Glycoside Hydrolase Gene Families Of Termite Hindgut Protists

by

Viola Sanderlin

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved June 2019 by the
Graduate Supervisory Committee:

Gillian Gile, Chair
Martin Wojciechowski
Taylor Weiss
Arul Mozhy Varman

ARIZONA STATE UNIVERSITY

August 2019

ABSTRACT

This project was completed to understand the evolution of the ability to digest wood in termite symbiotic protists. Lower termites harbor bacterial and protist symbionts which are essential to the termite ability to use wood as a nutritional source, producing glycoside hydrolases to break down the polysaccharides found in lignocellulose. Yet, only a few molecular studies have been done to confirm the protist species responsible for particular enzymes. By mining publicly available and newly generated genomic and transcriptomic data, including three transcriptomes from isolated protist cells, I identify over 200 new glycoside hydrolase sequences and compute the phylogenies of eight glycoside hydrolase families (GHFs) reported to be expressed by termite hindgut protists.

Of those families examined, the results are broadly consistent with Todaka *et al.* 2010, though none of the GHFs found were expressed in both termite-associated protist and non-termite-associated protist transcriptome data. This suggests that, rather than being inherited from their free-living protist ancestors, GHF genes were acquired by termite protists while within the termite gut, potentially via lateral gene transfer (LGT). For example one family, GHF10, implies a single acquisition of a bacterial xylanase into termite protists. The phylogenies from GHF5 and GHF11 each imply two distinct acquisitions in termite protist ancestors, each from bacteria. In eukaryote-dominated GHFs, GHF7 and GHF45, there are three apparent acquisitions by termite protists. Meanwhile, it appears prior reports of GHF62 in the termite gut may have been misidentified GHF43 sequences. GHF43 was the only GHF found to contain sequences from the protists not found in the termite gut. These findings generally all support the possibility termite-associated protists adapted to a lignocellulosic diet after colonization of the termite hindgut. Nonetheless, the poor resolution of GHF phylogeny and limited termite and protist sampling constrain interpretation.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Lignocellulosic Biomass

Currently, petroleum is the dominant feedstock in various fuel and material enterprises (Sanderson, 2011). As the global supply of crude oil is non-renewable and concerns mount regarding its contribution to global climate change, carbon-neutral alternatives are being investigated (Jensen *et al.*, 2017). First generation biofuel approaches focused on sugar, starch, and vegetable oils as feedstock, though their appeal is limited as they compete with food in the market and in land use (Scharf and Boucias, 2010). Using lignocellulosic materials, especially non-edible biomass, in place of fossil fuels is attractive as being potentially more eco-friendly and sustainable (Isikgor and Becer, 2015). Lignocellulosic biomass is the most abundant bio-renewable source of organic carbon on earth (Liu *et al.*, 2011). Non-food agro-industrial biomass is available as waste byproducts from a variety of industries, making it an appealing source of bio-sourced feedstock (Anwar *et al.*, 2014). These second-generation feedstocks include forestry remnants as well as the non-edible portions of crops (Serrano-Ruiz *et al.*, 2011; Soccol *et al.*, 2017). Agricultural residues like corn stover, sugarcane bagasse, coffee pulp, and corn cobs are often burnt for energy, sent to landfills, or sometimes used as cattle feed (Rodrigues Mota *et al.*, 2018; Van Wyk, 2001). Municipal solid wastes like paper, food, and yard scraps as well as sawdust and other waste from the pulp and paper industries are likewise underutilized (Saha, 2003).

Three types of carbon-based polymer make up lignocellulose: Cellulose, hemicellulose, and lignin along with various minerals and proteins. The hemicellulose and

lignin form a matrix around the cellulose microfibrils (Nieves *et al.*, 2015). Different sources such as hardwoods, softwoods, municipal, or agricultural waste have varying compositions of these component polymers (Jensen *et al.*, 2017). Lignocellulosic biomass can be treated to create many value-added chemicals; alternatives which can be used as the starting compounds for the biosynthesis of many synthetic polymers civilization has come to rely upon (Isikgor and Becer, 2015). The component monomers can be fermented to ethanol, acetic acid, or lactic acid; simple starting materials which can be chemically transformed into commodity chemicals used as building blocks for antifreeze, paints, pharmaceuticals, or polyester resins (Serrano-Ruiz *et al.*, 2011). They can also be further processed into value-added chemicals for polymer precursors to biodegradeable and biocompatible materials for packaging or prosthetics (Van Wyk, 2001).

To reduce lignocellulose into fermentable sugars, the substrate must be reduced in size, subjected to pretreatment, and enzymatically hydrolyzed (Zhang *et al.*, 2006). At present, converting lignocellulosic biomass to its more useful components in a cost-effective manner is a challenge for several reasons; recalcitrance, substrate heterogeneity, and feedback inhibition (Behera *et al.*, 2017). Cell wall recalcitrance refers to its remarkable resistance to chemical degradation, which has been shaped by long-range coevolution between photosynthetic organisms, herbivores and decomposers (Rodrigues Mota *et al.*, 2018). The strength of cellulosic materials is determined by its structural organization and chemical composition. Crystalline cellulose microfibrils provide the primary structure and strength to cell walls. The cellulose is embedded in a matrix of hemicellulose. Lignin coats and impregnates the polysaccharide networks, providing rigidity and strength (Jordan *et al.*, 2012). The lignin cross-linkages must be separated via a pretreatment step before the cellulose and hemicellulose can be processed into useful monomers (Brown and Chang, 2014).

Each pretreatment step, whether chemical, physical, or enzymatic increases the energy cost of utilizing woody feedstocks (Rashamuse *et al.*, 2017). After it has been mechanically ground down, pretreatment approaches include extreme physiochemical environments with some combination of high or low pH, high temperature, pressure, or salt concentration (Sanderson, 2011). These treatments can separate the lignin fraction, decrease crystallinity, and increase the surface area exposed for further treatment (Guerriero *et al.*, 2015). In addition to allowing access to the polysaccharides, delignification is appealing because lignin is emerging as a potential feedstock for the production of aromatic chemicals (Machas *et al.*, 2019). Lignin is a complex aromatic heterogeneous polymer which makes up approximately 30% of the fixed carbon in nature and makes up 15-25% of municipal solid waste (Raychoudhury *et al.*, 2013; Van Wyk, 2001). It has been considered a major industrial by-product for at least seventy years and has potential for additional applications (de Gonzalo *et al.*, 2016).

Though it is usually burned for its energy content, lignin utilization needs to increase in order to make bio-refineries more cost-competitive with petroleum (Varman *et al.*, 2016). Because lignin is made up of aromatic compounds, the derived products of its breakdown can be toxic to organisms and interfere with further hydrolysis and fermentation steps (Gírio *et al.*, 2010). The varied composition of different lignocellulosic feedstocks and the choice of pretreatment step means there is a wide variation in the amount of toxic phenols generated by the lignin separation. If organisms or enzymes are to be applied to these mixtures, they must be able to withstand industry-relevant production standards (Allgaier *et al.*, 2010).

In biological systems, some organisms respond to this stress using metabolic detoxification pathways or by producing laccase enzymes or peroxidases to oxidize the phenols (Nicolaou *et al.*, 2010). Other ways organisms cope with aromatic or lipophilic

compounds include upregulation of efflux pumps and other methods of membrane modification (Machas *et al.*, 2019). After the lignin is removed via a pretreatment process, the holocellulose fraction remains, made up of the polysaccharide components: hemicellulose and cellulose (Zhang *et al.*, 2006).

The carbohydrate potential of biowastes can be exploited after solubilization to its component sugars (Van Wyk, 2001). Saccharification is the process of breaking down polysaccharides such as starch or cellulose into its component sugars. Polysaccharides could be released via saccharification into the component sugars for fermentation and subsequent use as carbon-neutral biofuels or feedstocks for other bio-based biorefinery programs (Marriott *et al.*, 2015). The holocellulose requires a complex suite of enzymes to break down its components for further hydrolysis (Rashamuse *et al.*, 2017). Hemicellulose makes up about 15-35% of plant cell walls and has the potential to be a valuable source of fermentable sugars (Hongoh, 2011). It is a highly branched polysaccharide, comprised of varied pentose and hexose units, such as xylan, mannan and arabinose, along with sugar acids which together form cross-linking glycans to stabilize the cellulose microfibrils (Rodrigues Mota *et al.*, 2018). The diverse structure of hemicellulose requires a variety of enzymes to degrade it into sugars: core enzymes to cleave the backbones and ancillary enzymes to depolymerize the hemicellulose and relieve steric hindrances to saccharification. As depolymerization occurs, different linkages are exposed that glycoside hydrolases can access (Bhattacharya *et al.*, 2015).

After some combination of pretreatment and enzymatic hydrolysis, saccharification results in a mix of soluble sugars in solution. Depending on the specifics of the pretreatment, inhibitory compounds derived from hemicellulose can also be released, such as furfurals or weak acids (Behera and Ray, 2016). These molecules can interfere with fermentation or further processing of the product (Nicolaou *et al.*, 2010). Side-products notwithstanding, the sugar mixture resulting from enzymatic hydrolysis is

still a challenge to ferment efficiently, as commercial fermentation organisms do not easily metabolize pentoses (Marriott *et al.*, 2015). Some bacteria can utilize mixed sugars, producing acids or solvents rather than ethanol (Saha, 2003). Pentoses are challenging to ferment, as microbes in monoculture have been shown to suffer from hexose repression, wherein glucose is preferred over xylose metabolism and cells divert cellular resources toward one metabolic pathway over another (Bajwa *et al.*, 2009).

Cellulose is a homo-polymeric polysaccharide made up of strands of β (1→4) linked glucose molecules. The cellulose chains are arranged in parallel, with hydrogen bonds tightly linking them into the characteristic crystalline structure (Jensen *et al.*, 2017). Cellulose is challenging to break down; the glycosidic bonds and crystal structure of cellulose resists deconstruction (Juturu and Wu, 2014). Conversely, the homogeneity of cellulose allows it to be degraded into glucose units, which can be used as starting compounds for biosynthetic processes or fermented for biofuel (Van Wyk, 2001).

Physical and chemical approaches can be taken to pretreatment of the biomass. These treatments add to the cost of conversion and make the component polymers more accessible for cellulase enzymes, which are also costly to commercialize (Jordan *et al.*, 2012). These are the types of challenges that biological systems have impressively tackled to overcome these barriers, and studying these systems can aid efforts to likewise utilize these fuels. Investigating methods of enzymatic saccharification can supplement and maybe one day supplant mechanical and physical pretreatments.

1.2 Carbohydrate-active Enzymes

Due to the varied structures polysaccharides can take on, a multiplicity of enzymes has evolved to tackle the breakdown of lignocellulosic material. Glycoside hydrolases break bonds between sugars. Previously, enzyme classification was done by grouping enzymes by substrate and the type of reaction they performed into enzyme commis-

sion numbers (EC numbers). This system allows enzymes to be differentiated by the reactions they perform. Evolutionary convergence from different lineages to catalyze the same reaction results in non-homologous isofunctional enzymes to be grouped within the same EC number.

To supplement the EC system of classifying enzymes, glycoside hydrolases were grouped into families based on amino acid sequence similarities (Henrissat, 1991). This was proposed to include structural and mechanistic properties of related enzymes. As of May 2019, there are 165 GHFs, which are numbered sequentially upon discovery. This organization takes into account convergent or divergent evolutionary events. Each GHF represents a group of evolutionarily related enzymes. Though an enzyme may diverge from the reaction catalyzed by its homolog, it remains in the same family.

These families are further organized into phylogenetically related clans, ordered alphabetically from GH-A through GH-R. Each clan is united in the tertiary structure of the constituent families. The members of each clan share three-dimensional structural motif and catalytic mechanism (Durand *et al.*, 1997). Some of the larger families have been broken down into subfamilies. Information on these proteins and their classification is kept updated and available online at the Carbohydrate-Active Enzymes database (CAZy; HTTP://WWWcazy.org). This classification system has been found useful and expanded to accommodate other enzyme classes: glycosyltransferases, polysaccharide lyases, carbohydrate esterases, and auxiliary activities. As of May 2019 within glycoside hydrolases alone there are 645752 enzymes classified into families and an additional 10201 that have not yet been classified (Lombard *et al.*, 2014).

1.3 Termite Biology and Symbioses

Termites are eusocial insects, with specialized worker, soldier, and reproductive classes (Aanen and Eggleton, 2017). The termite lineage evolved eusociality before hymenopterids, but are much less well-studied (Korb *et al.*, 2015). Termites are relatives of cockroaches, having evolved from omnivorous ancestors and acquired a suite of protist symbionts which allowed them to specialize in a diet of cellulosic materials (Radek *et al.*, 2018). For convenience, the term "lower termite" refers to the deeper-branching families: Mastotermitidae, Archotermopsidae, Hodotermitidae, Stolotermitidae, Kalotermitidae, and Rhinotermitidae. These families share the characteristic of harboring symbiotic protists in their hindguts; it is thought that the family Termitidae subsequently lost their protist symbionts. Termitidae comprises a clade referred to as the "higher termites".

Termites are major decomposers of plant matter worldwide, being found on every continent except Antarctica (Brune, 2014). Termites are broadly categorized as drywood, dampwood, or humus-feeding. Various termites feed on wood, leaves, humus, fungi, dung or soil, though nearly all lower termites are wood feeders (Abdul Rahman *et al.*, 2015). Lower termites are drywood and dampwood feeding, with the exception of the family Hodotermitidae, which mainly eat dead grasses.

Termite biomass is highest in the tropics and subtropics, where they play a major role in soil engineering and carbon cycling (Jouquet *et al.*, 2011). The termitosphere offers many ecosystem services, though a few species are notorious pests of man-made buildings. Among the lower termites are found the most nuisance termite species which cause damage to woody buildings and structures. Family Mastotermitidae forms the deepest branch and retains the most ancestral features, and *Mastotermes darwiniensis* is among the world's most destructive termites (Watanabe *et al.*, 2006).

Coptotermes formosanus (family Rhinotermitidae) is a problem in Japan and the United States, though *Reticulitermes* (Rhinotermitidae) is the most destructive genus in the United States (Arakawa *et al.*, 2009; Baker and Marchosky Jr, 2005).

In the termite gut, cellulose and hemicellulose are degraded within 1 day, though lignin is excreted with little modification (Ke *et al.*, 2010). Mastication reduces wood into small pieces, serving as a mechanical pretreatment of the woody material. The salivary glands excrete termite-endogenous cellulases to begin the process of degradation. Depending on the species, the gut generally holds about 1μL of material, including the symbionts and ingested wood. In higher termites, the bacterial community handles H_2 metabolism, CO_2-reductive acetogenesis, and N_2 fixation (Warnecke *et al.*, 2007). The lower termites have a more simply formed gut, consisting of a foregut, midgut, paunch, colon, and rectum. The paunch is the dilated portion of the hindgut, densely populated with flagellates. The higher termites include examples that feed on grass, lichen, litter, soil, or wood, and lacking protist symbionts they have a longer, narrower gut with regions of distinct pH. (Hongoh, 2011).

Termites molt on a regular basis, shedding their gut microbiota along with their exoskeleton. Given the fundamental nature of the interaction between termites and their symbionts, it is important to maintain. The specialized protists inhabiting the lower termite hindgut are not known to encyst and are not believed to survive outside of their termite hosts, instead they are transferred between members of the same colony (Noda *et al.*, 2007). To replenish the symbionts, lower termites regularly inoculate themselves via proctodeal trophallaxis or coprophagy (Scharf *et al.*, 2017). Vertical inheritance strongly shapes the microbiome of both higher and lower termites (Abdul Rahman *et al.*, 2015).

Termites have traditionally been classified as an insect order, Isoptera, and are now understood to be phylogenetically nested within cockroaches. The sister lineage

to termites is the subsocial wood-feeding cockroach, *Cryptocercus*, which lives in family groups, feeding on woody materials (Chouvenc *et al.*, 2016). Together termites and cockroaches form Blattodea which includes major wood decomposers and pests worldwide (Evangelista *et al.*, 2019).

Distinguishing lower termites from higher termites (Termitidae), lower termites and the wood roaches are notable for having a symbiotic relationship with protist symbionts within their enlarged hindgut chamber. These flagellate protists are important for the breakdown of lower termites' cellulose-intensive diet (Noda *et al.*, 2018). This is an ancient association, dating back to the divergence of *Cryptocercus* and the lower termites (Bourguignon *et al.*, 2015). A Miocene termite amber fossil was found with the symbiont community of protists, spirochetes, and other bacteria having been preserved from 20 million years ago (König *et al.*, 2013). Symbionts can be obligate, requiring the host species to survive, or can be facultatively symbiotic, able to survive both inside the host and elsewhere (Bright and Bulgheresi, 2010). In general there is a symbiotic system comprised of the termite host, a core group of obligate flagellates and bacteria, as well as secondary facultative symbionts (Duarte *et al.*, 2018). The termite provides a habitat and a food source for its symbionts, while the symbionts assist with nitrogen cycling, immune function, and electron transfer in hindgut fermentations (Brune, 2014; Hussain *et al.*, 2013).

When a new colony is founded, the king and queen bring with them the full complement of symbionts that will be shared with their colony. Reproductive alates in *Nasutitermes*, a higher termite, were shown to have at least as diverse microbiota as the non-reproductive classes (Diouf *et al.*, 2018). However, compared to their workers, lower termite alates have been shown to have lower numbers of protists in their gut while preparing to swarm (Benjamino and Graf, 2016). This is a vertical method of transmission, from one generation to the next. With this method of transmission, it

is still unknown at what point termites acquired which parabasalid symbionts and associated lignocellulolytic abilities.

The lower termite microbiome includes bacteria, archaea, and protists. This complex symbiotic environment is understood to be the basis of their success in degrading woody material (Cleveland, 1924a). The protist symbionts are a major part of the lower termite microbiome, accounting for up to one-third the total insect volume (Inoue *et al.*, 2007). Indeed, a study on protist diversity in lower termites has suggested that greater protist diversity may be a driver of adaptation in invasive termites (Duarte *et al.*, 2018). It has also been proposed the failure of entomopathogens as a source of termite control may be due to the strong symbiotic collaboration between termites and their microbiome (Peterson and Scharf, 2016). Reflecting the vertical transmission, the particular suite of protists found within a termite species is characteristic of that species.

Phylogenetic studies have been undertaken to attempt to clarify the evolutionary history of the protists with respect to their termite hosts. An early study examining cospeciation in the triplex symbiosis involving protists in the genus *Pseudotrichonympha* and their hosts in the termite family Rhinotermitidae concluded the protists and their bacterial symbionts showed almost complete codivergence with their hosts (Noda *et al.*, 2007). A study examining the phylogeny of *Trichonympha* from various termite hosts, using small subunit rRNA gene sequences, did not provide strong support for strict co-speciation with their hosts (Boscaro *et al.*, 2017).

Protists are unicellular eukaryotes, found within all major branches of the Eukaryotic tree of life (Pawlowski *et al.*, 2012). Protists living in the termite hindgut are in the superorder Excavata, all from the phylum Parabasalia or the order Oxymonadida within the the phylum Preaxostyla (Hongoh, 2011). In the literature termite-associated protists are often referred to simply as flagellates. The relation-

ship between host and symbiont is further complicated by the historical reliance on morphology to build parabasalid phylogeny, and molecular evidence continues to improve understanding of their relationships. Though the phylogeny of Parabasalia is not fully resolved, it is clear that termite-associated protists are not monophyletic and there were multiple occurrences of free-living parabasalids colonizing the termite hindgut (Čepička *et al.*, 2017). It is unknown how and when select parabasalids evolved the ability to digest wood, as this ability is not seen in their free-living relatives. There are few protists identified to which a particular GHF enzyme has been attributed (see Table 1.1).

In general, parabasalid termite symbionts evolved to be very large compared to their free-living relatives, with multiple flagella. This is a hypermastigote phenotype, adapted for phagocytosis of wood particles. Long thought to be evolutionarily related, molecular evidence has shown hypermastigotes to be polyphyletic (Čepička *et al.*, 2017). The hypermastigote phenotype is understood to have evolved multiple times from smaller, structurally simpler parabasalids (James *et al.*, 2013). In contrast, the oxymonad termite symbionts tend to be either highly motile or attach themselves to the termite hindgut wall (Brune, 2014).

Within the anoxic center of the termite hindgut protists engulf the wood particles and release acetate, carbon dioxide and hydrogen gas. This anaerobic fermentation takes place within specialized organelles derived from mitochondria, called hydrogenosomes, which are found within of the parabasalid symbionts (Inoue *et al.*, 2007). The oxymonad symbionts lack hydrogenosomes and their metabolism is largely unexplored (Tamschick and Radek, 2013). Termite protists are understood to be are anaerobic, increasing the partial pressure of O_2 is used in studies to kill off the protists without harming the termites. Without their protist symbionts, termites continue to feed on wood but starve within a few days (Cleveland, 1924b).

Aerotolerant *Enterococcus* and *Lactococcus* bacteria have been isolated from termite and other insect hindguts, with high potential rates of O_2 reduction (Brune and Friedrich, 2000). The microbial community quickly metabolizes any O_2 that diffuses from the termite gut epithelium into the lumen (Ebert and Brune, 1997). A combination of methanogens and homoacetogens are spatially organized to make use of the nutrient and oxygen gradient within the lumen, being largely anoxic toward the center and increasingly oxic radially (Tholen and Brune, 2000). The methanogens are aerotolerant and are densely located on the gut epithelium, where they metabolize H_2 and CO_2 (Ohkuma, 2003). The bacterial community is persistent within species, but relative abundances can fluctuate with diet changes (Waidele *et al.*, 2017). *Treponema* spirochaete ectosymbionts of oxymonads have been shown to encode genes for reductive acetogenesis from H_2 and CO_2.

In this complex environment, the flagellates even support obligate mutualistic endosymbiotic bacteria which are not found outside the termite gut (Reuß *et al.*, 2015). Most gut protists have been shown to have endosymbiotic Bacteroidales within them or ectosymbiotic spirochetes firmly attached onto their cell surface (Noda *et al.*, 2009). Endosymbiotic methanogens are also found within termite gut protists (Noda *et al.*, 2009).

Bacterial inhabitants of termite and wood-feeding cockroaches include several phyla, including Actinobacteria, Actinomycetes, Bacteroidetes, Deltaproteobacteria, Firmicutes, Proteobacteria, Spirochetes, and Verrucomicrobia (Berlanga, 2015). There are also three candidate bacterial phyla, lineages which are found only within termites and have not been found elsewhere in surrounding environments (Ohkuma, 2008). It has been shown that gut bacteria tend to be specialized symbionts, consistent within a genus of termites (Hongoh *et al.*, 2005). Other studies have indicated that the bacterial termite gut inhabitants are more variable over evolutionary time

than the protist inhabitants (Waidele *et al.*, 2017). The bacterial inheritance is also an area of active study. For example, the phylogeny of *Cryptocercus* bacterial endosymbionts has been shown to be congruent with that of the host species (Che *et al.*, 2016). Different protist lineages acquired their bacterial endo- and ecto-symbionts independently from the host hindgut bacteria and once the association is established can likewise cospeciate (Tai *et al.*, 2016). Investigating the roles of microbes in higher and lower termite digestive activities allows us to examine the potential history of the gain of wood digestion in termite-associated protists and the loss of such protists in higher termites.

1.4 Glycoside Hydrolases in Lower Termites

In order to clarify the relationship between termites and their protists in relation to lignocellulosic digestion, this thesis examines several previous studies documenting glycoside hydrolases found in specific protists within the termite hindgut (see Table 1.1). Studies which rigorously identify and characterize the enzymes of interest allow a more confident interpretation of expression and functionality *in vivo*. The functionality and expression of the enzymes being examined is especially important when approaching a project that relies heavily on DNA and RNA sequences. Because the termite protists are so difficult to establish in culture, alternate methods have been utilized to determine which protists are responsible for specific enzyme activities in the lower termite hindgut. Given these data are so important in building and interpreting the phylograms, they will be reviewed here briefly.

In 2002, Watanabe *et al.* isolated endo-β-1,4-glucanases from the hindgut extract of *Coptotermes lacteus* and confirmed them to be different than the endogenous endoglucanases produced by the termite salivary glands. Using polymerase chain reaction (PCR)-based cloning on the hindgut and individual protists isolated from *Cop-*

Table 1.1: Protist-specific Glycoside Hydrolases

Author	GHF	Protist	Termite
Watanabe et al. 2002	GHF7	*Holomastigotoides mirabile* *Pseudotrichonympha grassii*	*(Coptotermes sp.)*
Nakashima et al. 2002	GHF7	*Pseudotrichonympha grassii*	*(Coptotermes sp.)*
Li et al. 2003	GHF45	*Deltotrichonympha nana*	*(M. darwiniensis)*
Inoue et al. 2005	GHF5	*Cononympha leidyi*	*(C. formosanus)*
Arakawa et al. 2009	GHF11	*Holomastigotoides mirabile*	*(C. formosanus)*

totermes formosanus hindgut, they determined the sequences were GHF7s from the trichonymphid *Pseudotrichonympha grassii* and the spirotrichonymphid *Holomastigotoides mirable*. These were then expressed in *Escherichia coli* and endoglucanase (EC 3.2.1.4) activity was verified by screening in sodium carboxymethylcellulose (CMC) in agarose, tetrazolium blue confirmed the release of glucose equivalents. Importantly, the mRNA recovered was poly-A-tailed, a trait specific to eukaryotes. They also ran a negative RT-PCR control to preclude bacterial contamination (Watanabe *et al.*, 2002).

Also in 2002, Nakashima *et al.* designed primers using the conserved catalytic center of GHF7 with the amino acid sequences of the N-terminus of the cellulase component of *Coptotermes formosanus* hindgut fluid and used these to amplify, clone, and sequence the cellulases. The poly-A-tail-based cloning method was used to exclude the possibility of extra- or intra-cellular archaea or bacteria associated with the flagellates in the hindgut. RT-PCR was run using using a single cell of each flagellate species as a template to determine the origin of the obtained genes, and only the *Pseudotrichonympha grassii* sample was amplified to the expected size. Multiple sequence alignments showed homology with GHF7 cellobiohydrolases. The sequence was then sub-cloned into an *Escherichia coli* vector and expressed and screened on CMC-LB plate and stained with Congo red to confirm expression (Nakashima *et al.*, 2002).

Using micro-manipulated nuclei from the cristamonad protists *Koruga bonita* and *Deltotrichonympha nana* isolated from the lower termite *Mastotermes darwiniensis* Li *et al.* 2003 used a nested RT-PCR approach to sequence GHF45 sequences. These were complete with a poly-A tail, a start and stop codon, and signal peptides similar to cellulases previously sequenced from *Reticulitermes speratus* protists (Li *et al.*, 2003; Ohkuma *et al.*, 2000).

The next GHF attributed to a protist species was in 2005. A GHF5, subfamily 2, was attributed to *Spirotrichonympha leidyi* from *Coptotermes formosanus* (Inoue *et al.*, 2005). Because recent molecular and morphological evidence demonstrates the type species *Spirotrichonympha* from *Reticulitermes* is phylogenetically distinct, previously described *Spirotrichonympha* isolated from host genera *Coptotermes* and *Heterotermes* has been reinstated to its original Japanese genus description *Cononympha* and will be referred to as such going forward (Jasso-Selles *et al.*, 2017). Inoue *et al.* 2005 purified poly-A-mRNA from the gut contents of *Coptotermes formosanus* and prepared a recombinant phage library. This library was screened for cellulolytic activity against CMC using Congo Red staining. The positive plaques were transferred to a plasmid vector for amplification and sequencing. The resulting sequence included a poly-A-tail and was used to make gene-specific primers, which were then applied to PCR-amplified genomic DNA for each protist, confirming the organismal source was *Cononympha leidyi*. This was further confirmed using whole-cell in situ hybridization, showing the enzyme localized within *Cononympha leidyi*. The enzyme was also heterologously expressed in *E. coli* and enzyme activity was characterized (Inoue *et al.*, 2005).

Arakawa *et al.* 2009 did a functional screening for xylanases from *Coptotermes formosanus* and verified their role in the gut via the elution profile. The N-terminal amino acid sequences were used to design primers, from which the corresponding

cDNA were successfully cloned. RT-PCR was used to confirm the xylanase was expressed by the spirotrichonymphid *Holomastigotoides mirabile* (Arakawa *et al.*, 2009).

Warnecke *et al.* 2007 undertook the first hindgut metagenomic study of a termite, the higher termite *Nasutitermes ephratae*, using the third proctodeal segment of its hindgut paunch. They also did a proteomic analysis of the clarified gut fluid for cellulose and xylan hydrolysis. These data provide a higher termite microbiota to contrast with the carbohydrate-active enzymes found in lower termites (Warnecke *et al.*, 2007).

1.5 Approach

This study sought useful and interesting new insights into this complicated symbiotic system, particularly in the areas of evolution, speciation, symbiosis and protein diversity. In particular, to investigate the evolutionary history of the termite-protist association and an exclusively wood-eating lifestyle, this thesis compares the glycoside hydrolases found in published termite-associated protists with new data generated in the Gile lab (see Table 1.2). Single-cell transcriptome sequencing was done on two protists: *Trichonympha* isolated from the hindgut of *Hodotermopsis*, a lower termite, and *Lophomonas* isolated from *Periplaneta*, the American cockroach. For comparison with a free-living relative a transcriptome from *Pseudotrichomonas* was sequenced from culture.

Trichonympha is part of the order Trichonymphida, comprised entirely of protists found in cockroaches and termites. For example, *Trichonympha* can be found in *Cryptocercus* cockroaches, as well as the lower termites *Hodotermopsis* and *Reticulitermes* (Cleveland *et al.*, 1934). *Lophomonas* is part of clade Lophomonadida, sister to Trichonymphida. Because *Lophomonas* is found in the omnivorous *Periplaneta* American cockroach, comparison provides an opportunity to contrast a protist

living in a lignocellulosic environment with one that potentially lacks that specialization (Koidzumi, 1921; Gile and Slamovits, 2012). *Pseudotrichomonas* is part of the clade Trichomonadida, distinct from both Lophomonadida and Trichonymphida. Unlike the other two single-cell transcriptomes investigated, *Pseudotrichomonas* is a free living protist and therefore has different selective pressures and requirements for survival. Seeking GHFs in these transcriptomes will supplement those GHF representatives previously identified in *Holomastigotoides* and *Cononympha*, both of which are in the termite-cockroach-symbiont order Spirotrichonymphida. Spirotrichonymphida is outgroup to Trichomonadida, Lophomonadida, and Trichonymphida (Čepička *et al.*, 2017).

Additionally, unpublished *Coptotermes formosanus* whole-gut shotgun-sequenced metagenome assembled reads were provided by Gillian Gile for analysis. The data was paired end 2 by 150, 2 lanes in hi seq. This dataset includes bacterial, archaeal, protist, and termite sequences. Though this study focuses on the protist role in lignocellulosic degradation, the bacterial presence is expected to shed light on the origins of cellulases and hemicellulases in termite protists, particularly in those cases where a GHF contains both bacterial and protist sequences.

Table 1.2: Unpublished Datasets

Author	Data	Termite
Gile unpublished	Metagenome	*Coptotermes formosanus*
Gile unpublished	Single-cell transcriptome	*Trichonympha (Hodotermopsis)*
Gile unpublished	Single-cell transcriptome	*Lophomonas (Periplaneta)*
Gile unpublished	Transcriptome	*Pseudotrichomonas (in mixed culture)*

Chapter 2

METHODS

2.1 Transcriptome Assembly

The raw reads were retrieved for transcriptome shotgun assembly (TSA) or short-read archive (SRA) from the National Center for Biotechnology Information (NCBI). 454 pyrosequencing metagenome data for *Coptotermes formosanus* was retrieved from SRA under accession SRX105331 (Xie *et al.*, 2012). For *Reticulitermes*, Illumina short reads were retrieved as follows: *R. flavipes*, SRA accession SRX565295 and SRX565296; *R. grassei*, SRA SRX565297-SRX565305; and *R. lucifugus*, SRA SRX565306 and SRX565307 (Dedeine *et al.*, 2015; Gayral *et al.*, 2013). *Coptotermes gestroi* 454 reads from soldier and worker were retrieved under SRX854076 and SRX854079, respectively (Franco Cairo *et al.*, 2016). Important datasets are given in Table 2.1.

Table 2.1: Lower Termite Hindgut Omic Databases

Author	Data	Termite
Todaka et al. 2010	Expressed Sequence Tags GHF5, GHF10, GHF7, GHF11, GHF8, GHF45, GHF43, GHF62	*Reticulitermes speratus* *Hodotermopsis sjostedti* *Neotermes koshunensis* *Mastotermes darwinensis* *Cryptocercus punctulatus*
Xie et al. 2012	Metagenome	*Coptotermes formosanus*
Hussain et al. 2013	Expressed Sequence Tags	*Coptotermes formosanus*
Dedine et al. 2015	Metatranscriptome	*Reticulitermes lucifugus* *Reticulitermes flavipes* *Reticulitermes grassei*
Franco Cairo et al. 2016	Metatranscriptome	*Coptotermes gestroi*

The reads in the FASTQ files were trimmed for quality with Trimmomatic version 0.38 using standard parameters, removing reads with a length of <25 bp and those with a mean quality <15 in a sliding window size of 4. For each set of SRA data, the appropriate primers for the sequencing method were specified for trimming (e.g. TruSeq2 primers for GAII machines, etc.) (Bolger *et al.*, 2014; MacManes, 2014). FastQC was run for quality checks before and after trimming to ensure quality after trimming was acceptable and no adapters remained in the data. The pooled reads for each species were assembled using Trinity version 2.4.0 (Haas *et al.*, 2014). The generated contigs and the remaining orphan sequences were used as databases in local blast searches.

2.2 Data Mining and Exploration

The protein sequences (accession numbers given in Todaka *et al.* 2010) were used for local BLAST searches of the assembled transcriptomic data; amino acid sequences were used as queries against the assembled nucleotide databases. Using NCBI Basic Local Alignment Search Tool (BLAST) command line application TBLASTN, potential homologs were retrieved with a cutoff E-value threshold of ≤1e-05. Results were compared against the NCBI non-redundant (NR) protein sequence database via reciprocal BLAST, to verify the best hit for each sequence was the GHF expected and aligned with an NCBI Conserved Domain Database entry for that GHF.

GHF transcript contig hits identified by reciprocal blast were inspected manually to check identity and reading frame, then were translated to amino acid sequences to be aligned. The amino acid sequences were trimmed to the annotated protein family (Pfam) motifs, where available in NCBI (Finn *et al.*, 2014). For those sequences annotated, Table 2.2 lists the Pfam motifs which sequences were trimmed to. Some alignments included long bacterial sequences lacking the Pfam motif annotation and

Table 2.2: Number of Taxa and Sites in GHF Alignments

	Taxa	Initial	Trimmed	Pfam Annotation
GHF5	233	2178	322	PF00150
GHF7	226	851	467	PF00840
GHF8	94	1104	245	PF01270
GHF10	173	3117	303	PF00331
GHF11	123	1593	197	PF00457
GHF43	161	1287	200	PF04616
GHF45	106	715	179	PF02015

so were aligned untrimmed, causing the overall alignment length to be inflated (see Table 2.2). The initial alignments also appear inflated because the inclusion of a few sequences with large insertions.

After translation, new and previously published sequences were aligned with MUSCLE version 3.8.1551 using default settings: a hydrophobic window size of 5; clustering with UPGMB, a combination of unweighted pair group method with arithmetic mean (UPGMA) and neighbor joining; iteration 1 using k-mer clustering distance measure "kmer6_6"; iteration 2 used the bipartite refinement distance "pctid_kimura"; tree scoring done by the sum-of-pairs score; with a maximum of 16 iterations (Edgar, 2004). The alignment was then then refined by eye in Aliview version 1.25 (Larsson, 2014). To remove uninformative insertions, excess gaps, and ambiguously aligned regions the aligned proteins were trimmed using trimal, "automated1" setting to automatically select the best trimming strategy for each alignment. This setting is a heuristic method which computes specific score thresholds based on cumulative graphs of column gap and similarity scores (Capella-Gutiérrez *et al.*, 2009).

2.3 Phylogenetic Analysis

To further confirm GHF membership and complete the initial exploration of the data mining results, minimum evolution (ME) trees were inferred using FastTree

version 2.1.10 with Shimodaira-Hasegawa (SH) tests providing local branch support (Price *et al.*, 2009). FastTree consists of three phases. Initial topology is done with a combination of fast neighbor-joining with relaxed neighbor-joining. The refining topology step is a balanced minimum evolution phase; mixing nearest-neighbor interchanges with subtree-prune-regraft moves. The final stage is approximating maximum likelihood. For the ML step, FastTree uses the Jones-Taylor-Thornton (JTT) model of amino acid evolution and uses a "CAT approximation" single rate of evolution for each site to account for the varying rates of evolution across sites (Price *et al.*, 2010). Trees were rooted using outgroups where family justified it, including diverse taxa where possible (Davies *et al.*, 2018). Figure generation was done with online visualization software Evolview version 3 (He *et al.*, 2016).

GHF11 and GHF45 were selected for further phylogenetic analysis to evaluate support for their multiple distinct clades from lower termite guts. For each family the amino acid sequences were aligned and trimmed as described above. Maximum likelihood (ML) phylogeny estimation was carried out using randomized axelerated maximum likelihood for high performance computing RAxML-HPC version 8.2.12 (Stamatakis, 2006). Support for ML topologies was assessed by percentage of 1000 total bootstrap replicates. ProtTest version 3.4.2 was used to determine the best amino acids replacement models and analysis parameters. ProtTest recommended WAG+G+F for GHF11 and LG+I+G for GHF45. For comparison, ML was also run using the substitution models selected by Todaka *et al.* 2010; LG+G+I+F and WAG+G for GHF11 and GHF45, respectively. The final ML optimization likelihoods for trees built from both models were compared and the higher scoring tree was used for each analysis; for GHF45 WAG+G, while for GHF11 WAG+G+F had the higher final ML optimization likelihood.

Bayesian inference (BI) analysis was carried out on amino acid alignments for GHF11 and GHF45 using Mr.Bayes version 3.2.7a (Ronquist and Huelsenbeck, 2003). BI analyses were run using a mixed substitution model estimation with four chains, three heated and one cold. Posterior probabilities were generated by sampling a tree every 100 generations. The first 25% of trees generated were discarded in the burn-in phase. Before generating a concensus tree, GHF11 was run for 20 million generations, GHF45 was run for 15 million generations. Convergence was assessed via average standard deviation of split frequencies (ASDSF) $\leq$0.02 due to the many taxa. This ASDSF is useful for mainly assessing the most well-supported parts of the tree (Ronquist *et al.*, 2012).

Chapter 3

RESULTS AND DISCUSSION

3.1 Glycoside Hydrolase Family 5

3.1.1 GHF5 Background and History

GHF5 is part of clan GH-A, which is the largest of the clans containing the most glycoside hydrolase member-families. GH-A contains two of families reported to be found in the lower termite hindgut, GHF5 and GHF10 which are investigated in this thesis. Clan GH-A also contains GHF1, GHF2, GHF17, GHF26, GHF30, GHF35, GHF39 GHF42, GHF50, GHF51, GHF53, GHF59, GHF72, GHF79, GHF86, GHF113, GHF128, GHF147, GHF148, GHF157, and GHF158 (Lombard *et al.*, 2014). They share a conserved $(\beta/\alpha)_8$ fold catalytic module and catalyze hydrolysis of glycosidic bonds, with two glutamate residues implicated as the catalytic nucleophile and proton donor, respectively. They take part in a double-displacement hydrolysis mechanism, resulting in the anomeric carbon retaining its stereochemistry (Zhang *et al.*, 2008). Though amino acid sequence can vary considerably between enzymes within a clan, the three dimensional structures are well-conserved (Henrissat and Davies, 1997).

GHF5 was the first glycoside hydrolase family described, being given the name glycoside hydrolase family A, but was later renamed under a numerical naming scheme (Henrissat *et al.*, 1989; Henrissat and Bairoch, 1993). GHF5 is the largest of all glycoside hydrolase families, as of May 2019 GHF5 contains 13792 members, of which 563 are biochemically characterized with EC numbers (Lombard *et al.*, 2014). Among these members, only seven amino acid residues are strictly conserved (Collins

et al., 2005). GHF5 is among the most diverse groups of glycoside hydrolases, with 51 subfamilies (Aspeborg *et al.*, 2012). GHF5 endo-acting and exo-acting glycoside hydrolases cover a variety of specificities; its members including cellulases, xylanases, arabinoxylanases, mannosidases, licheninases, and chitosanases.

Endo-acting hydrolases randomly cleave internal bonds of polysaccharides, creating new chain ends for other enzymes to act on. Many GHF5 members are endo-acting glucanases, which hydrolyze D-glucose polysaccharides, include examples of cellulase, laminarinase, licheninase, glucanohydrolase, and xyloglucan-specific glucanase. Other endo-acting GHF5 members work on xylose-, mannose-, arabinogalactose, or glucosamine-substituted polysaccharides. Exocellulases degrade cellulose from either the reducing or non-reducing ends of the polymer, generally releasing disaccharides or monosaccharides. Other exo-acting glycosyl hydrolases work on non-carbohydrate glycosyl-substituted molecules.

3.1.2 GHF5 in Termites and Across Life

Members of GHF5 are found within archaea, both anaerobic and aerobic bacteria, and within eukaryotes including fungi, nematodes, protists and insects (Aspeborg *et al.*, 2012). In termites, bacterial GHF5 cellulases have been identified within the metagenome of the higher termite *Nasutitermes*. The GHF5 found within *Nasutitermes* were confirmed to have cellulase activity via functional genomic screens and contained secretion signal peptides, which indicates the GHF cellulase activity in higher termites takes place within the luminal fluid (Warnecke *et al.*, 2007). These carbohydrate-active enzymes are being found widely across termite-protist symbioses as well, originating from the protists (Tartar *et al.*, 2009).

The first GHF5 found in a termite protist was found in *Cononympha leidyi* via functional screening and was subsequently cloned, expressed and its activity charac-

terized (Inoue *et al.*, 2005). Across termites and *Cryptocercus*, Todaka *et al.* 2010 indicated evidence of LGTs of bacterial GHF5s to termite-associated protists, determining GHF5 along with GHF7 are part of a "core enzyme set" acquired by early termite-associated protists. At the time, there were five sub-families within GHF5 and termite-associated protists were shown to have representatives nested within bacteria in GHF5 subfamilies GHF5_1, GHF5_2, and GHF5_4. Support for this topology was low, however alternative topologies without this nesting were rejected by SH tests (Todaka *et al.*, 2010a).

LGT of 16S pseudogenes has been documented between *Trichonympha* and its endosymbiotic bacteria, therefore LGT of useful cellulases is entirely plausible (Sato *et al.*, 2014). The genome of an ectosymbiont of the oxymonad protist *Dinenympha* was recently sequenced, containing GHF5_13 sequences. GHF5_13 currently lacks experimental characterization of enzyme activity (Yuki *et al.*, 2015). Additionally, Franco Cairo *et al.* 2016 found GHF5 transcripts from the xylophagous protists and bacteria inhabiting *Coptotermes gestroi* but did not make sub-family determinations within the larger GHF5 phylogeny.

GHF5 was rooted with GHF10 as they are both in clan GH-A, sharing a common ancestor and conserved $(\beta/\alpha)_8$ fold catalytic module (Cantarel *et al.*, 2009). GHF10 sequences of the following taxa comprise the root to the GHF5 phylogram: three fungi, two archaea, two bacteria, and two termite protists. GHF5 sequences were abundant in the *Trichonympha* single-cell transcriptome, and were found across several nodes of the GHF5 phylogeny (Figure 3.1). Along with previously published sequences, this work identified seven distinct sequences from the single-cell transcriptome of *Trichonympha*, along with a sequence from the assembled *Reticulitermes flavipes* metatranscriptome, and three sequences from the *Coptotermes formosanus* metagenome. *Trichonympha* GHF5 sequences were found in GHF5_1, GHF5_2, and

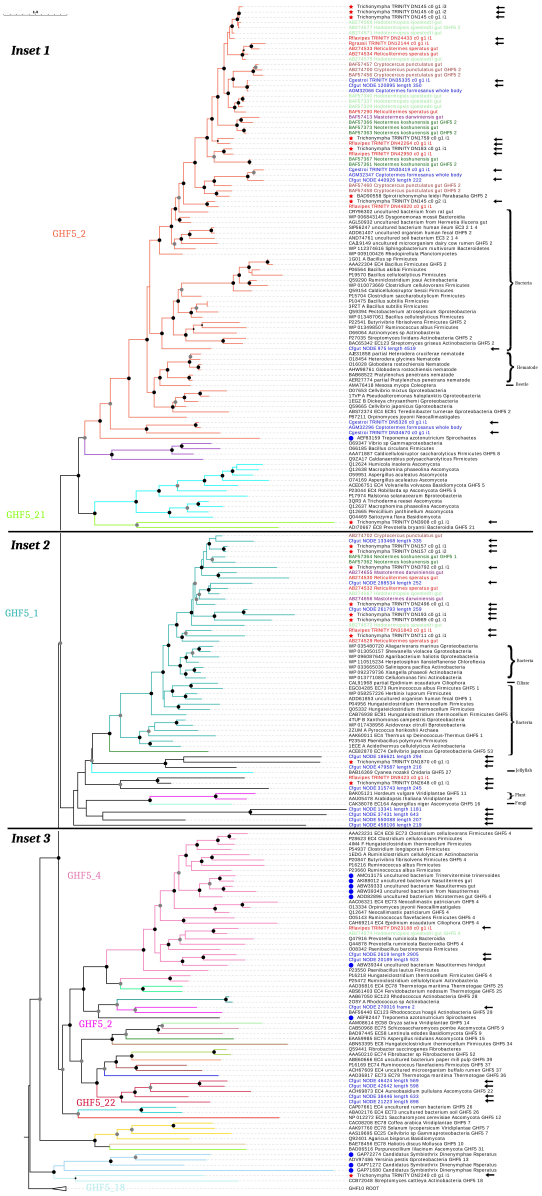

Figure 3.1: GHF5 Phylogeny Overview.
Overview of minimum evolution phylogram of GHF5 amino acid sequences, computed on FastTree v. 2.1.10. Black arrows indicate GHFs identified in this work. Font colors indicate sequences from the gut of lower termite genus: blue, Coptotermes; red, Reticulitermes; purple, Mastotermes; dark green, Neotermes; light green, Hodotermopsis; brown Cryptocercus. Dots at nodes indicate SH support values: gray dots 50-79; black dots 80-100. Leaf decorations indicate termite-associated organisms: red stars, protists; blue circles, termite-associated bacteria; brown squares, termite-associated fungi. Branch colors represent GHF5 subfamilies as determined by Aspeborg *et al.* 2012. Outgroup rooting was done with GHF10 sequences, indicated by a triangle representing the collapsed outgroup.

GHF5_4, forming monophyletic clades with previously identified termite protists in these subfamilies. This is consistent with the findings of Franco Cairo *et al.* 2016 and Todaka *et al.* 2010, finding three subgroups of protist GHF5s.

Termite protists in GHF5_1 were nested within bacterial sequences from Gammaproteobacteria and Actinobacteria (Figure 3.3). GHF5_2 had a termite protist clade nested within uncultured gut bacteria (Figure 3.2). The *Coptotermes formosanus* metagenome contributed two sequences to the termite protist clade and one sequence within the bacterial portion of GHF5_2. GHF5_2 also contains a clade of herbivorous nematodes.

Within the assembled *Reticulitermes flavipes* metatranscriptome there was an additional GHF5_4, which branched with the sole previously published termite protist sequence from *Hodotermopsis* (Figure 3.4). These termite protist sequences were nested within bacterial branches, including a clade of bacterial sequences from the higher termites *Nasutitermes*, *Trinervitermes*, and *Microtermes*. Elsewhere in GHF5_4, two sequences from the *Coptotermes formosanus* metagenome branched with a bacterial *Nasutitermes* sequence. GHF5_4 also includes a clade of anaerobic gut fungi (AGF), phylum Neocallimastigomycota.

In addition to those subfamilies already identified to contain protist termite GHF5s, the *Trichonympha* single-cell transcriptome contained a sequence branching with a *Streptomyces* GHF5_18 (Figure 3.4) and a sequence branching with a *Prevotella* GHF5_21 (Figure 3.2). Four *Coptotermes formosanus* metagenome sequences grouped with an Ascomycete GHF5_22 and one branched with an actinobacterial GHF5_29 (Figure 3.4). There were also some sequences falling within poorly resolved areas of the tree, unable to be assigned to any particular subfamily (Figure 3.3). This includes two sequences from *Trichonympha*, one from *Reticulitermes flavipes* and seven from *Coptotermes formosanus*. Notably, GHF5 transcripts were absent in the *Lophomonas*

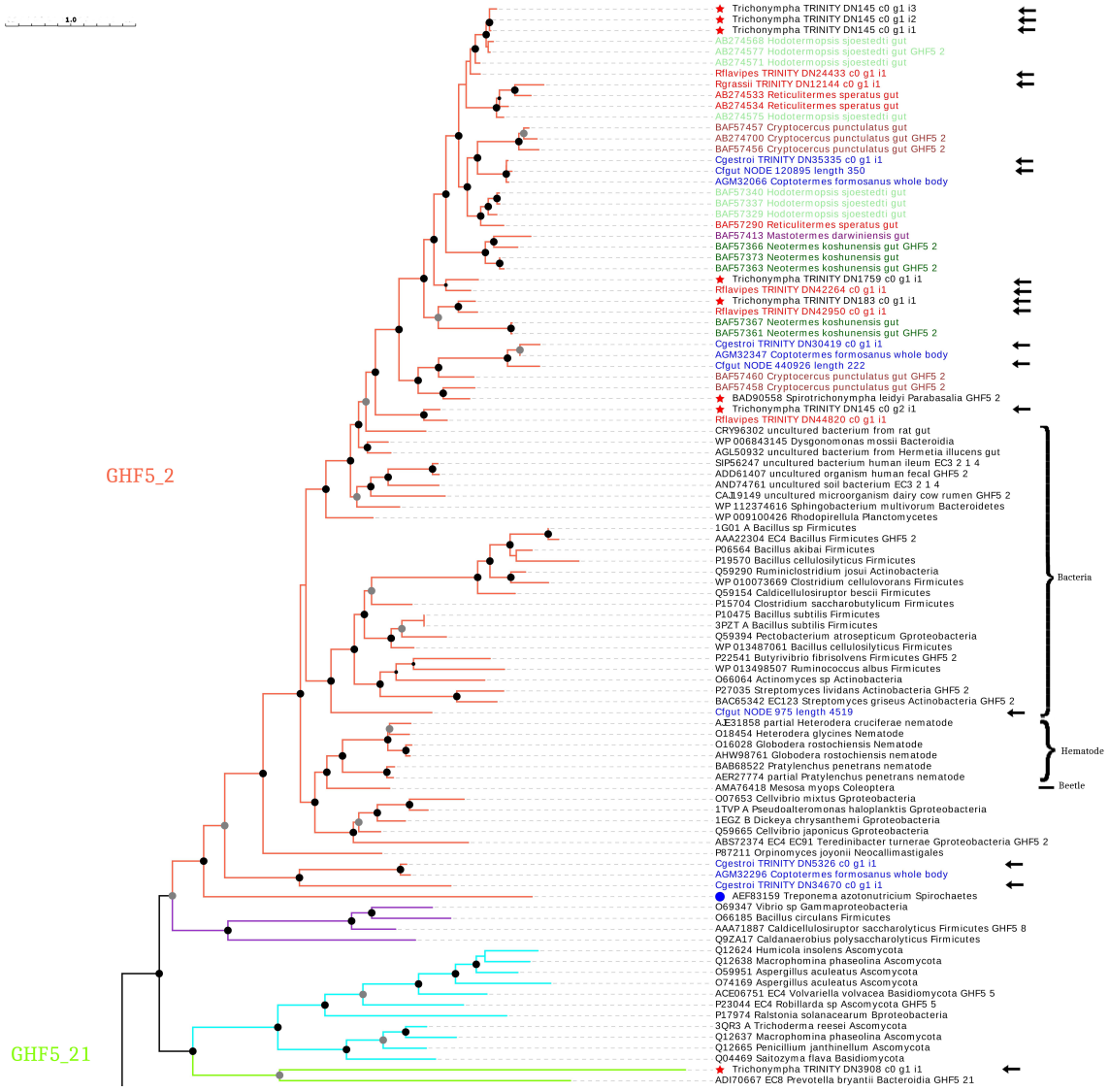

Figure 3.2: GHF5 Phylogeny Inset 1.
Inset 1 from figure 3.1. From bottom to top inset 1 includes colored branches indicating GHF5_21 (chartreuse), GHF5_5 (aqua), GHF5_8 (dark orchid), and GHF5_2 (tomato).

and *Pseudotrichomonas* transcriptomes, possibly indicating a LGT event in the protist class Trichonymphea early within the termite-protist relationship after divergence. As GHF5 has been found in all termite guts sampled to date, protist-originating GHF5 transcripts are likely to be found in other lower termite hindguts.

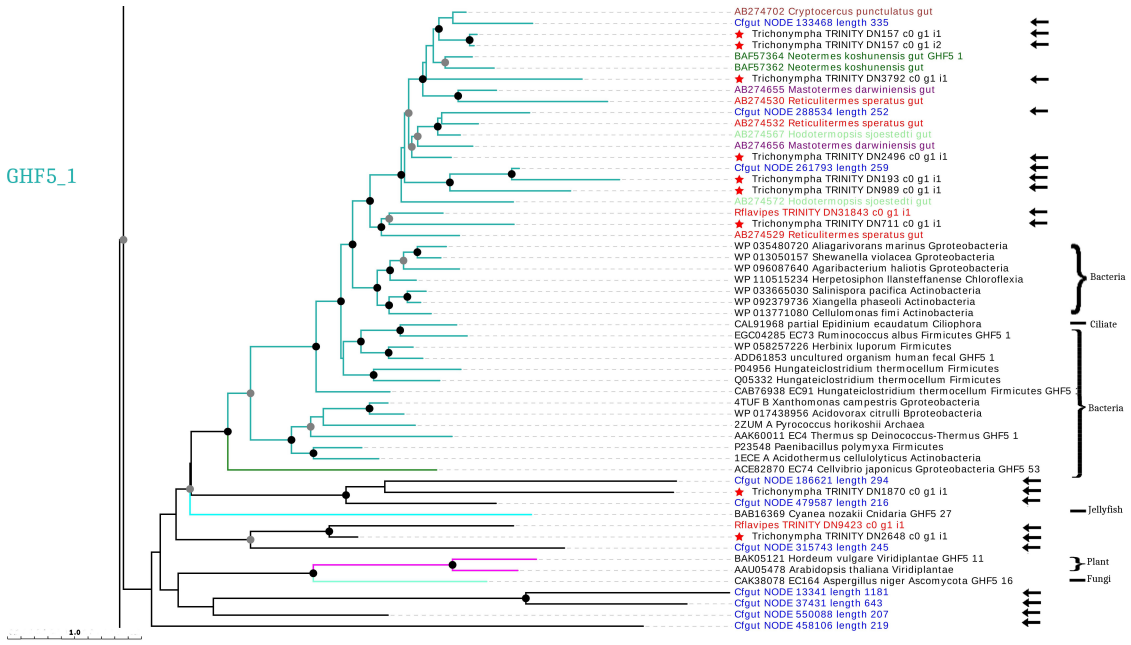

Figure 3.3: GHF5 Phylogeny Inset 2.
Inset 2 from figure 3.1. From bottom to top inset 2 includes colored branches indicating GHF5_16 (aquamarine), GHF5_11 (fuchsia), GHF5_27 (cyan), GHF5_53 (forest green), and GHF5_1 (light seagreen). Black branches were not able to be assigned to an established GHF5 subfamily.

3.2 Glycoside Hydrolase Family 7

3.2.1 GHF7 Background and History

GHF7 is one of the earliest discovered and largest families of glycoside hydrolases. The conserved 3D architecture is the β-jelly-roll structure which acts on polysaccharide main chains (Nagy *et al.*, 2016). Depending on the protein, GHF7 proteins function as a reducing-end cellobiohydrolase (CBH) or endoglucanase (EG). Unlike GHF5, the enzymatic activity catalyzed by members of this family follow phylogeny. The CBH and EG are the two major clades of GHF7, presumably resulting from a duplication and diversification event, wherein the CBHs have a tunnel-forming loop domain which allows them to bind to the ends of long-chain polysaccharides and act as exoglucanases.

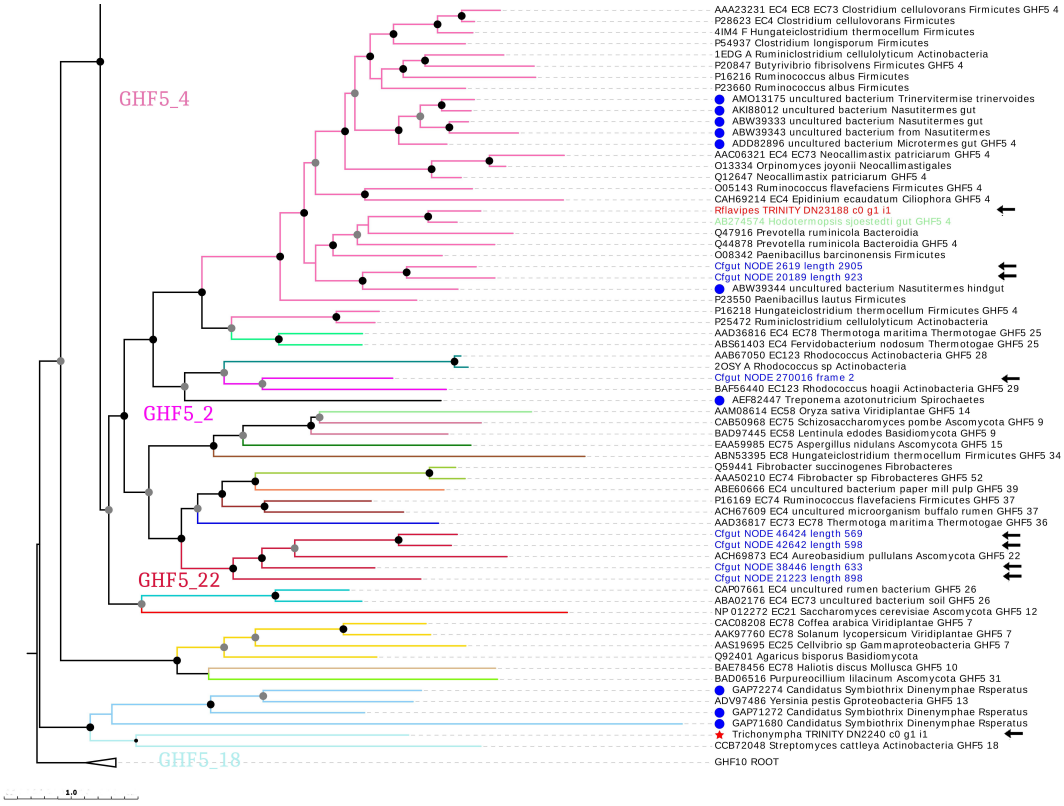

Figure 3.4: GHF5 Phylogeny Inset 3.
Inset 3 from figure 3.1. From bottom to top inset 3 includes colored branches indicating GHF5 subfamilies: GHF5_18 (pale turquoise), GHF5_13 (light skyblue), GHF5_31 (lawngreen), GHF5_10 (burlywood), GHF5_7 (gold), GHF5_12 (red), GHF5_26 (dark turquoise), GHF5_22 (crimson), GHF5_36 (blue), GHF5_37 (brown), GHF5_39 (coral), GHF5_52 (yellow-green), GHF5_34 (sienna), GHF5_15 (green), GHF5_9 (pale violet-red), GHF5_14 (light green), GHF5_29 (magenta), GHF5_28 (dark cyan), GHF5_25 (spring green), and GHF5_4 (hot pink). Black branches were not able to be assigned to an established GHF5 subfamily. Outgroup rooting was done with GHF10 sequences, indicated by a triangle representing the collapsed outgroup.

3.2.2 GHF7 in Termites and Across Life

GHF7 appears to be exclusively eukaryotic, with the few putatively bacterial members being identified in environmental sequencing results. GHF7s were first characterized to particular termite protists as EGs in *Pseudotrichonympha grassii* and *Holomastigotoides mirable* from *Coptotermes formosanus*. *Pseudotrichonympha grassii* CBHs were also found (Nakashima *et al.*, 2002; Watanabe *et al.*, 2002). To-

daka *et al.* 2010 found termite protist GHF7 sequences across all sampled lower termites and Cryptocercus, concluding it to be part of the "core enzyme set" upon which termite specialization in wood-eating derived. They found two distinct and supported clades dividing endoglucanases (EGs) and cellobiohydrolases (CBHs). As the closest relatives within both clades were fungal, the genes were considered to be innate genes rather than foreign acquisitions and that the termite-associated protist GHF7 proteins evolved concomitant with their cellulolytic symbiotic system (Todaka *et al.*, 2010a).

GHF7 was rooted with GHF11 as they both retain a β-jelly-roll structure (Collins *et al.*, 2005). GHF11 sequences of the following taxa comprise the root to the GHF7 phylogram: three bacteria, two termite protists, and one fungus. In line with previous studies there is a clade of termite-associated protists within each of these two major divisions (Figure 3.5). The *Trichonympha* transcriptome contained six distinct GHF7 EGs, spread throughout the termite protist clade. The published metatranscriptomes yielded GHF7 EGs; eleven from *Reticulitermes flavipes*, six from *Reticulitermes grassei*, and four from *Coptotermes gestroi* (Figure 3.2.2). The *Coptotermes formosanus* metagenomes had nine. Because GHF7 appears to be exclusively eukaryotic and these sequences are within the termite protist clade it can be safely assumed these are eukaryotic and of protist origin. The EG branch also contains Ascomycete fungi and *Daphnia* (a planktonic crustacean).

From the transcriptome data, the CBH GHF7 termite protist clade contained a *Trichonympha* sequence, eleven *Reticulitermes flavipes* sequences, seven *Reticulitermes grassei* sequences, and one *Coptotermes gestroi* sequence (Figure 3.2.2). From the *Coptotermes formosanus* metagenomes, three sequences were found in the termite protist clade. In addition to the termite-associated protist clade the CBH branch con-

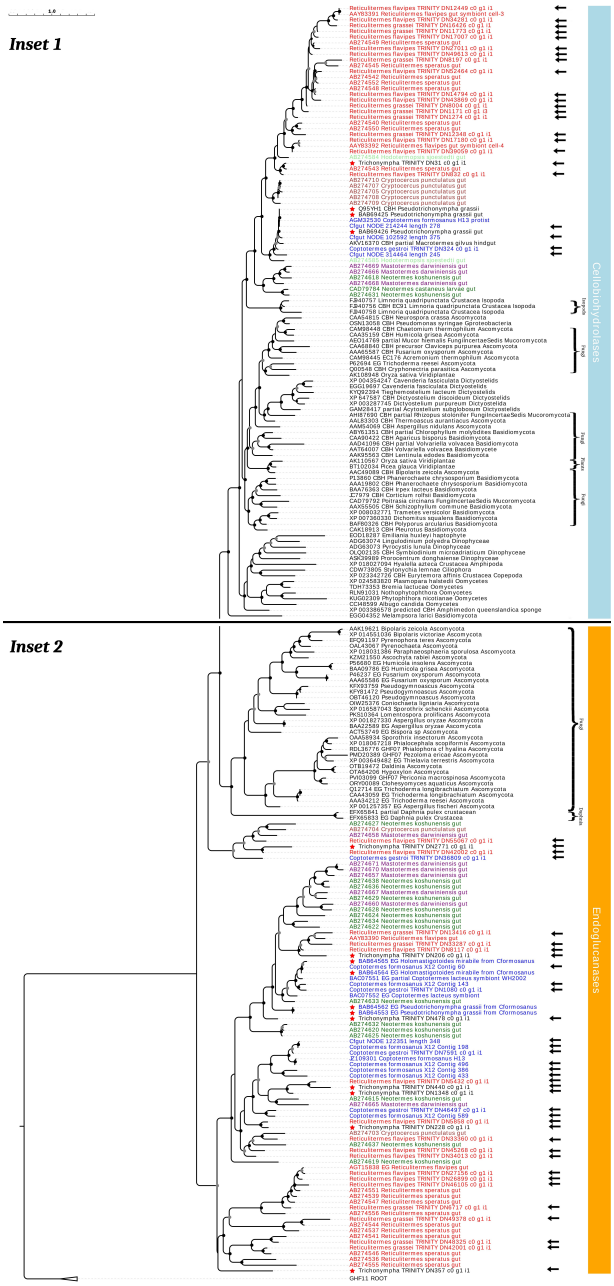

Figure 3.5: GHF7 Phylogeny Overview.
Minimum evolution phylogram of GHF7 amino acid sequences, computed on FastTree v. 2.1.10. Black arrows indicate GHFs identified in this work. Font colors indicate sequences from lower termite genus gut: blue, Coptotermes; red, Reticulitermes; purple, Mastotermes; dark green, Neotermes; light green, Hodotermopsis; brown Cryptocercus. Dots at nodes indicate SH support values: gray dots 50-79; black dots 80-100. Leaf decorations indicate termite-associated organisms: red stars, protists. Outgroup rooting was done with GHF11 sequences, indicated by a triangle representing the collapsed outgroup.

tains a more diverse set of taxa, including representatives of Oomycetes, Ciliophora, Haptophyta, Dinophyceae, Ascomycota, and Basidiomycota.

The transcriptomic data show *Trichonympha* throughout both the CBH and EG termite clades, while *Lophomonas* and *Pseudotrichomonas* were not found to have GHF7 transcripts (Figure 3.5). Termite gut metatranscriptomes from *Reticulitermes flavipes*, *R. grassei*, and *Coptotermes gestroi* all had GHF7 transcripts in the EG and CBH groups. The same was found of the *Coptotermes formosanus* gut metagenome. The incredible diversification and persistence of EG and CBH GHF7 enzymes in lower termites indicates it is an important enzyme family for this group. The rooted analysis shows the termite protist clade as the deepest branch with fungal, protist, and animal sequences and the termite protist CBHs branching later. These findings are consistent with an ancestral acquisition and duplication of a GHF7 protein that was retained in those taxa which have a need to deconstruct cellulose. However, given the lack of concordance with organismal phylogeny it is possible parabasalid GHF7 evolution involved lateral transfers.

3.3 Glycoside Hydrolase Family 8

3.3.1 GHF8 Background and History

Glycoside hydrolase family 8 was previously known as "Cellulase Family D" and is comprised of chitosanases, lichenases, cellulases and xylanases. GHF8 shares clan GH-M with GHF48, characterized by an $(\alpha/\alpha)_6$ fold catalytic module. Clan GH-M proteins use an inverting mechanism to hydrolyse polysaccharide bonds. It has been verified the catalytic proton donor in GHF8 is an asparagine residue, and the base is theorized to be a glycine.

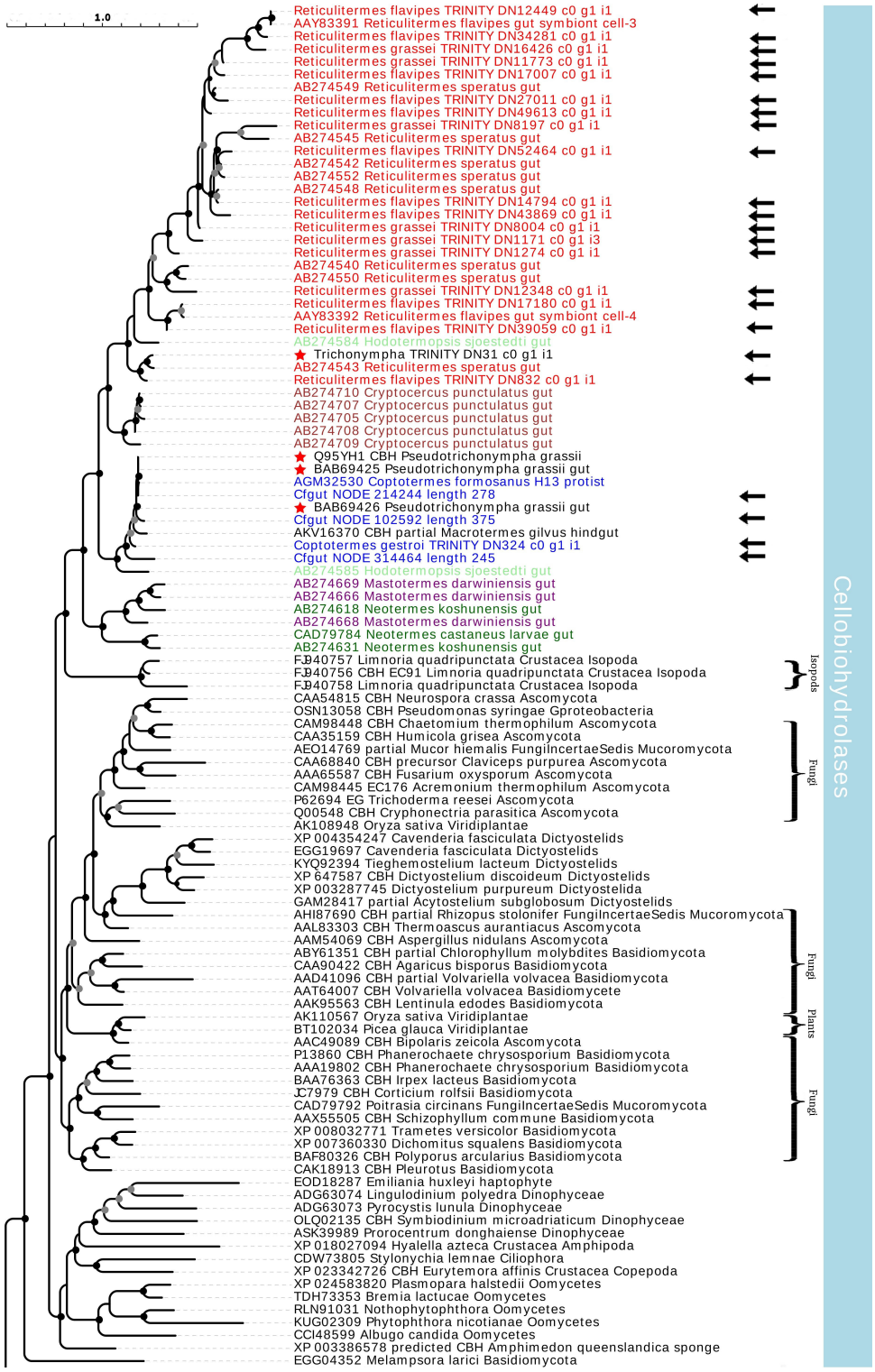

Figure 3.6: GHF7 Cellobiohydrolases.
Inset 1 from figure 3.5.

Figure 3.7: GHF7 Endoglucanases.
Inset 2 from figure 3.5.

3.3.2 GHF8 in Termites and Across Life

GHF8 is mostly encoded by prokaryotic organisms, however there is growing evidence of multiple LGT events of GHF8 early within the AGF lineages from Fibrobacter and Bacillales (Murphy *et al.*, 2019). Therefore it is possible there are other GHF8 proteins utilized by eukaryotes. Todaka *et al.* 2010 reported three GHF8 sequences in their EST analysis, one each from *Reticulitermes speratus*, *Neotermes koshunensis*, and *Mastotermes darwinensis*. However there were no GHF8 sequences among those ESTs uploaded and there was no phylogenetic tree or inference provided in the paper (Todaka *et al.*, 2010a). Other authors have reported GHF8 in lower and higher termite datasets, however these have not been subject to phylogenetic analysis.

GHF48 sequences served as the outgroup to GHF8 due to their being members of the same clan (Bourne and Henrissat, 2001). GHF48 sequences of the following taxa comprise the root to the GHF8 phylogram: three bacteria, two beetle, and one fungus. Like in other GHFs there is a group of diverse termite-associated protist, or possibly bacterial sequences (Figure 3.8). GHF8s were found in all three assembled *Reticulitermes* metatranscriptomes, the single-cell transcriptome from *Trichonympha*, both *Coptotermes formosanus* metagenomes, and the *Coptotermes gestroi* metatranscriptome. With the exception of some *Coptotermes* metagenomic sequences found in the bacterial portions of the tree, all GHF8 sequences mined formed one termite-associated clade.

There were three GHF8 transcripts from *Trichonympha*, which grouped with other termite-associated sequences (Figure 3.8). The *Trichonympha* GHF8s were found to group with poly-A selected metatranscriptome transcripts from *Reticulitermes flavipes* and *Reticulitermes grassei* within the termite protist clade. This group also contains metatranscriptome sequences from *Coptotermes formosanus* and *Coptoter-*

mes gestroi. The *Coptotermes* sequences were monophyletic within the termite-protist clade. The termite symbiotic clade is nested within a group of *Bacteroidetes* and *Prevotella*, groups associated with living in gut environments. This could point to an LGT event, however, the position of a *Reticulitermes flavipes* sequence near the base of this bacterial clade indicates this might not be the case.

The AGF clade is sister to the termite-associated GHF8s (Figure 3.8). Like the anaerobic gut fungi group, termite-associated protists likely benefited from a lateral gene transfer, allowing them to adapt to and exploit the hindgut environment. *Coptotermes formosanus* metagenome GHF8 grouped with *Treponema*, a known ectosymbiont of termite protists. The clade containing *Treponema* also groups with characterized reducing-end xylose releasing exo-oligoxylanases, indicating these sequences are likely bacterial and may catalyze the same reaction. Single-cell transcriptomics found no GHF8 transcripts within *Pseudotrichomonas* nor *Lophomonas*.

3.4 Glycoside Hydrolase Family 10

3.4.1 GHF10 Background and History

GHF10 was formerly known as "cellulase family F" prior to extensive enzymological characterization. GHF10 is a member of the Clan GH-A, as is GHF5. The members of Clan GH-A have a conserved $(\alpha/\beta)_8$ barrel, originally described for triosephosphate isomerase (TIM barrel) (Collins *et al.*, 2005). GH-A is a clan of retaining glycoside hydrolases, with a glutamate as the catalytic nucleophile and a glutamate as the general acid/base. GHF10 characterized members are exo-xylanases (EC 3.2.1.8), which are predominantly found in GHF10 and GHF11 though some have been found in GHF5, GHF8, GHF30, and GHF43 (Lombard *et al.*, 2014). Xylanases play an im-

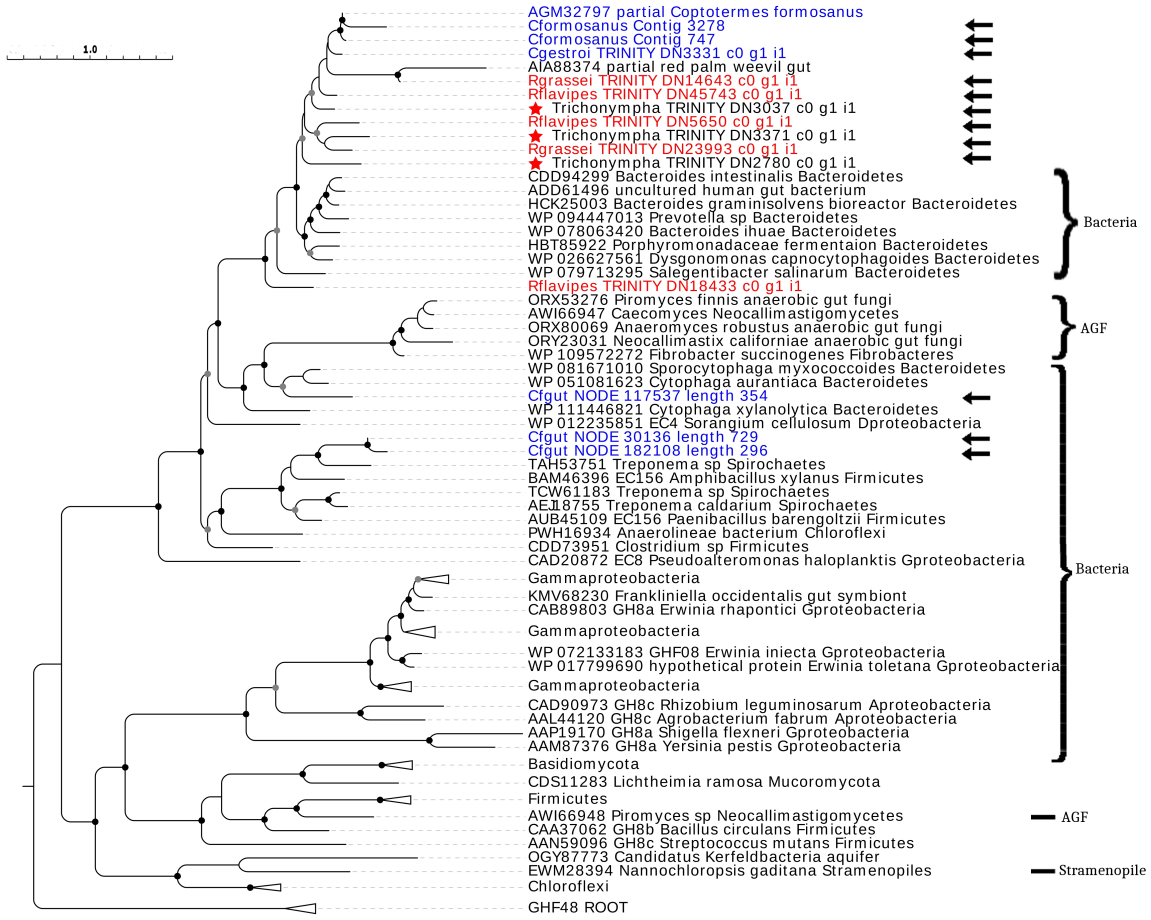

Figure 3.8: GHF8 Phylogeny.
Minimum evolution phylogram of GHF8 amino acid sequences, computed on FastTree v. 2.1.10. Black arrows indicate GHFs identified in this work. Font colors indicate sequences from lower termite genus gut: blue, Coptotermes; red, Reticulitermes. Dots at nodes indicate SH support values: gray dots 50-79; black dots 80-100. Leaf decorations indicate termite-associated organisms: red stars, protists. Outgroup rooting was done with GHF48 sequences, indicated by a triangle representing the collapsed outgroup.

portant role in depolymerizing hemicellulose. Interestingly, most xylanases of GHF10 and GHF11 are monospecific with one functional domain (Nguyen *et al.*, 2018).

3.4.2 GHF10 in Termites and Across Life

Characterized GHF10 sequences are broadly distributed in bacterial and fungal genomes, though there are characterized examples of GHF10 in Viridiplantae and

Fungi (Nguyen *et al.*, 2018). AGF have been demonstrated to have received their GHF10s from Clostridiales and unnested bacteria, the production of which assists in their survival on lignocellulosic foodstuffs within a gut environment (Murphy *et al.*, 2019). To handle dietary hemicellulose, higher termites are hypothesized to rely on increasing humification of their diet as well as their bacterial symbionts to produce the needed xylanases in their diet, as demonstrated in *Nasutitermes* (Dietrich *et al.*, 2014; Warnecke *et al.*, 2007). *Pseudacanthotermes militaris*, a fungus-farming higher termite has also been shown to harbor bacteria which produce GHF10 (Bastien *et al.*, 2013).

Lower termites are dependent on protist symbionts to break down xylan in the hindgut, though GHF10s have also been found in *Treponema azotonutricium*, a spirochaete bacteria found in lower termites (Rashamuse *et al.*, 2017). Regarding protists enzymes, Todaka *et al.* 2010 reported that GHF10 formed a monophyletic group including ESTs from all termites sampled except *Reticulitermes speratus*, which was absent. This lead them to propose GHF10 was shared by a recent common ancestor in lower termites but was secondarily lost at or just after *Reticulitermes* diverged from the main termite lineage. The termite gut-derived sequences formed a clade with the closest sequence from the bacterium *Rhodothermus marinus* (Todaka *et al.*, 2010a).

In this work, the GHF10 phylogeny was rooted with GHF5 sequences as its outgroup due to their shared evolutionary origins (Cantarel *et al.*, 2009). GHF5 sequences of the following taxa comprise the root to the GHF10 phylogram: five bacteria and four termite protists. For highly supported relationships, the overall topology was in congruence with the Todaka *et al.* 2010 topology, with the root being placed between Viridiplantae and termite protists (Figure 3.9). The addition of new taxa

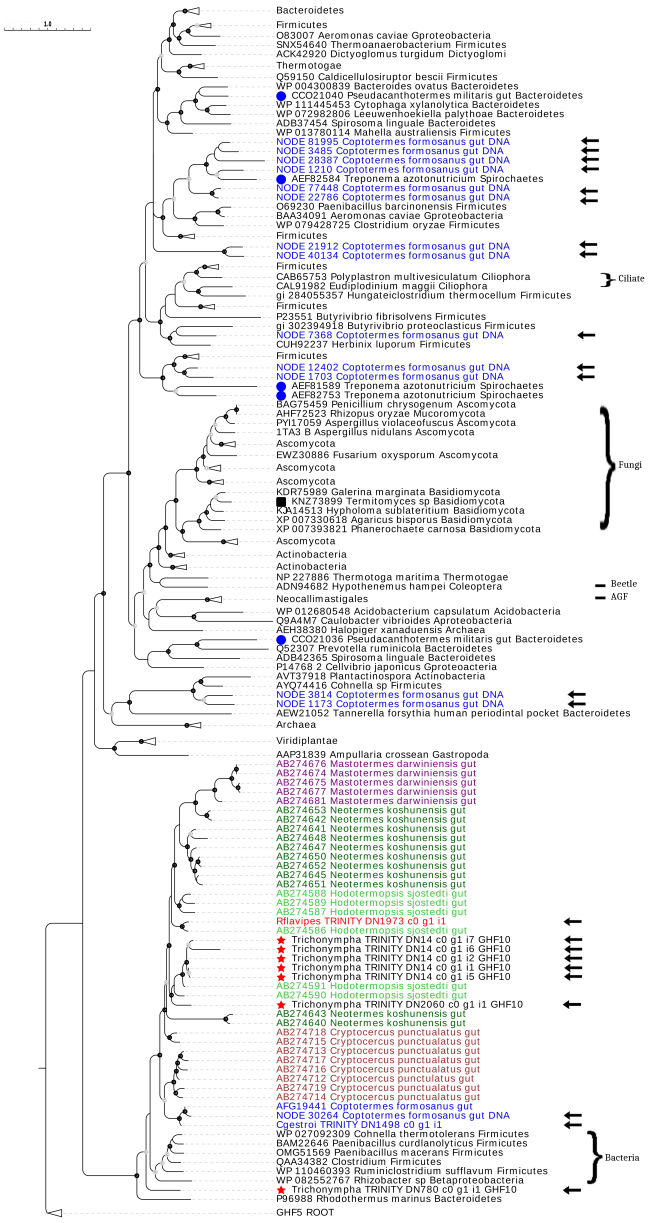

Figure 3.9: GHF10 Phylogeny.
Minimum evolution phylogram of GHF10 amino acid sequences, computed on FastTree v. 2.1.10. Black arrows indicate GHFs identified in this work. Font colors indicate sequences from lower termite genus gut: blue, Coptotermes; red, Reticulitermes; purple, Mastotermes; dark green, Neotermes; light green, Hodotermopsis; brown Cryptocercus. Dots at nodes indicate SH support values: gray dots 50-79; black dots 80-100. Leaf decorations indicate termite-associated organisms: red stars, protists; blue circles, termite-associated bacteria; black squares, termite-associated fungi. Outgroup rooting was done with GHF5 sequences, indicated by a triangle representing the collapsed outgroup.

and sequences did disrupt some aspects of the internal topology, particularly for long branches which are not well-supported.

A *Reticulitermes flavipes* sequence was found in the assembled metatranscriptome, its presence indicating *Reticulitermes* was not found in Todaka *et al.* 2010 either due to a lack of coverage or GHF10 has not actually been lost in all *Reticulitermes* (Figure 3.9). The *Reticulitermes flavipes* GHF10 branched with high SH support with *Hodotermopsis sjostedti*. One *Coptotermes formosanus* and one *Coptotermes gestroi* sequences were found within the protist clade sister to *Cryptocercus*, branching early within the termite protist clade. There were seven *Trichonympha* single-cell transcriptome transcripts, found only in the termite-associated protist clade with their host, *Hodotermopsis sjostedti*. This is notable because trichonymphids are also found in *Cryptocercus*, *Reticulitermes*, and *Coptotermes* but not in *Mastotermes* nor *Neotermes*. The termite-associated protist clade was nested within a Firmicutes-dominated bacterial group, which still included *Rhodothermus marinus* as the deepest branch. Additionally, there were thirteen *Coptotermes formosanus* metagenomic sequences found, distributed throughout the bacterial groups elsewhere in the tree. In the transcriptomes there were no *Lophomonas* nor *Pseudotrichomonas* GHF10 sequences, which would imply the GHF10 sequences in termite protists are not ancestrally derived (Figure 3.9).

3.5 Glycoside Hydrolase Family 11

3.5.1 GHF11 Background and History

GHF11 was among the earliest glycoside hydrolase families classified and was formerly known as family G. With GHF12, it forms clan GH-C, a retaining clan of glycoside hydrolases characterized by a β-jelly roll 3D structure. Because this

structure is retained, it allows the use of GHF7 members as an outgroup when rooting the GHF11 tree (Collins *et al.*, 2005). This can be done as their parent clans are both characterized by the β-jelly-roll fold (Naumoff, 2011). As of May 2019 there are 1019 bacterial GHF11s and 524 eukaryotic GHF11s in CAZy (Lombard *et al.*, 2014). GHF11 is a family of endo-β-1,4 and endo-β-1,3 xylanases with a glutamate serving as the catalytic acid and another glutamate serving as the catalytic nucleophile. Much like GHF10, GHF11 is important for depolymerizing the hemicellulose component of wood.

3.5.2 GHF11 in Termites and Across Life

GHF11s are found among bacteria, algae, fungi, protists, gastropods, and arthropods (Prade, 1996). GHF11 is important to many cellulolytic rumen eukaryotes, with several examples of its acquisition via LGT. The AGF Neocallimastigomycota GHF11s were obtained from the bacterial phyla Fibrobacter and Clostridiales, groups common in gut environments (Murphy *et al.*, 2019). Determining the donor in LGTs is always challenging and more evidence is needed to clear up the donors of the Ascomycete and Basidiomycete fungi (Álvarez-Cervantes *et al.*, 2016). Another eukaryotic group found in the rumen, ciliates, are thought to have acquired GHF11 xylanases from Firmicutes bacteria to adapt to the anaerobic, carbohydrate-rich gut environment (Ricard *et al.*, 2006).

Within the dual cellulolytic system of the lower termite GHF11 is contributed by protists (Tartar *et al.*, 2009). Of the few single-cell parabasalid transcriptomes published, GHF11 was identified within *Holomastigotoides mirable*, isolated from a *Coptotermes formosanus* (Arakawa *et al.*, 2009). Lower termite hindgut metatranscriptomes have identified GHF11 transcripts in *Coptotermes*, as well as in *Reticulitermes* and *Hodotermopsis* (Hussain *et al.*, 2013; Todaka *et al.*, 2010a). This is

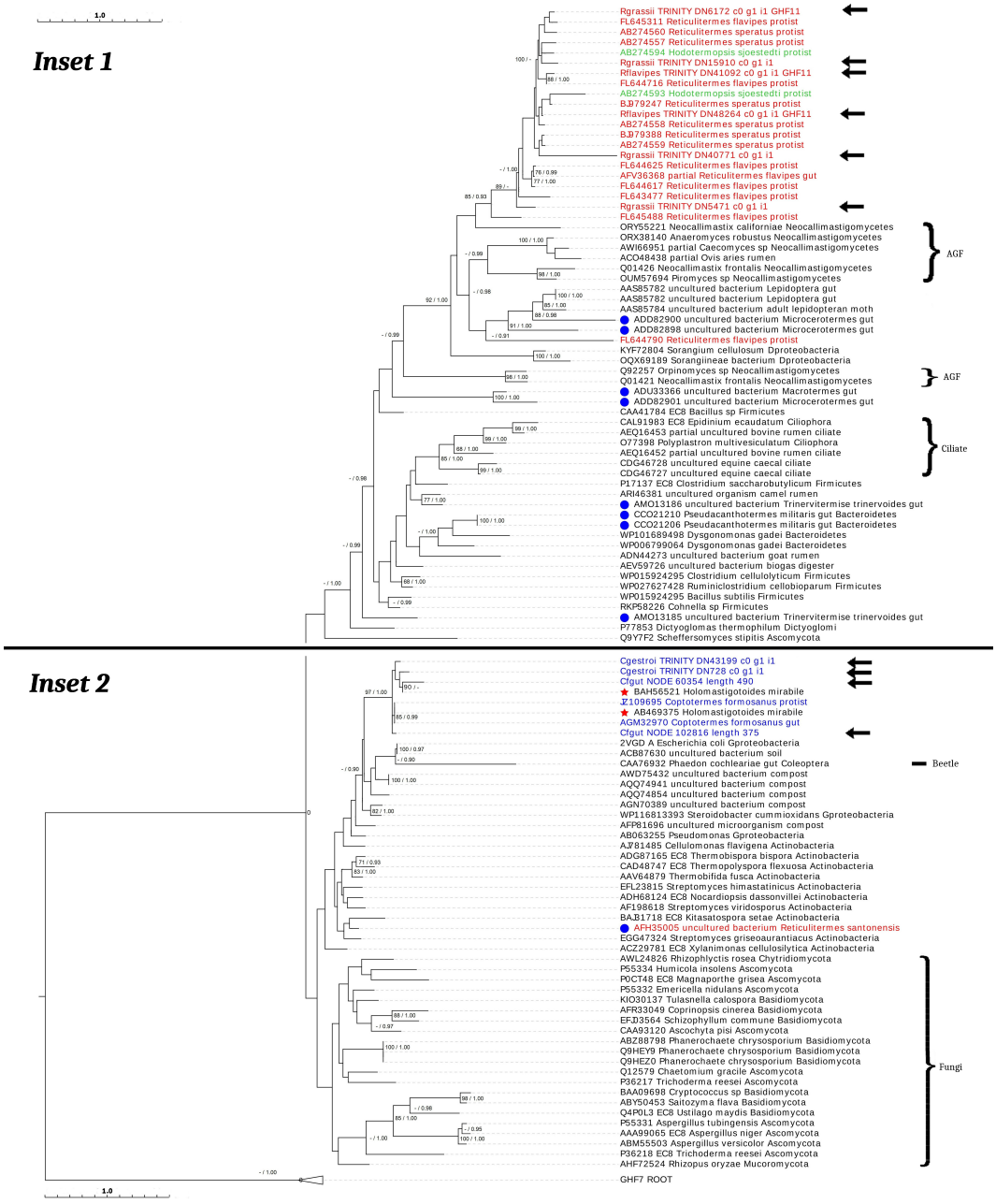

Figure 3.10: GHF11 Maximum Likelihood Phylogeny Overview.
Maximum Likelihood phylogram of GHF11 amino acid sequences, computed on RAxML v. 8.2.12 using WAG+G+F substitution. The scale bar represents branch length (as number of DNA substitutions/site). Black arrows indicate GHFs identified in this work. Font colors indicate sequences from lower termite genus gut: blue, Coptotermes; red, Reticulitermes; light green, Hodotermopsis. Numbers at nodes indicate bootstrap support values over 65 before the slash and posterior probability over 0.90 after the slash, where applicable. Leaf decorations indicate termite-associated organisms: red stars, protists; blue circles, bacteria. Outgroup rooting was done with GHF7 sequences, indicated by a triangle representing the collapsed outgroup.

consistent with Spirotrichonymphid symbionts in *Coptotermes*, *Reticulitermes*, and *Hodotermopsis*. These spirotrichonymphids are not found in other sampled lower termite lineages *Cryptocercus*, *Mastotermes*, or *Neotermes*. Because Todaka *et al.* 2010 identified GHF11s clustered together with low support in *Reticulitermes speratus* and *Hodotermopsis sjostedti* they hypothesized GHF11 to be found only in subterranean termites. Noting the short branch lengths, it was inferred that these genes share a recent common ancestor. It was also observed the termite protists formed a moderately supported clade with bacteria and AGFs as their closest relatives (Todaka *et al.*, 2010a).

GHF7 sequences of the following taxa comprise the root to the GHF11 phylogram: five fungi, two termite protists, a ciliate, and a dinoflagellate. In the recently assembled metatranscriptomes, four additional GHF11s were found in *Reticulitermes grassei* and three were found in *Reticulitermes flavipes* (Figure 3.10). These clustered with the previously described termite protist cluster with *Hodotermopsis sjostedti* and *Reticulitermes speratus* (Figure 3.11). In addition to the ME exploratory analysis, BI and ML analyses were run on the GHF11 data to further investigate the possibility of lateral gene transfer in this family. This clade is reasonably well-supported in the ME, ML, and BI analyses. Because GHF11 was not found in *Neotermes*, *Mastotermes*, or *Cryptocercus* these sequences likely belong to a protist only found within *Reticulitermes* and *Hodotermopsis*, such as such as the oxymonads *Pyrsonympha* or *Dinenympha* or the spirotrichonymphids *Spirotrichonympha* or *Microjoenia*. Deeper branching to this termite-associated protist clade was a clade containing anaerobic gut fungi and bacterial GHF11s from the higher termite, *Microtermes* (Figure 3.11).

Forming a second termite-associated group were two sequences from the *Coptotermes gestroi* metatranscriptome and two from the *Coptotermes formosanus* metagenomes which nested with the Spirotrichonymphid protist *Holomastigotoides mirable* sequences

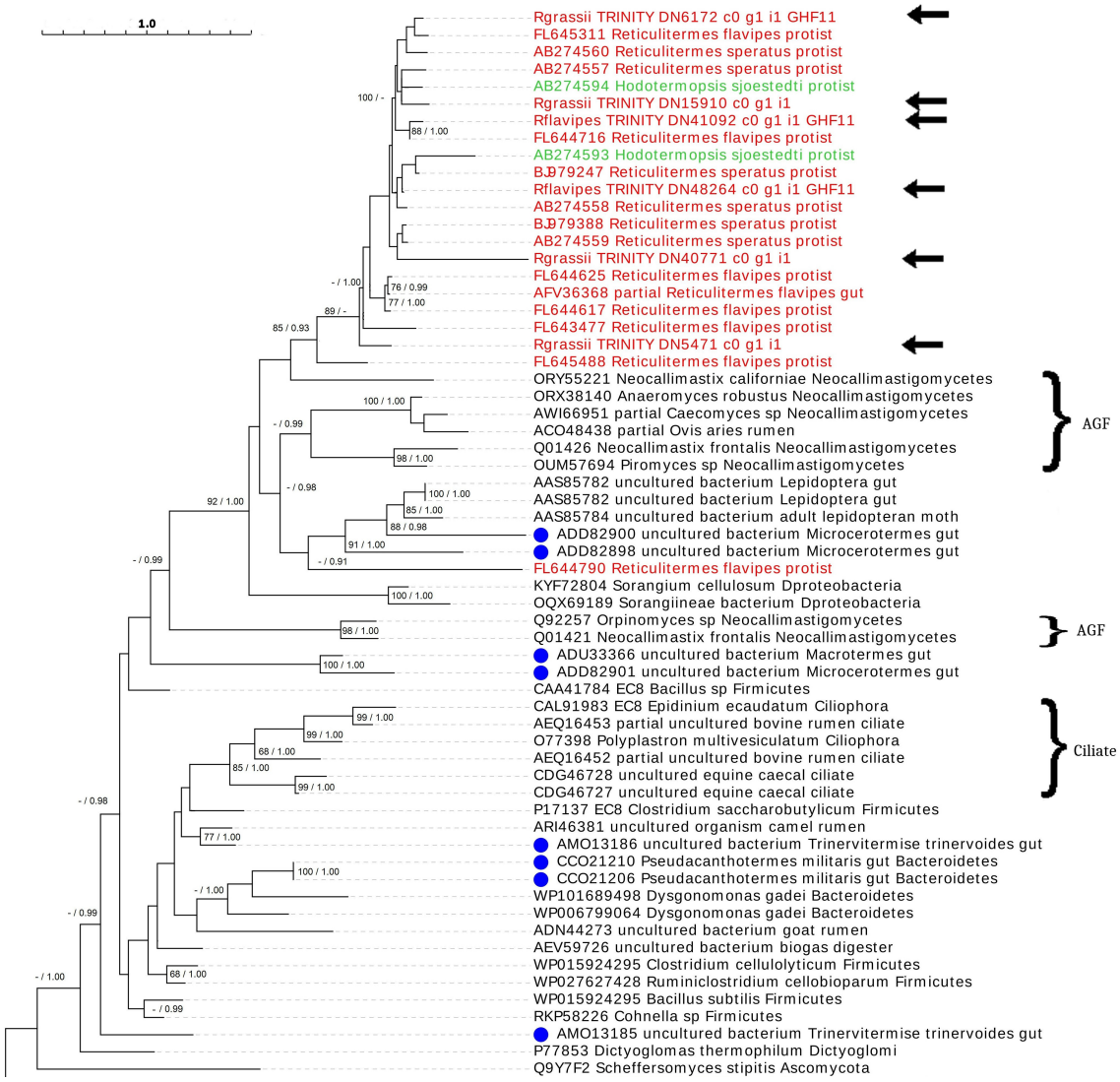

Figure 3.11: GHF11 Maximum Likelihood Phylogeny Inset 1.
Inset 1 of figure 3.10. The scale bar represents branch length (as number of DNA substitutions/site).

from *Coptotermes formosanus* (Figure 3.12). Sister to this clade is a group of bacterial sequences including bacterial GHF11s from compost samples. These results imply two LGT events from bacteria to early termite-associated protists. *Coptotermes* is also a subterranean termite, which would seem to support the hypothesis that GHF11 is found in the subterranean termites. The bootstrap and posterior probability scores are shown for highly supported branches in Figures 3.10, 3.11, and 3.12. The overall

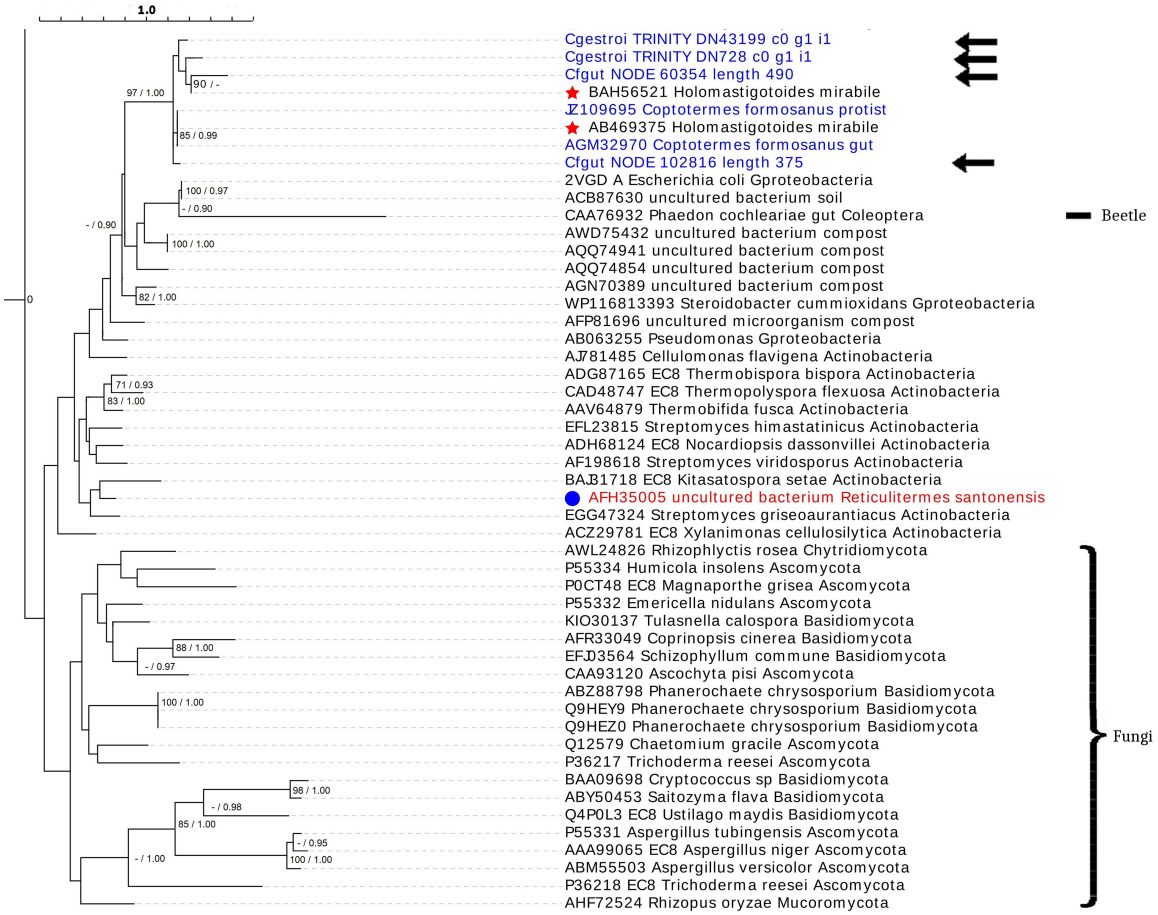

Figure 3.12: GHF11 Maximum Likelihood Phylogeny Inset 2.
Inset 2 of figure 3.10. The scale bar represents branch length (as number of DNA substitutions/site).

topology supports the ME phylogram. The *Coptotermes* clade is supported but lack of phylogenetic resolution prevented definitive identification of its closest relatives (Figure 3.12). GHF11 sequences were not found within the transcriptomes of *Trichonympha*, *Lophomonas*, nor *Pseudotrichomonas*, implying the GHF11 groups in the lower termite hindgut are not ancestrally derived, nor contributed by trichonymphids. These results support two independent acquisitions of GHF11 in the termite hindgut.

3.6 Glycoside Hydrolase Family 43 and 62

3.6.1 GHF43 Background and History

GHF43 is united with GHF62 with a conserved 5-fold β-propeller catalytic domain in clan GH-F (Nurizzo *et al.*, 2002; Naumoff, 2011). Clan GH-F is comprised entirely of inverting enzymes, inverting the stereochemistry of the anomeric carbon atom of its substrate. GHF43 is one of the largest glycoside hydrolase families, containing various debranching enzymes, particularly useful for deconstructing arabinoxylans in hemicellulose: xylanase, xylosidase, arabinase, arabinofuranosidase, and galactosidase (Mewis *et al.*, 2016). Arabinoxylans are the major component of grass and cereal hemicellulose (Gírio *et al.*, 2010). Xylans are the third most abundant biopolymers on earth, found in the cell walls of all grasses and the secondary cell walls of dicots. Endo-1,4-β-xylanase (EC 3.2.1.8) catalyzes endohydrolysis of xylans and can be found in GHF5, GHF8, GHF10, GHF11 and GHF43. Xylan 1,4-β-xylosidase (EC 3.2.1.37) hydrolyzes xylobiose, freeing monomeric xylose from the non-reducing termini of from short xylooligosaccharides during the final breakdown of plant cell-wall hemicellulose and are found in GHF1, GHF2, GHF3, GHF30, GHF39, GHF43, GHF51, GHF52, GHF54, GHF116, and GHF120 (Brüx *et al.*, 2006).

After D-xylose, L-arabinose is the second-most abundant pentose in hemicellulose and pectin (Seiboth and Metz, 2011). Arabinan endo-1,5-α-L-arabinanase (EC 3.2.1.99) catalyzes endohydrolysis of arabinans and thus far has only been found in GHF43. Non-reducing end α-L-arabinofuranosidase (EC 3.2.1.55) hydrolyze the terminal α-L-arabinofuranoside residues in α-L-arabinosides and is found in GHF2, GHF3, GHF43, GHF51, GHF54, GHF62, and GHF155. Galactan 1,3-β-galactosidase (EC 3.2.1.145) hydrolyzes terminal, non-reducing β-D-galactose residues from galactopyrans and is found in GHF16 and GHF43.

3.6.2 GHF62 Background and History

GHF62 is part of clan GH-F with GHF43, sharing a five-bladed β-propeller tertiary structure (Naumoff, 2012). Because of their shared origins, GHF62 is used as the outgroup to GHF43 (Lagaert *et al.*, 2014). Unlike the more diverse GHF43, GHF62 is comprised solely of non-reducing end α-L-arabinofuranosidases (EC 3.2.1.55). Arabinofuranosidases are xylan and arabinan debranching enzymes, catalyzing the release of arabinofuranosyl residues from lignocellulose or pectin. This enzyme activity is found in GHF2, GHF3, GHF43, GHF51, GHF54, and GHF62 though only GHF62 contains a single enzyme activity. The majority of GHF62 members are bacterial, and the eukaryotic members described so far are fungal, as of April 2019 there are 364 proteins total. Todaka 2007 *et al.* reported finding arabinofuranosidases of GHF62 in *Reticulitermes speratus* with an environmental cDNA library approach. In a meta-expressed sequence tag analysis in 2010 Todaka *et al.* reported GHF62 members in protist-enriched gut samples of *Reticulitermes speratus* and *Neotermes koshunensis* though the sequences were not among those uploaded to GenBank. Warnecke *et al.* 2007 reported not finding GHF62 in higher termite *Nasutitermes* metagenome, and Tartar *et al.* 2009 did not find GHF 62 in lower termite *Reticulitermes flavipes* nor in its protists.

3.6.3 GHF43 and GHF62 in Termites and Across Life

Given the diversity of function thus far characterized in GHF43, it may be unsurprising that the family has been split into 37 phylogenetically distinct subfamilies, 21 of which contain characterized members (Mewis *et al.*, 2016). The majority of GHF43 members are bacterial, with some eukaryotic. The multicellular eukaryotes are mostly

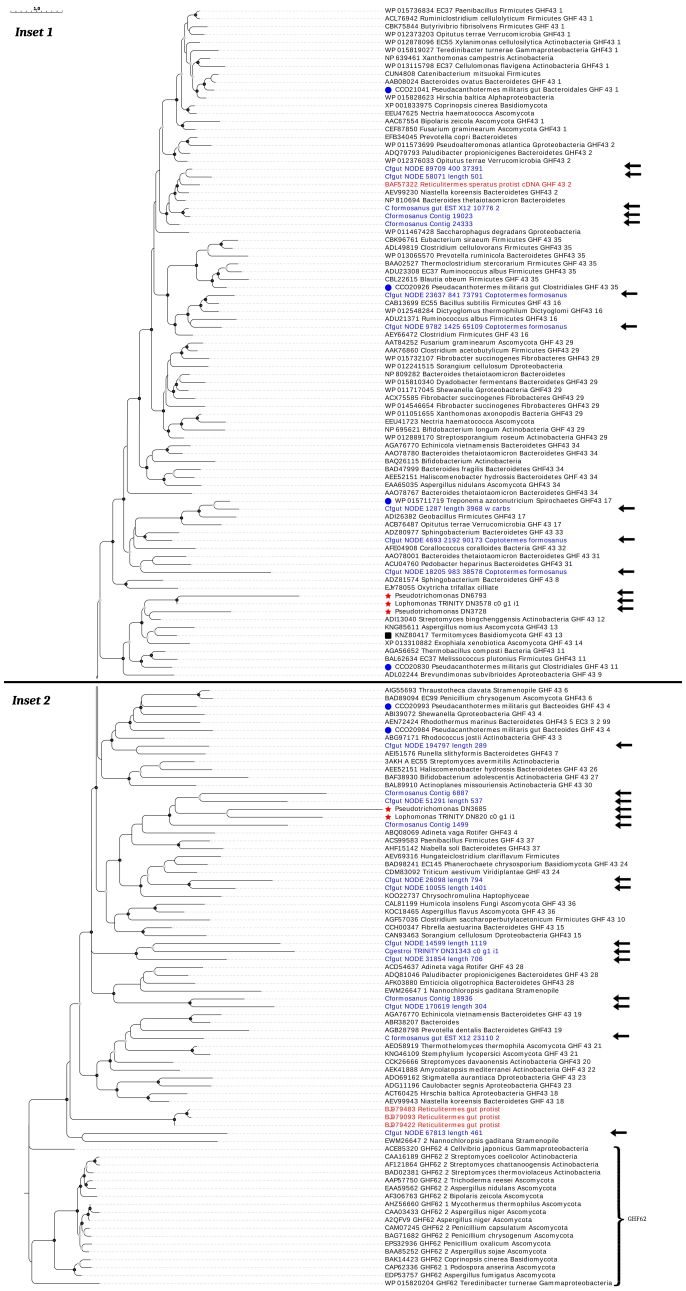

Figure 3.13: GHF43 and GHF62 Phylogeny Overview.
Minimum evolution phylogram of GHF43 amino acid sequences, computed on FastTree v. 2.1.10. Black arrows indicate GHFs identified in this work. Font colors indicate sequences from lower termite genus gut: blue, Coptotermes; red, Reticulitermes. Dots at nodes indicate SH support values: gray dots 50-79; black dots 80-100. Leaf decorations indicate termite-associated organisms: red stars, protists; blue circles, termite-associated bacteria; black squares, termite-associated fungi.

fungal with a few plant rotifer examples. The few GHF43 protist representatives are from ciliates, haptophytes, and stramenopiles, scattered among the subfamilies.

In a metagenome of the GI tract of the higher termite *Nasutitermes* GHF43 proteins made up 3% of identified glycoside hydrolases, and were attributed to treponemes (Warnecke *et al.*, 2007). A functional metagenomic screening of the higher termite *Pseudacanthotermes militaris* identified GHF43 in *Clostridiales* and *Bacteroides* sequences (Bastien *et al.*, 2013). Yuki *et al.* 2015 reported a GHF43 in a whole genome shotgun amplification of a bacterial ectosymbiont of the oxymonad protist *Dinenympha* from *Reticulitermes speratus*. In a meta-expressed sequence tag analysis, there were two GHF43 sequences in *Reticulitermes speratus* attributed to protists (Todaka *et al.*, 2010a). Xiao Jing Liu, 2016 reported EC 3.2.1.8, EC 3.2.1.37, and EC 3.2.1.55 in a metatranscriptome of the protistan community in *R. flaviceps*, however they were not classified into GHFs nor are the sequences available. While GHF43 representatives have been found in the lower termite hindgut metatranscriptome of *Coptotermes formosanus*, they have been ascribed to the bacterial inhabitants (Xie *et al.*, 2012; Zhang *et al.*, 2012).

GHF62 sequences were not in the transcriptomes of *Pseudotrichomonas*, *Trichonympha*, nor *Lophomonas* (Figure 3.13). GHF62s were also lacking in assembled metatranscriptomes of *R. flavipes*, *R. grassii*, *R. lucifugus*, and *C. gestroi*. The *C. formosanus* metagenome had two blast hits for GHF62, however both were too short to produce meaningful alignments. Therefore further GHF62 being reported in lower termite DNA or functional screening assays is unlikely.

GHF43 was located in the transcriptome data for *Pseudotrichomonas* and *Lophomonas* but lacking in the termite symbiote *Trichonympha* (Figure 3.13). After assembly, searching the previously published protist-enriched termite hindgut transcriptomes of *R. flavipes*, *R. lucifugus*, *R. grassei*, *C. formosanus*, and *C. gestroi* revealed no

Figure 3.14: GHF43 and GHF62 Phylogeny Inset 1.
Inset 1 of figure 3.13.

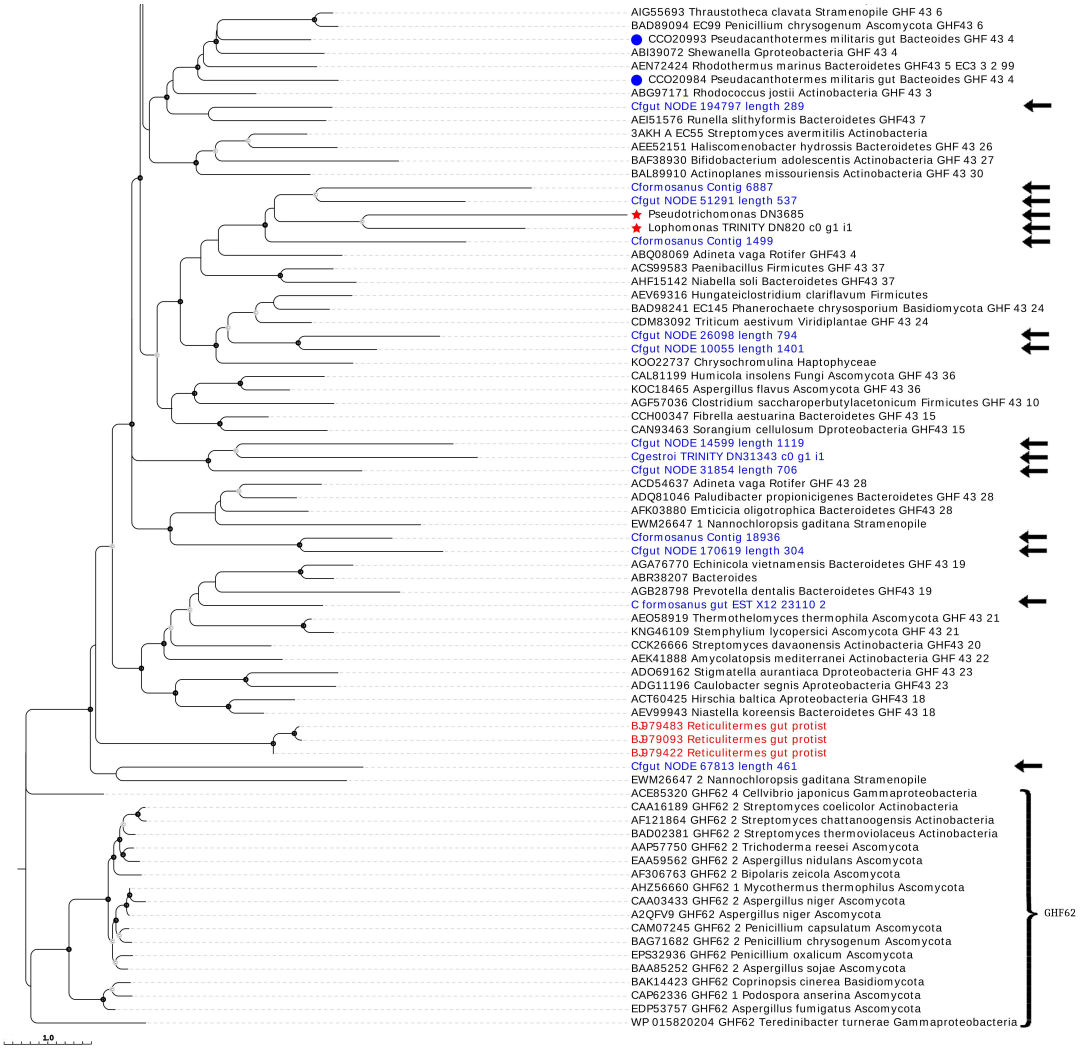

Figure 3.15: GHF43 and GHF62 Phylogeny Inset 2.
Inset 2 of figure 3.13.

additional GHF43 enzymes. When blasted against GHF43 representatives, the *C. formosanus* metagenome data contained many GHF43 hits, the majority of which are likely bacterial. A few of these hits branch with termite-associated representatives such *Treponema*, an intracellular endosymbiotic bacterium of termite protists (Figure 3.14 and 3.15). Another branched with *Reticulitermes* (Figure 3.14) and another with *Pseudotrichomonas* and *Lophomonas* (Figure 3.15). However, these

associations tended to have low support, potentially exacerbated by the differences when comparing metagenomic samples with metatranscriptome data.

Pseudotrichomonas is also in culture with bacteria so it is possible the transcripts are bacterial. Overall, the GHF43 tree is consistent with other phylogenetic examinations of this gene family, in which transcripts cluster with the subfamily designations and enzyme function. The relative paucity of GHF43 enzymes from the protist-enriched metatranscriptomes of wood-feeding lower termites may reflect the lack of grasses in their diet, whereas grass-eating and fungal farming termites would be expected to have them. Consistent with Su *et al.* 2016, it should be expected that the hemicellulose composition of a higher termite gut will more closely resemble the bacteria found in lower wood-feeding termites than that of their fungal farming relatives.

3.7 Glycoside Hydrolase Family 45

3.7.1 GHF45 Background and History

GHF 45 was formerly known as cellulase family K and was later renamed under the numerical naming scheme (Henrissat and Bairoch, 1993). The canonical structure of GHF45 is a six-stranded β-barrel domain with a seventh strand hydrogen bonded to the barrel. It is different to the jelly-roll structure of other glycoside hydrolases because the β-strands run both parallel and antiparallel (Davies *et al.*, 1993). The mechanism of hydrolysis is inverting, with aspartic acid residues serving as the catalytic nucleophile and the catalytic proton donor.

Within GHF45, there has been confirmed endo-β-1,4-glucanase activity (EC 3.2.1.4), mannan endo-β-1,4-mannanase (EC 3.2.1.78), and xyloglucan-specific endo-β-1,4-glucanase (EC 3.2.1.151) activities. As of May 2019 there are 408 GHF45 enzymes

listed in the CAZy database. Of these, 374 are eukaryotic. The family is distantly related to plant expansins, though no GHF45 members have been described within Viridiplantae (Lombard *et al.*, 2014).

3.7.2 GHF45 in Termites and Across Life

GHF 45 is largely found in ascomycete fungi so far, and to a lesser extent, basidiomycetes. There is some evidence the GHF45 was inherited from a common ancestor of Ascomycota and Basidiomycota, however the remaining phyla of fungi have not been as well sampled to date (Palomares-Rius *et al.*, 2014). Though Neocallimastigomycota have been shown to have received most of their GHFs via LGT from gut bacteria, GHF45 appears to be native within Fungi (Murphy *et al.*, 2019).

Within Metazoa, there are a few apparently isolated instances of GHF45 across invertebrates, being found in Mollusca, Nematoda, and Arthropoda (Busch *et al.*, 2019). It is thought that the mollusk and bacterial GHF45 genes are be a subfamily within GHF45 because they exhibit low sequence similarity to the rest of the GHF45 members while protist GHF45 are more closely related to insects. In nematodes, the GHF45s appears to have been acquired via LGT to fungus-feeding nematodes from Ascomycete fungi, class Sordariomycetes, and used to exploit the novel niche of plant parasitism (Palomares-Rius *et al.*, 2014). A recent study examined the function and phylogeny of GHF45 in Phytophaga beetles (order Coleoptera, superfamilies Chrysomeloidea+Curculionoidea) and across Opisthokonta. Consistent with Palomares-Rius *et al.*, Busch *et al.* found nematodes to be monophyletic. Of the arthropod-derived sequences, Phytophaga beetles form a monophyletic clade most closely related to budding yeast (Saccharomycetales), distinct from the nematode-related Sordariomycetes fungi. Oribatida mites formed a monophyletic clade separate from the beetles while Collembola springtails grouped together with a bacterial

sequence. They concluded that arthropods did not share an ancestral GHF45. They also showed rotifers and tardigrades as a clade with the AGF Neocallimastigomycota elsewhere in their GHF45 phylogram. Excitingly, they also undertook a functional screening of 37 GHF45 sequences from 5 species of beetle, being able to distinguish how amino acid substitutions result in altered substrate specificity. They concluded GHF45 is especially prone to substrate shifts and subsequent diversification within a lineage (Busch *et al.*, 2019).

The first glycoside hydrolase of protist origin was a member of GHF45 in *Reticulitermes speratus*, attributed to the trichonymphids *Trichonympha agilis* and *Teranympha mirabilis* (Ohtoko *et al.*, 2000). Shortly after that, GHF45 proteins were isolated and sequenced from *Koruga bonita* and *Deltotrichonympha nana*, hypermastigote Parabasalids within *Mastotermes darwiniensis* (Li *et al.*, 2003). It should be noted that *Koruga* and *Deltotrichonympha* are now considered likely synonyms (Čepička *et al.*, 2017). It has been previously reported *Coptotermes lacteus* may lack symbionts coding for GHF45 as they were not found in the hindgut extract when screened for endoglucanase activity (Watanabe *et al.*, 2002). Todaka *et al.* 2010 showed GHF45s from termite protists grouped with beetle GHF45s, separate from Fungi and nematodes (Todaka *et al.*, 2010a). Within the non-fungal clade they found GHF45 sequences in *Reticulitermes speratus*, *Hodotermopsis sjoestedti*, and *Cryptocercus punctulatus* branching together and *Mastotermes darwiniensis* being its own distinct clade. There were no GHF45 sequences found in the *Neotermes koshunensis* sampled.

With the data mined from assembled metatranscriptomes, there were three well-supported nodes within the protist cluster, with additional GHF45 sequences in *Reticulitermes*, *Coptotermes formosanus* and *Coptotermes gestroi*, as well as in the gut metagenome of *Coptotermes formosanus* (Figure 3.16). The failure of an earlier study to identify these transcripts in *Coptotermes* may be due to their screening for

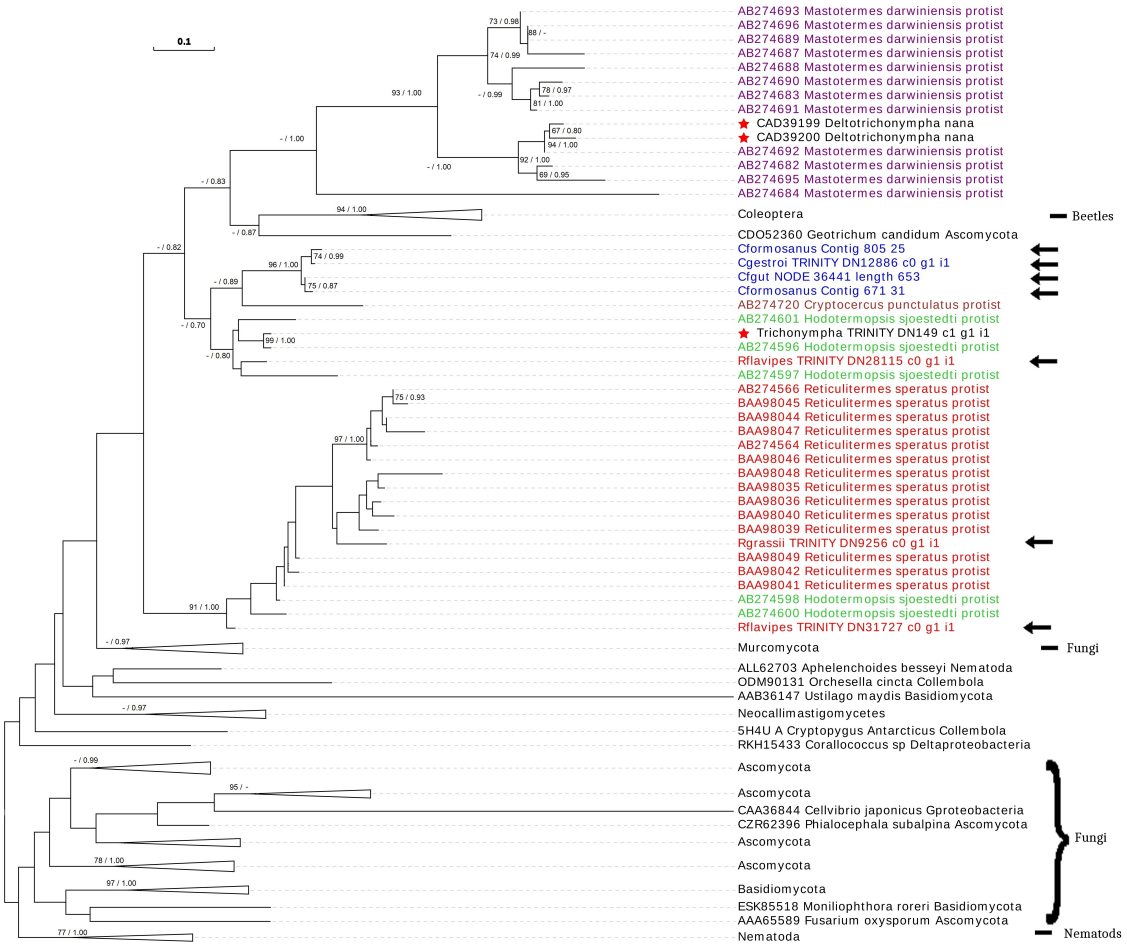

Figure 3.16: GHF45 Maximum Likelihood Phylogeny.
Unrooted maximum likelihood phylogram of GHF45 amino acid sequences, computed on RAxML v. 8.2.12 constructed with a WAG+G substitution model. The scale bar represents branch length (as number of DNA substitutions/site). Black arrows indicate GHFs identified in this work. Font colors indicate sequences from lower termite genus gut: blue, Coptotermes; red, Reticulitermes; purple, Mastotermes; light green, Hodotermopsis; brown Cryptocercus. Numbers at nodes indicate bootstrap support values. Leaf decorations indicate termite-associated organisms: red stars, protists.

endoglucanases, while the *Coptotermes* associated sequences might be glucomannases or glucoxylanases. The new *Coptotermes* GHF45s formed a monophyletic group with metatranscriptome sequences from *Cryptocercus*, *Hodotermopsis*, *Reticulitermes*. The placement of a *Trichonympha* GHF45 within this group indicates this clade is made

up of protists of the family Trichonymphidae, consistent with the identities of the host termites.

None of the new GHF45s found within this study are in the *Mastotermes* clade, the second major protist group within GHF45 (Figure 3.16). This may be due to the evolutionary distance between the cristamonad parabasalids inhabiting *Mastotermes* and those found within the other termites sampled. Given the distance between their flagellates, *Neotermes* would be expected to appear in this group, however Todaka *et al.* 2010 did not report any GHF45 sequences from *Neotermes koshunensis*, these results support GHF45 being secondarily lost in *Neotermes*. The third protist clade consisted of *Reticulitermes* and *Hodotermopsis* sequences but lacked *Trichonympha*. This is consistent with either a spirotrichonymphid parabasalid or possibly the oxymonads, *Pyrsonympha* or *Dinenympha*.

Similar to Todaka *et al.*, beetle sequences clustered within the termite-associated symbionts (Figure 3.16). Within the otherwise strictly protist group, Coleoptera was sister to *Geotrichum*, a saccharomycetes fungi. This was supported in both the BI and ME analyses. BI and ML analysis favored the WAG+G model of substitution, placing Coleoptera within the protist group. However, ML analysis did not show good bootstrap support for the position of this group, and showed disagreement between the WAG+G and LG+I+G substitution models. Given the uncertainty of the position of Coleoptera, it is possible GHF45 was acquired once by protists and subsequently by Coleoptera. However, if beetles acquired GHF45 from protists GHF45 should also be found in *Pseudotrichomonas* and *Lophomonas*.

Chapter 4

CONCLUSION

This study is a revised phylogenetic examination of the protist contribution to glycoside hydrolase enzymes found within the lower termite hindgut. As sequencing technology has become cheaper and more widespread the growing body of raw data available on the termite hindgut has grown. In the past decade genomes for more taxa have been sequenced and additional proteins have been characterized. Meanwhile, glycoside hydrolase families have grown and their evolution has become better understood. Increased computing power and molecular tools allow a methodical approach to surveying this new information. Specifically, data mining was done on metatranscriptomic data made available for three new species of *Reticulitermes* and two species of *Coptotermes*. In addition to the data available to the public through NCBI, this work was also able to include data from three protist transcriptomes; including the free-living parabasalid *Pseudotrichomonas*, the omnivorous cockroach symbiont *Lophomonas*, and the obligate wood-eating termite symbiote *Trichonympha*. The purpose of this study was to find glycoside hydrolases within these data and incorporate them into the existing body of knowledge and contextualize them within the scant protist-specific GHFs which had previously been identified, bolstered by new parabasalid transcriptomes.

Of the eight GHFs examined, the only family to contain sequences from *Pseudotrichomonas* and *Lophomonas* did not contain any sequences from *Trichonympha*. If the transcriptome coverage was good, the absence of GHF43 from *Trichonympha* could mean GHF43 was present in a shared ancestor of these three protists and secondarily lost in *Trichonympha*. This could explain the previously published *Reticulitermes*

GHF43 members, which could be from a related trichonymphid also found in *Reticulitermes*, *Teranympha*. The unpublished *Coptotermes formosanus* gut metagenome information shows some putative protist members within the termite protist clades, and the bacterial metagenome GHFs help shed light on the ecology of wood digestion in the *Coptotermes* hindgut.

Additional enzymes were found in all three GHF5 subfamilies previously known to contain termite protists and supports the bacteria-to-parabasalid LGT explanation for the origin of termite protist GHF5s (Todaka *et al.*, 2010a). GHF5 is a family that includes many bacterial and fungal members. While GHF5 was lacking in *Pseudotrichomonas* and *Lophomonas* single-cell transcriptomes, there were sequences from *Trichonympha* that fell within and others outside of the three previously known subfamilies known to contain termite protists. This hints at perhaps more, unrecognized GHF5 acquisitions in the ancestors of modern termite protists. Regardless of origin, it is clear GHF5 is important to termites, given its persistence and diversification within multiple termite lineages.

Much like GHF5, GHF7 shows a similar level of abundance and variety in the termite samples. This may point to an ancient dependence on GHF7 early in the termite lineage hypothesized in Todaka *et al.* 2010. Candidate close relatives were unable to be identified for termite GHF7s. GHF7 is dominated by fungal sequences. However, this phylogeny lacks concordance with organismal phylogeny makes inferences about origin from this dataset suspect. Inconclusive phylogenetic results like this highlight the need for additional approaches to the question of lateral gene transfer between eukaryotes. Though *Lophomonas* and *Pseudotrichomonas* lacked representation in all the glycoside hydrolase families, save GHF43, their absence from eukaryotic trees like GHF7 is especially resounding, as it indicates this gene was not inherited from a deeper eukaryotic shared ancestor.

Though GHF8 had not previously been subjected to phylogenetic examination with emphasis on the protists, these results support GHF8 presence within termite-associated protists. Rooting the tree with a member of the same clan helped to indicate the direction of diversification. Most of GHF8 is bacterial, much like GHF45. The phylogeny shows a termite-associated clade similar to what is seen in other GHFs. Three genera containing five species of termite harbored GHF8 sequences in this clade. Notably this protist clade also includes a monophyletic *Coptotermes* set of sequences within that clade. The monophyly of *Coptotermes* compared to the mixed *Reticulitermes* and *Hodotermopsis* topology perhaps reflects the intermingling of *Reticulitermes* and *Hodotermopsis* ancestral flora (Radek *et al.*, 2018). The location of a fermenting bacterial clade between the termite protist clade and a more basal *Reticulitermes* sequence makes inference of an LGT event less clear. Interpretation without additional data is challenging if the outlying *Reticulitermes* sequence is indeed of protist origin.

Typical of the glycoside hydrolases in termite-associated protists, GHF10 forms a diverse clade. The ME phylogeny provides some support for the the putative LGT of GHF10 to termite protists from bacteria. Previously thought to be secondarily lost in *Reticulitermes*, the discovered presence of a *Reticulitermes flavipes* GHF10 sequence indicates that not all Reticulitermes have lost their GHF10 xylanases. Because xylan is a sizable portion of lignocellulosic biomass it is unsurprising that it has been conserved. If other species of *Reticulitermes* are shown to have actually lost their GHF10 enzymes, this could be another instance of selective loss of protist symbionts within a species of termite. It is also possible that the presence of other gut xylanases from GHF11 relieves the selective pressure for protists to maintain their GHF10 enzymes.

The additional *Reticulitermes* GHF11s fell within the established termite protist clade with other *Reticulitermes* and *Hodotermopsis* sequences. Though these termite

genera are distantly related, they have closely related protist symbionts from the parabasalid families Trichonymphida and Spirotrichonymphida. The relatedness of their GHF11 sequences is consistent with a putative fauna transfer to a *Reticulitermes* ancestor (Radek *et al.*, 2018). This clade may belong to the spirotrichonymphid *Microjoenia*, which have been shown to be xylanolytic (Tarayre *et al.*, 2015). The placement of this group termite-associated GHF11s clustered with rumen fungal sequences and certain bacteria, lending further support to the association between AGFs and termite gut protists (Todaka *et al.*, 2010a). Also exciting is the additional *Coptotermes* sequences put into phylogenetic context with the *Holomastigotoides mirable* GHF11s. *H. mirable* is the only parabasalid GHF11 that has been confirmed to be produced by a particular protist. The phylogenetic analyses included BI, ML, and ME analyses; all of which pointed toward a second acquisition of a GHF11 from gut bacteria to protists with strong support for both termite protist clades. The addition of new termite gut and protist data updates the previous understanding where there was only one inferred acquisition of GHF11 by termite protists.

Between the six species of lower termite examined in these datasets, not one was able to produce a convincing example of a GHF62 of protist origin. While the possibility remains that there has not been enough coverage, it seems likely that GHF62 is absent in the termite gut. Because GHF62 and GHF43 are both in clan GH-F, it is possible GHF43 sequences had been misidentified as GHF62 due to similarity when previous termite protist studies were seeking family identification. The information available through NCBI has expanded dramatically in the past decade, allowing for more accurate identification of sequences.

Within GHF45 there has been a proliferation of new sequences available, particularly with beetles, nematodes, and AGF representatives. The expansion of eukaryotic representation available for this family provides tantalizing hints as to the acquisi-

tion and proliferation of specialized gene function. The protist representation within GHF45 does not contradict fungal origin of the family. Because the non-termite protists sampled lack GHF45, it appears there were three independent acquisitions of GHF45 in Cristamonadida, Spirotrichonymphida, and Trichonymphida.

Diverse, protist-specific clades were featured in GHF5, GHF7, GHF8, GHF10– nearly all the GHFs examined. These clades included as few as three and as many as all six of the termite genera examined, and the importance of these families is underscored by those instances where multiple *Reticulitermes* or multiple *Coptotermes* sequences were occurred together within one of these groups. In these cases the reinforced presence of a GHF member in multiple species in a genus also implies the data had good coverage across those datasets. GHF11, GHF43, and GHF62 did not exhibit these diverse protist groups. Given the sequence diversity we see in these termite protist clades, it is possible additional termite sampling could reveal hidden protist relatives in these groups.

The conclusions drawn rely on good coverage for the transcriptomes, metatranscriptomes, and genomes sampled here. Because data is drawn from a diversity of sources and technologies over time, inconsistencies are to be expected. Having a total of four species from *Reticulitermes* drawn from two different studies lends confidence in the clades and conclusions drawn here. Further, two species of *Coptotermes* were sampled from three datasets, one metatranscriptome and two metagenomes. This provides experimental as well as biological replication and the inclusion of the bacterial sequences from the genomes allow a more comprehensive examination of the microbial diversity within this complex symbiotic system.

In an effort to focus on the question of termite protists, this thesis only examined eight glycoside hydrolase families that were previously implicated as important in the lower termite hindgut. It is possible that in the intervening time since the last major

phylogenetic examination there have been additional or changed GHF classifications. There have been cases where GHF families or subfamilies are merged or split (Aspeborg *et al.*, 2012). Or there are situations as seen here with GHF62 where expanded sampling makes more accurate family determinations possible. This work helps to more accurately place, and in some cases recontextualize, GHFs that had previously been found in protists. The additional sequences from individual protists combined with metagenome mining allow for a more accurate determination of the role particular protists play within the cellulolytic system of the lower termite hindgut and helps to illuminate the origin of the GHF genes.

Aanen, D. K. and P. Eggleton, "Symbiogenesis: Beyond the endosymbiosis theory?", J. Theor. Biol. **434**, 99–103 (2017).

Abdul Rahman, N., D. H. Parks, D. L. Willner, A. L. Engelbrektson, S. K. Goffredi, F. Warnecke, R. H. Scheffrahn and P. Hugenholtz, "A molecular survey of Australian and North American termite genera indicates that vertical inheritance is the primary force shaping termite gut microbiomes", Microbiome **3**, 1, 5, URL `www.microbiomejournal.com/content/3/1/5` (2015).

Allgaier, M., A. Reddy, J. I. Park, N. Ivanova, P. D'Haeseleer, S. Lowry, R. Sapra, T. C. Hazen, B. A. Simmons, J. S. Vandergheynst and P. Hugenholtz, "Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community", PLoS ONE **5**, 1 (2010).

Álvarez-Cervantes, J., G. Díaz-Godínez, Y. Mercado-Flores, V. K. Gupta and M. A. Anducho-Reyes, "Phylogenetic analysis of β-xylanase SRXL1 of *Sporisorium reilianum* and its relationship with families (GH10 and GH11) of Ascomycetes and Basidiomycetes", Sci. Rep. **6**, April, 2–10, URL `http://dx.doi.org/10.1038/srep24010` (2016).

Anwar, Z., M. Gulfraz and M. Irshad, "Agro-industrial lignocellulosic biomass a key to unlock the future bio-energy: a brief review", J. Radiat. Res. Appl. Sci. **7**, 163–173 (2014).

Arakawa, G., H. Watanabe, H. Yamasaki, H. Maekawa and G. Tokuda, "Purification and molecular cloning of xylanases from the wood-feeding termite, *Coptotermes formosanus* Shiraki", Biosci. Biotechnol. Biochem. **73**, 3, 710–718, URL `www.tandfonline.com/doi/full/10.1271/bbb.80788` (2009).

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

Bajwa, P. K., T. Shireen, F. D'Aoust, D. Pinel, V. J. Martin, J. T. Trevors and H. Lee, "Mutants of the pentose-fermenting yeast *Pichia stipitis* with improved tolerance to inhibitors in hardwood spent sulfite liquor", Biotechnol. Bioeng. **104**, 5, 892–900, URL `www.doi.wiley.com/10.1002/bit.22449` (2009).

Baker, P. B. and R. J. Marchosky Jr, "Arizona termites of economic importance", Agriculture , June, 1–19, URL `cals.arizona.edu/pubs/insects/az1369.pdf` (2005).

Bastien, G., G. Arnal, S. Bozonnet, S. Laguerre, F. Ferreira, R. Fauré, B. Henrissat, F. Lefèvre, P. Robe, O. Bouchez, C. Noirot, C. Dumon and M. O'Donohue, "Mining for hemicellulases in the fungus-growing termite *Pseudacanthotermes militaris* using functional metagenomics", Biotechnol. Biofuels **6**, 1, 1–16, URL `www.ncbi.nlm.nih.gov/pubmed/23672637` (2013).

Behera, B., B. Sethi, R. Mishra, S. Dutta and H. Thatoi, "Microbial cellulases Diversity & biotechnology with reference to mangrove environment: A review", J. Genet. Eng. Biotechnol. **15**, 1, 197–210 (2017).

Behera, S. S. and R. C. Ray, "Solid state fermentation for production of microbial cellulases: Recent advances and improvement strategies", Int. J. Biol. Macromol. **86**, 656–669 (2016).

Benjamino, J. and J. Graf, "Characterization of the core and caste-specific microbiota in the termite, *Reticulitermes flavipes*", Front. Microbiol. **7**, 171 (2016).

Berlanga, M., "Functional symbiosis and communication in microbial ecosystems. the case of wood-eating termites and cockroaches", Int. Microbiol. **18**, 3, 159–169 (2015).

Bhattacharya, A. S., A. Bhattacharya and B. I. Pletschke, "Synergism of fungal and bacterial cellulases and hemicellulases: a novel perspective for enhanced bio-ethanol production", Biotechnol. Lett. **37**, 6, 1117–1129, URL `www.dx.doi.org/10.1007/s10529-015-1779-3` (2015).

Bolger, A. M., M. Lohse and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data", Bioinformatics (Oxford, England) **30**, 15, 2114–20 (2014).

Boscaro, V., E. R. James, R. Fiorito, E. Hehenberger, A. Karnkowska, J. Del Campo, M. Kolisko, N. A. A. T. A. T. T. Irwin, V. Mathur, R. H. Scheffrahn and P. J. Keeling, "Molecular characterization and phylogeny of four new species of the genus *Trichonympha* (Parabasalia, trichonymphea) from lower termite hindguts", Int. J. Syst. Evol. Microbiol. **67**, 9, 3570–3575 (2017).

Bourguignon, T., N. Lo, S. L. Cameron, J. Šobotnik, Y. Hayashi, S. Shigenobu, D. Watanabe, Y. Roisin, T. Miura and T. A. Evans, "The evolutionary history of termites as inferred from 66 mitochondrial genomes", Molecular Biology and Evolution **32**, 2, 406–421 (2015).

Bourne, Y. and B. Henrissat, "Glycoside hydrolases and glycosyltransferases: Families and functional modules", Current Opinion in Structural Biology **11**, 5, 593–600 (2001).

Bright, M. and S. Bulgheresi, "A complex journey : transmission of microbial symbionts", **8**, 3, 218–230 (2010).

Brown, M. E. and M. C. Y. Chang, "Exploring bacterial lignin degradation", Current Opinion in Chemical Biology **19**, 1, 1–7, URL `www.dx.doi.org/10.1016/j.cbpa.2013.11.015` (2014).

Brune, A., "Symbiotic digestion of lignocellulose in termite guts", Nat. Rev. Microbiol. **12**, 3, 168–180, URL `www.nature.com/doifinder/10.1038/nrmicro3182` (2014).

Brune, A. and M. Friedrich, "Microecology of the termite gut: Structure and function on a microscale", Curr. Opin. Microbiol. **3**, 3, 263–269 (2000).

Brüx, C., A. Ben-David, D. Shallom-Shezifi, M. Leon, K. Niefind, G. Shoham, Y. Shoham and D. Schomburg, "The structure of an inverting GH43 β-xylosidase from *Geobacillus stearothermophilus* with its substrate reveals the role of the three catalytic residues", J. Mol. Biol. **359**, 1, 97–109 (2006).

Busch, A., E. G. Danchin and Y. Pauchet, "Functional diversification of horizontally acquired glycoside hydrolase family 45 (GH45) proteins in Phytophaga beetles", BMC Evol. Biol. **19**, 1, 100 (2019).

Cantarel, B. L., P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard and B. Henrissat, "The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics.", Nucleic acids research **37**, D233–8 (2009).

Capella-Gutiérrez, S., J. M. Silla-Martínez and T. Gabaldón, "trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses", Bioinformatics **25**, 15, 1972–1973 (2009).

Čepička, I., M. F. Dolan and G. H. Gile, "Parabasalia", in "Handb. Protists", edited by J. A. et Al. (Springer International Publishing AG 2016, 2017), URL `http://link.springer.com/10.1007/978-3-319-32669-6`.

Che, Y., D. Wang, Y. Shi, X. Du, Y. Zhao, N. Lo and Z. Wang, "A global molecular phylogeny and timescale of evolution for *Cryptocercus* woodroaches", Mol. Phylogenet. Evol. **98**, April, 201–209, URL `www.dx.doi.org/10.1016/j.ympev.2016.02.005` (2016).

Chouvenc, T., H.-f. Li, J. Austin, C. Bordereau, T. Bourguignon, S. L. Cameron, E. M. Cancello, R. Constantino, A. M. Costa-leonardo, P. Eggleton, T. A. Evans, B. Forschler, J. K. Grace, C. Husseneder, J. Kreček, C.-Y. Lee, T. Lee, N. Lo, M. Messenger, A. Mullins, A. Robert, Y. Roisin, R. H. Scheffrahn, D. Sillam-Dussès, J. Šobotník, A. Szalanski, Y. Takematsu, E. L. Vargo, A. Yamada, T. Yoshimura and N.-Y. Su, "Revisiting *Coptotermes* (Isoptera: Rhinotermitidae): A global taxonomic road map for species validity and distribution of an economically important subterranean termite genus", Syst. Entomol. **41**, 2, 299–306, URL `http://doi.wiley.com/10.1111/syen.12157` (2016).

Cleveland, L. R., "Symbiosis between termites and their intestinal protozoa", Nature **114**, 2853, 22 (1924a).

Cleveland, L. R., "The physiological and symbiotic relationships between the intestinal protozoa of termites and their host, with special reference to *Reticulitermes flavipes* kollar", Biological Bulletin **46**, 5, 203–227, URL `www.jstor.org/stable/1536724` (1924b).

Cleveland, L. R., S. R. Hall, E. P. Sanders and J. Collier, "The wood-feeding roach *Cryptocercus*, its protozoa and the symbiosis between protozoa and roach", Mem. Am. Acad. Arts Sci. **17**, 2, 185–342 (1934).

Collins, T., C. Gerday and G. Feller, "Xylanases, xylanase families and extremophilic xylanases", FEMS Microbiol. Rev. **29**, 1, 3–23 (2005).

Davies, G., H. Gilbert, B. Henrissat, B. Svensson, D. Vocadlo and S. Williams, "Ten Years of CAZypedia: A living encyclopedia of carbohydrate-active enzymes", Glycobiology **28**, 1, 3–8 (2018).

Davies, G. J., G. G. Dodson, R. E. Hubbard, S. P. Tolley, Z. Dautert, K. S. Wilsont, C. Hjortt, J. M. Mikkelsent, G. Rasmussent and M. Schiileint, "Structure and function of endoglucanase V", Nature **365**, September 1993, 362–364 (1993).

de Gonzalo, G., D. I. Colpa, M. H. Habib and M. W. Fraaije, "Bacterial enzymes involved in lignin degradation", J. Biotechnol. **236**, 110–119 (2016).

Dedeine, F., L. A. Weinert, D. Bigot, T. Josse, M. Ballenghien, V. Cahais, N. Galtier and P. Gayral, "Comparative analysis of transcriptomes from secondary reproductives of three *Reticulitermes* termite species", PLoS ONE **10**, 12, 1–18, URL `www.dx.doi.org/10.1371/journal.pone.0145596` (2015).

Dietrich, C., T. Köhler and A. Brune, "The cockroach origin of the termite gut microbiota: Patterns in bacterial community structure reflect major evolutionary events", Appl. Environ. Microbiol. **80**, 7, 2261–2269 (2014).

Diouf, M., V. Hervé, P. Mora, A. Robert, S. Frechault, C. Rouland-Lefèvre and E. Miambi, "Evidence from the gut microbiota of swarming alates of a vertical transmission of the bacterial symbionts in *Nasutitermes arborum* (Termitidae, Nasutitermitinae)", Antonie van Leeuwenhoek **111**, 4, 573–587, URL `www.link.springer.com/10.1007/s10482-017-0978-4` (2018).

Duarte, S., T. Nobre, P. A. V. Borges and L. Nunes, "Symbiotic flagellate protists as cryptic drivers of adaptation and invasiveness of the subterranean termite *Reticulitermes grassei* Clément", Ecol. Evol. **8**, 11, 5242–5253, URL `www.doi.wiley.com/10.1002/ece3.3819` (2018).

Durand, P., P. Lehn, I. Callebaut, S. Fabrega, B. Henrissat and J.-P. Mornon, "Active-site motifs of lysosomal acid hydrolases: invariant features of clan GH-A glycosyl hydrolases deduced from hydrophobic cluster analysis", Glycobiology **7**, 2, 277–284 (1997).

Ebert, A. and A. Brune, "Hydrogen concentration profiles at the oxic-anoxic interface: a microsensor study of the hindgut of the wood-feeding lower termite *Reticulitermes flavipes* (Kollar).", Applied and environmental microbiology **63**, 10, 4039–4046 (1997).

Edgar, R. C., "MUSCLE: multiple sequence alignment with high accuracy and high throughput.", Nucleic Acids Res. **32**, 5, 1792–7 (2004).

Evangelista, D. A., B. Wipfler, O. Bethoux, A. Donath, M. Fujita, M. K. Kohli, F. Legendre, S. Liu, R. Machida, B. Misof, R. S. Peters, L. Podsiadlowski, J. Rust, K. Schuette, W. Tollenaar, J. L. Ware, T. Wappler, X. Zhou, K. Meusemann and S. Simon, "An integrative phylogenomic approach illuminates the evolutionary history of cockroaches and termites (Blattodea)", Proc. R. Soc. B **286**, URL `https://dx.doi.org/10.6084/m9.figshare.c.4358900` (2019).

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate and M. Punta, "Pfam: the protein families database.", Nucleic Acids Res. **42**, Database issue, D222–30 (2014).

Franco Cairo, J. P. L., M. F. Carazzolle, F. C. Leonardo, L. S. Mofatto, L. B. Brenelli, T. A. Goncalves, C. A. Uchima, R. R. Domingues, T. M. Alvarez, R. Tramontina, R. O. Vidal, F. F. Costa, A. M. Costa-Leonardo, A. F. Paes Leme, G. A. Pereira and F. M. Squina, "Expanding the knowledge on lignocellulolytic and redox enzymes of worker and soldier castes from the lower termite *Coptotermes gestroi*", Front. Microbiol. **7**, OCT, URL `https://www.frontiersin.org/articles/10.3389/fmicb.2016.01518/full` (2016).

Gayral, P., J. Melo-Ferreira, S. Glémin, N. Bierne, M. Carneiro, B. Nabholz, J. M. Lourenco, P. C. Alves, M. Ballenghien, N. Faivre, K. Belkhir, V. Cahais, E. Loire, A. Bernard and N. Galtier, "Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap", PLoS Genetics **9**, 4, e1003457, URL `www.dx.plos.org/10.1371/journal.pgen.1003457` (2013).

Gile, G. H. and C. H. Slamovits, "Phylogenetic position of *Lophomonas striata* Bütschli (parabasalia) from the hindgut of the cockroach *Periplaneta americana*", Protist **163**, 2, 274–283, URL `www.dx.doi.org/10.1016/j.protis.2011.07.002` (2012).

Gírio, F. M. F., C. Fonseca, F. Carvalheiro, L. C. L. Duarte, S. Marques and R. Bogel-Łukasik, "Hemicelluloses for fuel ethanol: A review", Bioresour. Technol. **101**, 13, 4775–4800, URL `www.dx.doi.org/10.1016/j.biortech.2010.01.088` (2010).

Guerriero, G., J. F. Hausman, J. Strauss, H. Ertan and K. S. Siddiqui, "Destructuring plant biomass: Focus on fungal and extremophilic cell wall hydrolases", Plant Sci. **234**, 180–193, URL `www.dx.doi.org/10.1016/j.plantsci.2015.02.010` (2015).

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, D. Philip, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. D. Macmanes, M. Ott, J. Orvis and N. Pochet, "*De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity", Nat Protoc. **8**, 8, 1–43 (2014).

He, Z., H. Zhang, S. Gao, M. J. Lercher, W.-H. Chen and S. Hu, "Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees", Nucleic Acids Res. **44**, W236–W241, URL `https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw370` (2016).

Henrissat, B., "A classification of glycosyl hydrolases based sequence similarities amino acid", Biochem. J. **280**, 309–316 (1991).

Henrissat, B. and A. Bairoch, "New families in the classification of glycosyl hydrolases based on amino acid sequence similarities", Biochem. J. **293**, 3, 781–788, URL `www.biochemj.org/lookup/doi/10.1042/bj2930781` (1993).

Henrissat, B., M. Claeyssens, P. Tomme, L. Lemesle and J. P. Mornon, "Cellulase families revealed by hydrophobic cluster analysis", Gene **81**, 1, 83–95 (1989).

Henrissat, B. and G. Davies, "Structural and sequence-based classification of glycoside hydrolases", Curr. Opin. Struct. Biol. **7**, 5, 637–644, URL `www.biomednet.com/elecref/0959440X00700637` (1997).

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

Hongoh, Y., P. Deevong, T. Inoue, S. Moriya, S. Trakulnaleamsai, M. Ohkuma, C. Vongkaluang, N. Noparatnaraporn and T. Kudo, "Intra- and interspecific comparisons of bacterial diversity and community structure support coevolution of gut microbiota and termite host.", Appl. Environ. Microbiol. **71**, 11, 6590–6599 (2005).

Hussain, A., Y. F. Li, Y. Cheng, Y. Liu, C. C. Chen and S. Y. Wen, "Immune-Related Transcriptome of *Coptotermes formosanus* Shiraki Workers: The Defense Mechanism", PLoS One **8**, 7, e69543 (2013).

Inoue, J.-I., K. Saita, T. Kudo, S. Ui and M. Ohkuma, "Hydrogen production by termite gut protists: characterization of iron hydrogenases of Parabasalian symbionts of the termite *Coptotermes formosanus*.", Eukaryotic cell **6**, 10, 1925–32 (2007).

Inoue, T., S. Moriya, M. Ohkuma and T. Kudo, "Molecular cloning and characterization of a cellulase gene from a symbiotic protist of the lower termite, *Coptotermes formosanus*", Gene **349**, 67–75 (2005).

Isikgor, F. H. and C. R. Becer, "Lignocellulosic biomass: a sustainable platform for the production of bio-based chemicals and polymers", Polym. Chem. **6**, 25, 4497–4559 (2015).

James, E. R., N. Okamoto, F. Burki, R. H. Scheffrahn and P. J. Keeling, "*Cthulhu macrofasciculumque* n.g., n.sp. and *Cthylla microfasciculumque* n.g., n. sp., a newly identified lineage of parabasalian termite symbionts", PLoS ONE **8**, 3, 9–13 (2013).

Jasso-Selles, D. E., F. De Martini, K. D. Freeman, M. D. Garcia, T. L. Merrell, R. H. Scheffrahn and G. H. Gile, "The parabasalid symbiont community of *Heterotermes aureus*: Molecular and morphological characterization of four new species and reestablishment of the genus *Cononympha*", Eur. J. Protistol. **61**, 48–63 (2017).

Jensen, C. U., J. K. Rodriguez Guerrero, S. Karatzos, G. Olofsson and S. B. Iversen, "Fundamentals of Hydrofaction^{TM}: Renewable crude oil from woody biomass", Biomass Conversion and Biorefinery **7**, 4, 495–509 (2017).

Jordan, D. B., M. J. Bowman, J. D. Braker, B. S. Dien, R. E. Hector, C. C. Lee, J. A. Mertens and K. Wagschal, "Plant cell walls to ethanol", Biochem. J. **442**, 2, 241–252 (2012).

Jouquet, P., S. Traoré, C. Choosai, C. Hartmann and D. Bignell, "Influence of termites on ecosystem functioning. Ecosystem services provided by termites", Eur. J. Soil Biol. **47**, 4, 215–222, URL www.dx.doi.org/10.1016/j.ejsobi.2011.05.005 (2011).

Juturu, V. and J. C. Wu, "Microbial exo-xylanases: a mini review", Appl. Biochem. Biotechnol. **174**, 1, 81–92, URL www.link.springer.com/10.1007/s12010-014-1042-8 (2014).

Ke, J., J.-Z. Sun, H. D. Nguyen, D. Singh, K. C. Lee, H. Beyenal and S.-L. Chen, "*In-situ* oxygen profiling and lignin modification in guts of wood-feeding termites", Insect Sci. **17**, 3, 277–290, URL www.doi.wiley.com/10.1111/j.1744-7917.2010.01336.x (2010).

Koidzumi, M., "Studies on the intestinal protozoa found in the termites of Japan", Parasitology **13**, 3, 235–309 (1921).

König, H., L. Li and J. Fröhlich, "The cellulolytic system of the termite gut", Appl. Microbiol. Biotechnol. **97**, 18, 7943–7962, URL www.link.springer.com/10.1007/s00253-013-5119-z (2013).

Korb, J., M. Poulsen, H. Hu, C. Li, J. J. Boomsma, G. Zhang and J. Liebig, "A genomic comparison of two termites with different social complexity", Front. Genet. **6**, 9 (2015).

Lagaert, S., A. Pollet, C. M. Courtin and G. Volckaert, "β − *Xylosidasesand*α − *L* − *arabinofuranosidases* : *Accessoryenzymesforarabinoxylandegradation*′′, *BiotechnologyAdvances***32**, 2, 316− −332(2014).

Larsson, A., "AliView: A fast and lightweight alignment viewer and editor for large datasets", Bioinformatics **30**, 22, 3276–3278 (2014).

Li, L., P. Pfeiffer and H. Ko, "Termite gut symbiotic Archaezoa are becoming living metabolic fossils", Eukaryotic cell **2**, 5, 1091–1098 (2003).

Liu, N., X. Yan, M. Zhang, L. Xie, Q. Wang, Y. Huang, X. Zhou, S. Wang and Z. Zhou, "Microbiome of fungus-growing termites: a new reservoir for lignocellulase genes.", Appl. Environ. Microbiol. **77**, 1, 48–56, URL www.ncbi.nlm.nih.gov/pubmed/21057022 (2011).

Liu, X. J., M. Che, L. Xie, S. Zhan, Z. H. Zhou, Y. P. Huang and Q. Wang, "Metatranscriptome of the protistan community in *Reticulitermes flaviceps*", Insect Sci. **23**, 4, 543–547 (2016).

Lombard, V., H. Golaconda Ramulu, E. Drula, P. M. Coutinho and B. Henrissat, "The carbohydrate-active enzymes database (CAZy) in 2013", Nucleic Acids Res. **42**, D1, 490–495 (2014).

Machas, M., G. Kurgan, A. K. Jha, A. Flores, A. Schneider, S. Coyle, A. M. Varman, X. Wang and D. R. Nielsen, "Emerging tools, enabling technologies, and future opportunities for the bioproduction of aromatic chemicals", J. Chem. Technol. Biotechnol. **94**, 1, 38–52, URL www.doi.wiley.com/10.1002/jctb.5762 (2019).

MacManes, M. D., "On the optimal trimming of high-throughput mRNA sequence data", Front. Genet. **5**, 13 (2014).

Marriott, P. E., L. D. Gómez and S. J. McQueen-Mason, "Unlocking the potential of lignocellulosic biomass through plant science", New Phytol. **209**, 4, 1366–1381, URL www.doi.wiley.com/10.1111/nph.13684 (2015).

Mewis, K., N. Lenfant, V. Lombard and B. Henrissat, "Dividing the large glycoside hydrolase family 43 into subfamilies: A motivation for detailed enzyme characterization", Appl. Environ. Microbiol. **82**, 6, 1686–1692, URL www.ncbi.nlm.nih.gov/pubmed/26729713 (2016).

Murphy, C. L., N. H. Youssef, R. A. Hanafy, M. B. Couger, J. E. Stajich, Y. Wang, K. Baker, S. S. Dagar, G. W. Griffith, I. F. Farag, T. M. Callaghan and M. S. Elshahed, "Horizontal gene transfer as an indispensible driver for Neocallimastigomycota evolution into a distinct gut-dwelling fungal lineage", bioRxiv Prepr. p. 487215, URL www.biorxiv.org/content/10.1101/487215v2 (2019).

Nagy, L. G., R. Riley, A. Tritt, C. Adam, C. Daum, D. Floudas, H. Sun, J. S. Yadav, J. Pangilinan, K. H. Larsson, K. Matsuura, K. Barry, K. Labutti, R. Kuo, R. A. Ohm, S. S. Bhattacharya, T. Shirouzu, Y. Yoshinaga, F. M. Martin, I. V. Grigoriev and D. S. Hibbett, "Comparative genomics of early-diverging mushroom-forming fungi provides insights into the origins of lignocellulose decay capabilities", Mol. Biol. Evol. **33**, 4, 959–970 (2016).

Nakashima, K., H. Watanabe and J.-I. Azuma, "Cellulase genes from the parabasalian symbiont *Pseudotrichonympha grassii* in the hindgut of the wood-feeding termite *Coptotermes formosanus*", Cell. Mol. Life Sci. **59**, 9, 1554–1560, URL www.link.springer.com/10.1007/s00018-002-8528-1 (2002).

Naumoff, D., "Hierarchical Classification of Glycoside Hydrolases", Biochem. **76**, 6, 622–635 (2011).

Naumoff, D., "Furanosidase superfamily: Search of homologues", Mol. Biol. **46**, 2, 322–327, URL www.link.springer.com/10.1134/S0026893312010153 (2012).

Nguyen, S. T., H. L. Freund, J. Kasanjian and R. Berlemont, "Function, distribution, and annotation of characterized cellulases, xylanases, and chitinases from CAZy", Appl. Microbiol. Biotechnol. **102**, 1629–1637, URL http://link.springer.com/10.1007/s00253-018-8778-y (2018).

Nicolaou, S. A., S. M. Gaida and E. T. Papoutsakis, "A comparative view of metabolite and substrate stress and tolerance in microbial bioprocessing: From biofuels and chemicals, to biocatalysis and bioremediation", (2010).

Nieves, L. M., L. A. Panyon and X. Wang, "Engineering Sugar Utilization and Microbial Tolerance toward Lignocellulose Conversion", Front. Bioeng. Biotechnol. **3**, February, 1–10 (2015).

Noda, S., Y. Hongoh, T. Sato and M. Ohkuma, "Complex coevolutionary history of symbiotic Bacteroidales bacteria of various protists in the gut of termites", BMC Evolutionary Biology **9**, 1, 1–12 (2009).

Noda, S., O. Kitade, T. Inoue, M. Kawai, M. Kanuka, K. Hiroshima, Y. Hongoh, R. Constantino, V. Uys, J. Zhong, T. Kudo and M. Ohkuma, "Cospeciation in the triplex symbiosis of termite gut protists (*Pseudotrichonympha* spp.), their hosts, and their bacterial endosymbionts", Mol. Ecol. **16**, 6, 1257–1266, URL `www.doi.wiley.com/10.1111/j.1365-294X.2006.03219.x` (2007).

Noda, S., D. Shimizu, M. Yuki, O. Kitade and M. Ohkuma, "Host-Symbiont Cospeciation of Termite-Gut Cellulolytic Protists of the Genera *Teranympha* and *Eucomonympha* and their *Treponema* Endosymbionts", Microbes and Environments **33**, 1, 26–33 (2018).

Nurizzo, D., J. P. Turkenburg, S. J. Charnock, S. M. Roberts, E. J. Dodson, V. A. McKie, E. J. Taylor, H. J. Gilbert and G. J. Davies, "*Cellvibrio japonicus* α-L-arabinanase 43A has a novel five-blade β-propeller fold", Nat. Struct. Biol. **9**, 9, 665–668, URL `www.nature.com/doifinder/10.1038/nsb835` (2002).

Ohkuma, M., "Termite symbiotic systems: efficient bio-recycling of lignocellulose", Appl. Microbiol. Biotechnol. **61**, 1, 1–9, URL `www.link.springer.com/10.1007/s00253-002-1189-z` (2003).

Ohkuma, M., "Symbioses of flagellates and prokaryotes in the gut of lower termites", Trends in Microbiology **16**, 7, 345–352 (2008).

Ohkuma, M., K. Ohtoko, T. Iida, M. Tokura, S. Moriya, R. Usami, K. Horikoshi and T. Kudo, "Phylogenetic identification of hypermastigotes, *Pseudotrichonympha*, *Spirotrichonympha*, *Holomastigotoides*, and Parabasalian symbionts in the hindgut of termites", Journal of Eukaryotic Microbiology **47**, 3, 249–259 (2000).

Ohtoko, K., M. Ohkuma, S. Moriya, T. Inoue, R. Usami and T. Kudo, "Diverse genes of cellulase homologues of glycosyl hydrolase family 45 from the symbiotic protists in the hindgut of the termite *Reticulitermes speratus*", Extremophiles **4**, 6, 343–349, URL `www.link.springer.com/10.1007/s007920070003` (2000).

Palomares-Rius, J. E., Y. Hirooka, I. J. Tsai, H. Masuya, A. Hino, N. Kanzaki, J. T. Jones and T. Kikuchi, "Distribution and evolution of glycoside hydrolase family 45 cellulases in nematodes and fungi", Proc. Natl. Acad. Sci. **14**, 69, URL `http://www.pnas.org/cgi/doi/10.1073/pnas.95.9.4906` (2014).

Pawlowski, J., S. Audic, S. Adl, D. Bass, L. Belbahri, C. Berney, S. S. Bowser, I. Cepicka, J. Decelle, M. Dunthorn, A. M. Fiore-Donno, G. H. Gile, M. Holzmann, R. Jahn, M. Jirků, P. J. Keeling, M. Kostka, A. Kudryavtsev, E. Lara, J. Lukeš, D. G. Mann, E. A. Mitchell, F. Nitsche, M. Romeralo, G. W. Saunders, A. G.

Simpson, A. V. Smirnov, J. L. Spouge, R. F. Stern, T. Stoeck, J. Zimmermann, D. Schindel and C. de Vargas, "CBOL Protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms", PLoS Biology **10**, 11, 1–5, URL `www.dx.plos.org/10.1371/journal.pbio.1001419` (2012).

Peterson, B. F. and M. E. Scharf, "Lower termite associations with microbes: Synergy, protection, and interplay", (2016).

Prade, R. A., "Xylanases: from biology to biotechnology", Biotechnol. Genet. Eng. Rev. **13**, 1, 101–132 (1996).

Price, M. N., P. S. Dehal and A. P. Arkin, "Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix", Molecular Biology and Evolution **26**, 7, 1641–1650 (2009).

Price, M. N., P. S. Dehal and A. P. Arkin, "FastTree 2–Approximately Maximum-Likelihood Trees for Large Alignments", PLoS One **5**, 3, e9490 (2010).

Radek, R., K. Meuser, J. F. H. Strassert, O. Arslan, A. Teßmer, J. Šobotník, D. Sillam-Dussès, R. A. Nink and A. Brune, "Exclusive gut flagellates of Serritermitidae suggest a major transfaunation event in lower termites: Description of *Heliconympha glossotermitis* gen. nov. spec. nov", J. Eukaryot. Microbiol. **65**, 1, 77–92 (2018).

Rashamuse, K., W. Sanyika Tendai, K. Mathiba, T. Ngcobo, S. Mtimka and D. Brady, "Metagenomic mining of glycoside hydrolases from the hindgut bacterial symbionts of a termite (*Trinervitermes trinervoides*) and the characterization of a multimodular β-1,4-xylanase (GH11)", Biotechnol. Appl. Biochem. **64**, 2, 174–186, URL `www.doi.wiley.com/10.1002/bab.1480` (2017).

Raychoudhury, R., R. Sen, Y. Cai, Y. Sun, V.-U. Lietze, D. Boucias and M. Scharf, "Comparative metatranscriptomic signatures of wood and paper feeding in the gut of the termite Reticulitermes flavipes (Isoptera: Rhinotermitidae)", Insect Mol. Biol. **22**, 2, 155–171, URL `www.doi.wiley.com/10.1111/imb.12011` (2013).

Reuß, J., R. Rachel, P. Kämpfer, A. Rabenstein, J. Küver, S. Dröge and H. König, "Isolation of methanotrophic bacteria from termite gut", Microbiological Research **179**, 29–37, URL `www.dx.doi.org/10.1016/j.micres.2015.06.003` (2015).

Ricard, G., N. R. McEwan, B. E. Dutilh, J. P. Jouany, D. Macheboeuf, M. Mitsumori, F. M. McIntosh, T. Michalowski, T. Nagamine, N. Nelson, C. J. Newbold, E. Nsabimana, A. Takenaka, N. A. Thomas, K. Ushida, J. H. P. Hackstein and M. A. Huynen, "Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment", BMC Genomics **7**, 1–13 (2006).

Rodrigues Mota, T., D. Matias de Oliveira, R. Marchiosi, O. Ferrarese-Filho and W. Dantas dos Santos, "Plant cell wall composition and enzymatic deconstruction", AIMS Bioeng. **5**, 1, 63–77 (2018).

Ronquist, F. and J. P. Huelsenbeck, "MrBayes 3: Bayesian phylogenetic inference under mixed models", Bioinformatics **19**, 12, 1572–1574 (2003).

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard and J. P. Huelsenbeck, "MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space", Syst. Biol. **61**, 3, 539–42 (2012).

Saha, B. C., "Hemicellulose bioconversion", J. Ind. Microbiol. Biotechnol. **30**, 5, 279–291 (2003).

Sanderson, K., "Lignocellulose: A chewy problem", Nature **474**, 7352 SUPPL. (2011).

Sato, T., H. Kuwahara, K. Fujita, S. Noda, K. Kihara, A. Yamada, M. Ohkuma and Y. Hongoh, "Intranuclear verrucomicrobial symbionts and evidence of lateral gene transfer to the host protist in the termite gut", ISME Journal **8**, 5, 1008–1019, URL `www.dx.doi.org/10.1038/ismej.2013.222` (2014).

Scharf, M. E. and D. G. Boucias, "Potential of termite-based biomass pre-treatment strategies for use in bioethanol production", Insect Sci. **17**, 166–174, URL `www.doi.wiley.com/10.1111/j.1744-7917.2009.01309.x` (2010).

Scharf, M. E., Y. Cai, Y. Sun, R. Sen, R. Raychoudhury and D. G. Boucias, "A meta-analysis testing eusocial co-option theories in termite gut physiology and symbiosis", Commun. Integr. Biol. **10**, 2, 1–13, URL `www.dx.doi.org/10.1080/19420889.2017.1295187` (2017).

Seiboth, B. and B. Metz, "Fungal arabinan and L-arabinose metabolism", Appl. Microbiol. Biotechnol. **89**, 6, 1665–1673, URL `www.link.springer.com/10.1007/s00253-010-3071-8` (2011).

Serrano-Ruiz, J. C., R. Luque and A. Sepúlveda-Escribano, "Transformations of biomass-derived platform molecules: from high added-value chemicals to fuels via aqueous-phase processing", Chem. Soc. Rev. **40**, 5266, URL `www.xlink.rsc.org/?DOI=c1cs15131b` (2011).

Soccol, C. R., E. S. F. da Costa, L. A. J. Letti, S. G. Karp, A. L. Woiciechowski and L. P. d. S. Vandenberghe, "Recent developments and innovations in solid state fermentation", Biotechnol. Res. Innov. **1**, 1, 52–71, URL `https://www.sciencedirect.com/science/article/pii/S2452072116300144` (2017).

Stamatakis, A., "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models", Bioinformatics **22**, 21, 2688–2690 (2006).

Su, L. J., L. L. Yang, S. Huang, X. Q. Su, Y. Li, F. Q. Wang, E. T. Wang, N. Kang, J. Xu and A. D. Song, "Comparative gut microbiomes of four species representing the higher and the lower termites", J. Insect Sci. **16**, 1 (2016).

Tai, V., K. J. Carpenter, P. K. Weber, C. A. Nalepa, S. J. Perlman and P. J. Keeling, "Genome evolution and nitrogen fixation in bacterial ectosymbionts of a protist inhabiting wood-feeding cockroaches", Applied and Environmental Microbiology **82**, 15, 4682–4695 (2016).

Tamschick, S. and R. Radek, "Colonization of termite hindgut walls by oxymonad flagellates and prokaryotes in *Incisitermes tabogae*, *I. marginipennis*and *Reticulitermes flavipes*", Eur. J. Protistol. **49**, 1, 1–14, URL `www.dx.doi.org/10.1016/j.ejop.2012.06.002` (2013).

Tarayre, C., J. Bauwens, C. Mattéotti, C. Brasseur, C. Millet, S. Massart, J. Destain, M. Vandenbol, E. De Pauw, E. Haubruge, F. Francis, P. Thonart, D. Portetelle and F. Delvigne, "Multiple analyses of microbial communities applied to the gut of the wood-feeding termite *Reticulitermes flavipes* fed on artificial diets", Symbiosis **65**, 3, 143–155, URL `http://link.springer.com/10.1007/s13199-015-0328-0` (2015).

Tartar, A., M. M. Wheeler, X. Zhou, M. R. Coy, D. G. Boucias and M. E. Scharf, "Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite *Reticulitermes flavipes*", Biotechnol. Biofuels **2**, 25 (2009).

Tholen, A. and A. Brune, "Impact of oxygen on metabolic fluxes and in situ rates of reductive acetogenesis in the hindgut of the wood-feeding termite *Reticulitermes flavipes*", Environ. Microbiol. **2**, 4, 436–449, URL `www.doi.wiley.com/10.1046/j.1462-2920.2000.00127.x` (2000).

Todaka, N., T. Inoue, K. Saita, M. Ohkuma, C. A. Nalepa, M. Lenz, T. Kudo and S. Moriya, "Phylogenetic analysis of cellulolytic enzyme genes from representative lineages of termites and a related cockroach", PLoS One **5**, 1, 1–10 (2010a).

Todaka, N., C. M. Lopez, T. Inoue, K. Saita, J.-i. Maruyama, M. Arioka, K. Kitamoto, T. Kudo and S. Moriya, "Heterologous expression and characterization of an endoglucanase from a symbiotic protist of the lower termite, *Reticulitermes speratus*", Appl. Biochem. Biotechnol. **160**, 4, 1168–1178, URL `www.link.springer.com/10.1007/s12010-009-8626-8` (2010b).

Todaka, N., S. Moriya, K. Saita, T. Hondo, I. Kiuchi, H. Takasu, M. Ohkuma, C. Piero, Y. Hayashizaki and T. Kudo, "Environmental cDNA analysis of the genes involved in lignocellulose digestion in the symbiotic protist community of *Reticulitermes speratus*", FEMS Microbiol. Ecol. **59**, 3, 592–599 (2007).

Van Wyk, J. P., "Biotechnology and the utilization of biowaste as a resource for bioproduct development", Trends Biotechnol. **19**, 5, 172–177 (2001).

Varman, A. M., L. He, R. Follenfant, W. Wu, S. Wemmer, S. A. Wrobel, Y. J. Tang and S. Singh, "Decoding how a soil bacterium extracts building blocks and metabolic energy from ligninolysis provides road map for lignin valorization", Proc. Natl. Acad. Sci. **113**, 40, E5802—-E5811, URL `www.pnas.org/lookup/doi/10.1073/pnas.1606043113` (2016).

Waidele, L., J. Korb, C. R. Voolstra, S. Künzel, F. Dedeine and F. Staubach, "Differential Ecological Specificity of Protist and Bacterial Microbiomes across a Set of Termite Species", Front. Microbiol. **8**, 2518 (2017).

Warnecke, F., P. Luginbühl, N. Ivanova, M. Ghassemian, T. H. Richardson, J. T. Stege, M. Cayouette, A. C. McHardy, G. Djordjevic, N. Aboushadi, R. Sorek, S. G. Tringe, M. Podar, H. G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N. C. Kyrpides, E. G. Matson, E. A. Ottesen, X. Zhang, M. Hernández, C. Murillo, L. G. Acosta, I. Rigoutsos, G. Tamayo, B. D. Green, C. Chang, E. M. Rubin, E. J. Mathur, D. E. Robertson, P. Hugenholtz and J. R. Leadbetter, "Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite", Nature **450**, 7169, 560–565, URL `www.nature.com/articles/nature06269` (2007).

Watanabe, H., K. Nakashima, H. Saito and M. Slaytor, "New endo-β-1,4-glucanases from the parabasalian symbionts, *Pseudotrichonympha grassii* and *Holomastigotoides mirabile* of *Coptotermes* termites", Cell. Mol. Life Sci. **59**, 1983–1992, URL `www.link.springer.com/10.1007/PL00012520` (2002).

Watanabe, H., A. Takase, G. Tokuda, A. Yamada and N. Lo, "Symbiotic "Archaezoa" of the primitive termite *Mastotermes darwiniensis* still play a role in cellulase production", Eukaryot. Cell **5**, 9, 1571–1576 (2006).

Xie, L., L. Zhang, Y. Zhong, N. Liu, Y. Long, S. Wang, X. Zhou, Z. Zhou, Y. Huang and Q. Wang, "Profiling the metatranscriptome of the protistan community in *Coptotermes formosanus* with emphasis on the lignocellulolytic system", Genomics **99**, 4, 246–255, URL `www.dx.doi.org/10.1016/j.ygeno.2012.01.009` (2012).

Yuki, M., H. Kuwahara, M. Shintani, K. Izawa, T. Sato, D. Starns, Y. Hongoh and M. Ohkuma, "Dominant ectosymbiotic bacteria of cellulolytic protists in the termite gut also have the potential to digest lignocellulose", Environ. Microbiol. **17**, 12, 4942–4953, URL `www.doi.wiley.com/10.1111/1462-2920.12945` (2015).

Zhang, D., A. R. Lax, B. Henrissat, P. Coutinho, N. Katiya, W. C. Nierman and N. Fedorova, "Carbohydrate-active enzymes revealed in *Coptotermes formosanus* (Isoptera: Rhinotermitidae) transcriptome", Insect Mol. Biol. **21**, 2, 235–245, URL `www.doi.wiley.com/10.1111/j.1365-2583.2011.01130.x` (2012).

Zhang, Y., J. Ju, H. Peng, F. Gao, C. Zhou, Y. Zeng, Y. Xue, Y. Li, B. Henrissat, G. F. Gao and Y. Ma, "Biochemical and structural characterization of the intracellular mannanase AaManA of *Alicyclobacillus acidocaldarius* reveals a novel glycoside hydrolase family belonging to clan GH-A", J. Biol. Chem. **283**, 46, 31551–31558, URL `www.ncbi.nlm.nih.gov/pubmed/18755688` (2008).

Zhang, Y. P., M. E. Himmel and J. R. Mielenz, "Outlook for cellulase improvement: Screening and selection strategies", Biotechnol. Adv. **24**, 5, 452–481, URL `https://www.sciencedirect.com/science/article/pii/S0734975006000413` (2006).