5-30-2017

Genomic and Transcriptomic Characterization of the Hindgut Symbiosis associated with the Eastern Subterranean Termite, Reticulitermes flavipes

Jacquelynn Benjamino
University of Connecticut - Storrs, jacquelynn.benjamino@uconn.edu

Genomic and Transcriptomic Characterization of the Hindgut Symbiosis associated with the Eastern Subterranean Termite, *Reticulitermes flavipes*

Jacquelynn Benjamino, Ph.D.

University of Connecticut, 2017

The termite, *Reticulitermes flavipes*, is a eusocial insect with the ability to utilize a nutrient-poor diet as its sole food source. The ability of this insect to survive on this diet is strongly dependent upon its microbiota residing in the hindgut. These symbionts are single-cell eukaryotic protists, bacteria, and archaea. Although the bacterial community in the hindgut is complex existing of hundreds of OTUs, it is robust and homogenous throughout a colony as a result of trophallaxis. To better understand the avenues of importance of this tripartite symbiosis, a series of metagenomics and metatranscriptomics studies were performed.

A 16S rRNA gene sequencing analysis of termites from multiple colonies and castes was performed. The core microbiota was found to consist of 17 taxa and consistent in relative abundance across multiple colonies, but the overall microbiota was different among multiple colonies. The bacterial and protist populations were compared among multiple castes and shown to differ, suggesting differences in caste diet causes change in the hindgut community.

A temporal dietary study was performed on termites with different food sources and showed a temporal shift in the hindgut microbiota with each food source causing a different shift. Artificial neural network (ANN) analysis on the bacterial abundance data was used to create a prediction model for the microbial community and determine highly correlated taxa. The ANN suggests

that it may be the low-abundant taxa driving the microbial community instead of the assumed higher-abundant organisms.

Single protist cells were isolated from the hindgut and after total nucleic acids amplification of the protists and associated bacteria, the metagenomes and metatranscriptomes were sequenced and analyzed. Metabolic expression was determined for endosymbiotic bacteria of *Pyrsonympha vertens* and *Trichonympha agilis* protists. Symbiotic-specific genes were found along with differential gene expression between the two bacteria, coinciding with different protist niches.

The data produced in this thesis adds to the knowledge of termite symbioses and provides the first description of a termite core microbiota, evidence for the driving force of low abundant bacteria through ANN, and a metatranscriptomics view into the protist-bacterial symbiosis within the overall termite symbiosis.

Genomic and Transcriptomic Characterization of the Hindgut Symbiosis associated with the Eastern Subterranean Termite, *Reticulitermes flavipes*

Jacquelynn Benjamino

B.S., University of Massachusetts-Boston, 2010

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of Doctor of Philosophy

at the

University of Connecticut

2017

2017

Approval Page

Doctor of Philosophy Dissertation

Genomic and Transcriptomic Characterization of the Hindgut Symbiosis associated with the Eastern Subterranean Termite, *Reticulitermes flavipes*

Presented by
Jacquelynn Benjamino, B.S.

Major Advisor

_____
Joerg Graf

Associate Advisor

_____
Daniel Gage

Associate Advisor

_____
J. Peter Gogarten

Associate Advisor

_____
Jonathan Klassen

Associate Advisor

_____
Ranjan Srivastava

University of Connecticut
2017

Acknowledgements

I would first like to thank Dr. Joerg Graf for his mentorship throughout the last six years. I will be forever grateful for the opportunities he has given me and for the guidance he has provided. I would also like to thank the members of my committee Dr. Gage, Dr. Gogarten, Dr. Klassen, and Dr. Srivastava. All of their help and questions during meetings and brainstorming sessions has made me become comfortable speaking about my research and think more critically about science.

I thank past lab members Dr. Lindsey Bomar, Dr. Sophie Colston, Dr. Michele Maltz, and Dr. Mike Nelson for their help and conversations. I thank current lab members Emily McClure, Meredith Mistretta, Susan Janton, Erin Breaker, and Manny Divinagracia. I especially thank lab members Dr. Jeremiah Marden and Lidia Beka for their countless hours of brainstorming, advice, and words of encouragement. I also thank Therese Tripler and Allison Kerwin for their friendship, advice, and venting sessions during the past six years.

I thank my undergraduate professors, Dr. Rachel Hirst and Professor Frances McCutcheon, for without them I would have never developed a passion for Microbiology. I thank my family and friends for all of their love and support. I especially thank my parents, Wayne and Christine and siblings Joey and Kayla for their support and encouragement not only pertaining to my graduate career, but all aspects of my life. Finally, I dedicate this dissertation to my Papa, Osmond Benjamino who has always taught me through example that I can fight for anything I want and still remain kind and humble (while also always making me feel like a rockstar!), and to all the men and women (past, present, and future) of the United States Military. Because of the selfless dedication of these men and women, we are able to pursue our dreams safely and freely.

Table of Contents

List of Tables

List of Figures

Appendix II

Chapter One

Introduction

Beneficial Symbioses

Bacteria can thrive in many states such as free-living organisms, pathogens and symbionts. Bacterial symbioses are important for many eukaryotic organisms ranging from bacteria aiding in the fertility of the tsetse fly (Pais et al., 2008) to the complexities of the human microbiome in maintaining human health (Human Microbiome Project Consortium, 2012; Quigley, 2017). The interactions within the gut is one important facet of host-bacterial symbiosis. Symbiotic gut bacteria aid in digestion and nutrient release of consumed foods and provide vitamins for the eukaryotic host (Gündüz & Douglas, 2009), and a dysbiosis in the gut microbiome has been shown to contribute to diseases such as asthma, diabetes, and Chron's disease (Quigley, 2017). It has been shown in humans, that different diets such as a high animal fat diet or high carbohydrate diet dictate the composition of bacterial community in the gut, consisting of high *Bacteroidetes* and high *Prevotella*, respectively(Wu et al., 2011). Invertebrates such as the medicinal leech, *Hirudo verbena*, also rely on symbiotic bacteria in the gut to aid in digestion and provide nutrients (Bomar & Graf, 2012; Maltz et al., 2014), which is especially important because the food source is a nutrient-poor blood meal. Most termites also feed exclusively on a nutrient-poor wood diet (Ohkuma, 2003).

Termite Biology

Termites are eusocial insects that live in large colonies made up of multiple castes, each caste performing a specific function for the colony as a whole. Termite eggs develop into nymphs, which can further develop into three casts: workers, soldiers, and reproductives. Upon

molting into a soldier, the termites primary function is to defend the colony from intruders such as predators and termites from other colonies. Reproductive termites develop wings and become alates, which exit the nest and can potentially develop into kings and queens. Alates do not feed once they develop wings and rely on their fat body for nutrition until a new colony is established and they molt into a king or queen (Costa-Leonardo et al., 2013). The king and reproductive queen are necessary for the establishment of a new colony. (Thorne et al., 2010) Nymphs that molt into workers perform the job of feeding on wood and feeding other members of the colony via stomodeal and proctodeal trophallaxis. This is especially important for soldiers because they cannot consume wood due to mandibles that are enlarged for defense purposes. The process of intra-colony feeding is also important when considering the gut microbiota; as termites exchange partially digested food, they are also exchanging their microbial partners, which creates a homogenous microbial population within a colony (Benjamino & Graf, 2016; Hongoh et al., 2005; Minkley et al., 2006). This homogenous microbial community can be used to distinguish between nestmates and termite intruders from other colonies (Matsuura, 2001).

There are a number of termite species, divided into two main phylogenetic categories, lower and higher termites that harbor distinct groups of organisms in their hindgut (Ohkuma, 2008). It has been found that termites descended from blattid cockroaches between 150-170 million years ago, when termites (Isoptera) acquired cellulolytic flagellates (lower termites) (Bourguignon et al., 2015). Around 50 million years ago, higher termites evolved that lost the ability to harbor flagellates but retained bacteria and archaea in the hindgut (Bourguignon et al., 2015). These higher termites adjusted feeding habits to accommodate the inability to digest cellulose and lignocellulose, feeding on fungus, soil humus, and leaf litter (Ohkuma, 2008). Higher termites are widely studied for their gut microbiota's ability to produce methane which is

due to the large population of methanogens living in the hindgut (Deevong et al., 2004). Methanogens are also present in lower termites, but at much lower numbers causing the termites to be less efficient at emitting methane (Brauman et al., 2001). Major bacterial phyla found in higher termites include Bacteroidetes, Firmicutes, Spirochaetes, Proteobacteria, TG3 phylum, Fibrobacteres, along with other low abundant taxa (Dietrich et al., 2014).

Lower termites have been well-studied in regards to the phylogenetic relationship between the termites and their protist symbionts. Cospeciation between the termite family *Rhinotermitidae* and the protist genus *Pseudotrichonympha* has been shown (Tai et al., 2015). Further, cospeciation was found between the *Pseudotrichonympha* protist and its endosymbiont, *Bacteroidales* (Noda et al., 2007). Another important facet of lower termite biology is the ability to utilize recalcitrant wood as a food source, which is made possible by the tripartite symbiosis, making the termite a dangerous pest as well as a valuable study tool. This ability draws researchers to the lower termite as a model for biotechnology purposes such as termiticides as well as a tool for purposes such as biofuel potential by studying lignocellulose degradation enzymes (Scharf, 2015b).

The Termite Digestive System

The termite digestive system is made up of three major sections, the foregut, midgut, and hindgut. The foregut and midgut are where termite enzymes partially breakdown the wood meal, whereas the hindgut is where the microorganisms reside and the majority of digestion occurs (Ohkuma, 2008). Due to the small size insect digestive tracts have a high surface area to volume ratio and because of this there is constant diffusion of oxygen into the hindgut lumen. However, flagellates, bacteria, and archaea that attach to the cuticle act as an oxygen sink, rendering the

lumen of the hindgut anoxic (Brune et al., 1995a; Tamschick & Radek, 2013). The archaea along the hindgut wall are anaerobic but contain catalases which allow them to live in the microaerophilic niche and are hypothesized to utilize the H_2 and CO_2 emitted from the flagellates to create methane (Leadbetter et al., 1999). The flagellates occupy the majority of space in the hindgut, either attached to the hindgut cuticle or in the anoxic paunch. Archaea are attached to the cuticle along with some bacteria. The remaining bacteria exist as free-living organisms in the anoxic paunch, endosymbionts inside the flagellates, ectosymbionts attached to the outside of flagellates, or in biofilms within the paunch.

Wood Digestion

The tripartite symbiosis is required for lower termites to digest a wood meal due to the inability to breakdown the meal on their own. The three components of the symbiosis work synergistically to digest the wood meal and provide nutrients to all members of the symbiosis. In order to fully understand the process from wood meal to nutrients, it is important to study the role each member of the microbiota plays in the hindgut. The first step is to determine the organisms present in the symbiosis, and how they contribute to the digestion of the wood meal. The termite begins the process of wood degradation with enzyme secretion in the foregut and the midgut. These enzymes include endo-β-1,4-glucanase which hydrolyzes cellulose chains, exo-β-1,4-cellobiohydrolase which removes cellobiose units from the end of the cellulose chain, and β-glucosidase which hydrolyzes glucose units from cellobiose or longer chains (Watanabe & Tokuda, 2001). These enzymes breakdown cellulose and cellobiose into smaller molecules such as glucose. Scharf et al. found that glucose non-competitively inhibits β-glucosidase (a termite digestive enzyme) by interacting with the enzyme-substrate complex (Scharf et al., 2011),

creating end-product inhibition for the termite digestive enzymes. By removing glucose the protist and prokaryotic symbionts prevent the inhibition and allow for the complete breakdown of the wood meal and the survival of the termite.

The Eukaryotic Symbionts

The protist population of lower termites belong to two major taxonomic groups, the phylum *Parabasalia* and the order *Oxymonadida*. *Parabasalia* are the more abundant gut flagellates in most lower termites; there are a total of six classes of parabasalids identified and three of them are found exclusively in termites (Brune & Dietrich, 2015). Both parabasalids and oxymonads can phagocytize wood particles and are adapted to the hindgut environment. They have many flagella which aid in movement and stability, along with the ability to attach to the hindgut wall (Tamschick & Radek, 2013). It is unknown exactly how many species of protists exist in the hindgut and reports have been variable depending on the termite species (Brune & Dietrich, 2015; Ohkuma, 2008). Protists in the hindgut have been found to produce enzymes belong to the glycosyl hydrolase families GHF7 (exoglucanases), GHF9 (endoglucanases), and GHF1 (β-glucosidases) (Brune, 2014; Tartar et al., 2009). Sethi et al. found three protist enzymes in the GHF7 family (GHF7-3, GHF7-5, GHF7-6) from the *Reticulitermes flavipes* hindgut that were homologous to protists found in the termite, *Coptotermes formosanus*. These enzymes were expressed between 10-20 fold higher in hindgut content than in hindgut tissue according to qRT-PCR, and 200-400 fold higher in the protist fraction of the hindgut than any other gut fraction (Sethi et al., 2013a).

The Prokaryotic Symbionts

The prokaryotic members of the lower termite hindgut outnumber the flagellates and inhabit many micro-niches. The archaeal members are from the genera *Methanobrevibacter* and *Methanobacterium* and have been found attached to the hindgut wall and inside some protists (Inoue et al., 2008; Leadbetter et al., 1998; Tokura et al., 2000). Leadbetter and Breznak estimated a density of 3.5×10^9 of methanogen cells per ml of hindgut fluid in *Reticulitermes flavipes* (Leadbetter & Breznak, 1996). Although methanogenic archaea are crucial for higher termites and contribute to a large portion of the worlds methane emissions, acetogenic organisms outcompete methanogens in the guts of lower termites (Brauman et al., 1992). The bacterial community of the hindgut is much more diverse and dense. Lower termites harbor bacteria from five major phyla: Spirochaetes, Firmicutes, Bacteroidetes, Proteobacteria, Elusimicrobia, along with other low abundant members (Dietrich et al., 2014). The bacterial population has been found to assist in wood-degradation with the discovery of prokaryotic cellulase, hemicellulase, chitinase, and alpha-carbohydrolase gene expression in the gut of *R. flavipes* (Tartar et al., 2009). However, the bacteria are potentially more valuable to the protists and termite by converting digestion products from a nutritionally poor food source into nutritional products for the eukaryotic members. As wood is low in usable nitrogen for the termite and protists, the bacteria in the hindgut fix nitrogen (N_2) to ammonia (NH_3) (Hongoh et al., 2008b; Lilburn, 2001). These bacteria can be free-living or ecto- and endosymbionts of the flagellate cells. Endosymbiotic bacteria also produce amino acids, co-factors, and other useful compounds for the protists and termite, such as acetate (Hongoh et al., 2008a). Acetate is the main nutrient source for the termite and it is hypothesized that the Spirochaetes in lower termites produce the majority of acetate for the host through reductive acetogenesis (Pester & Brune, 2006). The bacterial population

residing in the termite hindgut is comprised of mostly unculturable species, forcing researchers to find alternative techniques for studying the organisms and their interactions.

'Omics in the Termite System

Due to the inability to culture the vast majority of bacteria on the planet, researchers have utilized multi-omics approaches to address questions regarding the microscopic environment (Handelsman, 2005; Rappé & Giovannoni, 2003). 16S rRNA gene sequencing has rapidly become a popular technique to gain insight into the prokaryotic communities of many environments. The 16S rRNA gene encodes the 16S rRNA, which is a component of the 30S small subunit ribosome in prokaryotes. This gene is especially useful in identifying prokaryotes because of the conserved and hypervariable regions. The conserved regions are such among all prokaryotes, which allows for the design of PCR primers to detect all 16S rRNA genes. The hypervariable regions differ between bacterial OTUs and allow for the classification of organisms. (Woese, 1987) With the fairly recent technology of high-throughput sequencing it has become possible for many scientists to study the abundance of microorganisms using the 16S rRNA gene, resulting in over 169,000 studies published in scientific journals to date (according to Google Scholar, 4/23/2017). The microorganisms in the termite hindgut are especially difficult to culture due to the complexity of the microbiota and the anaerobic environment, and few bacteria have been isolated (Geissinger et al., 2009; Graber et al., 2004; Gupta et al., 2012). However, many 16S rRNA gene studies have characterized the bacterial communities associated with multiple genera of termites, castes, habitats, and diets (Boucias et al., 2013; Huang et al., 2013a; Karl & Scharf, 2015; Minkley et al., 2006; Scharf, 2015a). Metagenomics and metatranscriptomics are other methods for studying unculturable organisms in an environment in

which the entire genomic (DNA) or transcriptomic (mRNA/cDNA) content of the organisms in an environment is sequenced and analyzed on the basis of gene content and gene expression. Previous metagenomics studies have searched for genes present in the function of wood degradation (Mattéotti et al., 2011) and transcriptomic studies have shown expression of these microbial genes along with the expression of termite-specific genes (Tartar et al., 2009). Currently there are several published genomes of free-living and protist-symbiont bacteria associated with the termite. These published genomes include members of Elusimicrobia (Geissinger et al., 2009; Hongoh et al., 2008a; Zheng, Dietrich et al., 2016b), *Treponema* (Graber et al., 2004), Bacteroidales (Hongoh et al., 2008b), Desulfovibrio (Sato et al., 2009), and Deltaproteobacteria (Ikeda-Ohtsubo et al., 2016). While these genomes are great contributions to the understanding of the bacteria in the termite gut symbiosis, there remains much to be discovered.

Artificial Neural Networks

Artificial neural networks (ANNs) have been used as predictive models for a broad spectrum of areas such as sports, medicine, and industrial purposes (Hassan et al., 2017; Nevares et al., 2017; Wang et al., 2017). ANNs are a method for computing algorithms based on the structure of biological networks. Assumptions used in artificial neural network analysis are that information is processed in simple units (neurons), signals are passed between the neurons, each connection between neurons has a weight assigned by the user, and that an activation function is applied to each neuron (Fausett, 1994). The ANN is made up of a series of input nodes, output nodes, and a number of hidden layers that contain intermediate nodes between the input and output. ANNs can be created using unsupervised learning, reinforced learning, and supervised

learning (Gupta, 2013). Supervised learning is a method where the ANN is trained based on known output values, one method being back-propagation. For back propagation learning, the user provides input values, output values, and a desired measure of difference between the correct output and the calculated output. The calculated output values will be compared to the correct outputs, and if the difference measure is greater that the provided allowable difference, a cost is assigned and the ANN will go through another iteration until the difference reaches an allowable value (Rumelhart et al., 1988). The ability to train ANNs as predictive models for symbiotic systems may create an avenue for detection of keystone species, bacterial establishment, and dysbiosis prevention.

Scope of Dissertation

The scope of this thesis is to provide insight into the microbial dynamics within the *R. flavipes* hindgut. Many studies have reported 16S rRNA data on the microbiota within the gut from different species of termite species. These studies have spanned from comparing lower and higher termites to comparing the microbiota of termites fed on different nutrient sources, but sampled pooled guts at only one time-point (Boucias et al., 2013; Dietrich et al., 2014). This thesis provides data on the microbial community of the hindguts from single termites and observes the community temporally. In chapter 2, the bacterial community of termites from multiple colonies was compared and shown to be similar among the core and abundant taxa, but small populations were different between colonies. The microbiota from termites in different castes was also studied, showing that the two dominant bacterial symbionts, *Treponema* and *Endomicrobia* are reduced in the alate class, which also shows a reduction in the protist population. Chapter 3 includes a study of the microbiota of termites fed different substrates.

Previous studies have shown that diet plays an important role in shaping the hindgut microbiota, but the samples were pooled guts at a single time-point (Huang et al., 2013b; Mikaelyan et al., 2015). Although important, these studies provide only a glimpse into the full effect of the diet. In this study, single hindguts were sampled temporally over 56 days. It shows that diet plays a role in changing the hindgut community over time. The sequencing data was used in the creation of an artificial neural network (ANN) and was used to predict the community at a given day when fed a specific diet. The ANN was also used to create correlation networks between the bacterial taxa and determine organisms important in changing the community. Chapter 4 studies the bacterial communities associated with multiple protist species residing in different niches in the hindgut. For this, single protists were isolated and the DNA and cDNA was amplified using multiple displacement amplification (MDA). The amplified DNA and cDNA was used to create metagenome and metatranscriptome libraries for the protists and associated bacteria. Draft genomes of the endosymbionts of *Pyrsonympha ve*rtens and *Trichonympha agilis* were assembled (*Endomicrobium* sp. PV7 and *Endomicrobium* sp. TA21) and used as references for transcriptome mapping. Genes present in metabolic pathways for carbon utilization, amino acid biosynthesis, and vitamin B biosynthesis were found to be expressed in the endosymbionts of the protists *P. vertens* and *T. agilis*. Differences in gene content and expression were found between the bacterial species in regards to aerotolerance and hexuronate catabolism, which are suggestive of niche differences.

Chapter Two

Characterization of the Core and Caste-Specific Microbiota in the Termite, *Reticulitermes flavipes*[+*]

[+] Benjamino J and Graf J (2016) Characterization of the Core and Caste-Specific Microbiota in the Termite, *Reticulitermes flavipes*. Front. Microbiol. 7:171. doi: 10.3389/fmicb.2016.00171

frontiers in Microbiology

ORIGINAL RESEARCH
published: 17 February 2016
doi: 10.3389/fmicb.2016.00171

Characterization of the Core and Caste-Specific Microbiota in the Termite, *Reticulitermes flavipes*

*Jacquelynn Benjamino[1] and Joerg Graf[1,2]**

[1] *Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA,* [2] *Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA*

The hindgut of the termite *Reticulitermes flavipes* harbors a complex symbiotic community consisting of protists, bacteria, and archaea. These symbionts aid in the digestion of lignocellulose from the termite's wood meal. Termite hindguts were sampled and the V4 hyper-variable region of the 16S rRNA gene was sequenced and analyzed from individual termites. The core microbiota of worker termites consisted of 69 OTUs at the 97% identity level, grouped into 16 taxa, and together accounted for 67.05% of the sequences from the bacterial community. The core was dominated by *Treponema,* which contained 36 different OTUs and accounted for ∼32% of the sequences, which suggests *Treponema* sp. have an important impact on the overall physiology in the hindgut. Bray–Curtis beta diversity metrics showed that hindgut samples from termites of the same colony were more similar to each other than to samples from other colonies despite possessing a core that accounted for the majority of the sequences. The specific tasks and dietary differences of the termite castes could have an effect on the composition of the microbial community. The hindgut microbiota of termites from the alate castes differed from the worker caste with significantly lower abundances of *Treponema* and *Endomicrobia*, which dominated the hindgut microbiota in workers and soldiers. Protist abundances were also quantified in the same samples using qPCR of the 18S rRNA gene. *Parabasalia* abundances dropped significantly in the winged alates and the *Oxymonadida* abundances dropped in both alate castes. These data suggest that the changes in diet or overall host physiology affected the protist and bacterial populations in the hindgut. The in-depth bacterial characterization and protist quantification in this study sheds light on the potential community dynamics within the *R. flavipes* hindgut and identified a large and complex core microbiota in termites obtained from multiple colonies and castes.

Keywords: *Reticulitermes flavipes*, Core microbiota, Termite caste microbiota, 16S rRNA gene, Illumina amplicon sequencing

OPEN ACCESS

Edited by:
Mike Taylor,
The University of Auckland,
New Zealand

Reviewed by:
Irene Newton,
Indiana University, USA
Manpreet K. Dhami,
Stanford University, USA

****Correspondence:***
Joerg Graf
joerg.graf@uconn.edu

Specialty section:
This article was submitted to Microbial Symbioses, a section of the journal Frontiers in Microbiology

Received: *01 October 2015*
Accepted: *01 February 2016*
Published: *17 February 2016*

Citation:
Benjamino J and Graf J (2016) Characterization of the Core and Caste-Specific Microbiota in the Termite, Reticulitermes flavipes. Front. Microbiol. 7:171. doi: 10.3389/fmicb.2016.00171

INTRODUCTION

Termites have long been studied because of their uncommon diet and complex hindgut microbiota. Researchers study the termite symbiotic system for the discovery of lignocellulases to aid in biofuel production (Tartar et al., 2009; Sethi et al., 2013), understanding of the coevolution of the host and symbionts (Hongoh et al., 2005), and the ability to manipulate and study the structure and

function of a complex microbiota (Brauman et al., 2001). Termites are descendants of the wood-feeding cockroach *Cryptocercus,* and are separated into two groups, higher and lower termites (Dietrich et al., 2014). Lower termites contain an abundance and diversity of flagellate protozoa that aid them in the digestion of wood and higher termites have been reported not to harbor symbiotic protists (Ohkuma, 2003); however, recently, a low-abundant ciliate has been detected in the guts of several higher termites species (Rahman et al., 2015).

Reticulitermes flavipes, the Eastern subterranean termite, is indigenous to the northeastern United States and harbors a tripartite symbiosis in its hindgut consisting of protozoal, bacterial, and archaeal symbionts (Ohkuma, 2003). The digestive enzymes from *R. flavipes* cannot fully break down the lignocellulosic components of wood, the termites' sole food source, while the hindgut symbionts aid the digestion of these wood particles and provide acetate as a nutrient for the host (Ohkuma, 2003). The composition of the microbiota residing in the hindgut of *R. flavipes* has been previously investigated using culture-independent approaches as reviewed by Scharf (Scharf, 2015). Other studies investigated the *R. flavipes* hindgut microbiota by pooling DNA from several termites and sequencing a variable region of the 16S rRNA gene using 454 pyrosequencing (Ohkuma, 2008). Boucias et al. (2013) reported that the community is comprised of an estimated 581 bacterial operational taxonomic units, OTUs at the 97% identity level with approximately 80% of the symbionts belonging to the phyla *Spirochaetes*, *Elusimicrobia*, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria*. They also evaluated the maintenance and stability of the microbial community in the hindgut and discovered that after *R. flavipes* termites were fed either a lignocellulose or cellulose diet for seven days, 88% of the OTUs in the hindgut microbiota were preserved despite the different diets, while only 12% of the OTUs were variable (Boucias et al., 2013). Proctodeal feeding has been suggested as an important mechanism contributing to this stability of the microbial community wherein the worker caste feeds the other members in the colony via fecal transfer (Buczkowski et al., 2007).

The core microbiota is defined as the organisms shared across multiple samples obtained from the same host, which is likely to play crucial roles in the functionality of that habitat (Turnbaugh et al., 2007). The core community of any symbiotic system is important in the health and maintenance of the symbiosis. Many studies have found the presence of a core microbiota in a variety of hosts, either in the form of a taxonomic core or a functional core (encoded genes) (Huse et al., 2012; Shade and Handelsman, 2012; Turnbaugh and Gordon, 2013; Maltz et al., 2014). Knowing the composition of the core microbiota is important because it ensures the maintenance of functions within the habitat and serves as an anchor for community resistance and/or resilience (Huse et al., 2012; Shade et al., 2012). However, it should be noted that differences in the hindgut microbiota can be critical for nestmate recognition or various caste-related functions (Cleveland, 1925; Minkley et al., 2006). Determining the core in smaller animals such as insects can be more challenging as the samples can be very small and thus some tend to pool samples prior to DNA extraction. While these studies provide important insight into the complexity and stability of the community, pooling samples averages the signal and prevents detection of variation between individuals. As the resulting OTU data is averaged, determining the prevalence in the individuals comprising the sample is impossible, and thus the core microbiota cannot be accurately determined (Hamady and Knight, 2009). These studies still provide valuable information but are distinct from a "core" and we will refer to such conclusions as the common microbiota in this manuscript. The core microbiota in some insects is extremely small, for example the core consists of ten OTUs in the bed bug *Cimex lectularius* (Meriweather et al., 2013), two taxa in *Anopheles gambiae* (Wang et al., 2011), and nine taxa in the honey bee (Moran et al., 2012; Sabree et al., 2012). The common microbiota of the fungus-growing higher termite, *Macrotermitinae*, (Otani et al., 2014) non fungus-growing higher termites, lower termites and cockroaches is shared between eleven phyla residing in all four groups and the five most abundant phyla being *Firmicutes*, *Bacteroidetes*, *Spirochaetes*, *Proteobacteria,* and *Synergistes* (Dietrich et al., 2014; Otani et al., 2014).

We hypothesize that the feeding habits dictate protist abundance, which in turn affects the abundance of protist-associated bacteria such as *Endomicrobia* and *Treponema*. Caste specific micrbiotas have been shown in the honey bee where queens have a higher abundance of *Parasaccharibacter apium* (Kapheim et al., 2015), *Alphaproteobacteria* and a *Firmicute* (Firm-5) (Tarpy et al., 2015), while workers harbor a higher abundance of *Betaproteobacteria* and *Gammaproteobacteria* (Tarpy et al., 2015). Being a eusocial insect, *R. flavipes* colonies have a caste system made up of juveniles, workers, soldiers, and reproductives. The workers forage for food and return to the nest to feed other members. The soldiers' sole purpose is to defend the colony and these individuals have an enlarged mandible, which makes it impossible for them to masticate wood (Cleveland, 1925). Select members of the worker caste morph into alates (winged termites) and harbor a dramatically reduced number of protists in their gut while preparing to swarm (Shimada et al., 2013). After which they lose their wings, pair up and become reproductive termites that establish new colonies. The microbiota in the alates is of particular importance as these termites are the reproductive caste that found new colonies and presumably are the source of the key members of the hindgut microorganisms unless they can be acquired from the environment. The king and queen reproduce for colony growth, and are also fed by the worker caste. The soldier and reproductive castes are thought to have a reduced need for hindgut protists as they do not partake in the initial breakdown of wood into its constituent parts. Therefore, these castes have fewer protists in the hindgut, while kings and queens in mature colonies have no hindgut protists (Shimada et al., 2013).

While researchers are trying to understand the functions of the termite symbionts, fundamental aspects about the microbiota are not well known including the variability in the composition of microbiota between colonies, between different castes and between individual workers obtained from the same colony. We characterized the microbiota by sequencing the V4 region of the 16S rRNA gene from the hindguts of individual *R. flavipes*

obtained from multiple colonies. Our analysis revealed a stable microbial community within the hindgut of workers that is comprised by a large core community, while there were significant differences in the abundance of protists and in the composition of the bacterial community in different castes.

MATERIALS AND METHODS

Termite Collection and Maintenance

Reticulitermes flavipes termites were either collected using cardboard traps about a month after placement, captured directly from a rotting log, or purchased. The locations of termite colonies at time of collection are as follows: Mansfield CT (CT.A, October 2011 & CT.C, August 2012), Willington CT (CT.B, October 2013), Willimantic CT (CT.D, July 2013), Groton MA (MA.B, July 2013), Woods Hole MA (MA.C, July 2013), or purchased from Connecticut Valley Biological Supply Co. in Southampton MA (MA.A, June 2013). Additional alate termites used in the qPCR assays for the caste analyses were collected from an eighth colony (April 2014, Storrs, CT, USA), along with workers from the same colony for comparison. Once in the lab, the termites were placed in plastic containers with moist, autoclaved sand and spruce. Colonies were maintained at room temperature in the dark, and the sand was moistened with water every 3–4 weeks. Each colony, with the exceptions of CT.A and CT.C, were sampled upon collection. Colonies CT.A and CT.C were sampled for 4 months following the collection date. Termites sampled were assumed to have been initially collected from the natural habitat, as no evidence of reproduction was observed during the maintenance of the colonies in the lab.

Molecular Identification of Termites

Termite DNA from each colony was used for sequencing the Cytochrome Oxidase II (COII) gene to ensure the termites were *R. flavipes* (Supplementary Figure 1). Primers used for COII sequencing were a modified version of A-tLEU (5′-CAGATAAGTGCATTGGATTT-3′) and B-tLYS (5′-GTTTAAGAGACCAGTACTTG-3′) from Liu and Beckenbach (1992). Sequences were aligned in Geneious 6.1.7 using a MUSCLE alignment (Kearse et al., 2012). A neighbor-joining consensus tree was created with 100 bootstraps iterations, using COII sequences from this study along with sequences from multiple *Reticulitermes* species from NCBI (accession numbers: KR537205-12, JF7962324.1, KM245774-5, JF796221-2, AF262607.1, AY808093.1, EU253889.1, FJ806884.1, JQ280728-36, JX142171-72, JX142149-54) (Su et al., 2006; Legendre et al., 2008; Lim and Forschler, 2012; Perdereau et al., 2013).

Sample Collection and DNA Isolation

Hindguts were removed from the termite by pulling the thorax and anus apart with forceps (Matson et al., 2007) and placed in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). Samples consisted of single, whole hindguts with the exception of the data for colony CT.A where each data point represents pools from five hindguts (these samples were collected before we had established an efficient methods for single hindgut sampling). Seven colonies were sampled for all analyses, excluding the caste analysis. The number of samples per colony are as follows: CT.A (11 samples), CT.B (5), CT.C (8), CT.D (4), MA.A (9), MA.B (6), MA.C (2). For the caste analysis, 19 samples were taken from seven colonies; soldiers and alates were always matched with workers from the same colony. The number of samples per caste are as follows: workers (nine samples), soldiers (5), winged alates (2), de-winged alates (3). DNA was isolated immediately after collection using a modified (the starting lysis buffer was 500 μL and the final elution volume was 30 μL AE buffer) RBB+C isolation protocol as described by Yu and Morrison (2004). This method uses repeated bead beating along with chemical and high temperature cell lysis, and DNA precipitation followed by RNA and protein removal using a QIAmp DNA Mini Kit (Qiagen®, Germantown, MD, USA) column.

PCR Amplification of 16S rRNA Gene and Library Preparation

Hindgut samples were amplified using the V4 hyper-variable region of the 16S rRNA gene using primers developed by Caporaso et al. (2012). PCR reactions included Phusion® High-Fidelity PCR Master Mix with HF Buffer (New England Biolabs Inc., Ipswich, MA, USA) (50% of total volume), 10 μM forward and reverse primers, ∼10 ng DNA, and dH_2O to the final volume of 25 μL. All reactions were amplified in triplicate using the following parameters: 94°C for 3 min, followed by 30 cycles of 94°C (45 s), 50°C (60 s), and 72°C (90 s), with a final extension of 72°C for 10 min (Nelson et al., 2014). Triplicate reactions were pooled and each sample was tested for size by running a 1% agarose gel.

Amplicons were purified and size selected using Agencourt AMPure XP (Beckman Coulter Inc., Brea, CA, USA) magnetic beads (0.65 × μL of sample volume) to select for 400 bp amplicons according to manufacturer's protocol. Samples were then quantified using a Qubit® dsDNA HS Assay (ThermoFisher Scientific Inc.). Concentrations of each sample were calculated and then diluted to 4 nM. All samples were pooled in equimolar amounts for sequencing.

Sequencing and Data Processing

Samples were sequenced using an Illumina MiSeq (Illumina, San Diego, CA, USA) with custom sequencing primers added to the reagent cartridge (Caporaso et al., 2012) and sequenced 2 × 150 bp (CT.A & CT.C) or 2 × 250 bp. Both sequencing methods sequenced the entire V4 region of the 16S rRNA gene and the same merging and quality control parameters were used on both sets of data. The reads were processed as described previously by Nelson et al. (2014). Briefly, output reads were merged to create single reads spanning the entire 254 bp of the V4 hypervariable region using SeqPrep[1], and the PhiX control reads were removed by mapping to the PhiX genome. Data analysis was performed using QIIME (Caporaso et al., 2010). Low quality reads (less than Q30) were removed and operational taxonomic units (OTUs) were determined by clustering reads to the Greengenes reference

[1]https://github.com/jstjohn/SeqPrep

16S rRNA gene reference dataset (2013-08 release) (DeSantis et al., 2006) at a 97% identity, and then performing *de novo* OTU clustering on reads that failed to cluster to a reference (McDonald et al., 2011; Nelson et al., 2014). Chimeras were then removed and the dataset was filtered to remove singleton and doubleton OTUs and then OTUs present at less than 0.0005% (Bokulich et al., 2012). The data was rarified to 15,000 reads per sample in order to include all samples in this study.

Data Analysis

The core microbiota was determined using all of the samples collected from individual *R. flavipes* workers in this study (excluding colony CT.A). Using QIIME, with Greengenes (2013-08) and the DictDb database (Mikaelyan et al., 2015), we calculated the OTUs (at the 97% identity level) that were present in at least 95% of the samples (Huse et al., 2012). These OTUs were then paired with taxonomy to the lowest level of classification, and the sequence abundance of each core OTU was reported. The sequences in the DictDb database were shortened to only include the V4 hypervariable region and combined with the Greengenes database. The combined file was aligned to generate the aligned reference. Some sequences failed to align due to shortness in length and were removed from the unaligned reference file (5,845 sequences out of 55,394).

The samples used in the geographic analysis were all from the worker caste. After quality filtering and rarifying to 15,000 sequences per sample, alpha diversity (Shannon and Phylogenetic diversity) (Faith and Baker, 2006) and beta diversity metrics (Bray–Curtis) (Lozupone and Knight, 2005; Lozupone et al., 2006; Anderson et al., 2011) were performed using QIIME 1.8 and R 3.2.0 (R Development Core Team, 2005; Wickham, 2009; Oksanen et al., 2015). The PERMANOVA statistical analysis was performed to determine the significance of microbial community differences among the different colonies using the Bray–Curtis dissimilarity matrix in QIIME (Caporaso et al., 2010). This analysis was performed over 999 permutations and returned a Pseudo-F (f) statistic along with a *p*-value *(p)*.

Reticulitermes flavipes worker, soldier, and alate hindgut samples collected from various colonies were used in the analysis of the caste microbiota. For each non-worker (soldier, alate), a worker was collected at the same time from the same colony for comparison. The winged alates were collected on February 27, 2013, and de-winged alates were collected on May 31, 2013. The bacterial taxonomic abundances were averaged for each caste, and the averages were used in the analysis. Similar to the statistical analysis done on the microbial communities from different colonies, the PERMANOVA statistical analysis was used to determine the significance of the microbial community differences between workers and soldiers, workers and winged alates, and workers and de-winged alates. The sequences (OTU assignments using Greengenes) in the caste dataset were compared to the DictDb database, which is a curated database for microbes from termites and cockroaches that provides greater taxonomic resolution, using BLASTN at the 97% identity level (Altschul et al., 1990; DeSantis et al., 2006; Mikaelyan et al., 2015). *Spirochaete* sequences with 100% query coverage were assigned OTUs and taxonomy using the DictDb database. *Treponema* sequences that did not match reference sequences in the DictDb database or Greengenes database were designated as '*de novo*'. A one-way ANOVA with a Bonferroni post-test was performed for each taxonomic grouping of *Spirochaetes* using GraphPad Prism version 6.0f for Mac OSX[2] (GraphPad Software, San Diego, CA, USA).

Quantitative PCR (qPCR) of Protist Symbionts

Caste hindgut samples with 16S rRNA sequencing data were used for qPCR analysis. Additional alate samples (and workers from the same group) were added to the analysis, the number of samples per caste are: workers (19 samples), soldiers (3), winged alates (12), de-winged alates (6). Primer sets for the two groups of protists found in the hindgut were designed using 18S rRNA gene sequences from NCBI in Geneious (phylum *Parabasalia* and order *Oxymonadida*) (Kearse et al., 2012). The primers were then tested on hindgut contents, termite DNA, and bacterial DNA to ensure there was no amplification of termite or bacterial DNA. Primer sequences are as follows: Para361F-5′CGCGAAACTTACCCACTCG-3′, Para 510R-5′TTACCGCAGCTGCTGGC-3′ and Oxy161bF-5′CGGA TAGCCGTAGTAATTCTAGAGCT-3′, Oxy352bR-5′AACGTCA GGTTGATAGGTTAGAAATT-3′. All reactions were setup in a 10 μL volume including: SsoAdvanced SYBR Green Supermix (Bio-Rad Laboratories Inc., Hecrules, CA, USA) (50% of reaction volume), 10 μM forward and reverse primers (15% each of reaction volume), dH_2O (10% of reaction volume) and 1 μL of DNA template. Reactions were amplified in triplicate using a CFX96 Real-Time Thermocycler (Bio-Rad Laboratories Inc., Hercules, CA, USA) with the following parameters for the *Oxymonadida*: 95°C (3 min), followed by 40 cycles of 95°C (30 s), 64°C (30 s), and 72°C (30 s). The parameters for the *Parabasalia* were the same except that the annealing temperature was 67°C. Negative controls with no template added were prepared and tested with each set of reactions. Standard curves were generated for each primer set using 10^2–10^8 copies per reaction and the real-time data was normalized to the concentration of DNA added to the PCR reaction to calculate the C_t value, representing a single hindgut and then square-root transformed for statistical analyses. A one-way ANOVA with a Bonferroni post-test was performed for each caste in both protist groups using GraphPad Prism version 6.0f for Mac OSX[2] (GraphPad Software, San Diego, CA, USA).

Data Availability

The 16S rRNA gene sequence data was deposited in the European Nucleotide Archive (ENA) SRA under project ID PRJEB5527.

The COII gene sequence data was deposited in GenBank under accession numbers: KR537205-12.

[2]www.graphpad.com

RESULTS

Identifying the Core Microbiota of *R. flavipes*

Determining a core microbiota is important for any host-associated or environmental community because one can infer the composition of the "healthy" or undisturbed community and a diseased one (Turnbaugh et al., 2007; Shade and Handelsman, 2012). In our study, we defined the core microbiota as the OTUs at the 97% identity level that were present in 95% of the samples (Huse et al., 2012). Our data showed that the core of the worker termites consisted of 69 OTUs and accounted for 67.05% of the sequences of the hindgut microbiota (**Table 1**). Of these OTUs, the genus *Treponema,* contained 41 OTUs, 5 of them being classified to the species *T. primita*, and accounted for almost 41.43% of the total sequences in the hindgut. The class *Endomicrobia* (8 OTUs) and genus *Azobacteroides* (3 OTUs) had an abundance of 18.50 and 3.18%, respectively. The remaining 17 OTUs fell into 13 taxa and comprised 3.95% of sequences from the hindgut. 32.95% of OTUs found in the termite hindgut varied between individuals and were not considered to be part of the core microbiota. One OTU (0.16%) was identified from the DictDb (cockroach and termite specific) database and was unassigned. Box and whisker plots showing taxon abundances (**Figure 1**) were created for each colony, using the taxa found in the core microbiota. These data support the average taxon abundances shown in the core (**Table 1**). All colonies show similar abundance patterns for each of the core taxa.

TABLE 1 | The *R. flavipes* worker core hindgut microbiota.[a]

Core taxonomy			
Class	**Taxon[b]**	**Number of OTUs in taxon**	**Average abundance**
Spirochaetes	*Treponema (g)*	36	32.64%
Endomicrobia	*Endomicrobia(c)*	8	18.50%
Spirochaetes	*primitia (s)*	5	8.79%
Bacteroidia	*Azobacteroides (g)*	3	3.18%
Alphaproteobacteria	*Rickettsiales (o)*	4	0.74%
Mollicutes	*Mycoplasmataceae (f)*	1	0.61%
Epsilonproteobacteria	*Campylobacterales (o)*	1	0.44%
Bacilli	*Lactococcus (g)*	1	0.39%
Deltaproteobacteria	*Desulfovibrio (g)*	1	0.32%
Synergistia	*TG5 (g)*	1	0.25%
Bacteroidia	*Bacteroides (g)*	1	0.24%
Betaproteobacteria	*Propionivibrio (g)*	1	0.22%
Bacteroidia	*Dysgonomonas (g)*	1	0.19%
Opitutae	*HA64 (o)*	2	0.16%
Unassigned	*Unassigned (d)*	1	0.16%
Clostridia	*Ruminococcaceae (f)*	1	0.12%
Bacteroidia	*Bacteroidales (o)*	1	0.11%
	Total	**69**	**67.05%**

[a]*The R. flavipes core microbiota includes taxa in which the OTUs (at the 97% identity level) were found in at least 95% of the samples sequenced.*
[b]*Taxonomic level is designated in parentheses as: (d) domain, (p) phylum, (c) class, (o) order, (f) family, and (g) genus. Taxonomic level listed for each organism is the lowest classification available.*

Analysis of the Hindgut Microbiota Among Different Colonies

The microbiota of xylophagus insects, such as the wood-feeding cockroach, *Cryptocercus kyebangensis*, is shared between members of the colony through proctodeal trophalaxis (Park et al., 2002). This process is thought to create a homogenous microbial community throughout the colony, which may aid in digestion and colony health. To determine the homogeneity of the *R. flavipes* hindgut microbiota, individual and pooled worker hindguts were sampled from seven colonies originating in Massachusetts or Connecticut. Alpha diversity analyses were performed and the values for each sample within a colony averaged and a single value was presented for each grouping (**Table 2**). The Shannon index and equitability show that the microbial community is not evenly distributed. In the case of colonies CT.A and CT.C (sampled over 4 months in the

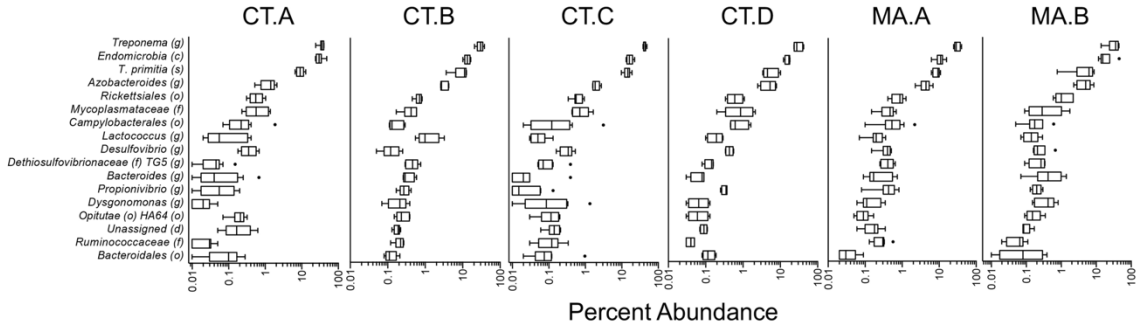


FIGURE 1 | The abundance of core taxa in the worker hindgut. A box and whisker plot for each colony is shown with the abundance of the taxa found in the core microbiota. The abundance of each taxon is consistent throughout the six. The number of samples in each colony are as follows: CT.A ($n = 11$), CT.B ($n = 5$), CT.C ($n = 8$), CT.D ($n = 4$), MA.A ($n = 9$), MA.B ($n = 6$). Colony MA.C was not used in this analysis due to its small sample size ($n = 2$). Colony CT.A contained five pooled hindguts per sample and was not used in determining the core microbiota, however, it showed the same pattern of core taxon abundances.

TABLE 2 | Bacterial alpha diversity of the *R. flavipes* worker hindgut among different colonies based on the 16S rRNA amplicon.

Colony*	n^+	Shannon index (H')	Shannon equitability (E_H)	Phylogenetic diversity (PD)
CT.A	11	3.54 ± 0.16	0.62 ± 0.02	48.25 ± 3.11
CT.B	5	4.57 ± 0.17	0.72 ± 0.02	88.26 ± 2.40
CT.C	8	3.62 ± 0.27	0.62 ± 0.03	55.16 ± 9.53
CT.D	4	4.56 ± 0.19	0.74 ± 0.03	72.40 ± 3.40
MA.A	9	4.70 ± 0.23	0.73 ± 0.03	92.22 ± 2.70
MA.B	6	4.39 ± 0.31	0.71 ± 0.04	70.40 ± 5.52
MA.C	2	4.48 ± 0.21	0.73 ± 0.02	79.30 ± 6.79

**Colony name denotes which state the colony derived from, Connecticut (CT) or Massachusetts (MA).*
$^+$n represents the number of termites sampled from each colony.

lab), the microbial community becomes less complex over time (One-way ANOVA; ($F_{(6,38)} = 36.54$, $p < 0.0001$)). The phylogenetic diversity differs among the microbial communities in different colonies, with some having a less diverse microbiota than others. At the phylum level, *Spirochaetes* dominate the hindgut community, with an average sequence abundance of 55% among the 45 *R. flavipes* workers. *Elusimicrobia* and *Bacteroidetes* are present at 24 and 10%, respectively. The remaining 11% of sequences belong to the phyla *Proteobacteria*, *Firmicutes*, *Tenericutes*, and *Synergistes*, or were unclassified *Bacteria* (**Figure 2**, Supplementary Figure 2 and Supplementary Table 1). In our study, the archaeal community was present at an abundance of less than 0.1%. This could be due to the specificity of the primers used for amplifying the V4 region of the 16S rRNA gene, but other studies using different approaches also reported that in lower termites, archaea are present at low abundances (Berchtold et al., 1999; Hongoh, 2010).

We wanted to assess whether the microbial community residing in the hindgut of one termite was more similar to other termites in the same colony than to termites from a different colony in similar geographic locations. The Bray–Curtis beta diversity analysis was performed to determine similarities and differences in the composition of the microbiota and was used to create a NMDS (non-metric multi-dimensional scaling) plot of the 45 hindgut samples (**Figure 3**). The microbial communities within a colony grouped significantly closer together than to communities from other colonies (PERMANOVA, $f = 8.62$, $p = 0.001$) and there was no clustering of samples according to the state from which they originated nor according to the COII sequence of the termite.

Comparison of the Microbiota Among Different Castes

A termite colony is composed of various castes, each with a unique function contributing that might influence the composition of the hindgut microbiota (Lewis and Forschler, 2013). Alpha diversity analysis of the members from each caste

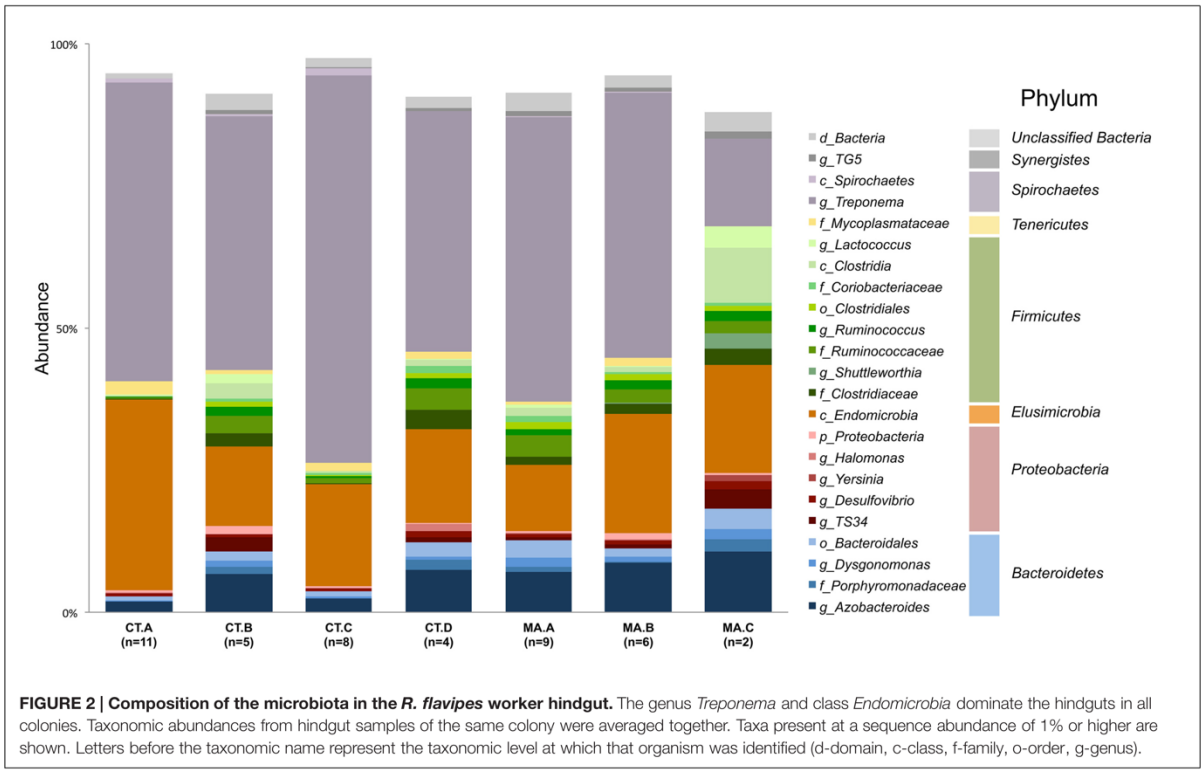


FIGURE 2 | Composition of the microbiota in the *R. flavipes* worker hindgut. The genus *Treponema* and class *Endomicrobia* dominate the hindguts in all colonies. Taxonomic abundances from hindgut samples of the same colony were averaged together. Taxa present at a sequence abundance of 1% or higher are shown. Letters before the taxonomic name represent the taxonomic level at which that organism was identified (d-domain, c-class, f-family, o-order, g-genus).

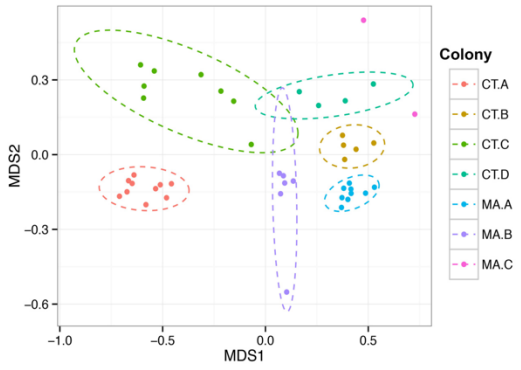


FIGURE 3 | Colony specificity of the *R. flavipes* worker hindgut microbiota. The hindgut microbiota shows more similarity between hindguts from the same colony than between colonies (PERMANOVA, $F = 6.6201$, $p = 0.001$). Nonmetric Multi-Dimensional Scaling was performed on the hindgut microbiota from different colonies using the Bray–Curtis dissimilarity metric. The ellipses were obtained from the standard deviation within a colony, and plotted using a 95% confidence level. The differences in the MA.C colony could be due to the coastal environment compared to the inland environment of the other colonies.

was performed, and the evenness and richness of the microbial community in each caste were similar (**Table 3**). While a trend of a difference in the phylogenetic diversity of each caste was noted, it was not statistically significant, perhaps future studies with a larger sample size might support this finding.

Each caste is known to have different diets and perform specialized functions in the colony, which suggest that the hindgut microbiota may reflect these differences (Lewis and Forschler, 2013). Averaging the sequence abundances at the taxonomic order level for each caste and comparing the values to the worker caste enabled a comparison of the microbial composition in different castes. We found differences in taxonomic abundances between alates and workers according to a PERMANOVA (winged: $f = 3.59$, $p = 0.001$; de-winged: $f = 2.27$, $p = 0.01$). Sequences representing the two dominating taxa, order *Spirochaetales* and class *Endomicrobia*, decreased in abundance in the winged alates from 48 and 22% to 11.6 and 1.1%, respectively, (**Figure 4** and Supplementary Figure 3). Sequences belonging to the order *Bacteroidales* were found in the workers and soldiers at an abundance of less than 11% while they were present at over 20% in both the winged and de-winged alates. The orders *Enterobacteriales*, *Flavobacteriales*, and *Pseudomonadales* were present at average abundances of 9.7, 13.6, and 9%, respectively, in the winged alates while they were below the limit of detection in workers, soldiers, and de-winged alates (0–0.06%). The *Spirochaete* sequences were further classified using OTU assignments and taxonomic classifications from the DictDb database (**Figure 5**) (Mikaelyan et al., 2015). The sequences were classified into five groups: *Treponema* Ia, Ib, Ic, Ig, II, and sequences that were not similar to any in the DictDb database were labeled as *de novo Treponema*. The subgroup *Treponema* Ia were the most abundant taxon in the hindgut with relative abundances reaching up to 33%. *Treponema* Ib were the least abundant and most consistent subgroup with average abundances of 1% for all castes. The abundances of *Treponema* Ia [$F_{(3,15)} = 9.331$, $p = 0.001$] and *Treponema* II [$F_{(3,15)} = 4.489$, $p = 0.0190$] were significantly lower in the winged alates when compared to the worker caste (one-way ANOVA with Bonferonni-corrected p-values), which may indicate that *Treponema* Ia and II are necessary in the digestion process of workers and associated with protists.

TABLE 3 | Bacterial alpha diversity of the *R. flavipes* hindgut among different castes based on the 16S rRNA amplicon.

Caste	n^+	Shannon index (H′)	Shannon equitability (E_H)	Phylogenetic diversity (PD)
Workers	9	4.14 ± 0.48	0.68 ± 0.06	68.44 ± 9.11
Soldiers	5	3.86 ± 0.25	0.65 ± 0.03	56.70 ± 6.88
Winged alates	2	3.17 ± 1.49	0.56 ± 0.19	46.46 ± 25.18
De-winged Alates	3	3.73 ± 0.12	0.63 ± 0.02	61.89 ± 3.25

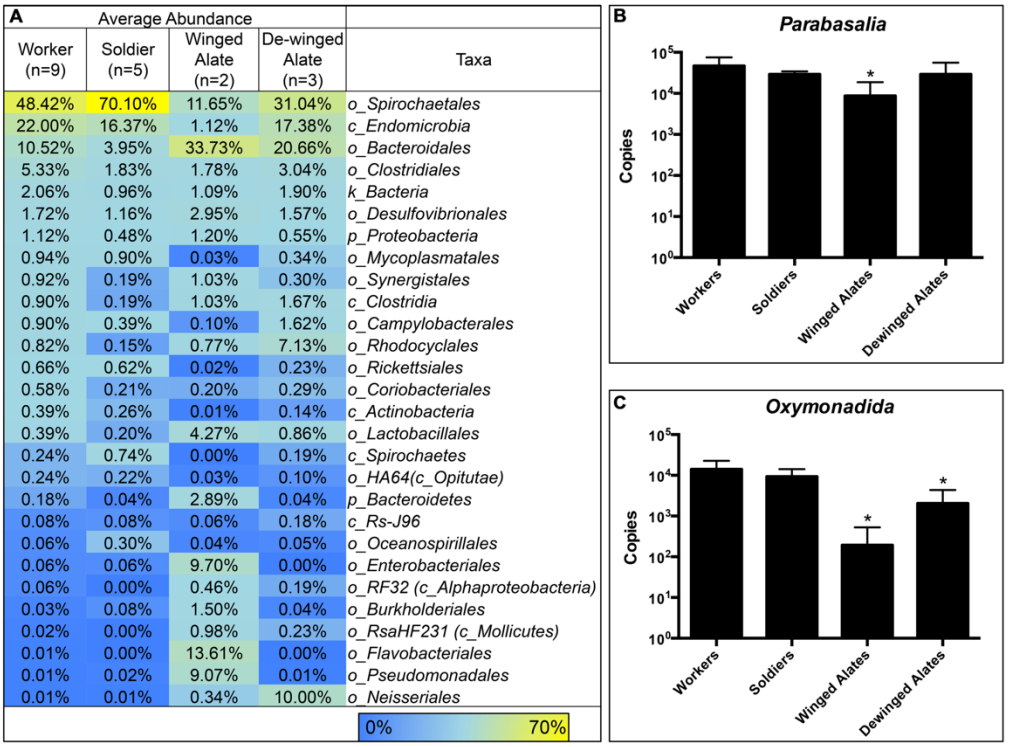
^+n *represents the number of termites sampled from each caste.*

Endomicrobia, along with some *Spirochaetes* are known protist symbionts in the termite hindgut (Iida et al., 2000; Stingl et al., 2005; Ikeda-Ohtsubo et al., 2007). The observed decrease in the abundance of these bacteria in winged alates led us to investigate the protist abundances in the same samples. Abundances of two groups of protists, phylum *Parabasalia* and order *Oxymonadida*, were determined using qPCR. The protist abundances in each caste were compared to the worker caste using a one-way ANOVA. *Parabasalia* abundances were 10-fold fewer in the winged alate class, compared to the worker caste [$F_{(3,36)} = 11.9$, $p < 0.0001$]. Protists belonging to the order *Oxymonadida* were less abundant in the winged alates and de-winged alates [$F_{(3,36)} = 36.94$, $p < 0.0001$]. We were interested in determining if the abundance of bacterial OTUs correlated with the abundance of protists and tested for this by calculating Pearson correlations between each protist group and the *Treponema* (*Spirochaetes*) or *Endomicrobia* OTUs. The two-tailed p-values were Bonferroni-corrected to account for the number of comparisons performed. Twenty-three of the 49 *Treponema* OTUs correlated with the *Oxymonadida* protists, and only ten of forty-nine correlated with the *Parabasalia* ($p < 0.001$, Supplementary Figure 4A). Of these nine *Treponema* OTUs were identical matches for sequences found in the DictDb database and belong to *Treponema* subgroups Ia, Ic, and Ig the OTU IDs are listed (Supplementary Figure 4A). Eight out of twelve *Endomicrobia* correlated with *Parabasalia*, and seven of 12 correlated with *Oxymonadida* protists ($p < 0.004$, Supplementary Figure 4B).

DISCUSSION

The *R. flavipes* Core Microbiota

The presence of a core microbiota and its composition provides insight into the structure of the microbial community in



A	Average Abundance			
Worker (n=9)	Soldier (n=5)	Winged Alate (n=2)	De-winged Alate (n=3)	Taxa
48.42%	70.10%	11.65%	31.04%	*o_Spirochaetales*
22.00%	16.37%	1.12%	17.38%	*c_Endomicrobia*
10.52%	3.95%	33.73%	20.66%	*o_Bacteroidales*
5.33%	1.83%	1.78%	3.04%	*o_Clostridiales*
2.06%	0.96%	1.09%	1.90%	*k_Bacteria*
1.72%	1.16%	2.95%	1.57%	*o_Desulfovibrionales*
1.12%	0.48%	1.20%	0.55%	*p_Proteobacteria*
0.94%	0.90%	0.03%	0.34%	*o_Mycoplasmatales*
0.92%	0.19%	1.03%	0.30%	*o_Synergistales*
0.90%	0.19%	1.03%	1.67%	*c_Clostridia*
0.90%	0.39%	0.10%	1.62%	*o_Campylobacterales*
0.82%	0.15%	0.77%	7.13%	*o_Rhodocyclales*
0.66%	0.62%	0.02%	0.23%	*o_Rickettsiales*
0.58%	0.21%	0.20%	0.29%	*o_Coriobacteriales*
0.39%	0.26%	0.01%	0.14%	*c_Actinobacteria*
0.39%	0.20%	4.27%	0.86%	*o_Lactobacillales*
0.24%	0.74%	0.00%	0.19%	*c_Spirochaetes*
0.24%	0.22%	0.03%	0.10%	*o_HA64(c_Opitutae)*
0.18%	0.04%	2.89%	0.04%	*p_Bacteroidetes*
0.08%	0.08%	0.06%	0.18%	*c_Rs-J96*
0.06%	0.30%	0.04%	0.05%	*o_Oceanospirillales*
0.06%	0.06%	9.70%	0.00%	*o_Enterobacteriales*
0.06%	0.00%	0.46%	0.19%	*o_RF32 (c_Alphaproteobacteria)*
0.03%	0.08%	1.50%	0.04%	*o_Burkholderiales*
0.02%	0.00%	0.98%	0.23%	*o_RsaHF231 (c_Mollicutes)*
0.01%	0.00%	13.61%	0.00%	*o_Flavobacteriales*
0.01%	0.02%	9.07%	0.01%	*o_Pseudomonadales*
0.01%	0.01%	0.34%	10.00%	*o_Neisseriales*

FIGURE 4 | Comparison of Bacterial Taxa and Protists in *R. flavipes* Castes. *Treponema*, *Endomicrobia*, and protist abundances are lower in the alate castes compared to the worker caste. Hindgut samples were taken from worker and non-worker termites in various colonies. **(A)** Abundances for each caste were averaged and compared to the worker caste. In winged alates, the abundance of the two dominating taxa, *Spirochaetales,* and *Endomicrobia*, are lower in abundance compared to the worker caste. *Enterobacteriales*, *Flavobacteriales*, *Pseudomonadales*, and *Neisseriales* are present in very low abundances in the workers and soldiers, but were more abundant in the alate caste. Protists belonging to the phylum *Parabasalia* **(B)** and the order *Oxymonadida* **(C)** were quantified using qPCR. Each caste was compared to the worker caste using a one-way ANOVA. * Indicates a significant value compared to the worker caste. **(B)** *Parabasalia* protists were less abundant in the winged alates [$F_{(3,36)} = 11.9$, $p < 0.0001$]. **(C)** *Oxymonadida* protists were less abundant in the winged and de-winged alates compared to workers [$F_{(3,36)} = 36.94$, $p < 0.0001$].

the habitat of interest, and suggests the metabolic potential and conserved functions of the community (Tap et al., 2009; Huse et al., 2012; Shade and Handelsman, 2012). In termites, many groups have researched the hindgut microbial population, and the presence of a core community has been suggested (Fisher et al., 2007; Boucias et al., 2013; Huang et al., 2013; Scharf, 2015), however, the actual core microbiota remains to be defined as previous groups relied on pools of animals in their analysis which hides the variation amongst individuals. The common microbiota present in nine species of fungus-growing termites was determined, which included 42 taxa comprising eight phyla. The majority of sequences were assigned to two taxa, *Bacteroidetes* and *Firmicutes* (78.6% of sequences) (Otani et al., 2014). In a large survey, Dietrich et al. pooled gut homogenates from 3 to 10 individuals and reported the similarities and differences of the gut microbiota among cockroaches, lower termites, and higher termites. That study found that between 77 and 79% of the sequenced reads, and between 50 and 87 genus-level taxa were assigned to the shared microbiota in each of the three groups tested, cockroaches, lower termites, and higher termites (Dietrich et al., 2014).

For determining the core microbiota, we optimized the DNA extraction protocol for working with individual hindguts. This optimization allowed us to determine variation between individuals when calculating the core microbiota. Determining the composition of the hindgut microbiota from 45 termites obtained from seven different colonies (up to 250 km apart) aided in determining a taxonomic and OTU-based core community (**Table 1**). The more abundant taxa that we identified to comprise the core are identical to the phyla previously reported in the hindgut of *R. flavipes: Spirochaetes*, *Elusimicrobia*, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria* (Fisher et al., 2007; Ohkuma, 2008; Boucias et al., 2013; Huang et al., 2013). Included in these phyla are the abundant taxa, *Treponema*, *Endomicrobia*, and *Azobacteroides*. Less abundant taxa comprising the core include *Desulfovibrio*, *Lactococcus*, *Bacteroidales*, which have been reported to be

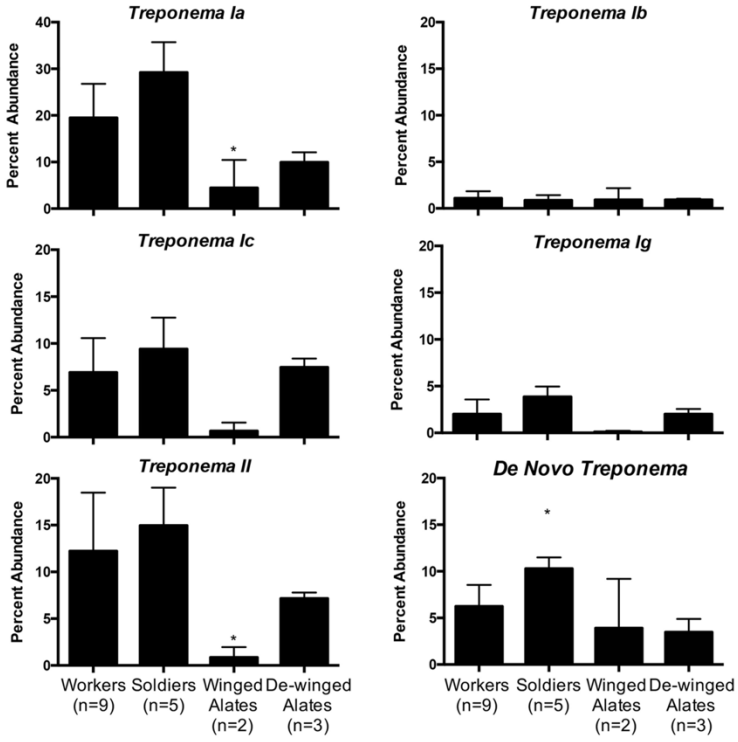


FIGURE 5 | Relative abundance of *Treponema* groups among different castes. Five groups of *Treponema* differ in abundance in workers, soldiers, winged alates, and de-winged alates. OTUs classified as *Treponema* were categorized into groups Ia, Ib, Ic, Ig, and II according to the DictDb database. OTUs that could not be classified in DictDb were processed *de novo* and classified to the genus *Treponema*. *Treponema Ia* is the most abundant group while *Treponema Ib* is the least abundant but most consistent throughout the castes. *Treponema Ia* and *II* have significantly lower abundances in winged alates compared to workers as indicated by a one-way ANOVA [$F_{(3,15)} = 9.331$, $p = 0.001$, and $F_{(3,15)} = 4.489$, $p = 0.019$]. *De Novo Treponema* have a significantly higher abundance in soldiers compared to workers as indicated by a one-way ANOVA [F(3,15)=7.043, p= 0.0035]. All significant data are indicated by an asterisk *.

present in the hindgut of termites (Scharf, 2015). While future studies that expand the geographic range could reduce the number of OTUs comprising the core, the large number of animals and geographic range sampled provides an excellent baseline.

The bacterial, protist, and archaeal populations in the wood-feeding termite hindgut are known to play an integral part in the digestion of the wood meal, and without the complex bacterial community, the termite cannot survive (Raina et al., 2004; Rosengaus et al., 2011). The presence of a core microbiota suggests that each member of the core fills a niche in the termite hindgut that is consistently present despite changes in habitat, geography or food source and thus are likely to contribute to the overall health of the termite. We hypothesize that in *R. flavipes*, the core microbiota is made up of 69 OTUs in 16 taxa and accounts for more than 67.05% of the sequences. The dominant taxa found in the core were *Treponema* (41.43%) and *Endomicrobia* (18.50%), and were reported as part of the common microbiota by Dietrich et al (Dietrich et al., 2014). *Treponema* is a protist ectosymbiont as well as a free-living bacterium in the lumen of the hindgut and is the primary producer of acetate via reductive acetogenesis, which is the main nutrient for the termite host (Leadbetter et al., 1999; Graber and Breznak, 2004). *Endomicrobia* have been found to exist as a protist endosymbiont and a free-living bacterium in the hindgut, providing vitamins, and amino acids (Hongoh et al., 2008; Ikeda-Ohtsubo et al., 2010). The genus *Azobacteroides* (3.44%) was also found in the core and has been previously identified as a protist symbiont and nitrogen fixer in the gut of the termite *Coptotermes formosanus* (Raina et al., 2004). Each of the less abundant taxa were comprised of 1-2 OTUs and each accounted for less than 1% of the total sequences but were present in over 95% of the samples. In general, the abundances of the different taxa comprising the core followed similar patterns in the different colonies analyzed in this study (**Figure 1**). The core taxa were determined using colonies CT.B, CT.C, CT.D, MA.A, MA.B, and MA.C because they were represented by single hindgut samples. Although colony CT.A was not used in the calculation of the core microbiota due to pooled sampling methods, the taxa found in the core were found in similar abundances (**Figure 1**). The consistent detection of 13 taxa

that each accounted for less than 1% of the sequences suggests an important, yet still undefined role for these low abundant organisms.

Sequencing depth will also affect the core size, as a greater sequencing depth increases the chance detection of less abundant taxa. By analyzing 15,000 sequences per sample, we had a greater likelihood of including these taxa in the core as opposed to utilizing fewer sequences, which may only detect low abundant taxa sporadically. This has important implications for previous studies of other microbial communities. The greater sampling depth provided by Illumina sequencing is likely to show a larger core microbial community by detecting less abundant sequences in different environments. Sixteen OTUs that we identified in this study were also present in the DictDb cockroach and termite symbiont reference database, which does not include data from *R. flavipes* but includes data from two related species, *R. chinensis* and *R. speratus* (Mikaelyan et al., 2015). A *Treponema* OTU was present at 3% in one sample, and another at 2% in one sample. The rest of the 14 OTUs were present at 1% or less in all samples. It is interesting that the most abundant OTUs were species-specific at least in this case. It has been reported that the hindgut protists vary depending on termite species (Ohkuma, 2008) and this may be the case for hindgut bacteria as well and has been suggested by Dietrich et al. (2014).

Analysis of the Hindgut Microbiota in Termites from Different Colonies

The microbiota has been shown to aid in host health when present in a symbiotic relationship. The maintenance of the bacterial community throughout a colony of bumble bees aids in the protection from the parasite *Crithidia bombi* (Koch and Schmid-Hempel, 2011). Rosengaus et al. (2011) reported the importance of the hindgut microbiota on host survival in the dampwood termite, *Zootermopsis angusticollis* and *R. flavipes*. In that study, a 64% reduction in bacterial diversity and a small, short-term reduction of gut protists occurred when the diet was supplemented with 0.005 g/mL of the antibiotic, Rifampin. Lower survival rates and a reduction of eggs, larvae, and soldiers were observed for both termite species and correlated with the reduced bacterial diversity (Rosengaus et al., 2011). In addition to proctodeal feeding, it has been shown that social grooming and deposition of fecal contents and saliva in foraging galleries spread termite hindgut bacteria throughout a colony. The bacteria (mostly *Actinobacteria* sp.) in these galleries have been shown to breakdown the cell walls of pathogenic fungi (Klassen, 2014) and possibly pathogenic bacteria (Carr et al., 2012). The maintenance of the *R. flavipes* hindgut microbiota within a colony could provide the termite with protection from microbial invaders in addition to the provision of nutrients.

Previous studies have characterized the bacterial taxonomic abundances in *R. flavipes* using the V1–V3 and V5–V6 region (Boucias et al., 2013; Huang et al., 2013; Nelson et al., 2014), and our study with samples from seven colonies in Connecticut and Massachusetts revealed similar taxonomic composition and abundances. The most abundant phyla represented were *Spirochaetes* (∼55%), *Elusimicrobia* (∼24%), and *Bacteroidetes* (∼10%), with lower abundant representatives from the phyla *Firmicutes*, *Proteobacteria*, *Tenericutes*, *Synergistes*, and unclassified *Bacteria*. The relative abundance of sequences in the termite hindgut varied slightly depending on which region of the 16S rRNA gene was sequenced, which is a known caveat of 16S rRNA studies (Janda and Abbott, 2007; Aird et al., 2011; Soergel et al., 2012; Nelson et al., 2014) but did not change the overall composition.

Different colonies in the same geographic area have been suggested to harbor slightly different hindgut microbiotas, which might allow termites to distinguish nest mates from invaders (Hongoh et al., 2005; Minkley et al., 2006). Beta diversity analyses in this study are consistent with this concept of nest specificity as termites within a colony show a greater similarity of the hindgut microbiota than to other colonies. Colonies CT.A and CT.C were sampled over 4 months and each grouped as a colony according to the NMDS plot. This shows the homogenous nature and stability of the hindgut microbiota within a colony that was transferred from the field and maintained in the laboratory. It was interesting to note, however, that these colonies had a lower Shannon Index (H') which is indicative of a lower OTU richness compared to the other colonies. The Shannon Equitability (E_H) was also lower in these two colonies which indicates more evenness compared to the other colonies. The lower richness and lower evenness may be due to the colonies being kept in the lab, whereas the other colonies were sampled directly from their natural habitat, which would be analogous to a "zoo" effect (Ley et al., 2008; Kohl and Dearing, 2014). A Bray–Curtis analysis of the samples in this study showed hindgut microbiotas from the same colony grouping together (**Figure 3**). Colony MA.C from Woods Hole, MA, shows the most differences according to relative abundances of multiple taxa among the colonies (**Figure 2**), however, our analysis of the COII sequences did not reveal a corresponding phylogenetic difference of the hosts. This may be due to environmental conditions such as higher salt concentration in the air, sand-rich soils, and the lack of dense forestry. Leadbetter and Breznak (1996) reported different morphotypes of methanogens found in *R. flavipes* hindguts in Michigan and Woods Hole, MA, USA, which coincides with our findings of differing bacterial taxa. While the core taxonomic abundances show very small differences, it can be that changes in the relative abundance of key taxa or the fluctuation of low abundance taxa between samples, which could be involved in colony recognition.

The Hindgut Microbiota Among Different Castes

While workers are the primary caste many researchers study, the soldier and alate castes play important roles in the colony. Soldiers protect the colony from invaders and cannot morph into any other caste. Alates, a form of reproductive termite, swarm to a new area to establish a new colony, wherein they

will become the primary reproductives (king and queen). During the transition to winged alates, the termites lose the majority of their gut protists and rely on lipids and glycogen stored in the fat body for nutrition (Costa-Leonardo et al., 2013). Alates shed their wings after swarming to an area to start a new colony, morph into primary reproductives and give rise to juvenile termites. The primary reproductives forage on wood, feed the first generation of juveniles until they are ready to provide for the colony and transmit the symbionts to the juveniles unless some symbionts are acquired from the environment (Shimada et al., 2013). Lewis et al demonstrated that protist abundances in the hindgut differ depending on the feeding habits of the caste, with protist abundances being lower in the alate and soldier castes of three *Reticulitermes* species (Lewis and Forschler, 2013). This finding leads to the question of whether or not different castes with different digestive functions harbor the same hindgut bacterial community.

The dramatic drop in abundance of the protozoal symbionts *Spirochaetales* and *Endomicrobia* in the winged alates coincides with the dramatic decrease in protist numbers during this morphing stage, as shown by Shimada et al. (2013). During this time, the termite is building up fat bodies and storing more nutrients in the fat bodies as the animals are preparing to swarm and establish a new colony (Shimada et al., 2013). We performed qPCR to quantify the two protist groups found in the hindgut, *Parabasalia* and *Oxymonadida* on the same hindgut samples that we sequenced the 16S rRNA gene from. These data show a drop in abundance of both protist groups in the winged (*Parabasalia* and *Oxymonadida*) and dewinged alates (*Oxymonadida*), when compared to the worker caste. When evaluating whether OTUs were potential protist symbionts, seven and eight of the 12 *Endomicrobia* OTUs correlated with *Oxymonadida* and *Parabasalia* protists, respectively (Supplementary Figure 4B). *Endomicrobia* exists in the hindgut as a strict endosymbiont in both *Oxymonadida* and *Parabasalia* protists (Ohkuma et al., 2007), which likely accounts for the large percentage of OTUs correlating with either protist group. The *Endomicrobia* OTUs that do not correlate with the abundance of either protist group could be due to sequencing errors, PCR sensitivity, being present inside protists that do not change abundance according to caste differentiation or exist without an obligate association with protists. The *Treponema* OTUs correlate with the *Parabasalia* and *Oxymonadida* as well (10/49 and 23/49, respectively). *Treponema* are known to exist in the hindgut as protist symbionts as well as free-living, which could account for less than half of the OTUs actually correlating with either protist group (Supplementary Figure 4A) (Leadbetter et al., 1999). Using the DictDb database, OTUs belonging to the *Treponema* taxon were further classified into subgroups (*Treponema Ia, Ib, Ic, Ig, II*) and this revealed that *Treponema Ia* was the most abundant taxa (**Figure 5**). *Treponema Ib* was the least abundant among the five groups and was also consistent among the castes, suggesting that this group may be a free-living spirochete. The abundance of *Treponema Ia* and *Treponema II* was significantly lower in winged alates compared to the worker caste, suggesting that these may be protist symbionts. Sequences corresponding to the order *Bacteroidales* nearly doubled in abundance in both the winged and de-winged alates as compared to workers and soldier microbiotas, which could be a result from a greater growth rate or of the *Spirochaetales* and *Endomicrobia* sequences dropping in abundance as these are not absolute but relative values. The spike of *Enterobacteriales*, *Flavobacteriales*, and *Pseudomonadales* sequences in the winged alates suggests the hindgut is in an altered state in the winged alates, which may reflect the physiological needs of the alate during swarming.

Overall Characterization of the *R. flavipes* Hindgut Microbiota

Studying the hindgut microbiota of individual termites from multiple colonies and castes has added to the understanding of the bacterial components of this complex symbiosis. The ability to sequence many hindgut samples has allowed for a more comprehensive comparison of different colonies and various castes, as well as the determination of a core microbiota in the *R. flavipes* species. Defining a core microbiota in the *R. flavipes* hindgut has revealed the presence of relatively constant and complex bacterial populations in the hindgut of workers. The differences in the composition of the bacterial and protist communities in the winged alates and de-winged alates suggest that major changes occur in the termite digestive tract physiology in this caste, perhaps related to the animals not feeding while relying on the fat body, which would lead to "starvation" of the protists, bacteria and archaea in the hind gut. Importantly, the community cannot be depleted of the core members, as it needs to be passed on to the workers from the new colony unless they are acquired from the environment. The maintenance of such a large core community is important as it suggests that a consistent group of microorganisms participates in the complex degradation of lignocellulose in the hindgut and the provision of nutrients that this simple diet is depleted in. The termite holobiont, or the combination of host and symbionts, which together form a functional unit, is complex and likely to be even more complex as insight is gained into viruses or fungi that may be present inside the termite in addition to the archaea, bacteria, and protists (Rohwer et al., 2002; Zilber-Rosenberg and Rosenberg, 2008; Bordenstein and Theis, 2015).

Studying the bacterial and protist populations in the *R. flavipes* hindgut throughout different life stages, colonies, and over time provides a cohesive representation of the community dynamics. The consistent presence of sequences at low percentages suggests that these analyses need to be done with sufficient sensitivity to detect the activities of these members as well, albeit technical caveats make the analysis of less abundant or even rare taxa more challenging (Reeder and Knight, 2009). As the bacteria in the hindgut are not easily cultured outside the host, the ability to manipulate the hindgut community as a whole *in vivo*, for example through environmental changes or dietary changes, allows for this host to become a model for complex symbioses by revealing

principles that are conserved among distantly related digestive tract symbioses (Ruby, 2008; Nelson and Graf, 2012; Maltz et al., 2014).

AUTHOR CONTRIBUTIONS

JB performed the laboratory work and statistical analysis. JB and JG contributed the experimental design and writing of the manuscript.

FUNDING

This research was funded by the National Science Foundation (NSF) division of Emerging Frontiers in Research and Innovation in Multicellular and Inter-kingdom Signaling. Award number 1137249 (R. Srivastava, D. Gage, J. Graf, L. Shor, B. Mustain, and J. Leadbetter).

ACKNOWLEDGMENTS

The authors wish to thank Dr. Michael C. Nelson for creating a script to centralize the QIIME processing, and Kendra Maas of the University of Connecticut Microbial Analysis Research and Services, MARS, Facility for assisting with the use of R. The authors also wish to thank Drs., Jared Leadbetter, Michael C. Nelson, Sophie Colston, and Emily McClure for their valuable feedback on this manuscript and Charles Bridges, Dr. Daniel Gage, Dr. Jared Leadbetter, and Emily McClure for collecting termites.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.00171

REFERENCES

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., et al. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12, 1–14. doi: 10.1186/gb-2011-12-2-r18

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

Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., et al. (2011). Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecol. Lett.* 14, 19–28. doi: 10.1111/j.1461-0248.2010.01552.x

Berchtold, M., Chatzinotas, A., Schonhuber, W., Brune, A., Amann, R., Hahn, D., et al. (1999). Differential enumeration and in situ localization of microorganisms in the hindgut of the lower termite Mastotermes darwiniensis by hybridization with rRNA-targeted probes. *Arch. Microbiol.* 172, 407–416. doi: 10.1007/s002030050778

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2012). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59. doi: 10.1038/nmeth.2276

Bordenstein, S. R., and Theis, K. R. (2015). Host biology in light of the microbiome: ten principles of holobionts and hologenomes. *PLoS Biol.* 13:e1002226. doi: 10.1371/journal.pbio.1002226

Boucias, D. G., Cai, Y., Sun, Y., Lietze, V.-U., Sen, R., Raychoudhury, R., et al. (2013). The hindgut lumen prokaryotic microbiota of the termite *Reticulitermes flavipes* and its responses to dietary lignocellulose composition. *Mol. Ecol.* 22, 1836–1853. doi: 10.1111/mec.12230

Brauman, A., Dore, J., Eggleton, P., Bignell, D., Breznak, J. A., and Kane, M. D. (2001). Molecular phylogenetic profiling of prokaryotic communities in guts of termites with different feeding habits. *FEMS Microbiol. Ecol.* 35, 27–36. doi: 10.1111/j.1574-6941.2001.tb00785.x

Buczkowski, G., Wang, C., and Bennett, G. (2007). Immunomarking reveals food flow and feeding relationships in the eastern subterranean termite, *Reticulitermes flavipes* (Kollar). *Environ. Entomol.* 36, 173–182. doi: 10.1603/0046-225X(2007)36[173:IRFFAF]2.0.CO;2

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624. doi: 10.1038/ismej.2012.8

Carr, G., Poulsen, M., Klassen, J. L., Hou, Y., Wyche, T. P., Bugni, T. S., et al. (2012). Microtermolides A and B from termite-associated *Streptomyces* sp. and structural revision of vinylamycin. *Org. Lett.* 14, 2822–2825. doi: 10.1021/ol301043p

Cleveland, L. R. (1925). The feeding habit of termite castes and its relation to their intestinal flagellates. *Biol. Bull.* 48, 295–308. doi: 10.2307/1536598

Costa-Leonardo, A. M., Laranjo, L. T., Janei, V., and Haifig, I. (2013). The fat body of termites: functions and stored materials. *J. Insect Physiol.* 59, 577–587. doi: 10.1016/j.jinsphys.2013.03.009

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05

Dietrich, C., Köhler, T., and Brune, A. (2014). The cockroach origin of the termite gut microbiota: patterns in bacterial community structure reflect major evolutionary events. *Appl. Environ. Microbiol.* 80, 2261–2269. doi: 10.1128/AEM.04206-13

Faith, D. P., and Baker, A. M. (2006). Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evol. Bioinformat.* 2, 121–128.

Fisher, M., Miller, D., Brewster, C., Husseneder, C., and Dickerman, A. (2007). Diversity of gut bacteria of *Reticulitermes flavipes* as examined by 16S rRNA gene sequencing and amplified rDNA restriction analysis. *Curr. Microbiol.* 55, 254–259. doi: 10.1007/s00284-007-0136-8

Graber, J. R., and Breznak, J. A. (2004). Physiology and nutrition of *Treponema primitia*, an H_2/CO_2-acetogenic spirochete from termite hindguts. *Appl. Environ. Microbiol.* 70, 1307–1314. doi: 10.1128/AEM.70.3.1307-1314.2004

Hamady, M., and Knight, R. (2009). Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res.* 19, 1141–1152. doi: 10.1101/gr.085464.108

Hongoh, Y. (2010). Diversity and genomes of uncultured microbial symbionts in the termite gut. *Biosci. Biotechnol. Biochem.* 74, 1145–1151. doi: 10.1271/bbb.100094

Hongoh, Y., Deevong, P., Inoue, T., Moriya, S., Trakulnaleamsai, S., Ohkuma, M., et al. (2005). Intra- and interspecific comparisons of bacterial diversity and community structure support coevolution of gut microbiota and termite host. *Appl. Environ. Microbiol.* 71, 6590–6599. doi: 10.1128/AEM.71.11.6590-6599.2005

Hongoh, Y., Sharma, V. K., Prakash, T., Noda, S., Toh, H., Taylor, T. D., et al. (2008). Genome of an endosymbiont coupling N_2 fixation to

cellulolysis within protist cells in termite gut. *Science* 322, 1108–1109. doi: 10.1126/science.1165578

Huang, X.-F., Bakker, M. G., Judd, T. M., Reardon, K. F., and Vivanco, J. M. (2013). Variations in diversity and richness of gut bacterial communities of termites (*Reticulitermes flavipes*) fed with grassy and woody plant substrates. *Microbial. Ecol.* 65, 531–536. doi: 10.1007/s00248-013-0219-y

Huse, S. M., Ye, Y., Zhou, Y., and Fodor, A. A. (2012). A core human microbiome as viewed through 16S rRNA sequence clusters. *PLoS ONE* 7:e34242. doi: 10.1371/journal.pone.0034242

Iida, T., Ohkuma, M., Ohtoko, K., and Kudo, T. (2000). Symbiotic spirochetes in the termite hindgut: phylogenetic identification of ectosymbiotic spirochetes of oxymonad protists. *FEMS Microbiol. Ecol.* 34, 17–26. doi: 10.1111/j.1574-6941.2000.tb00750.x

Ikeda-Ohtsubo, W., Desai, M., Stingl, U., and Brune, A. (2007). Phylogenetic diversity of 'Endomicrobia' and their specific affiliation with termite gut flagellates. *Microbiology* 153, 3458–3465. doi: 10.1099/mic.0.2007/009217-0

Ikeda-Ohtsubo, W., Faivre, N., and Brune, A. (2010). Putatively free-living 'Endomicrobia' – ancestors of the intracellular symbionts of termite gut flagellates? *Environ. Microbiol. Rep.* 2, 554–559. doi: 10.1111/j.1758-2229.2009.00124.x

Janda, J. M., and Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses. Perils, and Pitfalls. *J. Clin. Microbiol.* 45, 2761–2764. doi: 10.1128/JCM.01228-07

Kapheim, K. M., Rao, V. D., Yeoman, C. J., Wilson, B. A., White, B. A., Goldenfeld, N., et al. (2015). Caste-specific differences in hindgut microbial communities of honey bees (*Apis mellifera*). *PLoS ONE* 10:e0123911. doi: 10.1371/journal.pone.0123911

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformation* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Klassen, J. L. (2014). Microbial secondary metabolites and their impacts on insect symbioses. *Curr. Opin. Insect Sci.* 4, 15–22. doi: 10.1016/j.cois.2014.08.004

Koch, H., and Schmid-Hempel, P. (2011). Socially transmitted gut microbiota protect bumble bees against an intestinal parasite. *Proc. Natl. Acad. Sci. U.S.A.* 108, 19288–19292. doi: 10.1073/pnas.1110474108

Kohl, K. D., and Dearing, M. D. (2014). Wild-caught rodents retain a majority of their natural gut microbiota upon entrance into captivity. *Environ. Microbiol. Rep.* 6, 191–195. doi: 10.1111/1758-2229.12118

Leadbetter, J. R., and Breznak, J. A. (1996). Physiological Ecology of *Methanobrevibacter cuticularis* sp. nov.and *Methanobrevibacter curvatus* sp. nov., Isolated from the Hindgut of the Termite *Reticulitermes flavipes*. *Appl. Environ. Microbiol.* 62, 3620–3631.

Leadbetter, J. R., Schmidt, T. M., Graber, J. R., and Breznak, J. A. (1999). Acetogenesis from H_2 plus CO_2 by spirochetes from termite guts. *Science* 283, 686–689. doi: 10.1126/science.283.5402.686

Legendre, F., Whiting, M. F., Bordereau, C., Cancello, E. M., Evans, T. A., and Grandcolas, P. (2008). The phylogeny of termites (Dictyoptera: Isoptera) based on mitochondrial and nuclear markers: implications for the evolution of the worker and pseudergate castes, and foraging behaviors. *Mol. Phylogenet.* 48, 615–627. doi: 10.1016/j.ympev.2008.04.017

Lewis, J. L., and Forschler, B. T. (2013). Protist communities from four castes and three species of *Reticulitermes* (Isoptera: Rhinotermitidae). *Ann. Entomol. Soc. Am.* 97, 1242–1251. doi: 10.1603/0013-8746(2004)097[1242:PCFFCA]2.0.CO;2

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., et al. (2008). Evolution of mammals and their gut microbes. *Science* 320, 1647–1651. doi: 10.1126/science.1155725

Lim, S. Y., and Forschler, B. T. (2012). *Reticulitermes nelsonae*, a new species of subterranean termite (Rhinotermitidae) from the Southeastern United States. *Insects* 3, 62–90. doi: 10.3390/insects3010062

Liu, H., and Beckenbach, A. T. (1992). Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. *Mol. Phylogenet. Evol.* 1, 41–52. doi: 10.1016/1055-7903(92)90034-E

Lozupone, C., Hamady, M., and Knight, R. (2006). UniFrac – An online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformat.* 7:371. doi: 10.1186/1471-2105-7-371

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

Maltz, M. A., Bomar, L., Lapierre, P., Morrison, H. G., Mcclure, E. A., Sogin, M. L., et al. (2014). Metagenomic analysis of the medicinal leech gut microbiota. *Front. Microbiol.* 5:151. doi: 10.3389/fmicb.2014.00151

Matson, E., Ottesen, E., and Leadbetter, J. (2007). Extracting DNA from the gut microbes of the termite (Zootermopsis Angusticollis) and visualizing gut microbes. *J. Vis. Exp.* 4:e195.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., Desantis, T. Z., Probst, A., et al. (2011). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618. doi: 10.1038/ismej.2011.139

Meriweather, M., Matthews, S., Rio, R., and Baucom, R. S. (2013). A 454 survey reveals the community composition and core microbiome of the common bed bug (Cimex lectularius) across an Urban Landscape. *PLoS ONE* 8:e61465. doi: 10.1371/journal.pone.0061465

Mikaelyan, A., Kohler, T., Lampert, M., Rohland, J., Boga, H., Meuser, K., et al. (2015). Classifying the bacterial gut microbiota of termites and cockroaches: a curated phylogenetic reference database (DictDb). *Syst. Appl. Microbiol.* 38, 472–482. doi: 10.1016/j.syapm.2015.07.004

Minkley, N., Fujita, A., Brune, A., and Kirchner, W. H. (2006). Nest specificity of the bacterial community in termite guts (*Hodotermes mossambicus*). *Insect. Soc.* 53, 339–344. doi: 10.1007/s00040-006-0878-5

Moran, N. A., Hansen, A. K., Powell, J. E., and Sabree, Z. L. (2012). Distinctive gut microbiota of honey bees assessed using deep sampling from individual worker bees. *PLoS ONE* 7:e36393. doi: 10.1371/journal.pone.0036393

Nelson, M. C., and Graf, J. (2012). Bacterial symbioses of the medicinal leech *Hirudo verbana*. *Gut Microbes* 3, 322–331. doi: 10.4161/gmic.20227

Nelson, M. C., Morrison, H. G., Benjamino, J., Grim, S. L., and Graf, J. (2014). Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS ONE* 9:e94249. doi: 10.1371/journal.pone.0094249

Ohkuma, M. (2003). Termite symbiotic systems: efficient bio-recycling of lignocellulose. *Appl. Microbiol. Biotechnol.* 61, 1–9. doi: 10.1007/s00253-002-1189-z

Ohkuma, M. (2008). Symbioses of flagellates and prokaryotes in the gut of lower termites. *Trends Microbiol.* 16, 345–352. doi: 10.1016/j.tim.2008.04.004

Ohkuma, M., Sato, T., Noda, S., Ui, S., Kudo, T., and Hongoh, Y. (2007). The candidate phylum 'Termite Group 1' of bacteria: phylogenetic diversity, distribution, and endosymbiont members of various gut flagellated protists. *FEMS Microbiol. Ecol.* 60, 467–476. doi: 10.1111/j.1574-6941.2007.00311.x

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B., et al. (2015). *vegan: Community Ecology Package. R Package Version 2.2–1*. Available at: http://cran.r-project.org

Otani, S., Mikaelyan, A., Nobre, T., Hansen, L. H., Koné, N. G. A., Sørensen, S. J., et al. (2014). Identifying the core microbial community in the gut of fungus-growing termites. *Mol. Ecol.* 23, 4631–4644. doi: 10.1111/mec.12874

Park, Y. C., Grandcolas, P., and Choe, J. C. (2002). Colony composition, social behavior and some ecological characteristics of the korean wood-feeding cockroach (*Cryptocercus kyebangensis*). *Zoo. Sci.* 19, 1133–1139. doi: 10.2108/zsj.19.1133

Perdereau, E., Bagneres, A., Bankhead-Dronnet, S., Dupont, S., Zimmerman, M., Vargo, E., et al. (2013). Global genetic analysis reveals the putative native source of the invasive termite, *Reticulitermes flavipes*, in France. *Mol. Ecol. Res.* 22, 1105–1119. doi: 10.1111/mec.12140

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rahman, N. A., Parks, D. H., Wilnlner, D. L., Engelbrektson, A. L., Goffredi, S. K., Warnecke, F., et al. (2015). A molecular survey of the Australian and North American termite genera indicates that vertical inheritance is the primary force shaping termite gut microbes. *Microbiome* 3:5. doi: 10.1186/s40168-015-0067-8

Raina, A., Park, Y., and Lax, A. (2004). Defaunation leads to cannibalism in primary reproductives of the Formosan subterranean termite, *Coptotermes formosanus* (Isoptera: rhinotermitidae). *Ann. Entomol. Soc. Am.* 97, 753–756. doi: 10.1603/0013-8746(2004)097[0753:DLTCIP]2.0.CO;2

Reeder, J., and Knight, R. (2009). The 'rare biosphere': a reality check. *Nat. Methods* 6, 636–637. doi: 10.1038/nmeth0909-636

Rohwer, F., Seguritan, V., and Azam, F. (2002). Diversity and distribution of coral-associated bacteria. *Marine Ecol.* 243, 1–10. doi: 10.3354/meps243001

Rosengaus, R. B., Zecher, C. N., Schultheis, K. F., Brucker, R. M., and Bordenstein, S. R. (2011). Disruption of the termite gut microbiota and its prolonged consequences for fitness. *Appl. Environ. Microbiol.* 77, 4303–4312. doi: 10.1128/AEM.01886-10

Ruby, E. G. (2008). Symbiotic conversations are revealed under genetic interrogation. *Nat. Rev. Microbiol.* 6, 752–762. doi: 10.1038/nrmicro1958

Sabree, Z. L., Hansen, A. K., and Moran, N. A. (2012). Independent studies using deep sequencing resolve the same set of core bacterial species dominating gut communities of honey bees. *PLoS ONE* 7:e41250. doi: 10.1371/journal.pone.0041250

Scharf, M. E. (2015). Omic research in termites: an overview and a roadmap. *Front. Genet.* 6:76. doi: 10.3389/fgene.2015.00076

Sethi, A., Slack, J. M., Kovaleva, E. S., Buchman, G. W., and Scharf, M. E. (2013). Lignin-associated metagene expression in a lignocellulose-digesting termite. *Insect Biochem. Mol. Biol.* 43, 91–101. doi: 10.1016/j.ibmb.2012.10.001

Shade, A., and Handelsman, J. (2012). Beyond the Venn diagram: the hunt for a core microbiome. *Environ. Microbiol.* 14, 4–12. doi: 10.1111/j.1462-2920.2011.02585.x

Shade, A., Peter, H., Allison, S. D., Baho, D. L., Berga, M., Bürgmann, H., et al. (2012). Fundamentals of microbial community resistance and resilience. *Front. Microbiol.* 3:417. doi: 10.3389/fmicb.2012.00417

Shimada, K., Lo, N., Kitade, O., Wakui, A., and Maekawa, K. (2013). Cellulolytic protist numbers rise and fall dramatically in termite queens and kings during colony foundation. *Eukaryot. Cell* 12, 545–550. doi: 10.1128/EC.00286-12

Soergel, D. A. W., Dey, N., Knight, R., and Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J* 6, 1440–1444. doi: 10.1038/ismej.2011.208

Stingl, U., Radek, R., Yang, H., and Brune, A. (2005). "Endomicrobia": cytoplasmic symbionts of termite gut protozoa form a separate phylum of prokaryotes. *Appl. Environ. Microbiol.* 71, 1473–1479. doi: 10.1128/AEM.71.3.1473-1479.2005

Su, N., Ye, W., Ripa, R., Scheffrahn, R., and Giblin-Davis, R. (2006). Identification of Chilean *Reticulitermes* (Isoptera: rhinotermitidae) inferred from three mitochondrial gene DNA sequences and soldier morphology. *Ann. Entomol. Soc. Am.* 99, 352–363. doi: 10.1603/0013-8746(2006)099[0352:IOCRIR]2.0.CO;2

Tap, J., Mondot, S., Levenez, F., Pelletier, E., Caron, C., Furet, J.-P., et al. (2009). Towards the human intestinal microbiota phylogenetic core. *Environ. Microbiol.* 11, 2574–2584. doi: 10.1111/j.1462-2920.2009.01982.x

Tarpy, D. R., Mattila, H. R., and Newton, I. L. G. (2015). Development of the honey bee gut microbiome throughout the queen-rearing process. *Appl. Environ. Microbiol.* 81, 3182–3191. doi: 10.1128/AEM.00307-15

Tartar, A., Wheeler, M. M., Zhou, X., Coy, M. R., Boucias, D. G., and Scharf, M. E. (2009). Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite *Reticulitermes flavipes*. *Biotechnol. Biofuels* 2:25. doi: 10.1186/1754-6834-2-25

Turnbaugh, P. J., and Gordon, J. I. (2013). The core gut microbiome, energy balance and obesity. *J. Physiol.* 17, 4153–4158. doi: 10.1113/jphysiol.2009.174136

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Ligett, C., Knight, R., and Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* 449, 804–810. doi: 10.1038/nature06244

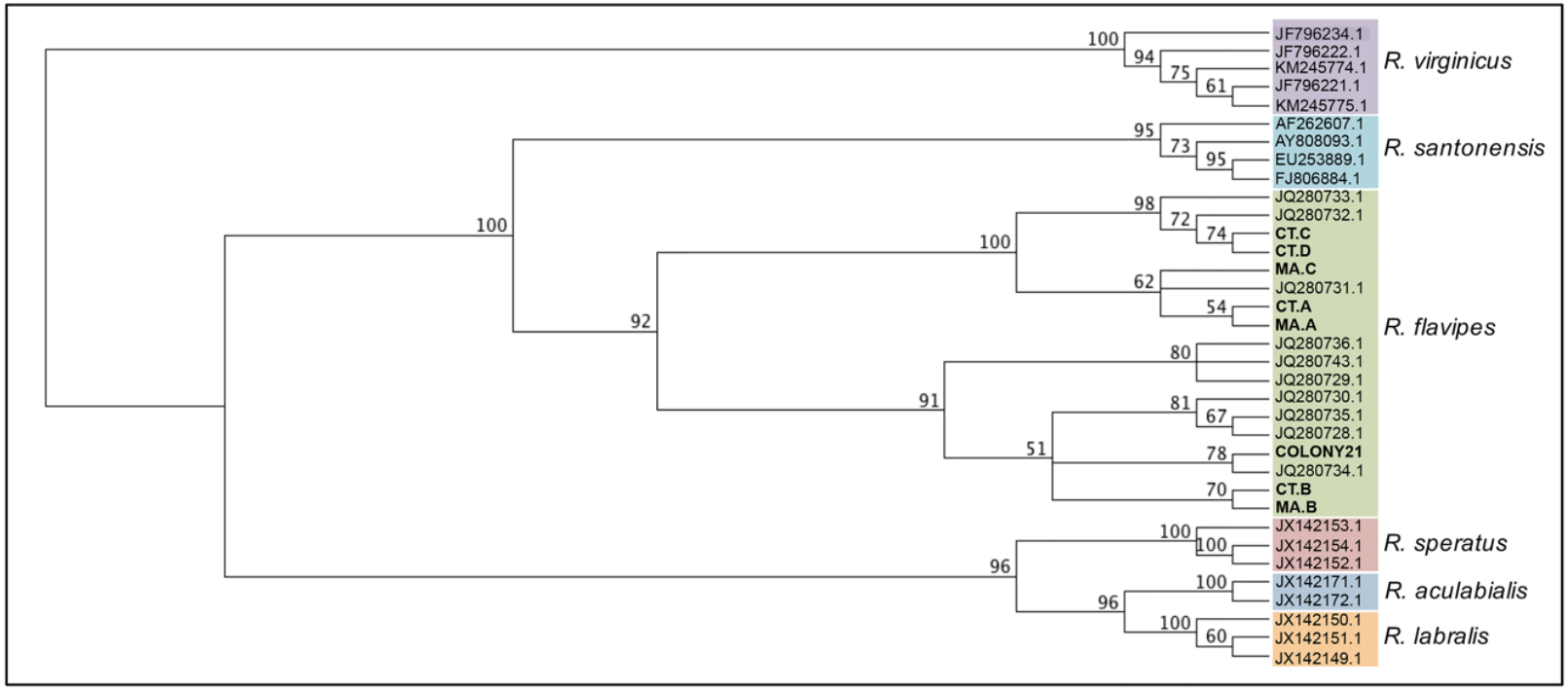
Wang, Y., Gilbreath, T. M., Kukutla, P., Yan, G., and Xu, J. (2011). Dynamic gut microbiome across life history of the malaria mosquito Anopheles gambiae in Kenya. *PLoS ONE* 6:e24767. doi: 10.1371/journal.pone.0024767

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

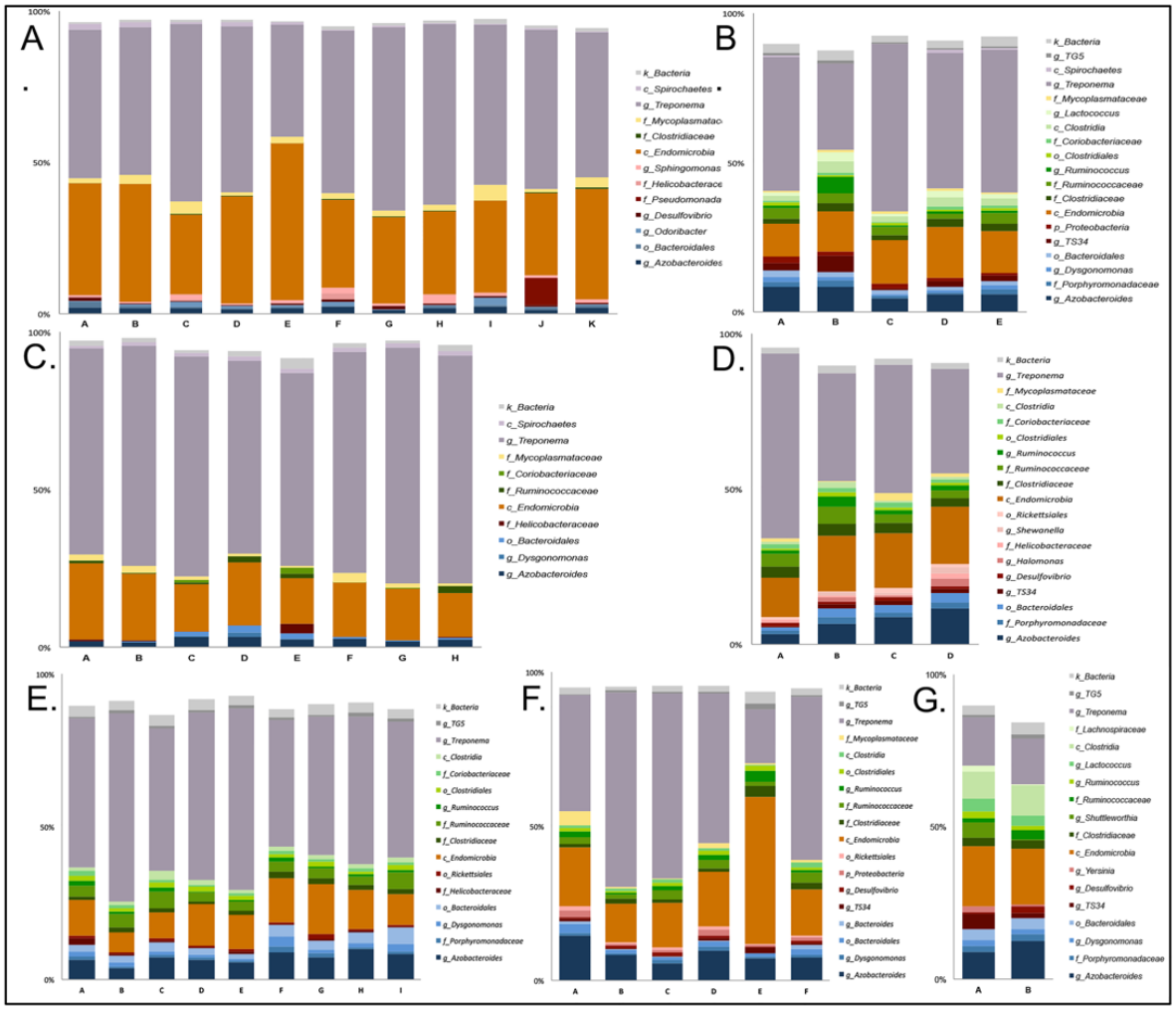
Yu, Z., and Morrison, M. (2004). Improved extraction of PCR-quality community DNA from digesta and fecal samples. *Biotechniques* 36, 808–812.

Zilber-Rosenberg, I., and Rosenberg, E. (2008). Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS Microbiol. Rev.* 32, 723–735. doi: 10.1111/j.1574-6976.2008.00123.x

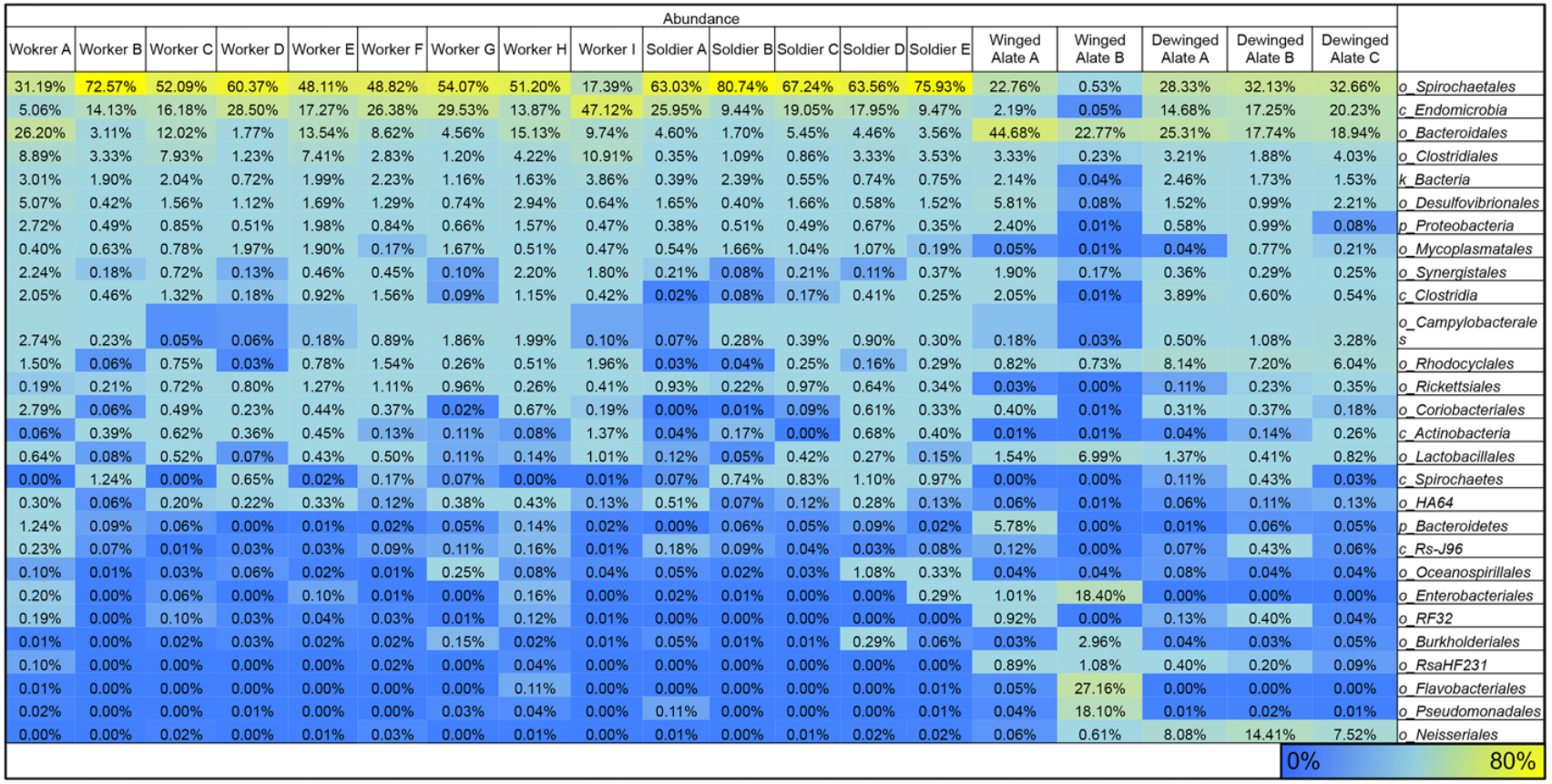
Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.



Supplementary Figure 1. Neighbor-joining (NJ) Analysis of Cytochrome Oxidase II (COII) Gene Sequences. COII DNA sequence data from termites used in this study and *Reticulitermes* species obtained from NCBI were used in NJ analysis with 100 bootstrap iterations. Sequences from R. virginicus were used to root the phylogram. Numbers above the nodes indicate NJ bootstrap support.



Supplementary Figure 2. Composition of the Hindgut Microbiota in Connecticut (CT) and Massachusetts (MA). Connecticut and Massachusetts colonies do not differ by state of collection, but vary slightly between colonies. Colonies, CT.A (A), CT.B (B), CT.C (C), and CT.D (D). Colonies CT.A and CT.C are lacking most of the *Firmicutes* phylum and a larger abundance of the *Spirochaetes*. These colonies were lab-maintained for over a year, while the other colonies were sampled immediately upon collection. MA.A (F), MA.B (F), and MA.C (G). Colony MA.C has a lower abundance of *Spirochaetes* than the other Massachusetts colonies and a greater taxonomic diversity.



Supplementary Figure 3. Taxonomic abundances of the hindgut microbiota in the worker, soldier, winged alate, and de-winged alate castes. Each termite used in the caste analysis (Figure 3) is shown.

A.

Treponema

OTU ID GG/DictDb[a]	Taxonomic Subgroup[b]	*Parabasalia* P- value[c]	*Parabasalia* R squared	*Oxymonadida* P- value[c]	*Oxymonadida* R squared
77905/UltTr169	*Treponema Ia*	0.0003	0.312	< 0.0001	0.4525
77925/UltTr173	*Treponema Ia*	0.0011	0.259	< 0.0001	0.4656
78075/UltSpi49	*Treponema Ia*	0.0277	0.1276	< 0.0001	0.3946
114589/UltB4827	*Treponema Ia*	0.0001	0.3367	< 0.0001	0.4822
114698/UltB4828	*Treponema Ia*	0.0063	0.1897	< 0.0001	0.4807
135860/UltTr348	*Treponema Ic*	< 0.0001	0.4258	< 0.0001	0.6102
4116/UltTr183-4	*Treponema Ig*	< 0.0001	0.4105	< 0.0001	0.6876
114560/UltB4833	*Treponema Ig*	0.0007	0.2786	< 0.0001	0.4183
143443/UltTr177	*Treponema Ig*	0.002	0.2362	< 0.0001	0.4175
77582		0.0006	0.2799	< 0.0001	0.4162
114660		0.0026	0.2246	< 0.0001	0.4323
141454		0.0046	0.2028	< 0.0001	0.4145
denovo136320		0.0002	0.3152	< 0.0001	0.3913
denovo15637		0.0089	0.1754	< 0.0001	0.4054
denovo193592		0.0218	0.1377	0.0001	0.336
denovo208664		0.0004	0.2939	< 0.0001	0.4391
denovo255921		0.0001	0.3387	< 0.0001	0.4327
denovo38396		0.0145	0.1548	0.0008	0.2727
denovo44029		0.0297	0.1247	0.0004	0.302
denovo53179		0.0003	0.3052	< 0.0001	0.5813
denovo67757		0.0694	0.08867	0.0004	0.2939
denovo84958		0.0071	0.1848	< 0.0001	0.3533
denovo94306		0.0124	0.1613	0.0001	0.3374

B.

Endomicrobia

OTU ID	*Parabasalia* P-value[d]	*Parabasalia* R squared	*Oxymonadida* P-value[d]	*Oxymonadida* R squared
114693	< 0.0001	0.427	< 0.0001	0.5298
114701	0.0024	0.2287	0.0005	0.2879
114552	< 0.0001	0.3531	0.0047	0.2011
108758	0.0162	0.1503	0.0085	0.1772
114047	0.0344	0.1183	0.0091	0.1742
15539	0.0203	0.1408	0.0065	0.1884
194402	0.0002	0.3315	< 0.0001	0.4248
30026	0.0007	0.2779	< 0.0001	0.365
36316	0.0016	0.2453	0.0002	0.3197
42156	0.0001	0.3405	0.0001	0.3402
87158	0.3475	0.02455	0.1896	0.0473
89039	0.0005	0.289	0.001	0.2613

[a]OTU ID from the Greengenes and DictDb databases
[b]Taxonomic classification based on the DictDb database
[c]A p-value <0.001 is significant of a correlation.
[d]A p-value <0.004 is significant of a correlation.
- Gray-shaded OTUs are not significantly correlated.

Supplementary Figure 4. Correlation of *Treponema* and *Endomicrobia* Bacterial OTUs to *Parabasalia* and *Oxymonadida* Protists. *Treponema* and *Endomicrobia* correlate with *Parabasalia* and *Oxymonadida* protists. *Treponema* and *Endomicrobia* OTUs from the core microbiota were compared to the protist qPCR data using a Pearson correlation. R squared values were calculated and the p-values were Bonferroni corrected. Areas shaded in gray represent OTUs that did not significantly correlate with a protist. The *Treponema* and *Endomicrobia* OTUs are considered significantly correlated if $p<0.001$ and $p<0.004$ respectively.

Chapter Three

Machine Learning and Microbiomes: Predicting Low-Abundant Bacteria as the Drivers of Gut Microbiomes after Alterations in Diet[+]

Contributions from other researchers

Stephen Lincoln contributed to this chapter. He trained and tested the artificial neural network (ANN) on the sequencing data I provided. He also ran a prediction through the ANN and developed the data that went into figures 1, 6, and 7.

[+] In preparation for submission, Benjamino, Lincoln, Srivastava, Graf.

Abstract

As the importance of beneficial bacteria becomes more apparent, understanding the dynamics of the symbiosis are a sought-after field. In many gut symbioses, it is essential to understand whether changes in the host diet plays a role in the bacterial community persisting in the gut. Although many advances have been made in genomic sequencing for analyzing the composition of microbiomes, very few studies have attempted to learn and model their dynamics. The challenge in learning and modeling microbiome dynamics is due to the complex, interdependent, and large number of highly non-linear interactions among members of a microbiome, as well as environmental factors. In this study, six diet sources were used to understand the effect of diet change on the termite hindgut microbiota. Termites were fed a mulch mixture, cardboard, spruce, oak, maple, and birch and hindguts were sampled at various time points over a 49-day period. DNA was extracted from each hindgut and 16S rRNA gene sequencing on the V4 hypervariable region was performed to determine the bacterial community at each time point. It was found that the sixteen core taxa (Benjamino & Graf, 2016)remained stable regardless of diet, and that many of the changes seen were in non-core taxa. The hindgut microbiota shifted on an OTU (operational taxonomic unit) level from the original day 0 samples when the termites fed on different wood sources. We also present a computationally tractable strategy using machine learning methods and stochastic optimization to characterize a microbiome. A deep backpropagation artificial neural network (ANN) is utilized to learn how the six different lignocellulose food sources affect the temporal composition of the hindgut microbiome of *Reticulitermes flavipes*, the eastern subterranean termite. These dynamics are then explored using a sensitivity analysis of the ANN to determine strength of taxon-taxon and taxon-

substrate interactions. The findings of the ANN are compared with 16S rRNA amplicon sequencing analysis.

Introduction

Symbioses are wide spread in nature, and beneficial digestive-tract symbioses have been shown to be critical for host health (Kostic et al., 2013). The benefits and contributions provided by the gut microbiota include enhancement of digestion efficiency, provision of nutrients and vitamins, and procurement of digestive enzymes (Graf, 2016). Members of the microbiome can contribute to host health by detoxifying allelochemicals from plants, such as tannins, flavonoids, and alkaloids, along with creating colony resistance against possible pathogens (Dillon & Dillon, 2004). When the host feeds on a nutrient-poor diet, the reliance on the physiological capabilities of the microbes is even greater for example the symbioses of aphids and tsetse flies. Aphids feed solely on plant phloem and require symbionts that synthesize amino acids that the aphid uses for nutrition (Douglas, 1998). The tsetse fly, which feeds exclusively on vertebrate blood has a symbiotic relationship with the bacterium, *Wigglesworthia glossinidia*. This bacterium contains genes for vitamin B production, which is an essential nutrient for the fly (Akman et al., 2002). Mammals that feed on a cellulose-rich diet require a bacterial community in the gut to generate energy for the host due to the inability of mammals to produce cellulases (Davison & Blaxter, 2005). Ruminant cows are an example of mammals that depend completely on their gut microbiota for nutrients due to their inability to digest grasses and hay (Jami & Mizrahi, 2012). In contrast, some insects can produce cellulases and sometimes harbor protist symbionts that are critical in the breakdown of the wood meal (Davison & Blaxter, 2005). These insects rely on the bacterial symbionts as a source of energy in the form of short-chained fatty acids (SCFAs) and

nutrients that are present in low amounts or absent in plant food sources such as nitrogen, amino acids, sterols, and many B vitamins. Examples of insects that produce cellulases are lower termites and Cryptocercus cockroaches (Varma et al., 1994). Another approach that higher termites such as *Macrotermes spp*. utilize is to maintain cellulolytic fungus gardens and consume the fungus and degraded wood for nutrients (Hyodo et al., 2003).

Insect feeding habits have been shown to partially dictate the microorganisms present in the gut. Cockroaches fed a low-protein and high-fiber diet show decreases of Streptococci and Lactobacilli in their gut. The decrease of these two organisms coincides with the reduction of acetate and lactate production (Kane & Breznak, 1991). The gut of the American cockroach, *Periplaneta americana*, is populated by a higher abundance of protozoa when feeding on a high-cellulose diet (Tinker & Ottesen, 2016). The house cricket, *Acheta domesticus*, shows reduced production of H_2, CO_2, and SCFAs when fed a high-protein diet compared to other diets (alfalfa, cricket chow, and sugarbeet root pulp) (Klug et al., 1998). A comprehensive study on higher termites showed that diet plays a role in shaping the microbiome composition of the gut. Bacteria with the ability to degrade cellulose were present in higher abundances in wood-feeding termites compared to diets without cellulose (Wang et al., 2016). Alternatively, termites that feed on humus and soil have a more alkaline gut content and bacteria that live in more alkaline environments were shown to be more abundant in these termites (Mikaelyan et al., 2015). *Reticulitermes flavipes*, the eastern subterranean termite, is a wood-feeding lower termite that harbors protist, bacterial, and archaeal symbionts. The protists are thought to aid in the breakdown of cellulose and lignocellulose, while the bacteria and archaea will utilize the breakdown products to produce nutrients for the symbiotic community and the host (Ohkuma, 2003; Scharf, 2015a). In a transcriptomic analysis, Sethi et al. showed that when fed wood, *R.*

flavipes termites had more protist-related transcript sequences in the gut compared to being fed paper (Sethi et al., 2013b). This suggests that protists are necessary for *R. flavipes* termites to digest wood. Elusimicrobia, a protist endosymbiont, was also more abundant in termites fed wood-diets compared to termites fed grassy diets. Grass diets also resulted in lower bacterial diversity in the hindgut (Huang et al., 2013b).

Although microbiomes are being studied extensively temporal studies are limited and predictive in silico modeling of microbiome dynamics is lacking. While the surge of metagenomics and gene sequencing technology have helped give insight to the composition of the microbiome, there have been only a few studies attempting to model the microbiome and learn dynamics between members of the community, as well as influences of environmental factors, such as diet of the host organism. Reasons behind the lack of microbiome models include inherent complexity of most communities leading to numerous interrelationships among community members, the computational difficulty of modeling highly nonlinear relationships, and other accounting for the effect of many external influences such as substrate, temperature, pH, micronutrient concentrations, etc. One of the few studies performed in this area involves using an artificial neural network with Bayesian network inference to predict relative abundance of microbial taxon in the English Channel as a function of its environment (Larsen et al., 2012). Although this method was successful at modeling how the environment shapes the microbiome, it did not answer the question of how to identify important taxa or environmental factors once the dynamics are learned. Along those lines, although previous studies found which organisms in the rhizosphere microbiome are important for disease protection in plants, there is no current method that would determine what taxa or environmental factors influence the growth or decline of these organisms based on a learned model (Mendes et al., 2011). Combining both the qualitative

knowledge of the bacterial members of a microbiome with quantitative in silico modeling of microbiomes is key for identifying influential organisms in the microbiome, as well as learning how members of a microbiome work symbiotically or antagonistically.

The hindgut microbiome of the lower termite, *R. flavipes*, is suitable for testing predictive modules because have a detailed understanding of the community member, we can readily identify them using 16S rRNA gene tag sequencing and this species is capable of feeding on different types of wood and can be readily maintained in the lab (Benjamino & Graf, 2016). An important aspect of the hindgut microbiome is that a large proportion of the taxa is consistently present in all the individual, these taxa are considered to be part of the "core" community. In this work, we monitored the composition of the hindgut microbiome of *R. flavipes* by sequencing the V4 region of the 16S rRNA over time following dietary changes. An algorithm inspired by Larsen et al. (Larsen et al., 2012) uses an artificial neural network to learn the dynamics of the microbiome not only from external influences, such as changes in the diet, but also changes in the relative abundance of the taxa in the microbiome from the 16S rRNA sequencing data. The artificial neural network was then trained on this data and a sensitivity analysis was performed to determine the accuracy of the model. When used in conjunction with microbial community analyses, the artificial neural network learned dynamics allows for an in-depth analysis of a microbiome to understand taxon-taxon and taxon-substrate interactions.

Methods

Experimental Design and Maintenance

The *R. flavipes* termites were purchased from Connecticut Valley Biological Supply Co. in Southampton MA and initially maintained on the mulch they were shipped with. These mulch-

fed termites were separated into colonies that received distinct diets. The colonies were kept in plastic containers with autoclaved sand and food. Termite colonies were maintained in a dark cabinet at room temperature and kept moist with water. The samples used in this study were all from the worker caste.

Colonies were fed either wood from spruce, oak, maple, or birch, cardboard or mulch (never changed from original food source), and one colony was starved. Termites were sampled on the day of arrival (Day 0) and on days 1, 2, 3, 7, 14, 21, 28, 35, 42, 49 after arrival. The starved colony was sampled through day 21 and the oak-fed colony was sampled through day 28, both due to the lack of termites available in the colony. The samples from Day 0 were previously published from Benjamino and Graf (Benjamino & Graf, 2016). The wood used was non-treated firewood, the cardboard was from shipping boxes, and the original mulch used was the material shipped with the termites from CT Valley Biological Supply Co. Termite DNA was used for COII sequencing to ensure the termites were *Reticulitermes flavipes*. Primers used for COII sequencing were a modified A-tLEU (5'-CAGATAAGTGCATTGGATTT-3') and B-tLYS (5'-GTTTAAGAGACCAGTACTTG-3') from Liu & Beckenbach (Liu & Beckenbach, 1992) and previously reported in Benjamino and Graf (Benjamino & Graf, 2016).

Sample Collection and DNA Isolation

Hindguts were removed from the termite and separated from the foregut/midgut and rectum. Single hindguts were collected in 1X TE buffer (10mM Tris-HCl, 1mM EDTA, pH 8.0). DNA was isolated immediately after collection using a modified (500µL starting lysis buffer, elution in 30µL AE buffer) RBB+C isolation protocol as described by Yu and Morrison (Benjamino & Graf, 2016; Yu & Morrison, 2004).

PCR Amplification and Library Preparation

Hindgut samples were amplified using the V4 hyper-variable region of the 16S rRNA gene using primers developed by Carporaso et al. (Caporaso et al., 2011). PCR reactions included Phusion High-Fidelity PCR Master Mix with HF Buffer (50% of total volume), 10μM forward and reverse primers (3% each of total reaction volume), ~10ng DNA, and dH2O to the final volume. All reactions were amplified in triplicate using the following parameters: 94°C for 3 minutes, followed by 30 cycles of 94°C (45 seconds), 50°C (60 sec), and 72°C (90 sec), with a final extension of 72°C for 10 minutes (Nelson et al., 2014).

Amplicons were purified and size selected using the GeneRead™ Size Selection Kit by Qiagen© to select for 400 bp amplicons according to manufacturer's protocol. Samples were then quantified using a Qubit® dsDNA HS Assay and diluted to 4 nM. All samples were pooled in equimolar amounts for sequencing.

Sequencing and Data Processing

Samples were sequenced using an Illumina MiSeq with custom sequencing primers added to the reagent cartridge (Caporaso et al., 2011) and sequenced 2x250bp. Output reads from the MiSeq were merged to create single reads spanning the entire 254 bp length of the V4 hypervariable region using SeqPrep (https://github.com/jstjohn/SeqPrep), and the PhiX control reads were removed by mapping to the PhiX genome (Nelson et al., 2014). Data analysis was performed on high quality reads (Q30 or greater) using Qiime (Caporaso et al., 2011). Operational taxonomic units (OTUs) were determined by clustering reads to the V4 hypervariable region of the DictDb 16S rRNA reference dataset at a 97% identity level (Benjamino & Graf, 2016; Mikaelyan et al., 2015). Reads that failed to cluster to the DictDb reference were clustered to the Greengenes reference 16S reference dataset (2013-08 release) at a

97% identity, and then de novo OTU clustering was performed on reads that failed to cluster to a reference (DeSantis et al., 2006). The dataset was checked for chimeras and filtered to remove singleton and doubleton OTUs and then OTUs present at less than 0.0005% (McDonald et al., 2011; Nelson et al., 2014).

Sequence Analysis

After quality filtering and rarifying to 18,000 sequences per sample, alpha diversity (Shannon Index and Equitability) (Faith & Baker, 2006) and the Bray Curtis beta-diversity metric (Anderson et al., 2011) were performed using Qiime 1.9. The Shannon Index and Equitability were graphed using GraphPad Prism version 6.0f for Mac OSX (GraphPad Software, San Diego, California USA, www.graphpad.com), and a one-way ANOVA with Bonferroni post-test analysis was performed for each. An NMDS plot using the Bray-Curtis metric was created in R 3.2.0 (Oksanen et al., 2009; Wickham, 2009). The PERMANOVA statistical analysis was performed to determine the significance of microbial community differences among the different food sources and temporally. The test used the Bray-Curtis dissimilarity matrix as the input and was performed over 999 permutations and returned a Pseudo-F (f) statistic along with a p-value (p). Each test compared the Day 0 samples to the last two days of samples in other diets.

Taxonomic abundance data was calculated using the percent abundance of OTUs present in the core microbiota. The relative abundance of each taxon, along with the non-core taxa, was combined for each diet and presented with the mean abundance of the temporal data. The non-core abundances were calculated by combining the remaining OTUs that were not present in the core.

Artificial Neural Network

The relative abundance of each OTU was grouped by taxonomic order as grouping by species for learning microbiome dynamics introduced a significant amount of noise and error to the algorithm. A deep backpropagation artificial neural network (ANN) was created using Fast Artificial Neural Network (FANN) (Nissen, 2003) with a network topology as seen in Figure 1. The goal of training the ANN was to learn dynamics of the microbiome based on substrate given to the colony and the influence of other community members. The network was trained by using the relative abundance of each order and the presence or lack of substrate given at a time period t for the input nodes. The output represented the relative abundances of each order for the time period t+1. A training data file for the ANN was generated from the 7 colony CSV files. The general algorithm is shown in Supplemental Figure 1. Since there was no target for the last time point (day 49), the last time point was never used as an input to the ANN. In addition, one random time point from each CSV (7 total) was left out of training and used for testing the ANN.

After the training data file was generated from the CSV files, the network used the standard backpropagation algorithm native to FANN to train on the dataset until the error threshold was met or the maximum number of epochs had passed. The activation function used in the first hidden layer was the hyperbolic tangent function (tanh), while the second hidden layer utilized the sigmoid function,

$$S(x)=11+e-x \quad \textbf{(1)}$$

The activation functions used were chosen based on the features of the system studied. The first hidden layer has more nodes than the second hidden layer, and therefore has more connections. Using an antisymmetric activation function has been shown to improve convergence for more connected networks than an asymmetric activation function, so tanh was

chosen as the first hidden layer's activation function (LeCun et al., 1991). However, the output of the ANN is the relative abundance of a taxon for each output node, which must be in the bounds [0,1], therefore, the sigmoid function was used for the second hidden layer's activation function.

After each epoch of training, the mean squared error was returned. If the error was under the error threshold, training was halted and the network was saved. The network was then tested using the seven time points left out of the training dataset. After testing, the ANN was subject to sensitivity analysis using a test time point from day 0. The parameters of the neural network are shown in Table 1.

Microbiome Dynamics Analysis

Once the network was sufficiently trained, a sensitivity analysis was performed on the ANN in order to determine how each taxon changes over time in response to a change in another taxon or substrate. This analysis was performed by using a test time point of relative abundance and substrate that was left out of the ANN training set. Since each input node is representative of a taxon or substrate and each output node is representative of the predicted relative abundance for a given taxon at the next time point, changing each input node individually and comparing the predicted outputs to the original outputs will allow us to discover the learned dynamics. Therefore, each input node was changed one at a time to a value in the range of 100 values ±5% of the original value of the input node for the test time point given. The ANN was then run for each new value of the input node while holding the other input values constant. After each run of the ANN at the new input node value, the new outputs, or predicted relative abundances of each taxon were recorded and compared to the original output. The percent change of each output

node was compared to the percent change of the input node for all values in the 100 values ±5% of the original input value using the following equation:

$$Percent\ Change = (New\ Output - Original\ Output) \times (New\ Input - Original\ Input) \quad \textbf{(2)}$$

This was repeated for each new value of the input node being tested. After all new values of the input node were tested, the average change of each output node was taken, which shows how each taxon (output node) changes with respect to a change in a certain taxon or substrate (input node). This was repeated for every input node in order to determine how the relative abundance of each taxon changes with respect to change in a certain taxon or substrate. The general algorithm is shown in Supplemental Figure 2. The end result of the sensitivity analysis was a matrix of relative change values. In other words, each row of the matrix is representative of a taxon or substrate, or an input node of the ANN, while each column is representative of a taxon, or an output node of the ANN. Each value in the matrix is the relative average percent change of the output node (column) with respect to the input node (row).

In order to visually display the results of the sensitivity analysis, a 2D influence heatmap was generated using Matplotlib in Python (Hunter, 2007). The heatmap displays the magnitude and direction (direct/inverse) of each relationship between each input node (taxon/substrate) and output node (taxon). A connectivity network was also generated using the graph-tool library in Python (Peixoto, 2014). The connectivity network was constructed using a vertice-edge plot where each vertex is a node in the ANN, or a taxon/substrate. An edge was drawn between two vertices if the value in the change matrix was more than three standard deviations above the absolute value of the average of the whole array. The top ten most connected vertices were highlighted and returned from the connectivity network.

Data Availability

The sequence data was deposited in the European Nucleotide Archive (ENA) SRA under project ID PRJEB20463.

Results

Effect of Dietary Changes on the Termite Hindgut Microbiota

The hindgut microbiota of termites supplies the host with energy and nutrients by fermenting the ingested lignocellulose. While it has been shown that this community structure changes when termites are fed different diets (Boucias et al., 2013; Huang et al., 2013b), it is not known how fast these changes occur. By dividing members from a single colony into different groups that were provided with different food sources and sampling multiple individuals over a seven-week period we were able to determine how a change in the food source affects the termite hindgut microbiota. Termites were transferred from the original mulch diet to either wood from spruce, oak, maple, or birch, cardboard or mulch, and one colony was starved. Individual animals were removed at different time points and the composition of the hindgut microbiome determined.

The overall composition of the hindgut community was compared by determining the Bray-Curtis beta diversity and depicted on an NMDS plot (Figure 2). All of the Day 0 samples were similar to each other (black points and ellipse). The microbiota of termites that were maintained on mulch, did not not exhibit a significant change in the overall composition of the community. The Mulch colony (magenta points and ellipse) was the most similar to the Day 0 samples, which is indicative of the temporal effect on the hindgut microbiota. There was no significant difference between the Day 0 samples and the last two days of the Mulch-fed colony according to a PERMANOVA with a Bonferroni correction for multiple comparisons (f=2.03,

p=0.029) (Figure 2). The microbiota of termites maintained on all of the other diets gradually moved away from the Day 0 samples over time. For the first seven days the communities were similar to Day 0, but samples from later dates differed significantly suggesting that dietary changes affect the composition the hindgut microbiota (PERMANOVA, f=4.18, p=0.001). This finding suggests that the hindgut microbial community shifts after about seven days of feeding on a new diet.

Observed differences in a microbial community can be due to instances in which OTUs are present or absent or by change in the abundance of sequences from any given OTU. The microbes present in the hindgut can divided into members of the core, which are consistently present and non-core taxa, that are present intermittently (Benjamino & Graf, 2016). For this analysis, the sequence counts for each OTU belonging into the same taxon were combined and the percent abundance in the entire community was calculated. The taxonomic abundance values for all time-points in each diet were plotted along with the mean. The percent abundance of each of the fifteen *R. flavipes* core taxa, along with the abundance of non-core taxa are shown in Figure 3. *Treponema* and *Endomicrobia* sequences accounted for >10% of the sequences for all diets excluding samples from starved termites. Members of these two genera are known to be associated with the hindgut protists, which decrease in number when the termite is starved, likely leading to concomitant decrease in their endo- and ectosymbionts (Ohkuma, 2008). The order, Bacteroidales, accounted for 1-10% of the sequences in all samples, with the lowest abundance detected in samples from birch-fed termites. The remaining taxa were present at abundances less than 3%, most accounting for <1%. Overall, sequences from the core taxa stayed at a constant level for each diet while sequences from the non-core taxa were more variable among diets.

Another characteristic of microbial communities is the number of species present or the richness that can be measured by the Shannon Index (H') and the evenness of the community that is measured by the Equitability (E_H) metrics (Figure 4). The average H' and EH values were 6.58 and 0.723 respectively. Termites fed birch and spruce showed the most variability within the colony, but there was no significant difference in the richness or evenness among the colonies when compared to Day 0, based on a one-way ANOVA analysis.

Learned Microbiome Dynamics

The robustness of the ANN was assessed by excluding one time-point from each of the seven diets from the training set. Each test time point was entered and the predicted relative abundances were recorded as an array. The array of predicted relative abundances were compared to the actual relative abundances by taking the root mean squared error and Bray-Curtis similarity. Specifically, the values of the time point preceding the time point excluded from the analysis were entered and used to predicted relative abundances for the excluded time points. The calculated and measured relative abundances were compared and the average RMSE and Bray-Curtis similarity were calculated across the seven time points tested. The RMSE was found to be 0.0153 while the Bray-Curtis Similarity was found to be 0.8576, showing that the network was sufficiently trained and was robust enough to predict the composition of the microbiome over time. The actual taxonomic abundance of the tested samples at the given time-points was compared to the values predicted by the ANN (Figure 5). A heatmap was created to show the difference in relative abundance (%) between the measured values and the predicted values. The average abundance of each taxon is also shown (excluding the starved time-point) along with the number of significant correlations for each taxon. Five out of the fifteen core taxa had greater than five significant correlations, while thirteen non-core taxa had greater than five

significant correlations. The majority of taxa with significant correlations were present at less than 1% average abundance.

Each taxon or substrate was given a specific number corresponding to the input node they were assigned to. This numbering scheme was kept constant throughout the ANN training and analysis. The influence of specific taxa on all other taxa was calculated using equation 2 in the methods section and used to generate a correlation network. A 2D heatmatp was created with the resulting data, showing the degree of direct and inverse correlations between taxa (Figure 6). The top ten most correlated taxa were also returned from the algorithm (Figure 7).

Discussion

Our study revealed that changing the food source for *R. flavipes* affected the overall composition of the hindgut microbiome without affecting the members of the core. The NMDS analysis revealed that after seven days on a new food source the overall community differed significantly from the community at Day 0. This change in the community structure was driven by the less abundant microorganisms that were not part of the core microbiota changed in abundance with changes in diet. In comparison, for the control group that was continuously maintained on mulch the samples collected after seven days, did not show a significant shift in the bacterial community when compared to Day 0 ($f=2.03$, $p=0.029$). An earlier study by Broderick et al. showed that the gypsy moth caterpillar gut microbiota is affected by diet. Moths fed Larch (a conifer) had a higher bacterial diversity than moths fed other diets, and moths fed Aspen retained the most unculturable bacteria. Although these dietary changes affected the gut community, there were also two bacterial taxa that were present in the larvae regardless of diet, *Enterococcus faecalis* and an *Enterobacter sp.* (NAB17) (Broderick et al., 2004). Another study

fed *R. flavipes* either a wood-substrate or paper-substrate, sampled the hindgut microbiota after seven days, and reported the similarities and differences between the hindgut bacterial communities (Boucias et al., 2013). They found that termites from the same colony and fed different diets were more similar to one another than termites from different colonies and fed the same diet. However, that study did not report data for later time points, which was when we observed the changes in the hindgut community. Our results are similar to those reported from wood-eating cockroaches and higher termites. In wood-eating cockroaches that are closely related to termites, the members of the core microbiome were shown to be stable during dietary changes. In higher termites, which do not harbor protist symbionts, it was shown that the dominant members of the hindgut microbiota remained stable and only less abundant members were affected. It is interesting that these three studies consistently determined that the core was not affected by dietary changes. This highlights the critical contributions the individual taxa provide to the overall function of the hindgut microbiota and animal host. The most abundant organisms in a symbiotic habitat have been shown to perform important functions within the environment, essentially securing their constant presence in the environment (Shade et al., 2012). A study of the gut microbiota of thirty-seven adults over five years showed that around 60-70% of the strains remained stable throughout the study and that the most stable organisms were also the most abundant, which supports our findings (Faith et al., 2013). The stability of the composition of the core microbiota is consistent with the animal and its microbe forming one functional unit on which selection acts, as has been proposed in the holobiont theory (Bordenstein & Theis, 2015). It was only members that did not belong to the core that changed in their abundance and contributed to shift in community structure when the food source was changed.

The one major exception to the stability of the core microbiota occurred in the group that was starved. Starved termites have been shown to lose the symbiotic protists in the hindgut, which are associated with *Treponema* and *Endomicrobia* that are extracellular and intracellular symbionts, respectively. Therefore, it was expected to see a decrease in the relative abundance of *Treponema* and *Endomicrobia*. While the *Treponema* exist not only as protist symbionts, but as free-living bacteria as well, the *Endomicrobia* are strict endosymbionts of the protists, which explains the larger drop in *Endomicrobia* in starved termites compared to the *Treponema* (Cleveland, 1925; Graber et al., 2004). It was interesting to note that the *Endomicrobia* are most abundant in the Spruce-fed termites, which was also shown by Huang et al. (pine) and may suggest that this substrate may create a hindgut environment that enriches for the protist and *Endomicrobia* populations (Huang et al., 2013b).

Another aspect that can be affected by changes in diet is the complexity of the overall community. The community composition can be described by the richness or number of taxa that are present, H', and the evenness, that describes how similar the proportions of all the taxa are, E_H. In this study, termites from a single colony were separated into experimental groups maintained on distinct diets, with one colony remaining on the original mulch diet. The richness and evenness of the bacterial community was not statistically different for any of the diets (Figure 2). However, there was a greater range of H' and E_H values in the Birch and Spruce fed termites. This distribution could be due to the richness and evenness (H' and E_H values) decreasing with time. A study by Huang et al. showed a lower microbial richness in *R. flavipes* termites fed corn-based diets compared to wood-based diets (Huang et al., 2013b).

We were interested in using an artificial neural network to generate an *in silico* model that would allow us to determine the effect a particular taxon or the diet has on the other taxa,

determine the connectivity of a particular taxon to other taxa and predict from what the composition of the community will be a particular time point. A temporal comparison of the bacterial OTUs from each diet showed that dietary changes influence the composition of the hindgut microbiota in *R. flavipes* (Figure 3). This was also predicted by the *in silico* model, as relationships between substrate and taxa were significantly weaker than taxon-taxon relationships (Figure 6). None of the substrates that were fed to the colonies were deemed significant, as seen in Figure 7. It was also reported that, using a divergence level of 0.05 for OTU identification, only 0.4% of the OTUs were variable between diets. Our analysis supports the finding that the microbiota remains stable throughout the change in diet, but the OTUs abundance shifts temporally. When grouped by order, the *in silico* model showed that the Spirochaetales, the order which the genus *Treponema* belongs, was the most connected order in the model with 19 total important correlations (Figure 7). The correlations can be divided into outbound, where the taxon affects other taxa and inbound, where it is affected by another taxon. All of the 19 correlations associated with Spirochaetales were found to be outbound, implying that Spirochaetales has the largest effect on the community as a whole. Along with Spirochaetales, the orders of Rhodocyclales, Bacteroidales, and two orders of Clostridiales (different classes) were among the most abundant orders found, and had 19, 14, 13, and 13 correlations, respectively. Rhodocyclales had 15 outbound correlations and 4 inbound correlations. Of the 14 correlations associated with Bacteroidales, 12 were outbound and 2 correlations were inbound. In the termite hindgut community, the most abundant organisms have the most outbound correlations, suggesting that they played important roles in shaping the overall structure of the community. The lack of *Endomicrobia* connections in the *in silico* model is of particular interest, as they are an abundant taxon and one would assume them to have a role

in determining the makeup of the microbiome. However, the *Endomicrobia* are strict endosymbionts of the protists, and if the protists along with their endosymbionts are required at a constant level for the degradation of the lignocellulose irrespective of its source, it could be that their abundance is decouple from auxilliary changes in the community structure. The order Methanobacteriales was the 8th most connected order with 13 total connections. However, all of the connections were inbound connections. Methanobacteriales also had the most inbound connections of all the orders present, suggesting that it is the most influenced order in the network. This is particularly interesting as this is a Euryarchaeota that reduces CO_2 with H_2 to methane (Banning et al., 2005). The H_2 in turn is released as a metabolic waste product by the protists and bacteria.

One particular powerful aspect of the neural network analysis is that after having trained the algorithm it can be used to predict future values. We evaluated the accuracy of the predictions by leaving out specific time points from the training set, using the relative abundance data from the time point immediately preceding the time point of interest. The community composition predicted by the neural network was compared to the actual values obtained (Figure 5) and showed that the ANN was able to predict the relative abundance of each taxon within less than 15% difference. The ANN predicted the majority of taxa within less than a 1% difference, and the taxa with higher discrepancies were highly abundant (>10% abundance) which would allow for a larger difference due to background noise. This suggests that even for a community as complex as that found in the termite hindgut and with a relatively sparse sampling frequency, we were able to very accurately predict the community composition. One limitation of this analysis is that due to complexity of the interactions, we only performed it on the order level. It would be interesting to perform this analysis on other time series, including human, and evaluate

the accuracy. If future implementations are able to use lower taxonomic levels, it could prove to be an important predictor for an unbalanced microbial community or dysbiotic state before it occurs. Understanding the effect of diet on a microbial population is valuable because it provides insight into the dynamics of the symbiotic niche. In a laboratory setting, it is necessary to consider the biological effect diet has on a host organism and its symbiotic bacteria. The ability to predict the taxonomic composition of a community is beneficial for forming hypotheses about the environment and can provide insight on the community dynamics within that environment.

Since the start of the high-throughput 16S sequencing revolution, scientists have reported the microbiome, focusing on the more abundant taxa or grouping the populations into phylotypes (Walker et al., 2011). In order to view the bacterial populations in numerous samples, it has been standard protocol to show organisms at abundances greater than 1% (Shang et al., 2017), or even group the low-abundant organisms into an "other" category (Tinker & Ottesen, 2016). While this is a widely-accepted method of determining correlations between healthy and diseased states and reporting microbial communities in general, researchers may be missing some key organisms that get lost in the low abundant population. In the hindgut community, only five out of the fifteen core taxa had greater than five significant correlations, while thirteen non-core taxa had greater than five significant correlations (Figure 5). Among the core and non-core taxa with significant correlations, only four were abundant at greater than 1% of the community. This suggests that although the core and highly abundant taxa perform important functions for the community and are conserved throughout diet changes and time, perhaps the non-core and less abundant taxa are the drivers of the community, and changes in these organisms would have higher impacts on community structure.

Table 1. ANN training parameters

Input Nodes	**70**
Hidden 1 Nodes	67
Activation 1	Hyperbolic Tangent
Hidden 2 Nodes	60
Activation 2	Sigmoid
Output Nodes	64
Error Threshold	10-5
Momentum	0.95
Learning Rate	0.20
Max Epochs	20000

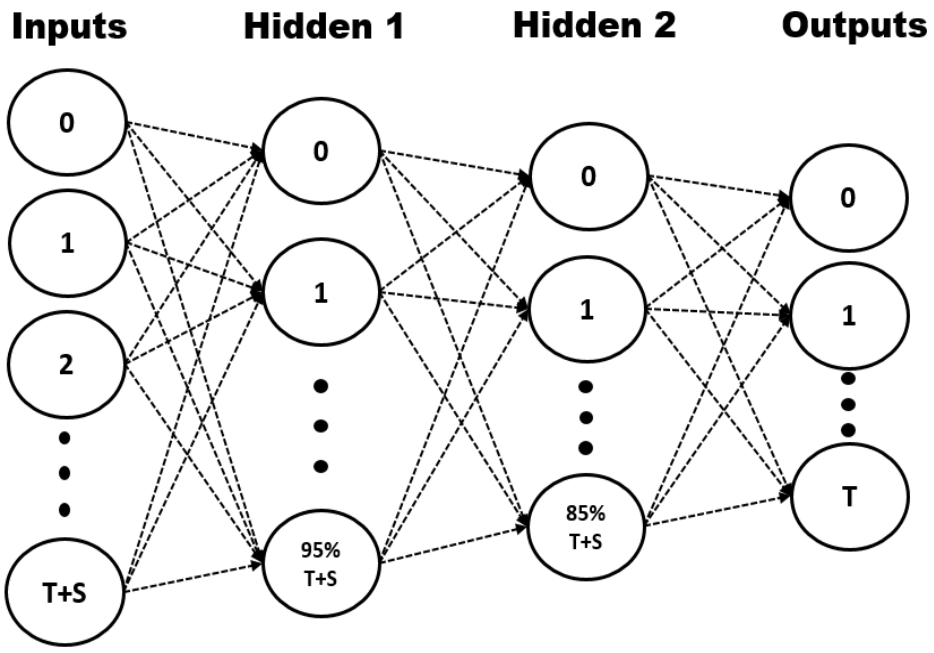


Figure 1. Topology of the ANN used to train on sequenced data. The number of input nodes used was set to the number of taxonomical orders (T) plus the number of environmental variables (E), which was the number of types of substrate fed to the colonies. Since there were 64 taxonomical orders and 6 substrates, there were 70 total input nodes for the ANN. The number of nodes in the first hidden layer was set to 95% of the total input nodes, whereas the number of nodes in the second hidden layer were set to the 85% of the number of input nodes. The number of output nodes was set to the number of taxonomic orders, as the goal of the network was to predict relative abundance changes over time for each taxon. The arrangement of taxonomic orders remained constant for each CSV file; therefore, each input node (up to the amount of taxa) and output node was representative of a specific taxonomic order. The remaining input nodes past the number of taxa represented an environmental variable.

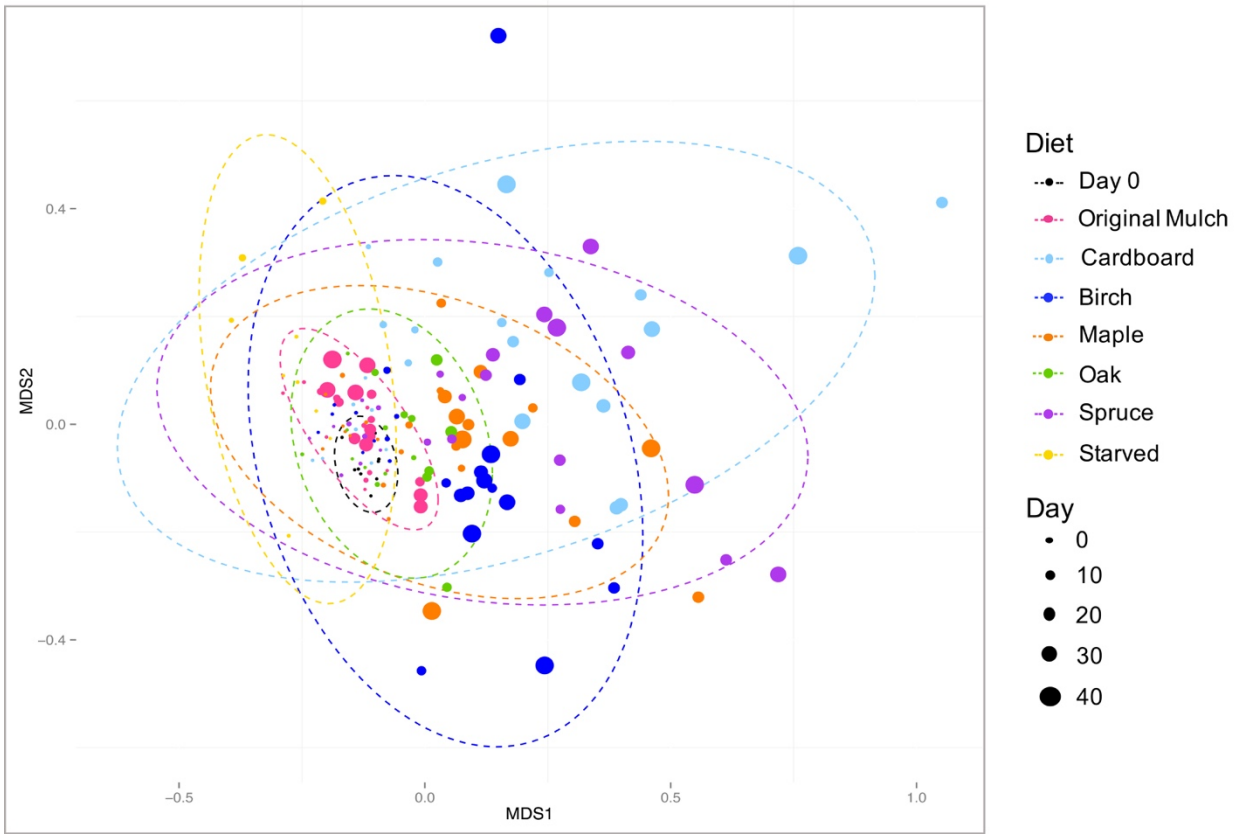


Figure 2. Diet change causes shifts in the hindgut microbiota. An NMDS plot was created using the Bray-Curtis dissimilarity metric based on OTU abundance. Each point represents the hindgut community at a specific day (denoted by size) and colored by diet. Ellipses circumscribe the 95% confidence interval for each diet. Day 0 samples are similar to each other as well as other diet samples early on (days 1-7). Samples from the original mulch colony remain most similar to the day 0 samples, indicating normal temporal/acclimation change of the termites and their microbiota. Other diets show shifts away from the day 0 samples (PERMANOVA, f=4.18, p=0.001).

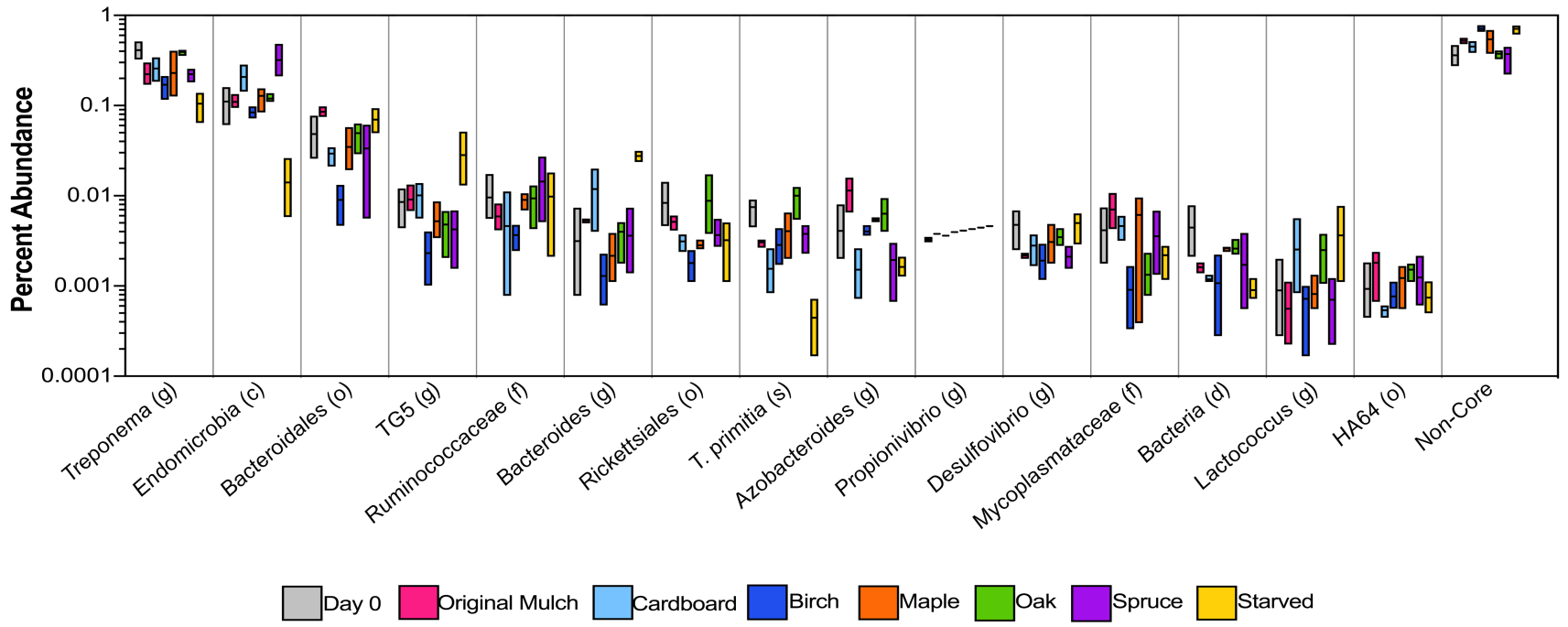


Figure 3. Effect of diet on the core and non-core taxa in the hindgut. The relative abundance of sequences from OTUs present in the core and non-core taxa were calculated for each diet. No diet had a significant effect on sequences from the core taxa as compared to the initial day 0 abundances. However, sequences for the non-core bacterial communities in the original mulch, cardboard, oak, and starved diets differed from the day 0 samples (one-way ANOVA ($F(1,7)=10.80$, $p<0.0001$).

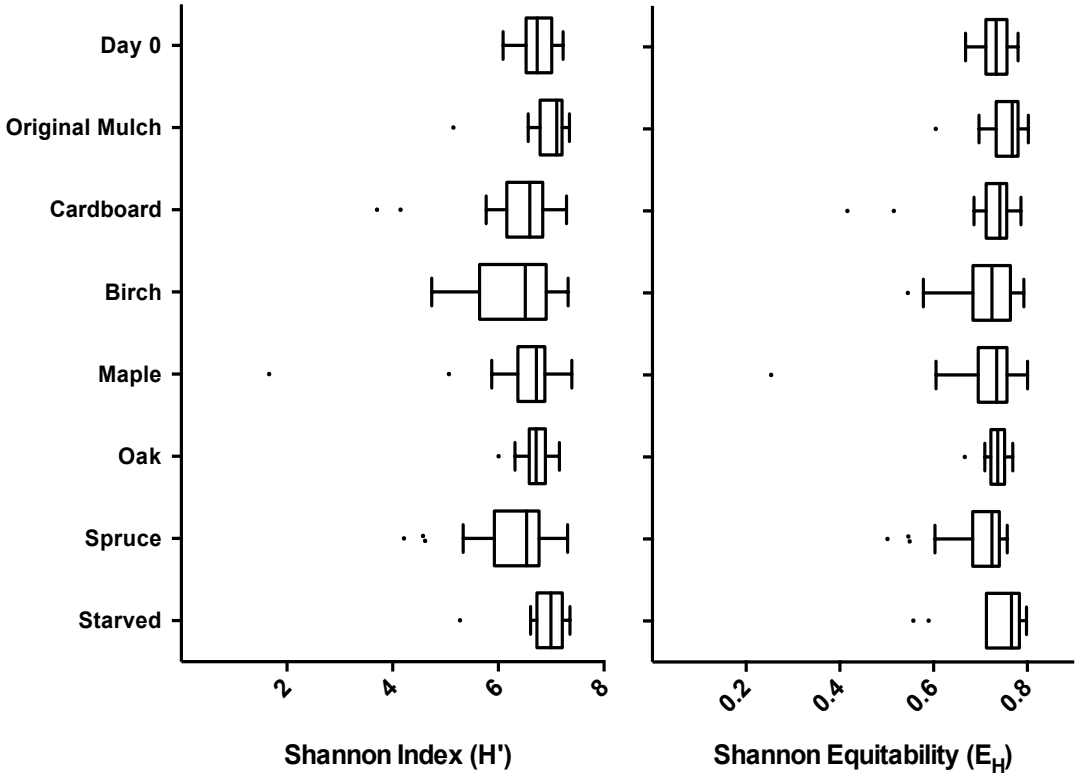


Figure 4. Alpha diversity of the *R. flavipes* hindgut fed multiple diets. The Shannon Index (H') and Shannon Equitability (EH) metrics were used to calculate the richness and evenness of the microbiota of the termite hindgut over 49 days when introduced to different diets and plotted using a box and whisker plot. A one-way ANOVA was used to compare the H' and EH values from each diet to day 0 and showed no significant difference ($p>0.05$). Day0 (n=9), original mulch (n=27), cardboard (n=26), birch (n=24), maple (n=28), oak (n=17), spruce (n=24), starved (n=10).


Corr.
Original Mulch
Cardboard
Birch
Maple
Oak
Spruce
Starved
Average (%)
Methanomicrobiales 4 0.39
Thermoplasmata-E2 10 0.01
Actinomycetales 5 0.89
Bifidobacteriales 0 0.00
Coriobacteriales 8 0.65
Bacteroidales 15 19.73
Flavobacteriales 0 0.04
Saprospirales 0 0.00
Sphingobacteriales 0 0.01
Deferribacterales 0 0.01
Elusimicrobiales 3 0.18
Endomicrobiales 0 14.35
Fibrobacteres-TG3 0 0.00
Bacilli 0 0.01
Lactobacillales 7 1.21
Turicibacterales 0 0.00
Clostridiales 11 14.03
Erysipelotrichales 0 0.02
Fusobacteriales 6 0.28
Plancomycetes 0 0.05
Planctomycetes-PeHg47 0 0.00
Pirellulales 6 0.21
Alphaproteobacteria 0 0.04
Caulobacterales 0 0.01
Alphaproteobacteria-RF32 0 0.03
Rhizobiales 0 0.00
Rhodospirillales 8 0.09
Rickettsiales 0 0.73
Sphingomonadales 0 0.00
Burkholderiales 1 0.15
Methylophilales 0 0.00
Neisseriales 0 0.01
Procabacteriales 0 0.00
Rhodocyclales 7 0.63
Desulfarculales 0 0.05
Desulfovibrionales 3 1.85
Aeromonadales 0 0.49
Alteromonadales 11 0.15
Enterobacteriales 13 0.07
Oceanospirillales 4 0.24
Pseudomonadales 0 0.08
Xanthomonadales 0 0.01
Proteobacteria-CV90 6 0.03
SR1 8 0.25
Leptospirales 6 0.12
Spirochaetales 17 39.79
Synergistales 1 0.74
Mycoplasmatales 0 0.56
Methylacidiphilales 11 0.02
Opitutae 0 0.00
Opitutae-HA64 2 0.15
Opitutales 0 0.01
Chthoniobacterales 0 0.05
Verrucomicrobiales 0 0.02
Rs-J96 0 0.09
Unassigned 1 1.48
0 5 10 15
Correlations/Difference (%)
0 10 20 30
Average Abundance (%)

Figure 5. Accuracy of the ANN to predict taxonomic abundances. In training the ANN, one sample per random time-point for each diet was left out and used to test the accuracy of the ANN. The measured abundances of taxa (Order) in the samples were compared to the abundances predicted by the ANN. The taxa represented in the core microbiota are highlighted in yellow and the average abundances are plotted in the right column. The difference between the actual values and predicted values were calculated and used to color the heatmap. The number of significant correlations for each taxon is also shown, in the left column. The ANN was able to predict the taxonomic abundance of each taxon within less than 16% of the measured value. The taxa with the largest differences were present at average abundances of >14%, therefore the differences could be due to background noise. The majority of predicted values were <1% different from the measured values.

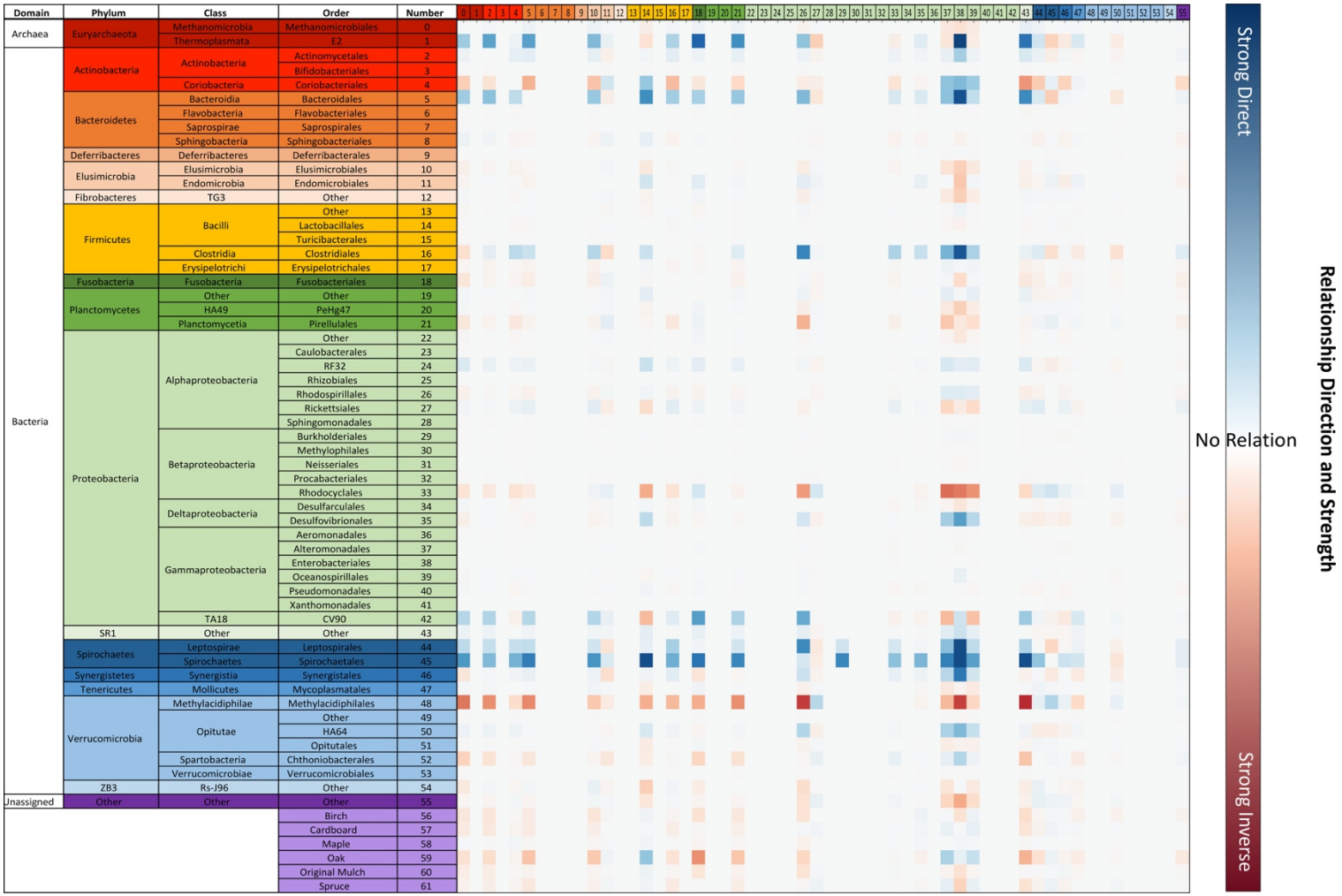


Figure 6. 2D heatmap of influences of taxa and substrates on other taxa. The abundance of each taxon/substrate (on the left of the heatmap) was changed by ±5% and the effect on the remaining taxa is shown in the heatmap. Direct correlations are shown in blue and inverse correlations are shown in red.

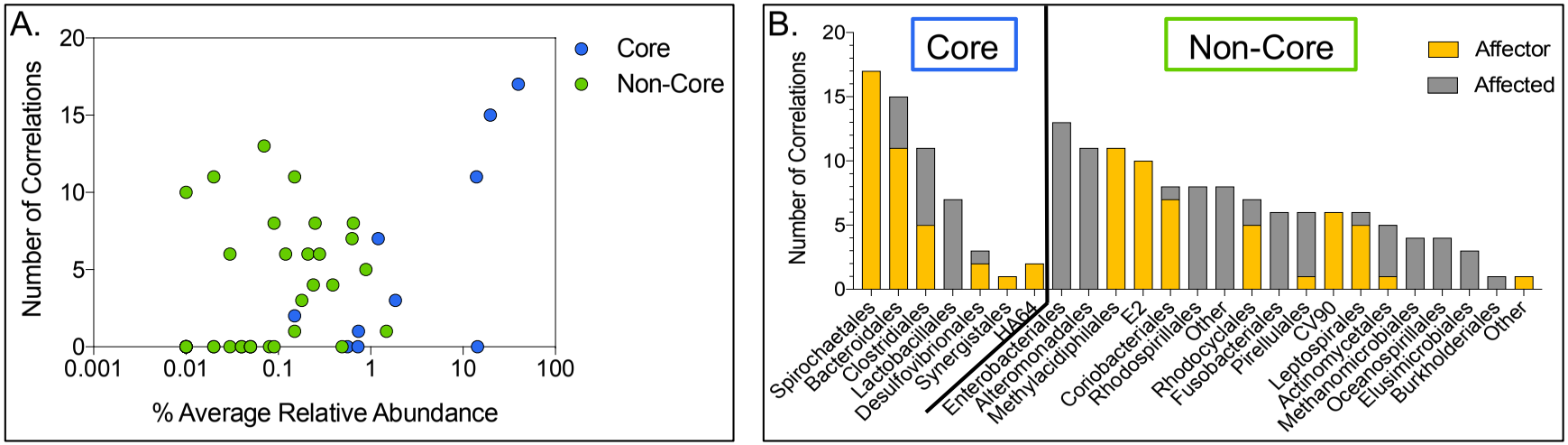


Figure 7. Significantly correlated taxa in the hindgut. A taxon was considered significantly correlated if the value in the heatmap (figure 6) was above three standard deviations of the absolute average of the percent change matrix generated. The majority of significant correlations are associated with taxa below 1% average relative abundance (A). The core and non-core taxa were grouped and the number of affector (yellow) and affected (gray) correlations are shown. Six of the core taxa have affector correlations, while 9 of the non-core taxa have affector correlations (B).

1. Foreach CSV do

2. Foreach time point – 1 in CSV do

3. Input ← time point

4. Target ← time point + 1

5. Write input, target to training file

6. End Foreach

7. End Foreach

8. While error > allowed_error and epochs < max_epochs do

9. Train network

10. End While

11. Save network

Supplemental Figure 1. Algorithm for training the ANN

1. Original_Outputs ← Run_ANN(Original_Inputs)

2. Foreach input_node in Original_Inputs:

3. Foreach value in 100 values ±5% of input_node:

4. New_Outputs ← Run_ANN(New_Inputs)

5. Changes ← Percent_Change(Input, Outputs)

6. End Foreach

7. Average_Change ← Average(Changes)

8. End Foreach

Supplemental Figure 2. Algorithm for sensitivity analysis.

Chapter Four

Insights into the Physiologies of Endosymbiotic Bacteria of Two Protist Species Living in the Hindgut of the Termite, *Reticulitermes flavipes*

Contributions from other researchers

Michael Stephens and Dr. Daniel Gage contributed to this chapter. Michael Stephens isolated single protist cells for the project and assembled draft genomes for *Endomicrobium* sp. PV7 and *Endomicrobium* sp. TA21, which were used in the study. Dr. Gage quality trimmed and filtered the raw sequencing data that was used for draft genome assembly and transcriptome analysis.

Introduction

A subset of symbiotic bacteria exists as endosymbionts living inside of eukaryotic cells and are essential to host survival, while the host provides a protective environment for the bacterium. These populations are especially important in insects, which feed on nutrient-poor substrates, endosymbionts are necessary to provide the host with nutrition, immunity, and fecundity (Wernegreen, 2002). The continued presence of the important intracellular symbionts in offsprings is ensured by the vertical transmission of these bacteria. Mealybugs (Pseudococcidae) have primary and secondary endosymbionts, both of which contribute to nutrient provision to the eukaryotic host (Dohlen et al., 2001). The phylogenies of mealybugs and its two endosymbionts suggest cospeciation of the primary symbiont (*Tremblaya*) and the mealybug host (Baumann, 2005). Endosymbiont genomes are highly reduced when compared to their free-living counterparts, and the host can play a role in genome reduction by providing a protective environment for the bacterium, acquiring genes from the bacterium, or acquiring a second bacterial symbiont to supplement the genes lost in the primary symbiont, ensuring the symbiosis remains stable (McCutcheon & Moran, 2011). Psyllids acquired secondary endosymbionts which attributed to gene loss in *Carsonella*, and that the endosymbiotic *Carsonella* co-evolved with the secondary endosymbiont to supplement the gene loss (Sloan & Moran, 2012).

Aphids, insects feeding strictly on phloem sap, have endosymbionts residing in the hemocel which provide the insect with essential amino acids and act as an ammonia sink. Studies have shown that aposymbiotic aphids fed substrates lacking one essential amino acid per substrate had lower survival rates than symbiotic aphids (Dadd & Krieger, 1968). Also, symbiotic aphids continually thrived better than aposymbiotic aphids regardless of an amino acid

enriched diet (Mittler, 1971). Wilkinson et al. used symbiotic and aposymbiotic *Acyrthosiphon pisum* aphids to study the ability of *Buchnera* species to assimilate host-derived ammonia into glutamine. In this study, aphids with *Buchnera* had higher concentrations of glutamine in the honeydew (sugar-rich liquid secretions) and tissue homogenates than aposymbiotic aphids (Wilkinson & Douglas, 1995). The genome of *Buchnera* sp. APS is highly reduced at only 640kbp and contains 54 genes involved in the biosynthesis of essential amino acids for the aphid host, but pathways for some essential and nonessential amino acids are incomplete or missing (Shigenobu et al., 2000). The ability of this bacterium to synthesize these nonessential amino acids is partly due to the missing genes being present in the aphid genome. Together, the aphid and its endosymbiont possess the suite of genes necessary for the production of tyrosine and methionine (Wilson et al., 2010). This finding adds to the knowledge of endosymbionts co-evolving with their eukaryotic hosts and the need for the host to maintain the bacteria.

The tsetse fly (Diptera: *Glossinidae*) relies on its endosymbiont *Wigglesworthia glossinidia,* which resides in epithelial cell bacteriocytes. As tsetse feeds on vertebrate blood, many of the required nutrients for survival come from the bacterial endosymbionts. The genome of *W. glossinidia brevipalpis* is 697kbp in length and has a GC-content of 22%, which is characteristic of endosymbiotic reduced genomes and low %GC. The genome contains genes involved in the biosynthesis of essential amino acids, multiple cofactors, and B vitamins (Akman et al., 2002). Tsetse fed antibiotic-spiked blood meals had significantly lower fecundity, showing that the symbiotic bacteria not only aided in nutrient provision, but also in fly fertility (Pais et al., 2008).

Bacterial endosymbionts are also found in single-celled eukaryotic protists in the hindguts of lower termites. Termites feed on nutrient-poor wood substrates and require a plethora

of gut bacteria to aid in the digestion of wood and nutrient provision for the termite. In lower termites, this process is enhanced by the presence of flagellate protists (Ohkuma, 2008). *Treponema* species have been found in the protist *Eucomonympha* sp and shown to perform acetogenesis and nitrogen fixation within the cell. *Eucomonypha* sp breakdown cellulose into acetate, H_2, CO_2, and N_2. The *Treponema* use H_2 and CO_2 for acetogenesis and fix N_2 into NH_4, which is released into the hindgut lumen (Ohkuma et al., 2015). The genome of an uncultured Bacteroidales endosymbiont of a protist (*Pseudotrichonympha grassii*) in the termite *Coptotermes formosanus* was sequenced and includes genes for nitrogen fixation and uric acid recycling (Hongoh et al., 2008b). Another protist endosymbiont is an *Endomicrobium* referred to as Rs-D17, isolated from a *Trichonympha agilis* cell in the hindgut of *Reticulitermes speratus*. It is estimated that around 4,000 cells od Rs-D17 are present in one *T. agilis* cell. The genome is larger than other endosymbionts, at a length of 1,125,857 bp and has a GC-content of 35%. The Rs-D17 genome contains genes for glycolysis and fermentation of sugars to acetate. The complete pathways for synthesis of 15 amino acids (and 5 incomplete amino acid pathways) suggest the importance for the endosymbiont to provide substrates for the protist and termite (Hongoh et al., 2008a).

Although *Endomicrobia* bacteria have been known as endosymbionts of protist cells, a free-living cell has been isolated and its genome sequenced. This bacterium, *Endomicrobium proavitum,* was found in the termite *Reticulitermes santonensis* when defaunated (fed starch) (Zheng et al., 2016b) and has also been found in higher termites (lacking protists) and ruminant. The existence of *E. proavitum* in higher termites and ruminant suggests that the ability to exist as an endosymbiont developed later. The genome of *E. proavitum* is 1,588,979 bp in length and has

a GC-content of 39%. This bacterium can only utilize D-glucose as a carbon source and produces lactate, acetate, H_2, and CO_2 (Zheng & Brune, 2015).

The termite hindgut environment varies spatially and can be microaerophilic at the gut tissue interface, or anaerobic in the center of the lumen. The hindgut paunch of *Reticulitermes flavipes* termites is around 800 μm in diameter and through microelectrode analysis, it was found an oxygen concentration of 50 μM at the hindgut wall. As the electrode went deeper into the paunch, the oxygen concentration dropped to 0 μM about 100 μm into the paunch (Brune et al., 1995a). The aposymbiotic midgut was also subjected to microelectrode analysis and it was found that the midgut never becomes anoxic, maintaining oxygen concentrations of 100-150 μM. This finding suggested that the microbes residing in the hindgut are responsible for the oxygen gradient, acting as a sink along the hindgut wall. A simultaneous study was performed to determine the rate of breakdown of lignocellulose-derived aromatic compounds by the hindgut community. Crude gut homogenates were kept in anaerobic conditions while the breakdown of aromatic compounds was measured. It was found that these compounds were not fully metabolized under anaerobic conditions, suggesting that oxygen was required (Brune et al., 1995b). As the oxygen concentration diminishes, the hydrogen concentration in the hindgut has been shown to increase in the lumens center to a partial pressure of 50 mbar (Ebert & Brune, 1997).

The different niches in the hindgut are home to different metabolically active microorganisms. Bacteria, archaea and protists have been found to colonize the gut wall of lower termites. Oxymonad protists belonging to the genus *Pyrsonympha* have been found to inhabit the microaerophilic niche, attaching to the hindgut wall (Tamschick & Radek, 2013), while other protist species inhabit the anaerobic lumen. It is hypothesized that the organisms living attached

to the gut wall utilize the low oxygen concentrations and perform different metabolic functions than the anaerobic organisms. Along with inhabiting different niches, *P. vertens* and *T. agilis* harbor different bacteria. *Trichonympha agilis* has both *Endomicrobium* endosymbionts and *Treponema* extosymbionts, while *P. vertens* organisms only have *Endomicrobium* endosymbionts. The differences in bacterial load may also have an impact on the metabolism of the protist and the endosymbiotic bacteria within them.

Many endosymbionts are unculturable due to the strict reliance on the eukaryotic host. The advancement in genome sequencing and single genome amplification using Φ29 DNA polymerase has enhanced the ability of researchers to study these organisms (Lasken & Raghunathan, 2005). In this study, five protist species were chosen based on morphology, isolated in triplicate and identified using 18S rRNA gene sequencing. The DNA and cDNA from these samples were amplified using multiple displacement amplification and sequenced. The genomes of *Endomicrobia* endosymbionts from multiple protist species were compared to a the free-living *E. proavitum* (Zheng & Brune, 2015) and to each other along with the comparison of gene expression. Full pathways for the utilization of various carbon sources were expressed in both bacteria, along with the biosynthesis of various amino acids and B vitamins. Differential gene content and expression was found between the bacteria with possible correlations to their different niches.

Methods

Protist Isolation

Materials used for protist isolation were submerges in RNAse Away™ overnight then rinsed with 70% ethanol and UV sterilized before use. Each termite was dissected separately to

remove the hindgut paunch, then the hindgut was ruptured into 500 µL of ice cold Trager's Solution U (TU) (Trager, 1934). After rupturing the paunch, the hindgut contents were centrifuged twice for 90 seconds at 3k RPM to enrich for protist cells. The protist cells were then diluted 1:5 in TU and kept on ice. Protist aliquots of 10 µL were added to a microscope slide and a single live protist was targeted and removed from the aliquot using an CellTram® Vario (Eppendorf). The isolated cell was placed in three separate drops of TU to be washed then added to a 10 µL drop of phosphate buffered saline (PBS). The cell in a 10 µL suspension was added to a PCR tube and immediately placed on dry ice, and then stored in a -80˚C freezer. This process was completed until four protist cells from each of five species was obtained. Negative control samples were prepared by adding 2 µL of TU from the last TU droplet on the microscope slide along with 8 µL of PBS on the same slide. The isolated samples were named by the protist genus, species, and number of isolation. Protists isolated for this study include *Trichonympha agilis* (TA) and *Pyrsonympha vertens* (PV).

Whole Genome and Whole Transcriptome Amplification

The genome (DNA) and transcriptome (cDNA) from the protist cells were amplified between 12-24 hours after isolation. Cell lysis and amplification was performed using the Repli-g WGA/WTA kit (Qiagen). Cells were lysed using a Qiagen lysis buffer and incubation step, immediately followed by incubation on ice. Aliquots from the same lysed cell were separated and used in simultaneous whole genome amplification and whole transcriptome amplification. The process was carried out per the kits standard protocol with exception of the addition of primers for amplification. For two samples per protist species, random hexamer primers were used to amplify mostly bacterial DNA and cDNA.

Library Preparation and Sequencing

DNA and cDNA was sheared using a Covaris M220 ultra-sonicator™ according to standard protocol. WGA samples were sheared to a 550 bp insert size using 200 ng of DNA and WTA samples were sheared to a 350 bp insert size using 100 ng of cDNA. Sequencing libraries were prepared using the TruSeq Nano DNA Library Prep kit from Illumina according to standard protocol. Each sample was prepared with a forward and reverse barcode for the ability to add all samples on the same sequencing run. The samples were sequenced on one Illumina NextSeq 1x150 Mid Output run and two NextSeq 1x150 High Output runs.

Quality Control of Metagenomes and Metatranscriptomes

Raw Illumina Read1 and Read2 files from all three runs were concatenated and then paired using BBMap (Bushnell, 2014). Paired reads were quality trimmed using BBDuk based on a phred score of Q15, removal of homopolymers, removal of Illumina adapters, and read length of >50 nt (www.jgi.doe.gov). Trimmed datasets were mapped to reference genomes of known contaminants and reads that mapped with 95% coverage to the reference were discarded. After trimming, rRNA reads were removed from the WTA samples and WGA samples were normalized and duplicate reads were removed.

Draft Genome Assemblies

Metagenomes of *Endomicrobium* sp. TA21 and *Endomicrobium* sp. PV7 from *Trichonympha agilis* (TA21) and *Pyrsonympha vertens* (PV7) respectively, were assembled using the A5 pipeline in Kbase (Tritt et al., 2012). Kbase starts by cleaning reads with

Trimmomatic (Bolger et al., 2014), followed by error correction with SGA (Simpson & Durbin, 2010), contig assembly using IDBA-UD (Peng et al., 2010), initial scaffolding with SSPACE (Boetzer et al., 2011), missasembly correction with BWA (Kelley et al., 2010), and final scaffolding using repaired broken contigs. Scaffolds were binned using VizBin (Laczny et al., 2015) by kmer frequency of 4. A blastn (Altschul et al., 1990) search was performed each scaffold present in the *Endomicrobia* bin, against a database of all *Endomicrobia* reference genomes. If there was a significant hit (E-value $<10^{15}$, 80% sequence identity, >1kb length) to a reference genome, the scaffold was entered into the draft genome. For scaffolds without a significant hit to a reference genome, a blastn search was performed against the non-redundant database, and scaffolds with a hit to *Endomicrobia* were considered for the draft genome. Blastn was also used to search the scaffolds not present in the original *Endomicrobia* bin, and any scaffolds hitting to the Endomicrobia reference genomes were also added to the draft genome. The 16S rRNA genes from a published *Endomicrobia* genome were used as a reference to find 16S rRNA genes in the draft genome. The program RNAmmer (Lagesen et al., 2007) was also used to find rRNA genes to be added to the draft genomes. The draft genomes were uploaded to the RAST (Rapid Annotation using Subsystem Technology, www.rast.nmpdr.org) server and annotated using RASTtk (Brettin et al., 2015).

Metatranscriptome Mapping

Metatranscriptomes from *Trichonympha agilis* and *Pyrsonympha vertens* samples were analyzed with CLC Genomics Workbench v.9 (https://www.qiagenbioinformatics.com), using the draft genomes for *Endomicrobium* sp. TA21 and *Endomicrobium* sp. PV7 for reference. The expression value of each gene was calculated using the following equation and $\log_2$ transformed:

$$Expression\ value = \frac{\#\ reads\ mapped\ to\ target}{(length\ of\ gene\ (in\ kb) \times (total\ reads\ mapped\ (in\ millions)}$$

Results

Metagenome and Metatranscriptome Sequencing of Two Protist Species

Two protist species were selected for this study based on the differences of the protists themselves (different lineages), differences of associated symbionts, and differences of environmental niches, *Trichonympha agilis* and *Pyrsonympha vertens*. Three protist cells were isolated and subjected to simultaneous whole genome and whole transcriptome amplification and the resulting samples were sequenced. One cell for each protist species was also amplified using oligo-dT primers to select for protist transcripts. Fourteen total libraries, metagenomes and metatranscriptomes, were sequenced. The sequences for each sample were quality filtered and shown in Table 1. While the protists have both endosymbionts and ectosymbionts, the draft genomes from the endosymbionts assembled into higher quality genomes, with the best quality genomes being for *Endomicrobium* sp. PV7 and *Endomicrobium* sp. TA21. These bacteria are endosymbionts of protists from two different taxonomic groups, Parabasalids (TA21) and Oxymonads (PV7), and inhabit different environmental niches in the termite hindgut (anaerobic and microaerophilic, respectively). *Trichonympha agilis* harbors endosymbionts along with ectosymbionts from the genus *Treponema*, while *P. vertens* only harbors endosymbionts. The draft genomes for these bacteria were assembled, checked for completeness, and were made up of 116 (PV7) and 25 (TA21) scaffolds (Table 2).

Comparison of *Endomicrobium* Genomes

Draft genomes were created for *Endomicrobium* sp. PV7 and *Endomicrobium* sp. TA21 and genome content was compared. The genome of *Endomicrobium* sp. PV7 is 1.25 Mb in size, consisting of 116 scaffolds with a GC content of 35.6%. Genomes were then added to a program (CheckM) that assesses genome completeness based on the presence of marker genes commonly found in closely related reference genomes. *Endomicrobium* sp. PV7 appears to be 92.6% complete and *Endomicrobium* sp. TA21, 96.6% complete and contain 1,334 and 1,309 genes, respectively. While the draft genomes were similar, it appears that the *Endomicrobium* symbionts from *Pyrsonympha vertens* and *Trichonympha agilis* are different species (Figure 1). A blast comparison using Blast Ring Image Generator (BRIG) was carried out with the two endosymbiont draft genomes and a non-intracellular ("free-living") symbiont, *Endomicrobium proavitum* and showed a ~20% reduction in the genomes of *Endomicrobium* sp. PV7 (22%) and TA21 (19%), or that E. proavitum gained genes (Alikhan et al., 2011). A comparison of genes differing between the free-living *E. proavitum* and the endosymbionts *Endomicrobium* sp. PV7 and TA21 (Figure 2). To correct for the draft genomes not being complete, a closed endosymbiont of the same genus, *Endomicrobium* Rs-D17 was also used in this comparison. Also, the transcriptomes of *Endomicrobium* sp. PV7 and TA21 were mapped to the genomes of Rs-D17 and *E. proavitum* to further confirm the genes were absent. *Endomicrobium proavitum* contains genes for the biosynthesis of proline, DNA repair, membrane transporters, protein biosynthesis, osmotic stress, and resistance to toxic compounds, all of which are absent from the three endosymbionts. The endosymbionts contain genes for arabinose and hexuronate catabolism, synthesis of B vitamins, restriction-modification systems, protein degradation, RNA

metabolism and modification, cold shock proteins, and carbon starvation response proteins. These gene categories may contribute to a free-living vs. endosymbiotic state of these bacteria.

Carbon Utilization in *Endomicrobium* sp. PV7 and *Endomicrobium* sp. TA21

Carbon utilization pathways were found in the draft genomes to the two endosymbiotic *Endomicrobia* and the expression data for the genes in the pathways was compared. The endosymbionts of both *Pyrsonympha vertens* and *Trichonympha agilis* have the ability to utilize D-glucose as a carbon source (Figure 3). All genes in the glycolysis pathway are present in the respective draft genomes and the bacteria both express all genes in the pathway. As this is the only carbon utilization pathway that both bacteria express, suggesting that D-glucose is the main carbon nutrient for these bacteria. Gulonate, a hexamer derived from pectin can be used by both species as well, with expression of genes catabolizing gulonate to glyceraldehyde-3P, however, *Endomicrobium* sp. PV7 doesn't possess one of the genes, Uronate Isomerase, to complete this catabolism. Both bacteria also contain the genes for xylulose import and catabolism to D-ribulose and β-D-fructose-6P. *Endomicrobium* sp. PV7 also contains genes for the metabolism of arabinose from xylulose. *Endomicrobium* sp. TA21 has the additional ability to catabolize cellulose and glucuronate (another derivative of plant pectin). Cellulose can be broken down using the cellulase β-1,4-glucanase, leading into glycolysis for further catabolism. Glucuronate is imported into the cell and catabolized to pyruvate. The ability for *Endomicrobium* sp. TA21 to breakdown cellulose and glucoronate may suggest that the *T. agilis* protist cell breaks down components of wood and plant pectins into these substrates making them readily available for the endosymbiont to utilize them.

Biosynthesis in *Endomicrobium* sp. PV7 and TA21

The ability for bacteria to synthesize compounds in the protist symbiosis is important due to the lack of these compounds in the wood meal, especially amino acids and B vitamins. Biosynthesis genes for fifteen amino acids were present in both *Endomicrobium* draft genomes and transcriptomes (Figure 3). Neither *Endomicrobium* sp. PV7 nor TA21 retained full pathways for lysine, threonine, or methionine biosynthesis. To determine if the genes were detected in the transcriptome, the draft genome from the opposite bacterium was used as the reference for RNA-Seq analysis (PV7 genome for TA transcriptomes, TA21 genome for PV transcriptomes). Some of the "missing genes" were present in the transcriptomes, which suggests an annotation error, but others were confirmed to be missing and marked with a gray 'X' in Figure 3. To determine whether transcriptome mapping to a different genome would not alter the expression value problem, six housekeeping genes (*dnaJ, dnaK, gyrA, gyrB, groEL, groES*) were compared (Figure 5/Table 3). There was no pattern of differential expression values in the six housekeeping genes. Both bacteria maintain the full biosynthesis pathways for isoleucine, leucine, valine, serine, glutamine, aspartate, and phenylalanine. *Endomicrobium* sp. PV7 has the ability to synthesize proline, while *Endomicrobium* sp. TA21 is missing these required genes. *Endomicrobium* sp. TA21 expresses genes for the synthesis of tryptophan, tyrosine, histidine, and arginine. However, *Endomicrobium* sp. PV7 only expresses about half of the genes for each pathway. Both bacteria can synthesize fumarate and malate but retain incomplete pathways for the citric acid cycle. *Endomicrobium* sp. TA21 expresses an Acetyl-coenzyme-A carboxyl transferase gene, leading to the production of acetate which is the major nutrient source of the termite. Biotin, a B-vitamin thought to be an important cofactor produced by *Endomicrobium* species is synthesized by both endosymbionts, along with pantothenate and folate. Partial

pathways are also expressed for thiamine, cobalamin, and riboflavin. Genes for the catabolism of riboflavin to FAD (flavin adenine dinucleotide) are expressed, which suggests that the incomplete pathway may be due to incomplete genome assembly or annotation. The biosynthesis pathways in *Endomicrobium* sp. PV7 and TA21 shed light on not only the metabolism of the bacteria, but may lead to hypotheses regarding the physiology of the protists as well.

Niche-specific Genes in *Endomicrobium* sp. PV7 and TA21

Pyrsonympha vertens and *Trichonympha agilis* protists inhabit different niches in the termite hindut. *Pyrsonympha. vertens* attaches to the hindgut wall residing in a microaerophilic environment with low hydrogen concentrations, while *T. agilis* cells persist in the anaerobic hindgut lumen, with high hydrogen concentrations. It has been hypothesized that each protist performs different functions due to the differences in the environment in which they live. Therefore, it is imaginable that the endosymbiotic bacteria within those protists also perform different functions. A comparison of the endosymbionts genomes showed 636 shared genes between the two bacteria, and each had about 100 unique genes. *Endomicrobium* sp. PV7 and *Endomicrobium* sp. TA21 express genes for a membrane aerotolerance protein complex in the *batI* operon. Interestingly, *Endomicrobium* sp. PV7 possesses the *batB* gene which is missing from the genome of *Endomicrobium* sp. TA21, and expresses the suite of *bat* genes at higher levels (Figure 4). In contrast, *Endomicrobium* sp. TA21 contains the complete pathway for glucoronate catabolism, whereas *Endomcirobium* sp. PV7 is missing the first two genes in the pathway (Figure 4). Along with the difference in gene content, *Endomicrobium* sp. TA21 also has higher expression of the remaining genes in the glucuronate metabolism pathway compared to *Endomicrobium* sp. PV7.

Discussion

Bacteria from the phylum Elusimicrobia (previously known as Termite Group 1 (TG1)) have been identified as specific to the guts of termites, and is represented thus far by three bacterial genomes. These bacteria have been thought to be strict endosymbionts of the protist cells living inside the termite hindgut. One organism, *Endomicrobium proavitum*, was cultured and sequenced from the termite *Reticulitermes santonensis* and is presumed to be a free-living bacterium (Zheng et al., 2016b). The genome of *E. proavitum* is 1.58Mbp in size with a GC-content of 39%. Two other *Endomicrobium* symbionts, *Endomicrobium* sp. Rs-D17 (1.25 Mbp, 35% GC) from protists in *Reticulitermes speratus* and *Endomicrobium minutum* (1.64 Mbp, GC not reported) from the scarab beetle *Pachnoda aphippiata* (Geissinger et al., 2009; Hongoh et al., 2008a). In comparison, the two endosymbionts from *R. flavipes* protists, *Endomicrobium* sp. PV7 and TA21 have 35% GC content and 1.25 Mbp/1.36 Mbp sized genomes, respectively. It is known that genome reduction and low %GC occurs in strict endosymbiotic bacteria (Sloan & Moran, 2012), which is indicative of the endosymbiont not requiring many genes because of the protective environment or supplementation of substrates from the host. *Carsonella*, an endosymbiont of psyllids, was shown to have active gene loss and supplemented by a secondary symbiont of the psyllid (Sloan & Moran, 2012). Both of the *Endomicrobium* genomes in this study appear to have reduced genomes and low GC-content, which are similar in size and %GC to the endosymbiont Rs-D17, however more analysis is required to confirm the reduced genomes as these are not closed.

Endomicrobium proavitum and the endosymbionts (*Endomicrobium* sp. PV7, TA21, and Rs-D17) have several groups of genes that differ, suggesting their roles may be specific to the symbiotic state of the bacterium. *Endomicrobium proavitum* contains genes for proline

biosynthesis which have been shown to be absent from Rs-D17 (Hongoh et al., 2008a) and are also absent from the endosymbionts in this study. This genome also has genes contributing to DNA repair which are absent from the endosymbionts, which has been shown in other endosymbiotic reduced genomes (Sloan & Moran, 2012). Protein synthesis genes are also present in *E. proavitum* along with membrane transporter genes, such as the *opp* (oligopeptide permease) gene cluster, which is important in nutrition and signaling (Nepomuceno et al., 2007). The final gene categories present in *E. proavitum* and absent in the endosymbionts are for osmotic stress and resistance to toxic compounds which may be more important for this bacterium compared to intracellular symbionts that are protected from the environment of the hindgut. The endosymbiotic bacteria also have several gene clusters specific to their genomes such as arabinose and hexuronate catabolism from hemicellulose and plant pectins, respectively. The presence of these genes in the endosymbionts and not in the free-living symbiont suggests that these bacteria require substrates from the protist after hemicellulose and plant pectin degradation. The endosymbionts also contain genes for the synthesis of multiple B vitamins such as thiamin, biotin, folate, and riboflavin, all of which are not present in the genome of *E. proavitum*. Multiple genes for a restriction-modification (RM) system is present in the endosymbionts which has been suggested as an important feature of termite protist endosymbiotic bacteria due to the ability of viruses and other bacteria entering the protist cytoplasm through ingestion of wood particles (Zheng et al., 2016a). Along with the RM system, genes for RNA metabolism/modification and protein degradation were also present in the endosymbionts. Finally, genes for cold shock proteins and carbon starvation response proteins were found in the endosymbionts. The presence and absence of genes in either the free-living or

endosymbiotic *Endomicrobia* suggest these gene categories as promising endosymbiont-specific genes, as well as provide further evidence on the physiology of the protists in which they live.

The gene expression data of *Endomicrobium* sp. PV7 and TA21 show that the bacteria are able to use D-glucose as a carbon source, which has also been suggested to be the main carbon source for other *Endomicrobium* species. *Endomicrobium proavitum* seems to only be able to metabolize D-glucose, while it is used by Rs-D17 and *E. minutum*, among other sources (Geissinger et al., 2009; Hongoh et al., 2008a; Zheng et al., 2016b). *Endomicrobium minutum* has been shown to utilize D-fructose, D-galactose, and N-acetyl-D-glucosamine as well (Geissinger et al., 2009). D-glucose is a component that makes up cellulose, which has been shown to be broken down by the protists of the hindgut (Tartar et al., 2009). *Endomicrobium* sp. TA21 also expresses genes for the breakdown of cellulose to cellobiose and the breakdown of gulonate to Glyceraldehyde-3P, which both feed into glycolysis. *Endomicrobium* sp. PV7 expresses the genes to import and breakdown xylulose to arabinose. Xylulose is a component of hemicellulose, which is a more complex, branched polymer (Kluepfel, 1988). It is not known which protist species express the genes for the degradation of different wood substrates, but perhaps the ability of *Endomicrobium* sp. PV7 to metabolize xylulose suggests that *P. vertens* contains enzymes for hemicellulose degradation while *T. agilis* may contain enzymes for cellulose and plant pectin degradation.

An important role for symbionts of organisms that feed on nutrient-poor diets, is the ability to biosynthesize essential amino acids, cofactors, and vitamins. Aposymbiotic aphid development is halted even when fed amino acids in addition to the normal diet of phloem sap, showing the endosymbionts are necessary for amino acid production (Mittler, 1971). As wood is a nutrient-poor diet, the termites require the symbionts to breakdown the substrate and provide

nutrition to the termite. *Endomicrobium* sp. PV7 and TA21 express genes for the synthesis of fifteen amino acids, including eight essential amino acids. Neither bacterium retains the ability to synthesize alanine, asparagine, cysteine, lysine, or methionine. The lysine and methionine pathways are represented by few genes in each bacterium, which provides evidence for a reduced genome and that the bacteria may need aid from the protist or other symbionts. Wilson et. al. identified three amino acid synthesis pathways that require both the endosymbiont *Buchnera aphidicola* and the pea aphid host *Acyrthosiphon pisum* (Wilson et al., 2010). *Endomicrobium* Rs-D17 has the ability to synthesize fifteen amino acids and must import asparagine, cysteine, glutamine, proline, and serine. Unlike Rs-D17, *Endomicrobium* sp. PV7 and TA21 have the ability to synthesize glutamine, proline and serine. The free-living *E. proavitum* has the ability to synthesize all amino acids needed for protein synthesis, which was determined by the ability of the bacterium to grow on minimal medium without amino acids added (Zheng et al., 2016b). *Endomicrobium* sp. PV7 and TA21 also express genes for the production of many B vitamins, such as biotin, folate, pantothenate, and partial B12 synthesis. Although these results coincide with findings from other endosymbiotic bacteria, the genomes of *Endomicrobium* sp. PV7 and TA21 are not closed so the results must be verified by other methods.

Along with the ability for *Endomicrobium* sp. TA21 to utilize D-glucose, cellulose, and gulonate as carbon sources, it also expresses genes for the pathway of glucuronate catabolism. Glucuronate is a product of plant pectins and has been shown to be degraded by *Erwinia chrysanthemi*, a bacterium responsible for soft-rot disease in plants (Hugouvieux-Cotte-Pattat & Baudouy, 1987). Pectin is first degraded by a pectin lyase (absent from *Endomicrobium* sp. TA21) then transported into the bacterial cell via a hexuronate transporter. From there, it can be further broken down to pyruvate. While *Endomicrobium* sp. PV7 contains the genes for half of

the catabolism pathway, it is missing the gene for transport into the cell (*exuT*) and the first gene (*uxaC*) for catabolism. The expression of genes for the import and catabolism of glucuronate but the absence of the pectin lyase suggests that the protist digests pectin to glucuronate which can then be broken down to pyruvate by the bacterium. The differences in carbon utilization of *Endomicrobium* sp. PV7 and *Endomicrobium* sp. TA21 suggests that the protist hosts contain different enzymes to breakdown cellulose, pectin, and hemicellulose, providing different substrates to the bacteria.

The genes present in the genomes of *Endomicrobium* sp. PV7 and *Endomicrobium* sp. TA21 were compared and the two bacteria share 636 genes and each bacterium contains 100 and 105 unique genes, respectively. Because *P. vertens* and *T. agilis* occupy different niches, it was hypothesized that the endosymbionts present within each protist would express different genes allowing them to thrive in the different environments of their hosts. *Pyrsonympha vertens* lives attached to the hindgut lumen in a microaerophilic environment, while *T. agilis* lives in the anaerobic hindgut lumen (Tamschick & Radek, 2013). It has been shown that the microorganisms attached to the gut wall act as an oxygen sink consuming the oxygen coming through the termite tissue, rendering the rest of the gut anaerobic (Brune et al., 1995a). As *P. vertens* consumes oxygen, it is likely that the endosymbionts will be exposed to oxygen radicals and need to express enzymes that protect against oxidative stress. A previous study found that specific transposon mutants retarded the ability for *Bacteroides fragilis* to infect tissue monolayers in an aerobic hood (Tang et al., 1999). *Bacteroides fragilis* is an anaerobic bacterium but has the ability to survive under oxygen conditions. The transposon mutants showed disruptions in the Bacteroides aerotolerance operon (*batI*). The *batI* operon consists of genes *batABCDE*, which are all hypothesized to encode for a membrane-bound protein that exports

components with reducing power or reduces periplasmic proteins. A transposon mutant in *batB* inhibited the ability of *B. fragilis* to grow on the tissue monolayer completely. *Endomcirobium* sp. TA21 contains the genes for *batACDE*, but is missing *batB*. *Endomicrobium* sp. PV7 contains all the genes in the *batI* operon and expresses *batCDE* at 2-fold higher levels than *Endomicrobium* sp. TA21, with the highest expression levels for *batB*. The presence and high levels of expression of the *batI* operon in *Endomicrobium* sp. PV7 suggests the ability to survive a microaerophilic environment presented by the host protist attachment to the hindgut wall.

Symbionts are known to provide their host organism with a plethora of substrates used for nutrition and cellular metabolism. Endosymbionts form a special, intimate relationship with their host organism as they reside inside of the hosts cells, and are especially important in insects with nutrient-poor diets (Aksoy, 2000; Sloan & Moran, 2012; Wilson et al., 2010; Wu et al., 2006). Termites harbor a vast number of symbionts consisting of protists, bacteria, and archaea, creating a complex gut environment. While there have been studies of the gut environment (Brune et al., 1995a), bacterial composition (Benjamino & Graf, 2016; Mikaelyan et al., 2015), archeal composition (Leadbetter & Breznak, 1996), whole gut metatranscriptomes (Tartar et al., 2009), and various genomes (Graber et al., 2004; Hongoh et al., 2008a; Hongoh, Sharma, Prakash, Noda, Toh, et al., 2008b), there is still much to learn about the termite hindgut. This study of the metagenomes and metatranscriptomes of the protist endosymbionts begins to shed light on the intimate relationships of this nested symbiosis. By studying the metabolism of the bacteria associated with the hindgut protists, we can begin to understand the habitat, food sources, and physiology of the protist host.

Table 1. Sequencing and quality control statistics of metagenome and metatranscriptome samples from single *Pyrsonympha vertens* and *Trichonympha agilis* cells.

Protist	Library	Sample	# Paired Reads	Quality Trimming	Contamination Removal	rRNA Removal	% of Starting Reads
Pyrsonympha vertens	WGA	PV1	29,366,568	24,807,098	24,559,780		84
		PV7	24,358,298	15,208,376	14,968,148		61
		PV11	26,373,686	21,907,976	16,157,802		61
	WTA	PV1	107,226,450	89,264,890	71,971,984	25,477,526	24
		PV7	13,608,006	1,070,048	905,076	446,406	3
		PV11	85,032,418	67,769,648	46,451,474	18,239,734	21
Trichonympha agilis	WGA	PV2	3,307,452	2,735,302	1,863,148	1,765,332	53
		TA16	31,040,896	23,612,512	16,311,954		53
		TA21	27,586,758	22,562,532	22,457,024		81
		TA26	25,222,580	18,193,434	17,692,692		70
		Enuc	4,282,992		2,310,304		54
		Nuc	4,074,550		2,161,744		53
	WTA	TA16	229,365,166	181,781,634	155,473,548	121,418,504	53
		TA21	144,520,930	128,477,248	114,311,390	7,697,882	5
		TA26	129,047,734	109,173,508	105,113,224	10,757,040	8
		TA17	4,983,122	3,530,500	2,545,730	1,346,328	27
		Enuc	8,486,798		4,544,876	1,471,410	17
		Nuc	8,147,202		3,969,944	240,910	3
Totals			906,031,606	710,094,706	623,769,842	188,861,072	

Table 2. Draft genome assembly metrics for *Endomicrobium* sp. PV7 and *Endomicrobium* sp. TA21.

Draft Genome	Associated Protist	Size (Mb)	GC%	# Scaffolds	N50 (kb)	Completeness (%)	Number of Genes	Coding Density (%)
Endomicrobium sp. PV7	*Pyrsonympha vertens*	1.25	35.6	116	14	92.6	1,334	81.2
Endomicrobium sp. TA21	*Trichonympha agilis*	1.36	36.5	25	164	96.6	1,309	75.8

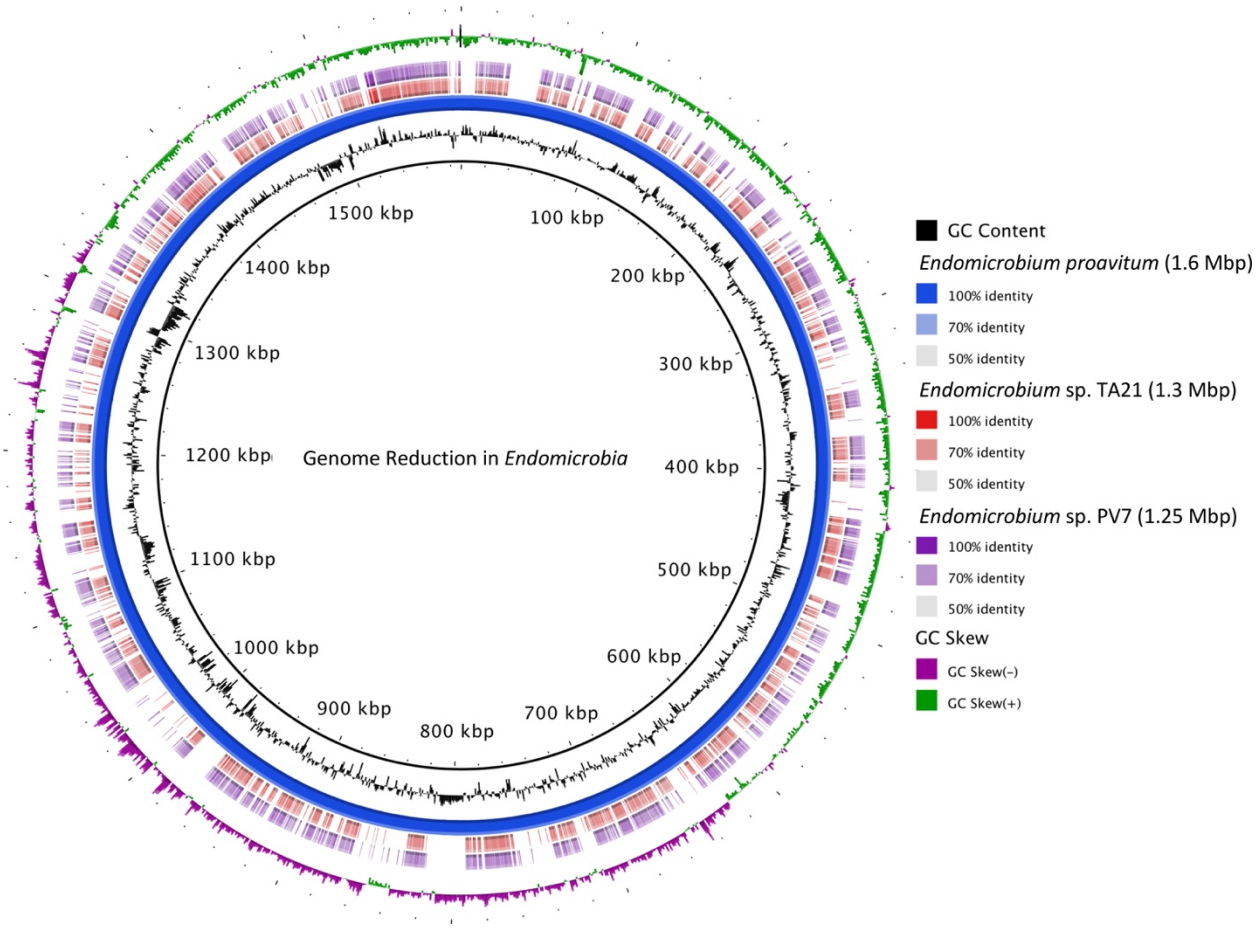


Figure 1. Comparison of *Endomicrobium* genomes. The draft genomes of *Endomicrobium* sp. PV7 (purple) and *Endomicrobium* sp. TA21 (red) were compared to the genome of the free-living *Endomicrobium proavitum* (blue ring). The gaps shown in the red and purple rings shows genome reduction from the free-living bacterial genome to the endosymbiotic genomes. The black lines show a relatively low GC content, while the purple line shows a negative GC skew (low %GC) and the green line shows a positive GC skew (high %GC).

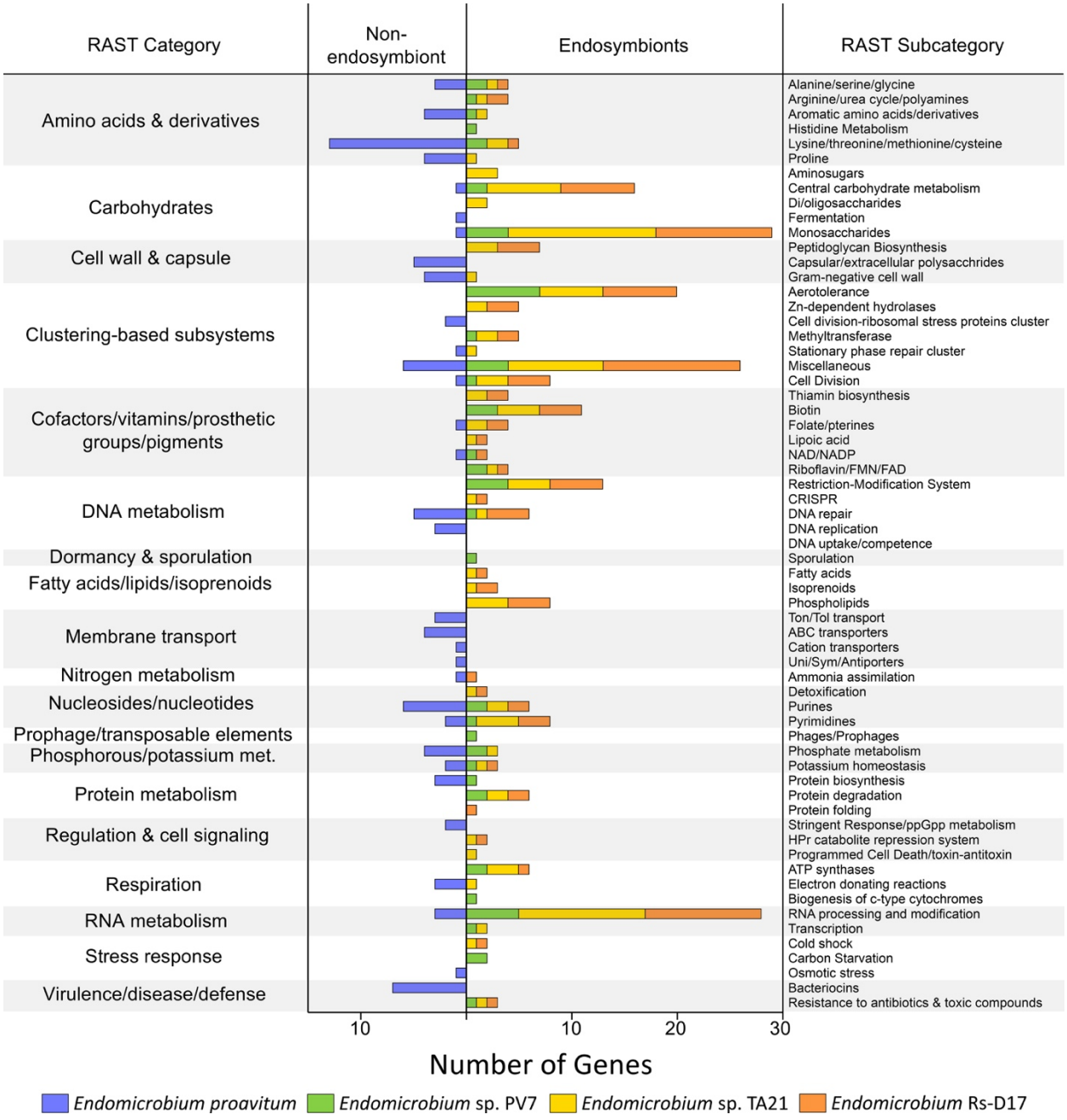

RAST Category
Non-endosymbiont
Endosymbionts
RAST Subcategory
Amino acids & derivatives
Alanine/serine/glycine
Arginine/urea cycle/polyamines
Aromatic amino acids/derivatives
Histidine Metabolism
Lysine/threonine/methionine/cysteine
Proline
Carbohydrates
Aminosugars
Central carbohydrate metabolism
Di/oligosaccharides
Fermentation
Monosaccharides
Cell wall & capsule
Peptidoglycan Biosynthesis
Capsular/extracellular polysacchrides
Gram-negative cell wall
Clustering-based subsystems
Aerotolerance
Zn-dependent hydrolases
Cell division-ribosomal stress proteins cluster
Methyltransferase
Stationary phase repair cluster
Miscellaneous
Cell Division
Cofactors/vitamins/prosthetic groups/pigments
Thiamin biosynthesis
Biotin
Folate/pterines
Lipoic acid
NAD/NADP
Riboflavin/FMN/FAD
DNA metabolism
Restriction-Modification System
CRISPR
DNA repair
DNA replication
DNA uptake/competence
Dormancy & sporulation
Sporulation
Fatty acids/lipids/isoprenoids
Fatty acids
Isoprenoids
Phospholipids
Membrane transport
Ton/Tol transport
ABC transporters
Cation transporters
Uni/Sym/Antiporters
Nitrogen metabolism
Ammonia assimilation
Nucleosides/nucleotides
Detoxification
Purines
Pyrimidines
Prophage/transposable elements
Phages/Prophages
Phosphorous/potassium met.
Phosphate metabolism
Potassium homeostasis
Protein metabolism
Protein biosynthesis
Protein degradation
Protein folding
Regulation & cell signaling
Stringent Response/ppGpp metabolism
HPr catabolite repression system
Programmed Cell Death/toxin-antitoxin
Respiration
ATP synthases
Electron donating reactions
Biogenesis of c-type cytochromes
RNA metabolism
RNA processing and modification
Transcription
Stress response
Cold shock
Carbon Starvation
Osmotic stress
Virulence/disease/defense
Bacteriocins
Resistance to antibiotics & toxic compounds
10
10
20
30
Number of Genes
Endomicrobium proavitum
Endomicrobium sp. PV7
Endomicrobium sp. TA21
Endomicrobium Rs-D17

Figure 2. Intracellular and extracellular symbiont-specific genes found in *Endomicrobium* species. Genomes from the free-living *E. proavitum* and endosymbiotic *Endomicrobium* sp. PV7, TA21, and Rs-D17 were scouted for intracellular or extracellular symbiont specific genes. The categories and subcategories of these genes were grouped by annotation on RAST. Categories of interest were DNA repair and osmotic stress genes present in *E. proavitum* and monosaccharide catabolism, B vitamin synthesis, and restriction-modification genes present in the endosymbionts.

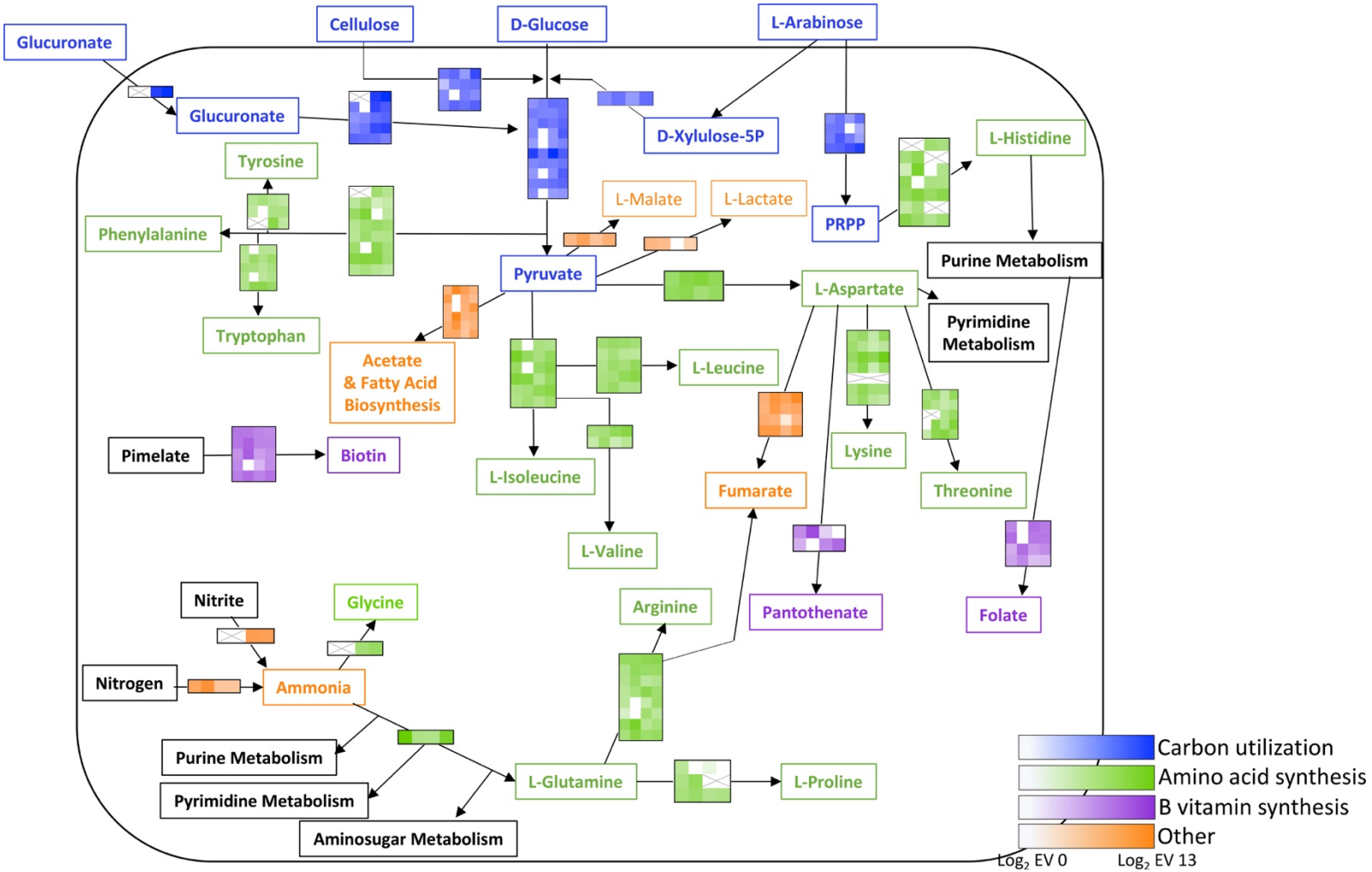


Figure 3. Metabolic pathways of *Endomicrobium* sp. PV7 and *Endomicrobium* sp. TA21. Metabolic pathways for both *Endomicrobium* species are mapped and the expression values of the genes in each bacterium are shown in the heatmaps (in order of pathway). The order of the samples in the heatmaps is as follows: PV1/PV7/TA21/TA26. Carbon utilization pathways are colored in blue, amino acid biosynthesis in green, B vitamin synthesis in purple, and other cofactors in orange. Genes that are not present in the genome or transcriptome have a gray X through the box.

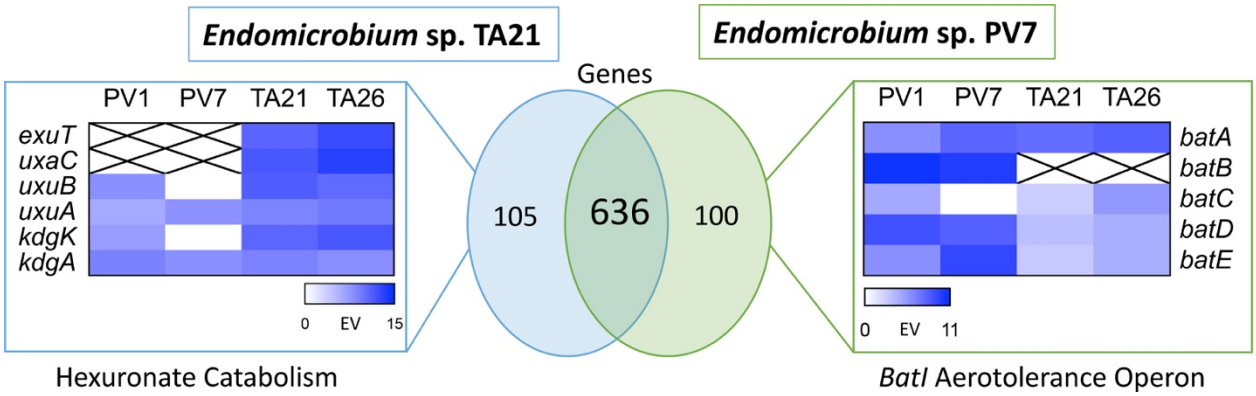


Figure 4. Differential expression of hexuronate catabolism and aerotolerance in *Endomicrobium* sp. PV7 and TA21. A comparison of the genes present in *Endomicrobium* sp. PV7 and *Endomicrobium* sp. TA21 showed that 636 genes were shared between the bacteria, and *Endomicrobium* sp. PV7 and TA21 have 100 and 105 unique genes, respectively. Among the different genes were the pathways for hexuronate catabolism in TA21 and the *BatI* aerotolerance operon in PV7. Expression values were $\log_2$ transformed and shown in the heatmaps. Genes absent from a genome are depicted as an X through the box.

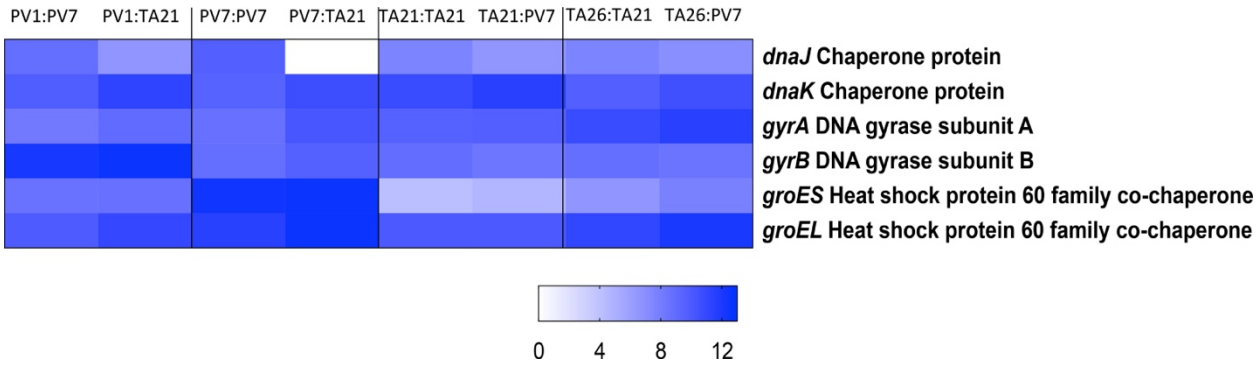


Figure 5. Expression data of six housekeeping genes mapping to different reference genomes. *Endomicrobium* sp. PV7 and Endomicrobium sp. TA21 genomes were used as a reference for mapping transcriptomes from the same species of protist. Alternatively, *Endomicrobium* sp. PV7 was used as a reference genome for transcriptome mapping of TA samples, and visa-versa. The expression values were $\log_2$ transformed and mapped. Overall patterns between the datasets were similar regardless of reference genome used.

Table 3. Standard deviation of expression values of transcriptomes mapped to two reference genomes.

PV1	PV7	TA21	TA26
1.55	6.84	0.73	0.39
1.01	0.84	0.46	0.63
0.53	1.05	0.09	0.46
1.19	0.49	0.38	0.29
0.08	1.79	0.38	0.86
0.79	1.10	0.01	0.61

[+]Standard deviations were calculated from the Log_2 transformed expression values of each transcriptome mapped to its respective genome compared with the same transcriptome mapped to the genome of the other *Endomicrobium* species.

Table 4. Intracellular and extracellular symbiont specific genes categorized by RAST.

Category	RAST Subcategory	RAST Role	EC Number	*Endomicrobium proavitum*	*Endomicrobium* sp. PV7	*Endomicrobium* sp. TA21	*Endomicrobium* Rs-D17
Amino Acids and Derivatives	Alanine, serine, and glycine	D-3-phosphoglycerate dehydrogenase	EC 1.1.1.95	✓			
		2-amino-3-ketobutyrate coenzyme A ligase	EC 2.3.1.29		✓		
		Glycine cleavage system transcriptional antiactivator GcvR		✓			
		Iron-sulfur cluster regulator IscR		✓			
		Serine transporter			✓	✓	✓
	Arginine; urea cycle, polyamines	Carboxyrspermidine dehydrogenase, putative	EC 1.1.1.-		✓	✓	✓
		Ornithine decarboxylase	EC 4.1.1.17				✓
	Aromatic amino acids and derivatives	Amino acid-binding ACT		✓			
		Biosynthetic Aromatic amino acid amitransferase beta	EC 2.6.1.57	✓			
		Indolepyruvate oxidoreductase subunit IorA	EC 1.2.7.8	✓			
		Indolepyruvate oxidoreductase subunit IorB	EC 1.2.7.8	✓			
		Prephenate dehydrogenase	EC 1.3.1.12		✓	✓	
	Histidine Metabolism	Phosphoribosyl-ATP pyrophosphatase	EC 3.6.1.31		✓		
	Lysine, threonine, methionine, and cysteine	Cystathionine beta-lyase	EC 4.4.1.8	✓			
		Cystathionine beta-synthase	EC 4.2.1.22	✓			
		Cystathionine gamma-lyase	EC 4.4.1.1	✓			
		Cysteine synthase	EC 2.5.1.47	✓			
		Homoserine O-acetyltransferase	EC 2.3.1.31	✓			
		L,L-diamipimelate amitransferase	EC 2.6.1.83				
		Lysine decarboxylase	EC 4.1.1.18		✓	✓	✓
		Lysine decarboxylase 2, constitutive	EC 4.1.1.18		✓	✓	
		Meso-diamipimelate D-dehydrogenase	EC 1.4.1.16				
		Methionine ABC transporter substrate-binding protein		✓			
		N-acetyl-L,L-diamipimelate deacetylase	EC 3.5.1.47	✓			
		N-succinyl-L,L-diamipimelate desuccinylase	EC 3.5.1.18				
		O-acetylhomoserine sulfhydrylase	EC 2.5.1.49	✓			
		O-succinylhomoserine sulfhydrylase	EC 2.5.1.48	✓			
		Pyruvate-flavodoxin oxidoreductase	EC 1.2.7.-	✓			
		S-ribosylhomocysteine lyase	EC 4.4.1.21	✓			
		Serine acetyltransferase	EC 2.3.1.30	✓			
		Sulfate permease		✓			
	Proline and 4-hydroxyproline	Gamma-glutamyl phosphate reductase	EC 1.2.1.41	✓			
		Glutamate 5-kinase	EC 2.7.2.11	✓			
		Pyrroline-5-carboxylate reductase	EC 1.5.1.2	✓			
		RNA-binding C-terminal domain PUA		✓			
		YggS, proline synthase co-transcribed bacterial homolog				✓	
Carbohydrates	Aminosugars	PTS system, N-acetylglucosamine-specific IIA component	EC 2.7.1.69			✓	
		PTS system, N-acetylglucosamine-specific IIB component	EC 2.7.1.69			✓	
		PTS system, N-acetylglucosamine-specific IIC component	EC 2.7.1.69			✓	
	Central carbohydrate metabolism	Acetaldehyde dehydrogenase	EC 1.2.1.10			✓	✓
		Acetyl-coenzyme A synthetase	EC 6.2.1.1				✓
		Glucokinase	EC 2.7.1.2	✓			
		Glycolate dehydrogenase GlcD	EC 1.1.99.14		✓	✓	
		NAD-dependent protein deacetylase of SIR2 family			✓		
		Phosphoglycerate mutase	EC 5.4.2.1			✓	✓
		Pyruvate:ferredoxin oxidoreductase, alpha subunit	EC 1.2.7.1			✓	✓
		Pyruvate:ferredoxin oxidoreductase, beta subunit	EC 1.2.7.1			✓	✓
		Pyruvate:ferredoxin oxidoreductase, delta subunit	EC 1.2.7.1			✓	✓
		Pyruvate:ferredoxin oxidoreductase, gamma subunit	EC 1.2.7.1			✓	✓
	Di- and oligosaccharides	PTS system, maltose and glucose-specific IIB component	EC 2.7.1.69			✓	
		PTS system, maltose and glucose-specific IIC component	EC 2.7.1.69			✓	
	Fermentation	L-lactate dehydrogenase	EC 1.1.1.27	✓			
	Monosaccharides	2-dehydro-3-deoxygluconate kinase	EC 2.7.1.45			✓	✓
		2-dehydro-3-deoxyphosphogluconate aldolase	EC 4.1.2.14			✓	✓
		Arabinose-proton symporter			✓		
		D-mannate oxidoreductase	EC 1.1.1.57			✓	✓
		D-mannonate oxidoreductase	EC 1.1.1.57			✓	
		Hexose phosphate uptake regulatory protein UhpC				✓	
		Hexuronate transporter				✓	✓
		L-arabinose isomerase	EC 5.3.1.4		✓		
		L-ribulose-5-phosphate 4-epimerase	EC 5.1.3.4		✓		
		Mannate dehydratase	EC 4.2.1.8			✓	✓
		Mannonate dehydratase	EC 4.2.1.8			✓	
		Manse-1-phosphate guanylyltransferase	EC 2.7.7.13				✓
		PTS system, fructose-specific IIA component	EC 2.7.1.69			✓	✓
		PTS system, fructose-specific IIB component	EC 2.7.1.69			✓	✓
		PTS system, fructose-specific IIC component	EC 2.7.1.69			✓	✓
		PTS system, manse-specific IIC component	EC 2.7.1.69			✓	✓
		PTS system, manse-specific IID component	EC 2.7.1.69			✓	
		Pyrimidine-nucleoside phosphorylase	EC 2.4.2.2	✓			
		Ribulokinase	EC 2.7.1.16		✓		
		Uronate isomerase	EC 5.3.1.12			✓	✓
	Polysaccharides	1,4-alpha-glucan branching enzyme, GH-13-type	EC 2.4.1.18	✓			
		Glycogen branching enzyme, GH-57-type, archaeal	EC 2.4.1.18	✓			
		Glycogen phosphorylase	EC 2.4.1.1	✓			
		Glycogen synthase, ADP-glucose transglucosylase	EC 2.4.1.21	✓			
	Peptidoglycan Biosynthesis	NAD(P)HX dehydratase					✓
		Soluble lytic murein transglycosylase precursor	EC 3.2.1.-			✓	✓
		UDP-N-acetylelpyruvoylglucosamine reductase	EC 1.1.1.158			✓	✓
		UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diamipimelate	EC 6.3.2.10			✓	✓

Cell Wall and Capsule	Capsular and extracellular polysacchrides	Alpha-L-Rha alpha-1,2-L-rhamsyltransferase	EC 2.4.1.-	✓			
		dTDP-4-dehydrorhamse 3,5-epimerase	EC 5.1.3.13	✓			
		dTDP-glucose 4,6-dehydratase	EC 4.2.1.46	✓			
		Glucose-1-phosphate cytidylyltransferase	EC 2.7.7.33	✓			
		Glucose-1-phosphate thymidylyltransferase	EC 2.7.7.24	✓			
	Gram-Negative cell wall components	ADP-heptose--lipooligosaccharide heptosyltransferase II	EC 2.4.1.-	✓			
		Beta-1,3-glucosyltransferase		✓			
		Beta-1,4-galactosyltransferase		✓			
		Phosphoheptose isomerase 1	EC 5.3.1.-	✓			
		UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	EC 3.5.1.108			✓	
Clustering-based subsystems	Aerotolerance	BatA (Bacteroides aerotolerance operon)			✓	✓	✓
		BatB			✓		✓
		BatC			✓	✓	
		BatD			✓	✓	✓
		BatE			✓	✓	✓
		hypothetical protein PA3071			✓	✓	✓
		MoxR-like ATPase in aerotolerance operon			✓	✓	✓
		TPR domain protein in aerotolerance operon					✓
	Bacterial RNA-metabolizing Zn-dependent hydrolases	Cell division protein FtsK				✓	✓
		Trk system potassium uptake protein TrkA				✓	✓
		YsaD/KaeT/QmY protein					✓
	Cell division-ribosomal stress proteins cluster	SSU ribosomal protein S1p		✓			
		Transcription accessory protein (S1 RNA-binding domain)		✓			
	Methyltransferase	Adenine-specific methyltransferase	EC 2.1.1.72		✓	✓	✓
		Protein-N(5)-glutamine methyltransferase PrmC				✓	✓
	Stationary phase repair cluster	Cell division protein FtsL				✓	
		FIG139438: lipoprotein B		✓			
	Miscellaneous	16S rRNA (guanine-N2)-methyltransferase	EC 2.1.1.171	✓			
		ATP phosphoribosyltransferase	EC 2.4.2.17			✓	✓
		Cytoplasmic axial filament protein CafA and Ribonuclease G	EC 3.1.4.-		✓	✓	✓
		FIG000325: clustered with transcription termination protein NusA				✓	✓
		FIG002344: Hydrolase (HAD superfamily)			✓	✓	✓
		FIG004453: protein YceG like		✓			
		FIG007959: peptidase, M16 family				✓	✓
		FIG116849: hypothetical protein					✓
		FIG131328: Predicted ATP-dependent endonuclease of the OLD family					✓
		FIG146085: oligoribonuclease A, Bacillus type					✓
		Magnesium and cobalt efflux protein CorC		✓			
		Membrane-bound lytic murein transglycosylase D precursor	EC 3.2.1.-	✓			
		Organic solvent tolerance protein precursor					✓
		Penicillin-binding protein 2 (PBP-2)		✓			
		Pyruvate,phosphate dikinase	EC 2.7.9.1	✓			
		Replicative DNA helicase	EC 3.6.1.-		✓	✓	
		Ribosomal RNA large subunit methyltransferase N	EC 2.1.1.-		✓	✓	✓
		Rod shape-determining protein MreC				✓	✓
		Serine/threonine protein kinase PrkC, regulator of stationary phase				✓	✓
		tRNA pseudouridine synthase B	EC 4.2.1.70				✓
	Cell Division	Cell division protein FtsQ		✓			
		GTPase and tRNA-U34 5-formylation enzyme TrmE				✓	✓
		Protein YidD			✓	✓	✓
		rRNA small subunit 7-methylguasine (m7G) methyltransferase GidB					✓
		tRNA uridine 5-carboxymethylamimethyl modification enzyme GidA				✓	✓
Cofactors, Vitamins, Prosthetic Groups, Pigments	Thiamin biosynthesis	Sulfur carrier protein adenylyltransferase ThiF				✓	✓
		Thiazole biosynthesis protein ThiG				✓	✓
	Biotin	8-amino-7-oxonaate synthase	EC 2.3.1.47			✓	✓
		Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	EC 2.6.1.62		✓	✓	
		Adesylmethionine-8-ami-7-oxonaate amitransferase	EC 2.6.1.62		✓	✓	✓
		Competence protein F homolog, phosphoribosyltransferase domain					✓
		Dethiobiotin synthetase	EC 6.3.3.3		✓	✓	✓
	Folate and pterines	Alcohol dehydrogenase	EC 1.1.1.1			✓	✓
		Hydroxyethylthiazole kinase	EC 2.7.1.50	✓			
		Thymidylate synthase thyX	EC 2.1.1.-			✓	✓
	Lipoic acid	Lipoate synthase				✓	✓
	NAD and NADP	ADP-ribose pyrophosphatase	EC 3.6.1.13	✓			
		C-terminal domain of CinA type S					✓
		Nicotinamide-nucleotide adenylyltransferase, NadM family	EC 2.7.7.1		✓		
	Riboflavin, FMN, FAD	8-amino-7-oxononanoate synthase	EC 2.3.1.47		✓	✓	
		Flavodoxin			✓		
		Molybdopterin binding motif, CinA N-terminal domain					✓
DNA Metabolism	Restriction-Modification System	Type I restriction-modification system, DNA-methyltransferase subunit M	EC 2.1.1.72		✓	✓	✓
		Type I restriction-modification system, restriction subunit R	EC 3.1.21.3		✓	✓	✓
		Type I restriction-modification system, specificity subunit S	EC 3.1.21.3		✓		✓
		Type III restriction-modification system DNA endonuclease	EC 3.1.21.5			✓	✓
		Type III restriction-modification system methylation subunit	EC 2.1.1.72		✓	✓	✓
	CRISPR	CRISPR-associated protein, Csn1 family				✓	✓
	DNA repair	ATP-dependent DNA helicase Rep					✓
		ATP-dependent DNA helicase UvrD/PcrA					✓
		DNA-cytosine methyltransferase	EC 2.1.1.37	✓			
		Endonuclease IV	EC 3.1.21.2	✓			
		Phosphoglycolate phosphatase	EC 3.1.3.18	✓			
		Regulatory protein RecX			✓	✓	✓
		SOS-response repressor and protease LexA	EC 3.4.21.88	✓			

		Uracil-DNA glycosylase, family 1		✓			
		Uracil-DNA glycosylase, family 4					✓
	DNA replication	ATP-dependent DNA helicase RecQ		✓			
		Replicative DNA helicase (DnaB)	EC 3.6.4.12	✓			
		Single-stranded-DNA-specific exonuclease RecJ	EC 3.1.-.-	✓			
	DNA uptake, competence	DNA topoisomerase III	EC 5.99.1.2			✓	
Dormancy and Sporulation	Sporulation	Protein of unkwn function identified by role in sporulation (SpoVG)			✓		
Fatty Acids, Lipids, and Isopreids	Fatty acids	Holo-[acyl-carrier protein] synthase	EC 2.7.8.7			✓	✓
	Isoprenoids	Dimethylallyltransferase	EC 2.5.1.1				✓
		Octaprenyl diphosphate synthase	EC 2.5.1.90			✓	✓
	Phospholipids	1-acyl-sn-glycerol-3-phosphate acyltransferase	EC 2.3.1.51			✓	✓
		Acyl-phosphate:glycerol-3-phosphate O-acyltransferase PlsY				✓	✓
		Phosphate:acyl-ACP acyltransferase PlsX				✓	✓
		Phosphatidylglycerophosphatase B	EC 3.1.3.27			✓	✓
Membrane Transport	Ton and Tol transport	Biopolymer transport protein ExbD/TolR		✓			
		MotA/TolQ/ExbB proton channel family protein		✓			
		Outer membrane lipoprotein omp16 precursor		✓			
	ABC transporters	Dipeptide transport system permease protein DppC	EC 3.A.1.5.2	✓			
		Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA	EC 3.A.1.5.1	✓			
		Oligopeptide transport system permease protein OppB	EC 3.A.1.5.1	✓			
		Oligopeptide transport ATP-binding protein OppD	EC 3.A.1.5.1	✓			
	Cation transporters	Copper-translocating P-type ATPase	EC 3.6.3.4	✓			
	Uni- Sym- and Antiporters	Sodium-dependent phosphate transporter		✓			
Nitrogen Metabolism	Ammonia assimilation	Glutamine synthetase type I	EC 6.3.1.2	✓			
		Glutamate synthase [NADPH] small chain	EC 1.4.1.13				✓
Nucleosides and Nucleotides	Detoxification	5-nucleotidase SurE	EC 3.1.3.5			✓	✓
	Purines	Adenine phosphoribosyltransferase	EC 2.4.2.7	✓			
		GMP synthase [glutamine-hydrolyzing]	EC 6.3.5.2		✓	✓	✓
		GMP synthase [glutamine-hydrolyzing], amidotransferase subunit	EC 6.3.5.2	✓			
		GMP synthase [glutamine-hydrolyzing], ATP pyrophosphatase subunit	EC 6.3.5.2	✓			
		Guanine deaminase	EC 3.5.4.3	✓			
		Phosphoribosylamiimidazole carboxylase catalytic subunit	EC 4.1.1.21		✓	✓	✓
		Phosphoribosylformylglycinamidine synthase, PurS subunit	EC 6.3.5.3	✓			
		Xanthine/uracil/thiamine/ascorbate permease family protein		✓			
	Pyrimidines	Cytidine deaminase	EC 3.5.4.5	✓			
		Cytidylate kinase	EC 2.7.4.25	✓			
		Dihydroorotate dehydrogenase electron transfer subunit	EC 1.3.3.1		✓	✓	✓
		Dihydroorotate dehydrogenase, catalytic subunit	EC 1.3.3.1			✓	✓
		Uracil phosphoribosyltransferase	EC 2.4.2.9			✓	✓
		Nucleotidase	EC 3.1.3.5			✓	
Phages, Prophages, Transposable elements, Plasmids	Phages, Prophages	Phage tail fiber protein			✓		
Phosphorus Metabolism	Phosphate metabolism	Alkaline phosphatase	EC 3.1.3.1	✓			
		Phosphate regulon sensor protein PhoR (SphS)	EC 2.7.13.3	✓			
		Phosphate regulon transcriptional regulatory protein PhoB (SphR)		✓			
		Predicted ATPase related to phosphate starvation-inducible protein PhoH		✓			
		Phosphate transport regulator (distant homolog of PhoU)				✓	
		Phosphoenolpyruvate phosphomutase	EC 5.4.2.9		✓		
		Phosphonopyruvate decarboxylase	EC 4.1.1.82		✓		
Potassium metabolism	Potassium homeostasis	Large-conductance mechasensitive channel		✓			
		Potassium efflux system KefA protein		✓			
		putative Glutathione-regulated potassium-efflux system protein KefB			✓	✓	✓
Protein Metabolism	Protein biosynthesis	LSU ribosomal protein L33p, zinc-dependent		✓			
		programmed frameshift-containing		✓			
		tRNA-Pro-CGG		✓			
		SSU ribosomal protein S18p, zinc-independent			✓		
	Protein degradation	ATP-dependent Clp protease ATP-binding subunit ClpX			✓	✓	✓
		ATP-dependent Clp protease proteolytic subunit	EC 3.4.21.92		✓	✓	✓
	Protein folding	Peptidyl-prolyl cis-trans isomerase PpiD	EC 5.2.1.8				✓
Regulation and Cell signaling	Stringent Response, (p)ppGpp metabolism	GTP pyrophosphokinase, (p)ppGpp synthetase II	EC 2.7.6.5	✓			
	Stringent Response, (p)ppGpp metabolism	Guasine-bis(diphosphate)-pyrophosphohydrolase	EC 3.1.7.2	✓			
	HPr catabolite repression system	Catabolite repression HPr-like protein Crh				✓	✓
	Programmed Cell Death and Toxin-antitoxin Systems	HigA protein (antitoxin to HigB)				✓	
Respiration	ATP synthases	ATP synthase F0 sector subunit b	EC 3.6.3.14		✓	✓	✓
		ATP synthase delta chain	EC 3.6.3.14			✓	
		ATP synthase epsilon chain	EC 3.6.3.14		✓	✓	
	Electron donating reactions	[NiFe] hydrogenase metallocenter assembly protein HypD		✓			
		[NiFe] hydrogenase metallocenter assembly protein HypE		✓			
		[NiFe] hydrogenase metallocenter assembly protein HypF		✓			
		Periplasmic [Fe] hydrogenase large subunit	EC 1.12.7.2			✓	
	Biogenesis of c-type cytochromes	Cytochrome c-type biogenesis protein CcdA (DsbD analog)			✓		
RNA Metabolism	RNA processing and modification	23S rRNA (guasine-O-) -methyltransferase rlmB	EC 2.1.1.-		✓	✓	✓
		FIG146085: 3-to-5 oligoribonuclease A, Bacillus type				✓	
		Peptidyl-prolyl cis-trans isomerase	EC 5.2.1.8	✓			
		Ribosomal large subunit pseudouridine synthase C	EC 4.2.1.70			✓	✓
		Ribosomal large subunit pseudouridine synthase D	EC 4.2.1.70			✓	✓
		Ribosomal protein S12p Asp88 (E. coli) methylthiotransferase				✓	✓
		RNA-PO4:RNA-OH ligase		✓			
		rRNA small subunit methyltransferase H				✓	✓
		tRNA (Guanine37-N1) -methyltransferase	EC 2.1.1.31		✓	✓	✓
		tRNA nucleotidyltransferase	EC 2.7.7.21			✓	✓
		tRNA-guanine transglycosylase	EC 2.4.2.29	✓			
		tRNA-i(6)A37 methylthiotransferase			✓	✓	✓

		tRNA-specific 2-thiouridylase MnmA			✓	✓	✓
		tRNA-specific adesine-34 deaminase	EC 3.5.4.-			✓	✓
		tRNA-t(6)A37 methylthiotransferase			✓	✓	✓
	Transcription	ATP-dependent DNA helicase RecG	EC 3.6.1.-		✓	✓	
Stress Response	Cold shock	Cold shock protein CspA				✓	✓
	Carbon Starvation	Carbon starvation protein A			✓		
		Starvation sensing protein RspB			✓		
	Osmotic stress	Aquaporin Z		✓			
Virulence, Disease and Defense	Bacteriocins	Inner membrane protein CreD-like protein		✓			
	Resistance to antibiotics and toxic compounds	Beta-lactamase	EC 3.5.2.6	✓			
		Cadmium efflux system accessory protein		✓			
		Cation efflux system protein CusA		✓			
		Cobalt-zinc-cadmium resistance protein		✓			
		Cobalt-zinc-cadmium resistance protein CzcA		✓			
		Macrolide-specific efflux protein MacA		✓			
		Macrolide export ATP-binding/permease protein MacB	EC 3.6.3.-		✓	✓	✓

Table S1. Gene expression values for carbon utilization pathways.

	Expression Values (Log2)			
Putative Product	PV1	PV7	TA21	TA26
Glycolysis				
ManX	8.25	7.84	7.80	8.78
Pgi	7.44	9.15	8.47	9.48
Pfk	7.62	0.00	6.65	10.17
FbaA	6.45	0.00	6.74	7.80
GapA	11.97	8.44	13.44	10.13
Pgk	7.65	8.07	7.92	8.65
Gpm	6.63	0.00	7.23	10.10
Eno	8.03	8.07	9.04	9.50
Pyk	7.29	0.00	6.40	6.89
Glucuronate Catabolism				
UxaC	0.00	0.00	12.25	13.70
UxuB	8.41	0.00	11.75	10.41
UxuA	6.86	7.77	9.31	9.85
KdgK	7.77	8.83	11.33	11.81
Eda	9.74	9.50	9.27	8.24
Xylulose Catabolism				
TktA	7.97	12.26	8.41	9.80
TktB	9.86	0.00	5.96	7.29
AraD	11.05	0.00	0.00	5.30
AraB	8.20	8.49	0.00	5.06
AraA	9.12	6.28	1.73	2.77
Rpe	7.23	8.48	6.27	8.90
AraB	8.20	8.49	0.00	5.06
RpiB	7.94	9.96	7.96	5.85
Prs	9.14	9.42	10.42	11.62
Cellulose Catabolism				
BglX	8.15	9.11	6.29	10.85
BcsZ	3.87	7.80	8.14	8.45
YcjU	7.26	0.00	7.86	7.07
Pgi	7.44	9.15	8.47	9.48

Table S2. Gene expression values for Acetate, Lactate, Malate, and Fumarate.

	Expression Values (Log2)			
Putative Product	PV1	PV7	TA21	TA26
Acetate				
PforA	7.42	0.00	8.20	10.59
PforB	7.15	12.02	7.72	8.19
PforD	10.07	0.00	8.33	6.68
PforG	7.84	11.71	9.69	7.83
Acs	8.11	10.78	6.49	7.87
AccA	8.58	8.84	5.99	9.36
Lactate				
LldD	7.58	6.95	0.00	4.85
Malate				
MaeB	7.79	9.75	7.11	8.74
Fumarate				
PyrB	8.10	10.76	9.52	11.69
ArgG	7.85	10.28	8.45	9.16
ArgH	8.37	7.96	2.24	7.48
PurB	10.48	10.48	7.46	10.33

Table S3. Gene expression values for Aspartate, Threonine, Ammonia, Glycine, Glutamine, Arginine, and Proline biosynthesis pathways.

	Expression Values (Log2)			
Putative Product	PV1	PV7	TA21	TA26
Aspartate				
PckA	8.78	10.29	10.77	9.59
AspC	9.27	9.94	8.13	9.16
Threonine				
ThrA1	7.91	8.88	5.54	9.33
Asd	8.59	6.80	5.19	6.48
ThrA2	0.00	0.00	6.79	9.04
ThrB	0.00	10.05	7.02	10.81
ThrC	5.18	9.90	10.55	8.64
Ammonia				
NirB	0.00	0.00	9.91	9.62
NifU	9.47	12.00	6.72	5.78
Glycine				
GcvT	0.00	0.00	8.16	9.06
Glutamine				
GlnA	15.22	8.79	8.80	13.74
Arginine				
GltB	9.13	9.27	9.77	9.96
ArgA	7.05	9.44	8.88	8.97
ArgB	6.90	8.86	6.07	6.76
ArgC	6.03	7.83	10.02	8.41
ArgD	6.15	0.00	5.36	8.64
ArgF	3.74	0.00	8.31	6.02
ArgH	9.30	11.51	6.51	7.84
ArgG	7.85	10.28	8.45	9.16
Proline				
ProB	6.36	0.00	1.96	0.00
ProA	7.29	9.33	0.00	0.00
ProC	7.75	8.75	6.87	7.34

Table S4. Gene expression values for Isoleucine, Valine, Leucine, Phenylalanine, Tryosine, and Tryptophan biosynthesis pathways.

	Expression Values (Log2)			
Putative Product	PV1	PV7	TA21	TA26
Isoleucine/Valine/Leucine				
IlvN	7.67	0.00	7.66	6.98
IlvB	11.56	9.24	7.49	9.41
IlvC	9.03	0.00	8.93	10.24
IlvD	10.77	10.53	6.62	9.24
IlvE	7.47	8.06	9.49	10.32
IleS	7.11	8.19	6.84	8.28
IlvE	7.47	8.06	9.49	10.32
ValS	5.27	8.58	4.58	6.45
LeuA	8.34	8.83	8.71	8.13
LeuD	6.44	7.87	6.34	5.99
LeuB	6.96	8.77	7.97	9.31
IlvE	7.47	8.06	9.49	10.32
LeuS	7.80	8.90	6.96	7.25
Phenylalanine/Tyrosine/Tryptophan				
AroF	0.00	0.00	8.28	6.61
AroB	6.06	8.33	4.19	5.36
AroD	8.96	11.03	0.00	5.91
AroE	6.59	8.28	8.93	8.11
AroK	6.41	8.74	6.87	7.15
AroA	8.66	0.00	9.94	7.88
AroC	10.29	10.65	10.45	8.03
PheA	6.65	7.78	6.47	5.71
TyrB	0.00	0.00	10.14	5.79
PheA	6.65	7.78	6.47	5.71
TyrA	5.87	0.00	7.24	7.40
TyrB	0.00	0.00	10.14	5.79
TrpE	6.68	0.00	6.50	6.16
TrpGD	6.32	8.84	7.22	8.81
TrpF	7.90	7.57	6.41	7.63
TrpC	6.82	0.00	6.05	6.97
TrpB	7.06	9.79	10.13	8.88

Table S5. Gene expression values for the Bacteroides Aerotolerance Operon and Hexuronate catabolism.

Gene	PV1	PV7	TA21	TA26	Avg. Fold Change
	Expression Values (Log_2)				
Bacteroides Aerotolerance					
batA	7.68	7.86	6.28	4.20	1.48
batB	10.87	9.87	-	-	NA
batC	6.25	7.89	4.01	6.18	1.39
batD	8.91	8.03	4.17	5.02	1.84
batE	6.71	9.29	3.49	13.50	0.94
Hexuronate Catabolism					
exuT	-	-	11.42	12.85	NA
uxaC	-	-	12.25	13.70	NA
uxuB	8.41	-	11.74	10.40	0.76
uxuA	6.85	7.76	9.31	9.84	0.76
kdgK	7.77	8.82	11.32	11.80	0.72
kdgA	9.73	9.49	9.27	8.23	1.10

Chapter Five

Conclusions and Future Directions

While animal-microbe symbiosis has become a large topic for research, the inability to culture most bacteria is hindering. Without being able to culture bacteria, researchers are unable to use genetic tools to study the interactions between the microbes and the host. The advances in high-throughput sequencing has enabled researchers to affordably study previously uncharted symbiotic territory. Prior to the start of this project, many research groups have utilized high-throughput 16S rRNA gene sequencing to discover the composition of the bacterial communities in many environments, including symbiotic systems (Clawson et al., 2004; Lazarevic et al., 2009). Some researchers have also started to utilize high-throughput 16S rRNA gene sequencing to study the prokaryotic communities in the termite hindgut (Fisher et al., 2007; Warnecke & Hugenholtz, 2007). The termite is an important organism to study because it has a complex community residing in the gut of protists, archaea, and bacteria, all of which aid in the digestion of wood, provide nutrients to the termite, and emit substrates of great interest to biofuel production (Brune, 2014; Zuroff & Curtis, 2012). The goal of my doctoral research was to study the symbiotic community of the termite, *R. flavipes*, to provide deeper insight into the roles of the symbionts.

Prior to this research, a core microbiota for a termite species was not discussed as all termite symbiotic studies have pooled termite hindguts to create libraries with enough DNA yield for sequencing. I was able to isolate enough DNA from single hindguts for sequencing along with the ability to sequence many samples together, which allowed us to study the microbiota across multiple colonies with replicates and determine a core. It was found that the microbiota within a colony was homogenous, which has previously been linked to a hypothesis

regarding nest-mate recognition, which states that the lower abundant microbes in the hindgut may produce substrates that allow a colony to tell apart nest-mates from intruders (Matsuura, 2001; Minkley et al., 2006). The core microbiota was determined within the hindgut of over 50 termites from multiple colonies in multiple areas and consisted of 69 OTUs in 17 taxa, all existing in similar relative abundances across all termites, suggesting that the differences between the microbiota of the termites was caused by the non-core organisms. It was previously shown that the microbes within termites belonging to different castes differ and thought to be due to the dietary differences (Cleveland, 1925), which has also been shown in honeybees where the microbiome clusters according to caste, i.e. males, foragers, nurses, and queens (Kapheim et al., 2015). The dominant bacteria of *R. flavipes*, *Treponema* and *Endomicrobia* (both protist associated), drop in abundance in the alate caste, possibly due to the individuals in this caste not actively feeding on wood. Quantitative PCR (qPCR) of the protist 18S rRNA gene was used to determine the relative abundance of the two groups of protists (Parabasalia and Oxymonadida) in the same hindgut samples. The lower abundance of both protist groups in the alate classes correlates with the decrease of protist-symbiotic bacteria. As the *Treponema* in the gut can be free-living and protist-associated, I tried to determine if there were separate *Treponema* OTUs present in the different caste samples. Four *Treponema* OTUs were shown to be present at lower abundances in the alate caste, suggesting their association with protists. Two OTUs were shown to have similar relative abundances across four castes, suggesting that they belong to the free-living population and were not affected by the lowering of protist abundance. This experiment resulted in the ability to calculate the core microbiota in the hindgut of *R. flavipes*, provided further evidence for colony homogeneity, and showed microbial reactions to termite dietary patterns.

Knowing that the microbiota of non-feeding castes changes, we wanted to determine whether the microbiota may also change when termites are fed different diets. In chapter three of this dissertation, I split a single termite colony into seven groups by diet, all differing in lignocellulose, hemicellulose, cellulose, and other chemical contents. During this experiment, other groups published results on the effect of diet on the microbiota, showing that the diet of multiple species of higher termites is the determining factor in hierarchical clustering of the microbiomes (Mikaelyan et al., 2015), and the microbiota of termites fed pine, sorghum, corn, and grass cluster differently on a principal coordinate plot according to diet (Huang et al., 2013b). As both of these groups show diet as a determining factor of the microbiota, they use pools of hindguts and limit the study to a single endpoint. The ability to sample single hindguts allowed us to view temporal changes in the microbiota throughout a period of 56 days. The core microbiota was shown to remain consistent over time regardless of diet (except the starved group which eventually died off), but each colony grouped according to diet on a principal coordinate analysis plot, suggesting the differences in the microbiota among diets are caused by the non-core organisms. Due to the robustness of the hindgut microbiota, of OTU richness and homogeneity, we were able to train an artificial neural network (ANN) on the system. The ANN was trained on each sample of the experiment, with one sample per diet left out for validation. After training, the ANN was used to try to predict the microbiota of a termite fed each diet on a particular day, and the results were compared to relative abundance values obtained from 16S rRNA gene sequencing. The ANN was able to predict the overall microbiota with a root means squared error (RMSE) value of 0.0153 and the majority of bacterial abundances (order level) within less than 1% difference. The ANN was also used to determine the number of correlations each taxonomic order had between the other orders in the hindgut by changing the abundance of

each taxon by ± 5%. Four core taxa had more than seven correlations with other taxa, meaning their rise or fall in abundance had an effect on the bacteria correlated with them. Nine taxa not belonging to the core had over seven correlations each, suggesting that the members outside of the core have a larger effect on the overall population. The low-abundant, non-core taxa had the largest number of correlations, suggesting these taxa may be drivers of the bacterial community.

The final aim of this dissertation was to study the simultaneously acquired metagenomes and metatranscriptomes of single protist cells living in the termite hindgut to shed light on the physiologies of the protists and associated bacteria in vivo. Prior to this research, a metatranscriptome of the entire hindgut, symbiotic contents and termite tissue, was performed and was able to show genes for lignocellulose degradation from the termite library and the symbiont library (Tartar et al., 2009). While this study is powerful and one of the first of its kind, it left unanswered questions as to which symbionts were performing these functions. There was also two studies which isolated the endosymbiotic bacteria from the supposed same species within a single protist, performed multiple displacement amplification of the cells, and sequenced the genomes (Hongoh et al., 2008a; Hongoh et al., 2008b). For this experiment, five protist species were chosen based on three criteria: the association with different endo- and ectosymbionts, different niches in the hindgut, and different taxonomic groups (Parabasalids and Oxymonads). The protists chosen were *Trichonympha agilis*, *Pyrsonympha vertens*, *Dinenympha gracilis*, *Dinenympha fimbriata*, and *Dinenympha* species II. Draft genomes of the endosymbiotic bacteria from *T. agilis* and *P. vertens* (*Endomicrobium* sp. TA21 and PV7) were assembled and compared to eachother and to the free-living *Endomicrobium proavitum* (genome previously published, (Zheng & Brune, 2015)). The metatranscriptomes of four protists (two from each species) were mapped to the respective genome and the expression data were

compared, showing four carbon utilization pathways being expressed in both species. Fifteen amino acid biosynthesis pathways were found and shown to be active with varying expression. One metabolic pathway shown to be different between the two species was the hexuronate catabolism pathway, breaking down plant pectins to pyruvate. *Endomicrobium* sp. PV7 is missing the first two genes, *exuT* for hexuronate transport into the cell, and *uxaC*, a uronate isomerase to begin catabolism. *Endomicrobium* sp. TA21 maintains these genes and has higher expression of all genes in the pathway compared to *Endomicrobium* sp. PV7, suggesting the ability of *Endomicrobium* sp. TA21 to utilize plant pectins as a carbon source. *Endomicrobium* sp. PV7 maintains the full Bacteroides aerotolerance operon, *batI,* while *Endomicrobium* sp. TA21 is missing the *batB* gene. Along with the missing *batB* gene, *Endomicrobium* sp. TA21 has close to two-fold lower average expression of the remaining genes compared to *Endomicrobium* sp. PV7. The presence of the entire *batI* operon and higher expression in *Endomicrobium* sp. PV7 may relate to the microaerophilic niche of the host protist.

There are many avenues for future work provided by this research. The data presented in chapter four is a small portion of the data generated from this dissertation. The genomes for *Endomicrobium* sp. PV7 and TA21 are of high quality and over 95% complete, but require more complete annotation, which was found when mapping the transcriptomes to the genomes. The protist, *T. agilis*, also has ectosymbiotic *Treponema* participating in the symbiosis and it has been shown through electron microscopy by Michael Stephens that there are multiple morphologically distinct *Treponema* attached to *T. agilis*, which creates problems with metagenome assembly. An immediate future direction is to test a previously performed metagenome assembly pipeline in attempt to pull out draft genomes of these *Treponema* (Maltz et al., 2014). Recently, Oxford Nanopore Technologies has put forth a protocol for whole genome amplification and sequencing

of small concentrations of DNA, which would allow long-read sequencing for these metagenomes and may improve assemblies of the ectosymbiotic genomes. I also created libraries enriched for protist transcripts by amplifying with oligo-dT primers, which will provide insight on the protist metabolism. I hypothesize that some of the genes missing or not expressed from the *Endomicrobium* symbionts may be expressed by the *Treponema* or even the protist itself, leading to metabolic complementarity which has been found in other endosymbiotic systems such as the pea aphid and its endosymbiont, *Buchnera* (Wilson et al., 2010). Beyond the two protists in this thesis, there are metagenomes and metatranscriptomes from three other protists species and their associated bacteria that require analysis.

The research performed in this dissertation provides a deeper understanding of the microbial community contributing to the betterment of the termite host. The data presented in chapter two is the first determination of a core microbiota in the termite hindgut. Chapter three presents the first temporal study of the termite hindgut microbiota in regards to diet with ANN prediction models showing low abundant bacteria as drivers of the community. The data presented in chapter four is the first simultaneous metagenome and metatranscriptome study of single protists and provides insight on the metabolic processes being performed by the protists and bacteria, a small-scale symbiosis within the larger, overall termite symbiosis. This dissertation also provides the groundwork for many studies of close-knit symbioses and I hope that others will use these techniques for other symbiotic systems.

Appendix I: Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys[+*]

Author Contributions

This collaborative manuscript was conceptually designed by Dr. Michael C. Nelson, Dr. Hilary G. Morrison (Marine Biological Laboratories, Woods Hole, MA.), and Dr. Joerg Graf. Experimental samples were collected by Dr. Michael C. Nelson and myself. I participated in the DNA extraction and isolation from the collected samples. I prepared the 16S rRNA amplicon libraries for the V4 and V4-V5 regions of each sample and sequenced them on an Illumina MiSeq. Sharon L. Grim (Marine Biological Laboratories, Woods Hole, MA.) performed the V4-V5 library preparation and sequencing for the samples on a Roche 454 pyrosequencer. Data analysis for all sequenced samples was performed by Dr. Michael C. Nelson and myself. Dr. Michael C. Nelson wrote the majority of the manuscript, with aid from Dr. Hilary G. Morrison, myself, and Dr. Joerg Graf.

[+] Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J (2014) Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys. PLoS ONE 9(4): e94249. doi: 10.1371/journal.pone.0094249

OPEN ACCESS Freely available online

PLOS ONE

Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys

Michael C. Nelson[1], Hilary G. Morrison[2], Jacquelynn Benjamino[1], Sharon L. Grim[2], Joerg Graf[1]*

1 Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut, United States of America, **2** Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, Massachusetts, United States of America

Abstract

The exploration of microbial communities by sequencing 16S rRNA genes has expanded with low-cost, high-throughput sequencing instruments. Illumina-based 16S rRNA gene sequencing has recently gained popularity over 454 pyrosequencing due to its lower costs, higher accuracy and greater throughput. Although recent reports suggest that Illumina and 454 pyrosequencing provide similar beta diversity measures, it remains to be demonstrated that pre-existing 454 pyrosequencing workflows can transfer directly from 454 to Illumina MiSeq sequencing by simply changing the sequencing adapters of the primers. In this study, we modified 454 pyrosequencing primers targeting the V4-V5 hypervariable regions of the 16S rRNA gene to be compatible with Illumina sequencers. Microbial communities from cows, humans, leeches, mice, sewage, and termites and a mock community were analyzed by 454 and MiSeq sequencing of the V4-V5 region and MiSeq sequencing of the V4 region. Our analysis revealed that reference-based OTU clustering alone introduced biases compared to *de novo* clustering, preventing certain taxa from being observed in some samples. Based on this we devised and recommend an analysis pipeline that includes read merging, contaminant filtering, and reference-based clustering followed by *de novo* OTU clustering, which produces diversity measures consistent with *de novo* OTU clustering analysis. Low levels of dataset contamination with Illumina sequencing were discovered that could affect analyses that require highly sensitive approaches. While moving to Illumina-based sequencing platforms promises to provide deeper insights into the breadth and function of microbial diversity, our results show that care must be taken to ensure that sequencing and processing artifacts do not obscure true microbial diversity.

Citation: Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J (2014) Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys. PLoS ONE 9(4): e94249. doi:10.1371/journal.pone.0094249

Editor: Markus M. Heimesaat, Charité, Campus Benjamin Franklin, Germany

Received December 4, 2013; **Accepted** March 12, 2014; **Published** April 10, 2014

Funding: This work was partially funded by an Emerging Frontiers in Research and Innovation – Multicellular and Inter-kingdom Signaling (EFRI-MIKS) grant awarded by the US National Science Foundation to Joerg Graf and NIH RO1 GM095390 to JG and HGM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors declare that no competing interests exist.

* E-mail: joerg.graf@uconn.edu

Introduction

The field of microbial ecology relies on knowledge about the structure and composition of microbial communities as a foundation for understanding their role and function. Culture-independent analyses, which allow the identification of species that are recalcitrant to cultivation, continue to have a large impact on our understanding of microbial communities since the first studies of 5S rRNA sequences by Stahl et al. in the mid 1980s [1,2]. While many consider full length sequences generated by Sanger sequencing of 16S rRNA clone libraries to be the gold standard for phylogenetic analysis, even the largest studies typically analyzed no more than a few hundred to a thousand sequences for each sample due to the costly and labor intensive process this method entails [3–5]. In the early 2000s, the development and commercial availability of high-throughput sequencing platforms capable of producing hundreds of thousands to millions of sequences per run at a significantly lower cost than Sanger sequencing led to a revolution in the field of microbial ecology.

Microbial ecologists quickly adopted high-throughput pyrosequencing instruments produced by Roche 454 Life Sciences for sequencing 16S rRNA genes, which led to the discovery of what has been termed the "rare biosphere" and provided a deeper and more thorough view of the composition of a vast number of microbial communities from a wide range of habitats [6–10]. Since its introduction, most investigators have preferred 454 pyrosequencing for microbial diversity projects due to the longer read lengths that the 454 pyrosequencing platform provided relative to competing sequencing instruments from Illumina and others. While capable of producing longer reads lengths than competing technologies, 454 pyrosequencing produces datasets that exhibit characteristic errors associated with insertions/deletions (indels) in stretches of identical nucleotides (homopolymers) [11]. These systematic errors must be removed or corrected using time consuming and computationally intensive software packages prior to further analysis [12–14].

Compared to 454 pyrosequencing, the Illumina sequencing-by-synthesis (SBS) methodology has a lower per-base error rate and is not as susceptible to indel errors in homopolymer stretches [15,16]. The significantly higher sequence quality of Illumina generated sequences, combined with a much lower cost per sequence compared to 454 pyrosequencing, has spurred a number of researchers to develop strategies to sequence 16S rRNA gene amplicons using Illumina systems [17–22]. Although initial studies suggested that Illumina-based 16S sequencing produced data of lower quality than 454 pyrosequencing [19], adjustments to the

library preparation and sequencing protocols have produced datasets with significantly higher quality than 454 pyrosequencing [18,22]. While Illumina instruments historically generated short sequences of 30–100 bp, increases in the maximum read length on the Illumina MiSeq platform [2×300 bp paired end sequencing as of this writing) allow the sequencing of amplicons of similar length to those traditionally used in 454 pyrosequencing studies. Additionally, the length and quality of Illumina sequenced amplicons can be increased by aligning and combining each set of paired end reads into a single contig, a process generally referred to as read merging. This allows researchers using the Illumina MiSeq to produce merged sequences with an average length similar to those generated by 454 pyrosequencing but of significantly higher quality and at a lower cost per sequence [17,22,23].

While some previous studies have compared the results of 454 pyrosequencing and Illumina sequencing for both metagenomic libraries and 16S amplicons [24–26], these studies mostly focused on comparing beta diversity measures to see if the two sequencing technologies produced similar comparisons between different samples. As such, finer details concerning whether Illumina-based 16S sequencing can serve as a replacement for users currently using 454 pyrosequencing have yet to be fully explored. In this study we generated amplicon libraries of the V4-V5 hyper-variable regions of the 16S rRNA gene for 6 natural microbial communities and a synthetic mock community using the same 16S rRNA gene template primers, which were sequenced using either a 454 GS FLX or Illumina MiSeq. Additionally, libraries for the V4 hyper-variable region alone were generated and sequenced on the MiSeq using the protocol described by Caporaso et al. [18] We examined multiple combinations of data processing methods involving OTU clustering and chimera detection to identify a combination that provides both processing efficiency and accuracy. Using this processing method we analyzed the resulting datasets and compared the results of alpha and beta diversity analyses to evaluate whether the choice of sequencing platform led to significant differences that could bias the interpretation of the results.

Materials and Methods

Sample descriptions

We chose five samples, representing diverse host-associated microbial communities, for analysis: stool from an adult human (sample Human), contents of the intestinum of the medicinal leech *Hirudo verbana* (sample Leech), contents from the small intestine of a healthy mouse (sample Mouse), the non-adherent microbial fraction obtained from rumen contents of a dairy cow (sample Rumen) and the hindgut contents from the eastern subterranean termite *Reticulitermes flavipes* (sample Termite). Mixed liquor from the municipal waste water treatment facility located on the University of Connecticut, Storrs campus (sample Sewage) was included as a complex, high-diversity environmental microbial community. We also included a synthetic mock community (sample Mock) which was developed by the Human Microbiome Project (HMP) and includes the following 20 bacterial species in equal concentration according to ribosomal copy number: *Acinetobacter baumannii* str. 5377, *Actinomyces odontolyticus* str. 1A.21, *Bacillus cereus* str. NRS 248, *Bacteroides vulgatus* str. NCTC 11154, *Clostridium beijerinckii* str. NCIMB 8052, *Deinococcus radiodurans* str. R1 (smooth), *Enterococcus faecalis* str. OG1RF, *Escherichia coli* str. K12 substr. MG1655, *Helicobacter pylori* str. 26695, *Lactobacillus gasseri* str. 63 AM, *Listeria monocytogenes* str. EGDe, *Neisseria meningitidis* str. MC58, *Propionibacterium acnes* str. KPA171202, *Pseudomonas aeruginosa* str. PAO1-LAC, *Rhodobacter sphaeroides* str. ATH 2.4.1, *Staphylococcus aureus* TCH1516, *Staphylococcus epidermidis* FDA str. PCI 1200, *Streptococcus agalactiae* str. 2603 V/R, *Streptococcus mutans* str. UA159, and *Streptococcus pneumoniae* str. TIGR4. For the 454 library, an earlier version of the HMP mock community was used that comprised the same 20 species plus *Porphyromonas gingivalis* str. 2561. The RBB+C protocol described by Yu and Morrison [27] was used to isolate high quality genomic DNA from all samples except the human stool and mock community. The mock community DNA was obtained from BEI Resources (catalog number HM-276D, Genomic DNA from Microbial Mock Community B, even concentration).

Vincent Young (University of Michigan) generously provided the human stool DNA from an anonymous female donor. The University of Connecticut (UConn) IRB committee determined that our research did not require IRB approval for our use of this sample as it was previously collected under an IRB approved protocol and the donor gave consent for its use in subsequent studies such as ours. The rumen and mouse samples were collected as part of IACUC approved studies being conducted at the University of Connecticut that are not a part of this current study. The UConn IACUC committee determined that this study did not require separate approval for the use of these samples as they were collected under approved protocols as part of ongoing research programs and not at the specific request of the authors. Leeches were purchased from Leeches USA, an approved supplier of medicinal leeches and termites were purchased from CT Valley Biological Supply. No specific permits or permissions were required for the acquisition of the sewage sample.

Library preparation

We used primers previously designed to amplify the V4-V5 hyper-variable regions of the 16S rRNA gene to generate the 454 and Illumina libraries using fusion primer designs appropriate for the respective sequencing platforms (Table 1) [28]. The 16S template binding sequence was identical between both sets of fusion primers, with the 454 fusion primers following the standard format used by the Marine Biological Laboratory (MBL) and the Illumina fusion primers using the format described by Bartram et al. [17]. Libraries for all seven samples were prepared and sequenced by 454 pyrosequencing at the MBL's Josephine Bay Paul Center according to their standard protocols on a GS FLX using Titanium sequencing chemistry [28].

The Illumina sequencing libraries were all prepared and sequenced at the University of Connecticut. We prepared two sets of V4-V5 Illumina libraries for the six natural community samples at two separate times. The first set of libraries was prepared following the same protocol used for the 454 pyrosequencing libraries, with the PCR product for each sample gel purified prior to pooling and sequencing. The PCR products for the second set of V4-V5 Illumina libraries and the mock community libraries were purified using a 0.6X PCR volume of AMPure XP magnetic beads following the manufacturer's instructions. Additionally, we prepared libraries for the V4 hyper-variable region according to the protocol described by Caporaso et al. [18]. The Illumina libraries were sequenced on separate runs of a MiSeq using a 2×250 bp paired end protocol.

Sequence pre-processing

The V4-V5 454 pyrosequencing datasets were pre-processed prior to QIIME analysis in accordance with the in-house processing pipeline used by the MBL for 454 pyrosequencing analysis. Sequences had to possess the full index and forward primer sequence with no errors present in either, have zero

Table 1. Library construction primer sequences.

Sequencing Instrument	16S Variable Region(s)	Name	Primer Sequence 5′-3′[A]	Length
Roche 454	V4-V5	454-518F	GCCTCCCTCGCGCCATCAGXXXXX**CCAGCAGCYGCGGTA**AN	41
GS FLX		454-926R-1	GCCTTGCCAGCCCGCTCAG**CCGTCAATTCNTTTRAGT**	37
		454-926R-3	GCCTTGCCAGCCCGCTCAG**CCGTCAATTTCTTTGAGT**	37
		454-926R-4	GCCTTGCCAGCCCGCTCAG**CCGTCTATTCCTTTGANT**	37
Illumina		Iv4v5-518F	CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT **CCAGCAGCYGCGG**TAAN	81
MiSeq		Iv4v5-926R-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT*NNNN* **CCGTCAATTCNTTTRAGT**	80
		Iv4v5-926R-3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT*NNNN* **CCGTCAATTTCTTTGAGT**	80
		Iv4v5-926R-4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT*NNNN* **CCGTCTATTCCTTTGANT**	80
	V4	Iv4-515f	AATGATACGGCGACCACCGAGATCTACACTATGGTAATTGT**GTGCCAGCMGCCGCGGTAA**	60
		Iv4-806r	CAAGCAGAAGACGGCATACGAGATXXXXXXXXXXXXAGTCAGTCAGCC**GGACTACHVGGG TWTCTAAT**	68

[A]For all primers sets, the 16S template specific sequences are given in bold. For the 454 primers, the Xs in the forward represent the 5 bp run-key defined by the MBL with the underlined portion representing the 454 Lib A (forward primer) or Lib B (reverse primers) adapter sequence. Underlined portions of the Illumina primers represent the full TruSeq adapter sequence (V4-V5 primers) or a truncated version (V4). The N-bases in italics for the V4-V5 primers represent the 4 base ambiguous mix in between the TruSeq adapter sequence and the 16S template sequence. The Xs in the V4-V5 forward primer represent the sequence of one of the 6 bp TruSeq indices defined by Illumina while in the V4 forward primer they represent the 12 base Golay encoding barcode as defined by Caporaso et al.
doi:10.1371/journal.pone.0094249.t001

ambiguous bases over the entire length of the read, and be longer than 300 bp after trimming of the index and forward primer sequences in order to be retained after demultiplexing with QIIME [29]. After demultiplexing, the 454 sequences were denoised using the QIIME Denoiser according to the QIIME standard protocol. The V4-V5 Illumina datasets were initially demultiplexed using MiSeq Reporter v2.0. The sequences corresponding to the forward and reverse primers were trimmed from the demultiplexed reads using cutadapt (http://code.google.com/p/cutadapt/) using similar stringency settings to those used for the 454 sequences. The trimmed read pairs were then merged into single contigs using SeqPrep (https://github.com/jstjohn/SeqPrep) followed by a length-filtering step prior to analysis with QIIME. The Illumina V4 read pairs were merged and length filtered in a similar manner as the V4-V5 reads to form single contigs prior to being demultiplexed with QIIME. Reads from all datasets were quality filtered using a *Q*20 minimum value during demultiplexing. In order to ensure an even treatment and comparison of all sequence datasets for the seven sample sources, the demultiplexed sequences for all datasets were combined and processed as a single bulk dataset for QIIME analyses.

QIIME analysis

We used QIIME versions 1.6 and 1.7 to perform OTU clustering and alpha and beta diversity analyses [29]. Reference-based OTU clustering was done using the parallel uclust_ref method while *de novo* OTU clustering was done with standard uclust, using the default options as implemented in QIIME for both methods at the 97% similarity level. For reference OTU clustering and *de novo* OTU alignment we used the V4-V5 section of the 97% clustered Greengenes reference OTU NAST alignment [30,31]. The 2012–10 Greengenes database release was used initially as this was the current version when analysis began. After the 2013–08 release became available, all processing was re-run with the new release, allowing us to examine the effect of the reference itself on data analysis and interpretation. Taxonomy assignments were made using the RDP Classifier after retraining against the above mentioned Greengenes reference sequences and their respective taxonomy files as recommended by Werner et al [32]. Chimera checking was performed using ChimeraSlayer with standard options as implemented in QIIME against the V4-V5 region of the Greengenes reference alignment.

A more detailed description of our creation of the V4-V5 specific Greengenes reference files and the different QIIME processing methods used is provided in the supplementary methods (File S1). The scripts (denovo.sh, Ref.sh, RDS.sh) used for QIIME analysis are also included the supplementary material (File S2).

Data availability

The sequence data generated and used in this study were deposited in the European Nucleotide Archive SRA under project ID PRJEB4688.

Results

We conducted a comparison of 454 pyrosequencing and Illumina sequencing of 16S amplicons by analyzing four different sequencing libraries for six different natural microbial community samples: the V4 hyper-variable region sequenced on an Illumina MiSeq (V4.I), a V4-V5 Illumina library that was gel-purified (V4V5.Ia), a second V4-V5 Illumina library that was AMPure purified (V4V5.Ib), and a V4-V5 454 pyrosequencing library (V4V5.454). We also analyzed one V4-V5 454 library and two replicate V4 and V4-V5 Illumina libraries for a synthetic mock community. As one of the stated advantages of Illumina sequencing is a lower error rate compared to 454 pyrosequencing [15,16], we first compared the overall quality of the sequences generated from each sequencing run. While these values represent predicted rather than absolute error rates, they are the most commonly used proxy for examining sequence quality and thus one of the primary metrics used in data pre-processing. The

median PHRED quality score (*Q*-score) for each base over the length of a read had an average value of *Q*39 in the V4V5.454 datasets (Figure 1) and represented the standard to which the Illumina datasets were compared. As the error rate of Illumina sequences increases at the 3′ ends of each read, as indicated by a drop in *Q*-scores, we merged the paired Illumina reads to form a single consensus contig prior to quality and QIIME analysis. This process serves to minimize the effects of sequencing errors by forming a consensus sequence from the overlapping ends of the reads as previously demonstrated [22]. The median *Q*-score for each base of the consensus contig after read merging was similar to or greater than that of the 454 dataset (Figure 1), demonstrating that by merging the paired Illumina sequencing reads we could produce single contigs of similar length as 454 pyrosequencing but of higher average quality. Additionally, improvements to Illumina's Real Time Analysis (RTA) base-calling software that occurred during this study have resulted in significantly higher *Q*-scores for bases later in a read, which correspond to greater confidence in base-calling. This improvement can be seen in the reads from the V4V5.Ib dataset, which have higher median *Q*-scores for bases in the overlap region than the V4V5.Ia dataset, which was sequenced using an earlier version of the RTA software (Figure 1). Additional improvements from Illumina regarding MiSeq read lengths and on-instrument data analysis now suggest that merging paired reads from longer amplicons, such as those covering the V1-V3 regions are now feasible. Overall, after read merging a greater proportion of reads from the Illumina sequencing runs was retained after demultiplexing compared to the V4V5.454 data when using the same quality threshold of *Q*20 (data not shown).

Low-levels of dataset contamination occur in Illumina sequencing

While we observed that overall read quality was higher in the Illumina datasets compared to the 454 pyrosequencing dataset, during our analyses we identified a small percentage of reads in the Illumina datasets that did not belong in the demultiplexed dataset for a given sample, a result that we did not observe in any of the 454 datasets. The source of these reads could be assigned to two separate issues that are particular to Illumina sequencing systems and especially for the MiSeq. The first source of these incorrect reads was the carry-over of samples from a previous sequencing run into the subsequent sequencing run. This occurs when samples from a previous run persist in the fluidics lines of the system and become mixed with new samples in subsequent sequencing runs [33]. If identical indices are used in consecutive sequencing runs, then the carry-over of reads from a previous library can artificially suggest the presence of low abundance OTUs that are not truly present in a subsequent sample.

The second source of incorrect reads that we identified was from other libraries that were sequenced during the same sequencing run. This was most noticeable for the V4V5.Ib datasets, which we sequenced at the same time as amplicon libraries created for non-ribosomal genes. Both the 16S V4V5.Ib and non-ribosomal libraries featured different six base TruSeq indices and we determined that ~0.06% of reads in the 16S libraries were sequences from the non-ribosomal amplicon libraries. This was the first time that the non-ribosomal libraries were sequenced, thus the contamination could not have been due to carry-over contamination of the fluidics lines from a previous run as noted above. In addition, the 16S and non-ribosomal libraries were prepared completely independently of each other and were only pooled immediately before loading into the MiSeq, eliminating the chances of contamination during library preparation. After consultation with Illumina representatives, we assume that this result is due to sequencing and/or image analysis errors during the index sequencing phase of the MiSeq run, which occurs as a separate step in the sequencing process, and likely caused a small number of amplicons from one library to be incorrectly

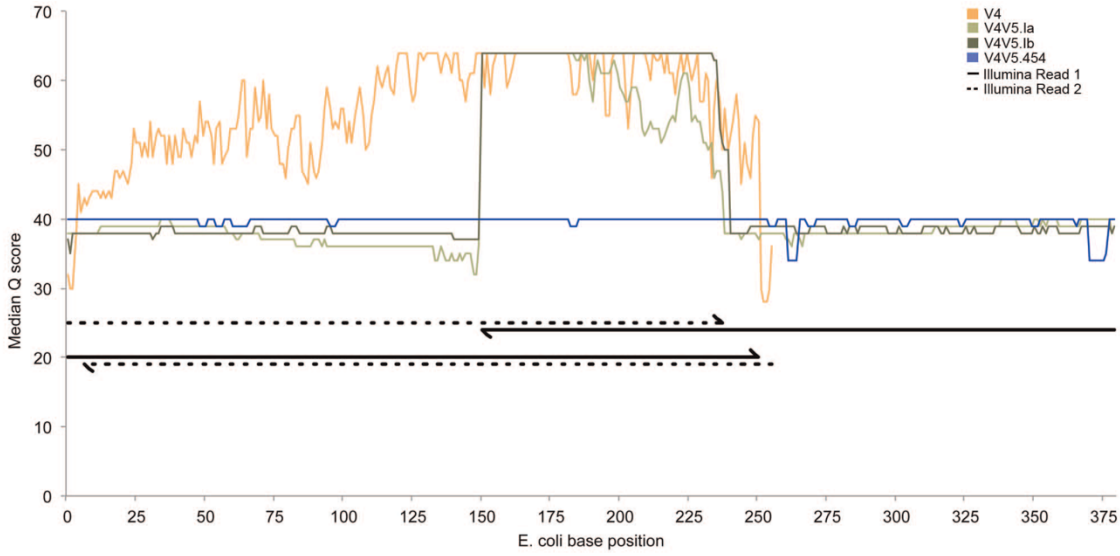


Figure 1. Comparison of 454 and Illumina sequence quality. Plot depicting the median per base PHRED quality scores (*Q* score) for the full length 454 and merged Illumina reads from the six natural community samples. The V4 data is shown in orange, the first V4-V5 Illumina run (V4V5.Ia) is in light green, the second run (V4V5.Ib) is in dark green, and the 454 data is in blue. The size and over-lapping regions of the V4 and V4-V5 Illumina amplicons is shown in black below the quality plots. Illumina sequencing read 1 is depicted as a solid line while read 2 is dashed, with arrow heads depicting the direction of the read in reference to the *E. coli* base position given along the X axis.
doi:10.1371/journal.pone.0094249.g001

assigned an index corresponding to another library. While we could adequately identify the contaminating non-ribosomal sequences and remove them from our 16S datasets before proceeding with downstream analyses, this finding suggests that a similar level of index misassignment could occur between different 16S libraries when sequenced on the same run, which will artificially inflate alpha diversity measures and bias the interpretation of results when investigating low abundance OTUs.

In addition to index misassignment and/or sample carryover of similarly indexed libraries, we also identified reads from the ΦX174 (phiX) genome in all of the raw Illumina datasets. These reads originated from the unindexed phiX control library that is added to Illumina sequencing runs as an inline control library and could not have resulted from contamination during library construction. Formerly, the sequencing of 16S amplicons on the MiSeq required phiX to comprise 50–90% of the run's throughput, as was done when we sequenced the V4.I and V4V5.Ia libraries. Upgrades to the MiSeq's RTA base-calling software (since version 2.2) have reduced the amount of phiX that is recommended to be added to amplicon sequencing runs to only 2–5%, however phiX reads were still observed in the raw V4V5.Ib datasets which we sequenced with the upgraded RTA software and only 2.5% phiX. In order to prevent the presence of phiX reads in the Illumina datasets from biasing our downstream analyses, we incorporated a step in our Illumina pre-processing pipelines to identify and remove these reads prior to analysis with QIIME.

Determining optimal OTU clustering method

As Illumina sequencing with the MiSeq generally produces at least 10 times more sequences than 454 pyrosequencing, recent publications discussing Illumina 16S amplicon sequencing have used and recommended reference-based OTU clustering methods to enable users to quickly process their data [18]. While reference OTU clustering has been used for the analysis of 454 data, many investigators still choose *de novo* OTU clustering methods for 454 data analysis as this method recovers OTUs not found in reference datasets. Thus we examined what effect these two clustering methods had on data analysis and the interpretation of the results. We performed *de novo* OTU clustering of the bulk dataset using standard QIIME methods for processing pyrosequencing data with uclust used for OTU clustering and chimera checking performed with ChimeraSlayer, while reference OTU clustering was performed with the parallel version of uclust_ref against the Greengenes 2012–10 reference as that was the current Greengenes release at the time. These two clustering methods yielded very different results, with the number of OTUs observed and the number of sequences assigned to an OTU being lower when performing reference-based clustering than *de novo* clustering for the same dataset (Table S1, Figure 2a).

One of the factors that contributed to the difference between the two clustering methods was that a large number of sequences failed to be assigned to a reference OTU (Table S1). On average, only 65% of the reads for a given dataset were assigned to a reference OTU, although the scale of this effect varied dramatically between the different samples. For example, over 90% of reads from each of the human stool datasets were assigned to a reference OTU, while for the termite sample only between 30% to 40% of reads from the V4-V5 datasets were assigned to a reference OTU (Table S1). As the Greengenes 2013 release occurred as we were performing our data analyses, we repeated the reference OTU clustering using this newer reference as the 2013 release includes a greater number of reference sequences than the 2012 release. When using the Greengenes 2013 release for reference OTU clustering, we observed that a greater number of sequences were assigned to a reference OTU and a greater number of OTUs per sample observed compared to using the 2012 version. Even with this improvement over the 2012 reference, we still did not observe the same number of OTUs as in the *de novo* clustered datasets (Table S1, Figure 2a). This difference in the number of OTUs based on reference or *de novo* OTU clustering carried over into the calculation of alpha diversity measures, although beta diversity analyses, particularly those using the phylogenetic tree based weighted UniFrac metric were less affected (data not shown).

Reference plus *de novo* OTU clustering with chimera checking

In order to more closely replicate the results of *de novo* OTU clustering while retaining the processing efficiency of reference OTU clustering, we developed an analysis pipeline that first performs parallel reference OTU clustering using the 97% Greengenes OTUs as reference, followed by *de novo* OTU clustering and chimera checking with ChimeraSlayer of the sequences that failed to be assigned to a reference OTU. As discussed below, this reference plus *de novo* OTU clustering with ChimeraSlayer pipeline, which we call RDS, produced similar alpha diversity measures, taxonomic composition, and beta diversity comparison as the chimera checked *de novo* OTU clustering method. Unlike the similar open reference clustering method of uclust_ref implemented in QIIME, which can only run on a single processing core, this split implementation takes advantage of the ability to perform reference OTU clustering across multiple processing cores, reducing the total time for analysis and thus is more amenable to processing large Illumina datasets.

We compared the number of observed OTUs and the Simpson (D), Shannon (H') and phylogenetic distance (PD) alpha diversity metrics generated by the RDS method using each of the Greengenes references to those obtained using *de novo* and reference OTU clustering. Using either Greengenes reference, the number of OTUs generated using the RDS method was more similar to the number of OTUs obtained by *de novo* clustering than for reference OTU clustering alone (Table 2, Figure 2). While the magnitude of the difference in the number of OTUs between processing methods varied for each dataset, on average the results of reference-based processing were 23.7% different from *de novo* while RDS processing was 12% different. When analyzed by ANOVA, the number of reference-clustered OTUs was significantly different from the results of *de novo* OTU clustering ($p<0.01$) but there was no statistical difference between the RDS method and *de novo* ($p>0.05$). Compared to *de novo* OTU clustering, the alpha diversity measures generated using the RDS clustering method had Pearson correlation coefficients closer to 1 and greater linear curve fits than the reference-clustered measures, indicating that the RDS method reproduces the results of *de novo* OTU clustering better than reference-based OTU clustering alone.

Even though the RDS processing method reproduced the results of *de novo* OTU clustering better than reference-based OTU clustering according to the alpha and beta diversity measurements we examined, there were a greater number of OTUs in nearly all of the Illumina datasets than has been reported for similar samples in the literature. In particular, the Illumina datasets of the mock community had between 25 to 125 times as many OTUs as expected based on an analysis of the available reference genome sequences. One factor contributing to this increase could be the above mentioned dataset contamination, which can be partially addressed using OTU filtering strategies to remove OTUs that

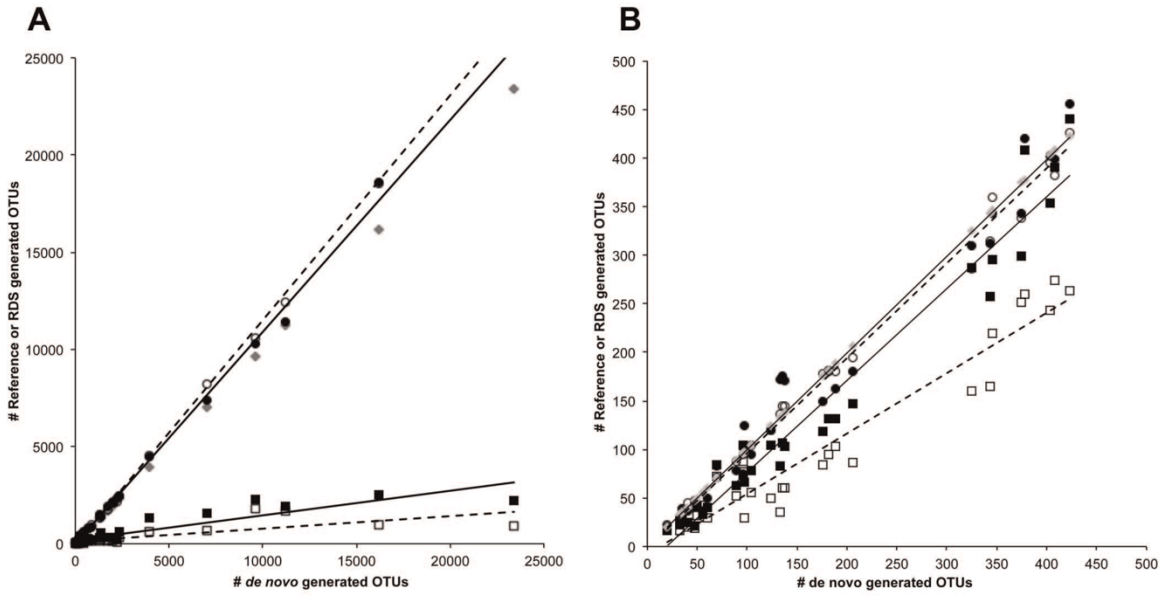


Figure 2. The RDS processing method replicates *de novo* OTU clustering better than reference-based clustering. The correlation between OTU clustering methods is shown by plotting the number of raw (a) and filtered (b) OTUs observed when using *de novo* OTU clustering versus reference or the RDS method. The reference-based OTU clustering results are depicted with squares while the RDS OTU clustering results are depicted with circles. Open markers indicate samples where the Greengenes 2012 reference was used while closed markers indicate samples where the Greengenes 2013 reference was used. *De novo* results are depicted as gray diamonds. Linear regression lines are shown for the reference and RDS datasets, with dash lines fitted to datasets processed using the Greengenes 2012–10 reference and solid lines fitted to datasets processed using the Greengenes 2013–08 reference.
doi:10.1371/journal.pone.0094249.g002

account for a low percentage of the total reads as recommended for Illumina datasets by Bokulich et al. [34] An analysis of different filtering methods and cutoffs showed that no single filtering value worked equally well across all samples, as cutoffs that reduced the number of OTUs in the mock community samples to reasonable numbers were overly restrictive for other samples (Table S2). Manual examination of the representative OTU sequences from the Illumina mock community datasets showed that a large proportion of them represented chimeras between two or more species from the community. Because the highly synthetic nature of the mock community is not very representative of the richness and evenness of natural samples, we chose to remove single and doubleton OTUs from the full OTU table as spurious reads, followed by filtering of OTUs representing fewer than 0.005% of all sequences as was recommended by Bokulich et al. [34] While the number of OTUs observed with reference clustering against the Greengenes 2013 reference was more similar to *de novo* after implementing the OTU filtering step, linear regression analysis showed that the RDS method still produced results more reflective of *de novo* OTU clustering (Figure 2b).

After processing the datasets using the RDS method and incorporating the OTU filtering step, the alpha diversity metrics for each of the Illumina datasets had more OTUs and a larger phylogenetic distance (PD) than the corresponding 454 dataset (Table 2). We observed a similar result when performing *de novo* OTU clustering of the datasets with the OTU filtering step. Except for one of the mock community datasets, all of the Illumina datasets for a sample had a greater number of input sequences than the corresponding 454 dataset. To prevent differences in sequencing depth from biasing our comparisons of 454 and Illumina sequencing, we normalized the number of sequences in the datasets for a sample by rarefying each dataset to the number of reads in the corresponding 454 dataset. The smallest mock community Illumina dataset was excluded from this analysis. After rarefication, the number of OTUs observed in the Illumina datasets was still greater than in the corresponding 454 dataset (Table 3) although the degree of difference was smaller for the higher diversity samples (rumen, sewage, termite) than the low diversity samples (human stool, leech, mouse).

When we compared the number of OTUs between the V4 and V4-V5 Illumina datasets of each sample, the V4 dataset consistently had fewer OTUs than for the corresponding V4-V5 Illumina datasets. Compared to the V4-V5 amplicons, the V4 amplicons are ~100bp shorter and cover only a single hyper-variable region. The greater number of OTUs for the Illumina V4-V5 datasets compared to the V4 after rarefaction suggests that the increased sequence information available for analysis by including the V5 hyper-variable region allowed for the discrimination of new OTUs that could not be differentiated based on the V4 region alone.

Beta diversity analysis

Beta diversity analysis of all datasets showed that each sample source represented a distinct microbiome irrespective of the processing method used. Each of the individual datasets clustered together on the basis of their original sample source as determined by principal coordinates analysis of the Bray-Curtis and UniFrac distances between each dataset (Figure 3). This clustering was independent of the hyper-variable regions chosen for sequencing, V4 or V4-V5, or the sequencing platform used, GS FLX or

Table 2. Comparisons of alpha diversity metrics produced from different processing methods.

			de novo				Reference				RDS			
Sample Source	**Library**	**Input # Seqs**	**# OTUs**	***D***	***H′***	**PD**	**# OTUs**	***D***	***H′***	**PD**	**# OTUs**	***D***	***H′***	**PD**
Human stool	H.v4.l	93769	96	0.778	3.098	18.35	73	0.786	3.057	9.72	75	0.784	3.030	13.08
	H.v4v5.l.a	32506	70	0.788	3.099	11.73	84	0.793	3.224	7.58	83	0.793	3.216	10.34
	H.v4v5.l.b	153159	97	0.754	2.922	19.89	104	0.757	3.033	10.61	103	0.756	3.015	16.35
	H.v4v5.454	7882	51	0.775	2.689	11.70	42	0.772	2.624	6.98	44	0.773	2.640	8.75
Leech intestinum	L.v4.l	118954	56	0.750	2.836	12.47	33	0.611	1.903	5.56	37	0.634	2.068	9.11
	L.v4v5.l.a	44230	41	0.580	1.801	7.97	25	0.509	1.377	3.88	37	0.578	1.788	6.93
	L.v4v5.l.b6	191369	105	0.623	2.076	21.13	78	0.549	1.626	10.23	95	0.620	2.049	18.07
	L.v4v5.l.b11	171969	90	0.615	2.006	20.77	63	0.550	1.628	10.33	78	0.613	1.986	16.96
	L.v4v5.454	10229	19	0.697	2.240	6.09	17	0.676	2.085	3.04	21	0.697	2.255	5.14
HMP Mock Even	Mock.v4.l.1	213043	141	0.932	4.436	15.19	107	0.932	4.353	6.65	176	0.937	4.635	16.51
	Mock.v4.l.105	240682	143	0.936	4.574	16.43	103	0.935	4.434	6.75	171	0.941	4.777	16.19
	Mock.v4v5.l.1	2484	99	0.932	4.588	12.16	66	0.930	4.487	5.53	125	0.946	5.140	13.09
	Mock.v4v5.l.11	90126	138	0.941	4.848	14.25	83	0.943	4.672	5.96	172	0.958	5.419	15.93
	Mock.v4v5.454	7386	36	0.930	4.073	8.80	28	0.930	4.059	5.15	39	0.931	4.106	8.07
Mouse small intestine	M.v4.l	45411	61	0.743	2.620	12.59	40	0.643	1.890	7.19	50	0.653	2.010	9.60
	M.v4v5.l.a	24061	47	0.766	2.739	9.01	21	0.711	2.159	4.28	47	0.772	2.859	7.16
	M.v4v5.l.b	155976	178	0.811	3.042	39.03	132	0.764	2.469	17.87	180	0.816	3.204	35.37
	M.v4v5.454	10453	33	0.761	2.432	9.02	22	0.749	2.255	5.49	30	0.761	2.431	7.21
Rumen content	R.v4.l	93881	402	0.986	7.292	67.79	390	0.988	7.297	29.91	399	0.987	7.282	57.91
	R.v4v5.l.a	44431	372	0.984	7.176	58.58	408	0.991	7.586	25.27	420	0.991	7.650	50.96
	R.v4v5.l.b	217371	417	0.985	7.284	68.63	440	0.992	7.721	29.59	456	0.992	7.776	59.21
	R.v4v5.454	35527	323	0.985	7.161	59.15	287	0.985	7.089	26.56	310	0.986	7.209	51.04
Municipal sewage	S.v4.l	117562	375	0.955	6.389	75.57	299	0.977	6.561	30.39	343	0.952	6.247	58.70
	S.v4v5.l.a	28971	346	0.973	6.814	69.85	295	0.982	6.831	28.80	343	0.973	6.827	57.48
	S.v4v5.l.b	160654	403	0.975	6.925	80.75	354	0.984	6.982	34.40	402	0.975	6.968	66.40
	S.v4v5.454	38227	343	0.979	6.884	73.67	257	0.985	6.811	31.30	312	0.979	6.876	59.17
Termite hindgut	T.v4.l	124664	182	0.941	5.107	30.63	132	0.926	4.619	13.84	163	0.935	4.925	23.31
	T.v4v5.l.a	31220	170	0.946	5.311	27.41	119	0.937	4.968	12.95	149	0.947	5.286	22.19
	T.v4v5.l.b	164780	198	0.915	4.921	33.24	147	0.904	4.606	16.74	180	0.918	4.947	27.57
	T.v4v5.454	7146	126	0.928	4.933	25.38	104	0.920	4.679	14.06	120	0.929	4.945	21.73

doi:10.1371/journal.pone.0094249.t002

Table 3. Alpha diversity measures of RDS processed samples after normalization.

Sample Source	Library	Normalized Seqs[A]	# OTUs	*D*	*H′*	PD
Human stool	H.v4.I	7737	56	0.785	2.999	9.13
	H.v4v5.I.a		72	0.793	3.232	8.95
	H.v4v5.I.b		72	0.756	3.006	9.23
	H.v4v5.454		44	0.773	2.640	8.75
Leech intestinum	L.v4.I	10213	26	0.628	2.055	6.21
	L.v4v5.I.a		28	0.575	1.767	5.40
	L.v4v5.I.b6		38	0.624	2.057	7.65
	L.v4v5.I.b11		34	0.613	1.981	6.31
	L.v4v5.454		21	0.697	2.255	5.14
HMP Mock Even	Mock.v4.I.1	7331	146	0.936	4.606	14.29
	Mock.v4.I.105		153	0.941	4.762	14.97
	Mock.v4v5.I.11		154	0.959	5.431	15.31
	Mock.v4v5.454		39	0.931	4.106	8.07
Mouse small intestine	M.v4.I	10350	34	0.650	1.993	6.37
	M.v4v5.I.a		46	0.769	2.838	7.14
	M.v4v5.I.b		55	0.815	3.196	10.30
	M.v4v5.454		30	0.761	2.431	7.21
Rumen content	R.v4.I	27672	386	0.987	7.275	55.75
	R.v4v5.I.a		420	0.991	7.650	50.96
	R.v4v5.I.b		426	0.992	7.751	53.03
	R.v4v5.454		310	0.986	7.210	51.04
Municipal sewage	S.v4.I	19354	311	0.953	6.253	52.53
	S.v4v5.I.a		343	0.973	6.827	57.48
	S.v4v5.I.b		349	0.975	6.951	58.00
	S.v4v5.454		302	0.979	6.869	58.15
Termite hindgut	T.v4.I	6850	127	0.935	4.897	18.68
	T.v4v5.I.a		136	0.949	5.338	20.10
	T.v4v5.I.b		139	0.916	4.909	19.90
	T.v4v5.454		120	0.929	4.945	21.73

[A]The normalized number of sequences represents the number of sequences that each dataset of a given sample were normalized to by rarefaction to allow for intra-sample comparisons of the datasets.

doi:10.1371/journal.pone.0094249.t003

MiSeq, indicating that these factors had no obvious effect on the interpretation of beta diversity analyses when comparing the diverse group of samples we used in this study. While the RDS method did not produce beta diversity results identical to those generated when using *de novo* OTU clustering, the overall interpretation of results was similar between the two methods. The primary difference that we observed was that when using the Bray-Curtis metric the human, mouse, and mock community samples were shown to be more similar to each other when the data was processed using the RDS method compared to using *de novo* OTU clustering. This result was similar to what we observed when performing reference OTU clustering only, and suggests that these three samples shared a greater percentage of OTUs as a result of the reference OTU clustering step of the RDS method than we observed with *de novo* clustering (Figure S1).

Because of the large overall dissimilarities between the seven samples as determined by principal coordinate analysis, we also performed beta diversity analyses of the datasets for each sample independently. In each case, the V4 Illumina dataset was consistently more different from the corresponding V4-V5 datasets than the V4-V5 datasets were from each other, indicating that choice of hyper-variable region had a greater effect on beta diversity than the choice of sequencing technology (Figure S2).

Effects of hyper-variable region and OTU clustering method on observed taxonomic diversity

While alpha and beta diversity measures provide important insights into the structure and relationship of microbial communities, a key aspect of generating hypotheses about the functional and physiological aspects of a microbial community is knowing its taxonomic composition. We determined the effect of the hyper-variable region chosen for sequencing and of the OTU clustering method used for analysis on the taxonomic composition of a sample by comparing the taxonomy summaries for each dataset when processed using *de novo*, reference-based, and the RDS OTU clustering methods. These comparisons revealed that for some samples there was a large effect on the observed taxonomic composition of the choice of hyper-variable regions sequenced or OTU clustering method used.

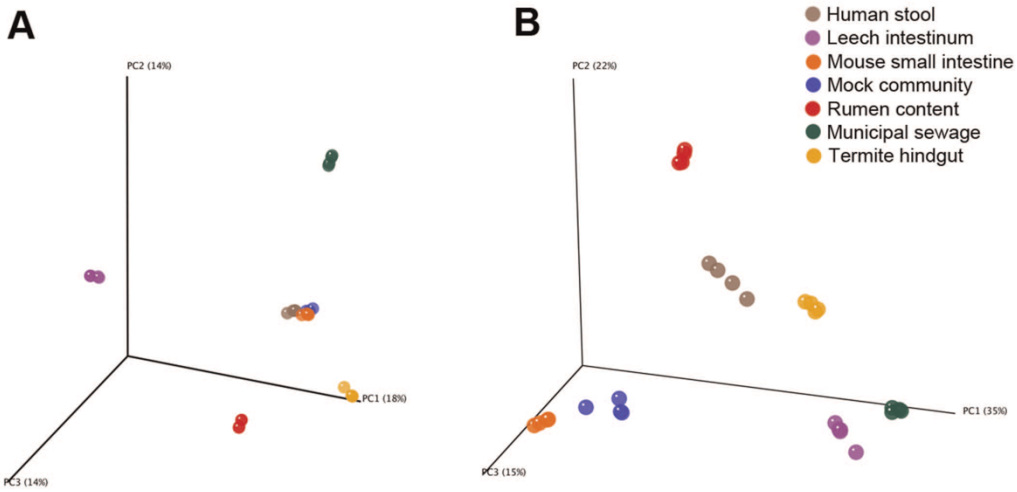


Figure 3. Beta diversity analysis of all datasets. Three dimensional principal coordinates analysis plots showing the relatedness of datasets using either the Bray-Curtis (A) and weighted UniFrac (B) metric. Individual datasets are represented at spheres which are colored according to their sample source as follows: human stool – brown, leech intestinum – purple, mouse small intestine – orange, mock community – blue, non-adherent rumen contents – red, mixed liquor – green, termite hindgut – gold.
doi:10.1371/journal.pone.0094249.g003

We tested the three processing pipelines using control DNA from a synthetic mock community created as part of the Human Microbiome Project (HMP) to determine if processing method alone introduced a source of bias [7]. The mock community DNA used for the Illumina libraries comprised 20 cultured bacterial species, while the DNA used for the 454 library also included *Porphyromona gingivalis*. None of the resulting datasets showed a taxonomic composition that was identical to the known composition of the mock community, however each of the three processing methods (*de novo*, reference, RDS) yielded a similar taxonomic composition for each of the three types of libraries (V4.I, V4V5.I, and V4V5.454, Figure 4).

The abundance of some taxa was affected dramatically by the type of library that was created and the processing method. In the V4 Illumina libraries, the genus *Propionibacterium* was almost completely absent while in the V4-V5 libraries it represented ~1.5% of the 454 dataset and 2.4–2.9% in the Illumina datasets. This result was likely due to primer specificity of the V4 primers compared to the V4-V5 primers, as there is a single base pair difference between the V4 forward primer and the annealing site

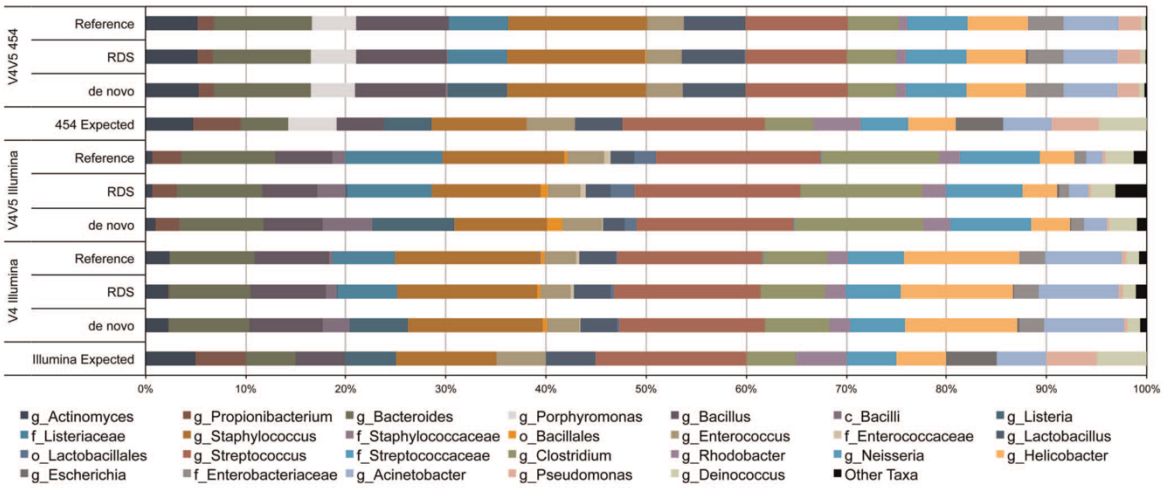


Figure 4. Effect of processing method on the taxonomic composition of the mock community datasets. Plot comparing the taxonomic composition of the mock community sample for the three different library types sequenced when processed three different ways. The replicate V4 and V4-V5 Illumina datasets were combined into one representative dataset for each library type. All taxonomic assignments were made using the RDP Classifier after retraining with the 2013-08 Greengenes reference. Taxonomic ranks are noted by letters preceding the taxon name as follows: genus – g, family – f, order – o.
doi:10.1371/journal.pone.0094249.g004

based on the *P. acnes* reference genome. Across all three library types, we consistently observed that the genus *Listeria* was only identified when using *de novo* OTU clustering, whereas the family *Listeriaceae* was instead observed when using the reference or RDS processing methods. Similarly, the genus *Escherichia* was only marginally identified in any of the datasets regardless of processing, with the family *Enterobacteriaceae* instead being the predominant taxonomic assignment for these OTUs. It is interesting to note that this result only occurred when the Greengenes 2013 release was used for taxonomy assignment, as OTUs were correctly classified as *Escherichia* when we used the 2012 version of the reference.

While the taxonomic composition of the mock community datasets showed little to no specific bias associated with the choice of hyper-variable regions sequenced or data processing method used, we did observe some distinct differences in the six natural microbial community samples that we analyzed. During our initial analyses using the Greengenes 2012 reference for OTU clustering and taxonomic assignment we observed that for certain samples the use of reference clustering alone often missed entire taxa. The most dramatic example of this was with the V4-V5 libraries for the termite sample, for which the class *Endomicrobia* was almost completely absent from the reference clustered datasets but comprised nearly 30% of the community when using *de novo* or the RDS processing methods (Figure 5). While this issue was largely resolved with the Greengenes 2013 release, the taxonomic composition of the RDS processed datasets were more similar to the *de novo* OTU clustered datasets than the reference clustered datasets were. The termite sample was also where differences between the V4 and V4-V5 libraries were the most apparent. While the V4-V5 Illumina and 454 libraries were not statistically different from each other, the V4 library was significantly different from both V4-V5 libraries (data not shown). In the V4-V5 libraries the genus *Treponema* comprised ~45% of the community but nearly 75% in the V4 library regardless of processing method (Figure 5). While we could not determine the exact cause of this discrepancy from the data, it is possible that primer amplification bias contributed to this result.

In the human stool sample datasets, the abundance of the two most abundant genera, *Bacteroides* and *Escherichia*, differed greatly between each of the three library types (Figure S2). While the differences between the V4 and V4-V5 libraries are likely due to the choice of different primers, the genus *Bacteroides* was more abundant in the V4-V5 Illumina libraries compared to the 454 (~65% vs. 55%), while *Escherichia* was much less abundant (~10% vs. ~28%). This difference in abundance between the two V4-V5 library types was observed even after normalization of the datasets by rarefaction, and thus does not directly represent a sampling depth bias. As noted above for the mock community datasets, the genus *Escherichia* was only identified in datasets processed using the Greengenes 2012 reference for taxonomy assignment, with a single exception of the V4-V5 Illumina datasets processed using the RDS method. In this case, the genus *Escherichia* was observed when using the 2013 reference but not at the same level as when using the 2012 (Figure S2). We observed similar differences in taxonomic composition of the sample corresponding to library type for the other five natural community samples that we analyzed, although these differences were minor for the rumen and sewage datasets which had the highest overall taxonomic diversity of the seven samples we analyzed.

Discussion

Illumina sequencing can faithfully supplant 454 pyrosequencing

The primary goal of this study was to examine how well Illumina sequencing could serve as a direct replacement for 454 pyrosequencing while using existing 16S sequencing primers and analysis workflows. To determine this we analyzed six natural microbial communities and a mock community by using both 454

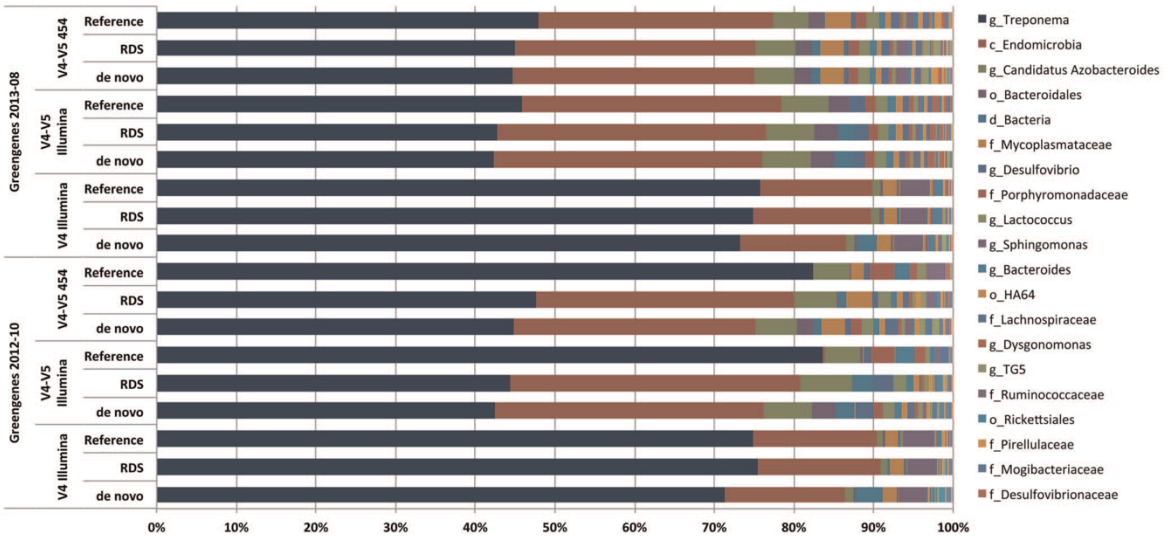


Figure 5. Effects of processing method and Greengenes database version on the taxonomic composition of the termite datasets. Plot comparing the taxonomic composition of the termite hindgut sample for the three different library types sequenced when processed using three different methods. The replicate V4 and V4-V5 Illumina datasets were combined into one representative dataset for each library type. Taxonomic assignments were made using the RDP Classifier after retraining with either the 2012–10 or 2013–08 Greengenes references. Taxonomic ranks are noted by letters preceding the taxon name as follows: genus – g, family – f, order – o, class – c, phylum – p, domain – d.
doi:10.1371/journal.pone.0094249.g005

pyrosequencing and Illumina sequencing of the V4-V5 hyper-variable region of the 16S rRNA gene. We additionally performed Illumina sequencing of the V4 region using the protocol developed by Caporaso et al. [18], which has been adopted as the standard protocol for Illumina 16S sequencing by researchers participating in the Earth Microbiome Project. Because the individual reads generated with the MiSeq are shorter than the single reads generated by the GS FLX, and previous studies [22] and our own analysis found that error rates increased towards the 3′ end of the reads, we utilized read merging of the paired Illumina reads to create single consensus Illumina reads with similar length to those generated by 454 sequencing. This pre-processing step for the Illumina datasets yielded merged reads that had a higher average quality than for the reads generated by 454 pyrosequencing (Figure 1), along with a greater number of reads per sample (Table 2).

When analyzing all of the datasets from the samples *en masse* we observed small differences in alpha diversity measures between the pyrosequencing and Illumina datasets for the high diversity samples while larger differences were observed for low diversity samples. Conversely, PCoA plots of beta diversity analyses showed that there was little to no apparent effect of the sequencing method used (454 or Illumina) or variable regions chosen (V4 or V4-V5), as each dataset from a given sample clustered together (Figure 3). Analysis of the individual datasets for each sample, however, did reveal that the V4 dataset was consistently more different from the V4-V5 datasets than the V4-V5 datasets were from each other (Figure S2). Part of this difference stems from the use of primers that anneal to different regions of the 16S rRNA gene for library creation, which likely have different amplification biases and template specificity [35,36]. This bias was apparent in examining the taxonomic composition of the mock community datasets which all had slightly different abundances for each taxon across the three library types we examined and the genus *Propionibacterium* nearly absent from the V4 libraries. While we did observe differences between the V4-V5 454 and V4-V5 Illumina datasets, these differences did not significantly affect the overall interpretation of beta-diversity analyses, although their effect on taxonomic composition varied according to sample. On the basis of our overall findings, we can conclude that researchers who wish to switch to Illumina sequencing from 454 pyrosequencing should be able to modify their existing primers by simply replacing the 454 adaptor sequences with Illumina TruSeq adaptor sequences. An additional option for researchers who do not need or wish to adapt a pre-established 454 workflow is to use one of the published V4 sequencing formats developed for Illumina sequencing by Caporaso et al. or Kozich et al.[18,22] While choosing a different hyper-variable region for analysis did affect the results in a sample-dependent manner, our analyses show that overall the V4 amplicons produced similar alpha and beta diversity measures as the V4-V5 amplicons.

One of the major differences between the Roche 454 GS FLX and Illumina MiSeq instruments is that the MiSeq is currently capable of generating well over 10 times as many sequence reads as the GS FLX in a single sequencing run. Combined with much lower operating costs, Illumina sequencing on the MiSeq provides researchers with the opportunity to sequence individual samples to a greater sampling depth than is feasible with the GS FLX and/or to include more samples in a single sequencing run through increased multiplexing of barcoded libraries. As sequencing depth increases however, a greater number of erroneous sequences can be incorporated into the resulting dataset, which will artificially bias estimates of alpha diversity through the generation of spurious OTUs. These erroneous sequences often arise due to chimera formation and PCR errors during library preparation, or are the result of sequencing errors that were not identified and removed during data processing. Protocols have been developed for 454 pyrosequencing to minimize the presence and effects of illegitimate sequences/OTUs on diversity analyses, and we incorporated these protocols as appropriate into our library preparation and data processing and analysis methods [37–39].

To minimize the effects of sequencing errors we first merged the paired Illumina reads to form a single consensus sequence prior to OTU clustering. This step results in a higher confidence that the base calls for the merged region are correct and thus reduces sequencing associated errors in the Illumina datasets (Figure 1). We also incorporated chimera checking with ChimeraSlayer as part of our RDS analysis pipeline. However, as demonstrated with the Illumina-sequenced mock community samples, not all chimeric OTUs were correctly identified and removed. One reason for this is that the chimera checking process typically depends on comparing differences in the sequence similarity of the two ends of a query sequence to two or more reference sequences derived either from a reference database such as Greengenes or chosen from within the dataset itself. This method poses a problem in detection as chimeras present in short sequences from closely related organisms are more difficult to identify than in longer sequences. Additionally, chimeric sequences originating from three or more parent sequences, such as those observed in the Illumina mock community datasets, may not be identified as chimeric but as novel sequences instead.

Reference OTU clustering can bias observed diversity

As the volume of sequence data generated by Illumina instruments is orders of magnitude greater than for the GS FLX, processing and analysis pipelines that were designed to handle pyrosequencing datasets have had to be modified to process Illumina data more efficiently. One such modification has been a shift from using *de novo* generation of OTUs for large sequencing datasets to the use of reference OTUs such as those from the Greengenes [30,31] or Silva [40] reference databases. The principal advantage of reference OTU clustering is that it is significantly faster than *de novo* OTU generation as it can be run in parallel across multiple processing cores, and the availability of reference datasets with pre-constructed phylogenetic trees and taxonomies allows for a simplified and more efficient analysis pipeline. However, with reference-based OTU clustering alone the observed microbial diversity of a sample can only be as diverse as the reference set itself, which can artificially limit the observed diversity for highly diverse or exotic environments whose microbial populations have few representative sequences in reference databases.

In this study we found that performing reference-based OTU clustering using either the Greengenes 2012 or 2013 references resulted in a reduction in the number of observed OTUs compared to *de novo* OTU clustering (Tables 2 and S1, Figure 2). The use of reference-based OTU clustering alone also had large effects on the observed taxonomy some samples, with certain taxa completely missing or misidentified when reference clustering was used compared to *de novo* (Figure 5). While the curators of the Greengenes database have made great efforts to expand their reference datasets to include more sequences from highly diverse and complex microbial communities, our results suggest that additional improvements are needed to provide better coverage for many non-human associated microbial environments. This is of particular importance as a greater number of researchers take advantage of the low cost of Illumina sequencing to characterize the microbial communities in many new and diverse environments that may not be well represented in current reference databases.

As we demonstrated, one option that researchers have is to perform reference OTU clustering and then to analyze the reduced number of sequences that did not match the reference data set using *de novo* OTU clustering, which we described above as the reference plus *de novo* with ChimeraSlayer, or RDS method. Our results demonstrated that the RDS method produces alpha (Table 2) and beta (Figure 2) diversity metrics and taxonomy summaries (Figure 3) that are more similar to *de novo* OTU clustering than reference-based clustering alone. While the current open reference picking implementation of uclust_ref allows for the creation of *de novo* OTUs from reads not assigned to a reference sequence, this process is limited to running on a single processing core. Our implementation of two separate steps for reference and *de novo* OTU clustering in the RDS method allows for reference clustering to be performed across multiple processing cores. This hybrid analysis method allows researchers to efficiently analyze large sequencing datasets generated with Illumina sequencing platforms while retaining the ability to identify novel OTUs that are not currently present in reference datasets.

Limitations of reference databases for taxonomy assignment

While not always feasible, *a priori* knowledge of the general composition of a microbial community can provide important checks for validating the results of high-throughput 16S sequencing surveys. Our inclusion of the mock community developed by the Human Microbiome Project partially served as such a control to identify potential issues with our library construction, sequencing and data analysis workflows. When using the Greengenes 2012 reference that was available when we began this study, we found that the taxonomic composition of the mock datasets differed greatly from expected, with many OTUs not being classified to the genus level but to higher taxonomic ranks instead. The release of the 2013–08 Greengenes reference database resolved many of these assignment issues, however the genus *Escherichia* was still not correctly identified when performing taxonomic assignment of OTUs with the Greengenes 2013 reference and the genus *Listeria* was only identified in the *de novo* processed datasets.

During our initial analysis of the leech intestinum samples using the Greengenes 2012 reference no OTUs were assigned to the genus *Aeromonas* for any of the datasets regardless of processing method, a finding inconsistent with previous culture and non-culture based studies we conducted of the leech intestinum [41,42]. We subsequently determined that this was due to a lack of any sequences in the Greengenes reference being annotated to the genus *Aeromonas*, with the lowest taxonomic rank being the family *Aeromonadaceae*. After communicating this and other findings to the Greengenes curators, an updated reference taxonomy was released (Greengenes 2013–08) that included additional genus and species level annotations compared to the previous release. However, even after performing taxonomic classifications with this updated reference only one OTU, representing less than 0.2% of all sequences in the V4-V5 datasets, was classified as *Aeromonas* when using the RDS method while all other OTUs were classified as *Aeromonadaceae* (data not shown). It is important to note that while this classification is not technically incorrect, it is less informative about the composition of the community and can potentially lead to inaccurate conclusions in situations where *a priori* knowledge of a microbial community is unknown.

This example also highlights the need for wider community efforts to ensure the highest possible accuracy of large reference datasets such as Greengenes. As the current version of the Greengenes database comprises over 1 million individual sequences, it is extremely challenging for the manual and automated curation steps to successfully identify and remove all potential chimeric sequences and ensure accurate taxonomic assignments for all sequences in the database. While this had a noticeable effect on the taxonomic composition of the leech intestinum, it appeared to have little to no effect on the composition of the human stool, rumen and sewage samples. Our findings suggest that researchers who rely on a reference dataset, such as for OTU clustering or taxonomy assignment as we do with the RDS processing method, should take caution in the interpretation of their results.

Low levels of cross-contamination in Illumina datasets

While our results show that overall Illumina and 454 pyrosequencing produced similar alpha and beta diversity results, we did observe cases of dataset contamination that appear to be specific to Illumina of 16S amplicons. For libraries sequenced at the same time, we also observed instances of index misassignment that resulted in a small percentage of reads from one library being incorrectly assigned an index sequence corresponding to a different library. This was most apparent when we sequenced the V4-V5.Ib libraries at the same time as non-ribosomal amplicon libraries. The source of index misassignment likely arises from image analysis errors during the index sequencing phase of the run, which may be addressed by future upgrades to the MiSeq software, hardware, or reagent kits. Reducing the target cluster density for amplicon libraries below Illumina's recommended values may reduce the occurrence of this error, while also improving read quality as previously discussed by Kozich et al. [22]. The use of dual indexing formats where indices are present at both ends of the amplicon being sequenced would likely decrease the occurrence of index misassignment, as errors would need to occur in both indices in order for a read to be assigned to the incorrect sample. We also observed a low percentage of reads from the phiX control library in all of the raw Illumina datasets we used in this study. While updates to the MiSeq's RTA base-calling software have reduced the potential for phiX reads to be incorrectly assigned a valid index sequence they did not eliminate it. We removed phiX reads from the datasets prior to QIIME analysis by applying the pre-processing methods detailed above. An additional concern with Illumina sequencing that we did not directly quantify with our datasets is low-levels of carryover contamination that occurs between consecutive MiSeq runs. This issue was acknowledged in a technical bulletin from Illumina which quantified the level of carryover contamination as typically being less than 0.1% of reads for a run being carried over into and contaminating a subsequent run [33].

The combination of index misassignment occurring at a rate of ~0.06% and carryover contamination between MiSeq runs of less than 0.1% can provide a baseline value that serves as a threshold to help distinguish which results stem from true biological signal and which may be due to noise. In order to mitigate index misassignment and sample carryover contamination for experiments that require high levels of sensitivity, we have begun to include one or more indexed control samples to more accurately quantify this occurrence. These control libraries can be created from a synthetic template, pure culture, or mock community and serve as inline controls for determining the level of index misassignment that occurs between different samples within a run and carryover contamination across separate sequencing runs. It is also recommended to alternate the indices used between runs to further reduce potential carryover contamination in high-sensitivity experiments. While researchers who are primarily concerned with identifying broad changes in microbial composition will typically not be affected by index misassignment and carry-over contamination, implementing the above listed

suggestions will improve the quality and accuracy of amplicon sequencing datasets produced on Illumina instruments. Researchers focused on examining the "rare biosphere" or the role of low abundant organisms in a community may need to implement additional precautions.

Our analysis shows that primers designed for Roche 454 instruments can be readily modified for use on Illumina instruments and produce consistent results. When we utilized the same template primers, the Illumina-produced datasets were more similar to the 454-produced datasets than when different template primers were used. The consistency between platforms was further improved by using the RDS processing pipeline, maximizing the quality of the sequences by merging of the paired Illumina reads, and minimizing artifacts due to the use of reference datasets and the inclusion of chimera checking. To account for and reduce the low levels of index misassignment and carryover contamination that we observed, we recommend the use of control libraries and alternating indices between consecutive sequencing runs when using the MiSeq. Overall our results show that Illumina sequencing of 16S rRNA genes is a cost effective approach that can readily supplant 454 pyrosequencing as the new standard analysis method for microbial populations.

Supporting Information

Figure S1 Effects of processing method on PCoA analysis using the Bray-Curtis metric.
(PDF)

Figure S2 Effect of processing method and Greengenes reference on taxonomic composition of the human stool sample.
(PDF)

Table S1 Comparison of the number of OTUs and retained reads using different processing methods.
(PDF)

Table S2 Comparison of OTU filtering cut-off values.
(PDF)

File S1 Supplementary QIIME processing methods.
(DOCX)

File S2 Archive containing the processing scripts used in for this study.
(ZIP)

Acknowledgments

The authors wish to thank Vince Young for providing the human stool DNA and Urs Boelsterli for providing the mouse intestinal contents for analysis.

Author Contributions

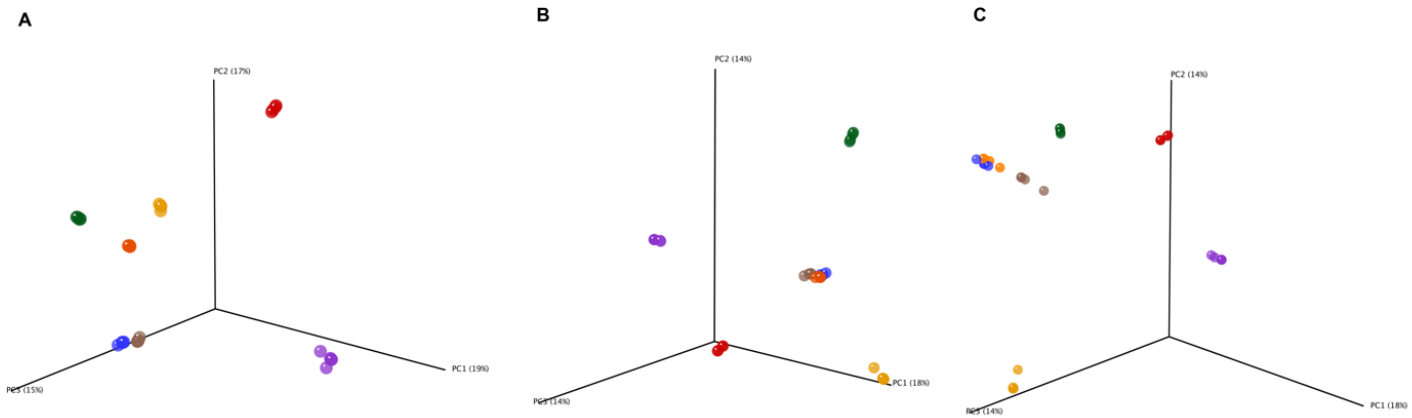
Conceived and designed the experiments: MCN HGM JG. Performed the experiments: MCN JB SLG. Analyzed the data: MCN JB. Wrote the paper: MCN HGM JB JG.

References

1. Stahl DA, Lane DJ, Olsen GJ, Pace NR (1984) Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. Science 224: 409–411.
2. Stahl DA, Lane DJ, Olsen GJ, Pace NR (1985) Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences. Appl Environ Microbiol 49: 1379–1384.
3. Rivière D, Desvignes V, Pelletier E, Chaussonnerie S, Guermazi S, et al. (2009) Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. ISME J 3: 700–714.
4. Nelson MC, Morrison M, Yu Z (2011) A meta-analysis of the microbial diversity observed in anaerobic digesters. Bioresour Technol 102: 3730–3739.
5. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, et al. (2005) Obesity alters gut microbial ecology. Proc Natl Acad Sci USA 102: 11070–11075.
6. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, et al. (2009) Bacterial community variation in human body habitats across space and time. Science 326: 1694–1697.
7. Jumpstart Consortium Human Microbiome Project Data Generation Working Group (2012) Evaluation of 16S rDNA-based community profiling for human microbiome research. PLoS ONE 7: e39315.
8. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc Natl Acad Sci USA 103: 12115–12120.
9. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457:480–484.
10. Human Microbiome Project Consortium (2012) A framework for human microbiome research. Nature 486: 215–221.
11. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8: R143.
12. Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol 12: 1889–1898.
13. Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, et al. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods 6: 639–641.
14. Reeder J, Knight R (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nat Methods 7: 668–669.
15. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30: 434–439.
16. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, et al. (2013) Updating benchtop sequencing performance comparison. Nat Biotechnol 31: 294–296.
17. Bartram AK, Lynch MDJ, Stearns JC, Moreno-Hagelsieb G, Neufeld JD (2011) Generation of Multimillion-Sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads. Appl Environ Microbiol 77: 3846–3852.
18. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, et al. (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6: 1621–1624.
19. Degnan PH, Ochman H (2011) Illumina-based analysis of microbial community diversity. ISME J 6: 183–194.
20. Gloor GB, Hummelen R, Macklaim JM, Dickson RJ, Fernandes AD, et al. (2010) Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products. PLoS ONE 5: e15406.
21. Ram JL, Karim AS, Sendler ED, Kato I (2011) Strategy for microbiome analysis using 16S rRNA gene sequence analysis on the Illumina sequencing platform. Syst Biol Reprod Med 57: 162–170.
22. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Appl Environ Microbiol 79: 5112–5120.
23. Eren AM, Vineis JH, Morrison HG, Sogin ML (2013) A filtering method to generate high quality short reads using illumina paired-end technology. PLoS ONE 8: e66643.
24. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, et al. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA 108 Suppl 1: 4516–4522.
25. Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, et al. (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res 38: e200.
26. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT (2012) Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. PLoS ONE 7: e30087.
27. Yu Z, Morrison M (2004) Improved extraction of PCR-quality community DNA from digesta and fecal samples. BioTechniques 36:808–812.
28. Marteinsson VT, Rúnarsson Á, Stefánsson A, Thorsteinsson T, Jóhannesson T, et al. (2013) Microbial communities in the subglacial waters of the Vatnajökull ice cap, Iceland. ISME J 7: 427–437.
29. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335–336.
30. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72: 5069–5072.
31. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, et al. (2011) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6: 610–618.

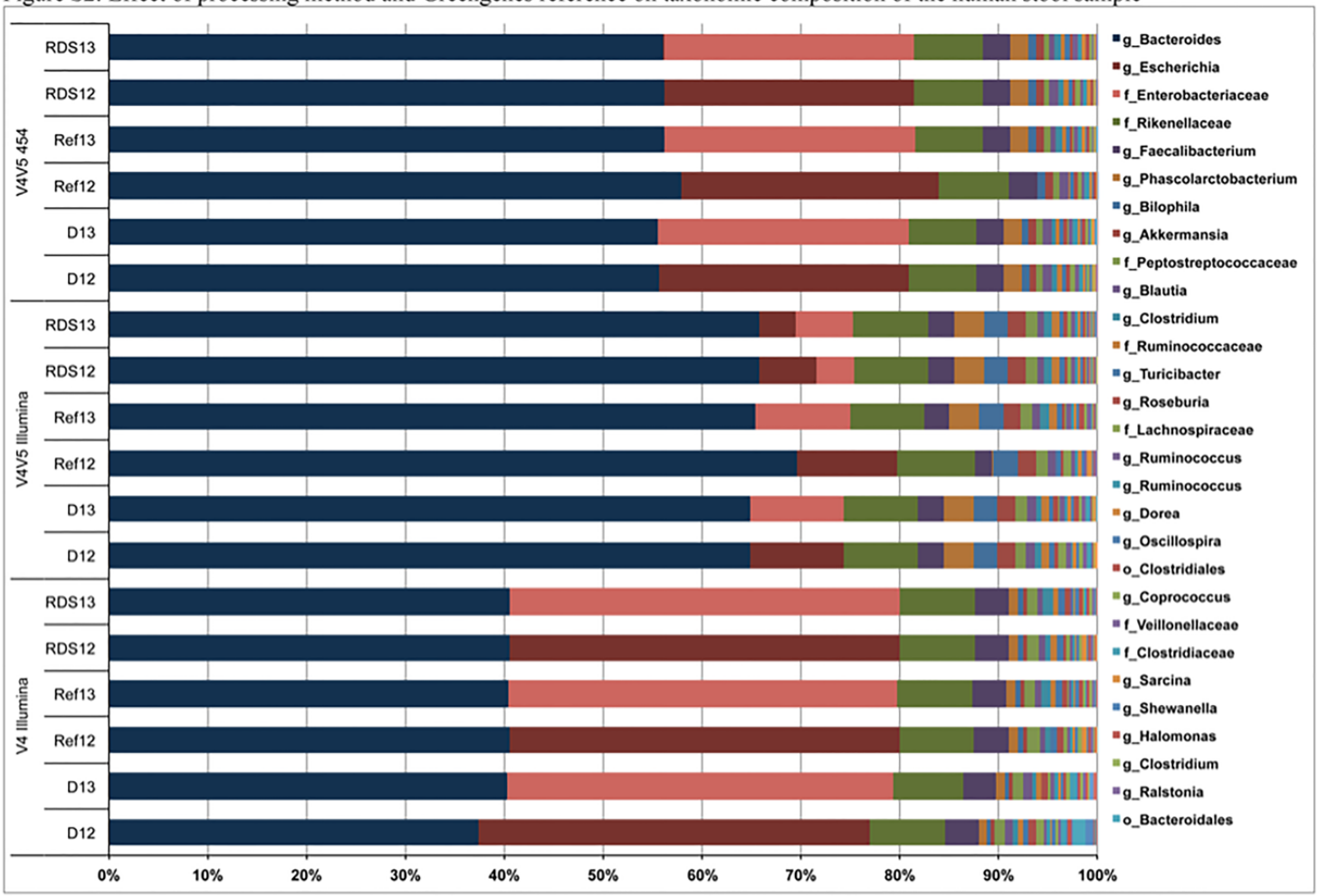
32. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, et al. (2011) Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. ISME J 6: 94–103.
33. Illumina (2013) Best Practices for High Sensitivity Applications: Minimizing Sample Carryover. Available: https://my.illumina.com/MyIllumina/Bulletin/DVzvSUldoEqh4oUyPaxoXA/best-practices-for-high-sensitivity-applications-m. Accessed 2014 Mar 18.
34. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, et al. (2012) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10: 57–59.
35. Reysenbach A-L, Giver LJ, Wickham GS, Pace NR (1992) Differential amplification of rRNA genes by polymerase chain reaction. Appl Environ Microbiol 58: 3417–3418.
36. Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, et al. (2011) The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biol Biochem 43: 1450–1455.
37. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27: 2194–2200.
38. Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, et al. (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J 4: 642–647.
39. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, et al. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 21: 494–504.
40. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188–7196.
41. Worthen PL, Gode CJ, Graf J (2006) Culture-independent characterization of the digestive-tract microbiota of the medicinal leech reveals a tripartite symbiosis. Appl Environ Microbiol 72: 4775–4781.
42. Laufer AS, Siddall ME, Graf J (2008) Characterization of the digestive-tract microbiota of Hirudo orientalis, a european medicinal leech. Appl Environ Microbiol 74:6151–6154.

Figure S1: Effects of processing method on PCoA analysis using the Bray-Curtis metric



Three dimensional principal coordinates analysis plots showing the relatedness of datasets using the Bray-Curtis metric after processing by *de novo* OTU clustering (A), RDS (B), or reference-based OTU clustering (C). Individual datasets are represented at spheres which are colored according to their sample source as follows: human stool – brown, leech intestinum – purple, mouse small intestine – orange, mock community – blue, non-adherent rumen contents – red, mixed liquor – green, termite hindgut – gold.

Figure S2: Effect of processing method and Greengenes reference on taxonomic composition of the human stool sample



Plot comparing the taxonomic composition of the termite hindgut sample for the three different library types sequenced when processed using three different methods: *de novo* (D), reference-based (Ref), and RDS. The replicate V4 and V4-V5 Illumina datasets were combined into one representative dataset for each library type. Taxonomic assignments were made using the RDP Classifier after retraining with either the 2012-10 or 2013-08 Greengenes references as denoted by the numbers 12 or 13 following the processing ID. Taxonomic ranks are noted by letters preceding the taxon name as follows: genus – g, family – f, order – o, class – c, phylum – p, domain – d.

Table S1: Comparison of the number of OTUs and retained reads using different processing methods

			de novo		Reference GG-2012		Reference GG-2013		RDS GG-2012		RDS GG-2013	
Sample Source	**Sample ID**	**Initial # Reads**	**% Reads Retained**	**# OTUs**	**% Reads Retained**	**# OTUs**	**% Reads Retained**	**# OTUs**	**% Reads Retained**	**# OTUs**	**% Reads Retained**	**# OTUs**
Human stool	H.v4.I	93769	99.65%	457	99.17%	303	99.24%	367	99.85%	564	99.92%	594
	H.v4v5.I.a	32506	99.79%	327	92.24%	138	98.63%	256	99.77%	429	99.88%	462
	H.v4v5.I.b	153159	99.98%	269	93.99%	179	99.84%	363	99.96%	406	99.99%	456
	H.v4v5.454	7882	99.96%	71	96.28%	40	99.47%	62	99.81%	70	99.92%	74
Leech intestinum	L.v4.I	118954	99.23%	301	95.19%	73	95.58%	90	99.41%	291	99.20%	284
	L.v4v5.I.a	44230	99.36%	135	63.27%	27	91.31%	42	99.24%	139	99.35%	148
	L.v4v5.I.b6	191369	99.28%	316	63.81%	80	90.07%	134	99.26%	386	99.29%	372
	L.v4v5.I.b11	171969	99.61%	275	64.61%	81	91.43%	127	99.58%	345	99.62%	329
	L.v4v5.454	10229	99.92%	23	74.89%	18	96.38%	21	99.90%	27	99.92%	26
Mouse small intestine	M.v4.I	45411	98.67%	268	96.49%	145	96.96%	166	99.08%	299	99.21%	288
	M.v4v5.I.a	24061	97.64%	188	52.39%	37	86.60%	59	97.93%	206	98.80%	206
	M.v4v5.I.b	155976	98.18%	696	58.26%	149	85.94%	286	97.82%	819	98.92%	785
	M.v4v5.454	10453	99.80%	46	74.71%	23	96.96%	32	99.24%	47	99.59%	47
HMP Mock Even	Mock.v4.I.1	213043	99.29%	1325	92.54%	152	92.97%	294	99.07%	1349	99.36%	1495
	Mock.v4.I.105	240682	98.98%	1708	90.13%	163	90.79%	312	98.73%	1766	99.10%	1933
	Mock.v4v5.I.1	2484	95.33%	449	65.26%	36	69.57%	96	95.57%	440	95.41%	472
	Mock.v4v5.I.11	90126	96.71%	2250	67.20%	64	71.27%	237	95.93%	2185	96.03%	2329
	Mock.v4v5.454	7386	100.00%	39	98.82%	28	98.86%	32	99.80%	50	99.96%	52
Rumen content	R.v4.I	93881	92.61%	11215	64.78%	1700	70.71%	1930	94.77%	12457	94.33%	11413
	R.v4v5.I.a	44431	93.03%	7051	42.29%	702	67.28%	1542	94.43%	8241	93.40%	7414
	R.v4v5.I.b	217371	91.35%	23398	37.86%	908	62.42%	2236	92.73%	27493	91.40%	25385
	R.v4v5.454	35527	98.11%	1339	55.68%	278	82.88%	584	98.52%	1474	97.81%	1413
Municipal sewage	S.v4.I	117562	93.03%	9628	57.94%	1797	62.21%	2271	95.01%	10605	94.83%	10271
	S.v4v5.I.a	28971	93.71%	3925	35.11%	644	59.29%	1352	95.01%	4526	94.95%	4496
	S.v4v5.I.b	160654	91.65%	16167	32.45%	961	54.70%	2534	93.05%	18573	92.78%	18629
	S.v4v5.454	38227	94.72%	2315	41.39%	292	65.11%	637	95.63%	2463	95.68%	2460
Termite hindgut	T.v4.I	124664	94.89%	2040	79.78%	235	85.59%	307	96.31%	2181	97.48%	2095
	T.v4v5.I.a	31220	98.72%	853	36.37%	117	85.73%	219	98.46%	960	99.17%	854
	T.v4v5.I.b	164780	98.67%	1949	31.09%	138	87.32%	323	98.30%	2111	99.02%	2087
	T.v4v5.454	7146	98.46%	193	40.39%	63	90.61%	139	98.10%	193	98.46%	196

Table S2: Comparison of OTU filtering cut-off values

Sample Source	Library	Raw	0.001%	0.005%	0.01%	0.05%	0.1%	0.5%
Human stool	H.v4.I	594	155	74	53	24	20	8
	H.v4v5.I.a	462	151	82	50	25	19	8
	H.v4v5.I.b	456	187	102	70	37	27	12
	H.v4v5.454	74	61	44	31	17	12	6
Leech intestinum	L.v4.I	284	66	36	31	18	12	5
	L.v4v5.I.a	148	72	37	30	17	9	6
	L.v4v5.I.b6	372	149	95	78	52	34	16
	L.v4v5.I.b11	329	126	77	66	41	27	14
	L.v4v5.454	26	25	21	20	16	12	8
HMP Mock Even	Mock.v4.I.1	1495	427	172	106	38	29	14
	Mock.v4.I.105	1933	430	167	103	37	28	14
	Mock.v4v5.I.1	472	245	124	76	30	23	11
	Mock.v4v5.I.11	2329	476	169	100	35	27	14
	Mock.v4v5.454	52	49	39	32	23	19	11
Mouse small intestine	M.v4.I	288	84	50	39	20	17	13
	M.v4v5.I.a	206	79	47	35	16	9	7
	M.v4v5.I.b	785	250	180	156	75	48	19
	M.v4v5.454	47	39	30	22	12	7	5
Rumen content	R.v4.I	11413	1258	392	214	43	21	2
	R.v4v5.I.a	7414	1308	410	216	38	17	3
	R.v4v5.I.b	25385	1403	446	247	59	31	12
	R.v4v5.454	1413	622	302	181	38	18	1
Municipal sewage	S.v4.I	10271	1006	336	197	47	28	10
	S.v4v5.I.a	4496	1006	338	194	42	20	6
	S.v4v5.I.b	18629	1153	396	236	69	40	17
	S.v4v5.454	2460	708	310	191	53	30	10
Termite hindgut	T.v4.I	2095	333	161	120	51	34	14
	T.v4v5.I.a	854	302	147	110	43	27	9
	T.v4v5.I.b	2087	358	178	137	57	37	11
	T.v4v5.454	196	166	119	89	36	22	7

Supplementary methods:

Creation of the V4-V5 specific Greengenes reference files:

Werner et al. [32] previously demonstrated that the use of reference sequences corresponding to the sequenced hyper-variable region(s) improved the accuracy of reference-based methods such as sequence alignment and taxonomic classification. To create a region specific reference for this study we took the NAST aligned 97% clustered reference OTUs of the Greengenes database and located the position of the forward and reverse primers within the alignment. Using a custom perl script, we cut the alignment based on these positions to excise the V4-V5 region from the full alignment, including 15 base-pairs up and down stream to ensure the sequenced amplicons would lie within the reference. Common gap columns were removed from the extracted alignment to reduce the size of the aligned V4-V5 reference sequence file, which improves computational efficiency for reference-based alignment and chimera checking. Alignment characters were stripped from the aligned V4-V5 reference file to produce the unaligned set of reference sequences that were used for reference-based OTU clustering and training of the RDP Classifier. This was done initially for the 2012-10 release and subsequently for the 2013-08 release when it was made available.

Description of QIIME analysis pipelines:

Multiple QIIME processing methods were analyzed in order to determine an analysis pipeline that optimized both accuracy and processing efficiency of large Illumina datasets. Final versions of the shell scripts that include the OTU filtering step that we used for running the *de novo* (denovo.sh), reference-based (Ref.sh), and RDS (RDS.sh) pipelines are provided as part of the supplementary material as a single compressed archive (.zip) file (File S2).

A brief description of each pipeline follows.

De novo processing pipeline: For our initial processing with *de novo* OTU clustering, the combined demultiplexed sequence dataset was clustered using uclust at the 97% similarity level. Representative sequences for each OTU were chosen and then aligned with NAST against the aligned V4-V5 reference file. The aligned representative sequences were then chimera checked using ChimeraSlayer against the aligned V4-V5 Greengenes reference described above. Any OTUs for which the representative sequence failed alignment or were marked as potential chimeras were excluded from further analysis. Taxonomic assignment of OTUs was made based on the representative sequence using the RDP Classifier, after retraining it with the unaligned V4-V5 Greengenes reference and the appropriate taxonomy file as recommended by Werner et al. [32] A phylogenetic tree was constructed using FastTree as implemented in QIIME from the aligned OTU representative sequences.

Reference-based processing pipeline: Reference-based OTU clustering was carried out using the parallel implementation of uclust_ref as implemented in QIIME and the unaligned V4-V5 Greengenes reference. Representative sequences for each OTU were chosen from the aligned Greengenes reference sequence file and alignment characters were stripped to create an unaligned set or representative sequences. No chimera checking was performed on the reference-based OTUs under the assumption that the reference sequences were non-chimeric. Taxonomic assignments were made similarly to the method used for the *de novo* OTU processing pipeline.

Reference plus *de novo* OTU clustering with chimera checking (RDS): As a large number of sequences failed to be assigned to a reference OTU, we analyzed a third processing pipeline that combined parallel reference-based OTU clustering with *de novo* OTU clustering of the reads that were not assigned to a reference OTU. The first step of this method conducts parallel OTU

clustering with uclust_ref against the V4-V5 Greengenes reference. Representative sequences for the reference-clustered OTU were selected from the aligned V4-V5 Greengenes reference, while all sequences that failed to be assigned to a reference OTU were then collected into a new file for *de novo* OTU picking. Representative sequences for the *de novo* OTUs were selected from the collected sequences unassigned to a reference OTU and aligned using NAST against the V4-V5 Greengenes reference and chimera checked using ChimeraSlayer as described above for As for the *de novo* pipeline, any *de novo* OTUs for which the representative sequence failed alignment were excluded from further analysis. The results of the reference-based and *de novo* OTU clustering steps were merged to create a single, unified set of OTUs and aligned and un-aligned representative sequences. Taxonomy assignment and phylogenetic tree construction were conducted as described for the *de novo* pipeline.

For each processing method, beta diversity analyses were conducted after first normalizing the filtered OTU table to the smallest dataset in the study, excluding the V4V5.I.1 mock community dataset. Per sample analyses of each microbial community were conducted after creating sample specific OTU tables from the original filtered OTU table. The sample specific OTU tables were then normalized to the smallest dataset of each sample, again excluding the V4V5.I.1 mock community dataset.

Appendix II

The Termite Hindgut Microbiota in Response to a Feeding-Starvation Cycle

Contributions from other researchers

Daniel Golden, an undergraduate researcher in the Graf lab, sequenced the hindgut samples from a termite feeding experiment that lead to appendix II.

Introduction

The termite, *Reticulitermes flavipes*, feeds exclusively on a diet of recalcitrant wood and relies on its hindgut microbiome for aid in digestion and provision of nutrients. The hindgut microbiota is made up of eukaryotic protists, bacteria, and archaea (Ohkuma, 2003). It has been shown that when termites do not feed (in the case of caste molting), they lose their gut protists and protist-associated bacteria (Benjamino & Graf, 2016; Cleveland, 1925). In chapter three of this thesis, it was shown that starved termites lost the majority of the dominant bacterial symbionts, *Endomicrobia* and *Treponema* and lower abundant taxa were becoming more prevalent over time. When plotted in an NMDS plot, it was shown that the starved bacterial community was different from the bacterial communities of other colonies, which were fed (chapter 3). This study was performed using three colony-types, constantly fed termites, constantly starved termites, and rescued termites (fed and starved). We hypothesize that when termites are starved, their microbiota will change, but that we could recover the microbiota by reintroducing a food source to the colony.

Methods

A single termite colony purchased from Connecticut Valley Biological Supply Co. was split into three colonies based on treatment, Fed control, Starved control, and Rescued. Hindguts were sampled immediately after delivery to develop a starting point for the experiment. The fed control was maintained on a spruce block and the starved control was maintained on autoclaved sand only. The rescued colony was immediately starved until day 3, when a block of spruce was introduced to the colony. Termites were sampled on day 7 and 14 (while fed) then the wood was removed on day 14. Termites were sampled on day 16 (starved) and then wood was introduced

on day 17. Termites were finally sampled on days 21 and 28. Fed and starved termites were sampled each time the rescued termites were sampled. 16S rRNA sequence libraries were prepared using the V4 hypervariable region according to the protocol in chapter 2 (Benjamino & Graf, 2016), sequenced on an Illumina MiSeq, and analyzed using Qiime.

Results and Discussion

The goal of this project was to determine whether feeding a starved colony of termites would recover the hindgut microbiota to that of constantly fed termites. The fed control group and the rescued experimental group was sampled through day 28, while the starved group was only sampled through day 10 (due to termite die-off). The graph of the OTU count in each control group (fed and starved) along with the rescued group separated into the days they were fed or starved (Figure 1). OTUs with only one read per the entire dataset were excluded due to the possibility of them being a sequencing error. Although there is no statistically significant different in the rescued microbiota between being fed or starved, there is a slight pattern where the mean OTUs in starved termites is lower than fed termites. The lower number of OTUs in the starved termites coincides with a simpler community, where the bacteria die off and are expelled from the hindgut due to the lack of a nutrient source.

A Bray-Curtis PCoA plot was created to show different OTUs and their abundances (Figure 2). The fed control termites are colored in blue, starved in yellow, and rescued in green. The rescued samples when fed are shown by a filled circle and the samples when starved are designated by an empty circle. The fed and starved controls are different from each other while the samples from rescued group lie amongst the fed and the starved control. The rescued termites when fed seem to be more like the fed samples, whereas the rescued termites when starved seem

more similar to the starved control samples. This shows that adding a food source to a colony after three days of starvation is sufficient to restore the hindgut microbiota to that of the constantly fed microbiota. These results are promising, but since it is a small study it would be advisable to replicate this experiment to determine if the results are sound.

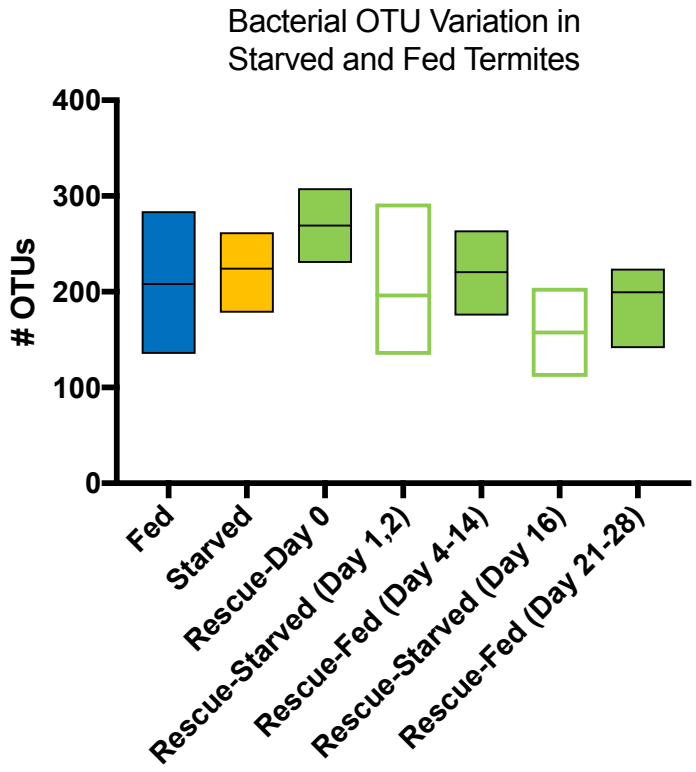


Figure 1. Bacterial OTU variation in starved and fed termites. Each symbol on the graph is representative of the number of OTUs in a single hindgut sample. Fed and starved control groups are plotted in blue and yellow, respectively. Rescued samples are designated by fed (filled in) and starved (empty). Although the differences between the groups are not significant, there is a pattern of starved experimental termites having a lower OTU mean than the fed experimental termites.

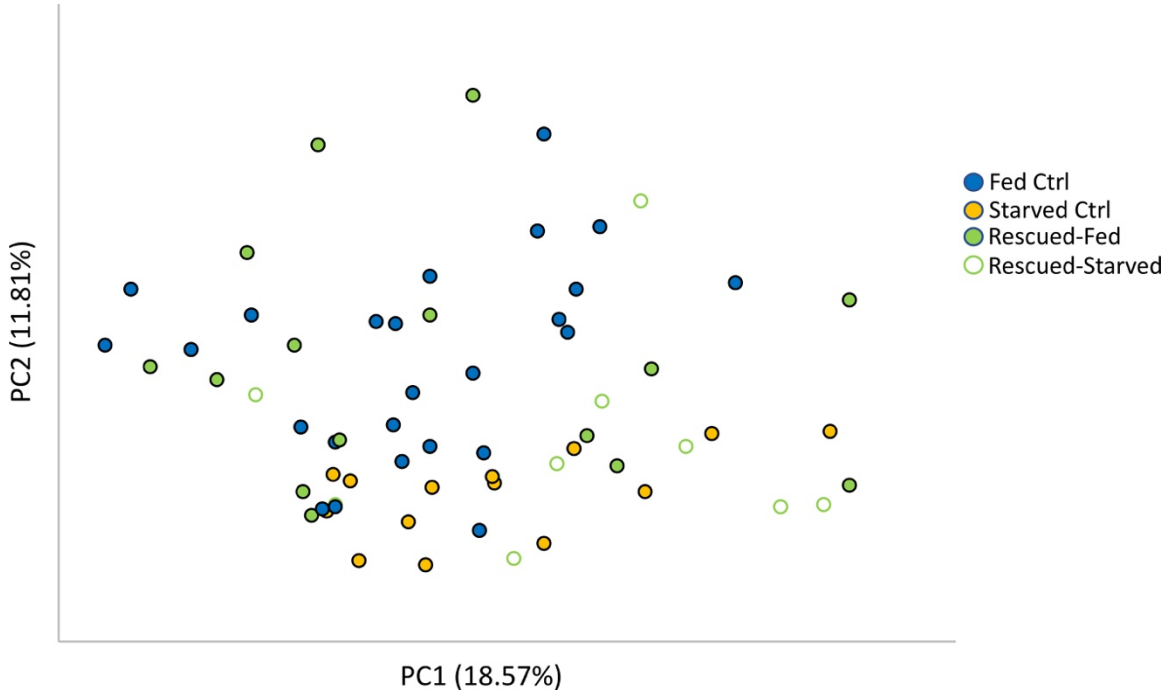


Figure 2. The hindgut microbiota is rescued when starved termites are given a food source. PCoA plot of the Bray-Curtis beta diversity metric was performed on fed control (blue), starved control (yellow) and rescued experimental (green) termites. The rescued group is separated into days when the termites were sampled after being starved (empty circles) and when termites were sampled after being fed (filled circles). While the fed and starved controls show separation, the rescued samples fall amongst the fed controls (when fed) and amongst the starved controls (when starved).

Appendix III

Metagenome and Metatranscriptome Sequencing of Five Isolated Protist Species and their Associated Bacteria

Contributions from other researchers

Michael Stephens isolated the protist cells used in this experiment. Dr. Daniel Gage quality filtered and trimmed the raw reads.

Introduction

It is estimated that over ten species of protists exist in the hindgut of the termite, *Reticulitermes flavipes*. The protists are associated with endosymbionts living inside the protist, and ectosymbionts living attached to the protist. The enzymatic breakdown of the wood meal by the protists is crucial to the survival of the termite (Tartar et al., 2009). The metabolism of cellulosic and lignocellulosic particles by the protist is followed by the metabolism of the associated bacteria. In Chapter four of this thesis, the metabolism of two endosymbionts from *Pyrsonympha vertens* and *Trichonympha agilis* was reported. This study also produced metagenome and metatranscriptome libraries from three other protist species: *Dinenympha species* (DS), *Dinenympha fimbriata* (DF), and *Dinenympha gracilis* (DG).

Methods

Four protist cells for each species were isolated and the DNA and cDNA was simultaneously amplified with random hexamer primers using the Repli-g WGA/WTA kit (Qiagen) according to standard protocol. For one protist from each species, oligo-dT primers were used in the amplification of cDNA to select for protist transcripts. Metagenomes were sheared to 550bp using a Covaris M220 ultra-sonicator™ and metatranscriptomes were sheared to 350bp. WGA and WTA samples from *D. gracilis* were too low for standard shearing and library preparation. These samples were sheared to 200bp and libraries were prepared using the NEBNext Ultra II Library Prep Kit for Illumina. Sequencing libraries were created using the Illumina TruSeq Nano DNA Library Prep Kit and sequenced on three runs of an Illumina NextSeq (2x150bp), yielding 290.09 Gbp of sequence data.

Raw Illumina reads were combined for each sample and paired. Illumina adapters and homopolymeric runs were trimmed from the paired reads, and a quality filtering using a Phred score of q15 was applied. Contaminating sequences were removed by mapping the raw reads to the genomes of possible contaminants. Ribosomal RNA (rRNA) reads were removed from the transcriptome datasets.

Results and Discussion

Thirty-five libraries in total were sequenced, fifteen metagenomes, fifteen metatranscriptomes, and five metatranscriptomes enriched with oligo-dT primers. Overall, the WGA and WTA libraries of each isolated protists yielded a high number of reads, with few samples being unusable. *Dinenympha gracilis* samples yielded the lowest number of reads and additional sequencing would be required for analysis. Two metagenomes resulted in high-quality draft genomes and transcriptome analysis was performed on four WTA samples from *Pyrsonympha vertens* and *Trichonympha agilis* (chapter 4). Draft genome assembly of the remaining samples and metatranscriptome analysis of the associated WTA samples will provide insight into the metabolic capability of these protists and their symbionts. Utilizing the oligo-dT enriched WTA samples will allow us to perform the first known gene expression analysis of the protists in the hindgut specifically.

Table. 1 Sequencing output of thirty-five samples from five protist species.

Protist	Library	Sample	# Paired Reads	Quality Trimming	Contamination Removal	rRNA Removal	% of Starting Reads
Dinenympha fimbriata	WGA	DF1	24,327,156	20,526,026	20,341,334		84
		DF6	28,573,162	20,856,604	20,443,478		72
		DF11	26,935,704	20,262,564	19,192,978		71
	WTA	DF1	144,354,524	116,909,278	97,347,314	48,381,512	34
		DF6	168,629,078	134,963,478	93,078,276	17,833,322	11
		DF11	107,759,180	70,474,682	53,567,220	7,964,358	7
		DF2	4,425,896	3,481,778	2,731,392	2,302,170	52
Dinenympha gracilis	WGA	DG1	21,262,596	8,401,672	7,700,350		36
		DG6	30,055,658	15,658,184	14,831,012		49
		DG11	27,212,990	19,535,944	18,585,356		68
	WTA	DG1	16,574,502	303,264	162,978	151,098	1
		DG6	7,127,242	25,208	23,742	20,428	0
		DG11	3,936	2,422	2,214	1,128	29
		DG2	14,329,560	651,768	549,878	546,230	4
Dinenympha species	WGA	DS1	20,232,284	15,102,444	15,072,736		74
		DS6	22,425,836	18,313,298	17,830,236		80
		DS12	30,804,474	23,312,968	23,261,694		76
	WTA	DS1	165,688	2,800	2,372	1,602	1
		DS6	128,178	3,146	2,690	1,974	2
		DS12	6,898,778	4,697,524	4,687,350	4,663,632	68
		DS2	907,440	661,104	643,310	638,462	70
Pyrsonympha vertens	WGA	PV1	29,366,568	24,807,098	24,559,780		84
		PV7	24,358,298	15,208,376	14,968,148		61
		PV11	26,373,686	21,907,976	16,157,802		61
	WTA	PV1	107,226,450	89,264,890	71,971,984	25,477,526	24
		PV7	13,608,006	1,070,048	905,076	446,406	3
		PV11	85,032,418	67,769,648	46,451,474	18,239,734	21
		PV2	3,307,452	2,735,302	1,863,148	1,765,332	53
Trichonympha agilis	WGA	TA16	31,040,896	23,612,512	16,311,954		53
		TA21	27,586,758	22,562,532	22,457,024		81
		TA26	25,222,580	18,193,434	17,692,692		70
		Enuc	4,282,992		2,310,304		54
		Nuc	4,074,550		2,161,744		53
	WTA	TA16	229,365,166	181,781,634	155,473,548	121,418,504	53
		TA21	144,520,930	128,477,248	114,311,390	7,697,882	5
		TA26	129,047,734	109,173,508	105,113,224	10,757,040	8
		TA17	4,983,122	3,530,500	2,545,730	1,346,328	27
		Enuc	8,486,798		4,544,876	1,471,410	17
		Nuc	8,147,202		3,969,944	240,910	3
Totals			1,609,165,468	1,204,240,862	1,033,827,752	545,245,590	

References

Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M., & Aksoy, S. (2002). Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nature Genetics*, *32*(3), 402–407. http://doi.org/10.1038/ng986

Aksoy, S. (2000). Tsetse – A Haven for Microorganisms. *Parasitology Today*, *16*(3), 114–118. http://doi.org/10.1016/S0169-4758(99)01606-3

Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, *12*(1), 14648. http://doi.org/10.1186/1471-2164-12-402

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. http://doi.org/10.1016/S0022-2836(05)80360-2

Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., et al. (2011). Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecology Letters*, *14*(1), 19–28. http://doi.org/10.1111/j.1461-0248.2010.01552.x

Banning, N., Brock, F., Fry, J. C., Parkes, R. J., Hornibrook, E. R. C., & Weightman, A. J. (2005). Investigation of the methanogen population structure and activity in a brackish lake sediment. *Environmental Microbiology*, *7*(7), 947–960. http://doi.org/10.1111/j.1462-2920.2004.00766.x

Baumann, P. (2005). Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. *Annual Review of Microbiology*, *59*(1), 155–189. http://doi.org/10.1146/annurev.micro.59.030804.121041

Benjamino, J., & Graf, J. (2016). Characterization of the Core and Caste-Specific Microbiota in the Termite, *Reticulitermes flavipes*. *Frontiers in Microbiology*, *7*(e1002226), 1–14. http://doi.org/10.3389/fmicb.2016.00171

Boetzer, M., Henkel, C. V., Jansen, H. J., & Butler, D. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, *27*.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*, 2114–2120. http://doi.org/10.1093/bioinformatics/btu170/-/DC1

Bomar, L., & Graf, J. (2012). Investigation into the physiologies of *Aeromonas veronii* in vitro and inside the digestive tract of the medicinal leech using RNA-seq. *The Biological Bulletin*, *223*(1), 155–166.

Bordenstein, S. R., & Theis, K. R. (2015). Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes. *PLoS Biology*, *13*(8), e1002226–23. http://doi.org/10.1371/journal.pbio.1002226

Boucias, D. G., Cai, Y., Sun, Y., Lietze, V.-U., Sen, R., Raychoudhury, R., & Scharf, M. E. (2013). The hindgut lumen prokaryotic microbiota of the termite *Reticulitermes flavipes* and its responses to dietary lignocellulose composition. *Molecular Ecology*, *22*(7), 1836–1853. http://doi.org/10.1111/mec.12230

Bourguignon, T., Lo, N., Cameron, S. L., ŠOBOTNÍK, J., Hayashi, Y., Shigenobu, S., et al. (2015). The evolutionary history of termites as inferred from 66 mitochondrial genomes. *Molecular Biology and Evolution*, *32*(2), 406–421. http://doi.org/10.1093/molbev/msu308

Brauman, A., Dore, J., Eggleton, P., Bignell, D. E., Breznak, J. A., & Kane, M. D. (2001). Molecular phylogenetic profiling of prokaryotic communities in guts of termites with different feeding habits. *FEMS Microbiology Ecology*, *35*(1), 27–36. http://doi.org/10.1111/j.1574-6941.2001.tb00785.x

Brauman, A., Kane, M. D., Labat, M., & Breznak, J. A. (1992). Genesis of acetate and methane by gut bacteria of nutritionally diverse termites. *Science*, *257*(5075), 1384–1387. http://doi.org/10.1126/science.257.5075.1384

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., et al. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, *5*(1), 1135–6. http://doi.org/10.1038/srep08365

Broderick, N. A., Raffa, K. F., Goodman, R. M., & Handelsman, J. (2004). Census of the Bacterial Community of the Gypsy Moth Larval Midgut by Using Culturing and Culture-Independent Methods. *Applied and Environmental Microbiology*, *70*(1), 293–300. http://doi.org/10.1128/AEM.70.1.293-300.2004

Brune, A. (2014). Symbiotic digestion of lignocellulose in termite guts. *Nature Reviews Microbiology*, *12*(3), 168–180. http://doi.org/10.1038/nrmicro3182

Brune, A., & Dietrich, C. (2015). The Gut Microbiota of Termites: Digesting the Diversity in the Light of Ecology and Evolution. *Annual Review of Microbiology*, *69*, 145–166. http://doi.org/10.1146/annurev-micro-092412-155715

Brune, A., Emerson, D., & Breznak, J. A. (1995a). The Termite Gut Microflora as an Oxygen Sink: Microelectrode Determination of Oxygen and pH Gradients in Guts of Lower and Higher Termites. *Applied and Environmental Microbiology*, *61*(7), 2681–2687.

Brune, A., Miambi, E., & Breznak, J. A. (1995b). Roles of oxygen and the intestinal microflora in the metabolism of lignin-derived phenylpropanoids and other monoaromatic compounds by termites. *Applied and Environmental Microbiology*, *61*(7), 2688–2695.

Bushnell, B. (2014). BBMap: A Fast, Accurate, Splice-Aware Aligner. *Berkeley National Laboratory*.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Presented at the PNAS. http://doi.org/10.1073/pnas.1000080107/-/DCSupplemental

Clawson, M. L., Bourret, A., & Benson, D. R. (2004). Assessing the phylogeny of Frankia-actinorhizal plant nitrogen-fixing root nodule symbioses with *Frankia* 16S rRNA and glutamine synthetase gene sequences. *Molecular Phylogenetics and Evolution*, *31*(1), 131–138. http://doi.org/10.1016/j.ympev.2003.08.001

Cleveland, L. R. (1925). The feeding habit of termite castes and its relation to their intestinal flagellates. *The Biological Bulletin*, *48*(5), 295–308.

Costa-Leonardo, A. M., Laranjo, L. T., Janei, V., & Haifig, I. (2013). The fat body of termites: functions and stored materials. *Journal of Insect Physiology*, *59*(6), 577–587. http://doi.org/10.1016/j.jinsphys.2013.03.009

Dadd, R. H., & Krieger, D. L. (1968). Dietary amino acid requirements of the aphid, *Myzus persicae*. *Journal of Insect Physiology*, *14*(6), 741–764. http://doi.org/10.1016/0022-1910(68)90186-8

Davison, A., & Blaxter, M. (2005). Ancient Origin of Glycosyl Hydrolase Family 9 Cellulase Genes. *Molecular Biology and Evolution*, *22*(5), 1273–1284.

http://doi.org/10.1093/molbev/msi107
Deevong, P., Hattori, S., Yamada, A., Trakulnaleamsai, S., Ohkuma, M., Noparatnaraporn, N., & Kudo, T. (2004). Isolation and Detection of Methanogens from the Gut of Higher Termites. *Microbes and Environments / JSME*, *19*(3), 221–226. http://doi.org/10.1264/jsme2.19.221
DeSantis, T. Z., Hugenholtz, P., Andersen, G. L., Larsen, N., Rojas, M., Brodie, E. L., et al. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072. http://doi.org/10.1128/AEM.03006-05
Dietrich, C., Köhler, T., & Brune, A. (2014). The cockroach origin of the termite gut microbiota: patterns in bacterial community structure reflect major evolutionary events. *Applied and Environmental Microbiology*, *80*(7), 2261–2269. http://doi.org/10.1128/AEM.04206-13
Dillon, R. J., & Dillon, V. M. (2004). The gut bacteria of insects: nonpathogenic interactions. *Annual Review of Entomology*, *49*(1), 71–92. http://doi.org/10.1146/annurev.ento.49.061802.123416
Dohlen, von, C. D., Kohler, S., Alsop, S. T., & McManus, W. R. (2001). Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. *Nature*, *412*(6845), 433–436. http://doi.org/10.1038/35086563
Douglas, A. E. (1998). Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria *Buchnera*. *Annual Review of Entomology*, *43*(1), 17–37. http://doi.org/10.1146/annurev.ento.43.1.17
Ebert, A., & Brune, A. (1997). 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.
Faith, D. P., & Baker, A. M. (2006). Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evolutionary Bioinformatics Online*, *2*, 121–128.
Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., et al. (2013). The long-term stability of the human gut microbiota. *Science*, *341*(6141), 1237439. http://doi.org/10.1126/science.1237439
Fausett, L. V. (1994). Fundamentals of neural networks.
Fisher, M., Miller, D., Brewster, C., Husseneder, C., & Dickerman, A. (2007). Diversity of gut bacteria of *Reticulitermes flavipes* as examined by 16S rRNA gene sequencing and amplified rDNA restriction analysis. *Current Microbiology*, *55*(3), 254–259. http://doi.org/10.1007/s00284-007-0136-8
Geissinger, O., Herlemann, D. P. R., Mörschel, E., Maier, U. G., & Brune, A. (2009). The ultramicrobacterium "*Elusimicrobium minitum*" gen. nov., sp. nov., the first cultivated representative of the termite group 1 phylum. *Applied and Environmental Microbiology*, *75*(9), 2831–2840. http://doi.org/10.1128/AEM.02697-08
Graber, J. R., Leadbetter, J. R., & Breznak, J. A. (2004). Description of *Treponema azotonutricium* sp. nov. and *Treponema primitia* sp. nov., the first spirochetes isolated from termite guts. *Applied and Environmental Microbiology*, *70*(3), 1315–1320. http://doi.org/10.1128/AEM.70.3.1315-1320.2004
Gupta, N. (2013). Artificial Neural Network. *Network and Complex Systems*, *3*, 24–28.
Gupta, P., Samant, K., & Sahu, A. (2012). Isolation of cellulose-degrading bacteria and determination of their cellulolytic potential. *International Journal of Microbiology*, *2012*(6), 578925–5. http://doi.org/10.1155/2012/578925
Gündüz, E. A., & Douglas, A. E. (2009). Symbiotic bacteria enable insect to use a nutritionally

inadequate diet. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1658), 987–991. http://doi.org/10.1098/rspb.2008.1476
Handelsman, J. (2005). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*. http://doi.org/10.1128/MMBR.69.1.195.2005
Hassan, A., Schrapf, N., Ramadan, W., & Tilp, M. (2017). Evaluation of tactical training in team handball by means of artificial neural networks. *Journal of Sports Sciences*, *35*(7), 642–647. http://doi.org/10.1080/02640414.2016.1183804
Hongoh, Y., Ekpornprasit, L., Inoue, T., Moriya, S., Trakulnaleamsai, S., Ohkuma, M., et al. (2005). Intracolony variation of bacterial gut microbiota among castes and ages in the fungus-growing termite *Macrotermes gilvus*. *Molecular Ecology*, *15*(2), 505–516. http://doi.org/10.1111/j.1365-294X.2005.02795.x
Hongoh, Y., Sharma, V. K., Prakash, T., Noda, S., Taylor, T. D., Kudo, T., et al. (2008a). Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Pnas*, *105*(14), 5555–5560. http://doi.org/10.1073/pnas.0801389105
Hongoh, Y., Sharma, V. K., Prakash, T., Noda, S., Toh, H., Taylor, T. D., et al. (2008b). Genome of an endosymbiont coupling N2 fixation to cellulolysis within protist cells in termite gut. *Science*, *322*(5904), 1108–1109. http://doi.org/10.1126/science.1165578
Huang, J., Sheng, X., He, L., Huang, Z., Wang, Q., & Zhang, Z. (2013a). Characterization of depth-related changes in bacterial community compositions and functions of a paddy soil profile. *Fems Microbiology Letters*, *347*(1), 33–42. http://doi.org/10.1111/1574-6968.12218
Huang, X.-F., Bakker, M. G., Judd, T. M., Reardon, K. F., & Vivanco, J. M. (2013b). Variations in diversity and richness of gut bacterial communities of termites (*Reticulitermes flavipes*) fed with grassy and woody plant substrates. *Microbial Ecology*, *65*(3), 531–536. http://doi.org/10.1007/s00248-013-0219-y
Hugouvieux-Cotte-Pattat, N., & Robert-Baudouy, J. (1987). Hexuronate catabolism in *Erwinia chrysanthemi*. *Journal of Bacteriology*, *169*(3), 1223–1231. http://doi.org/10.1128/jb.169.3.1223-1231.1987
Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214. http://doi.org/10.1038/nature11234
Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*. http://doi.org/10.1109/csx.2007.9.issue-3;subPage:string:Access
Hyodo, F., Tayasu, I., Inoue, T., Azuma, J. I., kudo, T., & abe, T. (2003). Differential role of symbiotic fungi in lignin degradation and food provision for fungus-growing termites (*Macrotermitinae*: Isoptera). *Functional Ecology*, *17*(2), 186–193. http://doi.org/10.1046/j.1365-2435.2003.00718.x
Ikeda-Ohtsubo, W., Strassert, J. F. H., Köhler, T., Mikaelyan, A., Gregor, I., McHardy, A. C., et al. (2016). "Candidatus *Adiutrix intracellularis*," an endosymbiont of termite gut flagellates, is the first representative of a deep-branching clade of Deltaproteobacteria and a putative homoacetogen. *Environmental Microbiology*, *18*(8), 2548–2564. http://doi.org/10.1111/1462-2920.13234
Inoue, J.-I., Noda, S., Hongoh, Y., Ui, S., & Ohkuma, M. (2008). Identification of Endosymbiotic Methanogen and Ectosymbiotic Spirochetes of Gut Protists of the Termite *Coptotermes formosanus*. *Microbes and Environments / JSME*, *23*(1), 94–97.
Jami, E., & Mizrahi, I. (2012). Composition and Similarity of Bovine Rumen Microbiota across Individual Animals. *PLoS ONE*, *7*(3), e33306. http://doi.org/10.1371/journal.pone.0033306
Kane, M. D., & Breznak, J. A. (1991). Effect of host diet on production of organic acids and

methane by cockroach gut bacteria. *Applied and Environmental Microbiology*.
Kapheim, K. M., Rao, V. D., Yeoman, C. J., Wilson, B. A., White, B. A., Goldenfeld, N., & Robinson, G. E. (2015). Caste-specific differences in hindgut microbial communities of honey bees (*Apis mellifera*). *PLoS ONE*, *10*(4), e0123911. http://doi.org/10.1371/journal.pone.0123911
Karl, Z. J., & Scharf, M. E. (2015). Effects of Five Diverse Lignocellulosic Diets on Digestive Enzyme Biochemistry in the Termite *Reticulitermes flavipes*. *Archives of Insect Biochemistry and Physiology*, *90*(2), 89–103. http://doi.org/10.1002/arch.21246
Kelley, D. R., Schatz, M. C., & Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome Biology*, *11*(11), R116. http://doi.org/10.1186/gb-2010-11-11-r116
Kluepfel, D. (1988). Screening of prokaryotes for cellulose- and hemicellulose-degrading enzymes. In *Biomass Part A: Cellulose and Hemicellulose* (Vol. 160, pp. 180–186). Elsevier. http://doi.org/10.1016/0076-6879(88)60118-2
Klug, M. J., Holben, W. E., Harris, D., Tiedje, J. M., & Domingo, J. S. (1998). Influence of diet on the structure and function of the bacterial hindgut community of crickets. *Molecular Ecology*, *7*(6), 761–767. http://doi.org/10.1046/j.1365-294x.1998.00390.x
Kostic, A. D., Howitt, M. R., & Garrett, W. S. (2013). Exploring host–microbiota interactions in animal models and humans. *Genes & Development*, *27*(7), 701–718. http://doi.org/10.1101/gad.212522.112
Laczny, C. C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H., et al. (2015). VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, *3*(1), 1–7. http://doi.org/10.1186/s40168-014-0066-1
Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108. http://doi.org/10.1093/nar/gkm160
Larsen, P. E., Field, D., & Gilbert, J. A. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods*, *9*(6), 621–625. http://doi.org/10.1038/nmeth.1975
Lasken, R. S., & Raghunathan, A. (2005). Multiple displacement amplification from single bacterial cells.
Lazarevic, V., Whiteson, K., Huse, S. M., Hernandez, D., Farinelli, L., Østerås, M., et al. (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *Journal of Microbiological Methods*, *79*(3), 266–271. http://doi.org/10.1016/j.mimet.2009.09.012
Leadbetter, J. R., & Breznak, J. A. (1996). Physiological ecology of *Methanobrevibacter cuticularis* sp. nov. and *Methanobrevibacter curvatus* sp. nov., isolated from the hindgut of the termite *Reticulitermes flavipes*. *Applied and Environmental Microbiology*, *62*(10), 3620–3631.
Leadbetter, J. R., Crosby, L. D., & Breznak, J. A. (1998). *Methanobrevibacter filiformis* sp. nov., A filamentous methanogen from termite hindguts. *Archives of Microbiology*, *169*(4), 287–292.
Leadbetter, J. R., Schmidt, T. M., Graber, J. R., & Breznak, J. A. (1999). Acetogenesis from H2 plus CO2 by spirochetes from termite guts. *Science*, *283*(5402), 686–689.
LeCun, Y., Kanter, I., & Solla, S. A. (1991). Second order properties of error surfaces: Learning time and generalization. *Advances in Neural Information.*

Lilburn, T. G. (2001). Nitrogen Fixation by Symbiotic and Free-Living Spirochetes. *Science*, *292*(5526), 2495–2498. http://doi.org/10.1126/science.1060281
Liu, H., & Beckenbach, A. T. (1992). Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. *Molecular Phylogenetics and Evolution*, *1*(1), 41–52.
Maltz, M. A., Bomar, L., Lapierre, P., Morrison, H. G., McClure, E. A., Sogin, M. L., & Graf, J. (2014). Metagenomic analysis of the medicinal leech gut microbiota. *Frontiers in Microbiology*, *5*, 151. http://doi.org/10.3389/fmicb.2014.00151
Matsuura, K. (2001). Nestmate recognition mediated by intestinal bacteria in a termite, Reticulitermes speratus. *Oikos*.
Mattéotti, C., Haubruge, E., Thonart, P., Francis, F., De Pauw, E., Portetelle, D., & Vandenbol, M. (2011). Characterization of a new β-glucosidase/β-xylosidase from the gut microbiota of the termite (*Reticulitermes santonensis*). *Fems Microbiology Letters*, *314*(2), 147–157. http://doi.org/10.1111/j.1574-6968.2010.02161.x
McCutcheon, J. P., & Moran, N. A. (2011). Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, *224*, 1209. http://doi.org/10.1038/nrmicro2670
McDonald, D., Price, M. N., Goodrich, J. K., Nawrocki, E. P., DeSantis, T. Z., Probst, A., et al. (2011). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, *6*(3), 610–618. http://doi.org/10.1038/ismej.2011.139
Mendes, R., Kruijt, M., de Bruijn, I., Dekkers, E., van der Voort, M., Schneider, J. H. M., et al. (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science*, *332*(6033), 1097–1100. http://doi.org/10.1126/science.1203980
Mikaelyan, A., Dietrich, C., Köhler, T., Poulsen, M., Sillam-Dussès, D., & Brune, A. (2015). Diet is the primary determinant of bacterial community structure in the guts of higher termites. *Molecular Ecology*, *24*(20), 5284–5295. http://doi.org/10.1111/mec.13376
Minkley, N., Fujita, A., Brune, A., & Kirchner, W. H. (2006). Nest specificity of the bacterial community in termite guts (*Hodotermes mossambicus*). *Insectes Sociaux*. http://doi.org/10.1007/s00040-006-0878-5
Mittler, T. E. (1971). Dietary amino acid requirement of the aphid *Myzus persicae* affected by antibiotic uptake. *Journal of Nutrition*.
Nelson, M. C., Morrison, H. G., Benjamino, J., Grim, S. L., & Graf, J. (2014). Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS ONE*, *9*(4), e94249. http://doi.org/10.1371/journal.pone.0094249
Nepomuceno, R. S. L., Tavares, M. B., Lemos, J. A., Griswold, A. R., Ribeiro, J. L., Balan, A., et al. (2007). The oligopeptide (opp) gene cluster of *Streptococcus mutans*: identification, prevalence, and characterization. *Oral Microbiology and Immunology*, *22*(4), 277–284. http://doi.org/10.1111/j.1399-302X.2007.00368.x
Nevares, I., Martínez-Martínez, V., Martínez-Gil, A., Martín, R., Laurie, V. F., & Del Álamo-Sanza, M. (2017). On-line monitoring of oxygen as a method to qualify the oxygen consumption rate of wines. *Food Chemistry*, *229*, 588–596. http://doi.org/10.1016/j.foodchem.2017.02.105
Nissen, S. (2003). Implementation of a fast artificial neural network library (fann). *Report*.
Noda, S., Kitade, O., Inoue, T., Kawai, M., Kanuka, M., Hiroshima, K., et al. (2007). Cospeciation in the triplex symbiosis of termite gut protists (*Pseudotrichonympha* spp.), their hosts, and their bacterial endosymbionts. *Molecular Ecology*, *16*(6), 1257–1266. http://doi.org/10.1111/j.1365-294X.2006.03219.x

Ohkuma, M. (2003). Termite symbiotic systems: efficient bio-recycling of lignocellulose. *Applied Microbiology and Biotechnology*, *61*(1), 1–9. http://doi.org/10.1007/s00253-002-1189-z

Ohkuma, M. (2008). Symbioses of flagellates and prokaryotes in the gut of lower termites. *Trends in Microbiology*, *16*(7), 345–352. http://doi.org/10.1016/j.tim.2008.04.004

Ohkuma, M., Noda, S., Hattori, S., Iida, T., Yuki, M., Starns, D., et al. (2015). Acetogenesis from H2 plus CO2 and nitrogen fixation by an endosymbiotic spirochete of a termite-gut cellulolytic protist. *Pnas*, *112*(33), 10224–10230. http://doi.org/10.1073/pnas.1423979112

Oksanen, J., Blanchert, F. G., Kindt, R., Legendre, P., Minchin, P. R., & Ohara, R. B. (n.d.). *vegan: Community Ecology Package*. Retrieved from http://cran.r-project.org

Pais, R., Lohs, C., Wu, Y., Wang, J., & Aksoy, S. (2008). The Obligate Mutualist *Wigglesworthia glossinidia* Influences Reproduction, Digestion, and Immunity Processes of Its Host, the Tsetse Fly. *Applied and Environmental Microbiology*, *74*(19), 5965–5974. http://doi.org/10.1128/AEM.00741-08

Peixoto, T. P. (2014). The graph-tool python library. figshare.

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2010). IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. In *Research in Computational Molecular Biology* (Vol. 6044, pp. 426–440). Berlin, Heidelberg: Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-12683-3_28

Pester, M., & Brune, A. (2006). Expression profiles of fhs (FTHFS) genes support the hypothesis that spirochaetes dominate reductive acetogenesis in the hindgut of lower termites. *Environmental Microbiology*, *8*(7), 1261–1270. http://doi.org/10.1111/j.1462-2920.2006.01020.x

Quigley, E. M. M. (2017). Gut microbiome as a clinical tool in gastrointestinal disease management: are we there yet? *Nature Reviews. Gastroenterology & Hepatology*, *14*(5), 315–320. http://doi.org/10.1038/nrgastro.2017.29

Rappé, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Review of Microbiology*, *57*(1), 369–394. http://doi.org/10.1146/annurev.micro.57.030502.090759

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*.

Sato, T., Hongoh, Y., Noda, S., Hattori, S., Ui, S., & Ohkuma, M. (2009). Candidatus *Desulfovibrio trichonymphae*, a novel intracellular symbiont of the flagellate *Trichonympha agilis* in termite gut. *Environmental Microbiology*, *11*(4), 1007–1015. http://doi.org/10.1111/j.1462-2920.2008.01827.x

Scharf, M. E. (2015a). Omic research in termites: an overview and a roadmap. *Frontiers in Genetics*, *6*, 76. http://doi.org/10.3389/fgene.2015.00076

Scharf, M. E. (2015b). Termites as targets and models for biotechnology. *Annual Review of Entomology*, *60*, 77–102. http://doi.org/10.1146/annurev-ento-010814-020902

Scharf, M. E., Karl, Z. J., Sethi, A., & Boucias, D. G. (2011). Multiple Levels of Synergistic Collaboration in Termite Lignocellulose Digestion. *PLoS ONE*, *6*(7), e21709–7. http://doi.org/10.1371/journal.pone.0021709

Sethi, A., Kovaleva, E. S., Slack, J. M., Brown, S., Buchman, G. W., & Scharf, M. E. (2013a). A GHF7 cellulase from the protist symbiont community of *Reticulitermes flavipes* enables more efficient lignocellulose processing by host enzymes. *Archives of Insect Biochemistry and Physiology*, *84*(4), 175–193. http://doi.org/10.1002/arch.21135

Sethi, A., Slack, J. M., Kovaleva, E. S., Buchman, G. W., & Scharf, M. E. (2013b). Lignin-

associated metagene expression in a lignocellulose-digesting termite. *Insect Biochemistry and Molecular Biology*, *43*(1), 91–101. http://doi.org/10.1016/j.ibmb.2012.10.001
Shade, A., Peter, H., Allison, S. D., Baho, D. L., Berga, M., Bürgmann, H., et al. (2012). Fundamentals of microbial community resistance and resilience. *Frontiers in Microbiology*, *3*, 417. http://doi.org/10.3389/fmicb.2012.00417
Shang, Y., Khafipour, E., Derakhshani, H., Sarna, L. K., Woo, C. W., Siow, Y. L., & O, K. (2017). Short Term High Fat Diet Induces Obesity-Enhancing Changes in Mouse Gut Microbiota That are Partially Reversed by Cessation of the High Fat Diet. *Lipids*, *101*(Suppl 4), 15718. http://doi.org/10.1007/s11745-017-4253-2
Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., & Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, *407*(6800), 81–86. http://doi.org/10.1038/35024074
Simpson, J. T., & Durbin, R. (2010). Efficient construction of an assembly string graph using the FM-index. *Bioinformatics (Oxford, England)*, *26*(12), i367–i373. http://doi.org/10.1093/bioinformatics/btq217
Sloan, D. B., & Moran, N. A. (2012). Genome Reduction and Co-evolution between the Primary and Secondary Bacterial Symbionts of Psyllids. *Molecular Biology and Evolution*, *29*(12), 3781–3792. http://doi.org/10.1093/molbev/mss180
Tai, V., James, E. R., Nalepa, C. A., Scheffrahn, R. H., Perlman, S. J., & Keeling, P. J. (2015). The role of host phylogeny varies in shaping microbial diversity in the hindguts of lower termites. *Applied and Environmental Microbiology*, *81*(3), 1059–1070. http://doi.org/10.1128/AEM.02945-14
Tamschick, S., & Radek, R. (2013). Colonization of termite hindgut walls by oxymonad flagellates and prokaryotes in *Incisitermes tabogae*, *I. marginipennis* and *Reticulitermes flavipes*. *European Journal of Protistology*, *49*(1), 1–14. http://doi.org/10.1016/j.ejop.2012.06.002
Tang, Y. P., Dallas, M. M., & Malamy, M. H. (1999). Characterization of the BatI (Bacteroides aerotolerance) operon in *Bacteroides fragilis*: isolation of a *B. fragilis* mutant with reduced aerotolerance and impaired growth in in vivo model systems. *Molecular Microbiology*, *32*(1), 139–149. http://doi.org/10.1046/j.1365-2958.1999.01337.x
Tartar, A., Wheeler, M. M., Zhou, X., Coy, M. R., Boucias, D. G., & Scharf, M. E. (2009). Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite Reticulitermes flavipes. *Biotechnology for Biofuels*, *2*(1), 25. http://doi.org/10.1186/1754-6834-2-25
Thorne, B. L., Traniello, J. F. A., Adams, E. S., & Bulmer, M. (2010). Reproductive dynamics and colony structure of subterranean termites of the genus *Reticulitermes* (Isoptera Rhinotermitidae): a review of the evidence from behavioral, ecological, and genetic studies. *Ethology Ecology & Evolution*. http://doi.org/10.1080/08927014.1999.9522833
Tinker, K. A., & Ottesen, E. A. (2016). The Core Gut Microbiome of the American Cockroach, Periplaneta americana, Is Stable and Resilient to Dietary Shifts. *Applied and Environmental Microbiology*, *82*(22), 6603–6610. http://doi.org/10.1128/AEM.01837-16
tokura, M., Ohkuma, M., & kudo, T. (2000). Molecular phylogeny of methanogens associated with flagellated protists in the gut and with the gut epithelium of termites. *FEMS Microbiology Ecology*, *33*(3), 233–240.
Trager, W. (1934). The Cultivation of a Cellulose-Digesting Flagellate, *Trichomonas termopsidis*, and of Certain Other Termite Protozoa. *The Biological Bulletin*.

Tritt, A., Eisen, J. A., Facciotti, M. T., & Darling, A. E. (2012). An Integrated Pipeline for de Novo Assembly of Microbial Genomes. *PLoS ONE*, *7*(9), e42304–9. http://doi.org/10.1371/journal.pone.0042304

Varma, A., Kolli, B. K., Paul, J., Saxena, S., & Konig, H. (1994). Lignocellulose degradation by microorganisms from termite hills and termite guts: A survey on the present state of art. *FEMS Microbiology Reviews*, *15*(1), 9–28. http://doi.org/10.1111/j.1574-6976.1994.tb00120.x

Walker, A. W., Ince, J., Duncan, S. H., Webster, L. M., Holtrop, G., Ze, X., et al. (2011). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *The ISME Journal*, *5*(2), 220–230. http://doi.org/10.1038/ismej.2010.118

Wang, Jeff, Kato, F., Yamashita, H., Baba, M., Cui, Y., Li, R., et al. (2017). Automatic Estimation of Volumetric Breast Density Using Artificial Neural Network-Based Calibration of Full-Field Digital Mammography: Feasibility on Japanese Women With and Without Breast Cancer. *Journal of Digital Imaging*, *30*(2), 215–227. http://doi.org/10.1007/s10278-016-9922-9

Wang, Ying, Su, L., Huang, S., Bo, C., Yang, S., Li, Y., et al. (2016). Diversity and resilience of the wood-feeding higher termite *Mironasutitermes shangchengensis* gut microbiota in response to temporal and diet variations. *Ecology and Evolution*, *6*(22), 8235–8242. http://doi.org/10.1002/ece3.2497

Warnecke, F., & Hugenholtz, P. (2007). Building on basic metagenomics with complementary technologies. *Genome Biology*, *8*(12), 231. http://doi.org/10.1186/gb-2007-8-12-231

Watanabe, H., & Tokuda, G. (2001). Animal cellulases. *Cellular and Molecular Life Sciences : CMLS*, *58*(9), 1167–1178. http://doi.org/10.1007/PL00000931

Wernegreen, J. J. (2002). Genome evolution in bacterial endosymbionts of insects. *Nature Reviews Genetics*, *3*(11), 850–861. http://doi.org/10.1038/nrg931

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*.

Wilkinson, T. L., & Douglas, A. E. (1995). Why pea aphids (*Acyrthosiphon pisum*) lacking symbiotic bacteria have elevated levels of the amino acid glutamine. *Journal of Insect Physiology*, *41*(11), 921–927. http://doi.org/10.1016/0022-1910(95)00063-Z

Wilson, A. C. C., Ashton, P. D., Calevro, F., Charles, H., Colella, S., Febvay, G., et al. (2010). Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*. *Insect Molecular Biology*, *19*(s2), 249–258. http://doi.org/10.1111/j.1365-2583.2009.00942.x

Woese, C. R. (1987). Bacterial Evolution, 1–51.

Wu, D., Daugherty, S. C., Van Aken, S. E., Pai, G. H., Watkins, K. L., Khouri, H., et al. (2006). Metabolic Complementarity and Genomics of the Dual Bacterial Symbiosis of Sharpshooters. *PLoS Biology*, *4*(6), e188–14. http://doi.org/10.1371/journal.pbio.0040188

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., et al. (2011). Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science*, *334*(6052), 105–108. http://doi.org/10.1126/science.1208344

Yu, Z., & Morrison, M. (2004). Improved extraction of PCR-quality community DNA from digesta and fecal samples. *BioTechniques*, *36*(5), 1–4.

Zheng, H., & Brune, A. (2015). Complete Genome Sequence of *Endomicrobium proavitum*, a Free-Living Relative of the Intracellular Symbionts of Termite Gut Flagellates (Phylum Elusimicrobia). *Genome Announcements*, *3*(4), e00679–15. http://doi.org/10.1128/genomeA.00679-15

Zheng, H., Dietrich, C., Hongoh, Y., & Brune, A. (2016a). Restriction-Modification Systems as Mobile Genetic Elements in the Evolution of an Intracellular Symbiont. *Molecular Biology and Evolution*, *33*(3), 721–725. http://doi.org/10.1093/molbev/msv264

Zheng, H., Dietrich, C., Radek, R., & Brune, A. (2016b). *Endomicrobium proavitum*, the first isolate of Endomicrobia class. nov. (phylum Elusimicrobia)--an ultramicrobacterium with an unusual cell cycle that fixes nitrogen with a Group IV nitrogenase. *Environmental Microbiology*, *18*(1), 191–204. http://doi.org/10.1111/1462-2920.12960

Zuroff, T. R., & Curtis, W. R. (2012). Developing symbiotic consortia for lignocellulosic biofuel production. *Applied Microbiology and Biotechnology*, *93*(4), 1423–1435. http://doi.org/10.1007/s00253-011-3762-9