinae	*Rhynchotermes nasutissimus*	Neo-tropical	litter	mgm4821363.3
BRA5	**301-35**	Termitidae	Syntermitinae	*Silvestritermes heyeri*	Neo-tropical	soil	mgm4821373.3
BRA11_2	**301-31**	Termitidae	Syntermitinae	*Syntermes grandis*	Neo-tropical	litter	mgm4821374.3
G13-112	**272-36**	Termitidae	Termitinae	*Cavitermes tuberosus*	Neo-tropical	soil	mgm4814056.3
NG49	**301-05**	Termitidae	Termitinae	*Protocapritermes odontomachus*	Oceania	soil	mgm4842691.3
G13-105	**272-35**	Termitidae	Termitinae	*Termes fatalis*	Neo-tropical	soil	mgm4814061.3
THAI096	**301-89**	Termitidae	Termitinae	*Termes rostratus*	Oriental	soil	mgm4815498.3
RDCT125	**301-71**	Termitidae	Termitinae	*Tuberculitermes bycanistes*	Afro-tropical	soil	mgm4815495.3

Suplementary table 2 cophylogeny analysis of CAZyme and host

ID cluster	number of sequences	paco p-value	Nye p-value	RF p-value	Percentage of CAZyme sequences belonging to each cluster
CBM22_cluster_1	21	0.0000	0.0000	0.0000	0.020600151
CBM9_cluster_1	123	0.0000	0.0000	0.0000	0.120658028
CBM9_cluster_2	67	0.0000	0.0000	0.0000	0.065724292
CE1_cluster_1	31	0.0000	0.0000	0.0000	0.030409747
CE1_cluster_10	50	0.0000	0.0000	0.0000	0.049047979
CE1_cluster_2	167	0.0000	0.0000	0.0000	0.163820249
CE1_cluster_3	33	0.0002	0.0001	0.0000	0.032371666
CE1_cluster_4	34	0.0000	0.0000	0.0000	0.033352626
CE1_cluster_5	52	0.0643	0.3184	0.0606	0.051009898
CE1_cluster_6	23	0.0094	0.0001	0.0001	0.02256207
CE1_cluster_7	37	0.0000	0.0000	0.0000	0.036295504
CE1_cluster_8	180	0.0000	0.0000	0.0000	0.176572723
CE1_cluster_9	29	0.0000	0.0000	0.0000	0.028447828
CE15_cluster_1	23	0.0015	0.0000	0.0000	0.02256207
CE15_cluster_2	143	0.0000	0.0000	0.0000	0.140277219
CE2_cluster_1	35	0.0000	0.0000	0.0000	0.034333585
CE2_cluster_2	209	0.0000	0.0000	0.0000	0.205020551
CE4_cluster_1	26	0.0008	0.0018	0.0007	0.025504949
CE4_cluster_5	140	0.0000	0.0000	0.0000	0.13733434
CE4_cluster_6	30	0.4737	0.0002	0.0001	0.029428787
CE4_cluster_7	24	0.0047	0.0009	0.0004	0.02354303
CE8_cluster_1	25	0.0000	0.0000	0.0000	0.024523989
CE9_cluster_1	73	0.0000	0.0000	0.0000	0.071610049
CE9_cluster_2	48	0.0000	0.0000	0.0000	0.04708606
CE9_cluster_3	394	0.0000	0.0000	0.0000	0.386498072
GH10_cluster_1	26	0.0000	0.0000	0.0000	0.025504949
GH10_cluster_10	252	0.0000	0.0002	0.0000	0.247201813
GH10_cluster_11	20	0.0005	0.0002	0.0000	0.019619191
GH10_cluster_12	25	0.0001	0.0000	0.0002	0.024523989
GH10_cluster_13	68	0.0000	0.0205	0.0000	0.066705251
GH10_cluster_14	24	0.0408	0.0000	0.0164	0.02354303
GH10_cluster_15	68	0.0000	0.0000	0.0000	0.066705251
GH10_cluster_16	72	0.0000	0.0000	0.0000	0.070629089
GH10_cluster_17	72	0.0000	0.0000	0.0000	0.070629089
GH10_cluster_18	24	0.0000	0.0000	0.0000	0.02354303

GH10_cluster_19	192	0.0000	0.0000	0.0000	0.188344238
GH10_cluster_20	282	0.0000	0.0000	0.0000	0.2766306
GH10_cluster_6	816	0.0000	0.0000	0.0000	0.800463013
GH10_cluster_7	26	0.0004	0.0000	0.0000	0.025504949
GH10_cluster_8	114	0.0000	0.0000	0.0000	0.111829392
GH10_cluster_9	36	0.0000	0.0000	0.0000	0.035314545
GH103_cluster_1	33	0.0000	0.0000	0.0000	0.032371666
GH103_cluster_2	28	0.0000	0.0000	0.0000	0.027466868
GH103_cluster_3	21	0.0000	0.0000	0.0000	0.020600151
GH103_cluster_4	37	0.0027	0.0002	0.0000	0.036295504
GH105_cluster_1	121	0.0000	0.0000	0.0000	0.118696109
GH105_cluster_10	23	0.0000	0.0000	0.0000	0.02256207
GH105_cluster_2	150	0.0000	0.0000	0.0000	0.147143936
GH105_cluster_3	39	0.0028	0.0000	0.0000	0.038257423
GH105_cluster_5	36	0.0001	0.0002	0.0000	0.035314545
GH105_cluster_8	43	0.0000	0.0014	0.0000	0.042181262
GH105_cluster_9	31	0.0129	0.0000	0.0010	0.030409747
GH106_cluster_1	138	0.0000	0.0000	0.0000	0.135372421
GH106_cluster_2	49	0.0000	0.0000	0.0000	0.048067019
GH106_cluster_3	34	0.0000	0.0217	0.0000	0.033352626
GH106_cluster_4	36	0.0014	0.0004	0.0026	0.035314545
GH106_cluster_5	27	0.0001	0.0000	0.0001	0.026485909
GH106_cluster_6	28	0.0000	0.0000	0.0000	0.027466868
GH110_cluster_1	27	0.0000	0.0001	0.0000	0.026485909
GH113_cluster_1	32	0.0000	0.0001	0.0000	0.031390706
GH113_cluster_3	24	0.0000	0.0000	0.0003	0.02354303
GH115_cluster_1	178	0.0000	0.0018	0.0000	0.174610804
GH115_cluster_3	24	0.0000	0.0000	0.0001	0.02354303
GH116_cluster_1	183	0.0000	0.0000	0.0000	0.179515602
GH116_cluster_2	24	0.0000	0.0003	0.0000	0.02354303
GH116_cluster_3	31	0.0016	0.0000	0.0000	0.030409747
GH125_cluster_1	39	0.0000	0.0000	0.0000	0.038257423
GH127_cluster_1	64	0.0063	0.0008	0.0000	0.062781413
GH127_cluster_2	35	0.0000	0.0002	0.0001	0.034333585
GH127_cluster_3	37	0.0056	0.0000	0.0000	0.036295504
GH128_cluster_1	78	0.0000	0.0264	0.0000	0.076514847
GH13_11_cluster_1	23	0.0000	0.0000	0.0000	0.02256207
GH13_11_cluster_2	20	0.0000	0.0000	0.0000	0.019619191
GH13_11_cluster_3	916	0.0000	0.0000	0.0000	0.89855897
GH13_13_cluster_1	54	0.0003	0.0000	0.0000	0.052971817
GH13_18_cluster_1	25	0.0000	0.0004	0.0000	0.024523989
GH13_18_cluster_2	38	0.0000	0.0000	0.0000	0.037276464

GH13_20_cluster_1	38	0.0002	0.0000	0.0000	0.037276464
GH13_20_cluster_2	122	0.0000	0.0000	0.0000	0.119677068
GH13_20_cluster_3	27	0.0001	0.0000	0.0000	0.026485909
GH13_20_cluster_4	417	0.0000	0.0000	0.0000	0.409060143
GH13_20_cluster_5	38	0.0008	0.0000	0.0000	0.037276464
GH13_20_cluster_6	27	0.0000	0.0000	0.0000	0.026485909
GH13_20_cluster_7	53	0.0000	0.0000	0.0000	0.051990857
GH13_21_cluster_1	32	0.0023	0.0003	0.0000	0.031390706
GH13_23_cluster_1	143	0.0000	0.0000	0.0000	0.140277219
GH13_31_cluster_2	304	0.0000	0.0001	0.0000	0.298211711
GH13_31_cluster_3	96	0.0000	0.0000	0.0000	0.094172119
GH13_36_cluster_2	63	0.0000	0.0002	0.0000	0.061800453
GH13_38_cluster_1	53	0.0000	0.0106	0.0000	0.051990857
GH13_9_cluster_1	36	0.0009	0.0515	0.0000	0.035314545
GH13_9_cluster_3	194	0.0000	0.0000	0.0000	0.190306157
GH13_cluster_1	110	0.0000	0.0001	0.0000	0.107905553
GH13_cluster_2	38	0.0017	0.0000	0.0000	0.037276464
GH130_cluster_1	32	0.0024	0.0000	0.0031	0.031390706
GH130_cluster_2	20	0.0000	0.0026	0.0000	0.019619191
GH130_cluster_3	24	0.0022	0.0000	0.0020	0.02354303
GH130_cluster_4	862	0.0000	0.0001	0.0000	0.845587153
GH139_cluster_1	23	0.0000	0.0000	0.0014	0.02256207
GH16_cluster_1	36	0.0003	0.0000	0.0000	0.035314545
GH16_cluster_2	85	0.0000	0.0000	0.0000	0.083381564
GH16_cluster_3	231	0.0000	0.0000	0.0000	0.226601662
GH16_cluster_4	172	0.0000	0.0000	0.0000	0.168725047
GH18_cluster_1	31	0.0000	0.0000	0.0000	0.030409747
GH18_cluster_2	37	0.0000	0.0000	0.0000	0.036295504
GH18_cluster_3	888	0.0000	0.0000	0.0000	0.871092102
GH18_cluster_4	22	0.0000	0.0008	0.0000	0.021581111
GH18_cluster_5	97	0.0000	0.0000	0.0000	0.095153079
GH18_cluster_7	24	0.0266	0.0044	0.0004	0.02354303
GH19_cluster_1	25	0.0001	0.0000	0.0000	0.024523989
GH2_cluster_1	21	0.0000	0.1786	0.0000	0.020600151
GH2_cluster_10	232	0.0000	0.0000	0.0000	0.227582621
GH2_cluster_11	176	0.0000	0.0023	0.0000	0.172648885
GH2_cluster_12	27	0.0115	0.0000	0.0210	0.026485909
GH2_cluster_13	213	0.0000	0.0429	0.0000	0.208944389
GH2_cluster_3	22	0.0004	0.0000	0.0000	0.021581111
GH2_cluster_4	25	0.0442	0.0000	0.0439	0.024523989
GH2_cluster_5	21	0.0002	0.3307	0.0882	0.020600151
GH2_cluster_6	21	0.0000	0.0006	0.0000	0.020600151

GH2_cluster_7	46	0.0000	0.0000	0.0000	0.04512414
GH2_cluster_8	62	0.0000	0.0000	0.0000	0.060819494
GH20_cluster_1	140	0.0000	0.0008	0.0000	0.13733434
GH20_cluster_2	25	0.0049	0.0000	0.0093	0.024523989
GH20_cluster_3	616	0.0000	0.0000	0.0000	0.604271098
GH20_cluster_4	25	0.0468	0.2059	0.0020	0.024523989
GH23_cluster_2	22	0.0002	0.0000	0.0000	0.021581111
GH23_cluster_3	56	0.0000	0.0000	0.0000	0.054933736
GH25_cluster_1	40	0.0000	0.0000	0.0000	0.039238383
GH25_cluster_2	29	0.0000	0.0000	0.0000	0.028447828
GH26_cluster_1	31	0.0058	0.0592	0.0001	0.030409747
GH26_cluster_2	45	0.0005	0.0000	0.0000	0.044143181
GH26_cluster_3	80	0.0000	0.0000	0.0000	0.078476766
GH26_cluster_4	322	0.0000	0.0000	0.0000	0.315868983
GH26_cluster_5	23	0.0014	0.0000	0.0125	0.02256207
GH26_cluster_6	195	0.0000	0.0000	0.0000	0.191287117
GH26_cluster_7	50	0.0001	0.0005	0.0000	0.049047979
GH27_cluster_1	49	0.0000	0.0000	0.0000	0.048067019
GH28_cluster_1	96	0.0000	0.0000	0.0000	0.094172119
GH29_cluster_1	30	0.0000	0.0000	0.0000	0.029428787
GH29_cluster_2	26	0.0003	0.0000	0.0001	0.025504949
GH29_cluster_4	47	0.0000	0.0435	0.0000	0.0461051
GH29_cluster_5	43	0.0000	0.0000	0.0000	0.042181262
GH3_cluster_1	74	0.0000	0.0000	0.0000	0.072591009
GH3_cluster_10	54	0.0000	0.0000	0.0000	0.052971817
GH3_cluster_11	135	0.0000	0.0000	0.0000	0.132429543
GH3_cluster_12	490	0.0000	0.0001	0.0000	0.480670192
GH3_cluster_14	110	0.0000	0.0000	0.0000	0.107905553
GH3_cluster_15	44	0.0009	0.0000	0.0000	0.043162221
GH3_cluster_16	58	0.0000	0.0000	0.0000	0.056895655
GH3_cluster_17	30	0.0000	0.0000	0.0000	0.029428787
GH3_cluster_18	424	0.0000	0.0299	0.0000	0.41592686
GH3_cluster_19	35	0.0000	0.0000	0.0000	0.034333585
GH3_cluster_2	674	0.0000	0.0000	0.0000	0.661166753
GH3_cluster_20	27	0.0000	0.0000	0.0000	0.026485909
GH3_cluster_4	110	0.0000	0.0000	0.0000	0.107905553
GH3_cluster_5	65	0.0000	0.0001	0.0000	0.063762372
GH3_cluster_6	809	0.0000	0.0001	0.0000	0.793596296
GH3_cluster_7	28	0.0000	0.0022	0.0000	0.027466868
GH30_1_cluster_1	22	0.0002	0.0000	0.0000	0.021581111
GH30_1_cluster_2	145	0.0000	0.0000	0.0000	0.142239138
GH30_2_cluster_1	55	0.0000	0.0000	0.0000	0.053952777

GH30_4_cluster_1	21	0.0010	0.0000	0.0000	0.020600151
GH30_8_cluster_1	31	0.0000	0.0000	0.0000	0.030409747
GH30_8_cluster_2	23	0.0000	0.0000	0.0000	0.02256207
GH30_8_cluster_3	360	0.0000	0.0000	0.0000	0.353145447
GH30_8_cluster_4	482	0.0000	0.0018	0.0000	0.472822515
GH30_cluster_1	30	0.0000	0.0000	0.0000	0.029428787
GH30_cluster_2	81	0.0121	0.0000	0.0000	0.079457726
GH30_cluster_3	249	0.0000	0.0000	0.0000	0.244258934
GH30_cluster_4	369	0.0000	0.0000	0.0000	0.361974083
GH31_cluster_2	23	0.0000	0.0000	0.0000	0.02256207
GH31_cluster_3	24	0.0000	0.0000	0.0000	0.02354303
GH31_cluster_4	165	0.0000	0.0000	0.0000	0.16185833
GH35_cluster_1	44	0.0000	0.0000	0.0000	0.043162221
GH38_cluster_1	96	0.0000	0.0000	0.0000	0.094172119
GH39_cluster_1	408	0.0000	0.0324	0.0000	0.400231506
GH39_cluster_2	31	0.0010	0.0000	0.0000	0.030409747
GH4_cluster_1	24	0.0515	0.0047	0.0000	0.02354303
GH4_cluster_2	161	0.0000	0.0000	0.0000	0.157934492
GH4_cluster_4	79	0.0000	0.0000	0.0000	0.077495806
GH4_cluster_5	198	0.0000	0.0000	0.0000	0.194229996
GH4_cluster_6	22	0.0003	0.0000	0.0029	0.021581111
GH4_cluster_8	70	0.0000	0.0000	0.0000	0.06866717
GH42_cluster_1	54	0.0000	0.0529	0.0000	0.052971817
GH42_cluster_2	21	0.0000	0.0000	0.0000	0.020600151
GH42_cluster_3	21	0.0027	0.0000	0.0011	0.020600151
GH42_cluster_4	24	0.0000	0.0126	0.0000	0.02354303
GH43_1_cluster_1	107	0.0000	0.0320	0.0000	0.104962674
GH43_1_cluster_2	146	0.0000	0.0000	0.0000	0.143220098
GH43_1_cluster_3	487	0.0000	0.0000	0.0000	0.477727313
GH43_10_cluster_1	27	0.0000	0.0000	0.0000	0.026485909
GH43_11_cluster_1	22	0.0003	0.0000	0.0137	0.021581111
GH43_12_cluster_1	131	0.0000	0.0000	0.0000	0.128505704
GH43_17_cluster_1	57	0.0000	0.0001	0.0000	0.055914696
GH43_17_cluster_2	32	0.0000	0.0000	0.0000	0.031390706
GH43_28_cluster_1	23	0.0096	0.0000	0.0000	0.02256207
GH43_28_cluster_2	21	0.0002	0.0000	0.0000	0.020600151
GH43_3_cluster_1	28	0.0000	0.0000	0.0000	0.027466868
GH43_30_cluster_1	23	0.0002	0.0000	0.0124	0.02256207
GH43_35_cluster_2	107	0.0000	0.0000	0.0000	0.104962674
GH43_35_cluster_3	40	0.0003	0.0000	0.0000	0.039238383
GH43_4_cluster_1	51	0.0003	0.0000	0.0000	0.050028938
GH43_cluster_1	64	0.0162	0.0046	0.0000	0.062781413

GH43_cluster_2	326	0.0000	0.0000	0.0000	0.319792821
GH44_cluster_1	28	0.0000	0.0000	0.0000	0.027466868
GH45_cluster_1	96	0.0004	0.0000	0.0000	0.094172119
GH45_cluster_2	352	0.0000	0.0000	0.0000	0.34529777
GH45_cluster_3	221	0.0000	0.0000	0.0000	0.216792066
GH45_cluster_4	52	0.0000	0.3340	0.0000	0.051009898
GH45_cluster_5	68	0.0000	0.0000	0.0000	0.066705251
GH45_cluster_6	389	0.0000	0.0000	0.0000	0.381593275
GH5_1_cluster_1	25	0.0000	0.0096	0.0000	0.024523989
GH5_10_cluster_1	45	0.0000	0.0000	0.0000	0.044143181
GH5_10_cluster_2	53	0.0000	0.0000	0.0000	0.051990857
GH5_10_cluster_3	223	0.0000	0.0000	0.0000	0.218753985
GH5_12_cluster_1	802	0.0000	0.0053	0.0000	0.786729579
GH5_2_cluster_1	189	0.0000	0.0000	0.0000	0.18540136
GH5_2_cluster_10	30	0.0000	0.0008	0.0000	0.029428787
GH5_2_cluster_11	302	0.0000	0.0000	0.0000	0.296249792
GH5_2_cluster_2	20	0.0172	0.0000	0.0034	0.019619191
GH5_2_cluster_3	28	0.0694	0.0000	0.0102	0.027466868
GH5_2_cluster_4	95	0.0000	0.0000	0.0000	0.09319116
GH5_2_cluster_5	38	0.0234	0.0000	0.0000	0.037276464
GH5_2_cluster_6	131	0.0000	0.0001	0.0000	0.128505704
GH5_2_cluster_7	181	0.0000	0.1096	0.0000	0.177553683
GH5_2_cluster_8	193	0.0000	0.0000	0.0000	0.189325198
GH5_2_cluster_9	50	0.0000	0.0000	0.0000	0.049047979
GH5_25_cluster_1	68	0.0000	0.0000	0.0000	0.066705251
GH5_36_cluster_1	54	0.0000	0.0000	0.0000	0.052971817
GH5_36_cluster_2	93	0.0000	0.4250	0.0000	0.09122924
GH5_36_cluster_3	30	0.0000	0.0168	0.0000	0.029428787
GH5_39_cluster_1	673	0.0000	0.0000	0.0000	0.660185794
GH5_4_cluster_1	59	0.0000	0.0000	0.0000	0.057876615
GH5_4_cluster_10	983	0.0000	0.0000	0.0000	0.964283262
GH5_4_cluster_11	37	0.0001	0.0000	0.0000	0.036295504
GH5_4_cluster_12	32	0.0001	0.0000	0.0000	0.031390706
GH5_4_cluster_2	363	0.0000	0.0000	0.0000	0.356088326
GH5_4_cluster_3	304	0.0000	0.0000	0.0000	0.298211711
GH5_4_cluster_4	866	0.0000	0.0000	0.0000	0.849510992
GH5_4_cluster_5	209	0.0000	0.0000	0.0000	0.205020551
GH5_4_cluster_6	29	0.0000	0.0000	0.0000	0.028447828
GH5_4_cluster_7	39	0.0000	0.0000	0.0000	0.038257423
GH5_4_cluster_8	240	0.0000	0.0000	0.0000	0.235430298
GH5_4_cluster_9	32	0.0000	0.0000	0.0000	0.031390706
GH5_40_cluster_1	25	0.0000	0.0000	0.0000	0.024523989

GH5_46_cluster_1	53	0.0000	0.0000	0.0000	0.051990857
GH5_46_cluster_3	20	0.0000	0.0000	0.0000	0.019619191
GH5_52_cluster_1	270	0.0000	0.0000	0.0000	0.264859085
GH5_52_cluster_2	546	0.0000	0.0000	0.0000	0.535603928
GH5_cluster_1	310	0.0000	0.2136	0.0000	0.304097468
GH50_cluster_1	33	0.0026	0.0000	0.0000	0.032371666
GH53_cluster_1	269	0.0000	0.0001	0.0000	0.263878126
GH53_cluster_2	118	0.0000	0.0000	0.0000	0.11575323
GH53_cluster_3	48	0.9536	0.0000	0.2756	0.04708606
GH53_cluster_4	28	0.3135	0.0007	0.0040	0.027466868
GH55_cluster_1	48	0.0000	0.0000	0.0000	0.04708606
GH57_cluster_1	44	0.0000	0.0000	0.0000	0.043162221
GH57_cluster_10	103	0.0000	0.0000	0.0000	0.101038836
GH57_cluster_11	182	0.0000	0.0000	0.0000	0.178534643
GH57_cluster_12	50	0.0000	0.1138	0.0000	0.049047979
GH57_cluster_13	780	0.0000	0.0000	0.0000	0.765148468
GH57_cluster_14	123	0.0000	0.2665	0.0000	0.120658028
GH57_cluster_2	35	0.0000	0.0018	0.0000	0.034333585
GH57_cluster_3	591	0.0000	0.0000	0.0000	0.579747109
GH57_cluster_5	40	0.0000	0.0000	0.0000	0.039238383
GH57_cluster_6	31	0.2842	0.0002	0.2492	0.030409747
GH57_cluster_7	61	0.0000	0.0000	0.0000	0.059838534
GH64_cluster_1	22	0.0483	0.0000	0.0694	0.021581111
GH65_cluster_1	67	0.0000	0.0000	0.0000	0.065724292
GH65_cluster_2	22	0.0000	0.0000	0.0000	0.021581111
GH67_cluster_1	72	0.0000	0.0000	0.0000	0.070629089
GH73_cluster_1	21	0.0003	0.0000	0.0148	0.020600151
GH73_cluster_2	24	0.1349	0.0000	0.0476	0.02354303
GH73_cluster_3	41	0.0000	0.0000	0.0000	0.040219343
GH73_cluster_4	26	0.0000	0.0000	0.0000	0.025504949
GH77_cluster_1	271	0.0000	0.0000	0.0000	0.265840045
GH77_cluster_10	174	0.0000	0.0000	0.0000	0.170686966
GH77_cluster_2	1080	0.0001	0.0000	0.0000	1.059436341
GH77_cluster_3	37	0.0000	0.0000	0.0000	0.036295504
GH77_cluster_4	23	0.0000	0.0000	0.0000	0.02256207
GH77_cluster_5	41	0.0000	0.0000	0.0000	0.040219343
GH77_cluster_6	265	0.0000	0.0000	0.0000	0.259954287
GH77_cluster_7	30	0.0000	0.0000	0.0000	0.029428787
GH77_cluster_9	46	0.0000	0.0000	0.0000	0.04512414
GH78_cluster_1	45	0.0000	0.0000	0.0000	0.044143181
GH78_cluster_2	25	0.0008	0.1456	0.0000	0.024523989
GH78_cluster_3	27	0.0003	0.0003	0.0000	0.026485909

GH8_cluster_10	79	0.0000	0.0000	0.0000	0.077495806
GH8_cluster_11	23	0.0000	0.0000	0.0000	0.02256207
GH8_cluster_12	139	0.0000	0.1055	0.0000	0.136353381
GH8_cluster_2	35	0.0010	0.0000	0.0000	0.034333585
GH8_cluster_3	117	0.0004	0.0000	0.0000	0.11477227
GH8_cluster_4	102	0.0000	0.0000	0.0000	0.100057877
GH8_cluster_5	156	0.0000	0.0000	0.0000	0.153029694
GH8_cluster_6	54	0.0000	0.0000	0.0000	0.052971817
GH8_cluster_8	138	0.0000	0.0000	0.0000	0.135372421
GH8_cluster_9	135	0.0000	0.0000	0.0000	0.132429543
GH88_cluster_1	26	0.0002	0.0000	0.0000	0.025504949
GH88_cluster_2	25	0.0000	0.0000	0.0000	0.024523989
GH9_cluster_1	135	0.0002	0.0001	0.0000	0.132429543
GH9_cluster_10	61	0.0000	0.0000	0.0000	0.059838534
GH9_cluster_11	97	0.0000	0.0000	0.0000	0.095153079
GH9_cluster_12	217	0.0000	0.0000	0.0000	0.212868228
GH9_cluster_2	120	0.0000	0.0000	0.0000	0.117715149
GH9_cluster_3	143	0.0000	0.0000	0.0000	0.140277219
GH9_cluster_4	428	0.0000	0.0000	0.0000	0.419850698
GH9_cluster_5	195	0.0000	0.0003	0.0000	0.191287117
GH9_cluster_6	25	0.0000	0.0003	0.0000	0.024523989
GH9_cluster_7	247	0.0000	0.0000	0.0000	0.242297015
GH9_cluster_8	21	0.0020	0.2765	0.0000	0.020600151
GH9_cluster_9	190	0.0000	0.0000	0.0000	0.186382319
GH92_cluster_1	25	0.0000	0.0000	0.0000	0.024523989
GH92_cluster_2	33	0.0002	0.0000	0.0001	0.032371666
GH94_cluster_10	74	0.0000	0.0000	0.0000	0.072591009
GH94_cluster_11	502	0.0000	0.0000	0.0000	0.492441706
GH94_cluster_2	42	0.0000	0.0000	0.0000	0.041200302
GH94_cluster_3	48	0.0000	0.0000	0.0000	0.04708606
GH94_cluster_4	44	0.0030	0.0149	0.0000	0.043162221
GH94_cluster_5	99	0.0000	0.0000	0.0000	0.097114998
GH94_cluster_6	123	0.0000	0.0001	0.0000	0.120658028
GH94_cluster_7	77	0.0000	0.0004	0.0000	0.075533887
GH94_cluster_8	119	0.0000	0.0000	0.0000	0.116734189
GH94_cluster_9	60	0.0000	0.0000	0.0000	0.058857574
GH95_cluster_1	62	0.0000	0.0000	0.0000	0.060819494
GH95_cluster_3	42	0.0000	0.0000	0.0000	0.041200302
GH95_cluster_4	28	0.0000	0.0000	0.0000	0.027466868
GH99_cluster_1	52	0.0381	0.0000	0.0001	0.051009898
GT10_cluster_1	22	0.0036	0.0000	0.0005	0.021581111
GT104_cluster_1	125	0.0000	0.0000	0.0000	0.122619947

GT104_cluster_2	22	0.0008	0.0001	0.0000	0.021581111
GT104_cluster_3	27	0.0000	0.0000	0.0000	0.026485909
GT11_cluster_1	32	0.0000	0.0000	0.0000	0.031390706
GT11_cluster_2	27	0.0000	0.0000	0.0001	0.026485909
GT14_cluster_3	33	0.0070	0.1741	0.0000	0.032371666
GT14_cluster_4	20	0.0000	0.0001	0.0000	0.019619191
GT17_cluster_1	39	0.0000	0.0004	0.0003	0.038257423
GT19_cluster_1	42	0.0000	0.0001	0.0000	0.041200302
GT19_cluster_10	79	0.0000	0.0000	0.0000	0.077495806
GT19_cluster_11	23	0.0000	0.0000	0.0000	0.02256207
GT19_cluster_12	21	0.0213	0.0000	0.0000	0.020600151
GT19_cluster_2	54	0.0000	0.0007	0.0000	0.052971817
GT19_cluster_3	129	0.0000	0.0000	0.0000	0.126543785
GT19_cluster_4	30	0.0000	0.0013	0.0000	0.029428787
GT19_cluster_6	22	0.0000	0.0001	0.0000	0.021581111
GT19_cluster_7	22	0.0000	0.0005	0.0000	0.021581111
GT19_cluster_8	21	0.0000	0.0000	0.0000	0.020600151
GT19_cluster_9	237	0.0007	0.0000	0.0000	0.232487419
GT26_cluster_1	1055	0.0000	0.0000	0.0000	1.034912351
GT26_cluster_2	31	0.0000	0.0000	0.0000	0.030409747
GT26_cluster_3	22	0.0012	0.0000	0.0000	0.021581111
GT28_cluster_1	60	0.0000	0.0000	0.0000	0.058857574
GT28_cluster_10	293	0.0000	0.0000	0.0000	0.287421155
GT28_cluster_11	35	0.0000	0.0000	0.0007	0.034333585
GT28_cluster_12	33	0.0007	0.0000	0.0000	0.032371666
GT28_cluster_13	32	0.0000	0.0000	0.0000	0.031390706
GT28_cluster_14	50	0.0012	0.0000	0.0000	0.049047979
GT28_cluster_15	181	0.0000	0.0016	0.0000	0.177553683
GT28_cluster_16	28	0.0000	0.0000	0.0000	0.027466868
GT28_cluster_17	90	0.0000	0.0011	0.0000	0.088286362
GT28_cluster_18	107	0.0000	0.0000	0.0000	0.104962674
GT28_cluster_19	354	0.0000	0.0000	0.0000	0.347259689
GT28_cluster_2	26	0.0025	0.0000	0.0000	0.025504949
GT28_cluster_20	649	0.0000	0.0000	0.0000	0.636642764
GT28_cluster_3	26	0.0000	0.0000	0.0000	0.025504949
GT28_cluster_4	30	0.0000	0.0000	0.0000	0.029428787
GT28_cluster_5	21	0.0025	0.0000	0.0000	0.020600151
GT28_cluster_6	39	0.0000	0.0358	0.0000	0.038257423
GT28_cluster_7	58	0.0000	0.0000	0.0001	0.056895655
GT28_cluster_8	75	0.0000	0.0066	0.0000	0.073571968
GT28_cluster_9	52	0.0071	0.0000	0.0000	0.051009898
GT30_cluster_1	43	0.0000	0.0000	0.0001	0.042181262

GT30_cluster_2	49	0.0000	0.0000	0.0000	0.048067019
GT35_cluster_1	672	0.0000	0.0000	0.0000	0.659204834
GT35_cluster_2	62	0.0000	0.0000	0.0000	0.060819494
GT35_cluster_3	38	0.0000	0.1008	0.0000	0.037276464
GT35_cluster_4	64	0.0000	0.0122	0.0000	0.062781413
GT35_cluster_5	41	0.0002	0.0014	0.0000	0.040219343
GT35_cluster_7	205	0.0000	0.0000	0.0000	0.201096713
GT35_cluster_8	238	0.0000	0.0000	0.0000	0.233468379
GT41_cluster_2	22	0.0000	0.0000	0.0000	0.021581111
GT51_cluster_1	45	0.0000	0.0000	0.0000	0.044143181
GT51_cluster_10	28	0.0000	0.0056	0.0000	0.027466868
GT51_cluster_11	112	0.0000	0.0000	0.0000	0.109867472
GT51_cluster_12	26	0.0002	0.0000	0.0000	0.025504949
GT51_cluster_13	24	0.0046	0.0413	0.0001	0.02354303
GT51_cluster_14	206	0.0000	0.0000	0.0000	0.202077672
GT51_cluster_15	254	0.0000	0.0000	0.0000	0.249163732
GT51_cluster_17	29	0.1512	0.0000	0.0107	0.028447828
GT51_cluster_18	36	0.0000	0.0000	0.0000	0.035314545
GT51_cluster_19	689	0.0000	0.0006	0.0000	0.675881147
GT51_cluster_2	168	0.0000	0.0244	0.0000	0.164801209
GT51_cluster_20	110	0.0000	0.0000	0.0000	0.107905553
GT51_cluster_21	183	0.0000	0.0000	0.0000	0.179515602
GT51_cluster_22	32	0.0631	0.0000	0.0320	0.031390706
GT51_cluster_23	21	0.0188	0.0000	0.0013	0.020600151
GT51_cluster_24	26	0.0000	0.0000	0.0000	0.025504949
GT51_cluster_3	22	0.0000	0.0000	0.0000	0.021581111
GT51_cluster_4	20	0.0001	0.0000	0.0100	0.019619191
GT51_cluster_5	100	0.0000	0.0000	0.0000	0.098095957
GT51_cluster_6	268	0.0000	0.0000	0.0000	0.262897166
GT51_cluster_7	71	0.0000	0.0000	0.0000	0.06964813
GT51_cluster_8	136	0.0000	0.0000	0.0000	0.133410502
GT51_cluster_9	119	0.0000	0.0000	0.0000	0.116734189
GT6_cluster_1	39	0.0005	0.0000	0.0000	0.038257423
GT8_cluster_1	289	0.0000	0.0000	0.0000	0.283497317
GT8_cluster_2	40	0.0000	0.0000	0.0000	0.039238383
GT8_cluster_3	65	0.0000	0.0000	0.0000	0.063762372
GT8_cluster_4	62	0.0000	0.0000	0.0000	0.060819494
GT81_cluster_1	22	0.0000	0.0000	0.0000	0.021581111
GT81_cluster_2	70	0.0000	0.0000	0.0000	0.06866717
PL1_2_cluster_1	31	0.0008	0.0000	0.0000	0.030409747
PL1_2_cluster_3	24	0.0000	0.0000	0.0000	0.02354303
PL1_cluster_1	26	0.0000	0.0000	0.0000	0.025504949

PL1_cluster_2	96	0.0000	0.0000	0.0000	0.094172119
PL1_cluster_3	127	0.0000	0.0000	0.0000	0.124581866
PL11_cluster_1	33	0.0000	0.0000	0.0000	0.032371666
PL11_cluster_2	208	0.0000	0.0000	0.0000	0.204039592
PL14_3_cluster_1	39	0.0000	0.0000	0.0000	0.038257423
PL14_3_cluster_2	26	0.0000	0.0000	0.0000	0.025504949
PL9_cluster_1	96	0.0000	0.0000	0.0000	0.094172119

8.2 Supplementary files

Supplementary file 1 – Cophylogeny script

```
#USAGE- Rscript [tree] [termitetree] [output_file1]
args <- commandArgs(TRUE)
tree <- args[1]
termitetree<-args[2]
output_file1 <- args[3]


library(plyr)
library(dplyr)
library(tidyr)
library(ape)


#read the symbiont tree-
symbiont_tree<-read.tree(file="noGTDB-CBM22_cluster_1.newick")
symbiont_tree2<-symbiont_tree

e<-data.frame(label=symbiont_tree2$tip.label)
e$label<-as.character(e$label)

#read the termite tree
termite_tree_new<-read.tree(file="newnames_rooted_ucetermitetree_57p_198samples.nwk")
#is.binary(termite_tree_new) # true
#is.ultrametric(termite_tree_new) #true
#is.rooted(termite_tree_new) #true

d<-data.frame(label=termite_tree_new$tip.label)
d$label<-as.character(d$label)

library(vegan)

H.matrix<-cophenetic(termite_tree_new)
E.matrix<-cophenetic(symbiont_tree2)

#generate the HE.matrix-
HE.matrix <- data.frame(matrix(ncol = nrow(e), nrow = nrow(d))) #columns=nrow(e), rows=nrow(d)
rownames(HE.matrix) <- d$label
colnames(HE.matrix)<-e$label

HE.matrix2<-HE.matrix
#remove the gene names from column names (everything after "--")
rownames(HE.matrix2) <- gsub(x = rownames(HE.matrix2), pattern = "--.*", replacement = "")
```

```
names(HE.matrix2) <- gsub(x = names(HE.matrix2), pattern = "--.*", replacement = "")
#names(HE.matrix2) <- gsub(x = names(HE.matrix2), pattern = "_", replacement = "-")


#match rownames and column names-
out <- outer(row.names(HE.matrix2), colnames(HE.matrix2), `==`) #find cells that are common
dimnames(out) <- dimnames(HE.matrix2)
out<-as.data.frame(out)
colnames(out)<-colnames(HE.matrix) #change the column names to original tip.labels
rownames(out)<-rownames(HE.matrix)
out <- as.matrix(1*out) #convert logical to numeric and into a matrix

#run paco-

library(paco)
D<-prepare_paco_data(H=H.matrix,P=E.matrix,HP=out)
D<-add_pcoord(D,correction="cailliez")
D<-PACo(D,nperm=10000,seed=12,method="backtrack",symmetric=FALSE) #"r0" algorithm is used when host maintains the symbionts evolution. If not known use "backtracking" or "swaps"
#symmetric=FALSE: one group is not assumed to track the evolution of the other.
D<-paco_links(D) #a jackknife procedure to estimate the degree of individual interactions
res<-as.data.frame(residuals_paco(D$proc)) #residuals of each interactions
links<-as.data.frame(D$jackknife)
D.pvalue<-D$gof #output D$gof$ss ->m^2xy=56.92 #gives the p-value of overall phylogenetic coevolution.
write.csv(D.pvalue,file="output_file1_test")

#---------beforetreedist----------#


#USAGE- Rscript [tree1] [output_file1] [output_file2]
args <- commandArgs(TRUE)
symbionttree <- args[1]
outputtree<-args[2] #symbionttree-${i} #my case out1-
outputfile<-args[3] #symbiontheader-${i} #my case out2-

##rename symbiont tree file with hostname_1
library(ape)
symbionttree<-read.tree(file="noGTDB-GT104_cluster_2.newick")

d<-data.frame(label=symbionttree$tip.label)
#d$label2<-gsub("d__.*$","",d$label)
#d$label2 <- sub("--[^--]+$", "", d$label2)
d$runnumber<-gsub("--.*$","",d$label)

hosttree<-read.tree("correct_newnames_rooted_ucetermitetree_57p_198samples.nwk")
e<-data.frame(label=hosttree$tip.label)
e$runnumber<-gsub("--.*$","",e$label)

d2<-merge(d, e, by="runnumber")

library(dplyr)
d2<-d2 %>%group_by(label.y) %>%mutate(onemore = paste(label.y,row_number(label.y),sep="_"))

##output- newtree, header
library(ggtree)
library(treeio)
symbionttree2<-rename_taxa(symbionttree, d2, label.y, onemore)
```

```
write.tree(symbionttree2,file="out1-GH103_cluster_4.newick")

d2<-as.data.frame(d2)
d3<-d2%>%select(onemore)
write.csv(d3,file="out2-GT104_cluster_2.txt")

##run "newhosttree.R" script to get hosttree with "0" branch lengths equal to symbionttree-

awk -F"," '{print $2}' symbiontheader-
d__Bacteria_p__Spirochaetota__p__UBP6__p__Deferribacterota__COG0552_tips_4.txt | sed 's/"//g' > 2-
symbiontheader-d__Bacteria_p__Spirochaetota__p__UBP6__p__Deferribacterota__COG0552_tips_4.txt
####working for me, before rename out2 and ad .txt or edit output from befortreedist###


#or
#for i in out2-*;do awk -F"," '{print $2}' ${i} | sed 's/"//g' > 2-${i};done

#my case symbiontheader = out2-

#IN_DIR="/flash/BourguignonU/Jigs/markers/cophylogeny"
#Rscript ${IN_DIR}/newhosttree.R

# 2-symbiontheader-d__Bacteria_p__Spirochaetota__p__UBP6__p__Deferribacterota__COG0552_tips_4.txt
${IN_DIR}/rna12-16S_202samples_newnames_feb2020.nwk hosttree-
d__Bacteria_p__Spirochaetota__p__UBP6__p__Deferribacterota__COG0552_tips_4.nwk
```

#---------**newhosttree**----------#

```
#!/usr/bin/env Rscript
#USAGE- Rscript [alignment] [hosttree] [output_file]
args <- commandArgs(TRUE)
alignment <- args[1]
hosttree <- args[2]
output_file <- args[3]

library(ape)
library(phytools)
#library(devtools)
#install_github("hmorlon/PANDA",ref="Benoit", dependencies = TRUE)
#install_github("BPerezLamarque/HOME", dependencies = TRUE)
library(HOME)

#alignment <- read.dna(file=alignment, format = "fasta", as.character = T)
alignment<-read.csv(file="2-out2-GH103_cluster_4.txt",header=FALSE) #this is the header of the alignment file
rownames(alignment)<-alignment$V1

host_tree <- read.tree(file="correct_newnames_rooted_ucetermitetree_57p_198samples.nwk")

# Add tree tips with close to zero branch lengths to match number of microbial sequences
add_host_tips <- function(host_tree, alignment){
 #host_tree$edge.length <- host_tree$edge.length/sum(host_tree$edge.length)
 tip_labels <- host_tree$tip.label[order(nchar(host_tree$tip.label), decreasing = TRUE)]
 list_reads <- c()
 for (i in 1:length(tip_labels)){
  reads <- grep(tip_labels[i], rownames(alignment))
  reads <- reads[!reads %in% list_reads]
```

```
   list_reads <- c(list_reads, reads)
   if (length(reads)>0){
    if (!tip_labels[i] %in% rownames(alignment)){ host_tree$tip.label[which(host_tree$tip.label==tip_labels[i])]
<-  rownames(alignment)[reads[1]]
    reference_tip <- rownames(alignment)[reads[1]] }else{ reference_tip <- tip_labels[i] }
    reads <- reads[which(!rownames(alignment)[reads] %in% host_tree$tip.label)]
    if (length(reads)>0){
     for (j in 1:length(reads)){
      host_tree <- bind.tip(host_tree, tip.label=rownames(alignment)[reads[j]], edge.length=NULL,
where=which(host_tree$tip.label==reference_tip), position=min(0.001,min(host_tree$edge.length)))
     }
    }
   }
 }
 host_tree <- drop.tip(host_tree, tip = host_tree$tip.label[!host_tree$tip.label %in% rownames(alignment)])
 host_tree$edge.length[host_tree$edge.length==0] <- 0.001
 return(force.ultrametric(host_tree,method = "extend"))
}

provided_tree <- add_host_tips(host_tree, alignment)

### ************************************************************
#*                    Note:                        *
#*   force.ultrametric does not include a formal method to    *
#*   ultrametricize a tree & should only be used to coerce    *
#*  a phylogeny that fails is.ultramtric due to rounding --   *
#*   not as a substitute for formal rate-smoothing methods.   *
#************************************************************
write.tree(provided_tree,file="hosttree-GH103_cluster_4.newick")

#------treedist.R--------#


#!/usr/bin/env Rscript
#USAGE- Rscript [tree1] [tree2] [output_file1] [output_file2]
args <- commandArgs(TRUE)
tree1 <- args[1]
tree2<-args[2]
output_file1 <- args[3]
output_file2<- args [4]

## run "before_treedist.R" to generate the tree1 and tree2 files
library(ape)
tree1<-read.tree("noGTDB-GH103_cluster_4.newick")
tree2<-read.tree("hosttree-GH103_cluster_4.newick")

#install.packages("rlang")
library(rlang) #0.4.10 version
library(usethis)
library(rJava) #0.9-13
library(htmltools) #0.5.1.1
#install.packages("TreeDist")
library(TreeDist)
library(TreeTools)
#install_github("ms609/TreeDistData")
library(TreeDistData)
library(TreeSearch)
```

```
#method1-generate random trees for the host tree-<USING GENERALIZED RF method>
tree1<-unroot(tree1) #based on https://github.com/ms609/TreeDist/issues/58
nRep <- 100000 # Use more replicates for more accurate estimate of expected value
randomTrees <- lapply(logical(nRep), function (x) RandomTree(tree1$tip.label))
randomDists <- ClusteringInfoDistance(tree1, randomTrees, normalize = TRUE)
 canexpectedCID <- mean(randomDists)


dist12 <- ClusteringInfoDistance(tree1, tree2, normalize = TRUE)
# Now count the number of random trees that are this similar to tree1
nThisSimilar <- sum(randomDists < dist12)
pValue <- nThisSimilar / nRep

write.csv(pValue,"output_file1_GH10_cluster_17_treedist.csv")

#-------------------------------------------------------------------------------
#method2-generate random trees for the host tree-<USING NYE method>
nRep <- 100000 # Use more replicates for more accurate estimate of expected value
randomTrees <- lapply(logical(nRep), function (x) RandomTree(tree1$tip.label))
randomDists <- NyeSimilarity(tree1, randomTrees, normalize = TRUE,similarity = FALSE)
expectedCID <- mean(randomDists)


dist12 <- NyeSimilarity(tree1, tree2, normalize = TRUE,similarity = FALSE)# Now count the number of random
trees that are this similar to tree1
nThisSimilar <- sum(randomDists < dist12)
pValue2 <- nThisSimilar / nRep

write.csv(pValue2,"output_file2_GH10_cluster_16_treedist.csv")
```