Isolation and characterization of novel enzymatic activities from gut metagenomes to support lignocellulose breakdown

Lê, Ngc Giang

2021

document version
Publisher's PDF, also known as Version of record

Isolation and characterization of novel enzymatic activities from gut metagenomes to support lignocellulose breakdown

Ngoc Giang Le

This study was partially financed by a grant from the BE-Basic Foundation (Biotechnology based Ecologically Balanced Sustainable Industrial Consortium) on behalf of the Dutch Ministry of Economic Affairs (grant number F07.003.05 and F07.003.07). The study was also carried out with the financial support of the Project “Metagenome of some potential mini-ecologies for mining novel genes encoding effective lignocellulolytic enzymes” code DTDLCN.15/14, managed by the Ministry of Science and Technology, Vietnam, in collaboration.

VRIJE UNIVERSITEIT

ISOLATION AND CHARACTERIZATION OF NOVEL ENZYMATIC ACTIVITIES FROM GUT METAGENOMES TO SUPPORT LIGNOCELLULOSE BREAKDOWN

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. C.M. van Praag,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètawetenschappen
op donderdag 21 oktober 2021 om 13.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door
Ngọc Giang Lê
geboren te Hanoi, Vietnam

promotoren:	prof.dr. N.M. van Straalen
	prof.dr. N.H. Truong
copromotor:	dr.ir. T.F.M. Roelofs

Table of contents

Chapter 1 - Introduction

1.1 Agriculture waste and renewable energy

To meet the needs of the growing world population, each year the global production of food has to increase. As an unavoidable side-effect, this results in increasing amounts of agricultural waste generated as byproducts, e.g. crop residues, straw, stubbles, seed hulls, etc. A large amount of organic carbon is retained in this waste, of which most is left to decompose or is burned, so contributing to uncontrolled carbon dioxide emission. Based on a report from the Food and Agriculture Organization of the United Nations (FAO), the greenhouse gases emissions of the world in 2018 from burning crop residues is 30,454.37 gigagrams (Gg), and these emissions increased by 21% since 2000. Asia, Americas and Africa are the top three continents with the largest emission. Maize, rice and wheat are the top three burned crops (Figure 1). These numbers are expected to increase as the world population increases.

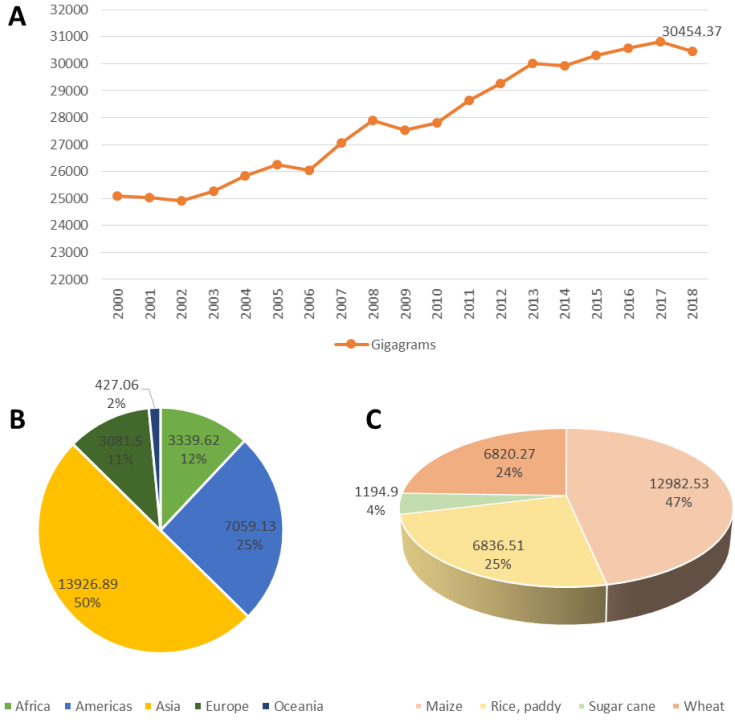


Figure 1: *A) The CO2 emission (Gg per year) increased from 2000 to 2018. B) The amount of emissions from the five continents. C) Burned waste classified by crop (Modified from FAO.org).*

Lignocellulosic biomass is the primary component of agricultural waste. Only a small fraction of this is recycled or reused as fertilizer, as a feedstock or in mushroom cultivation (Gaitán-Hernández, Cortés, and Mata 2014). In many countries these wastes are not efficiently managed and are often disposed of by burning (Ravindra, Singh, and Mor 2019; Cassou, Jaffee, and Ru 2018). This wasteful discarding of valuable resource contributes significantly to air pollution. In 2020 many Asian agricultural countries such as Vietnam and India suffered air pollution for weeks during the harvest period.

Vietnam is the second largest rice exporter in the world. About of 44,046,250 tons of paddy rice were produced in 2018. A study by Le, Phuong and Linh, published in 2020 showed that around the Red River Delta in 2018 about 3.24 million metric tons (Mt) of rice straw was burned with an emission of about 3.82 Mt of carbon dioxide and 301 Gg of carbon monoxide (Le, Phuong, and Linh 2020). The amount of CO_2 and CO accounted for 89.77% and 7.09% of all gaseous emissions, respectively. This is an appalling waste of valuable resources. Under proper treatment such biomass could be converted into raw materials and recycled for other carbon-based products.

One possible treatment of organic waste is fermentation and the production of bioethanol. This is an interesting product as it might substitute fuels derived from fossil carbon in the future. Ethanol is biodegradable, less toxic and causes less environmental pollution than gasoline (Balat 2011). When burned, ethanol produces carbon dioxide and water. Some of this carbon dioxide will be recycled by plants to create biomass, which can be used as raw material again for bioethanol production. In theory, such a process is a closed cycle with net zero carbon dioxide release to the atmosphere. Bioethanol is usually made in the sugar fermentation process called saccharification. Raw materials can be starch, sugar and lignocellulosic materials (Gírio et al. 2010; Kim 2018). When recycling organic carbon from agricultural waste, depending on the feedstock used, the global greenhouse gas emissions can be reduced by 30-85% (Saini, Saini, and Tewari 2015).

1.1.1 The mission of BE-Basic

The Biotechnology-based Ecologically Balanced Sustainable Industrial Consortium (BE-Basic) is an international public-private partnership (https://be-basic.org/). The foundation collaborates with industries and institutes from different countries to contribute to the transition towards a bio-based economy. In this economy, fossil fuel is replaced by biomass

from agricultural waste or non-edible plants. For a smooth transition, research in BE-Basic focuses on new technologies and insights into process of carbon release from waste. The philosophy of BE-Basic is to focus on enzyme-based reactions, so moving away from the classical chemistry which often comes with considerable energy inputs and emission of hazardous substances. Under the paradigm of "green chemistry", processes are designed that can generate valuable chemical products from waste in a sustainable way, without noxious emissions. One of the fields of research is to use high-throughput experimentation and (meta)genomic mining to identify enzymes and other products for improved properties (https://be-basic.org/research/high-throughput-experimentation-metagenomic-mining/). Enzymes can break down biomass into substrate for the fermentation of bioethanol.

1.2 Lignocellulose as renewable source

Plant biomass is one of the most viable renewable resources for biofuel and chemical feedstock (Bornscheuer, Buchholz, and Seibel 2014). Plant cell walls are composed of carbohydrate polymers such as cellulose, hemicellulose and aromatic polymers such as lignin. These structures consist of cross-linked matrices to protect the plant from physical and chemical damages (Jönsson and Martín 2016). Generally lignocellulose biomass is made up of about 10% - 20% lignin, 20% - 30% hemicellulose and 40% - 60% cellulose (Figure 2) (H. Chen 2014b). Hemicellulose mainly consists of pentoses such as xylose, arabinose and galactose. The proportion of these components varies between plants and species such as hardwood, softwood or grasses (Schutyser et al. 2017).

Cellulose is the most abundant polysaccharide on earth. It is synthesized by binding (1,4)-D-glucopyranose units via β-1,4 linkages. In nature, cellulose molecules join together to form microfibrils. Within this structure, the highly compact crystalline regions are separated by amorphous regions. The insoluble cellulose in the plant cell wall provides strength and flexibility to the cell wall, allowing turgor.

Hemicellulose is a short, amorphous and highly branched polymer. Sugar monomers include xylose, mannose, galactose, rhamnose and arabinose. Hemicellulose links to cellulose via hydrogen bonding and to lignin via ionic interaction.

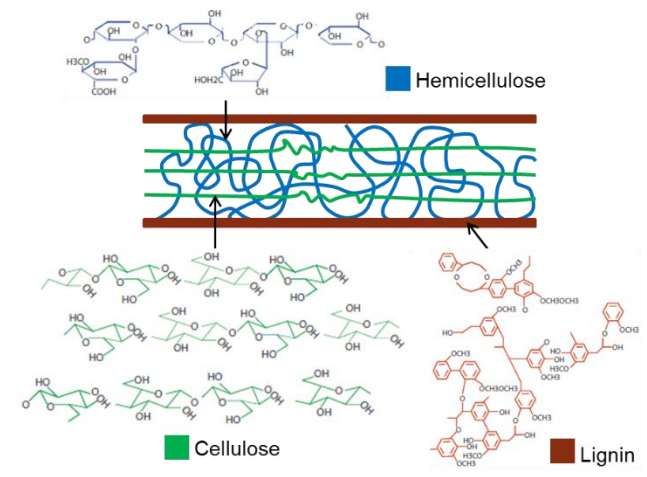


Figure 2: *The cell wall structure of a plant. Hemicellulose (blue) is entwined with cellulose (green) and lignin (brown). Together they create a matrix, forming a strong barrier to protect the plant. Modified from* (Bamdad, Hawboldt, and MacQuarrie 2018; Chitra Devi et al. 2020)

Lignin is a heterogeneous cross-linked aromatic polymer, which is composed of three phenyl propane monomers: coumaryl, coniferyl and sytingyl (Bornscheuer, Buchholz, and Seibel 2014; Kim 2018). Different alternative chemical forms of these monomers are found in different plants. Lignin fills the space between cellulose and hemicellulose. It gives the plant cell wall strength and rigidity as well as physical and chemical protection (Saini, Saini, and Tewari 2015; Maurya, Singla, and Negi 2015).

1.2.1 Current pretreatment methods

For bioethanol production, hemicellulose is the main source of material. However, it is well protected by lignin. When lignin is present as intact structure, a lower yield of monomeric sugar is obtained. A pretreatment method is needed to disrupt the lignin structure and loosen the plant cell wall for enzymatic access to cellulose (Fig 3). Such a pretreatment reduces the crystallinity and increases the amorphous state of cellulose. Currently, this is the most challenging step in the production of bioethanol from plant remains. A large number of investigations have been conducted to identify ways to efficiently break down the cellulosic component in lignocellulose. There are four main types of pretreatment methods (Maurya, Singla, and Negi 2015).

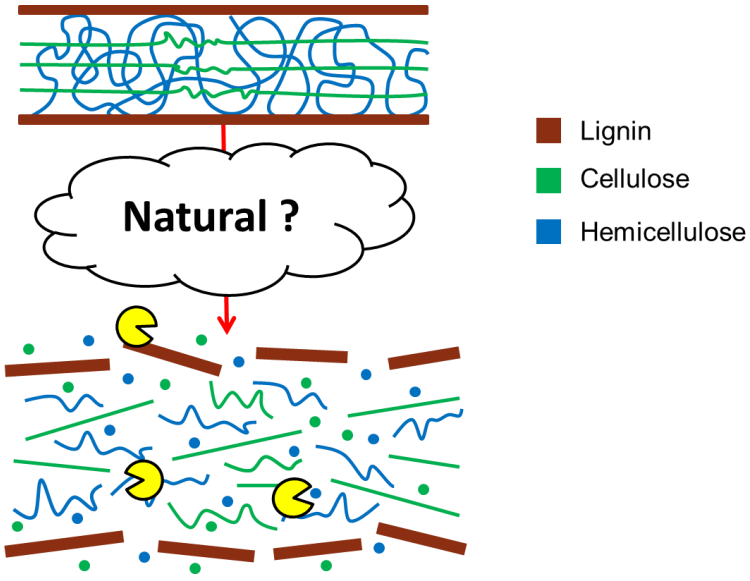


Figure 3: *Enzymes is the natural ways to breakdown of lignocellulose. This requires the least amount of energy and chemicals.*

1.2.1.1 Physical pretreatments

Lignocellulosic materials can be pulverized by grinding, shearing, milling or chipping. The particle size of the biomass is reduced to increase surface area. The disadvantages of this method are high power and high energy consumption (Wi et al. 2013; I. S. Choi et al. 2013).

1.2.1.2 Physico-chemical pretreatment

In case of steam explosion, biomass is subjected at high pressure to saturated steam for several minutes, then the pressure is released. The sudden pressure reduction causes the fibers to separate. Ammonia fiber expansion (AFEX) uses ammonia to cause fibers to swell and change formation. The biomass is treated with ammonia at 90-100°C for 30-60 min. Another method is wet oxidation: the biomass is treated with water and air at 120°C for 30 min. This process catalyzes the formation of acids from hydrolytic pressure and oxidative reactions. These pretreatments require high energy, chemicals and machines to create the high pressure and high temperatures (Ye et al. 2016; M. J. Taylor, Alabdrabalameer, and Skoulou 2019).

1.2.1.3 Chemical treatments

Different chemical agents are used such as acids and alkaline. Acid is used for the solubilization of hemicellulose and lignin. Alkaline is used for the removal of lignin from the biomass. It also removes acetyl and uronic acid from hemicellulose. Both methods are expensive and produce toxic compounds. The treated material has to be cleaned for further processing (Abedinifar et al. 2009).

1.2.1.4 Biological processes

The biological methods rely on enzymes that can release sugars from lignocellulose. Such enzymes are naturally present in many fungi that grow on dead or live wood (Zhao et al. 2013; Baldrian et al. 2016). Different fungal species are used: brown rot, white rot and soft rot fungi. This process requires little energy and proceeds under mesophilic conditions. However the yields are low and the process is time-consuming. A lot of investigations have been directed towards finding more efficient pretreatments. One strategy is to look for specialized enzymes present in micro-organisms (fungi and bacteria). The staggering diversity of the microbial world, which extends far beyond the microbial species that can be cultured in the laboratory, is considered to hold great promises for this new type of biotechnology. Therefore, many scientists are screening the genomes of microorganisms in search for novel genes that could be deployed for lignocellulose processing (Jönsson and Martín 2016; Kim 2018; Kucharska et al. 2018).

Carbohydrate-active enzymes

Enzymes that are involved with carbohydrate metabolism are grouped under the term carbohydrate-active enzymes (CAZys). It is a large and heterogenous group of proteins with the common property that they catalyze the degradation of carbohydrates and related molecules. The enzymes are grouped based on their sequence and activity profile. There are six groups, designated as glycoside hydrolases (GH), glucosyltransferases (GT), molecules with auxiliary activity (AA), carbohydrate esterases (CE), polysaccharide lyases (PL) and carbohydrate-binding molecules (CBM). Together, these six groups allow the complete breakdown of lignocellulose (Table 1).

Table 1: *Cocktail of enzymes for complete lignocellulose degradation* (Parisutham, Kim, and Lee 2014).

Cellulases	**Hemicellulases**	**Pectinolytic enzymes**	**Lignin degradation**	**Cell wall loosening enzymes**
Cellobiohydrolase Endoglucanase β-Glucosidase Phospho-β-glucosidase	Endoxylanase β-Xylosidase Acetyl xylan esterase Feruloyl esterase Glucuronoyl esterase Arabinofuranosidase Galactosidase Glucuronidase Mannanase Xyloglucan hydrolase	Polygalacturonases Pectin/pectate lyases Pectin methyl esterase	Lignin peroxidase Aryl-alcohol oxidase Laccase Glyoxal oxidase Cellobiose dehydrogenase	Expansin Swollenin Loosinin Cellulose induced protein

By way of example, we discuss one specific group of enzymes that is receiving special attention in this thesis. Alpha-arabinofuranosidases belong to the group of hemicellulases. They are exo-enzymes excreted by bacteria to hydrolyze the terminal nonreducing α-L-arabinofuranosyl side-chains from L-arabino-containing polysaccharides. This removal of side-chains eases the complete degradation of hemicellulose by endo-1,5-α-L-arabinanases that attack the hemicellulose backbone (Fig. 4).

Arabinofuranosidases have a potential application in agro-industrial processes: fruit, vegetables and cereals processing. The discovery of new and effective arabinofuranosidases might contribute to the sustainable conversion of hemicelluloses to bioethanol and organic chemicals.

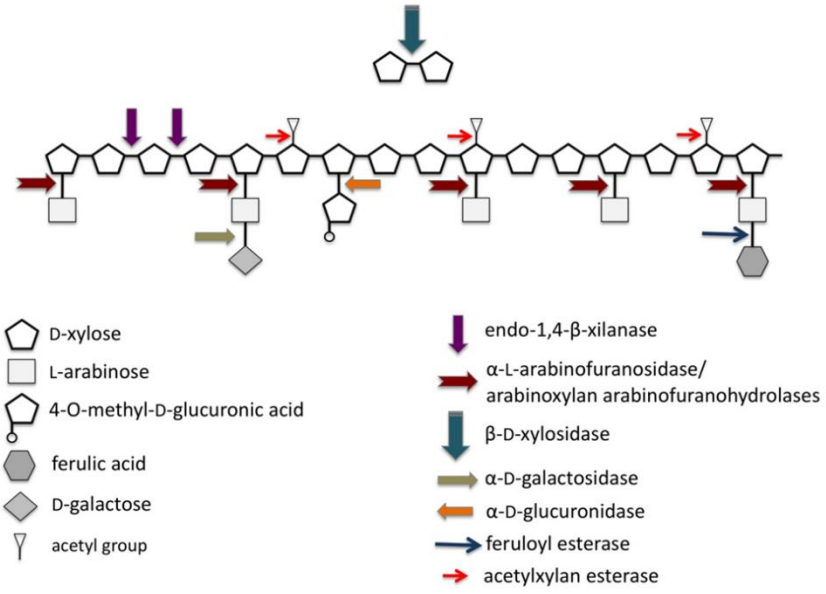


Figure 4: *Enzymes required for complete breakdown of hemicellulose. The points of attack of the different enzymes are indicated by arrows. Image from* (de Souza 2013).

1.3 Concept of animal guts to optimize catalytic functions

Decomposition of organic matter is a natural process, where materials are broken down by microbes to smaller building blocks. Animals usually do not have cellulolytic activity of their own, they rely on microbes in their gut. Of special interest are animals that feed on recalcitrant organic matter, as they must have gut microbes with special catalytic properties. Therefore, in this thesis I explored the microbiomes of three groups of animals, ruminants and detritivore arthropods, that are well-known for their capacity to digest recalcitrant plant materials (K. T. Lee et al. 2018; Do, Le, et al. 2018; Fountain and Hopkin 2005; Brune and Dietrich 2015). As the degradation of such materials is crucial for survival of these animals, they are relevant targets for the discovery of new cellulolytic activities.

1.3.1 Termites

Termites are a popular model for microbial biodiversity and lignocellulose degrading enzymes (Ni and Tokuda 2013). Worker termites are known for their diets containing high lignocellulosic fiber components. They can degrade 74–99% of cellulose and 65–87% of hemicellulose from biomass within hours after feeding (Geng et al. 2018; Hongjie Li et al.

2017b). The degradation of plant biomass is performed by host enzymes and microorganism inside the gut (Xie et al. 2014), Fig. 5.

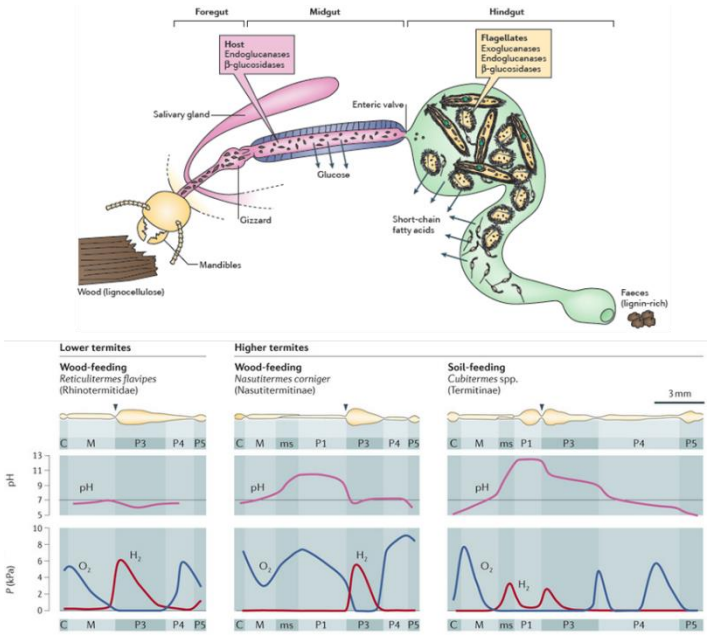


Figure 5: *The termite gut structure of lower and higher termite. The figure also shows the different carbohydrate enzymes found in each part of the gut. Differences in oxygen and hydrogen partial pressures (P) in kPa, intestinal pH between different groups of termites. C: crop, M: midgut; ms, mixed segment; P1–P5, proctodeal segments. Modified from* (Brune 2014; Hongjie Li et al. 2017b).

A discrimination is made between lower and higher termites based on their gut morphologies and microorganisms. Their anaerobic intestinal tract has different metabolic activities and microbial communities. It is composed of a foregut, a midgut and a hindgut (Fig. 5). Woody food sources are grinded in the foregut and transferred to the midgut. This is a secretion site, where nutrients are absorbed. Few microorganisms are found in the foregut and the midgut. The largest compartment is the hindgut, which is divided into several smaller segments. An abundance of microorganisms resides in this region, such as archaea, bacteria and protists. The lower termites have more protists in their hindgut than higher termites. The gut of higher termite has evolved to become longer and with a higher pH (Brune 2014; Brune and Friedrich 2000).

The termite gut systems are very similar to current pretreatment methods for lignocellulose. The microorganisms inside the gut are highly selective to be able to survive the high pressure, alkaline and anaerobic conditions. The termite gut is like a small lignocellulose degradation factory. The system is similar to the abiotic grinding, pressure and alkaline pretreatment method. Since the gut condition is quite specific, adapted species might have specific traits for their survival.

1.3.2 Goats

Through evolution animal guts have evolved to become specialized to degrade the food that the host consumes. Ruminant animals have coevolved with the microbial consortium that harness enzymatic hydrolysis to release fermentable sugar from plant cell wall polysaccharide. The released sugars are subsequently fermented by the microbes to short chain fatty acids as the main food source for the host (Fig 6).

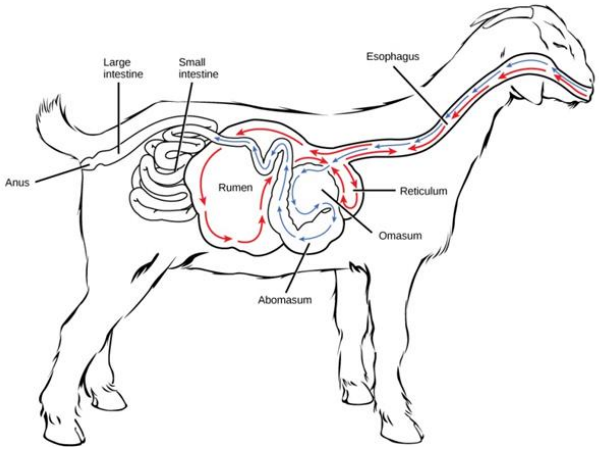


Figure 6: *The goat gut system. Image from* (Connie et al. 2016).

Goats are known for their abilities to adapt to harsh environments due to their behavioral, morphological, physiological as well as genetic properties (Berihulay et al. 2019). They also explore a very diverse diet including grass, plants, root, stems, and shoots. Plant fibers are broken down by fungi, bacteria and protists inside the goat rumen to generate monosaccharaides. Some bacteria and protists metabolized the monosaccharides to generate CO_2, NH_4, volatile fatty acids and H_2. Archaea use the H_2 to generate methane, which in turn eliminates the inhibiting effect of hydrogen on fermentation. As a result of this, ruminants are well-known for their ability to create methane (Agrawal, Karim, and Kumar 2014). The

released sugars are fermented by the microbes to short-chain fatty acids that serve as the main energy source for the host. Genomes of the plant cell wall degrading microbes in the rumen represent a rich source of novel and highly active plant cell wall degrading enzymes.

1.3.3 Springtails

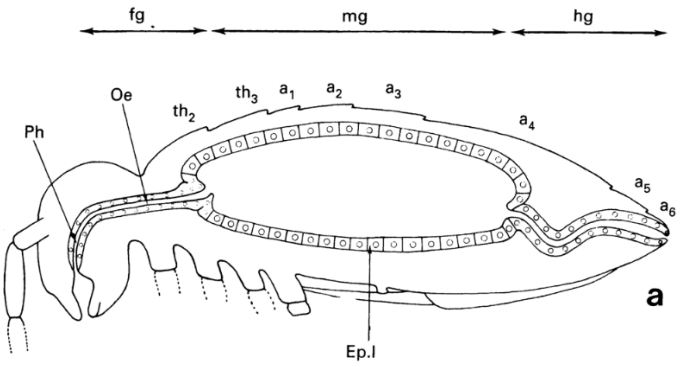


Figure 7: *Schematic drawing of the gut system of Sinella coeca, a collembolan; fg = foregut, mg = midgut, hg = hindgut, Ph = pharynx, Oe = oesophagus, th_2, th_3 = second and third thoracic segment, a_1 to a_6 = abdominal segments, Ep.1 = intestinal epithelium. Figure from Hopkin (1997).*

Springtails (Collembola) are hexapods belonging to the wingless branch of six-legged arthropods, a sister group of the insects. A great variety of species live in organic soils of forests, grasslands and agricultural fields (Fountain and Hopkin 2005). They are mostly unspecialized feeders, eating the mycelia of saprotrophic fungi and mycorryhizae, as well as dead organic matter. The gut of springtails consists of three compartments, like in termites, however, the main digestive compartment is not the hindgut but the midgut (Fig. 7). The midgut has an epithelial lining consisting of large digestive cells, which take up nutrients digested by microbes in the lumen. Collembola moult throughout their lives and with every moult the midgut lining is renewed as well. The recurrent regeneration of the midgut epithelium does not, however, prevent the build-up of a diverse community of microorganisms, the composition of which depends on the host strain as well as on the environment (Valeria Agamennone et al. 2015). The microbial communities in the gut are dominated by Proteobacteria, Actinobacteria, Bacteriodetes and Firmicutes (Valeria Agamennone et al. 2019).

1.4 Methodology

1.4.1 General strategy

Each animal host has a different gut system and so different mini eco-environments. Microorganisms form large populations in the guts and help the host with a suit of functions varying from decomposition of organic matters and mineralization, cycling of nutrients, defense against pathogens and metabolic digestion. These gut symbionts might contain interesting enzymes that can be used to break down biomass. Theoretically, this can be very beneficial for pretreatment plants as the natural enzymes have been prone to natural selection, thereby optimized and adapted to sometimes hard to process diets throughout the animal's evolution.

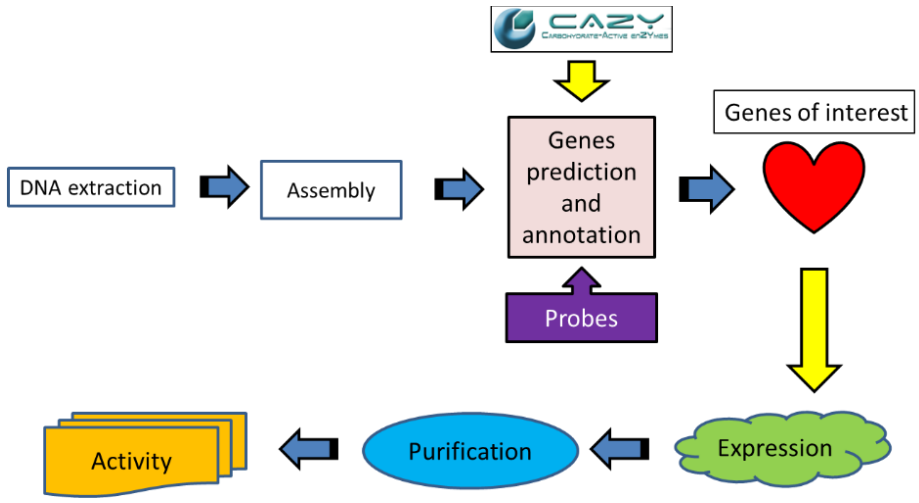


Figure 8: *The overall process of mining for active enzymes using bioinformatics. The DNA of the gut content from the host are extracted and sequenced. After passing a quality tests, reads are used for an assembly. The assembly creates large segments of DNA called contigs (sometimes a whole bacterial genome, but usually fragments of them). Bacterial genes are predicted from long contigs. The CAZy database is used to screen for enzymes with carbohydrate activities. Genes of interest are mined based on specific criteria. These genes are either cloned from the metagenome DNA or synthesized. They are ligated into an E. coli host for recombinant expression. Expressed proteins are collected and purified. Activity of the protein is tested using enzyme assays, usually reactions that can be monitored at a specific absorbance or color change.*

The traditional investigation method to identify active enzymes requires members of a microbiome to be cultured for further testing. However, the microorganisms that are culturable make up only a small fraction of the microbiome. The process makes it difficult to identify new species and novel genes (Wade 2002). The advancement in next-generation sequencing technologies, has led to taxonomic classification and functional metagenomics analysis of unculturable microorganisms at an unprecedented level. The ribosomal RNA genes from bacteria and fungi contain highly conserved regions, which can be targeted by primers for the identification of different species. This allows for identification of more variable regions to be used for taxonomic classification, where empirically defined divergence threshold have been agreed to link certain taxonomic levels such from strain/ species up to phylum level. Functional metagenomics is the study of genomes of all organisms in an environment sample (Schloss and Handelsman 2005). The process allows to study of microbiomes in their natural environments. The targets of study are culturable and unculturable microorganisms from the animal host, which can breakdown lignocellulose. This process can help us to understand more about the interaction between the microorganisms and the host. Figure 8 provides a general work flow.

1.4.2 DNA isolation, sequencing and assembly

First, the gut contents from the animals are isolated and DNA is extracted from the microbial community (Do, Dao, et al. 2018; Do, Le, et al. 2018; Valeria Agamennone et al. 2015). The quality of the DNA is checked before sending it out for sequencing. In metagenomics projects, the Illumina short read strategy is often used (Pearman, Freed, and Silander 2019). The raw reads are subjected to preprocessing. All general primer sequences used from next-generation sequencing are removed by trimming. Also, reads that do not meet specific quality such as singleton, too short or below Q20 are either trimmed or removed from the database (Valeria Agamennone et al. 2019).

A large part of the metagenome may be represented by contamination, especially from the host. On top of that, some animal species also contain (endo)symbionts, which can bias the overall analysis towards these organisms. For example, springtails contain endosymbiotic *Wolbachia* bacteria that dominate the microbial DNA isolated from the host. To have a clear view of the metagenome, the host, as well as the *Wolbachia* and some known virus contamination need to be removed from the data. Consequently, it reduces the number of

reads, but this reduced sequence output in turn speeds up the downstream assembly process. To compensate for reduced output, multiple samples are pooled together to improve on quality.

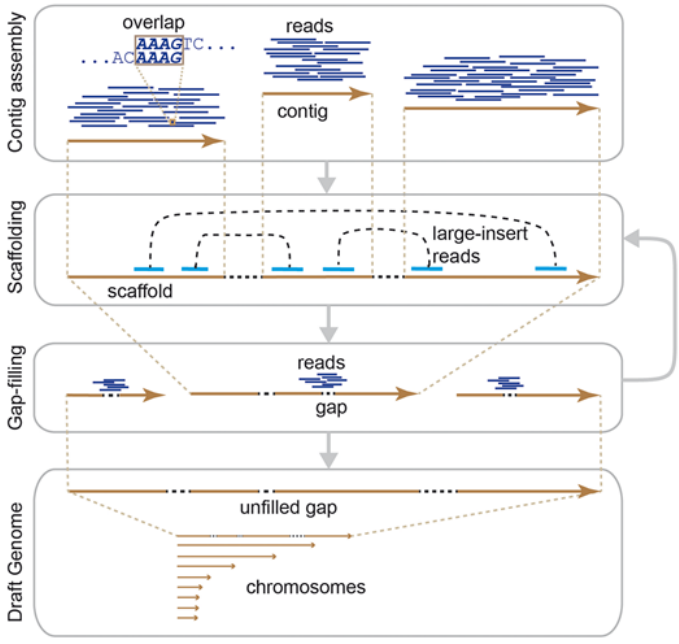


Figure 9: *Creation of draft genome from sequencing reads. Reads are aligned, where overlaps help to join multiple reads together to create contig. The scaffold are made up of large contigs with gap. Some of the reads are then used to fill out gaps from the large contigs to create draft genome. Image taken from* (Sohn and Nam 2018).

Obviously, a microbiome consists of a wide diversity of microorganisms. Hence, the reads derive from different organisms present in the microbial community. But for a better understanding of the community structure, reads of the same species need to be joined together, a process called assembly. Sequence reads that overlap help to join different reads to create longer contigs. A general work flow of microbial population assembly is depicted in figure 9.

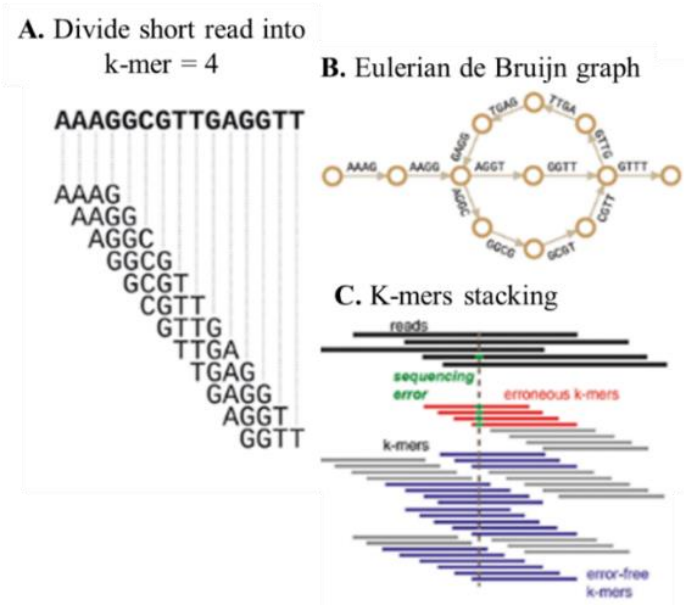


Figure 10: *Assembly using k-mers. A) The short read is divided into multiple instances with the example of k-mer =4. B) The graph of Eulerian de Bruijn graph. C) Kmers are stacked together to create contigs. Erroneous k-mers which appears in less reads are removed. Image modified from* (Sohn and Nam 2018).

The program SPADES uses a Eulerian de Bruijn graph approach for the assembly of metagenomes (Bankevich et al. 2012). For this method, a specific kmer's length is set. The short reads are then classified using this specific length. These k-mers are then stacked together to create a graph with multiple paths (Fig. 10). The most optimal path is selected, based on pre-set probability and accuracy tresholds? SPADES also has the option to combine multiple contigs from different k-mers to extend the contig length, by lowering the consensus accuracy. This process is often used for de novo assemblies (Sohn and Nam 2018).

1.4.3 Discovery of functional genes and assessment of activity

The contigs are subsequently used to identify open reading frames (ORF) from bacterial species (Hyatt et al. 2010). To that end, the ORFs are translated in predicted proteins and used as query to search for homology in online non-redundant protein databases with BLAST. The result of such homology searches shows similarities of predicted peptides is

with characterized proteins in databases, in the ideal case with known and proven functions (Altschul et al. 1990).

To identify proteins with carbohydrate activities, protein homologies are searched against the CAZy database. The dbCAN2 uses multiple programs for the identification CAZymes. DIAMOND aligns sequences of high similarity with the CAZymes conserve domains, Hotpep is used to identify of short, conserved protein sequence motifs and HMMER uses the statistical Hidden Markov Models for training and identifying of conserved domain of potential CAZymes proteins (Zhang et al. 2018; Busk et al. 2017; Bystroff and Krogh 2008). Sequences with homologies to known proteins, which pass certain thresholds are more likely to be active ones. However, picking proteins with lower similarities are more likely to exert potentially novel novel catalytic functions.

After screening for candidate genes, the protein is either cloned from the DNA metagenomics pool or *in vitro* synthesized. Primers are designed to create restriction sites to easily incorporate a gene into a plasmid. The plasmid is transformed into *E. coli* for expression. The plasmid is designed such that the expressed protein contains an N-terminal string of histidine residues, a so-called His-tag. The bacterial host is cultured to enhance the amount of cells containing recombinant protein A crude protein extract obtained after lysis of the host is loaded into a his-tag column. The proteins with the his-tag bind to the Cobalt beads due to their negative charge. A washing solution containing imidazol at low to high concentration is used to elute the column. The washing solution slowly removes all non his-tag protein. Imidazol competes for the protein of interest and at the correct concentration expels this protein from the beads (Spriestersbach et al. 2015). The protein of interest is collected from one of the fractions.

To study the activity of the isolated protein a variety of methods can be used. Because enzymes break down substrate into specific products, the activity assay is usually directed towards monitoring the rate at which product concentrations increase or substrate concentrations decrease, usually by spectrophometry. Activity assays are performed at different conditions (substrate concentration, pH, temperature, etc.), to characterize the enzyme's properties. Important properties of the enzyme are its substrate binding affinity (expressed as the parameter K_M in the Michaelis-Menten model) and the maximal reaction

rate V_{max}. These parameters are estimated by curve-fitting in a graph of reaction rate measurements at different substrate concentrations.

1.5 Scope, approach and outline of the thesis

In this thesis we are interested in exploring the microbial communities in the guts of animals to identify relevant enzymes with functions that can be used in industrial applications of lignocellulose breakdown. Animals were chosen that have a variety of biomass digestion strategies. Goats and termites can deal with hard woody materials and springtails with small decomposing wood items, dead organic matter and fungi. The microorganisms inside their guts are support digestion and subsequent nutrition as well as a variety of other functions including defense against pathogens and the production of antimicrobials. By analyzing the host-microbe interactions we also aim to increase our understanding of their co-evolution. The use of bioinformatics tools will help to identify and discover enzymes as well as the genes functional in the microorganisms.

Main question for the thesis are formulated as follows

1. What microbial communities are present in the three selected animal species, how do they compare to each other?
2. Which functionalities are encoded in the metagenomes of these communities (with emphasis on carbohydrate metabolism)?
3. What are the properties of metagenome-derived enzymes as possible candidates for bio-based degradation of organic waste?

In Chapter 2, we describe the gut composition of the goat gut in related to the carbohydrate active enzymes especially hemicellulose degrading enzymes. The research looks at the bacterial community diversity inside the goat gut.

In Chapter 3, we look at the springtail as a model of interest to explore the functional potential of the microbiome. Using bioinformatics tools we focus on carbohydrate metabolism functions, as well as antibiotic biosynthesis gene clusters.

In Chapter 4, we compare the gut microbiomes from three different invertebrates: springtails, isopods and termites. By expanding the animals using the same methodology from chapter 3, we look at the important functions such as the ability to break down carbohydrate for food

source as well as antibiotic resistance and the production of secondary metabolites. We observed that for each gut community, a specific set of carbohydrate enzymes are required even though they are not contributed by the same microorganisms.

In Chapter 5, we describe the activity of hemicellulases identified from the metagenome gut of termites and springtails. The scope of this chapter is to show the potential of using bioinformatics tools to mine for hemicellulase genes. To do this we expressed the genes of interest in *E. coli* against hemicellulose substrate. We observed activity as predicted by the bioinformatics tools.

In Chapter 6 we describe an α-L-arabinofuranosidase gene from the springtail similarly found in the gut of the termite. This could be a novel horizontal gene transfer from long time ago. The scope of this chapter was to show the evolutionary of this protein, which show activity like the one from chapter 5.

In the final Chapter 7, I present a general discussion of results from previous chapters in light of the research question of the thesis. I will also provide an outlook to future research on this topic.

Chapter 2 - Metagenomic insights into lignocellulose-degrading genes through Illumina based *de novo* sequencing of the microbiome in Vietnamese native goats' rumen

Received December 5, 2016; Accepted August 21, 2017; J-STAGE Advance publication date: March 12, 2018

Thi Huyen Do, Ngoc Giang Le, Trong Khoa Dao, Thi Mai Phuong Nguyen, Tung Lam Le, Han Ly Luu, Khanh Hoang Viet Nguyen, Van Lam Nguyen, Lan Anh Le, Thu Nguyet Phung, Nico M. van Straalen, Dick Roelofs, and Nam Hai Truong

2.1 Abstract

The scarcity of enzymes having an optimal activity in lignocellulose deconstruction is an obstacle for industrial-scale conversion of cellulosic biomass into biofuels. With the aim of mining novel lignocellulolytic enzymes, a ~9 Gb metagenome of bacteria in Vietnamese native goats' rumen was sequenced by Illumina platform. From the data, 821 ORFs encoding carbohydrate esterases (CEs) and polysaccharide lyases (PLs) serving for lignocellulose pre-treatment, 816 ORFs encoding 11 glycoside hydrolase families (GHs) of cellulases, and 2,252 ORFs encoding 22 GHs of hemicellulases, were mined. The carbohydrate binding module (CBM) was also abundant with 763 ORFs, of which 480 ORFs are located with lignocellulolytic enzymes. The enzyme modularity analysis showed that CBMs are usually present in endoglucanase, endo 1,3-beta-D-glucosidase, and endoxylanase, whereas fibronectin 3-like module (FN3) mainly represents in GH3 and immunoglobulin-like domain (Ig) was located in GH9 only. Every domain located in each ORF was analyzed in detail to contribute enzymes' modularity which is valuable for modelling, to study the structure, and for recombinant production. With the aim of confirming the annotated results, a mined ORF encoding CBM63 was highly expressed in *E. coli* in soluble form. The purified recombinant CBM63 exhibited no cellulase activity, but enhanced a commercial cellulase activity in the destruction of a paper filter.

2.2 Introduction

Lignocellulose waste comprising agro-industrial biomass is inexpensive, renewable, abundant, and provides a unique natural resource for enhancing bio-economy (Anwar, Gulfraz, and Irshad 2014) to substitute the fossil-based economy.

Overcoming the limitations of fossil-based economy, bio- based economy has the advantage to (i) be environmentally, economically and socially sustainable; (ii) decrease the dependence on fossil fuel; (iii) reduce atmospheric greenhouse gas emission, which is responsible for causing climate change; and (iv) stimulate regional and rural development (Jong et al. 2011). Lignocellulose can be converted into sugar molecules by microbial enzymes and the released sugars can be fermented into various high value products including bio-fuels, materials for food, bulk chemicals such as bioplastics, and value-added fine chemicals for pharmaceuticals and human health (Asgher, Ahmad, and Iqbal 2013; Iqbal, Kyazze, and Keshavarz 2013; Millati et al. 2011; Irshad et al. 2013). Therefore, lignocellulose biomass has recently gained

increasing research interest and special importance (Asgher, Ahmad, and Iqbal 2013; Baumann and Westermann 2016; Ofori-Boateng and Lee 2013).

The conversion of lignocellulose into higher-value products requires a multi-step process including (i) pre-treatment (e.g. mechanical, chemical, or biological), (ii) saccharification by enzymes, and (iii) fermentation into end products (Arumugam and Mahalingam 2015). A major obstacle to lignocellulose conversion in industry lies in the inefficient deconstruction of plant material owing to the retention of the natural lignocellulose structure. Also, currently available enzymes which can hydrolyze lignocellulose show a low and ineffective activity (Hess et al. 2011; Sebastian et al. 2013). In nature, individual enzymes interact synergistically, or are comprised of multi-modules (modularity), to degrade lignocellulose effectively. In modularity, besides the catalytic core, these enzymes also possess non-catalytic functionally-important domains, including carbohydrate-binding modules (CBMs), fibronectin 3-like modules (FN3s), dockerins, immunoglobulin-like domains (Ig), or functionally unknown "X" domains (Sweeney and Xu 2012). These domains are important for solubility, optimal activity (Ding et al. 2008; Wilson 2008), stability and even thermo-stability of the catalytic activity (Araki et al. 2006; X. Jia et al. 2016). In *Clostridia*, these enzyme modules are organized in so called cellulosomes through cohensin-dockerin complexes (Dou et al. 2015). Apparently, organisation and interaction of these microbial enzymes for the hydrolysis of lignocellulose are essential in the industrial development of lignocellulose breakdown, which is an important source for the green energy sector (M. Kumar, Varma, and Kumar 2016; Yang et al. 2014). Many recent studies have identified numerous potentially enzymatic pathways for biomass conversion, but less is known about the efficacy of catalytic activity of the enzyme modularity in biomass transformation and digestion (M. Kumar, Varma, and Kumar 2016). Thus, the discovery of novel enzyme modularity for lignocellulose saccharification is required.

Traditionally, functional microbial screening is applied to isolate genes involved in lignocellulose breakdown. More recently, metagenomics can identify candidate genes from environmental samples circumventing the need for culturing. This is important, since more than 99% of microorganisms from environmental samples are uncultivable and their functional significance is overlooked. Thus, next generation sequencing of whole metagenomic DNA from environmental samples with a high lignocellulose breakdown capacity is very powerful for the discovery of genes relevant in this process (M. Kumar,

Varma, and Kumar 2016; Sebastian et al. 2013). The digestive tract of termites (Do et al. 2014; M. Kumar, Varma, and Kumar 2016; Sebastian et al. 2013), and Korean goat rumen (Lim et al. 2013) represent rapid and efficient lignocellulose degradation environments, which make it more likely to discover enzymes that play an essential role in this process. Much emphasis has been given to investigating enzymes from microbiota that can hydrolyse cellulose, and hemicellulose substrates. However, much less information is available on the collocation of important domains (FN3, CBM and Ig) forming modules with catalytic domains to eventually create an efficient system for optimal lignocellulose degradation. Lim et al. (2013) reported nine CBM domains, dockerin-1, and FN3 domains, and these domains were collocated within cellulase and glycosyl hydrolase (GH) families, but lacked all information on genes for many hemicellulases and genes for lignocellulose pretreatment. In addition, most identified cellulases lacked a co-localized with CBM and/or FN3 domain (Lim et al. 2013).

Here, we report on the analysis of a large dataset generated by Illumina-based *de novo* sequencing of bacterial metagenomic DNA extracted from the rumen of native goats living in the natural high mountain at Ninh Binh and Thanh Hoa, Vietnam. These animals consume different plant materials with a high content of lignocellulose. Therefore, we hypothesize that the microbial digestive system of this animal has adapted to degrade substantial amounts of lignocellulose efficiently. A previous study used only one database to analyze goat rumen to identify potentially relevant enzymes (Lim et al. 2013). In this study, we have subjected all open reading frames (ORFs) to six available functional annotation tools. This integrated approach increased the number of identified cellulases and hemicellulases, and enzymes related to lignocellulose pretreatment. We have also analyzed the presence of collocated FN3, CBM, and Ig domains, thereby elucidating the potential of an enzyme to participate in modularity. This information may become necessary for the recombinant production of optimal enzyme cocktails.

2.3 Materials and Methods

2.3.1 Sampling and extraction of bacterial metagenomic DNA

The goat lines used in this study were a Vietnamese native breed (Co) and a hybrid (Bach Thao) generated by Beetal and Jamnapari long time ago. Adult Co animals, weigh approximately 30 to 35 kg (Fig. S1A) and live on natural hay in high rocky mountains at private goat farms at Ninh Binh and Thanh Hoa provinces in Vietnam. The domestic goat breed Co has a small body with brown or black hair, a large head, small short ears, and short horns. The breed Bach Thao is diverse in morphology and size (Fig. S1B). Three Co animals and two Bach Thao animals were sampled in Ninh Binh province (GPS coordinates 20.269002, 105.893267), while two Co animals and three Bach Thao animals were sampled in Thanh Hoa province (GPS coordinates 19.897450 105.795899). The diet of both goat lines consists of a variety of grasses, leaves of trees in the mountains, and also crop residues at night.

In total, ten selected goats were slaughtered at a local slaughter house. Rumen fluid from each goat was filtered through four layers of cheesecloth, and the remains was suspended in 2 liters of PBS buffer (137 mM NaCl, 2.0 mM KH_2PO_4, 10 mM Na_2HPO_4, and 2.7 mM KCl, pH 7.4). It was filtered through a new set of four layers of cheesecloth. The resulting fluids were centrifuged at 700 rpm (approximately 150–200 g) for 10 min to separate protozoa and plant debris from bacteria. This step was repeated twice. The bacteria in the supernatant were pelleted by centrifugation (4,500 *g* for 5 min), washed twice with PBS buffer, and resuspended in 500 ml of PBS buffer.

Genomic DNA was isolated from the bacteria-enriched fluid and purified using a PSP Spin Stool DNA Plus Kit (Stratec, Germany) according to the manufacturer's protocol. The extracted DNA was checked by agarose gel electrophoresis, quantified and quality-checked by NanoDrop ND-2000C (Implen, US) before storage. Equal amounts of DNA from the 10 goat rumens were mixed for sequencing. The mixed metagenomic DNA showed only slight degradation, and was concentrated to 132 mg/ml (OD260/280 value of 1.92). Total 10 mg of the DNA was sent to BGI-Hong Kong Co. Ltd. for sequencing.

2.3.2 Metagenome sequencing and assembly

The paired-end library was prepared as described elsewhere (Do et al. 2014). The metagenomic DNA was sequenced using next generation ultra high throughput sequencing system Illumina HiSeq2500 (Illumina, San Diego, CA, USA). The raw sequence data was analyzed using a standard bioinformatics approach as follows. Adaptor sequences and reads containing >10% "N" bases, and reads containing >50% low quality base scores (Q < 20), were removed from the raw data. The reads were then assembled by SOAPdenovo2 (Luo et al. 2012) with different k-mer sizes in parallel, and Rabbit tool (You et al. 2013) was used to extend the length of SOAPdenovo-derived contigs. Reads were then mapped back to the final contigs for each assembly in order to choose the most optimal k-mer size and to select the best assembly with regard to N50 and coverage. Contigs with a length of >200 bps were kept for open reading frame (ORF) prediction using

MetaGeneMark (W. Zhu, Lomsadze, and Borodovsky 2010). The predicted genes were further clustered using CD-hit (W. Li and Godzik 2006). The genes having a sequence identity ≥95% and alignment coverage ≥95% were merged and kept for functional annotation.

2.3.3 Functional annotation

All the predicted ORFs were blasted against public databases: (i) Swiss-Prot; (ii) Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2008); (iii) Non-supervised Orthologous Groups (eggNOG); (iv) Cluster of Orthologous Groups (COG) (Powell et al. 2012); (v) Carbohydrate-Active enZYmes (CAZy) database (Bernard et al. 2008); and (vi) Gene Ontology (GO) (Ashburner et al. 2000). A flow chart of the bioinformatics pipeline for analysis of the bacterial metagenomic DNA extracted from Vietnamese native goats' rumen is represented in Fig. S2. Top hits, those with an E-value lower than 10^{-5} and a sequence coverage >50%, and the highest sequence similarity, were used for further analysis. Lignocellulolytic enzyme families (Pfam, protein families) were predicted by performing Interproscan (https://www.ebi.ac.uk/interpro/).

For taxonomic classification, homology of the mapped ORFs was queried to previously characterized ORFs in the non-redundant (NR) database in NCBI. In this way, organism diversity was obtained in the goat rumens at the phylum level. For the annotation of species, the best matching ORFs, whose E value was lower than e^{-5}, were preserved in the classified

group for further analysis. The ORFs in the classified group were subjected to MEGAN (version 4.6) (Huson et al. 2007) for assignment into NCBI taxonomy using the lowest common ancestor (LCA) algorithm.

This project was deposited in the DNA Databank of Japan (DDBJ) with the accession ID PSUB006562.

2.3.4 CBM63 expression, purification and activity analysis

For assessment of the functional annotated results, the ORF 57,823 encoding CBM63 was chosen for *E. coli* gene expression and activity analysis.

This gene (858 bps) contains a 5' terminal sequence of 78 bps encoding a signal peptide and another sequence spanning the remaining 777 bps codes for a mature CBM63. The gene encoding mature CBM63 was synthesized by Genescript (USA) and cloned into pET22b(+) (Novagen) at *Nco*I and *Xho*I restriction sites. The obtained plasmid was introduced into *E. coli* BL21 (DE3) (Novagen). For protein expression, a single-colony transformant was inoculated into 10 ml Luria-Bertani broth (supplemented with 100 mg/ml ampicillin; LBA), and grown overnight at 37°C in a rotary shaker (200 rpm). The overnight culture (0.2 ml) was then transferred to 20 ml of fresh LBA and cultivated at 37°C, 200 rpm until the optical density (OD_{600}) reached 0.6–0.8. Subsequently, the cells were induced for CBM63 expression by adding 0.5 mM IPTG and continuously grown for 4 hours at 25°C. The cells were harvested by centrifugation at 6,000 rpm for 10 min at 4°C, and suspended in water to a density of OD_{600} = 10. The cells were disrupted by sonication on an ice bath (10 pulses, 30 s each at 100 W with 20 s intermission). The soluble fraction was separated from the pellet by centrifugation at 13,000 rpm for 10 min at 4°C. The expressed proteins in soluble and insoluble fractions were analyzed by SDS-PAGE. The recombinant CBM63 was purified by Immobilized Metal Affinity Chromatography (IMAC) with a 5 mL Ni-charged Sepharose Fast Flow column (HisTrap; GE Healthcare). Before loading the sample, the column was equilibrated with 10 column volumes (CV) of buffer (20 mM KH_2PO_4, 0.5 M NaCl, pH 7.4) containing 50 mM imidazole. After applying the soluble fraction to the column, it was washed with 5 CV of the same buffer containing 100 mM imidazole, and eluted by 10 CV of the buffer containing 500 mM imidazole. The protein concentration in the purified fractions was measured by NanoDrop ND-2000C (Implen, US) and was checked by electrophoresis SDS-PAGE and then desalted using a PD10 desalting column (GE Healthcare).

The purified CBM63 was used to check the activity with carboxymethyl cellulose (CMC, Sigma) and filter paper as substrates. Whatman No. 1 filter paper was cut into very small pieces by scissors. The total reaction volume was 0.5 ml containing: 10 mg of the filter paper (or 0.1 mg CMC); 0.05 ml of 0.05 M Na-citrate buffer, pH 6; and 0.3 mg purified CBM63 protein with, or without, 0.025 U of cellulase (Sigma). The reaction was performed at 50°C for exactly 90 min and then stopped immediately by adding 0.5 ml of dinitro salicylic (DNS) reagent. All the tubes were boiled for 5 min and the absorbance at 540 nm was measured. Each measurement was made in triplicate. The activity of the protein was calculated as the amount of reducing sugar (corresponding to mM glucose in this study) released (Miller 1959).

2.4 Results and Discussion

2.4.1 Sequencing analysis

Illumina sequencing of the metagenomic DNA yielded 89,964,640 reads. Of these, 84,625,346 reads (94.07%) were useful reads used for assembly to 172,918 contigs larger than 200 bp by a SOAPdenovo assembly tool using a k-mer size of 51. From the contigs, 164,644 ORFs were predicted (Table S1). The inventory of ORFs length distribution is shown in Fig. S3.

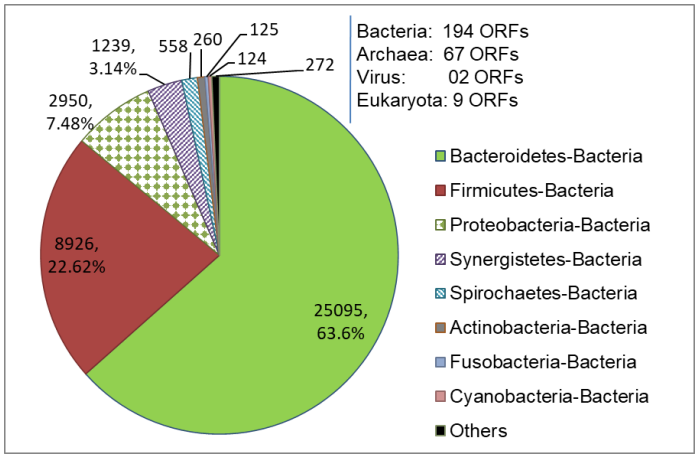


Figure 1: *Analysis of the goat rumen microbial community structure at the phylum level. The numbers of ORFs affiliated in each phylum and its percentage are indicated however percentage is not indicated for less than 3.14%.*

Similarity searches using BLAST against the non-redundant protein sequence database showed that 122,304 ORFs (74.3%) retrieved a Blast hit. The nrBLAST output was subjected to MEGAN (version 4.6) (Huson et al. 2007) for taxonomic assignment. Among 39,579 ORFs affiliated in taxonomic classification, most of the genes (99.8%) originated from bacteria. Only nine ORFs belonged to Eukaryota, two ORFs originated from viruses and 67 ORFs were classified to Archaea (Fig. 1). This confirms the enrichment of bacterial DNA during the metagenomic DNA extraction of goat rumen samples.

Among the bacterial genes, phylum Bacteroidetes was the most represented, accounting for 63.6%, followed by Firmicutes (22.6%), and Proteobacteria (7.5%) (Fig. 1). Also, these phyla are most abundantly represented in the microbial eco-system in Japanese goat rumens (Denman et al. 2015). Earlier studies showed that the dominance of Bacteroidetes is correlated to the presence of cellulolytic glycoside hydrolases (GH), which play an important role in lignocellulose degradation (Güllert et al. 2016; Han et al. 2015). The dominance of Bacteroidetes may reflect high lignocellulolytic degradation activity in the goat rumen.

For functional annotation, the ORFs were blasted against diverse databases. In total, 141,521 ORFs were annotated. Typical eukaryotic COG categories, such as RNA processing and modification, chromatin structure were almost not represented in our data set (Fig. S4). This result supports again the observation that our metagenomic DNA extraction was highly enriched for bacterial DNA.

The number of ORFs matching to each of the COG, eggNOG, KEGG, GO, CAZy, and Swiss-Prot, databases were 37,007 (Fig. S4), 134,843; 56,751; 86,693; 7,898 and 33,471 ORFs, respectively. However, in this study we are specifically interested in gene functions related to carbohydrate metabolism. As such, 3,642 genes could be annotated to the COG category carbohydrate transport and metabolism, while 17,984 ORFs received this annotation with eggNOG. Moreover, a subset of this gene set (11,999 ORFs) could be identified to be involved in the carbohydrate metabolism category in KEGG. As expected, almost all genes annotated by CAZy databases (7,898 ORFs) were valuable for mining carbohydrate degrading enzymes.

2.4.2 Functional annotation showed an abundance of ORFs encoding putative enzymes/proteins for lignocellulose degradation

The CAZy annotation exhibited mainly four kinds of catalytic domain related to carbohydrate degradations that comprised GHs (4,715 ORFs), glycosyl transferases (GTs: 1,956 ORFs), polysaccharide lyases (PLs: 229 ORFs), and carbohydrate esterases (CEs: 969 ORFs). The 4,715 ORFs classified in GHs categories were divided into 65 GH families (Table 1). According to CAZy, within these families, 11 GHs belonged to cellulases (Table S2), 22 GHs belonged to hemicellulases (Table S3), and 32 GHs contain other activity domains. Unfortunately, no enzyme responsible for lignin-degradation, such as Mn-peroxidase or laccase, was found. An earlier study showed that the lignin degradation process in ruminants is usually limited in rumen, due to the anaerobic conditions (Susmel and Stefanon 1993). Microflora in animal rumen is constituted by facultative anaerobic bacteria (1–10 x 10^9 per ml) protozoa and fungi, however fungi are found to play a predominant role in lignin degradation (Kasuya et al. 2007; Susmel and Stefanon 1993), while bacteria and protozoa are responsible for the efficient degradation of cellulose and hemicellulose (Moreira et al. 2013; Susmel and Stefanon 1993). In addition, lignin was revealed to have a positive function in the rumen to help maintain the reservoir of buffering exchangeable cations for feed digestion (Moreira et al. 2013).

Carbohydrate esterase families (CE and PL) are known to enhance lignocellulose pretreatment. The ORFs encoding these families were found in abundance in our data with a total of 821 ORFs. Most CEs were related to pectin esterase, only CE6 was suggested to be reductase and carboxylesterase. We identified 61 ORFs that contain both hemicellulase (GH10), as well as esterase (CE1), domains. In addition, another 19 ORFs encoded bifunctional domains. Within this group, 18 ORFs encoded both hemicellulase GH26 and esterase CE7, while 1 ORF encoded a protein with hemicellulase GH43 and esterase CE6 domains (Table S4). The enzyme with a bifunctional domain may be useful for application because, at the same time, two activities can be synergistically exhibited simultaneously for improving the substrate degradation (Neddersen and Elleuche 2015). The ORFs divided into PL groups in this study mostly have catalytic domains for pectin degradation (Table S4).

Table 1. *Summary of CAZy annotation of the genes from bacterial metagenomic DNA extracted from Vietnamese native goats' rumen.*

Name	ORFs	Name	ORFs	Name	ORFs	Name	ORFs	Name	ORFs	Name	ORFs
CBM:	763	CE:	969	GH:	4,715	GH2	372	GH5	192	GT:	1,956
CBM0	11	CE1	257	GH0	30	GH20	40	GH51	138	GT0	28
CBM2	13	CE10	163	GH1	16	GH23	105	GH53	79	GT1	7
CBM3	3	CE11	47	GH10	116	GH24	37	GH57	45	GT10	4
CBM4	11	CE12	104	GH103	3	GH25	109	GH63	1	GT11	40
CBM6	122	CE13	3	GH105	112	GH26	98	GH64	1	GT19	45
CBM9	2	CE14	2	GH106	46	GH27	19	GH66	2	GT2	933
CBM13	31	CE15	35	GH108	6	GH28	210	GH67	58	GT23	1
CBM20	66	CE2	33	GH109	4	GH29	67	GH73	62	GT26	19
CBM22	2	CE3	1	GH11	2	GH3	400	GH74	1	GT28	60
CBM25	2	CE4	66	GH112	1	GH30	16	GH77	115	GT3	35
CBM32	62	CE6	105	GH113	1	GH31	152	GH78	65	GT30	52
CBM34	6	CE7	68	GH115	121	GH32	61	GH8	48	GT32	20
CBM35	26	CE8	75	GH119	1	GH33	42	GH84	3	GT35	74
CBM37	56	CE9	10	GH120	4	GH35	75	GH88	7	GT4	397
CBM38	2	PL:	229	GH125	4	GH36	52	GH89	17	GT41	2
CBM41	2	PL1	108	GH127	62	GH38	1	GH9	46	GT5	63
CBM48	127	PL10	36	GH13	326	GH4	2	GH91	1	GT51	111
CBM50	205	PL11	76	GH130	44	GH42	1	GH92	37	GT8	6
CBM57	9	PL9	9	GH16	35	GH43	641	GH94	50	GT83	6
CBM61	4			GH18	14	GH44	2	GH95	115	GT9	53
CBM63	1			GH19	3	GH48	1	GH97	178		

For cellulose degradation, the functional annotation has assigned 816 ORFs encoding cellulases, which were categorized in 11 GHs (Table S2). While, according to CAZy, GH16, GH5, GH8, GH9 were related to endoglucanase, GH3 was beta-glucosidase, and GH16 was suggested to be glucan endo-1,3-beta-D-glucosidase and licheninase. For hemicellulose degradation, after integration of COG, KEGG and GO annotated results, a total of 2,252 ORFs were predicted to encode hemicellulases, including endo1,4-beta-xylanase, beta-xylosidase, and 20 kinds of branching enzymes (Table S3).

Besides the catalytic core, many of lignocellulases possess non-catalytic, but functionally important, domains for their activity. These domains include CBM, FN3, dockerins, Ig, and so-called unknown "X" domains. CBM has an affinity to an individual or bundled polysaccharide chains, as well as to single carbohydrate molecules. Thus, it anchors or directs host enzymes to targeted carbohydrate substrates (Guillén, Sánchez, and Rodríguez-Sanoja 2010). In some cases, CBM exerts the ability to disrupt crystalline cellulose microfibrils to assist cellulase reactions (Ding et al. 2008; Wilson 2008). In this study, 763 ORFs harbouring domains of 21 types of CBM, including a CBM63 (which may possess expansin activity to disrupt the crystal structure of lignocellulose), were mined (Table S5). In this, 15 types of CBMs (480 ORFs) were collocalized with cellulase (9 ORFs), and hemicellulase domains (241 ORFs) (Tables S2 and S3). Interestingly, all CBMs collocated with endoglucanase catalytic domain and endo 1,3-beta-D-glucosidase catalytic domain, but never co-localized with beta-glucosidase domain (which accounted for ~50% predicted cellulases). This suggests that, during cellulose degradation, endoglucanase first opens up the cellulose structure and subsequently digests cellulose into cellobiose and other small polysaccharides. Apparently, this enzyme needs CBM for more affinity to the substrate to function more optimally. Overall, CBM domains presented in 10% ORFs encoded hemicellulases and 1% ORFs encoded cellulases. In a previous study, Dai et al. (2012) also described 10% of the plant cell wall-targeting GH proteins carrying a CBM. CBM4 and CBM22 have the capacity to bind to xylan and beta-1,3/beta-1,4-glucans, while CBM22 has a thermo-stabilizing effect for catalytic domains (Araki et al. 2006). Interestingly, CBM4 domain was identified in CE1, and CBM22 was collocated with CE3. In the group of hemicellulases having CBM domains, endo-1,4-betaxylanase accounted for 30.6% (23 ORFs). Thus, the presence of CBM domain is clearly associated with enzymes participating in the first step of lignocellulose degradation for enhancing the enzyme affinity to more effectively decompose substrate.

The fibronectin-3-like module is known to loosen up the cellulose surface, and may separate cellulose chains and expose additional sites of cellulose for hydrolysis by the covalently-attached catalytic domain (Kataeva et al. 2002). In our study, 214 ORFs with FN3 domains were observed to be collocated with GH5 (1 ORF), and GH3 domains (213 ORFs for beta-glucosidase) (Table S2). This is in agreement with the finding in a previous study that beta-glucosidase did not harbour CBM but contained an FN3 domain (Sweeney and Xu 2012). Another previous study showed that bacterial FN3 sequences were identified only in extracellular matrix proteins (Kataeva et al. 2002). This suggests that many bacterial beta-glucosidases are secreted into the goat's rumen, playing an important role in the transformation of cellulose to glucose as a nutrient for the goat, rather than providing a carbon source for bacterial consumption.

In this study, we also identified Ig domains (30 ORFs) responsible for stabilization and enhanced thermo-stability of collocated catalytic domains, accompanied by only GH9 catalytic domains. This association is confirmed by another study, where Ig plays a vital role in activating GH9 enzymes (Kataeva et al. 2004).

2.4.3 Comparison of metagenomic data from Vietnamese and Korean goats' rumen in the emphasis of the ORFs for *putative lignocellulases*

We compared ORFs data encoding cellulases to the data published by Lim et al. (2013), and found that the endoglucanase GH8 was present in both datasets in comparable abundance, while GH44 and GH48 were also present in both datasets, although at a low abundance (Fig. 2) (Lim et al., 2013). This result may reflect the presence of a well-defined group of cellulase GHs that have evolved as a specific adaptation to the specific digestive circumstances in goat rumen. However, these two studies also differ considerably. For instance, GH9, GH44, and GH10 represent endoglucanases that were identified in both Korean and Vietnam goats' rumen, although at a lower abundance (~1/2 times) in our sample. The same pattern is observed in the case of cellulase PF00150, where a 37 times greater abundance was identified in the Korean goat rumen as compared with the Vietnamese goat rumen. In contrast, GH5, which is responsible for endoglucanase, showed a 6.4 times greater abundance in the Vietnam data (Fig. 2). Some GHs were only observed in the Korean goat rumen data, but were absent in the Vietnamese goat rumen data. Whereas, many GHs for cellulase activity were observed

in the Vietnamese goat rumen sequences, but were absent in Korean goat rumen data. For instance, endoglucanase GH45 and cellobiose hydrolase (CBH) were identified only in the Korean goat rumen, with a total of 241 genes. Meanwhile, we observed 453 ORFs assigned to the families GH1, GH16, GH3, GH64, and GH74, and comprising endoglucanase, beta-glucosidase, 6phospho-beta-glucosidase, beta-glucosidase-related glycosidase, cellobiose phosphorylase, glucan endo-1,3-betaD-glucosidase, and licheninase activities, only in the Vietnamese goat rumen data (Table S2, Fig. 2). Overall, the total genes annotated by the same databases (CAZy) for cellulases were 749 genes in our study and 687 genes in the Korean goat rumen data (Fig. 2).

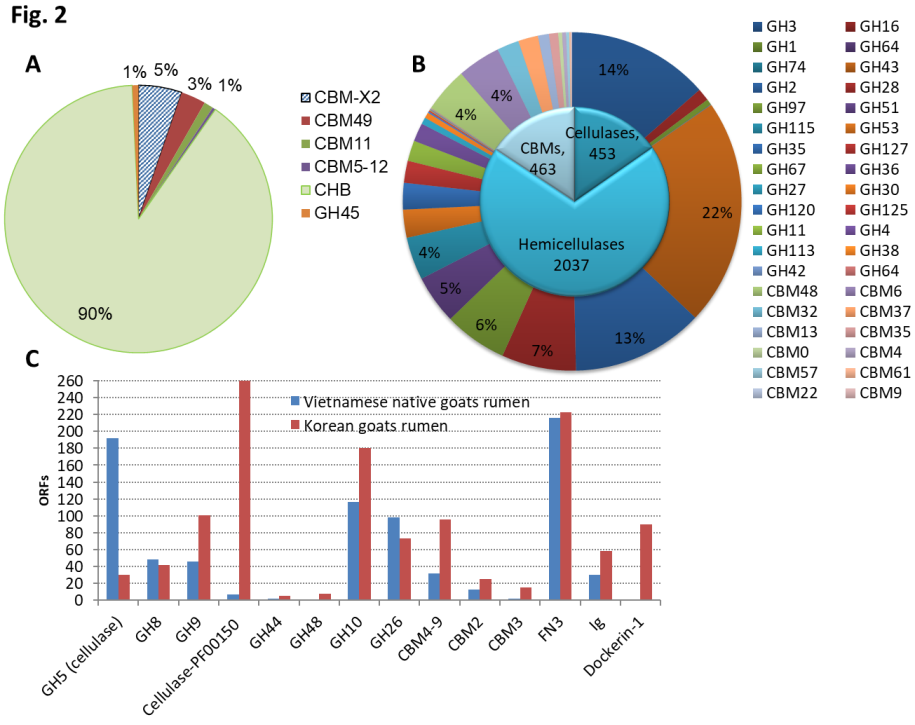


Figure 2: *Comparison of modules located in lignocellulolytic enzymes in the metagenomic data of Vietnamese native goats rumen and Korean goats rumen. Carbohydrate-binding modules (CBM), glycoside hydrolases (GHs) for cellulases and hemicellulases, and other unfunctional domains located in lignocellulolytic enzymes annotated by CAZy and Interpro, that were observed only in the metagenomic data of Korean goats' rumen (A) (Lim et al., 2013) or exhibited only in Vietnamese native goats' rumen (B) and found in both metagenomic data of Vietnamese and Korean goat rumen bacteria (C).*

The size of the Korean goat rumen data is 2.4 fold greater than the Vietnamese goat rumen metagenomic data. These results indicate that bacterial cellulase genes in the rumen of the Vietnamese native goats are more abundant than those in the Korean goats. In addition, in our study, some ORFs could not be annotated into the GH family, but were still predicted to have an activity linked to beta-glucosidase, cellulase M/endoglucanase, and endoglucanase. The total number of ORFs assigned from all databases for cellulases were 816 ORFs (Table S2). The difference in cellulase genes may lie in the bacterial sources. Several studies have provided evidence that the rumen microbiome can be influenced and shaped by the host genotype (An, Dong, and Dong 2005; Hess et al. 2011; Kittelmann and Janssen 2011; Nelson et al. 2003; Sundset et al. 2007), diet preference (Han et al. 2015; Tajima et al. 2001; Z. Zhu et al. 2014), as well as the habitat (Sundset et al. 2007). In the case of the host genotype, the Korean goats used for mining lignocellulase genes represent the Saanen hybrid line, and in this study we used genotypes derived from Co and Bach Thao hybrid lines. In general, these genotypes are omnivorous animals, feeding mainly on natural plants, leaves, agricultural waste such as straw, cornstalks, and sugarcane tops. However, we chose goats living in a mountainous area and feeding particularly on various plant and agriculture waste.

Although the overall abundance of cellulase genes is comparable to the rumen data of Vietnamese and Korean goat rumen data, the distribution of specific GH enzymes differs considerably between the two studies. This supports the notion that effective hydrolyzation of cellulases in any lignocelluose-degrading ecosystem is highly diverse and cannot be linked to a specific group of catalytic domains represented by a defined set of enzymes (Hu et al. 2013; M. Liu et al. 2013; Tiwari, Misra, and Sangwan 2013).

According to the CAZy annotation, 22 GHs having hemicellulase activities were found (Table S3). However, only GH10 and GH26 were observed in both metagenomic data from Korean and Vietnamese goat rumen. Overall, the absolute number of genes belonging to GH10 and GH26 in Korean goat rumen data (~256 ORFs) was slightly higher than in Vietnamese goat rumen data (214 ORFs). In contrast, the other 20 GHs, which accounted for 2,037 ORFs, were observed in our data but not described in the Korean dataset (Fig. 2). This suggests that bacteria in Vietnamese goat rumen adapted specifically to the digestion of diverse lignocellulose materials in the tree and dry crop residues, which may be harsher to digest when compared with digesting lignocellulose present in young leaves.

Of the 2,252 ORFs predicted to have hemicellulase activities, 20 kinds of branching enzymes were identified (Table S3). Remarkably, all the branching enzymes were absent in the Korean goat rumen dataset (Lim et al. 2013). The high abundance of hemicellulases in our metagenomic dataset may be explained by the specific diet requirements of Vietnamese native goat breeds.

The CEs and PLs were not represented at all in the bacterial metagenomic data from Korean goats rumen (Lim et al. 2013), indicating that the present dataset from Vietnam goat rumen is more diverse in the number and function of genes. The number of CEs and PLs genes affiliated to Bacteroidetes were approximately 16 times higher than that affiliated to Firmicutes. Detailed results will be published in the future.

The four most abundant CBMs (CBM6, CBM50, CBM48, CBM32) of the 21 CBM types in our data were also identified to be the four most abundant CBMs in cow rumen (Hess et al. 2011). When comparing our data with data from Korean goat rumen (Lim et al. 2013), CBM2, CBM3, and CBM4-9 were identified in both datasets, but their abundance was much lower among the Vietnamese sequences. Other CBM domains (CBM5-12, CBMX-2, CBM11, CBM19) were completely lacking in our data. In contrast, 12 CBMs among 463 ORFs were annotated only in the Vietnamese dataset. In total, 510 ORFs were annotated in our data, which was threefold higher compared with the CBM-containing genes in the Korean dataset (162 genes) (Fig. 2). Bacterial expansin is usually found in strains belonging to *Bacillus subtilis* (Kerff et al. 2008) and *Hahella chejuensis* (H. J. Lee et al. 2010), which are involved in disrupting the crystal structure of lignocellulose, enabling other cellulases to further depolymerize the liberated polysaccharides. After an extensive search in our data, we found only one gene for expansin that was annotated to be CBM63 according to CAZy. Finally, it is worth mentioning that expansin was not identified and described in Korean goat rumen (Lim et al. 2013), again indicating the more diverse and rich content of the Vietnamese goat rumen microbiome.

In agreement with the previous study in goat rumen, dockerin type I was only annotated in GH9, supporting previous observations (Borne et al. 2013; Hirano et al. 2015; Lim et al. 2013). Dockerin type I only exists in cellulosome modularity (Borne et al. 2013; Hirano et al. 2015). The low abundance of dockerin type I in this sample indicates that a cellulosome structure is not established in the Vietnamese goat rumen microbiome. This is supported by

the fact that we also did not find any cohensin, dockerin type II in this data, which is essential for cellulosome assembly.

In Korean goat rumen metagenomic data, no clear correlation was found between FN3 and a specific catalytic domain (Lim et al. 2013).

2.4.4 Expression of CBM63 for preliminary confirmation of annotated results

With regard to the confirmation of functional annotated results of the genes from metagenomic data, a nucleotide sequence of 777 bp encoding for mature CBM63 was expressed in *E. coli*. By MEGAN analysis, the CBM63 was assigned to be originated from *Ruminococcus flavefaciens*.

Fig. 3

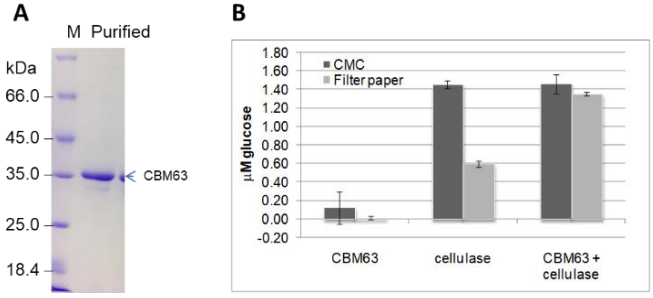


Figure 3: *Expression of CBM63 protein in E. coli. SDS-PAGE analysis of purified CBM63 from recombinant E. coli extract (A), and assessment of CBM63 ability to enhance cellulase activity by DNS method (B). M: Standard proteins (Fermentas).*

In the amino acid sequence, CBM63 was the most closely identical with expansin of *Clostridium* sp Marseille-P2415 NCBI (WP_077613372.1, 45%) and *Bacillus atrophaeus* NCBI (WP_061669738.1, 43%). CBM63 also possesses a conserved catalytic domain of endoglucanase at the C terminus. In *E. coli*, a substantial part of the expressed CBM63 (30 kDa) was soluble and highly accumulating in *E. coli* cells. The recombinant protein was successfully purified by His-tag affinity column (Fig. 3). The purified and desalted CBM63 did not exhibit endoglucanase activity to digest CMC but was capable of significantly enhancing commercial cellulase activity to convert filter paper (a typical crystal lignocellulose) into reducing sugars as detected by DNS reagent.

With the purpose of confirming the functional annotated results of the genes from metagenomic data, Hess et al. (2011) mined 27,755 putative carbohydrate-active genes from cow rumen's metagenomic data and expressed 90 candidate proteins which had an amino acid sequence identity to known carbohydrate-active proteins ranging from 26% up to 96%. They discovered that 57% recombinant proteins exhibited enzymatic activity. There was no link between enzymatic activity with the degree of amino acid sequence identity (Hess et al. 2011). In agreement with this study, we also expressed seven other cellulose-, hemicellulose-, pectin-active genes in *E. coli*, of which five showed enzymatic activities and the remaining enzymes were expressed at too low a level (data will be published elsewhere). This indicates that the majority of mined genes possess actual activity.

In conclusion, we were able to annotate a wide diversity of hemicellulase genes that are associated with CBMs in our samples. We also observed CBMs located in cellulases and enzymes for lignocellulose pretreatment, but to a much lesser extent. The FN3 domain was in high abundance and showed a clear association with GH3, while the Ig domain was more linked to GH9. This resource will be highly useful, when recombinant enzyme assays are needed to be applied as cocktail enzymes to accomplish a more optimal industrial degradation of lignocellulose.

Acknowledgments

We would like to acknowledge Dr. S. V. N. Vijayendra (Food Microbiology Dept., Central Food Technological Research Institute, Mysore 570020, India) for proofreading and correcting the English of this manuscript. The study was carried out with the financial support of the Project "Metagenome of some potential mini-ecologies for mining novel genes encoding effective lignocellulolytic enzymes" code DTDLCN.15/14, managed by the Ministry of Science and Technology, Vietnam, in collaboration with the Department of Ecological Science, Vrije Universiteit Amsterdam, The Netherlands, supported by the BE-BASIC consortium project numbers F07.003.05 and F07.003.07. We thank the National Key Laboratory of Gene Technology, Institute of Biotechnology, VAST, Vietnam, for the use of their facilities. We are also grateful to the Editor of JGAM for valuable comments to improve the manuscript.

Supplementary Materials

Figure S1: Vietnamese goats species

Figure S2: Bioinformatic pipeline

Figure S3: Contig lengths distribution

Figure S4: Metagenomic COG summary

Table S1: Assembly information

Table S2: List of cellulotic enzymes

Table S3: List of hemicellulotic enzymes

Table S4: List of lignocellulose enzymes

Table S5: List of carbohydrate binding model

Figure S1: *Goat breed Co (A) and Bach Thao (B) in Ninh Binh, Thanh Hoa, Vietnam*

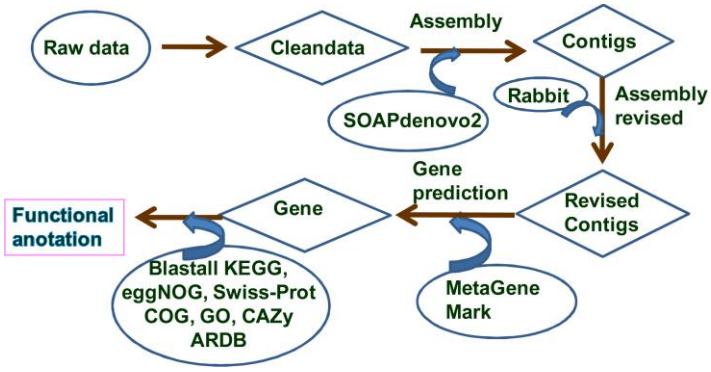


Figure S2: *Bioinformatics pipeline for analysis of metagenomic DNA data of bacteria extracted from Vietnamese native goats' rumen.*

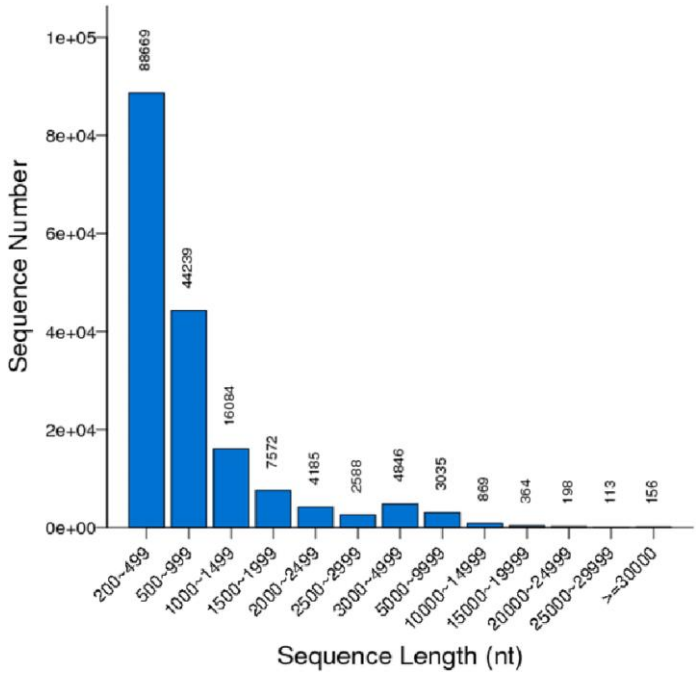


Figure S3: *Length distribution of the optimal assemblies. The horizontal axis corresponds to the contig length, and the vertical axis corresponds to the contig number*

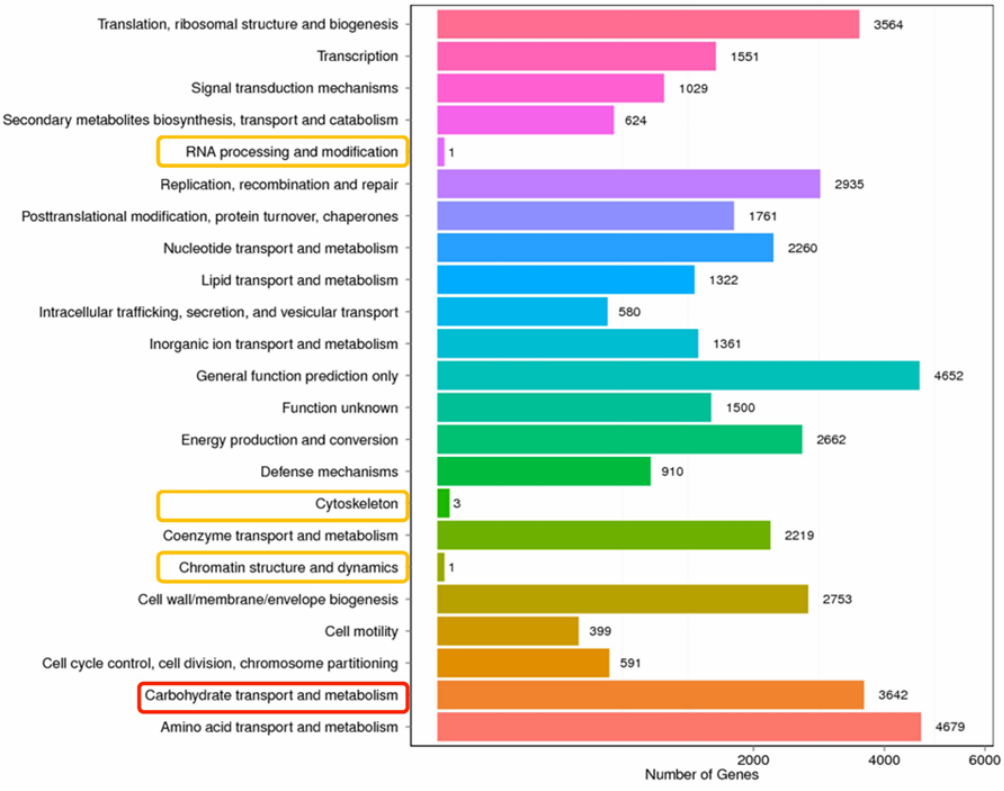


Figure S3: *Length distribution of the optimal assemblies. The horizontal axis corresponds to the contig length, and the vertical axis corresponds to the contig number*

***Table S1**: Illumina'sHiSeq sequencing and SOAPdenovo2 assembly metrics of thebacterial metagenomic DNA in the rumen of goats collected in Vietnam.*

Parameter	Metric
Total number of reads	89,964,640
Clean Data Size (bp)	8,462,534,600
Mean read length (nt)	100
Assembly length (bp)	178,514,501
Number of contigs	172,918
Largest contig (bp)	124,798
Shortest contig (bp)	200
Mean contig length (bp)	1,032
N50 contig length (bp)	1,879
N90 contig length (bp)	374
Number of ORFs	164,644

Table S2: *Inventory of putative genes encoding cellulolytic enzymes annotated byCAZy, COG, KEEG databases in Vietnamese native goats' rumen.*

GH (ORFs)	COG (ORFs)	COG/KEGG (ORFs number)	EC
GH1 (16)	COG2723 (10)	6-phospho-beta-glucosidase (13)	3.2.1.86
		Beta-glucosidase (1)	3.2.1.21
	COG2723 (1)	Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase (1)	
GH16 (33)	COG2273 (1)	Beta-glucanase (1)	
		Glucan endo-1,3-beta-D-glucosidase (5)	
		Licheninase (1)	
GH16-CBM4 (2)			
GH3 (238)	COG1472 (39)	Beta-glucosidase-related glycosidase (39)	3.2.1.21
		Beta-glucosidase (170)	3.2.1.21
		Beta-N-acetylhexosaminidase (25)	3.2.1.52
GH3-FN3 (202)	COG1472 (61)	Beta-glucosidase-related glycosidase (61)	3.2.1.21
		beta-glucosidase (38)	3.2.1.21
GH44 (2)			
GH48 (1)			
GH5 (190)	COG2730 (9)	Endoglucanase (9)	3.2.1.4
	COG2730 (1)	Endoglucanase (1)	
	COG3934 (24)	Endo-beta-mannanase (24)	
		Endoglucanase (26)	3.2.1.4
GH5-CBM2 (1)	COG2730 (1)	Endoglucanase (1)	3.2.1.4
GH5-CBM37 (1)	COG2730 (1)	Endoglucanase (1)	3.2.1.4
GH5-FN3 (1)	COG2730 (1)	Endoglucanase (1)	3.2.1.4
GH64-CBM6 (1)			
GH74 (1)			
GH8 (48)		endoglucanase (8)	3.2.1.4
	COG3405 (9)	Endoglucanase Y (9)	3.2.1.4
GH9 (11)		endoglucanase (7)	3.2.1.4
GH9-Ig (30)		endoglucanase (3)	3.2.1.4
GH9-dockerin type 1 (1)			
GH9-CBM3 (1)		Endoglucanase (1)	3.2.1.4
GH9-CBM3 (1)			
GH9-CBM37 (1)			
GH94 (50)	COG3459 (27)	Cellobiose phosphorylase (27)	2.4.1.20
FN3 (11)		beta-glucosidase (20)	3.2.1.21
	COG1363 (1)	Cellulase M/endoglucanase (1)	3.2.1.4
		endoglucanase (3)	3.2.1.4
CBM63 (1)	COG4305 (1)	Endoglucanase C-terminal domain (1)	
Total		**816 ORFs**	

Table S3: *Inventory of putative genes encoding hemicellulolytic enzymes annotatedby CAZy, COG, and KEEG databases in Vietnamese native goats' rumen.*

GHs (No. of ORFs)	COG (No. of ORFs)	COG/KEGG/GO (No. of ORFs)	EC
CE1 (1)	COG3693 (1)	Endo-1,4-beta-xylanase (1)	3.2.1.8
GH10 (54), GH10-CE1 (25)	COG3693 (23)	Endo-1,4-beta-xylanase (52)	3.2.1.8
GH10-CBM6 (16), GH10-CBM0 (9), GH10-CBM22 (1), GH10-CBM9 (1)		Endo-1,4-beta-xylanase (12)	3.2.1.8
GH11 (1)		Endo-1,4-beta-xylanase (1)	3.2.1.8
GH2 (358)	COG3250 (12)	Beta-galactosidase (179)	3.2.1.23
	COG3250 (1)	Beta-glucuronidase (30)	3.2.1.31
GH2-CBM32 (4)		Beta-galactosidase (1)	3.2.1.23
GH26 (48), GH26-CBM35 (17),	GH26-CE7 (42)	Mannan endo-1,4-beta-mannosidase (63)	3.2.1.78
GH27 (11), GH27-CBM35 (8)		Alpha-galactosidase (10)	3.2.1.22
		Alpha-N-acetylgalactosaminidase (1)	3.2.1.49
GH28 (194)	COG4225 (2)	Unsaturated glucuronyl hydrolase (2)	
	COG5434 (35)	Endopolygalacturonase (35)	3.2.1.67
		Endo-1,4-beta-xylanase (1)	3.2.1.8
		Galacturan 1,4-alpha-galacturonidase (3)	3.2.1.67
GH30 (16)		Glucosylceramidase (7)	3.2.1.45
GH35 (64), GH35-CBM32 (11)	COG1874 (11)	Beta-galactosidase (31)	3.2.1.23
GH36 (52)	COG3345 (2)	Alpha-galactosidase (52)	3.2.1.22
GH38-CBM32 (1)		Alpha-mannosidase (1)	3.2.1.24
GH4 (2)	COG1486 (1)	6-phospho-alpha-glucosidase (1)	3.2.1.122
	COG1486 (1)	Alpha-galactosidases/6-phospho-beta-glucosidase (1)	3.2.1.22
GH42 (1)		Beta-galactosidase (1)	3.2.1.23
GH43 (492)	COG3507 (4)	Beta-xylosidase (10)	3.2.1.37
	COG3940 (2)	Beta-xylosidase/alpha-N-arabinofuranosidase (2)	3.2.1.55
	COG3507 (5)	Beta-xylosidase/arabinan endo-1,5-alpha-L-arabinosidase (5)	3.2.1.99
	COG3507 (3)	Beta-xylosidase/alpha-N-arabinofuranosidase (3)	3.2.1.55
	COG3507 (17)	Beta-xylosidase (17)	
	COG2017 (35)	Galactose mutarotase/aldose 1-epimerase (35)	5.1.3.3
		Arabinan endo-1,5-alpha-L-arabinosidase (32)	3.2.1.99
		Alpha-N-arabinofuranosidase (56)	3.2.1.55
GH43-CBM13 (29)		Arabinan endo-1,5-alpha-L-arabinosidase (4)	3.2.1.99
GH51 (138)	COG3534 (62)	Alpha-L-arabinofuranosidase (93)	3.2.1.55
GH53 (76)	COG3867 (7)	Arabinogalactan endo-1,4-beta-galactosidase (75)	3.2.1.89
GH53-CBM61 (3)	COG3867 (3)	Arabinogalactan endo-1,4-beta-galactosidase (3)	3.2.1.89
GH67 (58)	COG3661 (34)	Alpha-glucuronidase (37)	3.2.1.139
GH97 (178)		Alpha-glucosidase (74)	3.2.1.20
	COG2731 (1)	Beta galactosidase (4)	
		Carboxylesterase type B (1)	3.1.1.1
Arabinogalactan beta-galactosidase (1), endo-beta-xylanase (2), alpha-n-arabinofuranosidase (1), endo-beta-xylanase cellulase (3), endo-beta-xylanase xyn5a (2), maltodextrin glucosidase (1), mannan endo-beta-mannosidase (2) xylan beta-xylosidase (1)			
GH10-CBM48-CE1 (1), GH113 (1), GH115 (121), GH120 (4), GH125 (4), GH127 (62), GH2-CBM57 (9), GH26-CBM13 (2), GH43-CBM0 (1), GH43-CBM32 (17),GH43-CBM6 (100), GH43-CBM61 (1), GH64-CBM6 (1)			
Total:	**2252 ORFs**		

Table S4: *Inventory of putative genes for the enzymes involved in lignocellulose pretreatment, annotated by CAZy, KEEG, COG, and GO databases in Vietnamese native goats' rumen.*

Enzyme family (ORFs)	COG (ORFs)	COG/KEGG/GO (ORFs)	EC
CE1 (161)	COG3693 (1)	Beta-1,4-xylanase/enterochelin esterase and related enzymes (1)	--
		Enterochelin esterase and related enzymes (122)	--
CE1-GH10 (60)		Enterochelin esterase and related enzymes	
CE12 (104)	COG2755 (22)	Lysophospholipase L1 and related esterases (22)	--
	COG4677 (4)	Pectin methylesterase/pectinesterase (4)	3.1.1.11
		Pectinesterase (24)	3.1.1.11
CE4 (66)	COG0726 (7)	Predicted xylanase/chitin deacetylase (7)	--
	COG0726 (1)	Xylanase/chitin deacetylase (1)	3.5.1.-
	COG0726 (3)	Xylanase/chitin deacetylase (3)	3.5.1.41
		Polysaccharide deacetylase (12)	--
CE6 (103)	COG0656 (2)	Aldo/keto reductases, related to diketogulonate reductase (2)	--
	COG2272 (5)	Carboxylesterase type B (28)	3.1.1.1
		enterochelin esterase and related enzymes (27)	--
CE8 (75)	COG4677 (11)	Pectin methylesterase/pectinesterase (11)	3.1.1.11
		Pectinesterase (37)	3.1.1.11
PL1 (108)	COG3866 (6)	Pectate lyase (6)	4.2.2.2
	COG3866	Pectate lyase (4)	
		Pectate lyase (32)	4.2.2.2
		Pectinesterase (1)	3.1.1.11
PL10 (36)		Pectate lyase (2)	4.2.2.2
		Pectate lyase (3)	
		Pectinesterase (9)	3.1.1.11
PL9 (8)		Pectate lyase (4)	4.2.2.2
CE1-CBM4 (7), CE13 (3), CE2 (33), CE3-GH11-CBM22 (1), CE6-GH43 (1), CE6-GH95 (1), CE7-GH26 (18), CE7 (29), PL9 (1)			
Pectate lyase (6)			
Total	**821 ORFs**		

Table S5: *Inventory of putative genes encoding carbohydrate-binding model (CBM) annotated by CAZy in Vietnamese native goats' rumen.*

CBM family	Briefly functinal description	ORFs
Bind to cellulose or hemicellulose		
CBM0	Carbohydrate-binding modules not yet assigned to a family.	11
CBM13	Cellulose-binding domain family (CBD) XIII.	31
CBM2	CBD II from bacteria.	13
CBM22	Xylan-binding domain has affinity with beta-1,3/beta-1,4-glucans and has a thermostabilizing effect.	2
CBM3	CBD III, cellulose/chitin-binding function	3
CBM32	Binding to galactose and lactose/ polygalacturonic acid /LacNAc, known as X56 modules and related to CBM6 mudules	62
CBM35	Xylan-binding domain and the interaction depends on calcium/ binds to decorated soluble mannans, mannooligosaccharides and beta-galactan	26
CBM37	Broad binding specificity to xylan, chitin, microcrystalline and phosphoric-acid swollen cellulose, as well as more heterogeneous substrates, such as alfalfa cell walls, banana stem and wheat straw, known as X94 modules	56
CBM4	Binding to xylan, beta-1,3-glucan, beta-1,3-1,4-glucan, beta-1,6-glucan and amorphous cellulose but not with crystalline cellulose, known as CBD IV.	11
CBM48	Glycogen-binding function, appended to GH13 modules (AMPK)	127
CBM57	Attached to various glycosidases.	9
CBM6	Cellulose-binding function on amorphous cellulose and beta-1,4-xylan/ also bind beta-1,3-glucan, beta-1,3-1,4-glucan, and beta-1,4-glucan,as CBD VI.	122
CBM63	The CBM63 module of *Bacillus subtilis* expansin EXLX1 has been shown to bind cellulose.	1
CBM61	Appended to GH16, GH30, GH31, GH43, GH53 and GH66 catalytic domains and binds to beta;-1,4-galactan	4
CBM9	Found maninly in xylanases and also binds to cellulose as CBD IX.	2
Bind to other polysaccharides		
CBM20	The granular starch-binding function has been demonstrated in several cases. Interact strongly with cyclodextrins. Often designated as starch-binding domains (SBD).	66
CBM25	Starch-binding function demonstrated in one case.	2
CBM34	Granular starch-binding function, known as X21 modules	6
CBM38	The inulin-binding function, known as X39 modules	2
CBM41	Bind to the alpha;-glucans amylose, amylopectin, pullulan, known as X28 modules	2
CBM50	Attached to various enzymes from families GH18, GH19, GH23, GH24, GH25 and GH73, i.e. enzymes cleaving either chitin or peptidoglycan and other enzymes targeting the petidoglycan such as peptidases and amidases, known as LysM domains	205
	Total	**763**

Chapter 3 - Antimicrobial activity and carbohydrate metabolism in the bacterial metagenome of the soil-living invertebrate *Folsomia candida*

Received: 11 October 2018, Accepted: 27 April 2019

Ngoc Giang Le, Valeria Agamennone, Nico M. van Straalen, Abraham Brouwer and Dick Roelofs

3.1 Abstract

The microbiome associated with an animal's gut and other organs is considered an integral part of its ecological functions and adaptive capacity. To better understand how microbial communities influence activities and capacities of the host, we need more information on the functions that are encoded in a microbiome. Until now, the information about soil invertebrate microbiomes is mostly based on taxonomic characterization, achieved through culturing and amplicon sequencing. Using shotgun sequencing and various bioinformatics approaches we explored functions in the bacterial metagenome associated with the soil invertebrate *Folsomia candida*, an established model organism in soil ecology with a fully sequenced, high-quality genome assembly. Our metagenome analysis revealed a remarkable diversity of genes associated with antimicrobial activity and carbohydrate metabolism. the microbiome also contains several homologs to *F. candida* genes that were previously identified as candidates for horizontal gene transfer (HGT). We suggest that the carbohydrate- and antimicrobialrelated functions encoded by *Folsomia*'s metagenome play a role in the digestion of recalcitrant soilborn polysaccharides and the defense against pathogens, thereby significantly contributing to the adaptation of these animals to life in the soil. Furthermore, the transfer of genes from the microbiome may constitute an important source of new functions for the springtail.

3.2 Introduction

Microorganisms inhabit every type of environment, and many live in association with eukaryotic hosts. These microbes can influence their host's ecology and evolution by contributing to a variety of processes such as digestion, immunity, and protection from pathogens (Engel and Moran 2013). Hexapods are good models to study host-associated microorganisms: they constitute the most diverse and abundant group of eukaryotic organisms on earth, and in many cases the establishment of specific microbial symbioses may have provided the key for their evolutionary success. Some hexapods depend on microbial symbionts for nutritional or defensive purposes (Douglas 2016; Kroiss et al. 2010), suggesting that a good understanding of their biology should include the study of their associated microbes. This has been described as a "new imperative for the life sciences" (McFall-Ngai et al. 2013).

The majority of microorganisms is not accessible through traditional culturing techniques (Rappé and Giovannoni 2003) and metagenomic sequencing is an appropriate tool to study the diversity of species and functions of microbes in different ecosystems (Streit and Schmitz 2004). Metagenomics of insect-associated microbial communities has provided important insights in the interactions between microorganisms and their hosts, including the discovery of metabolites with potential biotechnological applications. For example, metagenomics of a termite's gut microbiota has elucidated the mechanisms underlying wood degradation in this environment, while also identifying bacterial enzymes with interesting hydrolytic functions (Warnecke et al. 2007). Other studies have found that microbial symbionts of insects are important sources of novel antimicrobials (Lin Wang et al. 2015).

The springtail *Folsomia candida* Willem 1902 (Hexapoda: Collembola) is a small invertebrate living in soil environments, where it feeds on fungal hyphae, decaying organic material and microorganisms. This species is a commonly used test organism in ecotoxicology and in ecogenomics (Fountain and Hopkin 2005) and recently its genome and transcriptome have been sequenced (Faddeeva-Vakhrusheva et al. 2017). Approximately 2.8% of the genes in the genome of *F. candida* are of foreign origin, having been acquired from bacteria and fungi through HGT (Faddeeva-Vakhrusheva et al. 2017). Many of these genes are involved in carbohydrate metabolism, specifically in cell wall degradation; these functions may aid the animal in extracting nutrients from polysaccharides resulting from the degradation of plant and fungal biomass in the soil. In addition, several foreign genes are involved in antibiotic biosynthesis (Roelofs et al. 2013; Suring et al. 2017). These genes are strongly induced by stress exposure (Nota et al. 2008; Suring et al. 2016) and it is hypothesized that they may be involved in regulating the composition of gut microbial communities in *F. Candida* (Thimm et al. 1998), or in protecting the springtails from pathogens. In fact, *F. candida* has been shown to be non-susceptible to some microbial pathogens present in soil environments (Broza, Pereira, and Stimac 2001; Dromph and Vestergaard 2002).

Recently, we have shown that bacteria isolated from this springtail display inhibitory activity against a variety of pathogens, including entomopathogenic soil fungi (V. Agamennone et al. 2018). This suggests that the microbiota associated with *F. candida* may be a source of antimicrobial compounds, most likely involved in regulatory and defensive functions. Similar mechanisms have been observed in the honey bee: here, symbiotic lactic acid bacteria

(LAB) active against transient environmental microbes are suggested to play an important role in the establishment and maintenance of a normal gut microbiota through the production of various antimicrobial agents (Vásquez et al. 2012). Furthermore, the gut microbiota of *F. candida* may be involved in the breakdown of dietary component and in the uptake of nutrients. A nutritional role of gut microorganisms has been described for many other invertebrates and animals in general (Valdes et al. 2018; Engel and Moran 2013). Even though the exact role of the gut microbiota in *F. candida* and its potential nutritional and defensive functions still need to be elucidated, we suggest that gut bacteria are an important factor interacting with the springtail, and that they provide physiological traits advantageous to thrive in a microbe-dominated environment such as the soil.

In this paper, we provide the first functional description of the gut bacterial community of a springtail based on a whole-metagenome sequencing approach. We hypothesize that the gut microbiome may aid in nutrient uptake and pathogen defense of *F. Candida* (Engel and Moran 2013), thereby optimizing the fitness of the host. Furthermore, both functions are of potential interest for biobased applications: we identified a number of enzymes involved in lignocellulose break down and encoding compounds with predicted antimicrobial activity. Aside from constituting beneficial traits for an animal living in the soil environment, these functions may also represent good targets for drug discovery and for the development of biotechnological applications. Using a comparative analysis between genes of the gut microbiome and foreign genes in *F. candida*, we have identified functions possibly assimilated by the host through HGT.

3.3 Materials and Methods

3.3.1 Test organism

Folsomia candida individuals originated from a laboratory stock culture ("Berlin strain" VU University Amsterdam) that was originally established from specimens sampled in the field, and then maintained in stable laboratory conditions for several years. Springtails were cultured in plastic boxes with a bottom of plaster of Paris and charcoal. Cultures were kept in climate rooms at 20°C temperature, 75% humidity and a 12 hour light-dark cycle. The springtails were fed dry baker's yeast (Dr. Oetker, Bielefeld, Germany), and they were starved for 2 days prior to DNA isolation.

3.3.2 Sample preparation and DNA isolation

DNA was isolated from four different source samples. Two samples (Fc1 and Fc3) consisted of guts dissected from *F. candida* individuals; one sample (Fc4) consisted of whole springtails; one sample (Fc2) consisted of a mixture of whole animals and dissected guts. Dissected guts were rinsed in sterile PBS and whole springtails were rinsed three times in sterile water before processing. After the washing steps, DNA was directly isolated from two of the samples (Fc3 and Fc4) while additional steps were applied to prepare samples Fc1 and Fc2. For these two samples, we separated bacterial cells from *F. candida*'s cells by using the method described by Engel, Martinson, and Moran 2012, with modifications. The samples were crushed in PBS in a 1.5 ml microcentrifuge tube, using a plastic pestle. The samples were then gently vortexed, to encourage separation of cells, before being passed through 20 µm and 8 µm filters in succession. The filtered samples were centrifuged at 10 000 g for 30 min to harvest cells, and the pellet was resuspended in 200 µl TE buffer. For sample Fc2, an additional step with a density gradient was applied. An 80% Percoll solution in 0.15 mol l^{-1} NaCl was prepared. 1 ml of this solution was placed in a 2 ml microcentrifuge tube and spun at 20,000 g for 20 min to create a gradient. The 200 µl of TE buffer containing the cells was gently placed on top of the gradient, and the tube was centrifuged at 400 g for 20 min. Bacterial cells were then visible as a band and were collected using a pipette. The cells were centrifuged at max speed for 5 min and washed with TE buffer to remove residual Percoll solution. DNA was extracted from all samples using the PowerSoil DNA Isolation Kit (MOBIO Laboratories Inc., Carlsbad, CA, USA) and quantified using a Qubit 2.0 (Invitrogen, Carlsbad, CA, USA).

3.3.3 Library preparation and sequencing

Metagenomic libraries for the four samples were prepared using the TruSeq Nano DNA Library Preparation Kit (Illumina Inc., San Diego, CA, USA) with the following modifications. First, genomic DNA (250 ng) was sheared in a Covaris S2 (Covaris Inc., Woburn, MA, USA) with the following settings: duty cycle 10%, intensity 5.0, bursts per second 200, duration 300 s, mode frequency sweeping, power 23 W, temperature 5.5°C to 6°C. Fragmented DNA was cleaned using Agencourt AMPure XP beads (Beckman Coulter Inc., Brea, CA, USA) to remove short fragments. After end repair, cleaning was performed again to select the appropriate library size (180 bp). Then, 3′ end adenylation and adapter

ligation were performed, and the ligated fragments were subjected to two rounds of clean-up. PCR was used to enrich the ligated DNA fragments. The PCR program started with 3 min at 98°C, followed by eight amplification cycles (20 s at 98°C, 15 s at 60°C and 30 s at 72°C) and a final extension step of 5 min at 72°C. The amplified library was cleaned and its quality was assessed with a Bioanalyzer on a DNA 7,500 chip (Agilent Technologies, Santa Clara, CA, USA). Finally, libraries were equimolarly combined and the concentration of the final pool was checked using a High Sensitivity DNA chip. 10 pmol of barcoded DNA was sequenced on an Illumina HiSeq 2,500 using 125 base, paired end run mode.

3.3.4 Data analysis

Raw reads of the four samples obtained from the sequencer were trimmed using Trimmomatic version 0.36 (Bolger, Lohse, and Usadel 2014) to remove adapters and low quality reads, with the following options: ILLUMINACLIP:TruSeq3-PE. fa:2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:20, MINLEN:36. Metaphlan2 was used to characterize the taxonomic profile of the metagenome (Truong et al. 2015). Bowtie2 (Langmead and Salzberg 2012) was used to create reference genomes for *Folsomia candida* (BioProject accession: PRJNA299291) (Faddeeva-Vakhrusheva et al. 2017), *Wolbachia pipientis* (BioProject accession: PRJNA300838) (Faddeeva-Vakhrusheva et al. 2017), *Saccharomyces cerevisiae* (Assembly accession: ASM105121v1) and *Homo sapiens* (Assembly accession: GRCh38. p7), and to align and identify reads originating from these organisms in the metagenome. SAMtools was used to remove the reads aligned to the reference genomes of the above mentioned organisms from the metagenome. This program was also used to merge all the four sequencing samples together for comprehensive bioinformatic analysis (Heng Li et al. 2009). Only paired ends were extracted with Bedtools (Quinlan and Hall 2010). FastQC (Andrews and others 2010) was used to check the quality of the reads at different processing stages. Assembly was done using SPAdes version 3.9.0 with the (–meta) setting for metagenomic and k-mer values 21, 41, 65, 75, 87, 91, 95. This range of K-mer was found to give the best assembly result (Bankevich et al. 2012). The quality of contigs was checked with Quast 4.2 (Gurevich et al. 2013). Prodigal (version 2.6.3) was used for genes prediction with the option -m -p meta for predicting metagenomic genes with no gaps (Hyatt et al. 2010). Taxonomic assignment was done using Metaphlan2. The predicted proteins were uploaded to GhostKOALA webservice for KEGG assignment

(Kanehisa, Sato, and Morishima 2016). For functional annotation, blastp was performed against the Swiss-Prot, refseq and NR databases, with a threshold e-value of 1e-6. InterProScan5 was used with the addition of panther database to identify protein domains using HMM model (Quevillon et al. 2005). Blast2GO was used to integrate the blastp and interproscan results for further improving functional annotation (Götz et al. 2008). HMMER version 3.0 was used with CAZy database (version 6) using HMM model to identify carbohydrate-active genes (Lombard et al. 2014). These genes were subjected to filtering using an e-value threshold of 1e-5 for alignments over 80aa, and a threshold of 1e-3 for shorter alignments. The CARD database was used to identify resistance genes (B. Jia et al. 2017). All the amino acid sequences of anti-resistance proteins were merged and subjected to blastp with a threshold e-value of 1e-6. All the sequences with more than 60% identity with their top blast hit were collected. Descriptions of the ARO terms was obtained from the online database (https://card.mcmaster.ca/). The KEGG, Pfam and NR databases were used to confirm the accuracy of the functional annotations obtained with CAZY and CARD. Secondary metabolite biosynthetic gene clusters were identified for contigs larger than 3 000 bp using the antiSMASH2 program (Weber et al. 2015). To identify homologies and orthologies between the genome of *F. candida* and the metagenome, a reciprocal blast was performed. The metagenomic protein sequences were blasted against the host proteins, and vice versa. Sequences that were top hits of each other were extracted using a homemade script, and those matching *F. candida*'s foreign genes were identified (Faddeeva-Vakhrusheva et al. 2017). For a detailed explanation of the methods used to identify the foreign genes within the genome of the springtail, we refer to the publication from Faddeeva-Vakhrusheva *et al*. 2017. Phyre2 was used to predict the structure of the protein of the best reciprocal blast hits (L. A. Kelley et al. 2015).

3.3.5 Data deposition

The raw sequencing data was deposited in NCBI's Sequence Read Archive (SRA) under accesison number SRP149127. The Whole Genome Shotgun (WGS) project was deposited at DDBJ/ENA/GenBank under accession number QIRE00000000. The version described in this paper is version QIRE01000000.

3.4 Results

3.4.1 Sequencing results, assembly and annotation

Table 1 summarizes the sequencing results by indicating, for each sample, the preparation method used and the number of raw and filtered reads obtained. Approximately 90% of the reads passed the trimming step. Most of these reads (more than 97%) originated from the host *Folsomia candida*, and were removed during the next filtering step along with reads from *Wolbachia pipientis*, *Saccharomyces cerevisiae* (used as food source for *F. candida*, and therefore likely to contaminate the genomic libraries) and human DNA. The proportion of reads of prokaryotic origin was slightly higher in dissected gut samples compared to whole springtail samples (compare sample Fc3 to Fc4), and it was much higher in samples treated with the cell-separation method compared to untreated samples (compare sample Fc2 to Fc4, and Fc1 to Fc3). When combining dissection and cell-separation, the proportion of prokaryotic reads increased by a factor 5 (compare sample Fc1 to Fc4). The lowest proportion of *Wolbachia* was observed in the FC3 sample (untreated dissected guts).

Table 1. *Preparation method and number of raw and filtered reads obtained for each sample. For each sample, the number of raw reads and the numbers of reads surviving each processing step is indicated. The percentages in bracket indicate the numbers of reads after each step relative to the number of raw reads.*

Sample ID	**Sample type**	**Sample preparation method**	**Raw reads**	**Reads after trimming**	**Reads after bowtie**	**Filtered reads**
Fc1	Dissected guts (1 000)	Filter and DNA isolation	138,555,106	121,428,759 (87.6%)	3,605,008 (2.6%)	5,806,361 (1.23%)
Fc2	Whole springtails (300) and dissected guts (400)	Filter + Percoll and DNA isolation	133,586,006	116,187,374 (87%)	1,811,553 (1.36%)	
Fc3	Dissected guts (250)	Direct DNA isolation	103,864,717	93,503,412 (90%)	535,052 (0.52%)	
Fc4	Whole springtails (60)	Direct DNA isolation	94,686,416	84,746,773 (89.5%)	372,193 (0.39%)	

A total of 5,806,361 high quality paired reads were used for assembly, which resulted in 107,138 contigs with a total length of 69 Mb (Table 2). Prodigal predicted 147,851 protein-coding sequences (CDSs), 133,594 of which were annotated in Swiss-Prot (Supplementary File 1). 132,657 genes (99%) were of bacterial origin, 665 genes were annotated as Eukaryota, 209 as viruses, 33 as Archaea, 30 as vectors or uncultured microorganisms and 14,257 were unassigned. Supplementary Fig. 1 shows the length distribution of the contigs. The 20 longest contigs (more than 100,000 bp each) were assigned either to *Pseudomonas* or *Microbacterium*.

Table 2. *Results of assembly and annotation. N50 = the size of the contig that, together with the larger contigs, contains 50% of the total metagenome length; N75 = the size of the contig that, together with the larger contigs, contains 75% of the total metagenome length; L50 = number of contigs whose summed length is 50% of the metagenome size; L75 = number of contigs whose summed length is 75% of the metagenome size.*

Number of contigs	107,138
Largest contig (bp)	1,306,495
Total length	69,108,988
N50	2,514
N75	853
L50	1,835
L75	10,181
GC%	60.2%
Gene count	147,851
Genes with function prediction	133,500

3.4.2 Taxonomic classification

The dominant bacterial taxa in the metagenome of *F. candida* were Proteobacteria (50% of the reads), Actinobacteria (32%), Bacteroidetes (12%) and Firmicutes (6%) (Fig. 1). These phyla constituted 99.5% of all the reads. 35 additional phyla were found in the remaining 0,5% of reads. 826 bacterial genera (excluding singletons) were identified. 23 of these genera

covered 83% of the reads. The most abundant genus was *Microbacterium* (Actinobacteria, 13.1% of the reads), followed by *Paraburkholderia* (Betaproteobacteria, 7.2%), *Pseudomonas* (Gammaproteobacteria, 6.3%), *Staphylococcus* (Firmicutes, 5.6%), *Sphingopixis* (Alphaproteobacteria, 5.5%), *Stenotrophomonas* (Gammaproteobacteria, 5.4%), *Pseudoxanthomonas* (Gammaproteobacteria, 5.4%), *Gordonia* (Actinobacteria, 4.1%), *Burkholderia* (Betaproteobacteria, 3.4%) and 14 other genera each with a relative abundance higher than 1%. The overview of the identified taxonomic groups at the phylum, class and genus level is give in Supplementary Fig. 2.

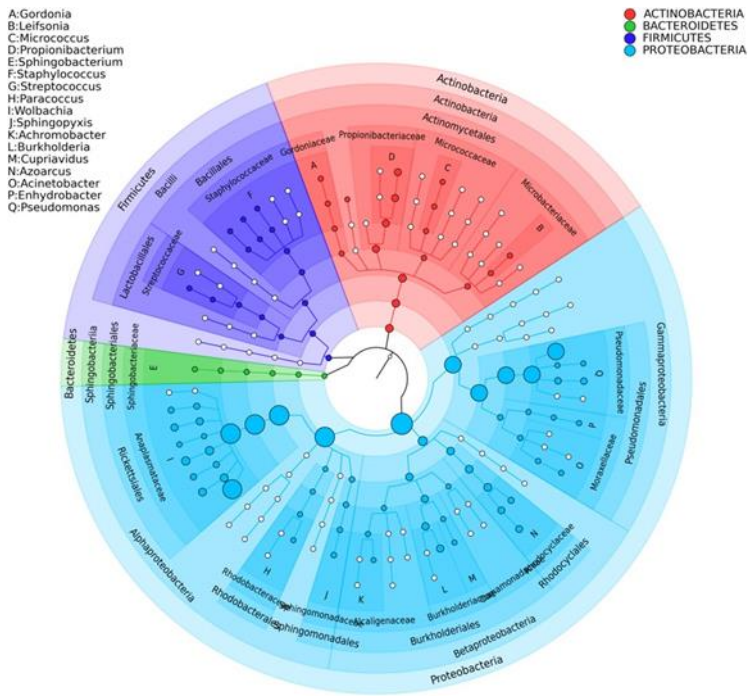


Figure 1: *Phylogenetic distribution of the bacterial community in the metagenome of F. candida. The size of the circles is proportionate to the abundance of the taxa. The phylogeny was built based using Metaphlan on high quality raw reads.*

3.4.3 Overall functional analysis

Comparison of the genes with the KEGG database recovered a number of functions. The most abundant functional categories were associated with membrane transport, signal

transduction, carbohydrate and amino-acid metabolism, and the genetic information processes replication and repair and translation (Fig. 2A).

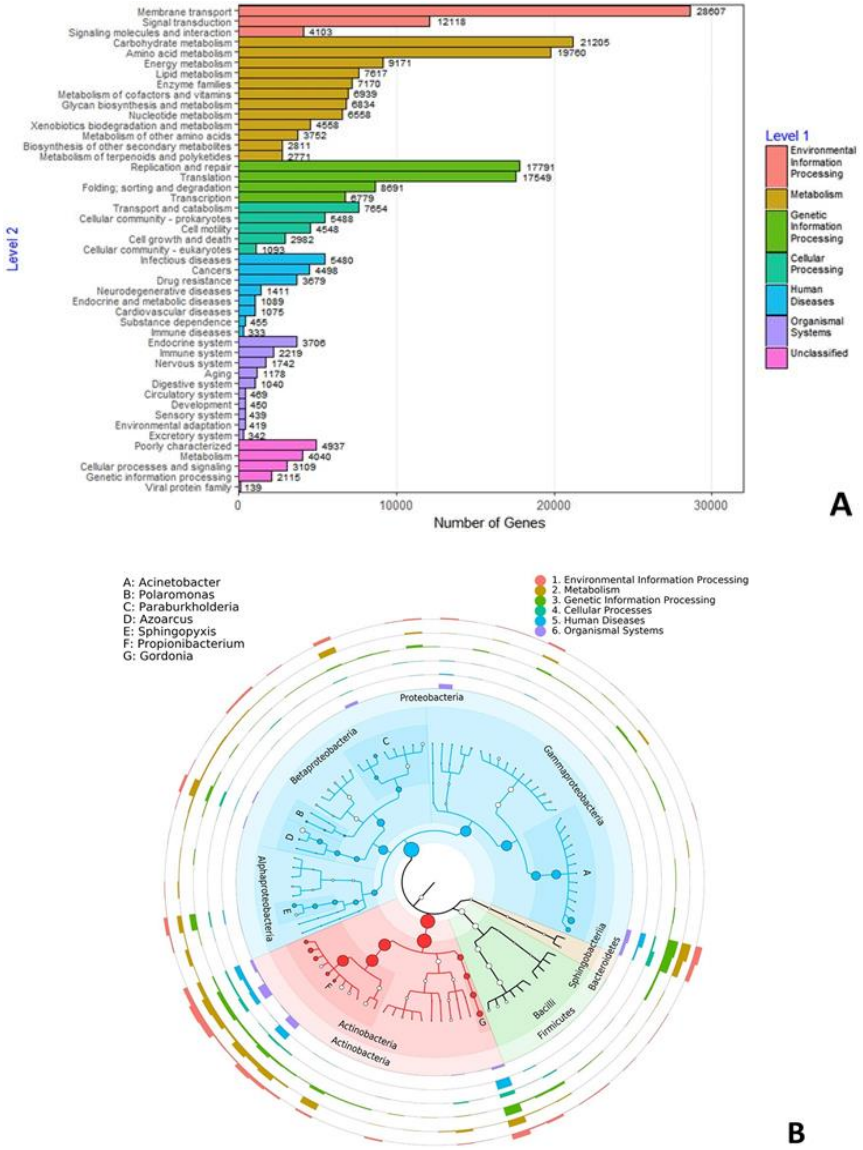


Figure 2: *Functional annotation. (A) Detailed representation of the functional classes belonging to six main functional categories. (B) Functions mapped on the phylogenetic tree. The heights of the bars represent the numbers of kegg terms found for each bacterial species and for each functional category, in proportion to the width of the rings surrounding the taxonomic tree. A bar as high as the ring represents 50 kegg terms.*

Mapping of the functions on the phylogenetic tree shows that most predicted genes within any functional category are assigned to few bacterial species, namely the Proteobacteria *Acinetobacter johnsonii*, *A. lwoffii*, *Pseudomonas stutzeri*, *Paraburkholderia phytofirmans*, *Azoarcus toluclasticus*, *Sphingopixis alaskensis*, the Actinobacteria *Gordonia araii*, *Cutibacterium acnes* and three *Propionibacterium* species, and the Firmicutes *Staphylococcus equorum* (Fig. 2B). The next sections present the functions related to carbohydrate metabolism, secondary metabolite production and antibiotic resistance identified in *F. candida*'s microbiome.

3.4.4 Carbohydrate metabolism

Carbohydrate metabolism was investigated by comparing predicted genes in *F. candida*'s microbiome with the carbohydrate-active enzymes (CAZY) database. 2,004 genes were predicted to code for enzymes involved in carbohydrate metabolism. 1,988 (99.2%) of these genes were of bacterial origin and they mostly originated from Proteobacteria (43%) and Actinobacteria (36%). The complete list of CAZymes is presented in Supplementary File 2, and an overview of the identified pathways involved in starch and sucrose metabolism is given in Supplementary Fig. 3.

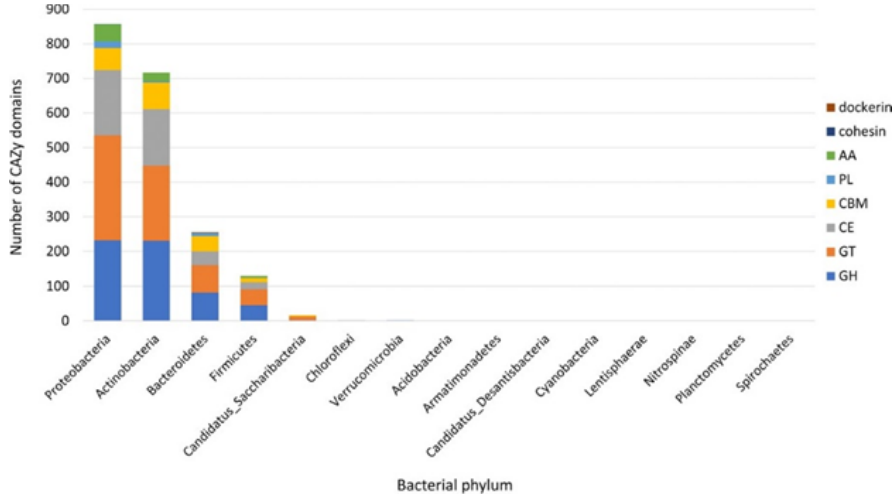


Figure 3: *Column chart indicating the distribution of Carbohydrate Active EnZyme (CAZy) domains among the bacterial phyla retrieved in the metagenome. CBM: carbohydrate-binding module; CE: carbohydrate esterase; GH: glycoside hydrolase; GT: glycosyltransferase; AA: auxiliary activity; PL: polysaccharide lyase.*

The carbohydrate-related genes were assigned to five CAZy classes and three modules (Fig. 3). 664 genes were identified as glycosyltransferases (GT, 33.1% of the total), 598 as glycoside hydrolases (GH, 30%), 420 as carbohydrate esterases (CE, 21%) and 206 as carbohydrate-binding modules (CBM, 10.1%). The GT, GH and CE CAZymes classes were overrepresented in the metagenome compared to the genome of *F. candida* (data not shown). Instead, enzymes with a carbohydrate-binding module (CBM) were more abundant in the genome of the host. 23 of the genes encoding carbohydrate-active enzymes had a best reciprocal blast hit against foreign genes in the genome of *F. candida*.

3.4.5 Secondary metabolites

We screened the gut microbiome for the presence of secondary metabolite biosynthesis pathways related to antimicrobial activity. In total, 166 pathways were identified, 96 of which are putatively involved in the production of an unknown type of secondary metabolite (Supplementary Table 1). 32 pathways are related to saccharide or fatty acid-containing metabolites, and one cluster showed similarity to metabolites with both a saccharide and fatty acid component. Thirteen clusters are represented by non-ribosomal protein synthases (NRPS), which encode multi-domain and multifunctional enzymes involved in the biosynthesis of a large class of biologically active natural products. Another group of ribosomally-synthesized antimicrobial peptides, bacteriocins, are represented by four biosynthetic clusters. We also identified known antibiotics classes among the antismash clusters, namely rifamycin, spectinomycin, chalcomycin, and the antifungal bacillomycin.

3.4.6 Antibiotic resistance

Predicted genes were mapped against the CARD database to determine the occurrence of antibiotic resistance genes (ARGs) in the gut microbiome of *F. candida* (B. Jia et al. 2017). The analysis recovered 811 genes, corresponding to 209 unique terms in the CARD database. Figure 4 provides an overview of the identified antibiotic resistance mechanisms and of the drug classes to which resistance is conferred. The complete list of genes with accession and classification in CARD is provided in Supplementary File 4. Most antibiotic resistance mechanisms retrieved involved antibiotic target alteration (52%), followed by efflux processes (33%) and antibiotic target replacement (8%). The most abundant class of antibiotics associated with resistance was that of fluoroquinolones (16%), followed by

aminocoumarins (10%), peptide antibiotics, lipopeptide antibiotics and tetracyclines (9% each), macrolides, beta-lactams and rifamycin (5% each). Several classes of ARGs involved in resistance to clinically relevant antibiotics, such as β-lactams and tetracycline, were identified (Figs 4 and 5).

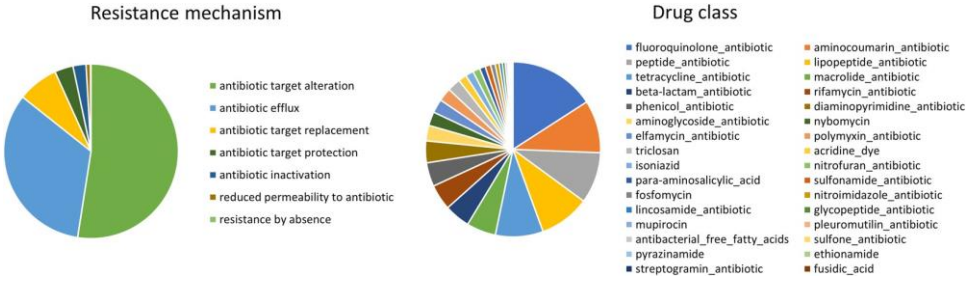


Figure 4. *Overview of the drug mechanisms (left) and classes (right) associated with antibiotic resistance recovered in the metagenome of F. candida. The data was obtained by mapping predicted genes against the CARD database* (B. Jia et al. 2017) *and by extracting the "resistance mechanism" and "drug class" categories from the results.*

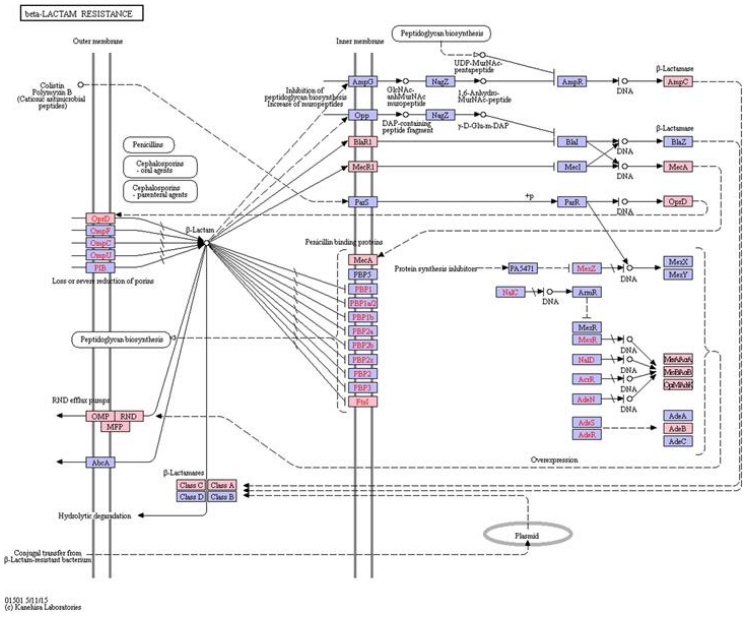


Figure 5: *Genes from the F. candida's metagenome predicted to be involved in β-lactam resistance are represented as colored items in KEGG's β-lactam resistance pathway. The pathway map was obtained from the KEGG database* (Kanehisa et al. 2019).

3.4.7 Host-microbiome interaction and horizontal gene transfer

A reciprocal blast was performed between the proteins in the *F. candida* genome and the predicted protein sequences in the metagenome, to identify orthologies between the springtails' genome and metagenome. The list of best reciprocal blast hits was then compared with the list of 809 horizontally transferred genes in the genome of *F.* candida. We hypothesize that the identified orthologs between gut microbiome and the host genome have undergone HGT from the gut microbiome into the host genome.

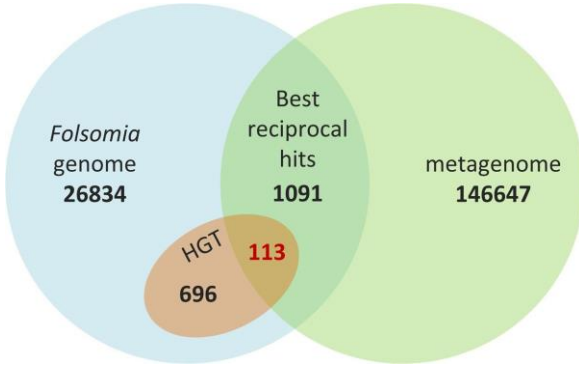


Figure 6. *Venn diagram showing overlap (best-reciprocal blast hits) between proteins from F. candida's genome (light-blue) and proteins from its gut microbiome (green). The red circle contains the horizontally transferred genes, and the number in red indicates the overlap with the gut microbiome.*

Within the gut microbiome, 1,204 predicted protein sequences showed a best reciprocal blast hit with predicted protein sequences in the host genome. Most of these genes are involved in basic metabolic functions that are highly conserved across most life forms, such as transcription, translation, fatty acid metabolism, chaperone activity, amino acid biosynthesis, nucleic acid biosynthesis and ATP biosynthesis. Of these 1,204 genes, 113 had a best reciprocal blast hit against one of the 809 foreign genes in *F. candida* (Fig. 6). The complete list of these 113 genes is given in Supplementary File 5. Taxonomic and functional annotation suggests that *Pseudomonas*, *Microbacterium* and *Gordonia* may be the potential donors of 26, 12 and 9 genes respectively, jointly accounting for almost 50% of them (Supplementary File 5).

Annotation analysis showed that 23 of the 113 genes are CAZymes. Supplementary Fig. 4 shows the predicted protein structures of both the metagenomic read and the animal contig

for the top three reciprocal blast hits, corresponding to a glycosidase, an arabinosidase, and an isocitrate lyase. We also identified a non-ribosomal peptide synthase potentially involved in bacteriocin synthesis, one polyketide synthase and several enzymes associated with detoxification (monooxygenases ABC transporters, gluthatione-S-transferases, and copper oxidase). Most of the 71 remaining annotated genes are related to basic metabolic processes. Because we did not conduct gene expression analysis on the gut microbiome, we are currently unable to verify whether these genes are transcribed and thus functional in the microbial community.

3.5 Discussion

In this study, we applied both dissection and a cell separation method to enrich the bacterial fraction of springtail samples, with the aim of increasing the proportion of bacterial reads after sequencing. The cell separation method was developed by Engel, Martinson, and Moran 2012 and it was more effective than dissection when applied to *F. candida*. Although dissection normally helps to effectively target the microbial component (Gontang et al. 2017), this may be more complicated in microarthropods such as springtails because of their small size. A combination of dissection and cell separation method proved to be most effective in increasing the proportion of prokaryotic reads. Still, more than 97% of the reads in any sample belonged to the host *Folsomia candida*: recovery of genetic material from symbiotic microorganisms can be problematic in microhabitats such as insect guts, due to the much higher abundance of host DNA (Paula et al. 2016).

Wolbachia can dominate the bacterial population in *F. candida* (Valeria Agamennone et al. 2015). By discarding organs containing high amounts of *Wolbachia* (brain and ovaries), dissection should be effective in reducing the occurrence of the endosymbiont in the samples. Indeed, sample FC3 (consisting of guts obtained through dissection) had the lowest proportion of *Wolbachia* reads. Cell separation is also expected to reduce the amount of *Wolbachia* DNA in the samples. Because of its intracellular location (gut epithelium, ovaries and brain), a method that separates the eukaryotic cells from the prokaryotic ones without lysing them should be effective in reducing the amount of host and *Wolbachia* DNA in the same step. However, in this study, a combination of dissection and filtering resulted in an increased amount of *Wolbachia* reads (9.26% in sample FC1 vs 3.03% in sample FC2). Because of the difference in size between prokaryotic genomes and the host genome

(resulting in sequencing biases), and because of possible lysis of host cells during the treatment of samples FC1 and FC2, resulting in the release of *Wolbachia* cells, it is difficult to conclude whether filtering was an effective strategy to reduce the representation of the endosymbiont in the metagenomic dataset.

The number of contigs and the total length after assembly are comparable to other soil invertebrate-associated metagenomes (Suen et al. 2010; Cheng et al. 2013; He et al. 2013). Although this was not attempted here, it may be possible to recover the genome of one or more species using the data collected in this study (Sangwan, Xia, and Gilbert 2016).

With 826 bacterial genera identified, the level of diversity in *F. candida* approaches that described in the hindgut of termites, wood-feeding insects that have one of the most complex microbiota of any animal group (Bourguignon et al. 2018). Other soil invertebrates are characterized by comparable or even higher levels of microbial diversity. For example, Pass et al. 2015, studied the microbiome of the earthworm *Lumbricus rubellus* and found no less than 9,120 host-specific OTUs. This very diverse community was dominated by Proteobacteria and Actinobacteria, very similar to the situation in *F. candida*. High diversity was also observed in the gut of two cockroach species, with approximately 1,000 OTUs (Berlanga et al. 2016), whereas slightly lower counts were detected in the ant *Cephalotes varians* (445 OTUs), in the compost worm *Eisenia fetida* (338 OTUs) and in the isopod *Armadillidium vulgare* (153 OTUs) (Kautz et al. 2013; D. Liu et al. 2018; Dittmer et al. 2016).

The bacterial community in *F. candida* was dominated by Proteobacteria species, and within this group the Gammaproteobacteria were particularly abundant (21% of the reads). Proteobacteria, a large taxon of functionally diverse bacteria, dominate the microbiome of terrestrial insects and other soil invertebrates such as earthworms, nematodes and isopods (Pass et al. 2015; Yun et al. 2014; Esposti and Romero 2017; M. Berg et al. 2016; Bouchon, Zimmer, and Dittmer 2016). *Pseudomonas*, one of the most abundant bacteria detected in *F. candida*, is commonly found in the microbiome of soil invertebrates like termites, ants and beetles, isopods and nematodes, as well as in their environment (D. Liu et al. 2018; Dittmer et al. 2016; Esposti and Romero 2017; Aylward et al. 2014). *Pseudomonas*, together with *Rickettsia* and *Chryseobacterium*, was also the most abundant OTU in the microbiome of the springtail *Orchesella cincta* (Bahrndorff et al. 2018). Another abundant bacterium in *F.*

candida was *Paraburkholderia*. This genus includes many soil species, a few of which are used as plant probiotics thanks to their growth-promoting and possibly defensive properties (X. Chen et al. 2018). Other members of the Proteobacteria identified in *F. candida*'s microbiota were *Sphingopixis*, *Stenotrophomonas*, *Pseudoxanthomonas*, *Burkholderia*, all of which were detected in soil invertebrates (worms, cockroaches, termites, ants and beetles) (Esposti and Romero 2017). The most abundant bacterium in *F. candida* was *Microbacterium*. Members of the Microbacteriaceae have been previously identified in different species of beetles (S. T. Kelley and Dobler 2011; Scully et al. 2013), and Actinobacteria in general (although in low amounts) have been found in cockroaches (Gontang et al. 2017) and in a few species of insects (ants, beetles and termites) characterized by nutritional symbioses with fungi (Kautz et al. 2013; Aylward et al. 2014). Actinobacteria are also one of the dominant bacterial groups in other soil invertebrates such as earthworms (Pass et al. 2015; D. Liu et al. 2018; L. Ma et al. 2017).

The observed bacterial diversity in *F. candida* is comparable to that previously detected by 16S high-throughput sequencing in the same lab-reared population of springtails (Valeria Agamennone et al. 2015). However, the taxonomic distribution between the two studies is very different. Based on 16S sequencing, *Pseudomonas* was the most abundant bacterial genus with 42% of the reads (Valeria Agamennone et al. 2015). Nine other dominant OTUs were identified, including *Bacillus* (19% of the reads), a member of the Actinomycetales (9%), *Escherichia sp*. (4%) and *Ochrobactrum sp*. (3%). *Microbacterium* accounted only for 0.3% of the read, and *Paraburkholderia* was not identified. This discrepancy can be explained by the difference in sequencing methods applied. High-throughput amplicon sequencing is subjected to PCR bias, with differences in the amplification efficiency of DNA from different bacterial species; in shotgun metagenomic sequencing, on the other hand, biases can be caused by the method chosen for taxonomic assignment, possibly leading to misidentifications (Tessler et al. 2017).

The majority of reads in *F. candida*'s metagenome originated from pathways involved in membrane transport, carbohydrate and amino acid metabolism, replication, translation and repair. The abundance of genes involved in carbohydrate and amino acid metabolism may suggest a nutritional role of the microbiota. The springtails used in this study were reared exclusively on baker's yeast (*Saccharomyces cerevisiae*), and specific microbial enzymes

could aid in the breakdown of components of the fungal cell wall, including various polysaccharides and glycoproteins (Manjula and Podile 2005). Natural populations of springtails may also benefit from the presence of such functions in their microbiome. In fact, carbohydrate-related functions are often enriched in the gut microbiome of different soil invertebrates, such as beetles, nematodes and isopods (Cheng et al. 2013; Bouchon, Zimmer, and Dittmer 2016; Scully et al. 2013; C. C. Smith et al. 2017; Brune and Dietrich 2015), some of which rely on symbiotic microbes for the breakdown of long polymers such as lignin, cellulose and other plant-derived products (Cheng et al. 2013; Brune and Dietrich 2015). *F. candida* is an euedaphic springtail species whose natural diet includes not only yeasts and other fungi, commonly occurring in the soil environment, but also decaying plant material. Recently, the microbiota of another springtail species, the epiedaphic *Orchesella cincta*, was studied, and some of the main functions predicted based on the microbial community structure were related to the breakdown of dietary components and of plant secondary metabolites (Bahrndorff et al. 2018). In a previous study we observed substantial overlap in the composition of the bacterial communities between a lab-reared and a field population of springtails (Valeria Agamennone et al. 2015). This suggests that similar carbohydrate-degrading functions may be present in both lab-reared and field populations of springtails.

Amino acid-related functions may also be beneficial for the host. Some intracellular endosymbionts biosynthesize essential amino acids that are lacking in the diet of their host (Douglas 2016) and gut bacteria may exert similar functions (Leitão-Gonçalves et al. 2017). A contribution to the host's nutrition may also explain the abundance of functions related to membrane transport in *F. candida*. Transport allows host-symbiont exchanges and therefore it constitutes one of the most important functions in the maintenance of the symbiosis with bacteria providing nutrients (Charles et al. 2011).

In accordance with the taxonomic assignment, most genes in the above discussed categories were predicted to belong to Proteobacteria and Actinobacteria species. Many genes were annotated to *Acinetobacter johnsonii*, a member of the Gammaproteobacteria that has been described as an opportunistic pathogens for animals as well as a possible reservoir of antibiotic resistance genes (Montaña et al. 2016; Tian et al. 2016). *Acinetobacter* was also a dominant genus in the microbiome of the earthworm *Eisenia fetida* (Dittmer et al. 2016) and it was identified in other soil invertebrates such as the *Longitarsus* beetle and the isopod *Armadillidium vulgare* (Dittmer et al. 2016; S. T. Kelley and Dobler 2011). Many functions

were also assigned to the genus *Propionibacterium*. This group of Actinobacteria includes species with good probiotic potential due to their capacity to modulate microbiota, gut metabolic activity and the immune system (Cousin et al. 2011). Interestingly, the immunomodulatory and anti-inflammatory properties of *Propionibacterium* have been observed not only in human and mouse models (Cousin et al. 2011), but also in soil invertebrates (Kwon, Lee, and Lim 2016). An abundance of genes was taxonomically assigned to a few other groups, among which *Gordonia*, a genus of Actinomycetes including many symbionts of terrestrial invertebrates (Sowani, Kulkarni, and Zinjarde 2018) and *Pseudomonas*, commonly found in soils and in soil invertebrates (Esposti and Romero 2017).

Carbohydrate-degrading enzymes are commonly found in the bovine rumen (Jose et al. 2017), in the gut of wood-feeding insects such as termites and woodwasps (Adams et al. 2011; Warnecke et al. 2007) and in the microbial community of fungus gardens associated with leaf-cutter ants (Aylward et al. 2012). These enzymes are often of microbial origin, suggesting that herbivorous animals can exploit the catalytic activities of microbial symbionts to access nutrients stored in plant biomass (Suen et al. 2010). In termites, the symbiotic relationship with a complex community of bacteria, archaea and protists in the gut enables the digestion of lignocellulose, conferring these insects a unique ecological position in tropical and subtropical ecosystems (Brune and Dietrich 2015). Whether similar relationships between Collembola and their microbiome exist is unknown at the moment, but microbial functions related to carbohydrate metabolism are likely to significantly contribute to the ecological role of springtails as members of the soil decomposer community.

Warnecke *et al* 2017 found 700 glycoside hydrolase (GH) catalytic domains corresponding to 45 CAZY families in the microbiome of wood-feeding termites (Warnecke et al. 2007). In the microbiome of *F. candida*, we identified a comparable number of genes encoding for enzymes with a capacity to break down long chain carbohydrates such as starch, lignin and cellulose. In nature, these enzymes may aid *F. candida* in extracting nutrients from the plant biomass that constitutes part of its diet, as was suggested for the springtail *O. cincta* (Bahrndorff et al. 2018).

A large number of glycoside hydrolases was also observed among *F. candida* foreign genes (Faddeeva-Vakhrusheva et al. 2017). Interestingly, some of the foreign genes that were also best reciprocal hits between the genome and the metagenome of *F. candida* were identified

as CAZymes (Supplementary File 2). HGT of cellulose-degrading enzymes has been previously observed in plant-feeding insects (Pauchet and Heckel 2013) and may be an important mechanism providing soil invertebrates with advantageous traits for living in the soil (Eyun et al. 2014).

The microbiome of *F. candida* contained several pathways responsible for the biosynthesis of secondary metabolites. This is a class of compounds that are often involved in competition and interaction between species, and they may contribute to the establishment and the maintenance of a stable gut microbiota through the exclusion of transient or pathogenic microbes (Richardson et al. 2015; Jousset, Scheu, and Bonkowski 2008). Secondary metabolites often find applications in the biotechnological and medical sector. The main contributors to the identified pathways seem to be *Gordonia*, *Pseudomonas fluorescens*, *Bacillus* and *Streptomyces*.

A few of the identified pathways were represented by NRPSs, a class of enzymes responsible for the biosynthesis of natural products with a broad range of biological activities and pharmaceutical properties. Cluster 10 and 28 show resemblance with an NRPS producing pyoverdines, siderophores well known for their high affinity for Fe^{3+} under low iron availability (Schalk and Guillon 2013). Another NRPS involved in the biosynthesis of the siderophore nocobactin was identified in cluster 95. Three clusters show homology to NRPSs involved in antibacterial and antifungal activity. Cluster 31 shows substantial similarity (47%) to an NRPS producing orfamide, a compound of bacterial origin with antifungal properties and with good potential as biocontrol agent against fungal pathogens (Z. Ma et al. 2016). Cluster 130 represents an NRPS involved in microsclerodermin biosynthesis, an antifungal compound produced by a marine sponge (Xiaohui Zhang et al. 2012). A recent study also showed that this compound has properties of pharmaceutical relevance, as it can inhibit NFkappaB transcription in a human pancreatic cell line leading to apoptosis (Guzmán et al. 2015). Finally, the NRPS identified in cluster 48 showed similarity to the NRPS involved in biosynthesis of the antibiotic caryoynencin, a compound originally isolated from a plant pathogen. Very recently it has been shown that this compound is produced by a symbiont of a herbivorous beetle, protecting its eggs against detrimental microbes (Flórez et al. 2017).

We also identified a number of bacteriocins, a class of compounds with potential as natural food preservative (Gálvez et al. 2007). Many bacteriocins are biosynthesized by lactic acid bacteria, and in *Folsomia*'s gut microbiome these clusters are homologous to *Pseudomonas fluorescens* and *Gordonia effusa*.

Several other interesting biosynthesis clusters with functions related to medical applications were found, such as lymphostin, a known immunosuppressant isolated from *Streptomyces* (Aotani, Nagata, and Yoshida 1997), and chartreusin, that exerts strong chemotherapeutic activity against various tumor cell lines (Z. Xu et al. 2005). We also identified a mangotoxin biosynthesis cluster. Mangotoxin causes apical necrosis of plant tissue, which may aid in food processing and digestion by the host (Arrebola et al. 2003). Biosynthesis of the volatile compound homoserine lactone (hserlactone) may be related to communication between fungi and bacteria (Shiner, Rumbaugh, and Williams 2005), while ectoine may serve as osmolyte conferring resistance to salt, dessication and temperature stress (Mosier et al. 2013).

The distribution of antibiotic resistance genes (ARGs) in microbiomes sampled across environments and organisms is still not well understood. A large-scale metagenomics study indicated that soils harbor most classes of ARGs (Nesme et al. 2014). In the gut microbiome of *F. candida*, we identified over 200 unique terms associated with antibiotic resistance distributed over more than 800 genes, more than twice the number detected in human microbiomes and almost eight times the number detected in the giant African snail *Achatina* (Fitzpatrick and Walsh 2016). This might be explained by the intimate association between the springtail and the soil ecosystem.

The presence of ARGs in the gut of *Folsomia* may have ecological relevance. It is noteworthy that we identified a substantial number of β-lactamases, probably resulting from the selective pressure caused by β -lactam production by the host itself (Suring et al. 2017). For example, *Bacillus toyonensis*, a member of *F. candida*'s microbiota, is highly resistant to β-lactams (Janssens et al. 2017). Furthermore, interactions between bacterial communities with antibiotic biosynthesis capacity and communities showing resistance to such antibiotics can also be expected. Observations from this and other studies indicate a potential for *Pseudomonas*, *Streptomyces* and *Gordonia* strains isolated from *F. candida* to synthesize antibiotics (see section above, Supplementary File 3 and (V. Agamennone et al. 2018)), while *Streptomyces*, *Enterococcus* and *Staphylococcus* are abundant among ARG-containing

bacterial strains in *Folsomia*'s gut (Supplementary File 4). This supports the notion that antibiotics regulate the homeostasis of microbial communities, and may even be beneficial for commensal bacteria in environments such as the animal gut (Linares et al. 2006). Finally, Engel & Moran (Engel and Moran 2013) suggested that this balance may be important in facilitating colonization resistance against parasites and bacteria pathogenic to the host. The data provided in this study will be highly relevant in formulating concrete hypotheses to investigate the ecological connectivity of antibiotic-biosynthetic and ARG-containing bacteria in gut microbiomes.

A previous study had identified 809 foreign genes in *F. candida*'s genome, which were validated by physical linkage with native genes (through PacBio long read single molecule sequencing), blast analysis and phylogenetic inference (Faddeeva-Vakhrusheva et al. 2017). Moreover, Faddeeva *et al*. used RNA sequencing to show that almost 60% of this gene set was actively transcribed, indicating functional relevance (Faddeeva-Vakhrusheva et al. 2017). Here, we applied best reciprocal blast analysis to identify microbial protein sequences orthologous to predicted protein-coding sequences in the genome of *F. candida*. We hypothesize that this would provide circumstantial evidence of horizontal gene transfer from members of the gut microbiome into the host genome. Indeed, within the gut microbiome we identified 113 best reciprocal blast hits with predicted protein sequences of foreign genes of the springtail, possibly indicating HGT from the gut microbiome to the host genome. The foreign genes without a best reciprocal blast hit within the gut microbiome may have been transferred from other microbial sources, for example the over 30% of foreign genes that conferred top blast hits with fungal donors (Faddeeva-Vakhrusheva et al. 2017). Alternatively, other genes may have been transferred to the host genome early in the evolution of *F. candida*. In that case, the accumulation of mutations over time would lead to low similarity with members of the microbiome, preventing the identification of the possible source of these genes through best reciprocal blast searches. A number of foreign genes with best reciprocal blast hit with genes in the microbiome were CAZymes, involved in the degradation of polymers such as cell wall components. Gene transfer of carbohydrate-active enzymes may optimize the capacity of *F. candida* to extract nutrients from their diet (Faddeeva-Vakhrusheva et al. 2016), thereby contributing to their adaptation to life in the soil.

Horizontal gene transfer from prokaryotes to eukaryotic host genomes has become a highly controversial topic. There are claims that gene transfer only occurs between hosts and mitochondria, plastids and endosymbionts, and that other HGT cases are the result of differential loss of ancestral genes, that originated prior to the last eukaryotic common ancestor (Martin 2017). However, this hypothesis overestimates gene contents of ancestral genomes, and is therefore unlikely (Leger et al. 2018). We suggest that the foreign genes in *Folsomia*'s genome are most likely acquired via horizontal gene transfer (Faddeeva-Vakhrusheva et al. 2017). Here, we propose that part of these HGT events could have taken place by interaction with the gut microbiota. In the gut environment host and microorganisms maintain an intimate physical association with many opportunities for interaction, thus increasing chances for gene transfer to occur (J. Huang 2013). Two recent studies provide evidence for bacterial DNA transfer into somatic human cells (Riley et al. 2013; Schröder et al. 2011) through bacterial type IV secretion system (T4SS). This system is known to mediate interbacterial conjugative DNA transfer and transkingdom protein transfer into eukaryotic host cells during bacterial pathogenesis. Schroder *et al*. showed that T4SS-dependent DNA transfer into host cells may occur naturally during human infection with *Bartonella* (Schröder et al. 2011). Furthermore, Ridley *et al*. identified a *Pseudomonas* strain as a donor of foreign DNA detected in human stomach carcinomas (Riley et al. 2013). It is still unclear why functions that can be provided by the microbiome would be incorporated and maintained in *F. candida*'s genome. In the case of foreign genes involved in lignocellulose breakdown, we speculate that such functions, when controlled by the host, could provide fitness advantage in terms of energy balance and nutrient acquisition. Similarly, transferred genes involved in detoxification may protect the host for natural toxins that are quite common in the soil. These and other hypotheses should be tested by conducting gene knockdown and other experiments.

We have provided an insight in the metagenome of a collembolan species, *F. candida*. Most bacterial diversity is attributed to four phyla, that are also representative for soil microbial ecosystems, possibly confirming the interaction of *F. candida* with its natural environment. A broad spectrum of gene functions was identified, most notably related to carbohydrate metabolism, antibiotic resistance and secondary metabolite production. These functions were presented and discussed in the context of their ecological relevance and in the light of potential biotechnological applications. Finally, we presented data suggesting that the gut microbiome may have been a source of genes acquired by the host through HGT. These genes

may have conferred a fitness advantage to the springtail, during adaptive evolution in the soil ecosystem.

Supplementary Materials

Supplementary Figure 1. Contig length distribution.

Supplementary Figure 2. Identified taxonomic groups at the phylum, class and genus level.

Supplementary Figure 3. Diagram of the pathways involved in starch and sucrose metabolism. Pink boxes indicate the genes identified in the microbiome.

Supplementary Figure 4. Predicted protein structures of the top three reciprocal blast hits between the metagenome and the genome of *F. candida*, corresponding to a glycosidase (A), an arabinosidase (B), an isocitrate lyase (C). The predicted structures of the microbial genes are on the left, the predicted proteins of the springtail are on the right.

Supplementary Table 1. Summary of antiSMASH results.

Supplementary File 1. List of all the predicted genes with the corresponding taxonomies (based on MetaPhlan) and functional annotations (based on NCBI protein database)*

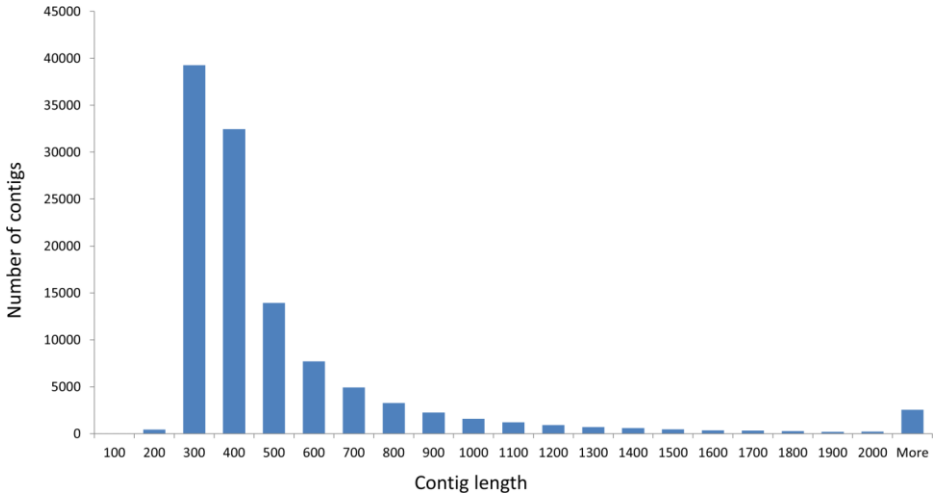
Supplementary File 2. Complete list of the 2004 genes predicted to code for enzymes involved in carbohydrate metabolism*

Supplementary File 3. Complete list of antiSMASH results. For each of the 166 contigs with a hit in the antiSMASH database*

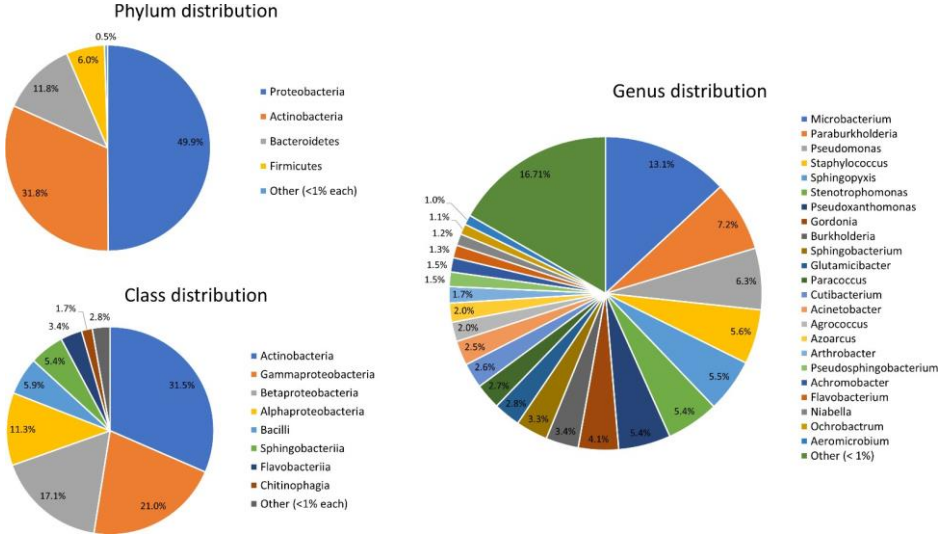
Supplementary File 4. Complete list of the predicted genes with a hit to antibiotic resistance in the Comprehensive Antibiotic Resistance Database (CARD) (B. Jia et al. 2017)(B. Jia et al. 2017)*

Supplementary File 5. List of all best reciprocal hits between *Folsomia candida*'s genome and metagenome that are also predicted foreign genes (HGT), including their taxonomic and functional annotation*

*available online at the VU University Library: www.ub.vu.nl

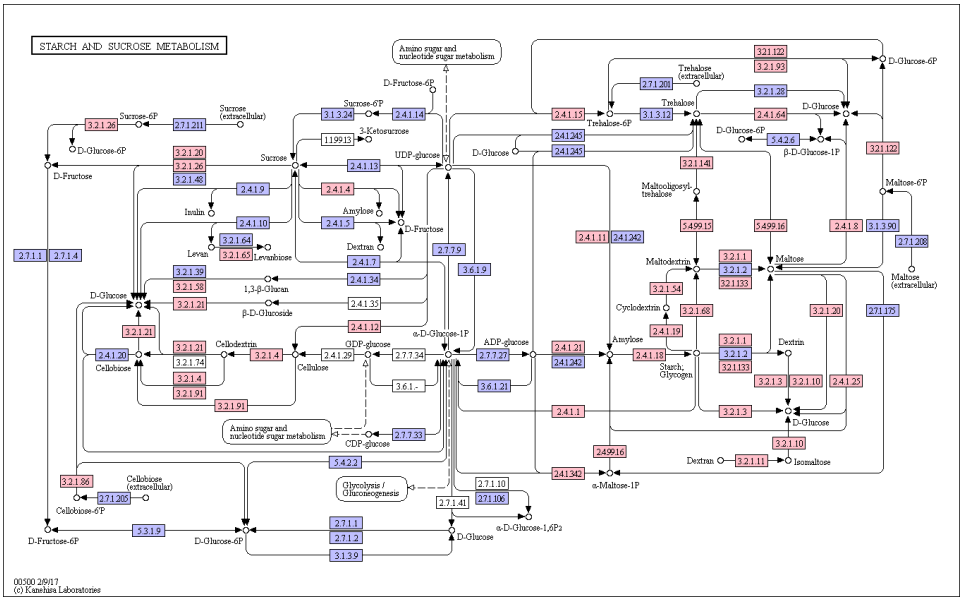


Supplementary Figure 1. *Contig length distribution.*

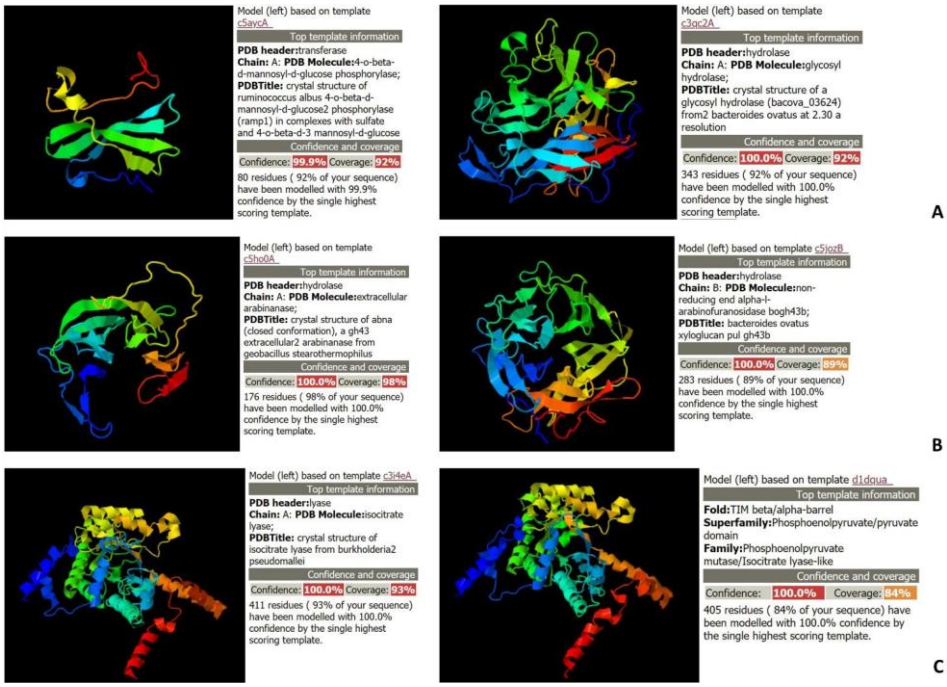


Supplementary Figure 2. *Identified taxonomic groups at the phylum, class and genus level.*

.



Supplementary Figure 3. *Diagram of the pathways involved in starch and sucrose metabolism. Pink boxes indicate the genes identified in the microbiome*



Supplementary Figure 4. *Predicted protein structures of the top three reciprocal blast hits between the metagenome and the genome of F. candida, corresponding to a glycosidase (A), an arabinosidase (B), an isocitrate lyase (C). The predicted structures of the microbial genes are on the left, the predicted proteins of the springtail are on the right.*

Supplementary Table 1. *Summary of antiSMASH results*

Cluster type	**Number of contigs**
Arylpolyene	1
Bacteriocin	4
Cf_fatty_acid	14
Cf_fatty_acid -Cf_saccharide	1
Cf_putative	96
Cf_saccharide	18
Cf_saccharide-Cf_fatty_acid	1
Ectoine	1
Hserlactone	3
Nrps	13
Nrps-Arylpolyene	1
Other	2
Siderophore	2
T1pks	1
T1pks-Nrps	1
T3pks	1
T3pks-Cf_saccharide	1
Terpene	5
TOTAL	**166**

For each type of secondary metabolite cluster, the number of contigs in F. candida's metagenome in which the cluster was detected is indicated. Cf indicates a putative cluster identified with the ClusterFinder algorithm. Pks = polyketide synthase. Nrps = non-ribosomal peptide synthetase. The complete output of the antiSMASH analysis is given in Supplementary File 3.

Chapter 4 - Genetic diversity of carbohydrate degradation, secondary metabolite production and antimicrobial resistance in the microbial metagenomes of three decomposer invertebrate animals

Ngoc Giang Le, Marius Bredon, Didier Bouchon, Bouziane Moumen, Abraham Brouwer, Nico M. van Straalen, Dick Roelofs

4.1 Abstract

Microorganisms associated with the guts of invertebrates represent a specialized community with a diversity of functions in the degradation of organic material or microbial interactions. These communities remain relatively little explored especially with reference to their role in the ecology of the host. In addition, the gut microbiome is a rich source of discovery of novel catalytic functions of possible relevance to biotechnology. In this paper, we report on a comparison of three species of invertebrates, a termite, a terrestrial isopod and a springtail, that represent three different functions in the decomposing community of a soil ecosystem: wood degradation, plant leaf degradation and fungivory. We analyze previously published metagenomes with respect to carbohydrate-associated enzymes (CAZy), microbial resistance (CARD) and the production of antimicrobial metabolites (Antismash). The three hosts differed significantly in the microbial community composition. Of all metagenomic contigs, 70% to 80% could be allocated to a bacterial phylum, *Proteobacteria* being the most dominant, followed by *Firmicutes*, *Bacteriodetes*, *Actinobacteria*, and *Spirochaeta*. We identified 162 CAZy families. The distribution of the main categories was similar among hosts, but each host had 10 -30 specific CAZy families and the most diverse were found in the termite. There was very little overlap between the hosts at the gene level. All three metagenomes had genes encoding functions in antimicrobial resistance. The isopod metagenome had most of them, especially with regard to antibiotic efflux transporters. The springtail was the least diverse in terms of antibiotic resistance genes in their microbiome. Regarding the production of secondary metabolites, a high diversity of non-ribosomal peptide synthetases was found in springtail metagenome and many bacteriocins in termite metagenome. The isopod metagenome had fewer genes encoding the production of secondary compounds. Comparing the three hosts, we conclude that each species has a microbiome that overlaps only with the other microbiomes on a high taxonomic level. When analyzed on a more detailed level, it turns out that each species is unique and has many functional genes not found in another species. This is all the more surprising as the soil invertebrate community is often lumped together as a single unit in soil ecosystem studies. We show that soil decomposer animals include a microbiome with unprecedented diversity and many unique functions.

4.2 Introduction

Many animals depend crucially on the microbiome they carry within and on their bodies. The symbiotic microorganisms are connected to their host through a variety of pathways, including the digestive, immune, circulatory, and neuro-endocrinological systems. Through these interactions, the microbiome influences the behaviour of animals and their ecological function. A better knowledge of the relationship between animals and their microbiome has been considered "an imperative for the life sciences" (McFall-Ngai et al. 2013).

This argument holds especially for invertebrates that live in the soil environment. Soil represents an enormous reserve of microbial communities of which the vast richness has penetrated since microbiologists began to sequence the environmental DNA (Fierer 2017). It is expected that the soil microbial communities also harbour many unknown functions that once revealed, could be used in biotechnology, such as new pathways of carbohydrate degradation, unknown antimicrobial agents and new catalytic functions for the synthesis of bio-based chemicals (Handelsman 2004; Riesenfeld, Goodman, and Handelsman 2004).

A special position is held by many species of soil invertebrates, earthworms, mites, springtails, termites, isopods and the like, and the microbiota associated with them. In estimates for global biodiversity it is assumed that every single invertebrate may contain several species of microbial symbionts that are not yet known. This multiplies present estimates of global biodiversity to 2 billion species, of which threequarters are bacteria (Larsen et al. 2017). Conversely, it may be expected that many invertebrates depend on their microbial communities with regard to food digestion, defense against pathogens and metabolic functions. This interdependence of microbes, invertebrates and ecosystem function is only beginning to be explored.

The metagenomics approach has been very helpful in accessing the unexplored richness of microbial communities associated with invertebrates. Metagenomics is the large-scale sequencing of microbial DNA of a community as a whole. Not only the species composition is the main interest of metagenomics, but also the collective set of functional genes active in a community. Using next-generation sequencing, a more or less complete overview of the functional potentials of an animal-associated microbiome can be obtained. This approach has been applied to model species such as aphids and honeybees (Engel, Martinson, and Moran

2012; Engel and Moran 2013). In this paper we apply a similar approach to three selected species of soil invertebrate.

Termites (infraorder Isoptera of the subphylum Hexapoda) are an order of insects well-known for their complicated social structure and caste system. The lower termites are known to harbour a complex community of bacteria, archaea and protists that allow them to digest food items such as lignocellulose that cannot be digested by almost all animals (Brune and Dietrich 2015). Terrestrial isopods or woodlice (order Isopoda of the arthropod subphylum Crustacea) are known leaf eaters, which through their activity contribute to the degradation of organic matter in soil. Their gut microbiome contains an unexpected richness of symbionts and parasitic microbes (Bouchon, Zimmer, and Dittmer 2016). Springtails (class Collembola of subphylum Hexapoda) are an abundant group of microarthropods present in any soil, mostly consisting of fungal grazers which are known for their remarkable resistance to pathogenic fungi. Their gut microbiome has been explored recently (Valeria Agamennone et al. 2019). Together these three groups capture a wide range of food habits and ecological functions and a comparison of their microbiomes may shed light on the relationship between gut microbial communities and ecological function.

In this paper we focus on three functional categories of genes in the metagenomes, which we believe are of crucial importance in the ecological function of soil invertebrates: (1) degradation of carbohydrates, more specifically the genes classified as carbohydrate-associated enzymes (CAZymes), (2) genes associated with microbial resistance catalogued in the Resistance Gene Identifier database RGI, and (3) genes associated with the biosynthesis of secondary metabolites, as revealed by comparison to the antiSMASH database. We compare the metagenomes of the three invertebrates with regard to these three functional gene categories in order to shed light on the relationship between microbial metagenomes and the ecological function of their hosts.

4.3 Material and Methods

We compared the microbiomes of three different soil invertebrates. The termite *Coptotermes gestroi* (Isoptera, Rhinotermitidae), also called Asian subterranean termite, is a common termite originally occurring in South-East Asia, but now spread across the world and considered a pest in many places. The woodlouse *Armadillidium vulgare* (Isopoda, Armadillidiidae) is a widely distributed species associated with dead leaves and wood in the

temperate regions, and also in anthropogenic habitats. The springtail *Folsomia candida* (Collembola: Isotomidae) is a common species of microarthropod associated with rich soils and compost heaps across the world. All animals were collected from their natural habitat and cultured in the laboratory in an attempt to remove the direct influence of microbial communities at their place of collection. For details of library preparation, sequencing and metagenomic assembly we refer to Do *et al.* (2014) for *C. gestroi*, to Bredon et al. (2018) for *A. vulgare*, and to Agamennone *et al.* (2015,2019) for *F. candida*.

The four samples of F. candida metagenomes based on different DNA extraction method were obtained from (Valeria Agamennone et al. 2015). They were processed individually and also pooled together using the same metagenomic assembly approach (Valeria Agamennone et al. 2019). The three species of *A. vulgare* metagenomes from the laboratory strain were combined together and subjected to CD-HIT (version 4.8.1) processing with the setting –c 1 –n 10 to remove contigs with 100% identity (Fu et al. 2012). Qualities of all three metagenome assemblies were checked using QUAST (version 4.6.1) (Gurevich et al. 2013). The contigs were analysed for genes associated with the production of secondary compounds by means of the antiSMASH server version 5.1.2 with full settings (Blin et al. 2019). Genes were predicted using Prodigal with the –m for metagenomics setting (Hyatt et al. 2010). Kraken2 was used to profile the metagenomes and so identify the bacterial composition (Wood, Lu, and Langmead 2019). Predicted genes with complete open reading frames and a stop codon were used for further analysis. For identifying genes associated with carbohydrate activity enzymes from all three metagenomes, we used the dbCAN2 with the CAZy database version 8 (Zhang et al. 2018). This program combines results from hidden Markov models, sequence aligner DIAMOND and short sequence predictor HOTPEP to cluster CAZy families based on family-specific domains. To identify and analyse genes related to antimicrobial resistance, the Resistant Gene Identifier (RGI) program (version 5.1.0) was used on the Comprehensible Antibiotic Resistance Database CARD version 3.0.5 (Alcock et al. 2020). To cover all possible antibiotic resistance genes (ARGs) the default settings from RGI was used. Hits with low coverage but high identity percentage were included for the analysis. The proteins identified were subjected to BLASTp (version 2.10.0) searches against the non-redundant (NR) database with the default setting for e-value to confirm their annotation (Altschul et al. 1990). The contigs, which contain proteins associated with CAZymes, antimicrobial resistance and secondary metabolites were mapped back to the

taxonomic result from Kraken2. Closely related taxonomic at the genus level were agglomerated. The taxonomy chart was drawn using GraPhlAn version 1.1.3 (Asnicar et al. 2015) using an custom script. For the stacked bar plot, taxonomic with the abundance below 1% were merged into a group called Others. Figures were generated using R and the ggplot2 package (Wickham 2009; R Core Team 2018). The CAZyme proteins from all three metagenomes were used to construct a protein BLASTp database and aligned against each other with the default e-value. Proteins with over 90% identity were kept for similarity analysis.

4.4 Results

The metagenome of *A.vulgare* lab strains consists of 123,613 contigs, of which 5,038 were 100% identical. Isopod, springtail and termite metagenomes contained 118,575, 106,798 and 79,262 contigs, respectively. As a first step in our comparison of metagenomes we explored the taxonomic diversity of bacteria associated to hosts. There were significant differences in taxonomic composition between the three gut communities of isopod, *Armadillidum vulgare* (Av), springtail, *Folsomia candida* (Fc) and termites, *Coptotermes gestroi* (Cg) as only 26.1%, 74.3% and 46.2% of contigs respectively, were taxonomically classified (Table 1). The communities are dominated by *Proteobacteria*, with 24,868 (80.59%), 45,666 (57.59%) and 17,061 contigs (46.65%) of the isopod, springtail and termite microbiomes, respectively. The second largest community in isopod and termite and the third largest in the springtail is *Firmicutes*, with 2,542 (8.24%), 7,348 (20.09%) and 5,359 (6.76%) contigs, respectively. Remarkably, the *Actinobacteria* with 22,156 contigs (27.94%) is the third largest group in the *F. candida* microbiome. Even though *Spirochaetes* were found in all gut communities, they are more abundant in the termite with 2,112 contigs (5.77%). Other common phyla are *Bacteroidetes* with 1,410 (4.57%), 4,836 (6.10%) and 3,343 (9.14%) contigs in, respectively, isopod, springtail and termite (Table 1).

Next, we considered the three functional gene categories of our interest: genes encoding carbohydrateactive enzymes for metabolism, genes encoding proteins related to antimicrobial resistance for defense and genes related to the production of secondary metabolites (Fig. 1).

Table 1: *Assembly metrics of the studied metagenomes*

	Armadillidium vulgare	*Folsomia candida*	*Coptotermes gestroi*
Total Contigs	118,575	106,798	79,262
Contigs (<5,000 bp)	116,411	105,611	77,422
contigs (>= 5,000 bp)	1,263	766	1,301
contigs (>= 10,000 bp)	634	315	407
contigs (>= 25,000 bp)	202	72	104
contigs (>= 50,000 bp)	65	34	28
Largest contig (bp)	435,086	1,306,495	183,852
Total length (bp)	91,552,032	69,056,649	90,150,744
Total length (< 5,000 bp)	34,831,365	15,817,019	59,385,192
Total length (>= 5,000 bp)	21,976,144	17,802,470	14,989,078
Total length (>= 10,000 bp)	17,579,018	14,693,322	9,065,738
Total length (>= 25,000 bp)	10,927,209	11,028,500	4,603,586
Total length (>= 50,000 bp)	6,238,296	9,715,338	2,107,150
N50	1,300	2,513	1,215
GC (%)	42.57	60.02	50.95
Taxonomic assignment (contigs)	30,857	79,289	36,575

A total of 1,392 contigs containing either one or all of the above functional groups were mapped backed into their taxonomic groups. An overview of the taxonomic groups contributing to all three functional gene categories is given in Fig. 1. There were 461, 303 and 628 contigs for isopod, springtail and termite, respectively. The same phyla as mentioned above were shown to contain genes contributing to carbohydrate metabolism, antimicrobial defense and secondary metabolism (Fig. 1A).

However, at the lower taxonomic level, there was considerable diversity within the phyla. Two large phyla, *Gammaproteobacteria* and *Alphaproteobacteria* (subdivisions of the large group of *Proteobacteria*) had different genera contributing to the functional metagenomes. The springtail metagenome had 20.13% *Alphaproteobacteria* and 8.25% *Gammaproteobacteria*. The isopod metagenome contained mostly *Gammaproteobacteria* at 88.50%, while the termite contained 24.52% (Fig. 1B).

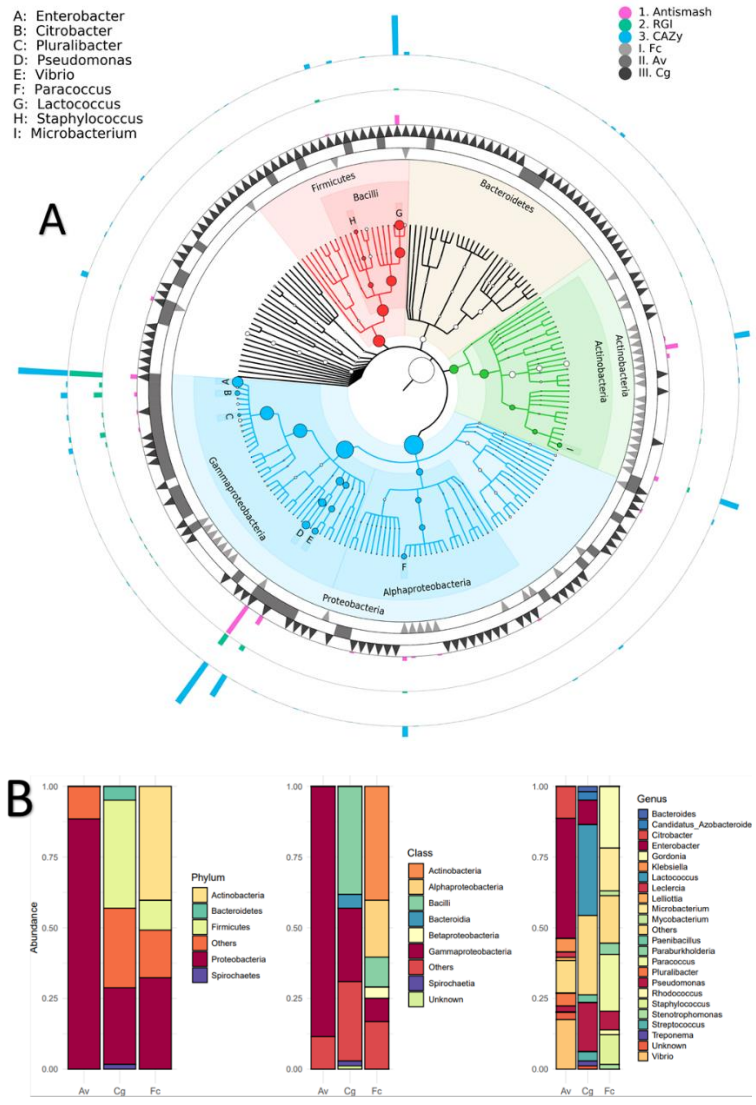


Figure 1A: *Phylogenetic distribution of bacterial taxa in the metagenomes of the three species of decomposer invertebrates. The centre circle shows a phylogeny of the contigs; the names of a few common genera are indicated by capital letters (see the upper left inset). The second ring indicates the names of the phyla to which they belong. The next three rings indicate in which hosts these genera were found (grey nabla symbol F. candida, dark grey A. vulgare and black triangle C. gestroi, see inset up right). The outer three circles indicate the number of contigs (by bar length) in the three functional categories (see inset up right): secondary compound biosynthesis (antiSMASH), antimicrobial resistance (RGI) and carbohydrate-active enzymes (CAZy).* ***B)*** *Taxonomic abundance at phylum, class and genus level. Taxonomic genera below 0.01% were merged together into the others group.*

The genus *Pseudomonas* contributed to all three functional groups in all three investigating gut microbiomes. The proportion of *Pseudomonas* contigs in the springtail, termite and isopod were 6.60%, 16.40% and 4.58%, respectively. Similarly, *Vibrio* was found in all functional groups and all metagenomes and was most abundant in the isopod (17.57%, Fig. 1B).

4.4.1 Carbohydrate-active enzymes

In total, 163 CAZy families were identified and divided into six functional classes. The overview of CAZymes in Fig. 2 shows that we identified 627, 905 and 648 full-length CAZy proteins for *A. vulgare*, *C. gestroi* and *F. candida*, respectively. Among the three species, the termite community possesses the greatest diversity of CAZymes in comparison to the isopod and springtail with 135 CAZy families. In all three metagenomes, *Proteobacteria* (669 contigs), *Actinobacteria* (148 contig) and *Firmicutes* (276 contigs) are the main CAZymes contributors (Fig. 2). These groups of bacteria are known to break down cellulose (López-Mondéjar et al. 2016). Within these phyla, the genera *Enterobacter*, *Pseudomonas*, *Vibrio*, *Lactococcus* and *Microbacterium* contain many CAZyme proteins.

The four main focus CAZyme classes for plant biomass degradation are glycosyl hydrolases (GH), carbohydrate esterases (CE), polysaccharide lyases (PL) and a group of enzymes classified as auxiliary activities (AA, redox enzymes that act in conjunction with CAZymes). Finally the non-catalytic carbohydrate binding molecules (CBM) can direct enzymes to the substrates and also help with cell-wall hydrolysis (Bernard et al. 2008; Biely 2012; Zhao et al. 2013; M. E. Taylor and Drickamer 2014).

A total of 43 common CAZy families were shared between the three metagenomes (Fig. 2). However, the PL class from springtail was low and did not share any CAZymes families with the other two. The top most common CAZy families detected in all three gut metagenomes were CBM50, GH1, GH13 and GH23 corresponding to 86, 67, 130 and 111 proteins, respectively. The GH1 family represents hemicellulose degradation activity, specifically through *beta-glucosidases* and *beta-galactosidases*. GH13 is one of the largest groups of glycosyl hydrolases which act on substrates containing α-glucoside linkages. This family is specialised in starch degrading and does not include cellulase or hemicellulase activities (López-Mondéjar et al. 2016). CBM50 can bind to N-acetylglucosamine residues in bacterial

peptidoglycans and chitin (Bussi and Gutierrez 2019). These peptidoglycans are then subject to lytic transglycosylases from family GH23 (Dik et al. 2017).

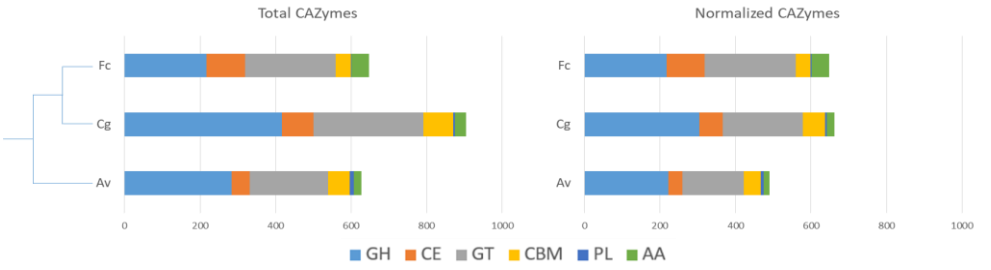


Figure 2: *Number of CAZymes classified by five categories, in the microbiome of the three arthropod host species. The evolutionary relationship between the hosts is given on the left. Fc = F. candida (springtail) Cg = Coptotermes gestroi (termite), Av = Armadillidium vulgare (isopod). GH = glucoside hydrolysases, CE = carbohydrate esterases, GT = glycosyl transferase, CBM = carbohydrate-binding molecules, PL = polysaccharide lyases, and AA = enzymes with auxiliary activities. The left graph provides the crude number, while in the right graph the numbers are normalized to the termite microbiome.*

Several CAZy families appeared to be host-species specific. Termites had 32 unique families, the largest being CBM9, GH106, GH113, GH29, GH32 and GH95. There were 3, 4, 3, 15, 6 and 3 proteins respectively. The second animal exhibiting the most diverse CAZymes is the springtail in which 16 of the 101 CAZy families are unique. The classes AA7 and CE5 had 7 and 6 proteins that were only found in springtails. In contrast, out of the 88 CAZy families found in the isopod, only 11 are unique and only GH127 appears to be isopod-specific. There were seven proteins in the GH127 family and all were predicted to be β-Larabinofuranosidase. Most of the host-specific glucoside hydrolases are hemicelluloses. Interestingly, springtails turned out to have the largest number of carbohydrate esterases (102 proteins, Fig. 3).

Enzymes from families CE8, PL1, PL2, PL9, GH28, GH78, and GH88 can break down pectin. Interestingly, most of these classes were not found in *F. candida*. This could indicate that pectin is not a main resource for the gut bacteria of the springtail, and the host may not rely on pectin as a carbon sources. Alternatively, the host could produce these enzymes itself and does not need the help of the microbiome. In contrast to springtails, the isopod metagenome appeared to contain multiple pectate lyases (PL1, 3 proteins), periplasmic

pectate lyases (PL2, 2 proteins) and pectinesterases (CE8, 5 proteins); the latter enzyme catalyzes the de-esterification of pectin to pectate and methanol. Only a single pectate lyase was found in the termite. So, it seems that isopods rely on pectin much more than termites and springtails (Supplement Table. 1).

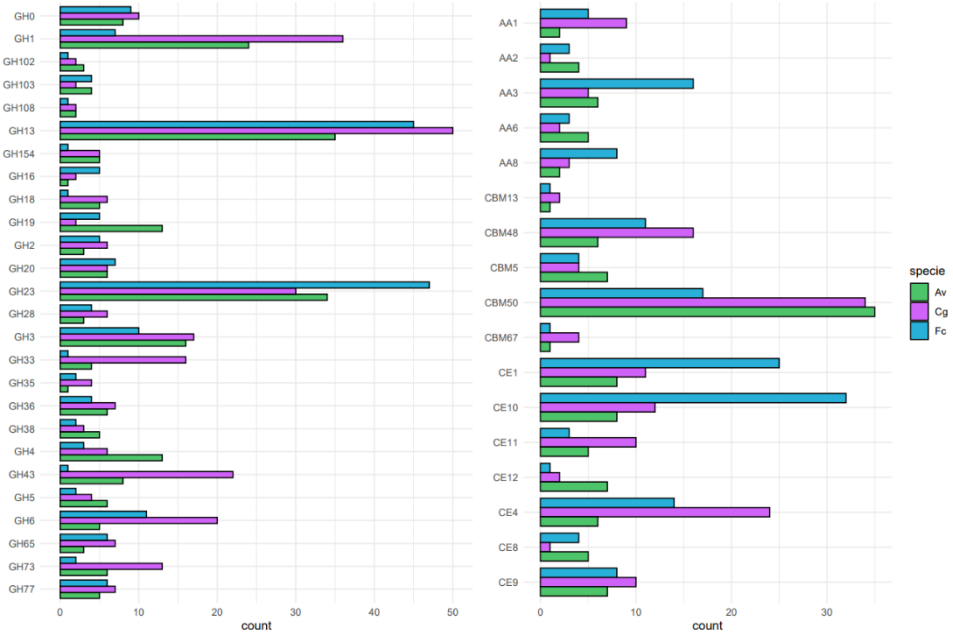


Figure 3: *Classification of carbohydrate-active enzymes identified in the microbial metagenomes of the three decomposer invertebrates, specified for four different functional categories. Righ: Enzymes with auxiliary activity (AA), carbohydrate-binding molecules (CBM), and carbohydrate esterases (CE). Lower graph: glucoside hydrolases (GH). The number of contigs falling into a functional group is given for each of the three hosts (stacked upon each other): Av= A. vulgare (isopod), Cg = C. gestroi (termite) and Fc = F. candida (springtail).*

Cellulase families GH1 (67 proteins), GH3 (43 proteins), GH5 (12 proteins) and GH6 (36 proteins) were found in all metagenomes in high abundance. This group of enzymes cleaves the β-1,4 bond in the cellulose chain. They are important for the breakdown of all dead plant biomass. There were 11 GH8 genes and 3 GH9 genes in both termites and isopods, respectively, but none of them in springtails. The lytic polysaccharide mono-oxygenase from the AA10 family, which can degrade cellulose, was found in springtails and termites (1 and 4 proteins, respectively).

Hemicelluloses such as xylans, xyloglucans, arabinoxylans and glucomannans can be broken down by a variety of enzymes. The enzymes from the families GH2 (14 proteins), GH16 (8 proteins) and GH43 (31 proteins) were found in all gut metagenomes. Xylanases from GH30, GH39 and GH11 were only found in the termite. Each of these CAZyme families only contain a single gene. The other xylanase families GH26 and GH42 were not found in the isopod, instead, a single gene from GH53 was isopod-specific. Carbohydrate esterases in CAZy families CE1 and CE10 contain 44 and 52 proteins respectively.

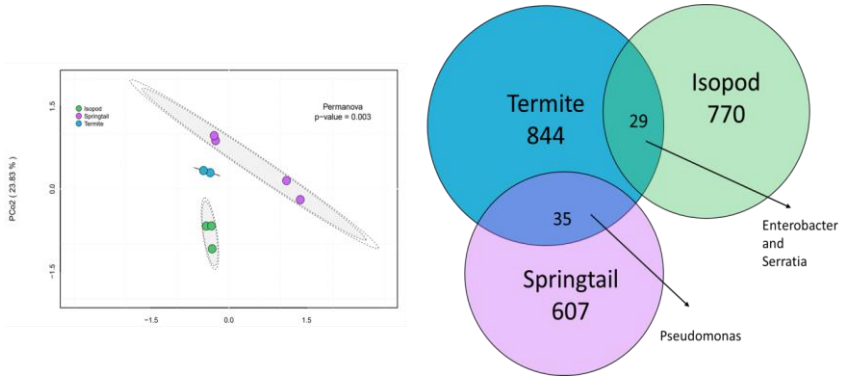


Figure 4A: *Bay-Curtis dissimilarity plot for microbiome samples colored by host. The significant was calculated using permanova with the p-value of 0.003. The first two axes explained 23.83% and 43.73% of the changes. The springtail was colored purple, termite blue and isopod green.* ***B:*** *Venn Diagram of bacterial genera found to be present with a carbohydrate-active gene with 90% similarity between in the metagenomes of three host species. Fc = F. candida (springtail), Cg = C. gerstoi (termite), Av = A. vulgare (isopod). Three common genera are indicated by name.*

Cutin is one of two waxy polymers that are the main components of the plant cuticle. The springtail is the only group that has 6 cutinases from the CE5 family, which attach to the ester bond to release cutin monomers.

The PCoA shows that for all samples from the three metagenomes clustered closely together (Fig. 4A). Using the dissimilarity matrix Bray-Curtis were generated with the permanova significant of 0.003. The sparseness of the springtail samples came from different DNA extraction methods (Valeria Agamennone et al. 2015; 2019). It is interesting to see that the termite is in the center of the isopod and the springtail. This was also observed in the HGT analysis, where the termite contain similar carbohydrate active genes to both the springtail and the isopod, but none shared similar genes together (Fig. 4B). Springtails and termites

share CAZy proteins from the genus *Pseudomonas*, while isopods and termites share CAZy proteins from *Enterobacter* and *Serratia*. There was no overlap at all between springtails and isopods. Even on the level of functional categories (with a more relaxed criterion for similarity at 90%), there is hardly any overlap. The unique character of the CAZys repertoire in each microbiome is remarkable.

4.4.2 Antibiotic resistance genes

The isopod, termite and springtail metagenomes were scanned with Resistance Gene Identifier (RGI) using the Comprehensive Antibiotic Resistance Database (CARD), with default settings. The CARD database (version 3.0.5) is one of the most well developed AR Ontology (ARO) available with 82 pathogens, 67,366 resistomes and 92,896 AMR allele sequences.

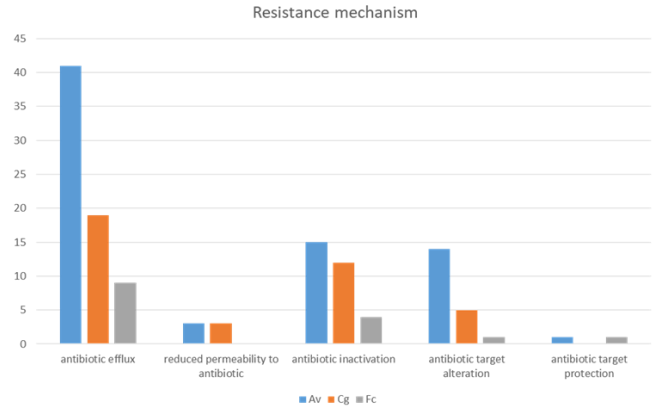


Figure 5: *Number of genes classified as antibiotic resistance genes, in the microbial metagenomes of the isopod (Armadillidium vulgare, Av), the termite (Coptotermes gestroi, Cv) and the springtail (Folsomia candida, Fc). The genes are classified according to five different functional categories from CARD database.*

The ARG profiles of the three gut metagenomes appeared to be quite diverse (Fig. 5). The isopod metagenome had 75 predicted ARGs, followed by the termite metagenome with 43 ARGs and 18 ARGs for springtail gut. With regard to antibiotic mechanisms the isopod is the most diverse in terms of the five functional groups described above. Out of these, antibiotic efflux and inactivation are the two most common mechanisms in all gut metagenomes. Genes encoding antibiotic efflux pumps were 44, 25 and 11 for isopod, termite

and springtail respectively. All gut metagenomes contain the resistance-nodulationcell division (RND) efflux pumps, but also the major facilitator superfamily (MFS) and ATP-binding cassette (ABC) antibiotic efflux pumps. Termites have multidrug and toxic compound extrusion (MATE) transporters which were not found in the metagenomes of the other hosts (Piddock 2006a).

Mechanisms relying on antibiotic inactivation cause changes to the antibiotic compound itself (degradation, binding), so that it can no longer affect the target (Hoffman 2001). The isopod contained 15 antibiotic inactivation ARGs, 12 were found in the termite and 4 in the springtail. Beta-lactamases (enzymes that degrade beta-lactam antibiotics) were most common. Interestingly, aminoglycoside was absent in the springtail, and only identified in isopod and termite guts.

A similar situation holds for antibiotic target alteration proteins. These were abundant in the isopod metagenome (12 genes), while the termite genome contained only 3 genes and none were found in the springtail (Fig. 6). However, the springtail metagenome was the only one to contain genes associated with antibiotic target replacement. This mechanism relies on the production of alternative proteins that function in a similar way as the principal antibiotic target proteins but through a slightly different structure. The more alternative proteins present, the less active antibiotics can reach the correct target.

The antibiotic target protection mechanism of the springtail metagenome is predicted to be directed against tetracycline. Different from springtails, the isopod gut metagenome had glycopeptides and other peptides as antibiotic target replacements, predicted to act against fluoroquinolone. The antibiotic target protection mechanism was absent from the termite metagenome. The springtail metagenome also contained the largest group of target replacement mechanisms addressing penicillin antibiotics (penam). In this resistance mechanism, a protein similar to the antibiotic target is produced, but with lower binding affinity to the active antibiotic so the cell is rescued from antibiotic inhibition. Both the termite and the isopod are lacking this mechanism. Finally, low permeability of the outer bacterial cell wall is another mechanism to become resistant. This mechanism can be observed in the termite and isopod metagenomes but is lacking in the springtail metagenome.

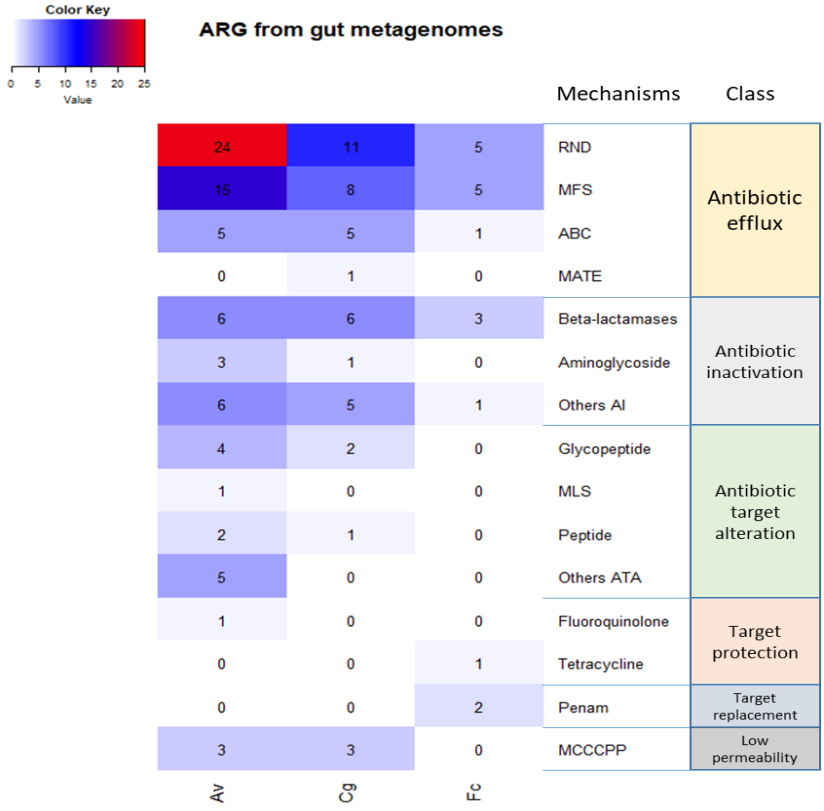


Figure 6: *Classification of antibiotic resistance genes in the metagenomes of the three invertebrate species, according to six functional categories. RND = resistance-nodulation-division efflux pumps, MFS = major facilitator superfamily efflux pumps, ABC = ATP binding cassette efflux pumps, MATE = multidrug and toxic compound extrusion efflux pumps. MLS = macrolide, lincosamide and streptogramin antibiotics. ATA = aurintricarboxylic acid. Penam = penicillin antibiotics, MCC-CPP = maleidomethyl-cyclohexane-carboxylate bound to cell penetrating peptides. Av = Armadillium vulgare (isopod), Cg = Coptotermes gestroi (termite), Fc = Folsomia candida (springtail).*

In summary, our survey of antibiotic resistance genes in the guts of the three invertebrates shows striking differences between the hosts, which are much more profound in comparison to the large taxonomic overlap of the microbial communities on the level of bacterial phyla (Table 2).

Table 2. *Bacterial composition (in %) (by BLAST within the contig) of microbial metagenomes associated with the three different invertebrate hosts.*

	Hosts		
Bacterial phylum	*Armadillidium vulgare*	*Folsomia candida*	*Coptotermes gestroi*
Actinobacteria	1.72	27.9	7.61
Bacteriodetes	4.57	6.10	9.14
Cyanobacteria	1.11	0.10	0.94
Firmicutes	8.24	6.76	20.1
Planctomycetes	0.13	0.13	1.22
Proteobacteria	80.6	57.6	46.7
Spirochaetes	0.45	0.02	5.77
Tenericutes	1.04	0.03	0.14
Not assigned	2.14	1.36	8.38

4.4.3 Gene clusters involved in secondary metabolite biosynthesis (antiSMASH)

The assembled sequences from gut metagenomes of all three soil invertebrates were mined using antiSMASH. A total of 17 types of biosynthetic gene clusters (BGCs) appeared in 115 contigs across three gut metagenomes (Fig. 7). Non-ribosomal peptide synthetases (NRPS), bacteriocin, arylpolyene and siderophores are among the most common secondary metabolites that appeared in all three organisms. The NRPS gene clusters function similarly to an assembly line, where multiple genes modify the metabolite. Non-ribosomal peptides (NRP) are the final products and have a wide variety of biological functions, from iron

acquisition, to insecticidal, nematicidal, phytotoxic, antimicrobial, and antiviral activities (Le Govic et al. 2019).

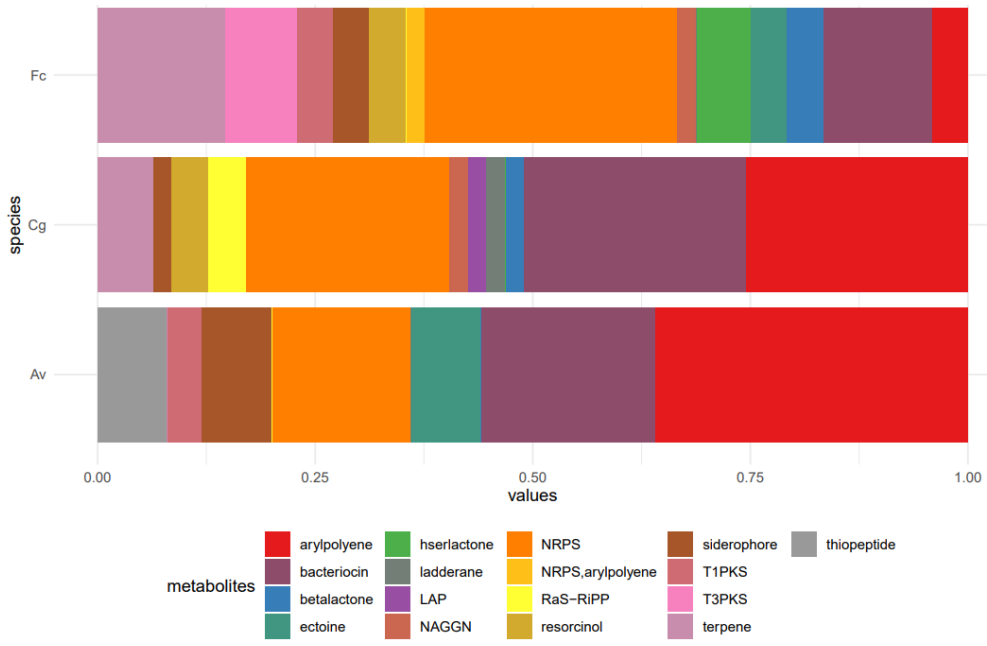


Figure 7: *Relative composition of biosynthetic gene clusters in the metagenomes of the three invertebrate species. The different colors indicate different structural categories identified using the AntiSMASH database. T1PKS = Type I polyketide synthase, NAGNN = N-acetylglutaminylglutamin, T3PKS = type III polyketide synthase, NRPS = non-ribosomal peptide synthase, Ras-RIPP = ribosomally synthesized and posttranslationally modified peptides, LAP = lingual antimicrobial peptide. The three hosts are Av = Armadillidium vulgare (isopod), Cg = Coptotermes gestroi (termite), and Fc = Folsomia candida (springtail).*

The springtail metagenome contained the most diverse set of such genes with 13 types of BGCs (Table 3). These include NRPSs, and clusters encoding the synthesis of bacteriocins, terpenes, homoserine lactone, siderophores, betalactone, type III polyketide synthases (T3PKS), and ectoine. An osmoregulation cluster called N-acetyl-L-glutaminyl-L-glutamine amide (NAGGN) and a siderophore biosynthetic gene cluster were found on one contig of 745 kbp long (Fig. 8A). Siderophores are used by bacteria to acquire iron from the environment; they are typically induced by microbial infection (Holden and Bachman 2015; Page 2019; Kramer, Özkaya, and Kümmerli 2020). The siderophore gene cluster is 53 kbp

long and has a 97 - 100% similarity to gene clusters from different *Pseudomonas* species (*Pseudomonas* sp. NFIX49 NZ_FOYE01000001_c1, *Pseudomonas* sp. GM25 NZ_AKJQ01000040_c2 and *Pseudomonas fluorescens* strain H24 NZ_LACH01000031_c3). This cluster shows 21 % similarity with a pyoverdin biosynthesis cluster in *Pseudomonas protegens* Pf-5 (BGC0000413 from MIBiG database) (Kautsar et al. 2020). Pyoverdin is a virulence factor. By regulating iron availability it can secure iron as a nutrient, but also regulate virulence factors and biofilm formation. Free iron is toxic and can have antimicrobial properties (Kang et al. 2018). Another large contig, of 650 kbp, contained two bacteriocin clusters (Fig. 8B). Bacteriocins are ribosomally synthesized peptides, which are produced to be active against various strains of bacteria (Yang et al. 2014; Chikindas et al. 2018). The most interesting contig with a length of 409 kbp has four different types of BGCs: bacteriocin, NRPS, NRPS with arylpolyene and siderophore biosynthesis genes (Fig. 8C). This contig is annotated to be homologous to *Pseudomonas* sp. The central NRPS has 68% similarity towards NZ_JTGH01000016_c3 from *Pseudomonas fluorescens*. About 71% of this gene cluster is similar to a lokisin biosynthesis cluster, which is a plant antifungal identified in *Pseudomonas* spp. (Omoboye, Oni, et al. 2019; Gu et al. 2020; Omoboye, Geudens, et al. 2019). The other NRPS, annotated as arylpolyene is predicted to produce rimosamide. This secondary metabolite acts against the antibiotic activity of blasticidin (McClure et al. 2016). These valuable BGCs have been found in other *Pseudomonas* strains as well, and help to protect the cell against infection of various pathogenic bacteria and fungi. Furthermore, four T3PKS (type III polyketide synthases) were only found in the springtail metagenome. The best predicted gene cluster is 43 kbp long containing multiple biosynthesis, regulatory and core genes. The whole cluster has 60% genes similarity toward *Microbacterium* species. However, this could be a new cluster as we could not find a good homolog in the databases; it shows only 4% similarity with regard to the formicamycins A-M biosynthetic gene cluster from *Streptomyces*. Formicamycin is also known to have antibacterial activity (Qin et al. 2020).

Table 3. *Number of secondary metabolite biosynthetic gene clusters (BCGs) assigned to 17 different metabolite categories in the gut microbiomes of the three species of soil invertebrate (Av = Armadillidium vulgare, woodlouse; Fc = Folsomia candida, springtail, and Cg = Coptotermes gestroi, termite). For explanation of gene cluster names, see the legend to Fig. 7.*

Secondary metabolite encoded by BGC	*Armadillidium vulgare*	*Folsomia candida*	*Coptotermes gestroi*
arylpolyene	9	2	12
thiopeptide	2	0	0
NRPS	4	14	11
bacteriocin	5	6	12
T1PKS	1	2	0
ectoine	2	2	0
siderophore	2	2	1
betalactone	0	2	1
NAGGN	0	1	1
T3PKS	0	4	0
NRPS, arylpolyene	0	1	0
terpene	0	7	3
hserlactone	0	3	0
resorcinol	0	2	2
RaS-RiPP	0	0	2
LAP	0	0	1
ladderane	0	0	1

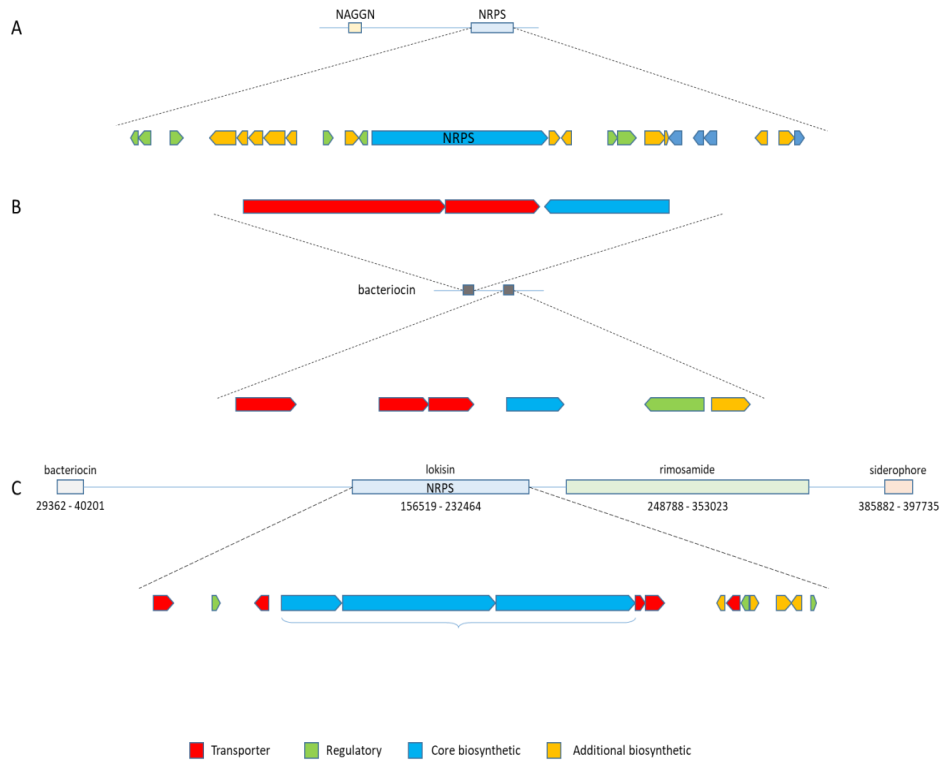


Figure 8: *Overview of a DNA segment in the contigs of the springtail gut metagenome:* ***A****: cluster of NAGGN and NRPS.* ***B****: two bacteriocins on the same contig.* ***C****: encoding four different biosynthetic gene clusters, a bacteriocin producing gene, an NRPS expected to synthesize lokisin, another one for rinosamide production, as well a siderophore encoding gene. These contigs are annotated to Pseudomonas.*

The termite gut metagenome contains NRPS clusters that are quite different from the two other host metagenomes. A cluster of biosynthesis genes of 46 kbp shows 100% identity toward *Pseudomonas fluorescens.* It is 50% similar to a bananamide 1-3 biosynthesis cluster from the same species. This metabolite confers antimicrobial activity against the oomycete (water mould) *Pythium myriotylum* and the ascomycete fungus *Pyricularia oryzae* (Omoboye, Geudens, et al. 2019). Interestingly, a cluster of NRPS genes predicted to biosynthesize ralsolamycin (40% similarity) is also present in the termite metagenome. This metabolite is an inducer of chlamydospore formation in fungi (Baldeweg et al. 2017). The termite also shows the highest number of bacteriocins. This could be related to the diversity of microorganisms in the termite gut (see above). Finally, the termite metagenome also

contains type III polyketide synthases (different from the ones in springtails), which are known to produce active enzymes synthesizing antimicrobials.

A cluster that was found in both springtail and termite metagenomes is linked to betalactome biosynthesis with 13% genes similar to fengysin biosynthesis genes in *Bacillus velezensis*, an antifungal compound. Another NRPS cluster shared between termite and springtail metagenomes contains a single NRPS gene, and shows 100% similarity to *Paraburkholderia rhizoxinica*. The gene contains conserved domains of gene clusters to produce rhizomide A-C. They were 1.7 kbp and 3.1 kbp long in the springtail and termite metagenomes, respectively. Rhizomide A is shown to have weak antitumor properties in human cell lines (X. Wang et al. 2018).

The isopod metagenome contains two thiopeptides gene clusters that are 63% and 68% identical to the homologous gene clusters in *Enterobacter cloacae* and *Pluralibacter gergoviae* respectively. This group of antibiotics is directed exclusively to Gram-positive bacteria and has no antibiotic effects on Gram-negative bacteria. Most of the secondary metabolite clusters found in the isopod metagenome are aryl polyene BGCs, which are responsible for pigmentation in Gram-negative bacteria.

4.5 Discussion

Our analysis revealed remarkably differences in functional genes of the metagenome of three soil invertebrate species. The functions explored are expected to be crucial to life in soil: carbohydrate degradation, antimicrobial defense and production of secondary metabolites. Our findings show that the termite contains the most diverse community of microorganisms followed by the springtail and the isopod. This result is consistent with our previous finding (Valeria Agamennone et al. 2019). The top five common prokaryotic genera in microbiota are *Paraburkholderia, Pseudomonas, Stenotrophomonas, Enterobacter* and *Microbacterium*. However, there are very large differences in community composition and only few genera are present in all hosts. *Pseudomonas* is the only common genus found in all three gut metagenomes. This contrasts with the relatively large similarity of microbiomes when classified by bacterial phylum.

Carbohydrate degradation is achieved by carbohydrate-active enzymes and include all proteins that bind to carbohydrates, hydrolyse glycoside bonds in polysaccharides, cleave off

specific side chains, etc. Obviously, the activity of such enzymes is crucial for the nutrition of soil invertebrates which often consume large amounts of organic material of plant and fungal origin. It may be expected that all invertebrates, like many higher animals, rely on their microbiomes to ensure appropriate nutrition. One of the most dominant enzymes in the metagenomes are glycosyl hydrolases, which hydrolyse the glycosidic bond between carbohydrates or between carbohydrates and non-carbohydrate moieties (protein or lipid). In addition, the cross linking of hemicellulose with lignin in plant biomass is weakened by carbohydrate esterases. These enzymes de-acetylate the polysaccharide side-chains. Pectin is a component of the plant cell wall with a function in cell adhesion and cell wall hydration. Glycosidic bonds between carbohydrates of glycosaminoglycans and pectin are broken down by polysaccharide lyases, using a non-hydrolysis mechanism (Xiao and Anderson 2013).

Antibiotics are widely used to combat bacterial infections in health care, agriculture and animal farming. However, the anthropogenic overuse of antibiotics constitutes a severe hazard since the number of resistant pathogens increases (Kraemer, Ramachandran, and Perron 2019). Microorganisms in the environment are known to evolve resistance against a large number of antibiotics. Antibiotic resistance can spread rapidly in a microbial community when the resistance genes are encoded on plasmids or mobile genetic elements such as integrons and transposons. Using the metagenomics approach can help to broaden knowledge regarding the type of antibiotic resistance as well as mechanisms, transmission and evolution of microorganisms from a specific mini ecosystem (Garmendia et al. 2012; Mullany 2014; Watford and Warrington 2018).

The third investigated gene category comprises gene clusters involved in secondary metabolite biosynthesis, which could have novel and interesting properties (Khater, Anand, and Mohanty 2016; Naughton et al. 2017; Zheng et al. 2019). By investigating key functional attributes of the microbial metagenomes, carbohydrate enzyme activity, antibiotic resistance and secondary metabolites, as we did in this paper, it may be possible to achieve a better understanding of the interaction of microorganisms with their hosts, including the host's lifestyles, food sources and ecological functions in the soil environment.

Below we compare the three different invertebrates with respect to the above-mentioned gene categories encoded in their metagenomes.

4.5.1 Termite

The number of CAZymes in the termite metagenome is much larger than the number in the springtail and the isopod, and it has the most diversified microbiome. It contain 32 unique CAZy families not found in the other species. It is known that the so-called lower termites such as *Coptotermes gestroi* harbour various groups of bacteria, archaea and protists for the digestion of lignocellulose (Tai et al. 2015). We found that *Firmicutes* and *Proteobacteria* are two major contributors to cellulases. They are also more prevalent than *Spirochaetes*, which are often found in large abundance in lower and higher wood eating termites. This could be related to environmental factors, diets or host genetics (X. F. Huang et al. 2013). The phylum of *Firmicutes* is known to have several cellulose fermentors, which are important to lignocellulose breakdown (Su et al. 2017). Some of them can work in alkaline solution (Husseneder 2010). Other well-known glycosyl hydrolase groups for cellulose and hemicellulose degradation (GH1, GH9) are also present. Similar observations were done in the gut microbiome of a higher termite from Brazil (Grieco et al. 2019). We also found multiple hemicellulose degrading CAZy groups solely in the termite: (endo-beta-1,4-xylanase, α-L-fucosidase, α-glucuronidase, beta-mannase, and beta-xylosidase). This shows the diversity of enzymes that the termite microbiome deploys to breakdown different types of hemicellulose.

Regarding antibiotic resistance, the termite gut metagenome is the only one that has genes encoding multidrug and toxic compound extrusion (MATE), which consist of Na+/H+ drug antiporters. This system is found in Gram-positive microorganisms (Piddock 2006a). The resistance gene diversity found in the termite is intermediate between the isopod and the springtail (Peterson and Scharf 2016a; 2016b).

Besides, the termite gut metagenome contained different clusters of genes encoding biosynthesis of secondary metabolites with anti-bacterial and fungal properties. The termite gut is a great place to identify novel antifungal as there are many bacteria protect the host from antagonistic fungi (Um et al. 2013; Benndorf et al. 2018). We identified a bananamide and ralsolamycin gene synthesis clusters, that have antifungal properties.

4.5.2 Springtail

Collembola include a wide variety of feeding habits, varying from root-eating, fungivory and detritivory to predation, with very little food specialization, although there are diverging views on this (M. P. Berg, Stoffer, and Van Den Heuvel 2004). *Folsomia candida* is usually considered a fungivore, although some claim it prefers nematodes over soil fungi (Q. Lee and Widden 1996). In the laboratory, they readily feed on yeast. A remarkable property of *F. candida* is that it is extremely resistant to entomopathogenic fungi known for quickly kill termites and ants (Broza, Pereira, and Stimac 2001), which suggest that a living with fungi is the prime lifestyle of this animal.

In our survey of CAZymes in the springtail metagenome, we found a low number of polysaccharide lyases and a large number of carbohydrate esterases, no pectinases but many cutinases. Since pectin is a typical constituent of plant cell walls and cutine is found in the plant cuticle, this would suggest that *F. candida* is better equipped to feeding on the surface of plant leaves than degrading the cell wall itself. Enzymes with chitinase activity, contributing to the degradation of fungal cell walls (such as CBM50 and GH23) are also found in the springtail metagenome, but these belong to CAZymes that are shared between the three hosts.

In previous studies we have shown that the mycorrhizal fungus (AMF) *Rhizophagus irregularis* (*Glomus intraradices*) is a food source for the springtail (Duhamel et al. 2013; Faddeeva-Vakhrusheva et al. 2017). Five genes from the group AA1 were found in the springtail's gut metagenome, which are known for their laccase activity. Another large group, AA3, contains cellobiose dehydrogenases, which oxidize cellobiose and cellodextrins to produce glucose (Sützl et al. 2018). For the breakdown of hemicellulose, the springtail gut metagenome contains large amounts of acetyl xylan esterases. There are more xyloglucanases (GH16) from the springtail than the other two gut metagenomes. Crystal cellulose-binding enzymes are more abundant in the springtail. Overall, this shows that the springtail contains some but not a very complex cocktail of enzymes to break down lignocellulose.

In terms of antibiotic resistance, the springtail microbiome contained fewer genes than identified in the other two invertebrate gut microbiomes. However, it is the only metagenome that encodes proteins for protection against tetracycline and penan. Also remarkable, the springtail gut microbiome contains the largest number of gene clusters encoding biosynthesis

of secondary metabolites, among the three host species microbiomes investigated. It shows a wide-ranging capacity for the production of antibiotics as well as antifungal metabolites within its metagenome. Many of the large assembled contigs are assigned to *Pseudomonas* spp., which contains bacteriocin and antifungal gene clusters. We reported on one of the longest contigs with four different types of BGCs: bacteriocin, lokisin, rimosamide and siderophore biosynthesis genes were present. Bacteriocins are multifunctional substances, which are produced by the ribosomes. At certain concentration, bacteriocins display antimicrobial activities and can stop biofilm formation through the inhibition of quorum sensing. Some can have additional properties such as interfering with cell division process as well as biological functions (Algburi et al. 2017; Chikindas et al. 2018). The lokisin cluster was shown to trigger systemic resistance and direct antagonism against *Magnaporthe oryzae* as well as inhibiting fungal growth (Hultberg et al. 2010; Omoboye, Oni, et al. 2019; Gu et al. 2020). The rimosamide and associated NRPS/PKS-type gene cluster contain a very similar structure and biosynthesis to detoxin family (Yonehara et al. 1968). This cluster was observed in *Streptomyces rimosus* and is capable of negating the antibiotic activity of blasticidin from *Bacillus cereus* (McClure et al. 2016). Another observed cluster is the siderophores pyoverdine. During an infection, *Pseudomonas aeruginosa* produce pyoverdine, which is a core set of virulence factors. It can also act as a siderophore for absorbing iron from the environment for biofilm formation (Kang et al. 2018; Bonneau, Roche, and Schalk 2020). This would allow the springtail to benefit from resistance to entomopathogenic fungi, although a causal relation has, of course, not yet been demonstrated. We have previously shown that some of the antibiotic-producing genes have migrated to the host's genome by horizontal gene transfer (Suring et al. 2017).

4.5.3 Isopod

For hemicellulose degrading enzymes, a large number of α-galactosidases from GH4 and GH31 were observed. The isopod metagenome also has more hemicellulase and/or cellulase genes from GH8 and GH9. Many of these enzymes are due to *Proteobacteria*. The other phyla such as *Spirochaetes* and *Firmicutes* are present but in low abundance (Bredon et al. 2018; 2020). Another remarkable property of the isopod metagenome is that it is represented by numerous pectin-degrading enzymes, which are totally absent from the springtail

metagenome. This suggests that isopods, much more than springtails are equipped to degrade the cell wall polysaccharides of plant leaves, which accordant with their dietary preference.

Interestingly, the isopod metagenome contains a large number of antibiotic resistance genes. To our knowledge, this is new finding as this has not been reported earlier in literature. The wide range of resistance genes encoding membrane pumps is particularly striking. The resistance/nodulation/division (RND) and major facilitator superfamily (MFS) are the two largest antibiotic efflux mechanisms in the isopod. The RND superfamily is found specifically in the Gram-negative microorganisms, where they form a tripartite complex across the two membranes. The MFS family is widely distributed in Gram-positive and Gram-negative bacteria. These pumps are activated to eliminate endogenous toxic compounds (Piddock 2006a; Nikaido 2010; Blanco et al. 2016; Piddock 2006b). By understanding these pumps, it is possible to design inhibitors to target efflux pumps of resistance microorganisms (Blanco et al. 2016).

The isopod gut metagenome also contains thiopeptide biosynthetic gene clusters, which affect Grampositive but not Gram-negative bacteria. They are macrocyclic peptide antibiotics and can be used clinically to combat pathogenic *Staphylococcus* and *Bacillus* infections. They are also valuable as they can inhibit the protein synthesis in Gram-positive bacteria of methicillin-resistant Staphylococcus aureus (MRSA) and vancomycin-resistant *Enterococcus faecium* (VRE). They have been shown to have antimalarial and anticancer activities (Rogers, Cundliffe, and McCutchan 1998; Donia, Ravel, and Schmidt 2008; Engelhardt, Degnes, and Zotchev 2010).

4.6 Conclusions

By investigating different functional gene groups encoded in the microbiomes of three different hosts it is possible to match some of the microbiome functionalities to the host environment as well as to their feeding habits. This functional diversity lies underneath an appreciable similarity in microbial community composition at high taxonomic levels. Different communities play similar but also different roles in different host animals. Our study illustrates the complexity of interactions between soil invertebrates, their microbiomes and the soil microbial community and the inappropriateness of lumping them together as simply “decomposers”.

Supplements

Supplemental Table 1: *Overview of CAZymes found in Armadillidium vulgare (Av), Folsomia candida (Fc) and Coptotermes gestroi (Cg)*

Supplemental Table 1. *Overview of CAZy families found in the metagenomes of Armadillidium vulgare (Av), Folsomia candida (Fc) and Coptotermes gestroi (Cg). The number of identified genes falling into each category is specified.*

	CAZy family	Known activities	Av	Fc	Cg
Lignin modifying enzymes	AA1	Laccase	2	5	9
	AA2	Manganese peroxidase; versatile peroxidase; lignin peroxidase	4	3	1
	AA3	Cellobiose dehydrogenase	6	16	5
Hemicellulases	CE1	Acetyl xylan esterase; feruloyl esterase	8	25	11
	CE3	Acetyl xylan esterase	2	0	0
	CE4	Acetyl xylan esterase	6	14	24
	CE5	Acetyl xylan esterase	0	6	0
	CE7	Acetyl xylan esterase	0	1	4
	CE12	Acetyl xylan esterase	7	1	2
	GH2	β-galactosidase; β-mannosidase; α-L-arabinofuranosidase	3	5	6
	GH4	α-galactosidase	13	3	6
	GH11	Endo-β-1,4-xylanase	0	0	2
	GH16	Xyloglucanase	1	5	2
	GH27	α-galactosidase	0	1	1
	GH29	α-L-fucosidase	0	0	15
	GH31	α-galactosidase; α-xylosidase	5	0	2
	GH35	β-galactosidase	1	2	4
	GH36	α-galactosidase	6	4	7
	GH39	β-xylosidase	0	0	1
	GH42	β-galactosidase	0	1	1
	GH43	β-xylosidase; α-Larabinofuranosidase; arabinanase; xylanase	8	1	22
	GH53	Endo-β-1,4-galactanase	1	0	0
	GH57	α-galactosidase	0	1	3
	GH67	α-glucuronidase	0	0	1
	GH113	β-mannanase	0	0	3
	GH116	β-xylosidase	0	0	2
	GH120	β-xylosidase	0	0	2
Hemicellulases and/or cellulases	GH1	β-glucosidase; β-galactosidase; exo-β-1,4-glucanase; βmannosidase; β-xylosidase	24	7	36
	GH3	β-glucosidase; exo-β-1, 4-glucanase; xylan 1,4-β-xylosi dase;α-L-arabinofuranosidase	16	10	17
	GH5	Endo-β-1,4-glucanase; βglucosidase; exo-β-1,4-glucanase; endo-β-1,4-xylanase; βmannosidase; endo-β-1,4-manno sidase;	6	2	4
	GH6	Endo-β-1,4-glucanase; cellobiohydrolase	11	20	5
	GH8	Endo-β-1,4-glucanase; endo-1, 4-β-xylanase	9	0	2
	GH9	Endo-β-1,4-glucanase; βglucosidase; exo-β-1,4-glucanase; cellobiohydrolase	2	0	1
	GH30	β-glucosidase; endo-β-1, 4-xylanase; β-xylosidase	0	0	2
	GH51	Endo-β-1,4-glucanase; endo-β-1,4-xylanase; β-glucosidase; β-xylosidase; α-L-arabinofuranosidase	0	1	4
	GH94	Cellobiose phosphorylase	1	0	1
Lignocellulose-binding modules	CBM6	Cellulose-binding	0	0	1
	CBM9	Crystal cellulose-binding	0	3	0
	CBM13	Xylan-binding	1	1	2
	CBM23	Mannan-binding	0	2	0
	CBM32	Galactose-binding	0	1	5
	CBM35	Xylan, mannans and β-galactan binding	2	0	0
	CBM51	Galactose-binding	1	0	3
	CBM67	L-rhamnose-binding	1	1	4
Pectin	PL1	Pectate lyases	3	0	1

Chapter 5 - Functional characterization of hemicellulose degrading enzymes from animal gut microbiomes

Ngoc Giang Le, Peter van Ulsen, Rob van Spanning, Tung Lam Le, Mohamed Aliawi, Abraham Brouwer, Nico M. van Straalen, Dick Roelofs, Thi Huyen Do and Nam Hai Truong

5.1 Abstract

In this study, we identified two hemicellulose degrading enzymes, an α-L-arabinofuranosidase (LAraf43) and an α-glucuronidase (PGluc67) in two different animal gut metagenomes. Carbohydrate Activity enZyme (CAZy) database was used to identify these enzymes. The LAraf43 is predicted to be a glycoside hydrolase of family 43 from the *Coptotermes* termite gut bacterium *Lactococcus lactis*. The PGluc67 protein sequence was deduced from goat gut metagenomes and predicted to be a glycoside hydrolase of family 67 from a non-characterized *Prevotella* species. Both enzymes were expressed and characterized in *Escherichia coli*. The activity assays with purified enzymes revealed that LAraf43 hydrolyzed synthetic p-nitrophenol-α-L-arabinofuranoside at 37°C and pH 7.4 with a K_m of 0.104 ± 0.05 mg/ml and a V_{max} of 12.4 ± 5.0 U/mg. The PGluc67 hydrolyzed aldutriouronic acid with a K_m of 3.92 ± 1.76 mM and a V_{max} of 55.0 ± 17.6 U/mg at 37°C and pH 7.4.

5.2 Introduction

The breakdown of lignocellulose through the action of recombinant enzymes is receiving increased attention, because it could facilitate a more sustainable way of generating organic building blocks for industrial use. Plant cell wall material is one of the most abundant carbon resources on earth and a possible source of bio-based chemicals. Lignocellulose is the main component of the dry woody part of a plant and consists of 23%-38% hemicellulose and 41%-51% cellulose depending on the plant species (H. Chen 2014a; Boonmee 2012). Currently, lignocellulose is broken down by chemical means to isolate fermentable monomeric sugars such as glucose, xylose and pentoses at a high cost of energy and/or waste (Das et al. 2012; Amin et al. 2017). Complete enzymatic degradation of lignocellulose is difficult as each component in the complex polysaccharide structure is made up of distinct precursors and linkages and so will require multiple enzymes to hydrolyze the various bonds (de Souza 2013; Bornscheuer, Buchholz, and Seibel 2014). Therefore, it is important to identify enzymes with novel functions or improved catalytic activities that may foster such bio-based recovery of fermentable sugars from lignocellulose.

The microbiome present in the gut of wood-feeding animals is a potential source of novel plant cell wall-degrading enzymes, because the associated microbial communities have coevolved with their hosts (Brune and Dietrich 2015; Puniya, Singh, and Kamra 2015; Valeria Agamennone et al. 2019). Potentially interesting enzymes that can break down

complex polysugars/polysaccharides such as hemicellulose have already been isolated from microbiomes of termites (Brune 2014; Hongoh 2011; Ni and Tokuda 2013; Breznak and Brune 2002) and goats (Do, Le, et al. 2018; K. T. Lee et al. 2018; Al-Masaudi et al. 2019; G. Wang, Luo, Meng, et al. 2011; G. Wang, Luo, Wang, et al. 2011). These enzymes cleave side-chains and glycosidic linkages in the polymeric backbone to release sugars that may be used by the microbes as an energy source. The two major mechanisms of such glycoside hydrolases are known as i) inverting, where the product has a stereochemistry opposite to the substrate and ii) retaining, where the anomeric configuration of the product is the same as that of the substrate (Sweeney and Xu 2012).

Hemicellulose consists of a heteropolymer of the pentoses D-xylose, L-arabinose and hexoses, such as D-mannose, D-glucose and D-galactose (Ravindran and Jaiswal 2016). It strengthens the lignocellulose matrix by linking cellulose microfibrils and lignin together (Gírio et al. 2010). Due to its heterogeneous structure, multiple de-branching enzymes, such as α-arabinofuranosidase, α-glucuronidase, acetyl-xylan esterase and phenolic acid esterase, are required for the complete degradation of hemicelluloses. This is a critical step, because their removal disrupts the lignocellulose matrix and makes the structure more accessible to other enzymes, such as endo-xylanases, eventually leading to complete hydrolysis into monomer sugars (Lagaert et al. 2014; C. C. Lee, Kibblewhite, Wagschal, Li, Robertson, et al. 2012).

The α-L-arabinofuranosidases hydrolyze the terminal α-L-arabinofuranosyl groups from L-arabino-containing polysaccharides and oligosaccharides. These enzymes are found in different CAZy families but are mostly abundant in glycoside hydrolase family 43 (GH43). All enzymes in this family are characterized by a 5-fold β-propeller structure and are of the so-called inverting type, yet their activities are diverse and range from β-xylosidase; α-L-arabinofuranosidase; xylanase; α-1,2-L-arabinofuranosidase; exo-α-1,5-L-arabinofuranosidase; exo-α-1,5-L-arabinanase; β-1,3-xylosidase; exo-α-1,5-L-arabinanase; endo-α-1,5-L-arabinanase; exo-β-1,3-galactanase to β-D-galactofuranosidase (www.cazy.org) (Maehara et al. 2014; Dimarogona and Topakas 2016).

Degradation of polysaccharides and oligosaccharides by α-L-arabinofuranosidases releases L-arabinose residues. This natural sweetener can be used as a food flavor and a source of pharmaceutical products (Fehér 2018). Feeding of L-arabinose and sucrose to rats showed

reduced insulin level in blood, as L-arabinose noncompetitively inhibits intestinal sucrose, resulting in slowing down of the glycemic response (Kaneko et al. 1998; Kotake et al. 2016). The same effect was observed in human and so L-arabinose has the potential to be used in diabetes treatments (Kaats et al. 2011). In addition, α-L-arabinofuranosidase can play a role in the production of bio-ethanol. Traditionally, hexoses were one of the main substrates for the production of bioethanol, however, with the discovery of microorganisms that can ferment pentoses, there has been increasing interest in α-L-arabinofuranosidase (Das et al. 2012).

Other side chains in lignocellulose, like the 4-*O*-methyl-D-glucuronic acid (MeGlcA) found in xylose, can prevent enzymatic hydrolysis of xylan and can be covalently cross-linked to lignin (C. C. Lee, Kibblewhite, Wagschal, Li, Robertson, et al. 2012). The α-glucuronidase from the glycoside hydrolase family 67 acts on such xylooligomers to release MeGlcA. This family comprises only two types of enzymes; α-glucuronidase and xylan α-1, 2-glucuronidase. Both are inverting enzymes and fold into a characteristic $(\beta/\alpha)_8$ barrel domain structure (Nurizzo, Nagy, et al. 2002). α-glucuronidase is currently applied in bio-bleaching of paper pulp, fermentation for animal feed and to remove MeGlcA after alkaline pretreatment of plant cell wall materials for bio-ethanol production (Septiningrum et al. 2015; Rhee et al. 2017; C. C. Lee, Kibblewhite, Wagschal, Li, and Orts 2012).

Metagenomics approaches supported by bioinformatics and subsequent biotechnology may help to recover and characterize novel genes from interesting ecosystems, where most of the species may be uncultivable. In this paper, we describe the identification of a novel α-glucuronidase (PGluc67) and an α-L-arabinofuranosidase (LAraf43). The genes encoding these enzymes were identified in the metagenomes from two gut microbiomes from termites and goat, respectively and cloned. Subsequently, they were expressed in *Escherichia coli* and the purified enzymes were biochemically characterized.

5.3 Materials and Methods

5.3.1 Selection of hemicellulose-degrading enzymes

Previously described metagenome assemblies from the *Coptotermes* termite gut metagenome (Do et al. 2014) and native Vietnamese goat rumens metagenome (Do, Le, et al. 2018; Do, Dao, et al. 2018) were used for the analysis. The bacterial open reading frames (ORFs) from these metagenomes were analyzed *in silico* as follows. A hidden Markov model (HMM)

database for the Carbohydrate Activity enZymes (CAZy) was obtained from dbCAN (http://csbl.bmb.uga.edu/dbCAN/) (Yin et al. 2012). Protein sequences were analyzed using HMMER 3.0 (Eddy 1998) to identify CAZymes candidates. The nucleotide and protein sequences of these candidates were analyzed using basic local alignment search tool (BLAST) software from the National Center for Biotechnology Information (NCBI) to establish sequence homologies (Altschul et al. 1997). For predicting the 3 dimensional (3D) structures model, the protein sequence was analyzed using the Phyre2 web server (http://www.sbg.bio.ic.ac.uk/phyre2/html) (L. A. Kelley et al. 2015) and SWISS-MODEL (https://swissmodel.expasy.org) (Waterhouse et al. 2018) with default settings. The nucleotide sequences were analyzed for the presence of signal peptides using gram negative and gram positive settings using the SignalP 4.1 server (http://www.cbs.dtu.dk/services/SignalP) (Almagro Armenteros et al. 2019). Bacterial promoters were identified using BPROM (http://www.softberry.com). Contig analysis, protein molecular weight and isoelectric point (pI) value calculations were done using Cloning Manager 9.0 (Sci-Ed Software, USA).

5.3.2 Plasmid construction for recombinant expression

A metagenomic DNA library of *Coptotermes* termite guts (Do et al. 2014) was amplified using the REPLIg kit (Qiagen, Germany) prior to be used as a template to PCR the gene sequences. Oligonucleotide primers were designed based on the predicted α-L-arabinofuranosidase (LAraf43) gene from the *Coptotermes* termite gut metagenome. The gene encoding LAraf43 was amplified using the 5'-primer (5'-GGG*CATATG*AGCAATTATACTGCACC-3'), which included the ATG translational start codon inside a *NdeI* restriction site (shown in italic) and 20 nucleotides of the ORF. The 3'-primer (5'-TTT*CTCGAG*CTATTGAATAGTAAATTTCTGAGGTT-3') included a stop codon (TAG), containing an *XhoI* restriction site and the preceding 26 nucleotides of the ORF. Three guanine and thymine residues were added at the 5'-end of the 5'-primer and 3'-primer, respectively, to create a good binding site for the respective restriction enzymes. The gene sequence was amplified using *Taq* polymerase and the product was purified on a 1% agarose gel. It was digested with *NdeI* and *XhoI* and ligated into *NdeI/XhoI*-digested pET16b vector, resulting in the plasmid pET16-LAraf43 with an N-terminal His-tag. The resulting plasmid was transformed into XL1-blue chemically competent cells. Successfully

transformed colonies were screened by restriction digestion and correct inserts were confirmed by DNA sequencing (Macrogen). After quality control, intact pET16-LAraf43 plasmid DNA was transformed in *E. coli* protein expression strain Rosetta2 (DE3) (Novagen).

The predicted sequence encoding an α-glucuronidase (PGluc67) was selected from the available goat rumen metagenome data (Do, Le, et al. 2018). The identified PGluc67 was also predicted to include a signal peptide. Codon usage of the ORF was optimized for enhanced expression in *E. coli*. The resulting ORF was chemically synthesized in such a way that the predicted signal peptide was omitted and replaced by the endogenous *pelB* signal sequence from pET22b resulting in the pET22b-pelB-PGluc67 vector (GenScript, USA). This is expected to direct the protein to the periplasm of *E. coli* cells expressing the construct (Singh et al. 2013). The resulting plasmid was transformed into *E. coli* Rosseta2 for protein expression.

5.3.3 Recombinant protein expression and purification of LAraf43

A 10 mL culture was inoculated from a glycerol stock of transformed Rosetta2 and diluted into 200 ml of LB medium with 100 mg/ml ampicillin at 37°C until the culture reached an optical density at 600nm (OD_{600}) of 0.6-0.8. At that point gene expression was induced by adding 50 µM isopropyl-beta-D-thiogalactopyranoside (IPTG) and cultures were further incubated for 2 hours at 37°C. Cells were harvested by centrifugation and suspended in 8 mL of phosphate buffer saline (PBS; pH 7.4). A cocktail of protease inhibitors cOmplete™, EDTA free (Roche) was added followed by two passages at ~1.7 k psi through a OneShot cell disruptor (Constant Systems Ltd) at room temperature. The cell extract was centrifuged at 586 × g for 10 min and 100,000 × g for 1 hour to remove debris and membrane fragments. The cleared cell extract was mixed with TALON Superflow resin (GE, Sweden), which had been pre-equilibrated with buffer A (50 mM potassium phosphate buffer, 500 mM sodium chloride, 10% glycerol, 10 mM Imidazole pH 7.5) and incubated at 4°C, for 1 hour. The mixture was transferred to a disposable 5 ml polypropylene column (Thermo Scientific) and then washed with 10 mL of buffer A. The His-tagged proteins were eluted from the beads by adding 10 mL of Buffer B (50 mM sodium phosphate buffer, 500 mM sodium chloride, 10% glycerol, 400 mM imidazole pH 7.5). Subsequently, a Vivaspin 20, MWCO 10 kDa column was used to concentrate the sample and remove salts. A volume of 20 mL PBS pH 7.4 was

added and centrifuged at 6000 x g and this was repeated five times after which the retentate was collected and aliquoted. The concentration of purified protein was measured using the BCA protein assay kit (Thermo Scientific) with bovine serum albumin (BSA) as the standard. Crude extracts or purified protein samples were denatured in sample buffer with dithiothreitol (DTT), boiled for 10 min and applied to 10% gradient sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE, BIORAD) along with the molecular weight marker to determine the molecular weight and purity. The gel was stained with 0.1 % Coomassie Blue as previously described by (Lämmli 1970).

5.3.4 Enzyme assays for LAraf43

Synthetic p-nitrophenyl-α-L-arabinofuranoside (pNP-α-L-Araf) was purchased from Megazyme International (Wicklow, Ireland). Alpha-L-arabinofuranosidases catalyze the release of *p*-nitrophenol (pNP) from pNP-α-L-Araf, which can be measured at 405 nm. Each assay mixture contained 10 µL of a 25 mM pNP-α-L-Araf solution with 88 µL of PBS buffer (pH 7.4) and 2 µL of enzyme solution. The reaction was carried out at 37°C for 6 hours measuring pNP every 15 minutes. A standard curve of pNP was generated to estimate the amount of pNP released from the reaction. One unit of enzyme activity was defined as the amount of enzyme releasing 1 µmol of PNP from PNP-α-LAraf per min under these conditions. The assay activity was performed in triplicate as mentioned above, unless otherwise stated.

The initial rate of hydrolysis to determine kinetic parameters were obtained from different pNP-α-L-Araf concentrations in the range 0 – 0.5 mM at pH 7.4 in PBS and at 37°C. The reaction was started by adding 2 µL LAraf43 to the reaction in a final volume of 100 µL. Initial rates were plotted against pNP-α-L-Araf concentrations and kinetic parameters were estimated by fitting the Michaelis-Menten equation, linearized by reciprocal transformations.

5.3.5 Enzyme assays for PGluc67

Alpha-glucuronidase catalyzes the conversion of aldotriouronic acid (Megazyme, Ireland) to 4-*O*-methyl-α-D-glucuronic acid, which is subject to oxidization by urinate dehydrogenase (UDH) to form glucarate coupled to reduction of NAD to NADH (Yoon et al. 2009). The activity of α-glucuronidase was determined by measuring NADH at 340 nm. The reaction was conducted in 50 mM PBS buffer pH 7.4, containing 0.375 to 6 mg/mL, stop buffer,

NAD^+ and urinate dehydrogenase (UDH). The Stop buffer, NAD^+ and UDH were used as provided by the manufacturer. Enzymes were incubated with each reagent for 5 min at 37°C followed by assaying the residual activity.

The kinetic activity of PGluc67 was assayed at 37°C after 0, 5, 10, 15 and 30 min. Kinetic parameters of the purified enzyme were estimated in duplicate and repeated in three separate days. V_{max} and K_m of the Michaelis-Menten model were estimated by linear regression after reciprocal transformation.

5.3.6 Effects of pH and temperature on PGluc67 activity and stability

Reactions were conducted at various pHs and temperatures. The effect of pH on the activity and stability of PGluc67 were determined in a series of different buffers with 0.5 mg/ml BSA: 100 mM sodium acetate (pH 3 - 5); 100 mM MES (pH 6); 100 mM MOPS (pH 7) and 100 mM HEPES (pH 8 - 9). The activity of PGluc67 was assayed as described above.

The effect of temperature on the activity and stability of PGluc67 was determined in Polymerase chain reaction machines at temperatures ranging from 10 to 70°C. Temperature stability was determined by incubating the enzyme for 30 min at various temperatures from 30 up to 70°C, followed by activity assaying as previously described.

5.3.7 Effects of chemical agents and metal cations

The effect of several metal ions and chemical agents on PGluc67 activity was determined. The Mn^{2+}, Ca^{2+}, Mg^+, K^+, Ni^{2+}, Zn^{2+} and Fe^{3+} metals ions were assayed at concentration of 10 mM in the reaction mixture at pH 7 in triplicate. Chemical agents such as urea, triton X-100 (Sigma-Aldrich, USA), and β-mercaptoethanol (MERCK, USA) were tested at 1 µM. Imidazole and tween-80 were tested at a concentration of 10 mM. The activity was determined as described above and presented as a percentage in comparison to the activity without the test compound. The reaction was carried out in triplicate.

5.3.8 Substrate specificity

The synthetic compounds 4-nitrophenyl-β-glucoside (pNPG) and 4-nitrophenol-β-D-xylopyranoside (pNPX) were tested with 400 µL enzyme, 100 µL substrate and 10 mM PBX 5x, pH 6.0 and incubated at 37°C for 1 hour. The reaction was stopped by adding 1M Na_2CO_3. Measurements were carried out at 405 nm.

5.3.9 Extending the contig containing the PGluc67-encoding operon

The original sequence data used to identify the PGluc67's ORF was assembled using a single k-mer value (Do, Le, et al. 2018). To obtain a better overview of the genomic region, we reprocessed the raw dataset to extend the contig length using multiple k-mer values. Low quality raw reads from the sequencer were removed using the programme bbduk with the following options: ktrim=r k=23 mink=7 hdist=1 tpe tbo qtrim=rl trimq=20 ftm=5 maq=20 minlen=36 (Bushnell 2017). A contamination library containing fungal, human, plant, protozoal and viral sequences was generated. Raw reads were aligned against this library using Kraken2 (Wood and Salzberg 2014). Any reads that matched to the contamination library were removed from the metagenome. For the assembly, MetaSPAdes version 3.13.0 with k-mer values 21, 33, 55, 77, 99 was used (Bankevich et al. 2012). Contig annotation tool (CAT) was used to remove contamination and uncharacterized contigs from the assembly (von Meijenfeldt et al. 2019). The resulting contigs were made into a nucleotide database using makeblastdb (Camacho et al. 2009). The original contig with the ORF encoding PGluc67 was aligned again this database and yielded a hit Contig with 100% identity and coverage that was used for further research.

5.4 Results

5.4.1 Sequence analysis of LAraf43

An ORF encoding LAraf43 was identified in the sequences obtained from a previously described gut metagenome of the *Coptotermes* termite (Do et al. 2014). The ORF was identified on scaffold4611_1 of 9,042 bp long. This contig was 99.12% identical to the corresponding region in the genome of *Lactococcus lactis* strain A106 (NCBI accession CP009472.1), 98.01% identical to that of *L. lactis* strain NCDO2118 (NCBI accession CP009054.1) and 98.01% identical to that of *L. lactis* strain 147 (NCBI accession CP001834.1) and thus apparently derived from a *L. lactis* species present in the termite gut. Strikingly, these three sequenced genomes of *L. lactis* strains were all isolated from plant sources and not from dairy. They all have an identical gene synteny that differs from what found in dairy related *L. lactis* isolates but similar in scaffold4611_derived from the termite gut (Passerini et al. 2013; Siezen et al. 2011). Specifically, these isolates contained a full-length functional ORF encoding an α-L-arabinofuranosidase with an average of 98% amino

acid identity with LAraf43 and, additionally, an ORF encoding *araT* encoding an arabinose-proton symporter, which was not found in genome sequences of dairy-derived *L. lactis* (Passerini et al. 2013; Siezen et al. 2011). However, the *araT* ORF in the scaffold4611_1 appeared truncated due to a single nucleotide deletion. We cannot exclude this derived from a sequencing error, but the remaining 3'-part of ORF encodes for putative transporter from the major facilitator superfamily (MFS) and could function as such in its truncated form. The BLASTP result revealed a MFS domain and the 3D structure was predicted to be a protein transporter (data not shown). General L-arabinose processing genes found in all *L. lactis* genomes were also detected on scaffold4611_1 and the three sequenced genomes. An L-arabinose operon (or *araBAD*) was observed upstream from LAraf43. Downstream of LAraf43, two ORFs were identified coding for a MFS transporter (*araP*), annotated to encode a disaccharide transporter and a GntR family transcriptional regulator (*araR*), respectively (Fig. 1).

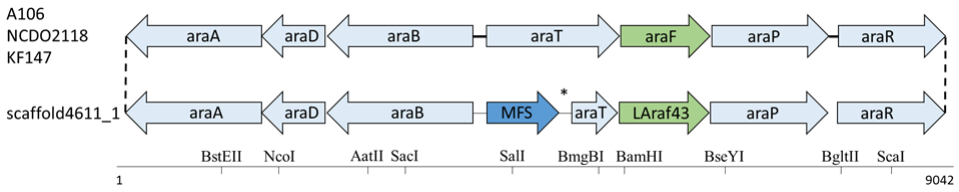


Figure 1: *Gene organization of the conserved region containing the α-arabinofuranosidase gene (green) in three sequenced L. lactis genomes (above) aligned to scaffold4611_1 of the metagenome of termite gut (below). The gene names are: araA, L-arabinose isomerase; araD, L-ribulose-5-phosphate 4-epimerase; araB, L-ribulokinase; araT, arabinose-proton symporter; araF, α-N-arabinofuranosidase; araP, disaccharide permease; araR, GntR family arabinose operon repressor; MFS, major facilitator superfamily membrane transporter (blue arrow). The asterisk showed the missing nucleotide on the scaffold4611_1.*

L. lactis is a species that belongs to the Firmicute phylum and is often found in insect gut microbiomes (Kaoutari et al. 2013; Do et al. 2014; Shannon et al. 2001; Shelomi et al. 2015). Addition of arabinoxylan oligosaccharides into the host dietary results in the expansion of *L. lactis* population (Geraylou et al. 2013). It has been reported that some *L. lactis* strains can grow using L-arabinose as carbon source (Golomb and Marco 2015; Passerini et al. 2013) and the contig-sequence shows synteny with plant-derived *L. lactis* genome regions, suggesting a similarity in metabolic capacities.

```
                                        *
L. brevis APU52333.1                12  IVQRADPYIYKHTDGYYYFTASVPAYNLIEIRRAKTLNGLANAAPRTIWRKHPDGSGAMS  71
W. cibaria/confusa APU52332.1       10  IIQRADPYIYKHTDGYYYFVASVPAYNLIELRRAKTIDGLAHAMPRTIWRKHDSGTGAQS  69
LlAraf43                             9  IIQRADPFIYKHTDGYYYFTASVPAYNLIELRRSKTIEGLAFAMPRTIWRKHDSGTGAQS  68
L. lactis WP 058219862.1             9  IIQRADPFIYKHTDGYYYFTASVPAYNLIELRRSKTIEGLAFAMPRTIWRKHDSGTGAQS  68
S. chartreusis BAA90772.1           55  VRQRADPHIHRHTDGRYYFTATAPEYDRIVLRRSRTLGGLSTAAESVIWRAHPTG--DMA  112
S. avermitilis BAC68753.1           42  AEKRADPHIFKHTDGYYYFTATVPEYDRIVLRRATTLQGLATAPETTIWTKHASG--VMG  99
                                            D       D

L. brevis APU52333.1                72  QLIWAPELHYIDGKWFIYFAASHTKEFDHNGMFQHRMYCIECDNPDPMRDEADWTEHGQI  131
W. cibaria/confusa APU52332.1       70  ELIWAPELHYTDGKWYVYYAASHTTAFDENGMFQHRMFAIECDAEDPMETEENWVEKGQI  129
LlAraf43                            69  ELIWAPEIHFIRGKWYVYYAASHTTEFDKNGMFQHRMFCIECENNNPMESEDNWVEKGQI  128
L. lactis WP 058219862.1            69  ELIWAPEIHFIRGKWYVYYAASHTTEFDKNGMFQHRMFCIECENINPMKSEDNWVEKGQI  128
S. chartreusis BAA90772.1          113  AHIWAPELHRIGGKWYVYFAAAPAE-----DVWRIRIWVLENSHPDPFK--GTWEEKGQV  165
S. avermitilis BAC68753.1          100  AHIWAPEIHFIDGKWYVYFAAGSTS-----DVWAIRMYVLESGAANPLT--GSWTEKGQI  152
                                           E                                E             E
                                              *
L. brevis APU52333.1               132  ETPLDTFALDATVFEAQKKLYYVWAQKDPAIKGNSNIYIAEMANPWTLKTKPVMLTKPEY  191
W. cibaria/confusa APU52332.1      130  ETHLDSFALDATSFELNDKLYYVWAQKDPEIKGNSNLYIAEMENPWTLKTAPVMLSKPEF  189
LlAraf43                           129  LTPMDSFALDATSLQLNGKLYYIWAQKDPNIRGNSNIYIAEMENPWTLKTKPILLSKPEY  188
L. lactis WP 058219862.1           129  LTPMDSFALDATSLQLNGKLYYIWAQKDPNIRGNSNIYIAEMENPWTLKTKPILLSKPEY  188
S. chartreusis BAA90772.1          166  RTAWETFSLDATTFTHRGARYLCWAQHEPGADNNTGLFLSEMANPWTLTGPQIRLSTPEY  225
S. avermitilis BAC68753.1          153  ATPVSSFSLDATTFVVNGVRHLAWAQRNPAEDNNTSLFIAKMANPWTISGTPTEISQPTL  212
                                              D
                                               *
L. brevis APU52333.1               192  DWETKIFWVNEGPAVLHRNGRFFLTYSASATDENYAMGMLTVAEDADLLDPTSWSKSETP  251
W. cibaria/confusa APU52332.1      190  DWETKIFWVNEGPAILKRNGKVFLTFSGSATDENYAMGMLWIEDDKDVLDAANWHKLDHP  249
LlAraf43                           189  DWETKIFWVNEGPAVLQRNGKLFLTYSASATDENYCMGMLTADENSNILDPKAWKKSSQP  248
L. lactis WP 058219862.1           189  DWETKIFWVNEGPAVLQRNGKLFLTYSASATDENYCMGMLTADENSNILDPKAWKKSSQP  248
S. chartreusis BAA90772.1          226  DWECVGYKVNEGPYALKRNGRIFLTYSASATDHHYCVGMFTADAGGNLMDPGNWSKSPIP  285
S. avermitilis BAC68753.1          213  SWETVGYKVNEGPAVIQHGGKVFLTYSASATDANYCLGMLSASASADLLNAASWTKSSQP  272
                                          E       E                   D
```

Figure 2: *Alignment of amino acid sequences of LAraf43 with α-L-arabinofuranosidase from Weissella (APU52332.1), the uncharacterized Lactococcus (WP_058219862.1), Lactobacillus (APU52333.1), exo-1,5-α-Larabinofuranosidase from Streptomyces avermitilis (BAC68753.1), and that of Streptomyces chartreusis (BAA90772.1). The alignment was restriced to the catalytic domain as given in CAZy and was generated using ClustalW. The asterisks show the putative catalytic residues within the family, with the amino acid residues (D14, D138 and E199 in LAraf43) highlighted in blue.*

The LAraf43 gene is 972 bp long and encodes a polypeptide of 324 amino acid residues with a molecular mass of 37.57 kDa and pI of 5.59. BLASTP results showed that LAraf43 contains a conserved amino acid motif (position 4 – 312) that shows high sequence homology to the Glycoside Hydrolase 43 (GH43) family (Fig. 2). When compared with known and characterized proteins in the GH43 family, the LAraf43 clusters into subgroup 26 (data not shown). Proteins in this subgroup often do not contain a carbohydrate-binding module (CBM) (Mewis et al. 2016). Further analysis indicated that the encoded protein did not include an N-terminal signal peptide, which renders the protein cytoplasmic.

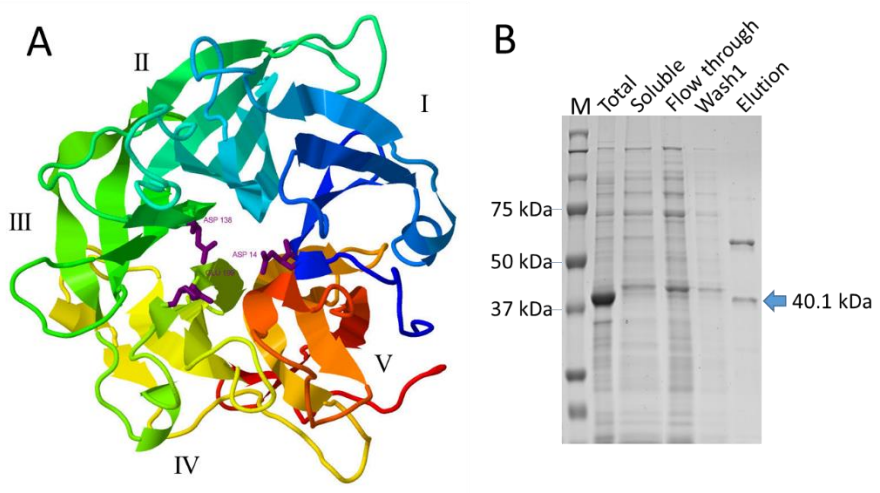


Figure 3 A: *Model of the 3D structure of LAraf43 generated by the Phyre2 website. The image was generated using Pymol (64).* ***B:*** *SDS-PAGE of samples obtained during purifiucation of LAraf43 enzyme. Lane M, molecular weight maker; Total, total cell sample; Soluble, Soluble fraction after ultra-centrifugation; Flow through, Flow through from the Nickel column, Wash1, First wash; Elution, purified α-L-arabinofuranosidase enzyme; two band are detected, one of them corresponds in size the His-tagged protein purified.*

Sequence analysis of LAraf43 showed that a protein with a predicted α-N-arabinofuranosidase activity from *L. lactis* was its closes orthologue with 99.38% identity (NCBI accession WP_058219862.1). No α-arabinofuranosidase gene from *L. lactis* has been functionally characterized. Similar enzymes that have been characterized are APU52332.1 from *Weissella*, AGT14430.1 from *L. brevis* DSM 20054, APU52333.1 from *L. brevis* DSMZ 1269, BAC68753.1 from *S. avermitilis* and BAA90772.1 from *S. chartreusis* GS901 with 75.86 %, 72.06 %, 72.06 %, 53.65 % and 49.31 % identity to LAraf43, respectively (Linares-Pastén et al. 2017; Michlmayr et al. 2013; Matsuo et al. 2000; Ichinose et al. 2008). Amino acid sequence of LAraf43 had higher similarity to dimer and tetramer proteins (Table 1). The solved structure of the exo-α-arabinofuranosidase protein from *L. brevis* DSMZ 1269 served as model to predict the 3D structure of LAraf43 (Fig. 3A) (Linares-Pastén et al. 2017; Fujimoto et al. 2010). The resulting model showed the characteristic five-bladed β-propeller domain of the GH43 family. The α-arabinofuranosidase enzymes from the GH43 family are further characterized by an active site comprised of three conserved amino acid residues, two aspartic acids and one glutamic acid (Pason et al. 2015), which all three are present in

LAraf43 at positions (D14, D138 and E199) that match with the known enzymes (Table 1; Fig. 2) (Michlmayr et al. 2013; Matsuo et al. 2000; Ichinose et al. 2008). This conserved motif is a general feature of enzymes that have an inverting function as their catalytic activity (Lagaert et al. 2014; Sweeney and Xu 2012). As noticed by Linares-Pasten *et al.* 2017, α-arabinofuranosidases with a long loop region within the blade V fold show exo-enzymatic activity. This region of the predicted model of LAraf43 was identical to SaAraf43A (data not shown) (Linares-Pastén et al. 2017). Exo enzymes can cleave both ends of a long-chain substrate instead of cleaving it in the middle. LAraf43 was further predicted to be a homodimer with exo-1, 5-α-L-arabinofuranosidase activity based upon sequence similarity to structurally of characterized enzymes.

5.4.2 Biochemical characterization of LAraf43

The LAraf43 gene was cloned from the *Coptotermes* termite gut metagenomes by PCR and inserted into a pET16b expression vector in such a way that a His-tag was fused to the N-terminus of the encoded protein. The plasmid was successfully transformed in *E. coli* Rosetta2. Upon induction of LAraf43 expression, high protein levels were detected in cell samples on coomassie brilliant blue-stained SDS-PAGE. However, when the cells were lysed and subsequent steps were performed to purify the His-tagged proteins the amount of soluble protein obtained was very low. Apparently, the majority of proteins that were expressed ended up in aggregates. Furthermore, the purified recombinant protein fraction showed two bands on the SDS-PAGE gel running at ~41 and ~56 kDa, respectively. The molecular weight of the histidyl-tagged N-terminus was calculated to be 40.1 kDa, which corresponds to the lower band. The extra band at 56 kDa could be a background protein that co-eluted in the purification process or an aggregation product of LAraf43 (Sagné et al. 1996) (Fig. 3B). LAraf43 was found to have enzymatic activity despite the presence of the His-tag, suggesting that the tag did not interfere with the protein structure and function (Carson et al. 2007). The purified protein also appeared stable, since it did not show loss of activity nor proteolysis during tests after 1 month of storage at 4°C.

The specific activity of the enzyme was determined based on the rate of pNP release as described in the Method section. Kinetic parameters estimated from the initial reaction rates were K_m of 2.70 ± 0.01 mM and a V_{max} of 0.08 ± 3.0 e-4 U/mg at 37°C and pH 7.4.

5.4.3 Sequence analysis of PGluc67

We identified a 10,772-bp contig from the metagenome of Vietnamese goat rumen that contained an ORF encoding a glycoside hydrolase of family 67 (GH67), which we named PGluc67. After reassembly of our raw sequencing data using multiple k-mer values, the resulting extended contig, called NODE_717, was 21,956 bp long and contained the original operon with 100% identity. BLASTN analysis of the complete contig showed that only small fragments aligned to sequences in the NCBI nucleotide database, suggesting low similarity to previously released DNA fragments. The highest similarity of 83.56 % nucleotide identity over a segment of 1,820 nucleotides was observed when the fragment was aligned to a 5,463 bp fragment of an uncultured bacterial clone obtained from a metagenome of a cow (NCBI accession JN684207.1) (C. C. Lee, Kibblewhite, Wagschal, Li, and Orts 2012). This segment included a predicted ORF encoding an α-glucuronidase. Importantly, however, the overall gene organization on the cow rumen contig is very different from that of NODE_717.

Sequence analysis of ORFs on NODE_717 showed that it encodes a sensor histidine kinase (*baeS*) downstream of PGluc67, which has been observed for other glycoside hydrolases as well (Rhee et al. 2017; Lingling Wang et al. 2013). Upstream of the ORF encoding PGluc67 lies a gene cluster encoding a polysaccharide outer membrane exporter *(wzA)*, a regulator for the chain length of O-antigen polysaccharides *(wzZ)*, an oligosaccharide repeats unit polymerase *(wzY)*, a transmembrane lipid transporter protein (flippase, *wzX*), two glycosyl transferases family 2 (*GT2*), a TDP-4-oxo6-deoxy-D-glucose transaminase (*wecE*), two *wcaJ*, UDP-glucose:undecaprenyl-phosphate glucose-1-phosphate transferase and the *wzI*, surface assembly of capsule (Fig. 4A). All genes are predicted to be involved in production of exopolysaccharides in Gram-negative bacteria (Fig. 4B) (Marolda et al. 2010; 2006; Vinés et al. 2005; Reid and Whitfield 2005; Furlong and Furlong 2013; Valvano, Furlong, and Patel 2011; Schmid, Sieber, and Rehm 2015). Further upstream are lipocalin and endonuclease proteins. This gene cluster is also linked with the response to envelope stress (Campanacci et al. 2006; 2004). Homology searches of these proteins showed their top hit were proteins from *Prevotella* species with 42.31% to 92.86% identity at the amino acid level. Similar result was obtained after CAT analysis (data not shown). This is a very common genus of *Bacteroidetes* that is also commonly found in animal guts (Lim et al. 2013; Flint and Bayer 2008;

Henderson et al. 2013), and suggested to be involved in colanic acid (CA) production (Dodd et al. 2010; Roberts and Whitfield 1999; Corbett and Roberts 2008).

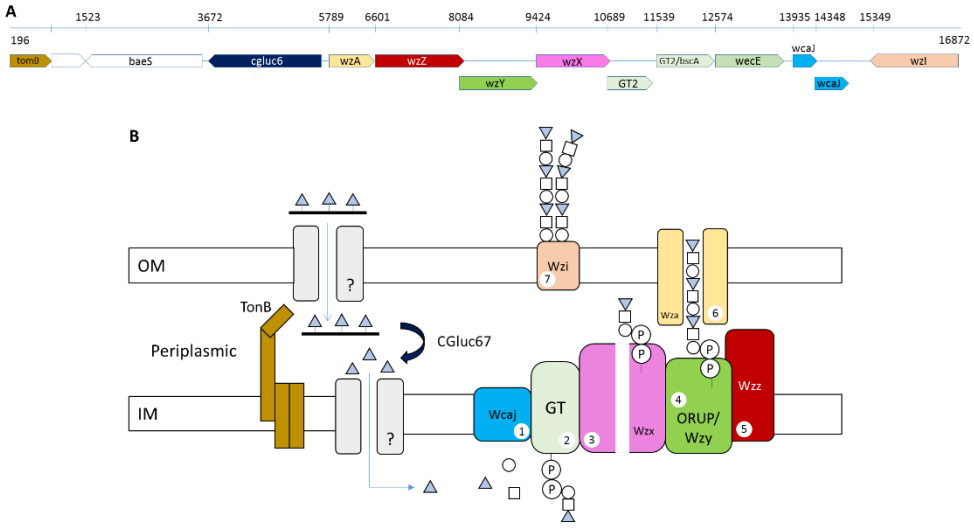


Figure 4 A: *Gene organization on NODE_717 from goat rumen metagenome containing the α-glucuronidaseencoding gene PGluc67. The other ORFs are baeS, encoding a sensor histidine kinase; wzA, polysaccharide outer membrane exporter; wzZ, chain length determinant protein; wzX, flippase, transmembrane lipid transporter; Wzy, oligosaccharide repeat unit polymerase; GT2, glycosyl transferase family 2; bcsA, bacterial cytoplasmic membrane cellulose synthase subunit A; wecE, TDP-4-oxo- 6-deoxy-D-glucose transaminase; wcaJ, UDP-Glc:Und-P Glc-1-P transferase and wzi, Outer membrane protein.* ***B:*** *predicting proteins involved in the CA production. PGluc67 cleaves to generate glucuronic acid as precursor for colanic acid. 1: The WcaJ protein initiates the capsule synthesis. 2: GTs add sugar to the exopolysaccharides. 3: The exopolysaccharide is flipped by Wzx. 4: Longer chain of exoploysaccharide is polymerized by Wzy. 5: The whole process is controlled by Wzz. 6: Capsule is transported to the outter membrane through Wza protein. 7: Exopolysaccharides are attached to the outer membrane surface with Wzi.*

The ORF of PGluc67 is 1,974 bp long and appears to encode a protein with a signal peptide targeting the Sec machinery for export out of the cytoplasm. It suggests the protein is located in the periplasm if the contig is derived from a Gram-negative bacterium, as the other contig ORFs suggested (Tsirigotaki et al. 2017; Dalbey, Wang, and van Dijl 2012). BLASTP indicated homology to the conserved motif of the GH67 family from amino acid 56 to 350 (data not shown). The PGluc67 protein without signal peptide was highly homologous to

several characterized α-glucuronidases enzymes AFE48530.1 from uncultured cow rumen bacterium, ADI70674.1 from *Prevotella bryantii* B14, ACE83468.1 from *Cellvibrio japonicus* Ueda107, AFJ94648.1 from uncultured compost bacterium, AAG09715.1 from *Geobacillus stearothermophilus* 236 and AGL48978.1 from *Thermotoga maritima* MSB8 with 87%, 53.6%, 50.4%, 44.6%, 43.12% and 42.3% similarity on amino acid level, respectively (C. C. Lee, Kibblewhite, Wagschal, Li, and Orts 2012; Dodd et al. 2010; Ruile, Winterhalter, and Liebl 1997; C. C. Lee, Kibblewhite, Wagschal, Li, Robertson, et al. 2012; I.-D. Choi, Kim, and Choi 2005; Nurizzo, Nagy, et al. 2002). Only α-glucuronidase proteins from *Prevotella* and *Bacteroides* contain a signal peptide. Crystal structures of *C. japonicus* and *G. stearothermophilus* were used for predicting the structure of PGluc67 (Nurizzo, Nagy, et al. 2002; Golan et al. 2004). The α-D-glucuronidase protein sequence contains two highly conserved amino acids (D332 and E360 in PGluc67), which act as critical catalytic residues for the inverting mechanism (Fig. 5A) (Zaide et al. 2001; Nurizzo, Nagy, et al. 2002). Similarly to other inverting group of enzyme, E360 acts ac catalytic acid and D332 acts as catalytic base (Fig. 5B) (Cuskin et al. 2015). PGluc67 was predicted to be homodimer with Calcium and Zinc binding site (data not shown).

The α-D-glucuronidase signal peptide was replaced by a PelB signal peptide and codon usage was optimized for expression in *E. coli*. When expressed from a pET22 plasmid the recombinant PGluc67 encoded a 667 amino acid sequence with *pelB* signal and a His-tag at the C-terminus and it was predicted to have 66 kD and a pI of 6.55. The *pelB* leader sequence improves solubility and transfers the protein to the periplasm (Singh et al. 2013; Freudl 2018; J. H. Choi and Lee 2004).

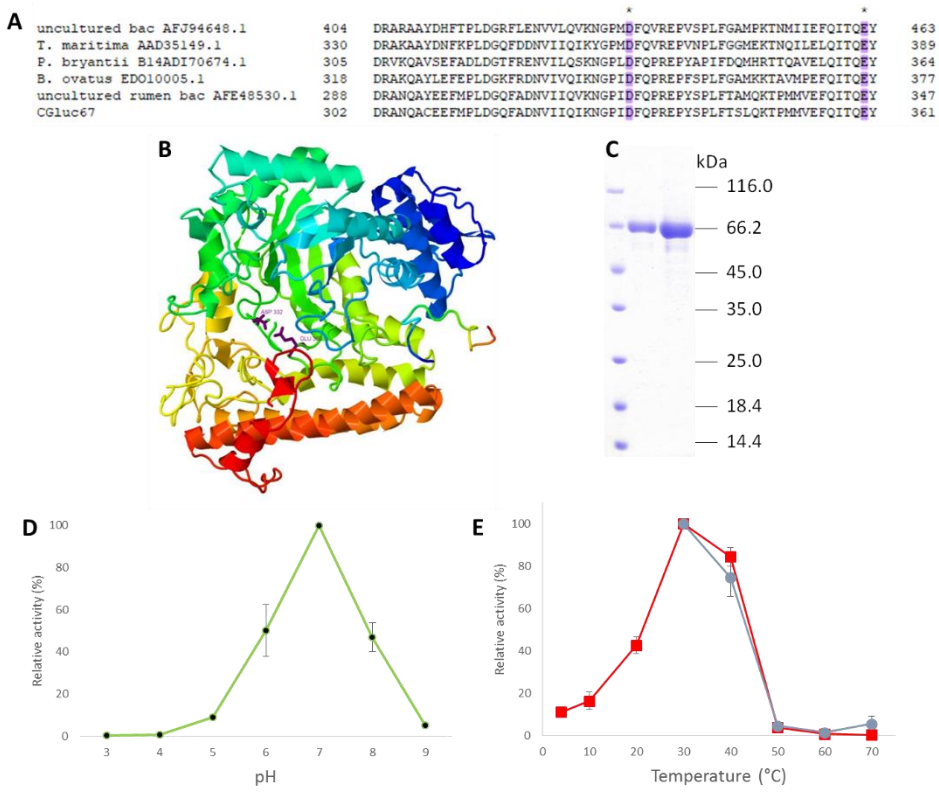


Figure 5 A: *Alignment of PGluc67 with other characterized α-glucuronidase from uncultured bacterium (AFJ94648.1 and AFE48530.1), Thermotoga maritima (AAD35149.1), Prevotella bryantii (B14ADI70674.1) and Bacteroides ovatus (EDO10005.1). The asterisks show amino acid residues that constitute the active site within the catalytic domain, which are highlighted in blue. The alignment was done using ClustalW.* ***B:*** *3D model of PGluc67 generated by Phyre2, with the active site shown by purple side chains.* ***C:*** *SDS-PAGE of purified samples of PGluc67. From left to right: molecular weight marker, and two concentrations of the sample.* ***D:*** *Effect of pH on PGluc67 activity.* ***E:*** *Effect of temperature (red squares) and heat stability (blue circles) at different temperatures. The activity at the optimal temperature was defined as 100%.*

5.4.4 Biochemical characterization of PGluc67

The PGluc67 was overexpressed and purified from *E. coli* Rosetta2 yielding 2.7 mg/mL of PGluc67 protein. The recombinant protein migrated between the 50 kDa and 68 kDA markers as predicted (Fig. 5C).

The purified protein was stored at 4°C for testing and did not show loss of activity nor proteolysis after 1 month of storage. As expected, PGluc67 did not show any β-glucosidase and β-xylosidase activities (data not shown). Activity measurements show hydrolase activity toward aldotriouronic acid with β-(1,4)-D-xylo-oligosaccharides and Dglucuronic acids as substrates, conducted according to the two-step essay described in the Methods section, revealed that PGluc67 has this activity, and it is optimal at pH 6-8 (Fig 5D). The optimal temperature for PGluc67 was 30°C.

Table 2: *Average effect on enzyme activity of PGluc67 at several concentrations of metals and chemical agents.*

Metal ions/ Chemicals	Concentration	Relative activity (%)
Control	-	100
Mn^{2+}	10 mM	116±5.3
Ca^{2+}	10 mM	106±6.7
Mg^{2+}	10 mM	98±19.7
K^{+}	10 mM	86±12.1
Ni^{2+}	10 mM	31±12.2
Zn^{2+}	10 mM	0.1±2.9
Fe^{3+}	10 mM	0
Urea	1 µM	107±11.4
Triton X-100	1 µM	106±25.4
2-Mercaptoethanol	1 µM	89±11.3
Imidazole	10 mM	81±7.1
Tween	10 mM	0

The relative activity dropped to 43% at 20°C and 84% at 40°C. The activity decreased drastically at 50°C and above. After 30 min incubation at a temperature range from 30°C and 40°C, PGluc67 retained 100% and 75% of its activity respectively (Fig 5E).

The kinetic parameters of PGluc67 towards aldotriouronic acids were calculated based on Michaelis-Menten analysis. Different concentrations of aldotriouronic acid were used to

generate kinetic curves. The kinetic parameters as derived from initial rates for hydrolysis for at 50 mM PBS pH 7.5 were K_m = 3.92 ± 1.76 mM and V_{max} = 55.0 ± 17.6 U/mg.

Since GH67 enzymes have metal ions as co-factors, the effect various metal ions on PGluc67 activity was evaluated. The activity of PGluc67 for aldotriouronic acid increased to 115.7 % for Mn^{+}. The same metal ion showed a similar effect when tested with α-glucuronidase from a mixed culture (C. C. Lee, Kibblewhite, Wagschal, Li, Robertson, et al. 2012) and from *Thermotoga maritima* (Suresh et al. 2002). The activity of PGluc67 was not affected by Ca^{2+}. PGluc67 enzyme activity was, however, adversely affected at 10 mM of Mg^{2+}, K^{+} and Ni^{2+}, where the activity decreased to 97.8 %, 86.3 % and 30.7 % respectively. PGluc67 showed no activity in the presence of 10 mM of Zn^{2+} and Fe^{3+} (Table 2).

The effect of adding different putatively inhibiting reagents was also tested. Urea, Triton X-100, 2-mercaptoethanol, imidazole and tween-80 were added to the reaction mixture of the first step at various concentrations. At 1 µM a small increase of activity was found for urea (107.2 %) and Triton X-100 (106.3 %). Concentrations of 1 µM 2mercaptoethanol and 10 mM imidazole reduced PGluc67 to 89 % and 80.7% respectively. Tween-80 10 mM caused the enzyme to lose its activity completely (Table 2).

5.5 Discussion

The use of lignocellulose from agricultural waste is on the rise. Recycling of biomass provides great benefits to the environment as this carbon source may be used to generate bio-based materials which are presently derived from fossil fuels (Kalia et al. 2017). However, an environmentally friendly way of extracting fermentable carbohydrates from lignocellulose is still a challenge and often requires chemical treatments that cost energy and generate waste. The use of natural enzymes could be an advantage, since these have evolved to become specialized and efficient in breaking down biomass to generate building blocks and carbon sources. In the recent past, metagenomics approaches have been used to identify applicable enzymes from ecosystems that are specialized in biomass processing (Lingling Wang et al. 2013; Hongjie Li et al. 2017a; Joynson et al. 2017; Warnecke et al. 2007). Within the biomass, hemicelluloses are cross-linked with microfibrils and lignins to strengthen the plant cell wall to protect it from chemicals and physical damage. The diversity of hemicelluloses creates random linkages and makes them very difficult to process. To access the carbon-rich cellulose, hemicelluloses need to be removed. On top of that, degradation of hemicellulose

also releases monomers for different chemical applications (X. Liu and Kokare 2016; Asghar et al. 2019). We report here on the use of bioinformatics tools and molecular approaches to identify and characterize two hemicellulases.

The plant specific pentose L-arabinose is one of the abundant sugars in hemicellulose, and it accounts for 5-10% of cell wall sugar in rice (*Oryza sativa*) and Arabidopsis (*Arabidopsis thaliana*) (Kotake et al. 2016; Olofsson, Bertilsson, and Lidén 2008). The enzyme α-L-arabinofuranosidase hydrolyzes arabinoxylan to produce L-arabinose (Ichinose et al. 2008; Linares-Pastén et al. 2017). In bacteria, this enzyme is classified into families GH2, GH3, GH43, GH5, GH54 and GH62, with GH43 being the most abundant (www.cazy.org). An α-L-arabinofuranosidase GH43 called LAraf43 was bioinformatically identified from a gut-derived *L. lactis genome*. LAraf43 is predicted to be a part of L-arabinose processing operon homologous to similar operons found in plant-derived *L. lactis*. This species can survive using L-arabinose as energy source (Golomb and Marco 2015; Passerini et al. 2013). The plant-derived *L. lactis* possess gene clusters specialized in breaking down hemicellulose to generate L-arabinose. The highly active α-arabinofuranosidase from this cluster would also give a gut-derived *L. lactis* a competitive edge over other gut bacteria.

The degradation of arabinoxylan by *L. lactis* is regulated by the araR protein in response to substrate availability. Extracellularly released L-arabinose is transported into the cell via an arabinose-proton symporter, AraT, while disaccharides are taken up via AraP. In our metagenomics assembly, the AraT encoding ORF is truncated due to a single-nucleotide deletion. Although it needs to be investigated whether this deletion is real or a sequencing error, the remaining ORF encodes a protein with the characteristics of an MFS membrane transporter. It is tempting to speculate that this MFS protein can transport arabinoxylan oligosacharides into the cell (Tauzin et al. 2016). This would be required for their degradation, since LAraf43 lacks a signal peptide and is not expected to be active externally. It seems that the catalytic function and the transport function, which are combined in one large protein in AraT of plant-derived *L. lactis*, are encoded by two separate ORFS in our metagenome. However, we are not sure whether the second ORF is expressed at all, since it appears to lack a promoter and a functional ribosome binding site. In any case, imported oligosaccharides could be hydrolyzed by LArafa43 to L-arabinose and then are expected to be further processed by *araBAD*-encoded proteins to produce D-xylulose-5-phosphate (Passerini et al. 2013; Gírio et al. 2010; Kuge and Teramoto 2015).

Comparative sequence analysis and 3D modeling showed that LAraf43 has all activity requirements. Its catalytic domain includes an aspartic acid residue at position 138 which acts as pKa modulator of the catalytic glutamic acid residue at position 199 and also ensures the correct orientation of the substrate. A water molecule oriented by D138 is activated by aspartic acid at position 14 to allow nucleophile attack on the anomeric carbon in the substrate. In addition the catalytic acid E199 donates a proton to the anomeric carbon resulting in breaking the glycosidic bond while inverting the anomeric configuration (Linares-Pastén et al. 2017; Maehara et al. 2014; Till et al. 2014). Similar to other α-L-arabinofuranosidase from GH43 sub 26, Laraf43 is likely to be an exo-α-1,5-L-arabinofuranosidase.

The LAraf43 gene was cloned from a termite gut metagenome. Previously, Margolles and De los Reyes-Gavilán (2003) described the expression of an α-L-arabinofuranosidase from *Bifidobacterium longum* in *L. lactis*. However, the α-L-arabinofuranosidase from *L. lactis* was not characterized (Margolles and De los Reyes-Gavilán 2003). Our study is the first to demonstrate activity of *L. lactis*-like arabinofuranosidase against pNP-α-l-Araf. The enzyme follows Michaelis-Menten kinetics and has the highest specific activity among all characterized GH43 subfamily 26 members. The specific activity of LAraf43 is 0.08 U/mg, which is considerably lower than that of the previously characterized α-L-arabinofuranosidases from *Weissella* sp. strain 142, *L. brevis, S. chartreusis* and *S. avermitilis* which are 5.4 U/mg, 1.94 U/mg, 3.16 U/mg and 2.92 U/mg, respectively (Linares-Pastén et al. 2017; Matsuo et al. 2000; Ichinose et al. 2008). LAraf43 specific activity was tested at pH 7.4 and at a temperature of 37°C, which is similar to the termite gut environment (Brune, Emerson, and Breznak 1995; Brune 2014). The optimal pH condition for the characterized α-L-arabinofuranosidases was found to be from 5.5 to 7 with an optimum ranging between 37 and 45°C (Table 1). For *Lactobacillus brevis* DSM1269 (LbAraf43) and *Weissella* strain 142 (WAraf43), specific activities towards pNP-α-l-Araf were measured at pH 5.5 and a temperature of 37°C, which were lower than their optimal values. This could reduce the specific activity of these enzymes (Linares-Pastén et al. 2017). The α-L-AFase II protein from *Streptomyces chartreusis* GS901, has an optimal pH at 7.0 but is very unstable at temperature lower than 40°C (Ichinose et al. 2008).

Another critical residue in hemicellulose is 4-*O*-methyl-D-glucuronic acid, which is cross-linked to lignin thereby preventing xylan hydrolysis. This residue can be cleaved by α-glucuronidase from families GH4, GH67 and GH115. We have identified an α-glucuronidase from the GH67 family called PGluc67 and showed that its closest homologue is an enzyme encoded by an uncultured bacterium from a cow rumen.

By studying the operon, it is possible to predict the taxonomy as well as understanding the gene function. Based on gene organization, it is predicted that the pathway identified from NODE_717 is similar to the Wzx/Wzy dependent pathway for biosynthesis of colanic acid. This acid is produced by gram-negative bacteria to form a protective capsule to shield the bacterial cell surface (Hanna et al. 2003; Furlong and Furlong 2013). It is made up of repeat units of D-glucose, D-fructose, D-glucuronic and D-galactose in different compositions that vary between strains and species. The wzx/wxy dependent pathway for the production of colanic acid in *E. coli* is the most well-studied and contains 19 genes (Schmid, Sieber, and Rehm 2015; Stevenson et al. 2006).

The gene organization found on the goat rumen contig NODE_717 has not been observed in *Prevotella* as well as other gram-negative bacteria species, but since homology of encoded proteins appeared highest to *Prevotella* proteins, it could derive from a *Prevotella* bacterium. The predicted functionality for the genes in the segment is based upon the functions of known orthologues. Under cell wall stress, the sensory kinase BaeS protein would be activated as part of a BaeSR two-component system known for being responsive to such conditions (Leblanc, Oates, and Raivio 2011; Vinés et al. 2005). It is predicted that carbohydrates are transported to the periplasm by a BtuB transporter, with the help of a tonB protein. The periplasmic PGluc67 releases D-glucuronic acid from carbohydrates as the precursor for CA production, which can be transported into the cell via another transporter such as an ATP Binding Casette (ABC) transporter, which all were not encoded by ORFs on NOD_717 but these could be located on another operon.

The Wcaj protein initiates the start of CA production by transferring the first glucose unit to the lipid II carrier need for CA synthesis. Other glycosyl transferase proteins such as BcsA and GT2, with the transaminase WecE then attach other sugars to the chain (Schmid, Sieber, and Rehm 2015; Marolda et al. 2006; Whitney and Howell 2013). This small repeating unit is translocated across the inner membrane through the flippase protein (Wza) (Hong, Liu,

and Reeves 2018). Polymerization of multiple individual repeats is then subsequently carried out by the periplasmic Wzy protein, which functions as an oligosaccharide repeat unit polymerase add more repeat units to the chain. The length of these chains is regulated by Wzz protein. The polymerized repeats are then transported to the cell surface via a Wza transporter and attached to the cell surface by means of a Wzi protein (Furlong and Furlong 2013; Schmid, Sieber, and Rehm 2015; Bentley et al. 2006).

Comparing to other known CA operons, the fragment lacks genes encoding the Wzb and Wzc proteins (Fig. 6). They are needed for the production of the capsule and to release polysaccharides extracellularly. These enzymes could be a part of a different operon elsewhere on the genome. However it has also been reported that alternative transport routes can act in the absence of Wzc and Wzb proteins (Bentley et al. 2006; Y. T. Huang et al. 2018; Pereira et al. 2018).

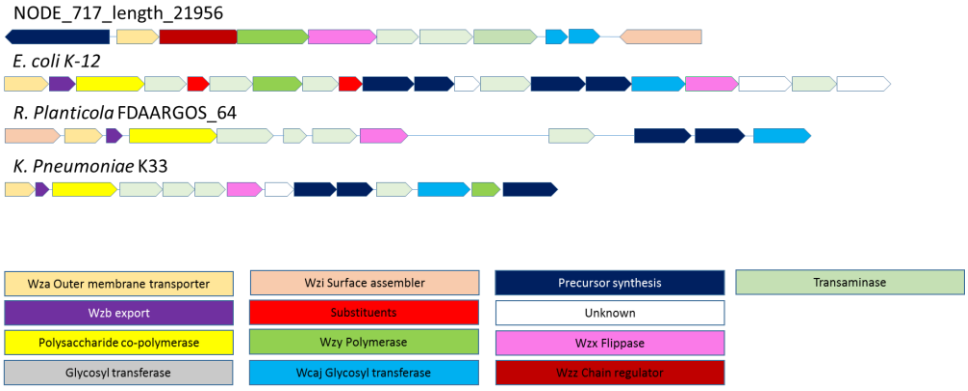


Figure 6: *Comparison of different functional Wzx gene cluster from different bacterium including function. (Modified from Schmid et al. 2015 and Huang et al. 2018)*

The presence of PGluc67 in the operon strengthens our conclusion that these proteins are likely to be functional. Sequence analysis revealed two conserved amino acid residues, aspartic acid D332 (general base) and glutamic acid E360 (general acid). The catalytic action, which is accompanied by a stereochemic inversion, happens when the general base deprotonates a water molecule and at the same time the general acid donates a proton to the anomeric carbon of the glycoside (Zaide et al. 2001; Nurizzo, Nagy, et al. 2002). The α-glucuronidase was predicted to have glycosyl hydrolase activity towards xylose to produce MeGlcA residues.

The kinetic activity of PGluc67 at 37°C and pH 6 – 8 and optimum temperature from 20°C – 40°C is in agreement with the gut conditions from which the metagenome was assembled. The α-glucuronidases from another uncultured bacterium, *Thermotoga maritima* and GH115's from *Bacteroides ovatus* also show an optimal pH around this range (C. C. Lee, Kibblewhite, Wagschal, Li, Robertson, et al. 2012; C. C. Lee, Kibblewhite, Wagschal, Li, and Orts 2012; Rogowski et al. 2014). Most of the characterized α-glucuronidase proteins are acidic and so not very suitable for usage after alkaline pretreatment. Only the uncultured bacterium AFJ94648.1 has an optimal pH range from 5.5 to 9.5 (C. C. Lee, Kibblewhite, Wagschal, Li, Robertson, et al. 2012). *Thermotoga maritima* is a marine hyperthermophilic bacterium, which can grow at a temperature of 90°C and above. Its α-glucuronidase is more stable at the range 60°C – 100°C (Ruile, Winterhalter, and Liebl 1997). Multiple factors can lead to hyperthermophilic activity such as salt bridges, hydrogen bonds, specific amino acids and α helices (S. Kumar, Tsai, and Nussinov 2002). The specific activity of PGluc67 is 46.8 ± 2.10 U/mg and is higher than α-glucuronidase from *G. stearothermophilus* 236 but lower than the xylan α-1,2-glucuronidase / α-glucuronidase from *Cellvibrio japonicas* with 15.3 U/mg and 61.3 U/mg respectively (Nagy et al. 2002; I.-D. Choi, Kim, and Choi 2005).

The activity of PGluc67 was tested in the presence of various metal cations and chemicals. Calcium and zinc were predicted to bind to PGluc67 (data not shown). Indeed, PGluc67 activity was slightly increased when calcium ions was present; the same was observed with α-glucuronidase from the fungal *Aspergillus niger* (Kiryu et al. 2005). The α-glucuronidase from *T. maritima* is slightly activated by Mn^{2+} ions, and lost its activity in the presence of Zn^{2+} (Ruile, Winterhalter, and Liebl 1997). In the presence of Mg^{2+}, K^{+} and Ni^{2+} ions, PGluc67 lost some activity similarly to α-glucuronidases from *G. stearothermophilus* and *Paenobacillus curdlanolyticus*. Metal ion such as zinc inhibited PGluc67, similarly to effects found in *G. stearothermophilus*, *P. curdlanolyticus* and *S. degradans* (Zaide et al. 2001; Septiningrum et al. 2015; I.-D. Choi, Kim, and Choi 2005).

While we have characterized two new enzymes involved in hemicellulose degradation, we realize that multiple enzymes are needed for its total degradation. The xylan is broken down using xylanase into xylo-oligosaccharides with ferulic and acetic acid as byproducts in combination with acetylxylan esterase and feruloyl esterase. Esterases then remove the acetyl group to loosen the bond for xylanases. The structure is further broken down into by α-L-arabinofuranosidase to produce arabinose while α-glucuronidase generates glucuronic acid.

Oligosaccharides are converted to xylose sugars with the help of xylosidases (Dodd and Cann 2009; X. Liu and Kokare 2016). By combining multiple enzymes such as xylanase, α-L-arabinofuranosidase, and α-glucuronidase, it is possible to rapidly release xylose sugars. Multiple experiments have been reported regarding this approach. McKee et al. 2016 reported a cocktail of xylanase, α-glucuronidase, α-l-arabinofuranosidase and β-xylosidase, which were selected to efficiently break down glucuronoarabinoxylan and to generate arabinofuranose, xylopyranose and MeGlcA monosaccharides (McKee et al. 2016). A similar approach using a high temperature resistance enzyme cocktail was tested for the biotechnology industry. A hyperthermophilic α-glucuronidase from *T. maritima* was used in combination with β-xylosidase. Similar results were obtained showing xylose, xylobiose and 4-O-methylglucuronic acid as products (Zhou et al. 2018). This short summary illustrates the complexity of complete degradation of hemicellulose.

We show in this paper that bioinformatic tools are of great value to explore CAZymes in functional metagenomes and identify potentially valuable enzymes for industrial applications. It is possible to create a pipeline for high-throughput candidate gene selection. The two new hemicellulases that we have identified were showing high activity, which could be used for lignocellulose degradation. Further research must show their usefulness in an industrial set-up.

Table 1: Comparison of different characterized α-L-arabinofuranosidase.

Name	Species	% id LAraf43	kDA	pH	Temp	Substrates	Homology	Accession	Specific activity pNP-a-L-Araf	Ref
Abf3	Lactobacillus brevis DSM 20054	72.06	38	5.5	37	pNP-a-L-Araf, 1,5-a-L-Arabinobiose,1,5-a-L-Arabinotriose	tetramer	AGT14430.1	1.79 U/mg	(59)
LbAraf4	Lactobacillus brevis DSMZ 1269	72.06	40	6	45	pNP-a-L-Araf, 1,5-a-L-Arabinobiose,1,5-a-L-Arabinotriose	tetramer	APU52333.1	1.94 U/mg	(58)
AFase II	Streptomyces chartreusis GS901	49.31	37	7	50	Arabinoxylan, Arabinogalactan,Arabinan,Debranched Arabinan,methyl 5O-α-Larabinofuranosyl-α-Larabinofuranoside	monomer	BAA90772.1	3.16 U/mg	(60)
SaAraf43A	Streptomyces avermitilis NBRC 14893	53.65	52	6	45	pNP-a-L-Araf	monomer	BAC68753.1	2.92 U/mg	(61)
LAraf43	Lactococcus Lactis	100	40	7	37	pNP-a-L-Araf	Dimer	none	12 U/mg	This article
WAraf43	Weissella sp.strain 142	75.86	40	6	45	pNP-a-L-Araf, 1,5-a-L-Arabinobiose,1,5-a-L-Arabinotriose	Dimer	APU52332.1	5.4 U/mg	(58)

Chapter 6 - A functional carbohydrate degrading enzyme potentially acquired by horizontal gene transfer in the genome of the soil invertebrate *Folsomia candida*

Ngoc Giang Le, Peter van Ulsen, Rob van Spanning, Abraham Brouwer, Nico M. van Straalen, Dick Roelofs

6.1 Introduction

In regular transmission genetics, a genome is passed from the parents to the offspring and its DNA sequence reflects the evolutionary history of the organism. However, this is not always the case as genomes are changing and can be altered through loss of genes, expansion or contraction of non-coding or selfish elements. Different loci can have different evolutionary rates due to unequal selection pressures. Genes can be gained through duplication or acquired from foreign sources by horizontal gene transfer. Horizontal gene transfer (HGT) is a mechanism by which organisms may acquire functions that can hardly be obtained by selection on standing genetic variation. The frequency of successful HGT depends on the ability of the host to take in the foreign DNA, the ease at which the foreign DNA can recombine with the host DNA, the access to the germline and the frequency of the donor in the environment (Husnik and McCutcheon 2018).

Once integrated, the newly acquired DNA is subjected to selection. Only DNA that can transcribed and translated into proteins for the host to gain new functionalities or to contribute to existing functions are maintained. In addition, such genes are often adapted to the host. Non-beneficial genes are lost over time. After an HGT event, organisms will experience different evolutionary pressures. HGT is an important mechanism for evolutionary innovation and the exploitation of new habitats (Soucy, Huang, and Gogarten 2015; Husnik and McCutcheon 2018).

There are several ways by which genes can be transferred from one organism to another: transformation, transduction, bacterial conjugation and gene transfer agents. Conjugation implies that donor and recipient are in physical contact and genetic material is exchanged through a conjugation pilus. This process if often found among bacteria. *Agrobacterium* spp. uses this HGT mechanism to transfer T-DNA to plant cells. Transformation implies that environmental DNA is taken up by the recipient. Transduction is a mechanism in which phages or viruses deliver genetic material to the recipient. All of these mechanisms are seen in archaea and bacteria and have been crucial for the evolution for both of these organisms

HGT is not common among eukaryotes, or between prokaryotes and eukaryotes. When sequencing genomes of little-known animals several authors have claimed examples of HGT from bacteria into eukaryotes. Some of these examples have not withstood further investigation and may be due to contamination. However, that does not mean that such events

do not occur. Several solid cases for HGT in invertebrates have been made, including nematodes, tardigrades, rotifers and springtails (Mayer et al. 2011; Faddeeva-Vakhrusheva et al. 2017; Gladyshev, Meselson, and Arkhipova 2008). In animal genomes, due to their complexity it is difficult to identify HGT events. In the case of a bacterial donor, due to the high frequency of HGT among bacteria themselves, the donor DNA might or might not have the same evolutionary history as most of the genes from the donor bacterial genome. Other factors include bias in phylogenetic trees due to long-branch attraction, genes loss and shortage of samples to infer the donor of the DNA (Husnik and McCutcheon 2018).

The hexapod class Collembola has been shown to be a hot spot of horizontal gene transfer. In the genome of the model species *Folsomia candida* the percentage of open reading frames due to horizontal gene transfer after thorough validation was estimated as 2.8% (Faddeeva-Vakhrusheva et al. 2017). Since springtails live in close proximity with soil microbial communities and because they evolved as an ancestral group of hexapods, the opportunity for HGT is realistic. The class Collembola includes several species capable of anhydrobiosis, a mechanism of extreme droughttolerance that includes dissolution of the nuclear membrane and partial fragmentation of the genome. Anhydrobiosis has been suggested as a mechanism of HGT in nematodes and bdelloid rotifers (Husnik and McCutcheon 2018).

A recurrent question in HGT cases is how the (usually prokaryotic) donor DNA can be not only inserted but also expressed in a eukaryotic environment. Only in a few cases it has been demonstrated that sequences acquired by HGT from prokaryotic donors are actually expressed in the host genome, potentially contributing to the enhancement of its fitness. These events have occurred in the evolutionary past and most likely continue to occur to shape eukaryotic genomes. Among the various genes acquired by HGT in springtails, biosynthesis clusters for beta-lactam antibiotics are one of the most striking (Roelofs et al. 2013). However, another important functional contribution of HGT is related to carbohydrate-active enzymes. Carbohydrates are needed for multiple biological purposes such as energy storage, signal transduction and intracellular trafficking (Mewis et al. 2016). They are also the future of renewable fuel. This is why it is important to identify enzymes that can breakdown biomass. Obviously, the degradation of recalcitrant biomass is an extremely important capacity for any detritivore soil invertebrate (Bredon et al. 2018). All soil invertebrates rely on a microbial gut community.

The carbohydrate active enzymes (CAZymes) form a diverse group of enzymes and other proteins, all with a function in carbohydrate metabolism. The CAZymes database is the largest and most well-known of all sequence-based classification systems. It is made up of glycosyl hydrolases (GHs), glycosyltransferases (GTs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), carbohydrate binding modules (CBMs) and auxiliary activities (AAs). These enzymes jointly are responsible for the breakdown of lignocellulose, an abundant carbohydrate resource in soil ecosystems. We previously applied this database to identify carbohydrate-active enzymes in the metagenome of the springtail (Faddeeva-Vakhrusheva *et al.*, 2017; Agamennone *et al.*, 2019; Le, submitted). We were able to link several HGT CAZy genes in the host genome to a putative microbial donor. We also showed that most of them are transcriptionally active. Here, we further characterize one of these HGT CAZymes: α-L-arabinofuranosidase.

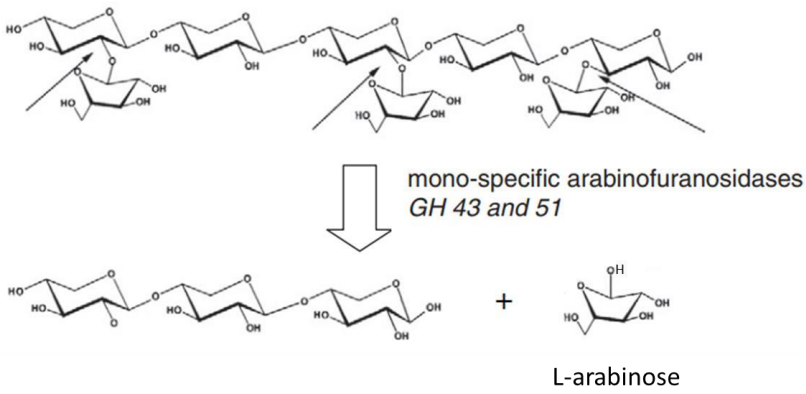


Figure 1: *Catalytic activity of a-L-arabinofuranosidase: the cleavage (indicated by arrows) of an L-arabinose side-group from the hemicullose backbone.*

This enzyme catalyzes the cleavage of L-arabinose side chains in hemicellulose, an important step in the final breakdown of this poly-sugar compound into monomeric sugars (Fig. 1). In this paper we provide functional evidence of its activity in vitro. In addition, we show that this gene, although of prokaryotic origin, shows adaptive evolution: it underwent eukaryotization and acquired an eukaryotic signal peptide, most likely to ensure extracellular action of the enzyme in the gut lumen of the host.

6.2 Methods

6.2.1 Gene annotation

Previously described transcripts from *Folsomia candida* was used for the analysis (Faddeeva-Vakhrusheva et al. 2017). Prodigal was used on the transcript to predicted bacterial Open Reading

Frames (ORF). Proteins with start and stop codons were scanned against the Carbohydrate Activity enZymes (CAZy) database using the hidden Markov model (HMM) model with the default settings (Yin et al. 2012). The CAZyme candidate genes were further explored using the basic local alignment search tool (BLAST) software from the National Center for Biotechnology Information (NCBI) to establish sequence homologies (Altschul et al. 1997). The 3-dimensional (3D) structures model and the binding sites were predicted using the Phyre2 web server (L. A. Kelley et al. 2015) and SWISS-MODEL (Waterhouse et al. 2018) with default settings. Nucleotide sequences were analyzed for the presence of signal peptides using gram negative and gram-positive settings using the SignalP4.1 server (Almagro Armenteros et al. 2019). Protein molecular weight and isoelectric point (pI) value calculations were done using Cloning Manager 9.0 (Sci-Ed Software, USA). The protein was blasted again the springtail proteins and transcript at https://collembolomics.nl/ using default settings (Faddeeva-Vakhrusheva et al. 2017).

6.2.2 Plasmid construction for recombinant expression

Total RNA from whole springtails was extracted using the SV Total RNA isolation system according to manufacturer's protocol (Promega, Wisconsin, US). Subsequently, messenger RNA was converted to cDNA using oligo dT(15)-guided reverse transcription with AMV reverse transcriptase according to manufacturer's instructions (Promega, Wisconsin, US). PCR was performed on cDNA by applying the following oligonucleotide primers designed on the predicted α-L-arabinofuranosidase (FcAraf43) gene from the ORF of *Folsomia* transcript: 5'-primer (5'GGGCATATGGCTTTCACAAAAATATTG-3'), which included the ATG translational start codon inside a *Nde*I restriction site (shown in italic) and 20 nucleotides of the ORF; The 3'-primer (5'- AAACTCGAGTTATTCCCCACTTGGAAC-3') included a stop codon (TAG), containing an *Xho*I

restriction site and the preceding 26 nucleotides of the ORF. Three guanine and thymine residues were added at the 5'-end of the 5'-primer and 3'-primer, respectively, to create a good binding site for the respective restriction enzymes. The gene sequence was amplified using *Taq* and *Pfu* polymerases and the product was purified on a 1% agarose gel. It was digested with *Nde*I and *Xho*I and ligated into *Nde*I/*Xho*I-digested pET16b vector, resulting in the plasmid pET16-FcAraf43 with an N-terminal His-tag. The resulting plasmid was transformed into XL1-blue chemically competent cells. Successfully transformed colonies were screened by restriction digestion and correct inserts were confirmed by DNA sequencing (Macrogen). After quality control, intact pET16-FcAraf43 plasmid DNA was transformed in *E. coli* expression strain Rosetta2 (DE3) (Novagen).

6.2.3 Recombinant protein expression and purification of FcAraf43

A glycerol stock of transformed Rosetta2 was used to inoculate into 200 ml of LB medium and 100 µg/ml ampicillin at 37°C. Cells were cultured until the optical density at 600 nm (OD_{600}) reached 0.6-0.8. The cultures were induced by adding 50 µM isopropyl-beta-D-

thiogalactopyranoside (IPTG) for gene expression and further incubated for 2 hours at 37°C. After centrifugation, the cells were harvested and suspended in 8 ml of phosphate buffer saline (PBS; pH 7.4). Protease inhibitors cOmplete™, EDTA free (Roche) cocktail, was added followed by two passages at ~1.7 k psi through a OneShot cell disruptor (Constant Systems Ltd) at room temperature. The debris and membrane fragments were removed from the cell extract after centrifugation at 586 g for 10 min and 100,000 g for 1 hour. TALON Superflow resin (GE, Sweden) premixed with buffer A (50 mM potassium phosphate buffer, 500mM sodium chloride, 10% glycerol, 10 mM imidazole pH 7.5) was added to the cleared cell extract and mixed. The mixture was incubated at 4°C, for 1 hour and transferred to a disposable 5 ml polypropylene column (Thermo Scientific) to be washed with 10 ml of buffer A. Several wash solutions with increasing imidazole concentration up to 200 mM were used. The His-tagged proteins were eluted from the beads by adding 10 ml of buffer B (50 mM sodium phosphate buffer, 500 mM sodium chloride,

10% glycerol, 400 mM imidazole pH 7.5). To concentrate the sample and remove salts, a Vivaspin 20, MWCO 10 kDa column was used. About 20 ml PBS pH7.4 was added and centrifuged at 6,000 g for five times after which the retentate was collected and aliquoted. The BCA protein assay kit (Thermo Scientific) with bovine serum albumin (BSA) as the

standard was used to measure the concentration of purified protein. For displaying protein, the crude extracts or purified protein samples were denatured in sample buffer with dithiothreitol (DTT), boiled for 10 min and applied to 10% gradient sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE, BIORAD) along with the molecular weight marker to determine the molecular weight and purity. The gel was stained with 0.1% coomassie blue as previously described by Lämmli (Lämmli 1970).

6.2.4 Cell-free protein expression

About 25 µL of plasmid was extracted and used with the PURE protein expression system (New England Biolabs, US). The mixture was incubated at 37°C overnight. The proteins collected were ran on SDS-PAGE. The rest of the proteins were washed and condensed for activity testing.

6.2.5 Enzyme assays for FcAraf43

Synthetic p-nitrophenyl-α-L-arabinofuranoside (pNP-α-L-Araf) was purchased from Megazyme International (Wicklow, Ireland). Alpha-L-arabinofuranosidases catalyze the release of *p*nitrophenol (pNP) from pNP-α-L-Araf, which can be measured at 405 nm (Biotek USA). Each assay mixture contained 10 µl of a 25 mM pNP-α-L-Araf solution with 88 µl of PBS buffer (pH 7.4) and 2 µl of enzyme solution. The reaction was carried out at 37°C and measuring pNP overnight. Readouts in blanks were subtracted from sample reads. As a positive control, an active α-Larabinofuranosidase gene from *Lactococcus lactis* was used. The assay activity was performed in triplicate as mentioned above, unless otherwise stated.

6.2.6 Phylogenetic analysis

Sequences of characterized GH43 enzymes from www.cazy.org were used to create a phylogenetic tree using ngphylogeny.fr with the advance FastTree analysis applying 1,000 bootstrap replicates (Lemoine et al. 2019). Clustal omega was used to align these sequences together (Sievers et al. 2011).

6.3 Results:

Results from collembolomic web server, show that the FcAraf43 open reading frame of 1,029 bp length was mapped back to scaffold 4 (Fcan01_Sc004) of the genome of the *Folsomia candida* with 100% identity. The FcAraf43 protein of 343 amino acids was predicted from

the Fcan01_09776-PA transcript (Faddeeva-Vakhrusheva et al. 2017). This gene has underwent the process of changing to the host genome as there is a predicted polyadenylation signal (AATAAA) 79 bp downstream of FcAraf43 ORF (Fig. 2).

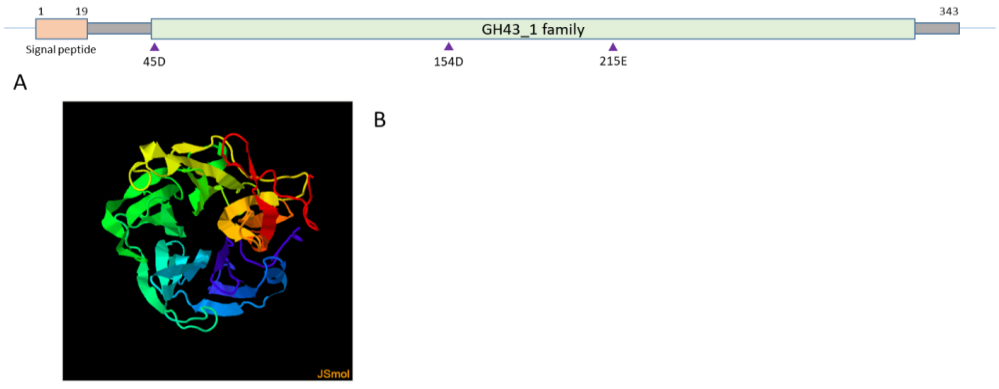


Figure 2: *Genomic context and 3D protein structure of FcAraf43. 2A: Genomic context and putative 3D structure of Fc Araf43. The signal peptide is the first 19 amino acids. The green segment is the conserved region characteristic for the CAZy GH43 family 1. The purple triangles indicates the locations of residues contributing to the active catalytic site.* ***2B:*** *The predicted 3D structure of the α – L – arabinofuranosidase. The five-blade spatial structure is characteristic of enzymes in the GH43 CAZy group.*

The core of the protein matches glycosyl hydrolase group 43 group 1 in the CAZy database. Alignment of FcAraf43 against the non-redundant protein database using blastp show that it is close to a similar sequence in the sister species *Orchesella cincta* (ODM95222.1), the midge *Bradysia coprophila* (XP_037044875.1) and the bacterium *Thermoanaerobacterium thermosaccharolyticum* (WP_015311875.1) with 68.31%, 51.69% and 40.17%, respectively. This enzyme was predicted to be a case of putative HGT, because a very similar gene in the sister genome of *O. cincta* had only microbial sequences in the top blast hits. The sequence in the *Folsomia* genome, and the protein predicted from the Fcan01_18043-PA transcript was 78% identical, with an e-value smaller than 0.01, to the *O. cincta* sequence. This latter protein was annotated as an α-L-arabinofuranosidase from *Streptomyces chartreusis* (https://collembolomics.nl/).

Figure 3: *Protein alignment and phylogenetic relation of FcAraf43: 2A alignment, each color is linked to particular amino acid. The asterisks show the common amino acids. 2B Phylogenetic tree of FcAraf43 with fungi and bacterial genes from classified CAZy family 43. The outgroup is Micromonospora. The number shows the bootstrap after 1,000 repeats. The Araf43 is clustered with Streptomyces chartreus. The fungal genes include Humicola, Magnaporthe, Chrysosporium and Penicillium.*

An N-terminal signal peptide of 19 amino acid was detected in the FcAraf43 gene. This shows that this protein is targeted for excretion to the extracellular environment. Further analysis shows that the predicted three-dimensional structure consists of five bladed-beta propellers found in the GH43 group. Two of the three active sites corresponding to amino acid positions 45 and 154 are aspartic acid while the third, at amino acid position 215 is a glutamic acid. These are conserved residues. The two aspartates act as general acid and pKa modulators. The glutamate acts as a general acid. Together they ensure the inverting glycoside hydrolase reaction, characteristic for GH43 CAZymes. The same conserved regions were also identified in GH43 enzymes of *Cellvibrio japonicas* (Nurizzo, Turkenburg, et al. 2002).

Alignment analysis with similar proteins from bacteria as well as fungi shows that both aspartates are conserved (Fig. 3A). The glutamate position, however, is different between prokaryote and eukaryote versions of the gene. As seen in the gene tree, FcAraf43 does fall within the bacterial clade of arabindonfuranisidases, and shows some resemblance with *Streptomyces chartreus*. This suggests a bacterial origin of the HGT. However, unique branch length of FcAraf43 is quite long and therefore indicative of a long evolutionary history of the HGT event (Fig. 3B).

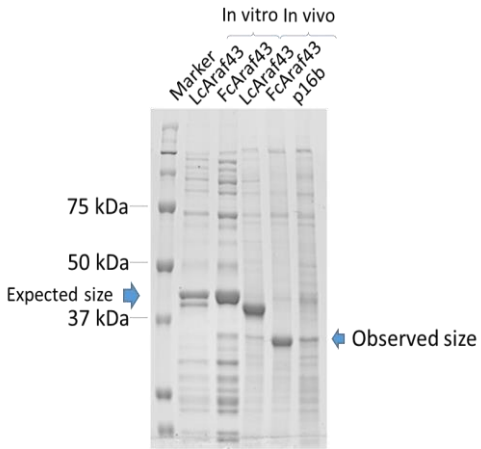


Figure 4: *Protein gel of cell free expression of FcAraf43 . Lane 1: in vitro FcAraf43 expressed using cell-free system. Lane 1: Molecular weight standard. Lane 2: Positive control α-arabinofuranosidase enzyme from a Lactococcus strain through the in vitro method. Lane 3: FcAraf43 in vitro. Lane4: Positive control cell culture LcAraf43. Lane5 Cell culture FcAraf43. Lane 6: Negative control empty pet16 vector.*

The FcAraf43 protein was expressed in *E. coli*. However, a protein band at the expected size was not found (Fig. 4). The gene was transferred into a new vector pGEMT (Promega, USA) and expressed in the cell-free protein expression system PURE express (BioLabs, USA). The total protein was used for the activity testing.

Analysis against the positive control showed that even under the cell free system little protein was expressed. However, the small amount of protein showed activity even though the amount was not as high as in the positive control from *Lactococcus lactis*.

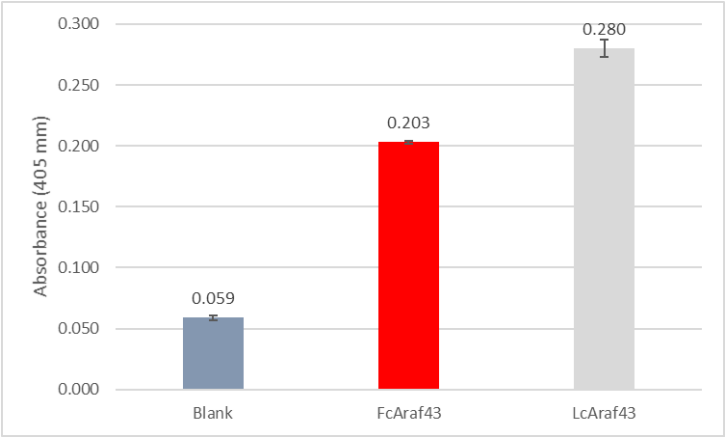


Figure 5: *Activity of Folsomia candida arabinofuranosidase 43 (FcAraf43), expressed in a cell-free extract after recombinant expression in E. coli. The absorbance at 405 mm of the F. candida geneis compared the activity of L. lactis arabinofuranosidase 43 (LcAraf43).*

Further analysis of the protein on the gel was performed. The start and end peptide of the protein were found to be intact (Fig. 4). The bands appearing at 37 KDa were identified as a housekeeping genes from *E. coli* (P. Hensbergen, personal communication).

The absorbance of the FcAraf43 at 405 mm is 0.203 (Fig 5). This shows that FcAraf43 is active (center), but at a lower rate than the *Lactococcus* positive control (right), which was recorded at 0.280. This could be due to the fact that the gene is from a eukaryotic origin and so difficult to be expressed in the bacterial host *E. coli*.

6.4 Discussion

We have identified a functionally novel gene in *Folsomia candida* that has α-arabinofuranosidase activity, that potentially evolved in the host after horizontally gene transfer.

The enzyme α-L-arabinofuranosidase is used in many industries such as food, animal feed and wine (Numan and Bhosle 2006; Yaru Wang et al. 2015; Thakur, Sharma, and Goyal 2019). By identifying a new variant of this protein it is possible to understand more about the enzyme and its evolution as well as potential function in combination with other CAZymes to break down biomass.

The gene was identified and cloned through the analysis of the active transcriptome of the springtail. The gene was predicted to be belong to the glycosyl hydrolase group 43. The predicted 3D structures from FcAraf43 shows a 5-bladed β-propeller structure with variable binding mechanisms and amino acids in the catalytic centre that are common to CAZymes in this group. The enzyme contains two aspartic acid and a two glutamic acid residue, which essentially contribute to the active site (Nurizzo, Turkenburg, et al. 2002; Vandermarliere et al. 2009; Jiang et al. 2012).

The gene shows multiple characteristics of HGT. It is small, only 1,029 bp. Small functional inserted DNA fragments are often tolerated better in a host genome than longer ones (Husnik and McCutcheon 2018). The GC content of FcAraf43 is 47% and is higher than the average value of 37.5% in the springtail nuclear genome (Faddeeva-Vakhrusheva et al. 2017). It is expected that over time, the gene will change, for instance in GC content to become equal to the host (Husnik and McCutcheon 2018). FcAraf43 is predicted an old HGT gene as it shows long unique branch length after splitting off from a potentially bacterial ancestor (Fig. 2B) . Moreover, it is located on scaffold 4 of the F. candida genome assembly (Faddeeva-Vakhrusheva et al. 2017), which is a gene-rich region with most abundant density of HGT genes. This scaffold is also rich in DNA transposon and retrotransposon sequences, potentially facilitating HGT (Husnik and McCutcheon 2018).

Phylogenetic analysis shows that the outgroup is the bacteria *Micromonospora echinospora*. The closest gene to FcAraf43 is *Streptomyces chartreusis*. There is a 55% confidence for the clade.

However, further blastp analysis shows that there is only 32.10% protein identity with *Streptomyces_avermitilis* and 29.82% with *Streptomyces_chartreusis*. The species *Streptomyces* seem to be closely cluster into clades, which contain fungi *Magnaporthe*, *Chrysosporium* and

Penicillium. This species is an ancient group of bacteria of about 380 million years old, which have ~ 300,000 gene transfer event as well as large number of point mutations (McDonald and Curriea 2017). They contain many bacterial conjugative elements. These elements can integrate modular mobile genetic elements into a host genome (Stewart et al. 2017). However, as pointed out by Mcdonald and Currie 2017, the HGT events are quite rare and due to different evolution rate it is difficult to locate the event.

One of the reasons why the gene was maintained in the host genome may be due to its advantage to the host nutrition as well as exploring the host into new niches such as in helping digesting hemicellulose (Ricard et al. 2006). Biomass is an abundance source of energy. Effectively degradation of this energy source will be very beneficial to the host. Having this gene helps the springtail to breakdown hemicellulose and decrease its dependence on the gut microbiome. There are other cases where HGTs gene have been found in relation to herbivorous insect. Even though these events are very rare, however they do occurs. A functional mananase was found in the coffee berry borer beetle, *Hypothenemus hampei*. This gives the beetle advantages in degrading the polysaccharide of the coffee seeds (Acuña et al. 2012). Endoglucanases and pectinases were also observed in plant parasitic nematodes (Scholl et al. 2003). Another common feature found is most of these cell wall or other glycan degrading enzymes is that they are secreted. This was observed in other host such as nematode (Danchin et al. 2010), spider mite (Grbić et al. 2011) and parasitic wasp (Di Lelio et al. 2019). The signal peptide was found was the Sec/SPI secretory signal peptides. It is transported by the Sec translocon and cleaved by Signal Peptidase I (Lep) (Owji et al. 2018; Almagro Armenteros et al. 2019).

FcAraf43 seems to be, however, incompletely optimized and so more testing needs to be done for a better understanding of the enzymes. It might also be used in a cocktail in to breakdown biomass.

6.5 Conclusion:

We have identified an active novel HGT α-L-arabinofurnosidase from the *Folsomia candida*. The HGT gene helps the springtail to digest hemicellulose from plant biomass. This in turn helps the springtail from utilizing the energy sources of arabinose and help it to survive in the niche environments.

Chapter 7 - Discussion

7.1 Introduction

Microorganisms are found everywhere on Earth and are an important part of nature. They live in communities depending on the environment they are in. In the recent years, researches focusing on gut microorganisms have demonstrated the significant impact they have on their hosts (Belkaid and Hand 2014). Symbionts in the gut have multiple functions, such as breaking down food to provide nutrients, modulate the immune systems as well as influence the host development (McFall-Ngai et al. 2013). We know that in humans, babies born by caesarean section lack key microbes present in naturally born infants (Callaway 2019; Casaburi et al. 2019). As the children get older, the gut microbiome continues to change. Wilmanski *et al* (2020) showed that the core microbiome becomes increasingly unique as humans grow up healthy. In unhealthy individuals the core microbiome was maintained up to high age, but did not show the pattern of increasing uniqueness (Wilmanski et al. 2020). More and more researches are showing the importance and impact of the gut microbiome on the host such as healthy aging and survival (Oliphant and Allen-Vercoe 2019; Wilmanski et al. 2020).

The gut bacteria can break down specific nutrients and can provide valuable and sometimes essential resources to the host (Rowland et al. 2018). On top of that, these organisms need to compete with other microbes and produce toxins, metabolites and organic compounds. These secondary metabolites and/or compounds are not required for the growth of the species but essential for the survival of the organism. Resistance genes and production of antimicrobial products can protect the microorganisms from other competitive community members (Baron, Diene, and Rolain 2018; Casals-Pascual, Vergara, and Vila 2018). Microbiomes that have multiple functions, as mentioned above, are more likely to be selected and stay in the gut. These properties can be determined by looking at the metagenomes. A complex system of genes is required for the organism to live in the gut environment. By studying these, it is possible to identify novel genes and also to understand the interaction between the host and its symbionts.

The gut main function is to extract nutrients from food. The intestinal tract is populated by mostly beneficial microorganisms. The host must strike a balance between possible assistance from symbionts and protecting itself from pathogen invasion. This is why the microbiome residing in an animal gut is defined by the environmental conditions in the gut

(pH, anoxia, host immune factors, reactive oxygen species), but also the diet and lifestyle of the host (Thursby and Juge 2017; Moran, Ochman, and Hammer 2019; Rinninella et al. 2019; Wilkins et al. 2019). The microorganisms adapt and provide benefits to the host for example by supporting digestion, nutrient synthesis, toxin metabolism, pathogen protection and metabolism of toxins (Moran, Ochman, and Hammer 2019).

In this thesis, we investigated the gut metagenomics of goats and compared the guts of three invertebrates, termite, springtail and isopod. All hosts have the capacity of degrading recalcitrant biomass, with the termite and goat being the best adapted to deploy this function. We used bioinformatics tools to investigate functional enzymes to breakdown hemicellulose. The methodology is strengthened by applying the same approach to different animal guts. First, we used whole genome sequencing and bioinformatics techniques to study the composition of goat guts in Vietnam (Chapter 2). A similar approach was carried out for the springtail (Chapter 3). The guts of the invertebrates were compared with regard to their biomass degrading, antibiotic resistance and secondary metabolites capabilities (Chapter 4). Finally using metagenome data, we mined and characterized two hemicellulases: α-L-arabinofuranosidase and glucuronidase from the goat and termite guts (Chapter 5). Further investigations were conducted into the evolution of a hemicellulase, α-L-arabinofuranosidase from the springtail, which was demonstrated to be active in the eukaryotic genome after horizontal gene transfer from a prokaryote (Chapter 6).

In the Discussion below, I will first try to answer the three main questions that were posed in the introduction of the thesis:

1. What microbial communities are present in the different animal hosts, and how do they compare to each other?
2. Which functionalities are encoded in the metagenomes of these communities (with emphasis on carbohydrate metabolism)?
3. What are the properties of metagenome-derived enzymes as possible candidates for biobased degradation of organic waste?

7.2 Different species with similar functionalities

The termite appears to have the most diverse collection of microbial species in its gut, due to the large number of identified and unidentified taxa. Along the three compared invertebrates, the isopod has the least number of contigs from prokaryotic species. The common phyla found in the guts of the three invertebrates were Proteobacteria, Bacteriodetes, Firmicutes and Actinobacteria. Their abundance is also different from host to host. The Actinobacteria are most represented in the springtail, which is commonly observed among gut microbial genomes from invertebrates living in the soil (Pass et al. 2015). Agamennone *et al (*2018) already showed that this group contributes most to the biochemical functions to *Folsomia*'s gut microbiome. Firmicutes and Bacteroides can perform anaerobic digestion process (Flint and Bayer 2008; Campanaro et al. 2016; Güllert et al. 2016). Some species of the Proteobacteria are capable of both aerobic and anaerobic metabolism (Mhuantong et al. 2015; N. Zhu et al. 2016). Other phyla such as Spirochaetes and Planctomycetes were higher in termite than in springtail or isopod. Further analysis showed that genera common to springtail, termite and isopod were *Pseudomonas*, *Enterobacter*, *Lactococcus*, *Staphylococcus* and *Microbacterium*. In the isopod, the Tenericutes phylum is higher than the other gut metagenomes. This phylum might contain pathogens and/or mutualistic symbionts and plays a role in degrading recalcitrant carbon sources in the gut of their hosts (Yong Wang et al. 2020). However, not all of these species are in high abundance in the metagenomes. Some low abundance species could also contribute and drive the gut composition (Benjamino et al. 2018).

Another soil living organism, the nematode *Caenorhabditis elegans* contains the core families of *Burkholderiaceae*, *Pseudomonadaceae, Xanthomonadaceae* and *Enterobacteriaceae* (Proteobacteria), and *Bacillaceae* (Firmicutes) in its gut (M. Berg et al. 2016). Similarly, the goat gut was also populated with Bacteroidetes, Firmicutes, Proteobacteria, Spirochaetes and Cyanobacteria (Do, Dao, et al. 2018; Do, Le, et al. 2018; Moran, Ochman, and Hammer 2019). This suggests that some form of commonality exists among communities in the gut, across a wide range of animals.

In *Drosophila*, the dominant phyla are Firmicutes and Proteobacteria. Interestingly, Actinobacteria, Bacteriodetes and Cyanobacteria appear to be related to the egg and larval stages of the fruit fly and influence development. However, in adult flies, and as the flies get

older, the number of gut bacteria decreases (Wong, Ng, and Douglas 2011; Bost et al. 2018). Another insect, the honey bee, also contains Firmicutes, Bacteroidetes, Betaproteobacteria and Gammaproteobacteria as dominant phyla, although their species compositions are very specific. Their alpha diversity is only 5-10 (Moran, Ochman, and Hammer 2019), much lower than in the gut microbiomes of termites, springtails and woodlice. The diversity of species seems to be driven by environmental factors such as the food source, as fruit flies feed on microbes in fermenting fruits and honey bee predominantly on pollens. The less diverse diet of honey bees might not require a diverse array of enzymes to breakdown the structure of biomass.

Taken together, this illustrates the complexity of interactions between the gut microbiome and the host. One thing that needs to be considered is that not all species are identified, which opens up possibilities of finding novel microbiome species/interactions in new hosts. This is in line with Larsen *et al.* (2017), who estimated that there are still a large number of new microbial species in every new host genome amenable to deeper investigation (Larsen et al. 2017; Douglas 2019). For example, new genome information was obtained for the *Verminephrobacter* strains, symbiotic bacteria living in the nephridia of earthworms. These bacteria possess crucial genes and pathways and play a pivotal role in micronutrient delivery, such as nitrogen fixation, to the host (Arumugaperumal et al. 2020).

7.3 Functionalities of gut metagenomes

It is now clear that the structure of the gut microbial community is related to variation in diet, gut structure, and immune system among animals (Moran, Ochman, and Hammer 2019). The importance of microbes to the host depends on how much the host relies on the nutrients and/or other functions provided by the microbes. The conditions in the guts and host diets makes the gene pool dynamic and highly adaptive. Consequently, animals contain extremely different functional gut communities. For the bacteria to be able to survive in the gut, they need a number of functional genes.

Clearly, diet is an important driving factor shaping gut communities. The termites and goats are well known their abilities to digest wood (Do, Le, et al. 2018) and isopod and springtails are decomposers, which feed on decaying biomass and fungal materials on the soil (Fountain and Hopkin 2005; Bouchon, Zimmer, and Dittmer 2016). Their gut microbiome communities were investigated for the enzymes that can breakdown polymeric carbohydrates, antibiotic

resistance genes and gene clusters involved with the synthesis of secondary metabolites. Carbohydrates are the source of energy for all heterotrophic organisms. Biomass is made up of complex polymeric molecules, which need to be broken down before any animal can use them.

In my thesis I focused on lignocellulose as major carbohydrate source for decomposers. Removal of lignin will release hemicellulose and cellulose. To fully break down biomass multiple groups of enzymes are required directed towards cellulose, hemicellulose and lignin. The glycoside hydrolase (GH) enzymes break down cellulose and hemicellulose. Esterases (CE) cleave ester bonds while uronic acid-containing polysaccharides are cleaved by polysaccharide lyase (PL). Finally, glycosyl transferases (GT) move sugar moieties to saccharide and nonsaccharide acceptors (Breton et al. 2006; Kameshwar and Qin 2017).

A large number of carbohydrate-active enzymes (CAZymes) have been identified in different hosts. To centralize the ever-increasing information on this class of enzymes, an online database system has been created named CAZy (http://www.cazy.org/) (Lombard et al. 2014). From a gut microbiome perspective, I showed that the termite has the largest number of CAZymes for breaking down lignin, hemicellulose and cellulose, more than the other two invertebrates. They contain the highest number of laccases (AA1) for breaking down lignin, and multiple enzymes from groups CE4, GH1, GH3, GH36, GH43 for breaking down hemicellulose and cellulose. When comparing the CAZyme content among gut microbiomes of termite species, major variation is observed associated with their ability to breakdown different types of biomass (Warnecke et al. 2007; Brune and Dietrich 2015; Grieco et al. 2019). Surprisingly, the springtail contains a large number of cellobiose dehydrogenases, which enable lignin degradation, next to hemicellulases and cellulases from the groups CE1, CE4, GH3 and GH6. Whether or not these CAZyme groups vary among springtail species needs further elucidation, since our studies are among the first to investigate this. In terms of carbohydrate-active enzymes the isopod has the lowest total number. However, they have the largest number of pectate lyases, more than the other two gut metagenomes. They also have many α-galactosidases (GH4), beta-glucosidases (GH1, GH3), and endo-beta-1,4glucanases (GH6). Again, data on variation of CAZyme gene content among gut microbiomes of isopod species is currently lacking.

Further analysis focusing on CAZyme contents among the species investigated in this thesis, suggests that there may be a core CAZyme group shared among animals. Many of these core enzymes represent GH1, GH3 and GH6 CAZymes, which code for glucosidase, galactosidase, as well as xylanase and arabinofuranosidase activity. Despite this functional similarity, the specific bacterial species that contribute to this core CAZyme activity differ from host to host. The three host species seem to deploy different taxonomic groups of microorganisms to fully degrade biomass with similar overall activity. In other words, it doesn't seem to matter who provides the CAZyme functionality as long as its activity is maintained in the gut. I also speculate that a high level of functional redundancy exists in the environment with respect to CAZyme functions, providing animals with the capability to incorporate these functions despite living in totally different ecological niches, while thriving on comparable carbohydrate (*e.g.* lignocellulose) sources. This could explain why the composition of their gut microbial communities may be functionally convergent, despite their taxonomic divergence.

The dynamics of gut bacteria depend not only on the host but also on temporal shifts in environmental resources consumed. Zhu *et al.* (2016) showed how a change of resources leads the changes in the abundance of enzymes over multiple days (N. Zhu et al. 2016). However, despite the taxonomic flexibility of the microbiome, the core enzyme constitution remained more or less stable, only fluctuating in abundance. The same phenomenon can be observed in humans. Even though the gut microbiome becomes more unique with age, the metabolic functions still retain similar traits (Wilmanski et al. 2020).

As a case study, I focused in my thesis on a unique carbohydrate-active enzyme function, identified in springtails. The glycosyl hydrolase α-L-arabinofuranosidase was subjected to further analysis. From a phylogenetic comparison of its sequence to databases it turned out that this gene was most likely acquired by *F. candida* through horizontal gene transfer (Chapter 6) (Faddeeva-Vakhrusheva et al. 2017). The enzyme catalyzes the cleavage of L-arabinose side-chains in hemicellulose, an important step in the final breakdown of this poly-sugar compound into monomeric sugars. In Chapter 6 we provided functional evidence of its activity. In addition, we show that the gene, although of prokaryotic origin, shows adaptive evolution: it underwent eukaryotization by acquiring a eukaryotic signal peptide, most likely

to ensure extracellular action of the enzyme in the gut lumen of the host. This example illustrates the intense interaction between the metagenome in the gut and the host genome.

Folsomia candida is among the animals (together with tardigrades, rotifers and nematodes) with the highest percentage of HGT genes in its genome (Faddeeva-Vakhrusheva et al. 2017). Earlier, we showed that also a cluster of genes associated with the production of beta-lactam biosynthesis was acquired by *F. candida* through HGT (Roelofs et al. 2013; Suring et al. 2017). Whether HGT is also a property of the other invertebrate genomes (termite, isopod) is not well investigated, because these organisms have not yet been subjected to detailed genomic analysis. It might be speculated that the occurrence of anhydrobiosis in Collembola, although not known for *F. candida* itself, might have played a role in acquiring novel gene functions through HGT in its evolutionary ancestors (J. Huang 2013). Anhydrobiosis is also known in rotifers, tardigrades and nematodes, where HGT has been indicated as an evolutionary scenario for acquiring novel gene functions (Abad et al. 2008; Flot et al. 2013; Yoshida et al. 2017).

7.4 Enzymes for the bio-based economy

When novel enzymes, recovered from invertebrate microbiomes, are considered as candidates for biotechnological application, their biosynthesis and function must be optimized with respect to the specific conditions for which it is intended. This is commonly done by heterologous expression and testing of the recombinantly expressed enzyme under various conditions. Several heterologous expression systems have been specifically optimized for increased CAZyme production, so that they may become more viable for industrial and biotechnological applications. For instance, *Zymomonas mobilis* confers an alternative glucose metabolism pathway compared to the conventional glycolytic pathway. As a consequence, it exerts high sugar uptake, lower cellular biomass yield and high alcohol formation, and has been successfully used as an efficient host for heterologous expression of extracellular cellulases (Linger, Adney, and Darzins 2010). Also, engineering its amino acid composition based on molecular modelling might be an relevant option (Chettri, Verma, and Verma 2020). Optimization could be achieved with respect to:

- *Specificity*. Many enzymes have multiple functions. The activity with respect to the prime function could be enhanced by engineering the enzyme towards that particular function.

- *Substrate dependence*. The enzyme could be optimized to show the highest reaction rate under the substrate concentrations that are prevalent in the intended application.

- *pH dependence*. In some cases, enzymes are used to release sugar from polymeric carbohydrates obtained after alkaline or acid treatments of lignocellulose. In these cases, the novel enzyme needs optimization with respect to its pH-dependency.

- *Temperature*. Enzymes isolated from ectothermic invertebrates are expected to show a temperature optimum at relatively low temperature. In biotechnology applications, the desired temperature of the reaction mixture might be higher and so an optimization is needed.

Two engineering methods, supported by computational biology, are in general applied to improve the above-mentioned properties for CAZymes (Figure 1). Directed protein evolution comprises a way of natural selection that can be implemented in a controlled manner under laboratory conditions to be applied on modification and testing towards more optimized CAZyme peptide structures. Variation in amino acid constitution is generated by random mutagenesis through, for instance, error-prone PCR on CAZyme cDNA amplification. The synthesized variants are subsequently screened and selected for the desired property. The advantage of directed evolution is that the method is less dependent on pre-knowledge about peptide structure. Many studies have been published on directed evolution on CAZymes targeting diverse functional modifications. For instance, Adesioye *et al.* (2018) subjected a an acetyl xylan esterase to directed evolution, which resulted in more thermostable determinants of carbohydrate esterase 7 family members (Adesioye et al. 2018).

A second approach (Figure 1) is rational design, where site-directed mutagenesis is applied to achieve pre-determined amino acid mutations, based on the structural and catalytic information of the enzyme of interest (Chettri, Verma, and Verma 2020). Following this approach, a laccase of *Bacillus* sp. HR03 was engineered by replacing a negative residue with hydrophobic residues on the surface of the protein, thereby enhancing thermo-resistance as well as solvent stability (Rasekh et al. 2014).

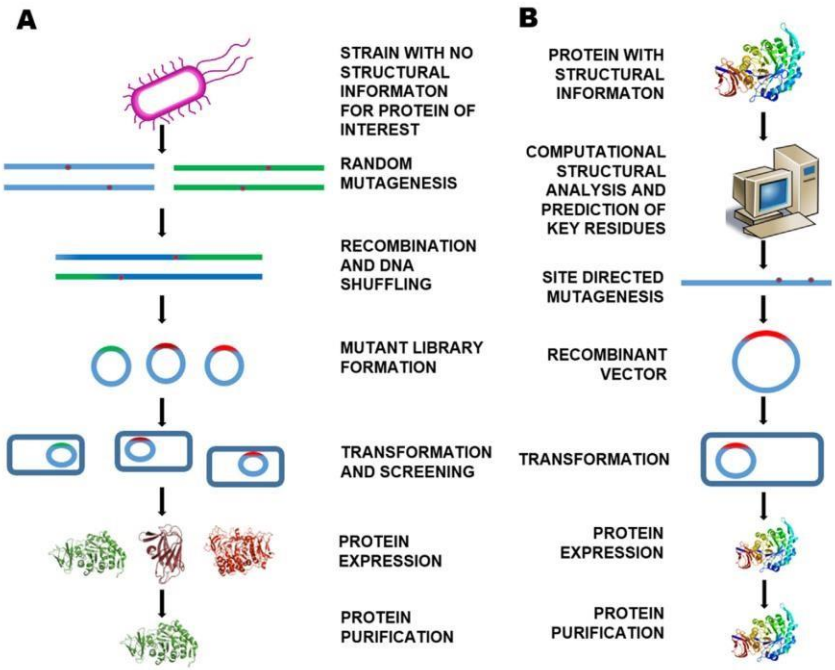


Figure 1: *Two approaches to optimizing CAZymes, A directed evolution B rational design. Figure courtesy of Chettri et al. 2020.*

Nevertheless, the exponentially growing number of predicted CAZymes from an exploding number of metagenomics studies has not been accompanied by a systematic and accurate attribution of function. Only a tiny fraction has been experimentally verified, which has become apparent in my thesis work as well. A potential solution for this is provided by Helbert (Helbert et al. 2019). They tried to better explore the sequence-to-function relationships of CAZymes by applying a strategy based on a rational bioinformatic selection of CAZyme targets from the existing database, followed by synthetic protein synthesis, and subsequent screening of recombinant proteins on a wide diversity of carbohydrate substrates. Only 14% exhibited actual enzymatic activity, but they found three new types of enzyme activities that had not been described previously (Helbert et al. 2019). This shows the power of combining a bioinformatic approach with a high throughput wet-lab testing and validation approach in the process of discovering new and more optimal enzyme functions.

These considerations illustrate that the isolation of enzymes from invertebrate metagenomes is only the first step in the development of novel activities for the bio-based industry. In this

thesis I have not yet been able to point out specific enzymes for direct biotechnological use. Still, my work is a possible contribution to a better insight into the molecular space of novel activity. If we know more about the sequences of gut enzymes in relation to the functions they have in different hosts, the basis for biotechnological optimization will be broadened. By looking at the space of possibilities provided by nature, solutions in biotechnology will be obtained more rapidly.

7.5 The future of the bio-based industry

Currently, we are urgently looking for the replacement of the traditional fossil fuel to more sustainable resources. The longer it takes for the transition to take place, the more will pollution, damage to the environment and human health become a problem. The bio-economy (BE) is a circular production of renewable biological resources as well as their conversion into food, feed, bio-based products and bioenergy. Multiple industries such as agriculture, forestry, fisheries, food and others are part of this bio-economy. Biorefineries provide the alternative sustainable solution for the production of bioproducts and bioenergy (Aristizábal-Marulanda and Cardona Alzate 2019). The biomass from agricultural waste is one of the preferred resources for supporting the urgent transition from fossil-based to sustainable energy (Chettri, Verma, and Verma 2020). Biorefineries breakdown plant materials into cellulose, hemicellulose and lignin. These structures can be further hydrolyzed to sugar monomers for fermentation or to various end products (Fernando et al. 2006). This approach benefits not only waste management but also reduces greenhouse gas emissions. The biorefinery market was estimated to be $714,6 billion by 2021 (Chettri, Verma, and Verma 2020).

As shown in my thesis, the animal guts are somewhat similar to biorefineries. In the animal guts, the biomass has a short retention time. The gut microorganisms are under selective pressure and competing against each other for the breaking down of biomass. Under different environmental conditions different bacteria and other microorganisms would contribute comparable enzymes for the biomass degradation. It emphasizes the notion that the origin of a certain trait is not so much an issue, provided that the trait can be delivered by any microbe that can be picked up by the host. This leads to a high level of functional convergence among multiple enzymes from different organisms (Chettri, Verma, and Verma 2020).

However, despite the large number of investigations in this field, there are still many unknown parameters with regard to information about CAZymes producers and genes involved. Biomass is difficult to breakdown and need tailor treatments. Traditional methods for pretreatment require pressure and heat to destroy the carbohydrates. The logical subsequent step is to further treat the resulting biomass with enzyme cocktails. Different types of pretreatments, physical and chemical, would need different treatment methods for downstream processes (Fig. 2). The condition of the fermentation process requires enzymes to be optimized for a specific pH and temperature. An effective way would be to have a range of cocktails, with multiple synergistic enzymes working together to generate different products and/or monomers (Merino and Cherry 2007; Bredon et al. 2018; Lopes, Ferreira Filho, and Moreira 2018).

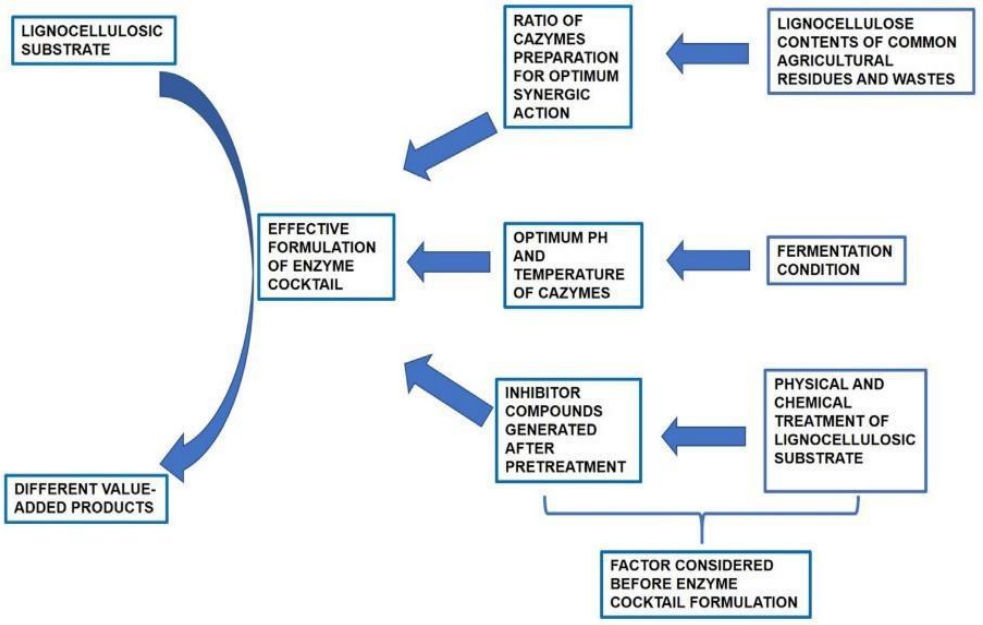


Figure 2: *Multiple factors such as composition of biomass, pre-treatment methodology, inhibitors generation, different CAZymes ratio and optimum activity conditions can impact the effectiveness of the cocktail (Chettri, Verma, and Verma 2020).*

Currently, there are some commercially available cocktails. The core enzymes of the cocktail include endoglucanases, exoglucanes, and beta glucosidase (Gao et al. 2014). These enzymes have previously been reported in the chapter 4 of my thesis. Additional enzymes such as

polysaccharide monooxygenases cleave the glycosidic bonds of cellulose via oxidation (Foreman et al. 2003) and hemicellulases can help to open up the structure and allows more cellulose to be accessed (Várnai et al. 2011).

The resulting products are furthermore suitable for later process such as saccharification (Horn et al. 2012). On top of that, the enzymatic enzymes approach help to maximum yield from biomass recycling and also produce alternative bioproducts. Enzymes such as laccase (Dao et al. 2021), lignin-peroxidases and manganases are used for the breaking down of dye (Yadav and Yadav 2015). Cellulases are used in textile industries for dye removal, fabric softening, and biopolishing; they are also used for biopulping in the paper industry. In animal feed production, cellulases are used to increase the nutritional value. They are even used as biopesticides due to their ability to degrade the cell wall of plant pathogens (Garron and Henrissat 2019).

With improved sequencing, metagenomics is becoming increasingly popular due to the ability to investigate uncultivated microorganisms and their functions in the target mini-ecosystem (Chettri, Verma, and Verma 2020). Metagenomics helps to look into the 99% of uncultured bacteria where there are potential novel enzymes. The development of bioinformatics tools helps to understand these microorganisms and their interactions/roles as well as their contribution toward the breaking down of biomass. The third-generation sequencing methods of Pacific Bioscience (PacBio) and Oxford Nanopore's MinION are capable of generating long read sequences. With the introduction of long read sequencing, more researches are using this technique for metagenomics (Haro-Moreno, López-Pérez, and Rodríguez-Valera 2020; Moss, Maghini, and Bhatt 2020; Maguire et al. 2021). Long read sequencing helps to identify complete open reading frames which is instrumental in annotation (Haro-Moreno, López-Pérez, and Rodríguez-Valera 2020).

By combining different omics approaches, the metabolic pathways involved in plant degradation can be mapped completely (Heyer et al. 2017). Meta-transcriptomics looks the gene expression from all microorganisms in the community. This approach can help to identify active genes as well as the microbe-microbe interactions ((Morita et al. 2011; F. Liu et al. 2012; Stark, Giersch, and Wünschiers 2014). Similarly, metaproteomics is the study of all the proteins in the studied community. The approach reveals the functional traits of microorganisms (Chettri, Verma, and Verma 2020).

On top of that, a set of more advanced high throughput techniques should also be employed. For example, combining metagenomics with expression of environmental DNA using multiple hosts growing on carbohydrate specific medium can help to screen for substrate-specific enzymes. Van Dijk et al (2020) reported a high throughput screening to identify the growth and ethanol production of a lignocellulose hydrolyzing strain of *Saccharomyces cerevisiae*. Multiple enzymes activity could also be investigated via automation systems (Bonowski et al. 2010; Navarro et al. 2010). When screening a large number of organisms and enzymes, it benefit greatly to have a large diversity of samples. Databases can help to solve problems or predict trends. Machine learning model when given a database to study can help to improve its prediction. The eCAMI is a k-merbased application that can be used for identification, classification, and genome annotation of CAZymes using a bipartite network algorithm (Jing Xu et al. 2020). Recently, Alphafold 2 was able to predict protein folding with a high accuracy using known protein crystals as the training model (Callaway 2020). Bioinformatic tools could also be used to predict random and/or direct mutations sites to make enzymes more thermostable, as explained in section 3 of this chapter. Newly modified enzymes can have activities up to 14 times the activity of the wild type (Anbar et al. 2012; M. A. Smith et al. 2012; Yoav et al. 2019). As more enzymes and microorganisms and biomass degrading cocktails are identified and improved, the fraction of products that are suitable for refineries of the bio-based economy will increase.

For the biorefineries to be functional, the technical, practicalities, as well as social aspect of biorefineries should also be addressed. There needs to be an effort from the government, scientists and farmers and public-private partnerships at national and local level to work together. For example, in Vietnam, burning of agricultural waste was common practice for a long time. For the farmers there is no incentive to recycle as there are no economic benefit. Other factors such as program conditions, incentive offered, famer's environmental preference, cultural characteristics and agricultural trends can affect the adoption of waste recycling (Piñeiro et al. 2020). The government and the scientific community could provide information as well as paying for the raw material. This would help the farmers to understand the important of recycling because they will get an extra source of income. Reports show that for short term adoption, economic benefit is essential. For the long term, the positive outcome for the environment and/or the farm would be the driver (Piñeiro et al. 2020).

In the future biorefineries recycling organic waste will generate benefits associated with energy conservation, food security, and mitigation of climate change at the same meeting societal demands for bio-products, chemicals and substances.

Bibliography

Abad, Pierre, Jérôme Gouzy, Jean Marc Aury, Philippe Castagnone-Sereno, Etienne G.J. Danchin, Emeline Deleury, Laetitia Perfus-Barbeoch, et al. 2008. “Genome Sequence of the Metazoan Plant-Parasitic Nematode Meloidogyne Incognita.” *Nature Biotechnology* 26 (8): 909–15. https://doi.org/10.1038/nbt.1482.

Abedinifar, Sorahi, Keikhosro Karimi, Morteza Khanahmadi, and Mohammad J. Taherzadeh. 2009. “Ethanol Production by Mucor Indicus and Rhizopus Oryzae from Rice Straw by Separate Hydrolysis and Fermentation.” *Biomass and Bioenergy* 33 (5): 828–33. https://doi.org/10.1016/j.biombioe.2009.01.003.

Acuña, Ricardo, Beatriz E. Padilla, Claudia P. Flórez-Ramos, José D. Rubio, Juan C. Herrera, Pablo Benavides, Sang Jik Lee, et al. 2012. “Adaptive Horizontal Transfer of a Bacterial Gene to an Invasive Insect Pest of Coffee.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (11): 4197–4202. https://doi.org/10.1073/pnas.1121190109.

Adams, Aaron S., Michelle S. Jordan, Sandye M. Adams, Garret Suen, Lynne A. Goodwin, Karen W. Davenport, Cameron R. Currie, and Kenneth F. Raffa. 2011. “Cellulose-Degrading Bacteria Associated with the Invasive Woodwasp Sirex Noctilio.” *ISME Journal* 5 (8): 1323–31. https://doi.org/10.1038/ismej.2011.14.

Adesioye, Fiyinfoluwa A., Thulani P. Makhalanyane, Surendra Vikram, Bryan T. Sewell, Wolf Dieter Schubert, and Don A. Cowana. 2018. “Structural Characterization and Directed Evolution of a Novel Acetyl Xylan Esterase Reveals Thermostability Determinants of the Carbohydrate Esterase 7 Family.” *Applied and Environmental Microbiology* 84 (8): 2695–2712. https://doi.org/10.1128/AEM.02695-17.

Agamennone, V., D. Roelofs, N. M. van Straalen, and T. K.S. Janssens. 2018. “Antimicrobial Activity in Culturable Gut Microbial Communities of Springtails.” *Journal of Applied Microbiology*. https://doi.org/10.1111/jam.13899.

Agamennone, Valeria, Dennis Jakupović, James T. Weedon, Wouter J. Suring, Nico M. van Straalen, Dick Roelofs, and Wilfred F.M. Röling. 2015. “The Microbiome of Folsomia Candida: An Assessment of Bacterial Diversity in a Wolbachia-Containing Animal.” *FEMS Microbiology Ecology* 91 (11). https://doi.org/10.1093/femsec/fiv128.

Agamennone, Valeria, Ngoc Giang Le, Nico M. van Straalen, Abraham Brouwer, and Dick Roelofs. 2019. “Antimicrobial Activity and Carbohydrate Metabolism in the Bacterial Metagenome of the Soil-Living Invertebrate Folsomia Candida.” *Scientific Reports* 9 (1): 7308.

https://doi.org/10.1038/s41598-019-43828-w.

Agrawal, AR, SA Karim, and Rajiv Kumar. 2014. “Sheep and Goat Production: Basic Differences, Impact on Climate and Molecular Tools for Rumen Microbiome Study.” *... . J. Curr. Microbiol. App. ...*. Vol. 3. http://ijcmas.com/vol-3-1/A.R.Agrawal, et al.pdf.

Al-Masaudi, Saad, Abdessamad El Kaoutari, Elodie Drula, Elrashdy M. Redwan, Vincent Lombard, and Bernard Henrissat. 2019. “A Metagenomics Investigation of Carbohydrate-Active Enzymes along the Goat and Camel Intestinal Tract.” *International Microbiology* 22 (4): 429–35. https://doi.org/10.1007/s10123-019-00068-2.

Alcock, Brian P., Amogelang R. Raphenya, Tammy T.Y. Lau, Kara K. Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, et al. 2020. “CARD 2020: Antibiotic Resistome Surveillance with the Comprehensive Antibiotic Resistance Database.” *Nucleic Acids Research* 48 (D1): D517–25. https://doi.org/10.1093/nar/gkz935.

Algburi, Ammar, Saskia Zehm, Victoria Netrebov, Anzhelica B. Bren, Vladimir Chistyakov, and Michael L. Chikindas. 2017. “Subtilosin Prevents Biofilm Formation by Inhibiting Bacterial Quorum Sensing.” *Probiotics and Antimicrobial Proteins* 9 (1): 81–90. https://doi.org/10.1007/s12602-016-9242-x.

Almagro Armenteros, José Juan, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2019. “SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks.” *Nature Biotechnology* 37 (4): 420–23. https://doi.org/10.1038/s41587-019-0036-z.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Altschul, Stephen F, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. 1997. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” *Nucleic Acids Research*. Vol. 25. Oxford University Press. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC146917/pdf/253389.pdf.

Amin, Farrukh Raza, Habiba Khalid, Han Zhang, Sajid U Rahman, Ruihong Zhang, Guangqing Liu, and Chang Chen. 2017. “Pretreatment Methods of Lignocellulosic Biomass for Anaerobic Digestion.” *AMB Express* 7 (1): 72. https://doi.org/10.1186/s13568-017-0375-4.

An, Dengdi, Xiuzhu Dong, and Zhiyang Dong. 2005. “Prokaryote Diversity in the Rumen of Yak (Bos Grunniens) and Jinnan Cattle (Bos Taurus) Estimated by 16S RDNA Homology Analyses.”

Anaerobe 11 (4): 207–15. https://doi.org/10.1016/j.anaerobe.2005.02.001.

Anbar, Michael, Ozgur Gul, Raphael Lamed, Ugur O. Sezerman, and Edward A. Bayer. 2012. "Improved Thermostability of Clostridium Thermocellum Endoglucanase Cel8A by Using Consensus-Guided Mutagenesis." *Applied and Environmental Microbiology* 78 (9): 3458–64. https://doi.org/10.1128/AEM.07985-11.

Andrews, Simon, and others. 2010. "FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010." *Https://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Anwar, Zahid, Muhammad Gulfraz, and Muhammad Irshad. 2014. "Agro-Industrial Lignocellulosic Biomass a Key to Unlock the Future Bio-Energy: A Brief Review." *Journal of Radiation Research and Applied Sciences* 7 (2): 163–73. https://doi.org/10.1016/j.jrras.2014.02.003.

Aotani, Yumiko, Hiroyuki Nagata, and Mayumi Yoshida. 1997. "Lymphostin (LK6-A), a Novel Immunosuppressant from Streptomyces Sp. KY11783: Structural Elucidation." *Journal of Antibiotics* 50 (7): 543–45. https://doi.org/10.7164/antibiotics.50.543.

Araki, Rie, Shuichi Karita, Akiyoshi Tanaka, Tetsuya Kimura, and Kazuo Sakka. 2006. "Effect of Family 22 Carbohydrate-Binding Module on the Thermostability of Xyn10B Catalytic Module from Clostridium Stercorarium." *Bioscience, Biotechnology and Biochemistry* 70 (12): 3039–41. https://doi.org/10.1271/bbb.60348.

Aristizábal-Marulanda, Valentina, and Carlos A. Cardona Alzate. 2019. "Methods for Designing and Assessing Biorefineries: Review." *Biofuels, Bioproducts and Biorefining* 13 (3): 789–808. https://doi.org/10.1002/bbb.1961.

Arrebola, Eva, Francisco M. Cazorla, Victoria E. Durán, Eugenia Rivera, Francisco Olea, Juan C. Codina, Alejandro Pérez-García, and Antonio De Vicente. 2003. "Mangotoxin: A Novel Antimetabolite Toxin Produced by Pseudomonas Syringae Inhibiting Ornithine/Arginine Biosynthesis." *Physiological and Molecular Plant Pathology* 63 (3): 117–27. https://doi.org/10.1016/j.pmpp.2003.11.003.

Arumugam, N., and P.U. Mahalingam. 2015. "Lignocellulose Plant Biomass ; an Emerging Alternative Fuel Resource." *Everyman's Science* XLIX (5): 291–95.

Arumugaperumal, Arun, Sayan Paul, Saranya Lathakumari, Ravindran Balasubramani, and Sudhakar Sivasubramaniam. 2020. "The Draft Genome of a New Verminephrobacter Eiseniae Strain: A Nephridial Symbiont of Earthworms." *Annals of Microbiology* 70 (1): 3. https://doi.org/10.1186/s13213-020-01549-w.

Asghar, Ali, Rastegari Ajar, Nath Yadav, and Arti Gupta. 2019. “Prospects of Renewable Bioprocessing in Future Energy Systems.” Edited by Ali Asghar Rastegari, Ajar Nath Yadav, and Arti Gupta. Vol. 10. Biofuel and Biorefinery Technologies. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-14463-0.

Asgher, Muhammad, Zanib Ahmad, and Hafiz Muhammad Nasir Iqbal. 2013. “Alkali and Enzymatic Delignification of Sugarcane Bagasse to Expose Cellulose Polymers for Saccharification and Bio-Ethanol Production.” *Industrial Crops and Products* 44 (January): 488–95. https://doi.org/10.1016/j.indcrop.2012.10.005.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology.” *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Asnicar, Francesco, George Weingart, Timothy L. Tickle, Curtis Huttenhower, and Nicola Segata. 2015. “Compact Graphical Representation of Phylogenetic Data and Metadata with GraPhlAn.” *PeerJ* 2015 (6): e1029. https://doi.org/10.7717/peerj.1029.

Aylward, Frank O., Kristin E. Burnum, Jarrod J. Scott, Garret Suen, Susannah G. Tringe, Sandra M. Adams, Kerrie W. Barry, et al. 2012. “Metagenomic and Metaproteomic Insights into Bacterial Communities in Leaf-Cutter Ant Fungus Gardens.” *ISME Journal* 6 (9): 1688–1701. https://doi.org/10.1038/ismej.2012.10.

Aylward, Frank O., Garret Suen, Peter H.W. Biedermann, Aaron S. Adams, Jarrod J. Scott, Stephanie A. Malfatti, Tijana Glavina Del Rio, et al. 2014. “Convergent Bacterial Microbiotas in the Fungal Agricultural Systems of Insects.” *MBio* 5 (6). https://doi.org/10.1128/mBio.02077-14.

Bahrndorff, Simon, Nadieh De Jonge, Jacob Kjerulf Hansen, Jannik Mørk Skovgaard Lauritzen, Lasse Holt Spanggaard, Mathias Hamann Sørensen, Morten Yde, and Jeppe Lund Nielsen. 2018. “Diversity and Metabolic Potential of the Microbiota Associated with a Soil Arthropod.” *Scientific Reports* 8 (1). https://doi.org/10.1038/s41598-018-20967-0.

Balat, Mustafa. 2011. “Production of Bioethanol from Lignocellulosic Materials via the Biochemical Pathway: A Review.” *Energy Conversion and Management* 52 (2): 858–75. https://doi.org/10.1016/j.enconman.2010.08.013.

Baldeweg, Florian, Hirokazu Kage, Sebastian Schieferdecker, Caitilyn Allen, Dirk Hoffmeister, and Markus Nett. 2017. “Structure of Ralsolamycin, the Interkingdom Morphogen from the Crop Plant Pathogen Ralstonia Solanacearum.” *Organic Letters* 19 (18): 4868–71. https://doi.org/10.1021/acs.orglett.7b02329.

Baldrian, Petr, Petra Zrůstová, Vojtěch Tláskal, Anna Davidová, Věra Merhautová, and Tomáš Vrška. 2016. “Fungi Associated with Decomposing Deadwood in a Natural Beech-Dominated Forest.” *Fungal Ecology* 23 (October): 109–22. https://doi.org/10.1016/j.funeco.2016.07.001.

Bamdad, Hanieh, Kelly Hawboldt, and Stephanie MacQuarrie. 2018. “A Review on Common Adsorbents for Acid Gases Removal: Focus on Biochar.” *Renewable and Sustainable Energy Reviews*. Elsevier Ltd. https://doi.org/10.1016/j.rser.2017.05.261.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” *J. Comput. Biol.* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Baron, Sophie A., Seydina M. Diene, and Jean Marc Rolain. 2018. “Human Microbiomes and Antibiotic Resistance.” *Human Microbiome Journal*. Elsevier Ltd. https://doi.org/10.1016/j.humic.2018.08.005.

Baumann, Ivan, and Peter Westermann. 2016. “Microbial Production of Short Chain Fatty Acids from Lignocellulosic Biomass: Current Processes and Market.” *BioMed Research International* 2016. https://doi.org/10.1155/2016/8469357.

Belkaid, Yasmine, and Timothy W. Hand. 2014. “Role of the Microbiota in Immunity and Inflammation.” *Cell*. Cell Press. https://doi.org/10.1016/j.cell.2014.03.011.

Benjamino, Jacquelynn, Stephen Lincoln, Ranjan Srivastava, and Joerg Graf. 2018. “Low-Abundant Bacteria Drive Compositional Changes in the Gut Microbiota after Dietary Alteration.” *Microbiome* 6 (1): 86. https://doi.org/10.1186/s40168-018-0469-5.

Benndorf, René, Huijuan Guo, Elisabeth Sommerwerk, Christiane Weigel, Maria Garcia-Altares, Karin Martin, Haofu Hu, et al. 2018. “Natural Products from Actinobacteria Associated with Fungus-Growing Termites.” *Antibiotics* 7 (3): 1–25. https://doi.org/10.3390/antibiotics7030083.

Bentley, Stephen D., David M. Aanensen, Angeliki Mavroidi, David Saunders, Ester Rabbinowitsch, Matthew Collins, Kathy Donohoe, et al. 2006. “Genetic Analysis of the Capsular Biosynthetic Locus from All 90 Pneumococcal Serotypes.” *PLoS Genetics* 2 (3): 0262–69. https://doi.org/10.1371/journal.pgen.0020031.

Berg, Matty P., Mirjam Stoffer, and Harry H. Van Den Heuvel. 2004. “Feeding Guilds in Collembola Based on Digestive Enzymes.” In *Pedobiologia*, 48:589–601. Elsevier GmbH. https://doi.org/10.1016/j.pedobi.2004.07.006.

Berg, Maureen, Ben Stenuit, Joshua Ho, Andrew Wang, Caitlin Parke, Matthew Knight, Lisa Alvarez-

Cohen, and Michael Shapira. 2016. “Assembly of the Caenorhabditis Elegans Gut Microbiota from Diverse Soil Microbial Environments.” *ISME J.* 10 (8): 1998–2009. https://doi.org/10.1038/ismej.2015.253.

Berihulay, Haile, Adam Abied, Xiaohong He, Lin Jiang, and Yuehui Ma. 2019. “Adaptation Mechanisms of Small Ruminants to Environmental Heat Stress.” *Animals* 9 (3): 1–9. https://doi.org/10.3390/ani9030075.

Berlanga, Mercedes, Carlos Llorens, Jaume Comas, and Ricardo Guerrero. 2016. “Gut Bacterial Community of the Xylophagous Cockroaches Cryptocercus Punctulatus and Parasphaeria Boleiriana.” *PLoS ONE* 11 (4). https://doi.org/10.1371/journal.pone.0152400.

Bernard, Thomas, Brandi I. L. Cantarel, Bernard Henrissat, Vincent Lombard, Pedro M. Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, and Bernard Henrissat. 2008. “The Carbohydrate-Active EnZymes Database (CAZy): An Expert Resource for Glycogenomics.” *Nucleic Acids Research* 37 (SUPPL. 1): D233–38. https://doi.org/10.1093/nar/gkn663.

Biely, Peter. 2012. “Microbial Carbohydrate Esterases Deacetylating Plant Polysaccharides.” *Biotechnology Advances*. Elsevier. https://doi.org/10.1016/j.biotechadv.2012.04.010.

Blanco, Paula, Sara Hernando-Amado, Jose Reales-Calderon, Fernando Corona, Felipe Lira, Manuel Alcalde-Rico, Alejandra Bernardini, Maria Sanchez, and Jose Martinez. 2016. “Bacterial Multidrug Efflux Pumps: Much More Than Antibiotic Resistance Determinants.” *Microorganisms* 4 (1): 14. https://doi.org/10.3390/microorganisms4010014.

Blin, Kai, Simon Shaw, Katharina Steinke, Rasmus Villebro, Nadine Ziemert, Sang Yup Lee, Marnix H. Medema, and Tilmann Weber. 2019. “AntiSMASH 5.0: Updates to the Secondary Metabolite Genome Mining Pipeline.” *Nucleic Acids Research* 47 (W1): W81–87. https://doi.org/10.1093/nar/gkz310.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Bonneau, Anne, Béatrice Roche, and Isabelle J. Schalk. 2020. “Iron Acquisition in Pseudomonas Aeruginosa by the Siderophore Pyoverdine: An Intricate Interacting Network Including Periplasmic and Membrane Proteins.” *Scientific Reports* 10 (1): 1–11. https://doi.org/10.1038/s41598-019-56913-x.

Bonowski, Felix, Ana Kitanovic, Peter Ruoff, Jinda Holzwarth, Igor Kitanovic, van Ngoc Bui, Elke Lederer, and Stefan Wölfl. 2010. “Computer Controlled Automated Assay for Comprehensive

Studies of Enzyme Kinetic Parameters." Edited by Vladimir Brusic. *PLoS ONE* 5 (5): e10727. https://doi.org/10.1371/journal.pone.0010727.

Boonmee, Atcha. 2012. "Hydrolysis of Various Thai Agricultural Biomasses Using the Crude Enzyme from Aspergillus Aculeatus Iizuka FR60 Isolated from Soil." *Brazilian Journal of Microbiology* 43 (2): 456–66. https://doi.org/10.1590/S1517-83822012000200005.

Borne, Romain, Edward A. Bayer, Sandrine Pagès, Stéphanie Perret, and Henri Pierre Fierobe. 2013. "Unraveling Enzyme Discrimination during Cellulosome Assembly Independent of Cohesin - Dockerin Affinity." *FEBS Journal* 280 (22): 5764–79. https://doi.org/10.1111/febs.12497.

Bornscheuer, Uwe, Klaus Buchholz, and Jürgen Seibel. 2014. "Enzymatic Degradation of (Ligno)Cellulose." *Angewandte Chemie - International Edition* 53 (41): 10876–93. https://doi.org/10.1002/anie.201309953.

Bost, Alyssa, Vincent G. Martinson, Soeren Franzenburg, Karen L. Adair, Alice Albasi, Martin T. Wells, and Angela E. Douglas. 2018. "Functional Variation in the Gut Microbiome of Wild Drosophila Populations." *Molecular Ecology* 27 (13): 2834–45. https://doi.org/10.1111/mec.14728.

Bouchon, Didier, Martin Zimmer, and Jessica Dittmer. 2016. "The Terrestrial Isopod Microbiome: An All-in-One Toolbox for Animal-Microbe Interactions of Ecological Relevance." *Frontiers in Microbiology* 7 (SEP). https://doi.org/10.3389/fmicb.2016.01472.

Bourguignon, Thomas, Nathan Lo, Carsten Dietrich, Jan Šobotník, Sarah Sidek, Yves Roisin, Andreas Brune, and Theodore A. Evans. 2018. "Rampant Host Switching Shaped the Termite Gut Microbiome." *Current Biology* 28 (4): 649-654.e2. https://doi.org/10.1016/j.cub.2018.01.035.

Bredon, Marius, Jessica Dittmer, Cyril Noël, Bouziane Moumen, and Didier Bouchon. 2018. "Lignocellulose Degradation at the Holobiont Level: Teamwork in a Keystone Soil Invertebrate 06 Biological Sciences 0605 Microbiology." *Microbiome* 6 (1): 1–19. https://doi.org/10.1186/s40168-018-0536-y.

Bredon, Marius, Benjamin Herran, Joanne Bertaux, Pierre Grève, Bouziane Moumen, and Didier Bouchon. 2020. "Isopod Holobionts as Promising Models for Lignocellulose Degradation." *Biotechnology for Biofuels* 13 (1). https://doi.org/10.1186/s13068-020-01683-2.

Breton, Christelle, Lenka Šnajdrová, Charlotte Jeanneau, Jaroslav Koča, and Anne Imberty. 2006. "Structures and Mechanisms of Glycosyltransferases." *Glycobiology* 16 (2): 29R-37R. https://doi.org/10.1093/glycob/cwj016.

Breznak, John A, and Andreas Brune. 2002. "Role of Microorganisms in the Digestion of

Lignocellulose by Termites." *Annual Review of Entomology* 39 (1): 453–87. https://doi.org/10.1146/annurev.ento.39.1.453.

Broza, Meir, Roberto M. Pereira, and Jerry L. Stimac. 2001. "The Nonsusceptibility of Soil Collembola to Insect Pathogens and Their Potential as Scavengers of Microbial Pesticides." *Pedobiologia* 45 (6): 523–34. https://doi.org/10.1078/0031-4056-00104.

Brune, Andreas. 2014. "Symbiotic Digestion of Lignocellulose in Termite Guts." *Nature Reviews Microbiology* 12 (3): 168–80. https://doi.org/10.1038/nrmicro3182.

Brune, Andreas, and Carsten Dietrich. 2015. "The Gut Microbiota of Termites: Digesting the Diversity in the Light of Ecology and Evolution." *Annual Review of Microbiology* 69 (1): 145–66. https://doi.org/10.1146/annurev-micro-092412-155715.

Brune, Andreas, David Emerson, and John A Breznak. 1995. "The Termite Gut Microflora as an Oxygen Sink : Microelectrode Determination of Oxygen and PH Gradients in Guts of Lower and Higher Termites . The Termite Gut Microflora as an Oxygen Sink : Microelectrode Determination of Oxygen and PH Gradients in Guts Of." *Applied and Environmental Microbiology* 61 (7): 2681–87.

Brune, Andreas, and Michael Friedrich. 2000. "Microecology of the Termite Gut: Structure and Function on a Microscale." *Current Opinion in Microbiology*. Current Biology Ltd. https://doi.org/10.1016/S1369-5274(00)00087-4.

Bushnell, Brian. 2017. "BBTools User Guide - DOE Joint Genome Institute." Joint Genome Institute. 2017. http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/.

Busk, P. K., B. Pilgaard, M. J. Lezyk, A. S. Meyer, and L. Lange. 2017. "Homology to Peptide Pattern for Annotation of Carbohydrate-Active Enzymes and Prediction of Function." *BMC Bioinformatics* 18 (1): 214. https://doi.org/10.1186/s12859-017-1625-9.

Bussi, Claudio, and Maximiliano G Gutierrez. 2019. "From Hairballs to Hypotheses–Biological Insights from Microbial Networks." *FEMS Microbiology Reviews* 43 (4): 341–61. https://doi.org/10.1093/FEMSRE.

Bystroff, Christopher, and Anders Krogh. 2008. "Hidden Markov Models for Prediction of Protein Features." In *Protein Structure Prediction*, 173–98. Humana Press. https://doi.org/10.1007/978-1-59745-574-9_7.

Callaway, Ewen. 2019. "C-Section Babies Are Missing Key Microbes." *Nature*, September. https://doi.org/10.1038/d41586-019-02807-x.

———. 2020. "'It Will Change Everything': DeepMind's AI Makes Gigantic Leap in Solving Protein

Structures." *Nature*. NLM (Medline). https://doi.org/10.1038/d41586-020-03348-4.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

Campanacci, Valérie, Russell E. Bishop, Stéphanie Blangy, Mariella Tegoni, and Christian Cambillau. 2006. "The Membrane Bound Bacterial Lipocalin Blc Is a Functional Dimer with Binding Preference for Lysophospholipids." *FEBS Letters* 580 (20): 4877–83. https://doi.org/10.1016/j.febslet.2006.07.086.

Campanacci, Valérie, Didier Nurizzo, Silvia Spinelli, Christel Valencia, Mariella Tegoni, and Christian Cambillau. 2004. "The Crystal Structure of the Escherichia Coli Lipocalin Blc Suggests a Possible Role in Phospholipid Binding." *FEBS Letters* 562 (1–3): 183–88. https://doi.org/10.1016/S0014-5793(04)00199-1.

Campanaro, Stefano, Laura Treu, Panagiotis G. Kougias, Davide De Francisci, Giorgio Valle, and Irini Angelidaki. 2016. "Metagenomic Analysis and Functional Characterization of the Biogas Microbiome Using High Throughput Shotgun Sequencing and a Novel Binning Strategy." *Biotechnology for Biofuels* 9 (1): 26. https://doi.org/10.1186/s13068-016-0441-1.

Carson, Mike, David H Johnson, Heather Mcdonald, Christie Brouillette, and Lawrence J Delucas. 2007. "Biological Crystallography His-Tag Impact on Structure." *Research Papers Acta Cryst* 63: 295–301. https://doi.org/10.1107/S0907444906052024.

Casaburi, Giorgio, Rebbeca M. Duar, Daniel P. Vance, Ryan Mitchell, Lindsey Contreras, Steven A. Frese, Jennifer T. Smilowitz, and Mark A. Underwood. 2019. "Early-Life Gut Microbiome Modulation Reduces the Abundance of Antibiotic-Resistant Bacteria." *Antimicrobial Resistance and Infection Control* 8 (1): 131. https://doi.org/10.1186/s13756-019-0583-6.

Casals-Pascual, Climent, Andrea Vergara, and Jordi Vila. 2018. "Intestinal Microbiota and Antibiotic Resistance: Perspectives and Solutions." *Human Microbiome Journal*. Elsevier Ltd. https://doi.org/10.1016/j.humic.2018.05.002.

Cassou, Emilie, Steven M. Jaffee, and Jiang Ru. 2018. *The Challenge of Agricultural Pollution: Evidence from China, Vietnam, and the Philippines*. *The Challenge of Agricultural Pollution: Evidence from China, Vietnam, and the Philippines*. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1201-9.

Charles, Hubert, Séverine Balmand, Araceli Lamelas, Ludovic Cottret, Vicente Pérez-Brocal, Béatrice Burdin, Amparo Latorre, et al. 2011. "A Genomic Reappraisal of Symbiotic Function in the

Aphid/Buchnera Symbiosis: Reduced Transporter Sets and Variable Membrane Organisations." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0029096.

Chen, Hongzhang. 2014a. *Biotechnology of Lignocellulose: Theory and Practice*. *Biotechnology of Lignocellulose: Theory and Practice*. https://doi.org/10.1007/978-94-007-6898-7.

———. 2014b. "Chemical Composition and Structure of Natural Lignocellulose." In *Biotechnology of Lignocellulose*, 25–71. Springer Netherlands. https://doi.org/10.1007/978-94-007-6898-7_2.

Chen, Xing, Christopher Karl Yost, João Marcelo Pereira Alves, Gehong Wei, Xiangchen Li, Wenjun Tong, Lina Wang, Siddiq Ur Rahman, and Shiheng Tao. 2018. "A Novel Strategy for Detecting Recent Horizontal Gene Transfer and Its Application to Rhizobium Strains." *Frontiers in Microbiology | Www.Frontiersin.Org* 1: 973. https://doi.org/10.3389/fmicb.2018.00973.

Cheng, Xin Yue, Xue Liang Tian, Yun Sheng Wang, Ren Miao Lin, Zhen Chuan Mao, Nansheng Chen, and Bing Yan Xie. 2013. "Metagenomic Analysis of the Pinewood Nematode Microbiome Reveals a Symbiotic Relationship Critical for Xenobiotics Degradation." *Scientific Reports* 3. https://doi.org/10.1038/srep01869.

Chettri, Dixita, Ashwani Kumar Verma, and Anil Kumar Verma. 2020. "Innovations in CAZyme Gene Diversity and Its Modification for Biorefinery Applications." *Biotechnology Reports* 28. https://doi.org/10.1016/j.btre.2020.e00525.

Chikindas, Michael L., Richard Weeks, Djamel Drider, Vladimir A. Chistyakov, and Leon MT Dicks. 2018. "Functions and Emerging Applications of Bacteriocins." *Current Opinion in Biotechnology* 49 (February): 23–28. https://doi.org/10.1016/j.copbio.2017.07.011.

Chitra Devi, V., S. Mothil, R. Sathish Raam, and K. Senthilkumar. 2020. "Thermochemical Conversion and Valorization of Woody Lignocellulosic Biomass in Hydrothermal Media." In , 45–63. Springer, Singapore. https://doi.org/10.1007/978-981-15-0410-5_4.

Choi, Il-Dong, Hwa-Young Kim, and Yong-Jin Choi. 2005. "Gene Cloning and Characterization of α-Glucuronidase of Bacillus Stearothermophilus No. 236." *Bioscience, Biotechnology, and Biochemistry* 64 (12): 2530–37. https://doi.org/10.1271/bbb.64.2530.

Choi, In Seong, Jae Hoon Kim, Seung Gon Wi, Kyoung Hyoun Kim, and Hyeun Jong Bae. 2013. "Bioethanol Production from Mandarin (Citrus Unshiu) Peel Waste Using Popping Pretreatment." *Applied Energy* 102 (February): 204–10. https://doi.org/10.1016/j.apenergy.2012.03.066.

Choi, J. H., and S. Y. Lee. 2004. "Secretory and Extracellular Production of Recombinant Proteins Using Escherichia Coli." *Applied Microbiology and Biotechnology* 64 (5): 625–35.

https://doi.org/10.1007/s00253-004-1559-9.

Connie, Rye, Robert Wise, Vladimir Jurukovski, Jean DeSaix, Jung Choi, and Yael Avissar. 2016. "Biology." OpenStax. 2016. https://bio.libretexts.org/Bookshelves/Introductory_and_General_Biology/Book%3A_General_Biology_(Boundless)/34%3A_Animal_Nutrition_and_the_Digestive_System/34.1%3A_Digestive_Systems/34.1A%3A_Digestive_Systems.

Corbett, David, and Ian S. Roberts. 2008. "Chapter 1 Capsular Polysaccharides in Escherichia Coli." *Advances in Applied Microbiology* 65 (08): 1–26. https://doi.org/10.1016/S0065-2164(08)00601-1.

Cousin, Fabien J., Denis D.G. Mater, Benoît Foligné, and Gwénaël Jan. 2011. "Dairy Propionibacteria as Human Probiotics: A Review of Recent Evidence." *Dairy Science and Technology*. https://doi.org/10.1051/dst/2010032.

Cuskin, Fiona, Arnaud Baslé, Simon Ladevèze, Alison M. Day, Harry J. Gilbert, Gideon J. Davies, Gabrielle Potocki-Véronèse, and Elisabeth C. Lowe. 2015. "The GH130 Family of Mannoside Phosphorylases Contains Glycoside Hydrolases That Target β-1,2-Mannosidic Linkages in Candida Mannan." *Journal of Biological Chemistry* 290 (41): 25023–33. https://doi.org/10.1074/jbc.M115.681460.

Dai, Xin, Yaxin Zhu, Yingfeng Luo, Lei Song, Di Liu, Li Liu, Furong Chen, et al. 2012. "Metagenomic Insights into the Fibrolytic Microbiome in Yak Rumen." *PLoS ONE* 7 (7). https://doi.org/10.1371/journal.pone.0040430.

Dalbey, R. E., P. Wang, and J. M. van Dijl. 2012. "Membrane Proteases in the Bacterial Protein Secretion and Quality Control Pathway." *Microbiology and Molecular Biology Reviews* 76 (2): 311–30. https://doi.org/10.1128/mmbr.05019-11.

Danchin, Etienne G.J., Marie Noëlle Rosso, Paulo Vieira, Janice De Almeida-Engler, Pedro M. Coutinho, Bernard Henrissat, and Pierre Abad. 2010. "Multiple Lateral Gene Transfers and Duplications Have Promoted Plant Parasitism Ability in Nematodes." *Proceedings of the National Academy of Sciences of the United States of America* 107 (41): 17651–56. https://doi.org/10.1073/pnas.1008486107.

Dao, Anh T.N., Sander J. Loenen, Kees Swart, Ha T.C. Dang, Abraham Brouwer, and Tjalf E. de Boer. 2021. "Characterization of 2,3,7,8-Tetrachlorodibenzo-p-Dioxin Biodegradation by Extracellular Lignin-Modifying Enzymes from Ligninolytic Fungus." *Chemosphere* 263 (January): 128280. https://doi.org/10.1016/j.chemosphere.2020.128280.

Das, Saprativ P, Rajeev Ravindran, Shadab Ahmed, Debasish Das, Dinesh Goyal, Carlos M G A Fontes, Arun Goyal, D Goyal, and C M G A Fontes. 2012. "Bioethanol Production Involving Recombinant C. Thermocellum Hydrolytic Hemicellulase and Fermentative Microbes." *Appl Biochem Biotechnol* 167: 1475–88. https://doi.org/10.1007/s12010-012-9618-7.

Denman, Stuart E., Gonzalo Martinez Fernandez, Takumi Shinkai, Makoto Mitsumori, and Christopher S. McSweeney. 2015. "Metagenomic Analysis of the Rumen Microbial Community Following Inhibition of Methane Formation by a Halogenated Methane Analog." *Frontiers in Microbiology* 6 (OCT): 1087. https://doi.org/10.3389/fmicb.2015.01087.

Dijk, Marlous van, Ignis Trollmann, Margarete Alice Fontes Saraiva, Rogelio Lopes Brandão, Lisbeth Olsson, and Yvonne Nygård. 2020. "Small Scale Screening of Yeast Strains Enables High-Throughput Evaluation of Performance in Lignocellulose Hydrolysates." *Bioresource Technology Reports* 11 (September): 100532. https://doi.org/10.1016/j.biteb.2020.100532.

Dik, David A., Daniel R. Marous, Jed F. Fisher, and Shahriar Mobashery. 2017. "Lytic Transglycosylases: Concinnity in Concision of the Bacterial Cell Wall." *Critical Reviews in Biochemistry and Molecular Biology*. Taylor and Francis Ltd. https://doi.org/10.1080/10409238.2017.1337705.

Dimarogona, M., and E. Topakas. 2016. *Regulation and Heterologous Expression of Lignocellulosic Enzymes in Aspergillus*. *New and Future Developments in Microbial Biotechnology and Bioengineering: Aspergillus System Properties and Applications*. Elsevier B.V. https://doi.org/10.1016/B978-0-444-63505-1.00012-9.

Ding, Shi You, Qi Xu, Michael Crowley, Yining Zeng, Mark Nimlos, Raphael Lamed, Edward A. Bayer, and Michael E. Himmel. 2008. "A Biophysical Perspective on the Cellulosome: New Opportunities for Biomass Conversion." *Current Opinion in Biotechnology* 19 (3): 218–27. https://doi.org/10.1016/j.copbio.2008.04.008.

Dittmer, Jessica, Jérôme Lesobre, Bouziane Moumen, and Didier Bouchon. 2016. "Host Origin and Tissue Microhabitat Shaping the Microbiota of the Terrestrial Isopod Armadillidium Vulgare." *FEMS Microbiology Ecology* 92 (5). https://doi.org/10.1093/femsec/fiw063.

Do, Thi Huyen, Trong Khoa Dao, Khanh Hoang Viet Nguyen, Ngoc Giang Le, Thi Mai Phuong Nguyen, Tung Lam Le, Thu Nguyet Phung, Nico M. van Straalen, Dick Roelofs, and Nam Hai Truong. 2018. "Metagenomic Analysis of Bacterial Community Structure and Diversity of Lignocellulolytic Bacteria in Vietnamese Native Goat Rumen." *Asian-Australasian Journal of Animal Sciences* 31 (5): 738–47. https://doi.org/10.5713/ajas.17.0174.

Do, Thi Huyen, Ngoc Giang Le, Trong Khoa Dao, Thi Mai Phuong Nguyen, Tung Lam Le, Han Ly

Luu, Khanh Hoang Viet Nguyen, et al. 2018. “Metagenomic Insights into Lignocellulose-Degrading Genes through Illumina Based de Novo Sequencing of the Microbiome in Vietnamese Native Goats’ Rumen.” *Journal of General and Applied Microbiology* 64 (3): 108–16. https://doi.org/10.2323/jgam.2017.08.004.

Do, Thi Huyen, Thi Thao Nguyen, Thanh Ngoc Nguyen, Quynh Giang Le, Cuong Nguyen, Keitarou Kimura, and Nam Hai Truong. 2014. “Mining Biomass-Degrading Genes through Illumina-Based de Novo Sequencing and Metagenomic Analysis of Free-Living Bacteria in the Gut of the Lower Termite Coptotermes Gestroi Harvested in Vietnam.” *Journal of Bioscience and Bioengineering* 118 (6): 665–71. https://doi.org/10.1016/j.jbiosc.2014.05.010.

Dodd, Dylan, and Isacc K. O. Cann. 2009. “Enzymatic Deconstruction of Xylan for Biofuel Production.” *GCB Bioenergy* 1 (1): 2–17. https://doi.org/10.1111/j.1757-1707.2009.01004.x.

Dodd, Dylan, Young Hwan Moon, Kankshita Swaminathan, Roderick I. Mackie, and Isaac K.O. Cann. 2010. “Transcriptomic Analyses of Xylan Degradation by Prevotella Bryantii and Insights into Energy Acquisition by Xylanolytic Bacteroidetes.” *Journal of Biological Chemistry* 285 (39): 30261–73. https://doi.org/10.1074/jbc.M110.141788.

Donia, Mohamed S., Jacques Ravel, and Eric W. Schmidt. 2008. “A Global Assembly Line for Cyanobactins.” *Nature Chemical Biology* 4 (6): 341–43. https://doi.org/10.1038/nchembio.84.

Dou, Tong Yi, Hong Wei Luan, Guang Bo Ge, Ming Ming Dong, Han Fa Zou, Yu Qi He, Pan Cui, et al. 2015. “Functional and Structural Properties of a Novel Cellulosome-like Multienzyme Complex: Efficient Glycoside Hydrolysis of Water-Insoluble 7-Xylosyl-10-Deacetylpaclitaxel.” *Scientific Reports* 5 (September). https://doi.org/10.1038/srep13768.

Douglas, Angela E. 2016. “How Multi-Partner Endosymbioses Function.” *Nature Reviews Microbiology* 14 (12): 731–43. https://doi.org/10.1038/nrmicro.2016.151.

———. 2019. “Simple Animal Models for Microbiome Research.” *Nature Reviews Microbiology*. Nature Publishing Group. https://doi.org/10.1038/s41579-019-0242-1.

Dromph, Karsten M., and Susanne Vestergaard. 2002. “Pathogenicity and Attractiveness of Entomopathogenic Hyphomycete Fungi to Collembolans.” *Applied Soil Ecology* 21 (3): 197–210. https://doi.org/10.1016/S0929-1393(02)00092-6.

Duhamel, Marie, Roel Pel, Astra Ooms, Heike Bucking, Jan Jansa, Jacintha Ellers, Nico M. Van Straalen, Tjalf Wouda, Philippe Vandenkoornhuyse, and E. Toby Kiers. 2013. “Do Fungivores Trigger the Transfer of Protective Metabolites from Host Plants to Arbuscular Mycorrhizal Hyphae?” *Ecology* 94 (9): 2019–29. https://doi.org/10.1890/12-1943.1.

Eddy, Sean R. 1998. “Profile Hidden Markov Models.” *Bioinformatics* 14 (9): 755–63. https://doi.org/10.1093/bioinformatics/14.9.755.

Engel, Philipp, Vincent G. Martinson, and Nancy A. Moran. 2012. “Functional Diversity within the Simple Gut Microbiota of the Honey Bee.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (27): 11002–7. https://doi.org/10.1073/pnas.1202970109.

Engel, Philipp, and Nancy A. Moran. 2013. “The Gut Microbiota of Insects - Diversity in Structure and Function.” *FEMS Microbiology Reviews* 37 (5): 699–735. https://doi.org/10.1111/1574-6976.12025.

Engelhardt, Kerstin, Kristin F. Degnes, and Sergey B. Zotchev. 2010. “Isolation and Characterization of the Gene Cluster for Biosynthesis of the Thiopeptide Antibiotic TP-1161.” *Applied and Environmental Microbiology* 76 (21): 7093–7101. https://doi.org/10.1128/AEM.01442-10.

Esposti, Mauro Degli, and Esperanza Martinez Romero. 2017. “The Functional Microbiome of Arthropods.” *PLoS ONE* 12 (5). https://doi.org/10.1371/journal.pone.0176573.

Eyun, Seong Il, Haichuan Wang, Yannick Pauchet, Richard H. Ffrench-Constant, Andrew K. Benson, Arnubio Valencia-Jiménez, Etsuko N. Moriyama, and Blair D. Siegfried. 2014. “Molecular Evolution of Glycoside Hydrolase Genes in the Western Corn Rootworm (Diabrotica Virgifera Virgifera).” *PLoS ONE* 9 (4). https://doi.org/10.1371/journal.pone.0094052.

Faddeeva-Vakhrusheva, Anna, Martijn F.L. Derks, Seyed Yahya Anvar, Valeria Agamennone, Wouter Suring, Sandra Smit, Nico M. van Straalen, and Dick Roelofs. 2016. “Gene Family Evolution Reflects Adaptation to Soil Environmental Stressors in the Genome of the Collembolan Orchesella Cincta.” *Genome Biology and Evolution* 8 (7): 2106–17. https://doi.org/10.1093/gbe/evw134.

Faddeeva-Vakhrusheva, Anna, Ken Kraaijeveld, Martijn F.L. Derks, Seyed Yahya Anvar, Valeria Agamennone, Wouter Suring, Andries A. Kampfraath, et al. 2017. “Coping with Living in the Soil: The Genome of the Parthenogenetic Springtail Folsomia Candida.” *BMC Genomics* 18 (1). https://doi.org/10.1186/s12864-017-3852-x.

Fehér, Csaba. 2018. “Novel Approaches for Biotechnological Production and Application of L-Arabinose.” *Journal of Carbohydrate Chemistry*. https://doi.org/10.1080/07328303.2018.1491049.

Fernando, Sandun, Sushil Adhikari, Chauda Chandrapal, and Naveen Murali. 2006. “Biorefineries: Current Status, Challenges, and Future Direction.” *Energy and Fuels*. American Chemical Society . https://doi.org/10.1021/ef060097w.

Fierer, Noah. 2017. “Embracing the Unknown: Disentangling the Complexities of the Soil Microbiome.” *Nature Reviews Microbiology*. Nature Publishing Group. https://doi.org/10.1038/nrmicro.2017.87.

Fitzpatrick, David, and Fiona Walsh. 2016. “Antibiotic Resistance Genes across a Wide Variety of Metagenomes.” *FEMS Microbiology Ecology* 92 (2): 1–8. https://doi.org/10.1093/femsec/fiv168.

Flint, Harry J, and Edward A Bayer. 2008. “Plant Cell Wall Breakdown by Anaerobic Microorganisms from the Mammalian Digestive Tract.” In *Annals of the New York Academy of Sciences*, 1125:280–88. https://doi.org/10.1196/annals.1419.022.

Flórez, Laura V., Kirstin Scherlach, Paul Gaube, Claudia Ross, Elisabeth Sitte, Cornelia Hermes, Andre Rodrigues, Christian Hertweck, and Martin Kaltenpoth. 2017. “Antibiotic-Producing Symbionts Dynamically Transition between Plant Pathogenicity and Insect-Defensive Mutualism.” *Nature Communications* 8. https://doi.org/10.1038/ncomms15172.

Flot, Jean François, Boris Hespeels, Xiang Li, Benjamin Noel, Irina Arkhipova, Etienne G.J. Danchin, Andreas Hejnol, et al. 2013. “Genomic Evidence for Ameiotic Evolution in the Bdelloid Rotifer Adineta Vaga.” *Nature* 500 (7463): 453–57. https://doi.org/10.1038/nature12326.

Foreman, Pamela K., Doug Brown, Lydia Dankmeyer, Ralph Dean, Stephen Diener, Nigel S. Dunn-Coleman, Frits Goedegebuur, et al. 2003. “Transcriptional Regulation of Biomass-Degrading Enzymes in the Filamentous Fungus Trichoderma Reesei.” *Journal of Biological Chemistry* 278 (34): 31988–97. https://doi.org/10.1074/jbc.M304750200.

Fountain, Michelle T., and Steve P. Hopkin. 2005. “Folsomia Candida (Collembola): A ‘Standard’ Soil Arthropod.” *Annual Review of Entomology* 50 (1): 201–22. https://doi.org/10.1146/annurev.ento.50.071803.130331.

Freudl, Roland. 2018. “Signal Peptides for Recombinant Protein Secretion in Bacterial Expression Systems.” *Microbial Cell Factories* 17 (1): 1–10. https://doi.org/10.1186/s12934-018-0901-3.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics* 28 (23): 3150–52. https://doi.org/10.1093/bioinformatics/bts565.

Fujimoto, Zui, Hitomi Ichinose, Tomoko Maehara, Mariko Honda, Motomitsu Kitaoka, and Satoshi Kaneko. 2010. “Crystal Structure of an Exo-1 , 5- α - L -Arabinofuranosidase from Streptomyces Avermitilis Provides Insights into the Mechanism of Substrate Discrimination between Exo- and Endo-Type Enzymes in Glycoside Hydrolase Family 43 *.” *The Journal of Biological Chemistry*

285 (44): 34134–43. https://doi.org/10.1074/jbc.M110.164251.

Furlong, Sarah E, and Sarah Ellen Furlong. 2013. “Structure-Function Analysis of UDP-Sugar : Polyisoprenyl Phosphate Sugar-1-Phosphate Transferases,” no. September.

Gaitán-Hernández, Rigoberto, Norberto Cortés, and Gerardo Mata. 2014. “Improvement of Yield of the Edible and Medicinal Mushroom Lentinula Edodes on Wheat Straw by Use of Supplemented Spawn.” www.sbmicrobiologia.org.br.

Gálvez, Antonio, Hikmate Abriouel, Rosario Lucas López, and Nabil Ben Omar. 2007. “Bacteriocin-Based Strategies for Food Biopreservation.” *International Journal of Food Microbiology* 120 (1–2): 51–70. https://doi.org/10.1016/j.ijfoodmicro.2007.06.001.

Gao, Dahai, Carolyn Haarmeyer, Venkatesh Balan, Timothy A. Whitehead, Bruce E. Dale, and Shishir P.S. Chundawat. 2014. “Lignin Triggers Irreversible Cellulase Loss during Pretreated Lignocellulosic Biomass Saccharification.” *Biotechnology for Biofuels* 7 (1): 175. https://doi.org/10.1186/s13068-014-0175-x.

Garmendia, L., A. Hernandez, M. B. Sanchez, and J. L. Martinez. 2012. “Metagenomics and Antibiotics.” *Clinical Microbiology and Infection* 18 (SUPPL. 4): 27–31. https://doi.org/10.1111/j.1469-0691.2012.03868.x.

Garron, Marie Line, and Bernard Henrissat. 2019. “The Continuing Expansion of CAZymes and Their Families.” *Current Opinion in Chemical Biology*. Elsevier Ltd. https://doi.org/10.1016/j.cbpa.2019.08.004.

Geng, Alei, Yanbing Cheng, Yongli Wang, Daochen Zhu, Yilin Le, Jian Wu, Rongrong Xie, Joshua S. Yuan, and Jianzhong Sun. 2018. “Transcriptome Analysis of the Digestive System of a Wood-Feeding Termite (Coptotermes Formosanus) Revealed a Unique Mechanism for Effective Biomass Degradation.” *Biotechnology for Biofuels* 11 (1): 24. https://doi.org/10.1186/s13068-018-1015-1.

Geraylou, Zahra, Caroline Souffreau, Eugene Rurangwa, Gregory E. Maes, Katina I. Spanier, Christophe M. Courtin, Jan A. Delcour, Johan Buyse, and Frans Ollevier. 2013. “Prebiotic Effects of Arabinoxylan Oligosaccharides on Juvenile Siberian Sturgeon (Acipenser Baerii) with Emphasis on the Modulation of the Gut Microbiota Using 454 Pyrosequencing.” *FEMS Microbiology Ecology* 86 (2): 357–71. https://doi.org/10.1111/1574-6941.12169.

Gírio, F M, C Fonseca, F Carvalheiro, L C Duarte, S Marques, and R Bogel-Łukasik. 2010. “Hemicelluloses for Fuel Ethanol: A Review.” *Bioresource Technology*. https://doi.org/10.1016/j.biortech.2010.01.088.

Gladyshev, Eugene A., Matthew Meselson, and Irina R. Arkhipova. 2008. “Massive Horizontal Gene Transfer in Bdelloid Rotifers.” *Science* 320 (5880): 1210–13. https://doi.org/10.1126/science.1156407.

Golan, Gali, Dalia Shallom, Anna Teplitsky, Galia Zaide, Smadar Shulami, Timor Baasov, Vivian Stojanoff, Andy Thompson, Yuval Shoham, and Gil Shoham. 2004. “Crystal Structures of Geobacillus Stearothermophilus α-Glucuronidase Complexed with Its Substrate and Products.” *Journal of Biological Chemistry* 279 (4): 3014–24. https://doi.org/10.1074/jbc.M310098200.

Golomb, Benjamin L., and Maria L. Marco. 2015. “Lactococcus Lactis Metabolism and Gene Expression during Growth on Plant Tissues.” *Journal of Bacteriology* 197 (2): 371–81. https://doi.org/10.1128/jb.02193-14.

Gontang, Erin A., Frank O. Aylward, Camila Carlos, Tijana Glavina Del Rio, Mansi Chovatia, Alison Fern, Chien Chi Lo, et al. 2017. “Major Changes in Microbial Diversity and Community Composition across Gut Sections of a Juvenile Panchlora Cockroach.” *PLoS ONE* 12 (5). https://doi.org/10.1371/journal.pone.0177189.

Götz, Stefan, Juan Miguel García-Gómez, Javier Terol, Tim D. Williams, Shivashankar H. Nagaraj, María José Nueda, Montserrat Robles, Manuel Talón, Joaquín Dopazo, and Ana Conesa. 2008. “High-Throughput Functional Annotation and Data Mining with the Blast2GO Suite.” *Nucleic Acids Research* 36 (10): 3420–35. https://doi.org/10.1093/nar/gkn176.

Govic, Yohann Le, Nicolas Papon, Solène Le Gal, Jean Philippe Bouchara, and Patrick Vandeputte. 2019. “Non-Ribosomal Peptide Synthetase Gene Clusters in the Human Pathogenic Fungus Scedosporium Apiospermum.” *Frontiers in Microbiology* 10. https://doi.org/10.3389/fmicb.2019.02062.

Grbić, Miodrag, Thomas Van Leeuwen, Richard M. Clark, Stephane Rombauts, Pierre Rouzé, Vojislava Grbić, Edward J. Osborne, et al. 2011. “The Genome of Tetranychus Urticae Reveals Herbivorous Pest Adaptations.” *Nature* 479 (7374): 487–92. https://doi.org/10.1038/nature10640.

Grieco, Maria B., Fabyano A.C. Lopes, Louisi S. Oliveira, Diogo A. Tschoeke, Claudia C. Popov, Cristiane C. Thompson, Luna C. Gonçalves, et al. 2019. “Metagenomic Analysis of the Whole Gut Microbiota in Brazilian Termitidae Termites Cornitermes Cumulans, Cyrilliotermes Strictinasus, Syntermes Dirus, Nasutitermes Jaraguae, Nasutitermes Aquilinus, Grigiotermes Bequaerti, and Orthognathotermes Mirim.” *Current Microbiology* 76 (6): 687–97. https://doi.org/10.1007/s00284-019-01662-3.

Gu, Yilin, Yi Nan Ma, Jing Wang, Zhenyuan Xia, and Hai Lei Wei. 2020. “Genomic Insights into a

Plant Growth-Promoting Pseudomonas Koreensis Strain with Cyclic Lipopeptide-Mediated Antifungal Activity." *MicrobiologyOpen*, June. https://doi.org/10.1002/mbo3.1092.

Guillén, Daniel, Sergio Sánchez, and Romina Rodríguez-Sanoja. 2010. "Carbohydrate-Binding Domains: Multiplicity of Biological Roles." *Applied Microbiology and Biotechnology* 85 (5): 1241–49. https://doi.org/10.1007/s00253-009-2331-y.

Güllert, Simon, Martin A. Fischer, Dmitrij Turaev, Britta Noebauer, Nele Ilmberger, Bernd Wemheuer, Malik Alawi, et al. 2016. "Deep Metagenome and Metatranscriptome Analyses of Microbial Communities Affiliated with an Industrial Biogas Fermenter, a Cow Rumen, and Elephant Feces Reveal Major Differences in Carbohydrate Hydrolysis Strategies." *Biotechnology for Biofuels* 9 (1): 121. https://doi.org/10.1186/s13068-016-0534-x.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. "QUAST: Quality Assessment Tool for Genome Assemblies." *Bioinformatics* 29 (8): 1072–75. https://doi.org/10.1093/bioinformatics/btt086.

Guzmán, Esther A., Kelly Maers, Jill Roberts, Hilaire V. Kemami-Wangun, Dedra Harmody, and Amy E. Wright. 2015. "The Marine Natural Product Microsclerodermin A Is a Novel Inhibitor of the Nuclear Factor Kappa B and Induces Apoptosis in Pancreatic Cancer Cells." *Investigational New Drugs* 33 (1): 86–94. https://doi.org/10.1007/s10637-014-0185-3.

Han, Xufeng, Yuxin Yang, Hailong Yan, Xiaolong Wang, Lei Qu, and Yulin Chen. 2015. "Rumen Bacterial Diversity of 80 to 110-Day- Old Goats Using 16s RRNA Sequencing." *PLoS ONE* 10 (2): e0117811. https://doi.org/10.1371/journal.pone.0117811.

Handelsman, Jo. 2004. "Metagenomics: Application of Genomics to Uncultured Microorganisms." *Microbiology and Molecular Biology Reviews* 68 (4): 669–85. https://doi.org/10.1128/mmbr.68.4.669-685.2004.

Hanna, Andrea, Michael Berg, Valerie Stout, and Anneta Razatos. 2003. "Role of Capsular Colanic Acid in Adhesion of Uropathogenic Escherichia Coli." *Applied and Environmental Microbiology* 69 (8): 4474–81. https://doi.org/10.1128/AEM.69.8.4474-4481.2003.

Haro-Moreno, Jose M., Mario López-Pérez, and Francisco Rodríguez-Valera. 2020. "Long Read Metagenomics, the next Step?" *BioRxiv*. bioRxiv. https://doi.org/10.1101/2020.11.11.378109.

He, Shaomei, Natalia Ivanova, Edward Kirton, Martin Allgaier, Claudia Bergin, Rudolf H. Scheffrahn, Nikos C. Kyrpides, Falk Warnecke, Susannah G. Tringe, and Philip Hugenholtz. 2013. "Comparative Metagenomic and Metatranscriptomic Analysis of Hindgut Paunch Microbiota in Wood- and Dung-Feeding Higher Termites." *PLoS ONE*.

https://doi.org/10.1371/journal.pone.0061126.

Helbert, William, Laurent Poulet, Sophie Drouillard, Sophie Mathieu, Mélanie Loiodice, Marie Couturier, Vincent Lombard, et al. 2019. “Discovery of Novel Carbohydrate-Active Enzymes through the Rational Exploration of the Protein Sequences Space.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (13): 6063–68. https://doi.org/10.1073/pnas.1815791116.

Henderson, Gemma, Faith Cox, Sandra Kittelmann, Vahideh Heidarian Miri, Michael Zethof, Samantha J. Noel, Garry C. Waghorn, and Peter H. Janssen. 2013. “Effect of DNA Extraction Methods and Sampling Techniques on the Apparent Structure of Cow and Sheep Rumen Microbial Communities.” Edited by Stefan Bertilsson. *PloS One* 8 (9): e74787. https://doi.org/10.1371/journal.pone.0074787.

Hess, Matthias, Alexander Sczyrba, Rob Egan, Tae Wan Kim, Harshal Chokhawala, Gary Schroth, Shujun Luo, et al. 2011. “Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen.” *Science* 331 (6016): 463–67. https://doi.org/10.1126/science.1200387.

Heyer, Robert, Kay Schallert, Roman Zoun, Beatrice Becher, Gunter Saake, and Dirk Benndorf. 2017. “Challenges and Perspectives of Metaproteomic Data Analysis.” *Journal of Biotechnology*. Elsevier B.V. https://doi.org/10.1016/j.jbiotec.2017.06.1201.

Hirano, Katsuaki, Satoshi Nihei, Hiroki Hasegawa, Mitsuru Haruki, and Nobutaka Hirano. 2015. “Stoichiometric Assembly of the Cellulosome Generates Maximum Synergy for the Degradation of Crystalline Cellulose, as Revealed by in Vitro Reconstitution of the Clostridium Thermocellum Cellulosome.” *Applied and Environmental Microbiology* 81 (14): 4756–66. https://doi.org/10.1128/AEM.00772-15.

Hoffman, Stacey B. 2001. “Mechanisms of Antibiotic Resistance.” *Compendium on Continuing Education for the Practicing Veterinarian* 23 (5): 464–72. https://doi.org/10.1128/microbiolspec.vmbf-0016-2015.

Holden, Victoria I., and Michael A. Bachman. 2015. “Diverging Roles of Bacterial Siderophores during Infection.” *Metallomics* 7 (6): 986–95. https://doi.org/10.1039/c4mt00333k.

Hong, Yaoqin, Michael A. Liu, and Peter R. Reeves. 2018. “Progress in Our Understanding of Wzx Flippase for Translocation of Bacterial Membrane Lipid-Linked Oligosaccharide.” *Journal of Bacteriology* 200 (1): 1–14. https://doi.org/10.1128/JB.00154-17.

Hongoh, Yuichi. 2011. “Toward the Functional Analysis of Uncultivable, Symbiotic Microorganisms in the Termite Gut.” *Cellular and Molecular Life Sciences* 68 (8): 1311–25.

https://doi.org/10.1007/s00018-011-0648-z.

Horn, Svein Jarle, Gustav Vaaje-Kolstad, Bjørge Westereng, and Vincent G.H. Eijsink. 2012. "Novel Enzymes for the Degradation of Cellulose." *Biotechnology for Biofuels*. BioMed Central. https://doi.org/10.1186/1754-6834-5-45.

Hu, Jinguang, Valdeir Arantes, Amadeus Pribowo, and Jack N. Saddler. 2013. "The Synergistic Action of Accessory Enzymes Enhances the Hydrolytic Potential of a 'Cellulase Mixture' but Is Highly Substrate Specific." *Biotechnology for Biofuels* 6 (1): 1–12. https://doi.org/10.1186/1754-6834-6-112.

Huang, Jinling. 2013. "Horizontal Gene Transfer in Eukaryotes: The Weak-Link Model." *BioEssays* 35 (10): 868–75. https://doi.org/10.1002/bies.201300007.

Huang, Xing Feng, Matthew G. Bakker, Timothy M. Judd, Kenneth F. Reardon, and Jorge M. Vivanco. 2013. "Variations in Diversity and Richness of Gut Bacterial Communities of Termites (Reticulitermes Flavipes) Fed with Grassy and Woody Plant Substrates." *Microbial Ecology* 65 (3): 531–36. https://doi.org/10.1007/s00248-013-0219-y.

Huang, Yao Ting, Wei Yao Chuang, Bing Ching Ho, Zong Yen Wu, Rita C. Kuo, Mengwei Ko, and Po Yu Liu. 2018. "Comparative Genomics Reveals Diverse Capsular Polysaccharide Synthesis Gene Clusters in Emerging Raoultella Planticola." *Memorias Do Instituto Oswaldo Cruz* 113 (10): e180192. https://doi.org/10.1590/0074-02760180192.

Hultberg, M., T. Alsberg, S. Khalil, and B. Alsanius. 2010. "Suppression of Disease in Tomato Infected by Pythium Ultimum with a Biosurfactant Produced by Pseudomonas Koreensis." *BioControl* 55 (3): 435–44. https://doi.org/10.1007/s10526-009-9261-6.

Husnik, Filip, and John P. McCutcheon. 2018. "Functional Horizontal Gene Transfer from Bacteria to Eukaryotes." *Nature Reviews Microbiology* 16 (2): 67–79. https://doi.org/10.1038/nrmicro.2017.137.

Huson, Daniel H., Alexander F. Auch, Ji Qi, and Stephan C. Schuster. 2007. "MEGAN Analysis of Metagenomic Data." *Genome Research* 17 (3): 377–86. https://doi.org/10.1101/gr.5969107.

Husseneder, Claudia. 2010. "Comparison of the Bacterial Symbiont Composition of the Formosan Subterranean Termite from Its Native and Introduced Range." *The Open Microbiology Journal*. Vol. 4. https://doi.org/10.2174/1874285801004010053.

Hyatt, Doug, Gwo Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11. https://doi.org/10.1186/1471-2105-11-119.

Ichinose, Hitomi, Makoto Yoshida, Zui Fujimoto, and Satoshi Kaneko. 2008. “Characterization of a Modular Enzyme of Exo-1,5-α-L- Arabinofuranosidase and Arabinan Binding Module from Streptomyces Avermitilis NBRC14893.” *Applied Microbiology and Biotechnology* 80 (3): 399–408. https://doi.org/10.1007/s00253-008-1551-x.

Iqbal, Hafiz Muhammad Nasir, Godfrey Kyazze, and Tajalli Keshavarz. 2013. “Advances in the Valorization of Lignocellulosic Materials by Biotechnology: An Overview.” *BioResources* 8 (2): 3157–76. https://doi.org/10.15376/biores.8.2.3157-3176.

Irshad, Muhammad, Zahid Anwar, Hamama Islam But, Amber Afroz, Nadia Ikram, and Umer Rashid. 2013. “The Industrial Applicability of Purified Cellulase Complex Indigenously Produced by Trichoderma Viride through Solid-State Bio-Processing of Agro-Industrial and Municipal Paper Wastes.” *BioResources* 8 (1): 145–57. https://doi.org/10.15376/biores.8.1.145-157.

Janssens, Thierry K.S., Tjalf E. De Boer, Valeria Agamennone, Niels Zaagman, Nico M. Van Straalen, and Dick Roelofs. 2017. “Draft Genome Sequence of Bacillus Toyonensis Vu-Des13, Isolated from Folsomia Candida (Collembola: Entomobryidae).” *Genome Announcements* 5 (19). https://doi.org/10.1128/genomeA.00287-17.

Jia, Baofeng, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, et al. 2017. “CARD 2017: Expansion and Model-Centric Curation of the Comprehensive Antibiotic Resistance Database.” *Nucleic Acids Research* 45 (D1): D566–73. https://doi.org/10.1093/nar/gkw1004.

Jia, Xiaojing, Weibo Qiao, Wenli Tian, Xiaowei Peng, Shuofu Mi, Hong Su, and Yejun Han. 2016. “Biochemical Characterization of Extra- and Intracellular Endoxylanse from Thermophilic Bacterium Caldicellulosiruptor Kronotskyensis.” *Scientific Reports* 6 (February). https://doi.org/10.1038/srep21672.

Jiang, Daohua, Junping Fan, Xianping Wang, Yan Zhao, Bo Huang, Jianfeng Liu, and Xuejun C. Zhang. 2012. “Crystal Structure of 1,3Gal43A, an Exo-β-1,3-Galactanase from Clostridium Thermocellum.” *Journal of Structural Biology* 180 (3): 447–57. https://doi.org/10.1016/j.jsb.2012.08.005.

Jong, Ed de, Adrian Higson, Patrick Walsh, and Maria Wellisch. 2011. “Task 42 Biobased Chemicals - Value Added Products from Biorefineries.” *A Report Prepared for IEA Bioenergy-Task*, 36.

Jönsson, Leif J., and Carlos Martín. 2016. “Pretreatment of Lignocellulose: Formation of Inhibitory by-Products and Strategies for Minimizing Their Effects.” *Bioresource Technology* 199: 103–12. https://doi.org/10.1016/j.biortech.2015.10.009.

Jose, V. Lyju, Ravi P. More, Thulasi Appoothy, and A. Sha Arun. 2017. “In Depth Analysis of Rumen Microbial and Carbohydrate-Active Enzymes Profile in Indian Crossbred Cattle.” *Systematic and Applied Microbiology* 40 (3): 160–70. https://doi.org/10.1016/j.syapm.2017.02.003.

Jousset, A., S. Scheu, and M. Bonkowski. 2008. “Secondary Metabolite Production Facilitates Establishment of Rhizobacteria by Reducing Both Protozoan Predation and the Competitive Effects of Indigenous Bacteria.” *Functional Ecology* 22 (4): 714–19. https://doi.org/10.1111/j.1365-2435.2008.01411.x.

Joynson, Ryan, Leighton Pritchard, Ekenakema Osemwekha, and Natalie Ferry. 2017. “Metagenomic Analysis of the Gut Microbiome of the Common Black Slug Arion Ater in Search of Novel Lignocellulose Degrading Enzymes.” *Frontiers in Microbiology* 8 (NOV): 2181. https://doi.org/10.3389/fmicb.2017.02181.

Kaats, Gilbert R, Samuel C Keith, Patti L Keith, Robert B Leckie, Nicholas V Perricone, and Harry G Preuss. 2011. “A Combination of L-Arabinose and Chromium Lowers Circulating Glucose and Insulin Levels after an Acute Oral Sucrose Challenge.” *Nutrition Journal* 10 (1): 42. https://doi.org/10.1186/1475-2891-10-42.

Kalia, Vipin Chandra, Yogesh S. Shouche, Hemant J. Purohit, and Praveen Rahi. 2017. *Mining of Microbial Wealth and Metagenomics*. *Mining of Microbial Wealth and MetaGenomics*. https://doi.org/10.1007/978-981-10-5708-3.

Kameshwar, Ayyappa Kumar Sista, and Wensheng Qin. 2017. “Metadata Analysis of Phanerochaete Chrysosporium Gene Expression Data Identified Common CAZymes Encoding Gene Expression Profiles Involved in Cellulose and Hemicellulose Degradation.” *International Journal of Biological Sciences* 13 (1): 85–99. https://doi.org/10.7150/ijbs.17390.

Kanehisa, Minoru, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, et al. 2008. “KEGG for Linking Genomes to Life and the Environment.” *Nucleic Acids Research* 36 (SUPPL. 1). https://doi.org/10.1093/nar/gkm882.

Kanehisa, Minoru, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. 2019. “New Approach for Understanding Genome Variations in KEGG.” *Nucleic Acids Research* 47 (D1): D590–95. https://doi.org/10.1093/nar/gky962.

Kanehisa, Minoru, Yoko Sato, and Kanae Morishima. 2016. “BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences.” *Journal of Molecular Biology* 428 (4): 726–31. https://doi.org/10.1016/j.jmb.2015.11.006.

Kaneko, Satoshi, Mitsue Arimoto, Misako Ohba, Hideyuki Kobayashi, Tadashi Ishii, and Isao

Kusakabe. 1998. “Purification and Substrate Specificities of Two α-L- Arabinofuranosidases from Aspergillus Awamori IFO 4033.” *Applied and Environmental Microbiology* 64 (10): 4021–27. http://www.ncbi.nlm.nih.gov/pubmed/9758835.

Kang, Donghoon, Daniel R. Kirienkoa, Phillip Webster, Alfred L. Fisher, and Natalia V. Kirienko. 2018. “Pyoverdine, a Siderophore from Pseudomonas Aeruginosa, Translocates into C. Elegans, Removes Iron, and Activates a Distinct Host Response.” *Virulence* 9 (1): 804–17. https://doi.org/10.1080/21505594.2018.1449508.

Kaoutari, Abdessamad El, Fabrice Armougom, Jeffrey I Gordon, Didier Raoult, and Bernard Henrissat. 2013. “The Abundance and Variety of Carbohydrate-Active Enzymes in the Human Gut Microbiota.” *Nature Reviews Microbiology* 11 (7): 497–504. https://doi.org/10.1038/nrmicro3050.

Kasuya, Natsuki, Itsuko Wada, Mimune Shimada, Hiroshi Kawai, and Hisao Itabashi. 2007. “Effect of Presence of Rumen Protozoa on Degradation of Cell Wall Constituents in Gastrointestinal Tract of Cattle.” *Animal Science Journal* 78 (3): 275–80. https://doi.org/10.1111/j.1740-0929.2007.00435.x.

Kataeva, Irina A., Ronald D. Seidel, Ashit Shah, Larry T. West, Xin Liang Li, and Lars G. Ljungdahl. 2002. “The Fibronectin Type 3-like Repeat from the Clostridium Thermocellum Cellobiohydrolase CbHa Promotes Hydrolysis of Cellulose by Modifying Its Surface.” *Applied and Environmental Microbiology* 68 (9): 4292–4300. https://doi.org/10.1128/AEM.68.9.4292-4300.2002.

Kataeva, Irina A., Vladimir N. Uversky, John M. Brewer, Florian Schubot, John P. Rose, B. C. Wang, and Lars G. Ljungdahl. 2004. “Interactions between Immunoglobulin-like and Catalytic Modules in Clostridium Thermocellum Cellulosomal Cellobiohydrolase CbhA.” *Protein Engineering, Design and Selection* 17 (11): 759–69. https://doi.org/10.1093/protein/gzh094.

Kautsar, Satria A., Kai Blin, Simon Shaw, Jorge C. Navarro-Muñoz, Barbara R. Terlouw, Justin J.J. Van Der Hooft, Jeffrey A. Van Santen, et al. 2020. “MIBiG 2.0: A Repository for Biosynthetic Gene Clusters of Known Function.” *Nucleic Acids Research* 48 (D1): D454–58. https://doi.org/10.1093/nar/gkz882.

Kautz, Stefanie, Benjamin E.R. Rubin, Jacob A. Russell, and Corrie S. Moreaua. 2013. “Surveying the Microbiome of Ants: Comparing 454 Pyrosequencing with Traditional Methods to Uncover Bacterial Diversity.” *Applied and Environmental Microbiology*. https://doi.org/10.1128/AEM.03107-12.

Kelley, Lawrence A., Stefans Mezulis, Christopher M. Yates, Mark N. Wass, and Michael J.E.

Sternberg. 2015. “The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis.” *Nature Protocols* 10 (6): 845–58. https://doi.org/10.1038/nprot.2015.053.

Kelley, Scott T., and Susanne Dobler. 2011. “Comparative Analysis of Microbial Diversity in Longitarsus Flea Beetles (Coleoptera: Chrysomelidae).” *Genetica* 139 (5): 541–50. https://doi.org/10.1007/s10709-010-9498-0.

Kerff, Frédéric, Ana Amoroso, Raphaël Herman, Eric Sauvage, Stéphanie Petrella, Patrice Filée, Paulette Charlier, et al. 2008. “Crystal Structure and Activity of Bacillus Subtilis YoaJ (EXLX1), a Bacterial Expansin That Promotes Root Colonization.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (44): 16876–81. https://doi.org/10.1073/pnas.0809382105.

Khater, Shradha, Swadha Anand, and Debasisa Mohanty. 2016. “In Silico Methods for Linking Genes and Secondary Metabolites: The Way Forward.” Synthetic and Systems Biotechnology. 2016. https://doi.org/10.1016/j.synbio.2016.03.001.

Kim, Daehwan. 2018. “Physico-Chemical Conversion of Lignocellulose: Inhibitor Effects and Detoxification Strategies: A Mini Review.” *Molecules* 23 (2). https://doi.org/10.3390/molecules23020309.

Kiryu, Takaaki, Hirofumi Nakano, Taro Kiso, and Hiromi Murakami. 2005. “Purification and Characterization of a Novel α-Glucuronidase from Aspergillus Niger Specific for O -α- D -Glucosyluronic Acid α- D -Glucosiduronic Acid.” *Bioscience, Biotechnology, and Biochemistry* 69 (3): 522–29. https://doi.org/10.1271/bbb.69.522.

Kittelmann, Sandra, and Peter H. Janssen. 2011. “Characterization of Rumen Ciliate Community Composition in Domestic Sheep, Deer, and Cattle, Feeding on Varying Diets, by Means of PCR-DGGE and Clone Libraries.” *FEMS Microbiology Ecology* 75 (3): 468–81. https://doi.org/10.1111/j.1574-6941.2010.01022.x.

Kotake, Toshihisa, Yukiko Yamanashi, Chiemi Imaizumi, and Yoichi Tsumuraya. 2016. “Metabolism of L-Arabinose in Plants.” *Journal of Plant Research* 129 (5): 781–92. https://doi.org/10.1007/s10265-016-0834-z.

Kraemer, Susanne A., Arthi Ramachandran, and Gabriel G. Perron. 2019. “Antibiotic Pollution in the Environment: From Microbial Ecology to Public Policy.” *Microorganisms* 7 (6). https://doi.org/10.3390/microorganisms7060180.

Kramer, Jos, Özhan Özkaya, and Rolf Kümmerli. 2020. “Bacterial Siderophores in Community and Host Interactions.” *Nature Reviews Microbiology*. Nature Research.

https://doi.org/10.1038/s41579-019-0284-4.

Kroiss, Johannes, Martin Kaltenpoth, Bernd Schneider, Maria Gabriele Schwinger, Christian Hertweck, Ravi Kumar Maddula, Erhard Strohm, and Ale Svatos. 2010. “Symbiotic Streptomycetes Provide Antibiotic Combination Prophylaxis for Wasp Offspring.” *Nature Chemical Biology* 6 (4): 261–63. https://doi.org/10.1038/nchembio.331.

Kucharska, Karolina, Piotr Rybarczyk, Iwona Hołowacz, Rafał Łukajtis, Marta Glinka, and Marian Kamiński. 2018. “Pretreatment of Lignocellulosic Materials as Substrates for Fermentation Processes.” *Molecules* 23 (11). https://doi.org/10.3390/molecules23112937.

Kuge, Takayuki, and Haruhiko Teramoto. 2015. “AraR , an L -Arabinose-Responsive Transcriptional Regulator in Corynebacterium Glutamicum ATCC 31831 , Exerts Different Degrees of Repression Depending on the Location of Its Binding Sites Within.” *Journal of Bacteriology* 197 (24): 3788–96. https://doi.org/10.1128/JB.00314-15.Editor.

Kumar, Manoj, Ajit Varma, and Vivek Kumar. 2016. “Ecogenomics Based Microbial Enzyme for Biofuel Industry.” *Science International* 4 (1): 1–11. https://doi.org/10.17311/sciintl.2016.1.11.

Kumar, Sandeep, Chung-jung Tsai, and Ruth Nussinov. 2002. “Factors Enhancing Protein Thermostability.” *Protein Engineering, Design and Selection* 13 (3): 179–91. https://doi.org/10.1093/protein/13.3.179.

Kwon, Gayeung, Jiyun Lee, and Young Hee Lim. 2016. “Dairy Propionibacterium Extends the Mean Lifespan of Caenorhabditis Elegans via Activation of the Innate Immune System.” *Scientific Reports* 6 (1): 1–11. https://doi.org/10.1038/srep31713.

Lagaert, Stijn, Annick Pollet, Christophe M. Courtin, and Guido Volckaert. 2014. “β-Xylosidases and α-L-Arabinofuranosidases: Accessory Enzymes for Arabinoxylan Degradation.” *Biotechnology Advances*. Elsevier Inc. https://doi.org/10.1016/j.biotechadv.2013.11.005.

Lämmli, U. K. 1970. “Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4.” *Nature* 227 (5259): 680–85. https://doi.org/10.1038/227680a0.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Larsen, Brendan B., Elizabeth C. Miller, Matthew K. Rhodes, and John J. Wiens. 2017. “Inordinate Fondness Multiplied and Redistributed: The Number of Species on Earth and the New Pie of Life.” *Quarterly Review of Biology* 92 (3): 229–65. https://doi.org/10.1086/693564.

Le, Hoang Anh, Do Minh Phuong, and Le Thuy Linh. 2020. “Emission Inventories of Rice Straw Open Burning in the Red River Delta of Vietnam: Evaluation of the Potential of Satellite Data.”

Environmental Pollution 260 (May): 113972. https://doi.org/10.1016/j.envpol.2020.113972.

Leblanc, Shannon K.D., Christopher W. Oates, and Tracy L. Raivio. 2011. "Characterization of the Induction and Cellular Role of the BaeSR Two-Component Envelope Stress Response of Escherichia Coli." *Journal of Bacteriology* 193 (13): 3367–75. https://doi.org/10.1128/JB.01534-10.

Lee, Charles C., Rena E. Kibblewhite, Kurt Wagschal, Ruiping Li, George H. Robertson, and William J. Orts. 2012. "Isolation and Characterization of a Novel GH67 α-Glucuronidase from a Mixed Culture." *Journal of Industrial Microbiology and Biotechnology* 39 (8): 1245–51. https://doi.org/10.1007/s10295-012-1128-7.

Lee, Charles C, Rena E Kibblewhite, Kurt Wagschal, Ruiping Li, and William J. Orts. 2012. "Isolation of α-Glucuronidase Enzyme from a Rumen Metagenomic Library." *Protein Journal* 31 (3): 206–11. https://doi.org/10.1007/s10930-012-9391-z.

Lee, Hee Jin, Saeyoung Lee, Hyeok Jin Ko, Kyoung Heon Kim, and In Geol Choi. 2010. "An Expansin-like Protein from Hahella Chejuensis Binds Cellulose and Enhances Cellulase Activity." *Molecules and Cells* 29 (4): 379–85. https://doi.org/10.1007/s10059-010-0033-z.

Lee, Kyung Tai, Sazzad Hossen Toushik, Jin Young Baek, Ji Eun Kim, Jin Sung Lee, and Keun Sung Kim. 2018. "Metagenomic Mining and Functional Characterization of a Novel KG51 Bifunctional Cellulase/Hemicellulase from Black Goat Rumen." *Journal of Agricultural and Food Chemistry* 66 (34): 9034–41. https://doi.org/10.1021/acs.jafc.8b01449.

Lee, Queena, and Paul Widden. 1996. "Folsomia Candida, a 'fungivorous' Collembolan, Feeds Preferentially on Nematodes Rather than Soil Fungi." *Soil Biology and Biochemistry* 28 (4–5): 689–90. https://doi.org/10.1016/0038-0717(95)00158-1.

Leger, Michelle M., Laura Eme, Courtney W. Stairs, and Andrew J. Roger. 2018. "Demystifying Eukaryote Lateral Gene Transfer (Response to Martin 2017 DOI: 10.1002/Bies.201700115)." *BioEssays* 40 (5): 1700242. https://doi.org/10.1002/bies.201700242.

Leitão-Gonçalves, Ricardo, Zita Carvalho-Santos, Ana Patrícia Francisco, Gabriela Tondolo Fioreze, Margarida Anjos, Célia Baltazar, Ana Paula Elias, Pavel M. Itskov, Matthew D.W. Piper, and Carlos Ribeiro. 2017. "Commensal Bacteria and Essential Amino Acids Control Food Choice Behavior and Reproduction." *PLoS Biology* 15 (4). https://doi.org/10.1371/journal.pbio.2000862.

Lelio, Ilaria Di, Anna Illiano, Federica Astarita, Luca Gianfranceschi, David Horner, Paola Varricchio, Angela Amoresano, Pietro Pucci, Francesco Pennacchio, and Silvia Caccia. 2019. "Evolution of

an Insect Immune Barrier through Horizontal Gene Transfer Mediated by a Parasitic Wasp." *PLoS Genetics* 15 (3). https://doi.org/10.1371/journal.pgen.1007998.

Lemoine, Frédéric, Damien Correia, Vincent Lefort, Olivia Doppelt-Azeroual, Fabien Mareuil, Sarah Cohen-Boulakia, and Olivier Gascuel. 2019. "NGPhylogeny.Fr: New Generation Phylogenetic Services for Non-Specialists." *Nucleic Acids Research* 47 (W1): W260–65. https://doi.org/10.1093/nar/gkz303.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Hongjie, Daniel J. Yelle, Chang Li, Mengyi Yang, Jing Ke, Ruijuan Zhang, Yu Liu, et al. 2017a. "Lignocellulose Pretreatment in a Fungus-Cultivating Termite." *Proceedings of the National Academy of Sciences* 114 (18): 4709–14. https://doi.org/10.1073/pnas.1618360114.

———. 2017b. "Lignocellulose Pretreatment in a Fungus-Cultivating Termite." *Proceedings of the National Academy of Sciences of the United States of America* 114 (18): 4709–14. https://doi.org/10.1073/pnas.1618360114.

Li, Weizhong, and Adam Godzik. 2006. "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences." *Bioinformatics* 22 (13): 1658–59. https://doi.org/10.1093/bioinformatics/btl158.

Lim, Sooyeon, Jaehyun Seo, Hyunbong Choi, Duhak Yoon, Jungrye Nam, Heebal Kim, Seoae Cho, and Jongsoo Chang. 2013. "Metagenome Analysis of Protein Domain Collocation within Cellulase Genes of Goat Rumen Microbes." *Asian-Australasian Journal of Animal Sciences* 26 (8): 1144–51. https://doi.org/10.5713/ajas.2013.13219.

Linares-Pastén, Javier A., Peter Falck, Khalil Albasri, Sven Kjellström, Patrick Adlercreutz, Derek T. Logan, and Eva Nordberg Karlsson. 2017. "Three-Dimensional Structures and Functional Studies of Two GH43 Arabinofuranosidases from Weissella Sp. Strain 142 and Lactobacillus Brevis." *FEBS Journal* 284 (13): 2019–36. https://doi.org/10.1111/febs.14101.

Linares, J. F., I. Gustafsson, F. Baquero, and J. L. Martinez. 2006. "Antibiotics as Intermicrobiol Signaling Agents Instead of Weapons." *Proceedings of the National Academy of Sciences of the United States of America* 103 (51): 19484–89. https://doi.org/10.1073/pnas.0608949103.

Linger, Jeffrey G., William S. Adney, and Al Darzins. 2010. "Heterologous Expression and Extracellular Secretion of Cellulolytic Enzymes by Zymomonas Mobilis." *Applied and Environmental Microbiology* 76 (19): 6360–69. https://doi.org/10.1128/AEM.00230-10.

Liu, Dianfeng, Bin Lian, Chunhao Wu, and Peijun Guo. 2018. “A Comparative Study of Gut Microbiota Profiles of Earthworms Fed in Three Different Substrates.” *Symbiosis* 74 (1): 21–29. https://doi.org/10.1007/s13199-017-0491-6.

Liu, Fanghua, Amelia Elena Rotaru, Pravin M. Shrestha, Nikhil S. Malvankar, Kelly P. Nevin, and Derek R. Lovley. 2012. “Promoting Direct Interspecies Electron Transfer with Activated Carbon.” *Energy and Environmental Science* 5 (10): 8982–89. https://doi.org/10.1039/c2ee22459c.

Liu, Min, Jiali Gu, Wenping Xie, and Hongwei Yu. 2013. “Directed Co-Evolution of an Endoglucanase and a β-Glucosidase in Escherichia Coli by a Novel High-Throughput Screening Method.” *Chemical Communications* 49 (65): 7219–21. https://doi.org/10.1039/c3cc42485e.

Liu, Xiangyang, and Chandrakant Kokare. 2016. *Microbial Enzymes of Use in Industry*. *Biotechnology of Microbial Enzymes: Production, Biocatalysis and Industrial Applications*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-803725-6.00011-X.

Lombard, Vincent, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M. Coutinho, and Bernard Henrissat. 2014. “The Carbohydrate-Active Enzymes Database (CAZy) in 2013.” *Nucleic Acids Research* 42 (D1): D490–95. https://doi.org/10.1093/nar/gkt1178.

Lopes, A. M., E. X. Ferreira Filho, and L. R.S. Moreira. 2018. “An Update on Enzymatic Cocktails for Lignocellulose Breakdown.” *Journal of Applied Microbiology* 125 (3): 632–45. https://doi.org/10.1111/jam.13923.

López-Mondéjar, Rubén, Daniela Zühlke, Dörte Becher, Katharina Riedel, and Petr Baldrian. 2016. “Cellulose and Hemicellulose Decomposition by Forest Soil Bacteria Proceeds by the Action of Structurally Variable Enzymatic Systems.” *Scientific Reports* 6 (1): 1–12. https://doi.org/10.1038/srep25279.

Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. 2012. “SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de Novo Assembler.” *GigaScience* 1 (1). https://doi.org/10.1186/2047-217X-1-18.

Ma, Lili, Yuwei Xie, Zhihua Han, John P. Giesy, and Xiaowei Zhang. 2017. “Responses of Earthworms and Microbial Communities in Their Guts to Triclosan.” *Chemosphere* 168: 1194–1202. https://doi.org/10.1016/j.chemosphere.2016.10.079.

Ma, Zongwang, Niels Geudens, Nam P. Kieu, Davy Sinnaeve, Marc Ongena, José C. Martins, and Monica Höfte. 2016. “Biosynthesis, Chemical Structure, and Structure-Activity Relationship of Orfamide Lipopeptides Produced by Pseudomonas Protegens and Related Species.” *Frontiers in*

Microbiology 7 (MAR). https://doi.org/10.3389/fmicb.2016.00382.

Maehara, Tomoko, Zui Fujimoto, Hitomi Ichinose, Mari Michikawa, Koichi Harazono, and Satoshi Kaneko. 2014. "Crystal Structure and Characterization of the Glycoside Hydrolase Family 62 α-L-Arabinofuranosidase from Streptomyces Coelicolor." *Journal of Biological Chemistry* 289 (11): 7962–72. https://doi.org/10.1074/jbc.M113.540542.

Maguire, Meghan, Julie A. Kase, Dwayne Roberson, Tim Muruvanda, Eric W. Brown, Marc Allard, Steven M. Musser, and Narjol González-Escalona. 2021. "Precision Long-Read Metagenomics Sequencing for Food Safety by Detection and Assembly of Shiga Toxin-Producing Escherichia Coli in Irrigation Water." Edited by Pina Fratamico. *PLoS ONE* 16 (1): e0245172. https://doi.org/10.1371/journal.pone.0245172.

Manjula, K., and A. R. Podile. 2005. "Production of Fungal Cell Wall Degrading Enzymes by a Biocontrol Strain of Bacillus Subtilis AF 1." *Indian Journal of Experimental Biology* 43 (10): 892–96.

Margolles, Abelardo, and Clara G. De los Reyes-Gavilán. 2003. "Purification and Functional Characterization of a Novel α-L-Arabinofuranosidase from Bifidobacterium Longum B667." *Applied and Environmental Microbiology* 69 (9): 5096–5103. https://doi.org/10.1128/AEM.69.9.5096-5103.2003.

Marolda, Cristina L., Bo Li, Michael Lung, Mei Yang, Anna Hanuszkiewicz, Amanda Roa Rosales, and Miguel A. Valvano. 2010. "Membrane Topology and Identification of Critical Amino Acid Residues in the Wzx O-Antigen Translocase from Escherichia Coli O157:H4." *Journal of Bacteriology* 192 (23): 6160–71. https://doi.org/10.1128/JB.00141-10.

Marolda, Cristina L., Laura D. Tatar, Cristina Alaimo, Markus Aebi, and Miguel A. Valvano. 2006. "Interplay of the Wzx Translocase and the Corresponding Polymerase and Chain Length Regulator Proteins in the Translocation and Periplasmic Assembly of Lipopolysaccharide O Antigen." *Journal of Bacteriology* 188 (14): 5124–35. https://doi.org/10.1128/JB.00461-06.

Martin, William F. 2017. "Too Much Eukaryote LGT." *BioEssays*. https://doi.org/10.1002/bies.201700115.

Matsuo, Noriki, Satoshi Kaneko, Atsushi Kuno, Hideyuki Kobayashi, and Isao Kusakabe. 2000. "Purification, Characterization and Gene Cloning of Two α-l-Arabinofuranosidases from Streptomyces Chartreusis GS901." *Biochemical Journal*. Vol. 346. https://doi.org/10.1042/bj3460009.

Maurya, Devendra Prasad, Ankit Singla, and Sangeeta Negi. 2015. "An Overview of Key Pretreatment

Processes for Biological Conversion of Lignocellulosic Biomass to Bioethanol." *3 Biotech* 5 (5): 597–609. https://doi.org/10.1007/s13205-015-0279-4.

Mayer, Werner E., Lisa N. Schuster, Gabi Bartelmes, Christoph Dieterich, and Ralf J. Sommer. 2011. "Horizontal Gene Transfer of Microbial Cellulases into Nematode Genomes Is Associated with Functional Assimilation and Gene Turnover." *BMC Evolutionary Biology* 11 (1): 13. https://doi.org/10.1186/1471-2148-11-13.

McClure, Ryan A., Anthony W. Goering, Kou San Ju, Joshua A. Baccile, Frank C. Schroeder, William W. Metcalf, Regan J. Thomson, and Neil L. Kelleher. 2016. "Elucidating the Rimosamide-Detoxin Natural Product Families and Their Biosynthesis Using Metabolite/Gene Cluster Correlations." *ACS Chemical Biology* 11 (12): 3452–60. https://doi.org/10.1021/acschembio.6b00779.

McDonald, Bradon R., and Cameron R. Curriea. 2017. "Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces." *MBio* 8 (3). https://doi.org/10.1128/mBio.00644-17.

McFall-Ngai, Margaret, Michael G Hadfield, Thomas C G Bosch, Hannah V Carey, Tomislav Domazet-Lo, Angela E Douglas, Nicole Dubilier, et al. 2013. "Animals in a Bacterial World, a New Imperative for the Life Sciences." *Proc. Natl. Acad. Sci.* 110 (9): 3229–36. https://doi.org/10.1073/pnas.1218525110.

McKee, Lauren S., Hampus Sunner, George E. Anasontzis, Guillermo Toriz, Paul Gatenholm, Vincent Bulone, Francisco Vilaplana, and Lisbeth Olsson. 2016. "A GH115 α-Glucuronidase from Schizophyllum Commune Contributes to the Synergistic Enzymatic Deconstruction of Softwood Glucuronoarabinoxylan." *Biotechnology for Biofuels* 9 (1): 1–13. https://doi.org/10.1186/s13068-015-0417-6.

Meijenfeldt, F A Bastiaan von, Ksenia Arkhipova, Diego D Cambuy, Felipe H Coutinho, and Bas E Dutilh. 2019. "Robust Taxonomic Classification of Uncharted Microbial Sequences and Bins with CAT and BAT." *BioRxiv*, 530188. https://doi.org/10.1101/530188.

Merino, Sandra T., and Joel Cherry. 2007. "Progress and Challenges in Enzyme Development for Biomass Utilization." *Advances in Biochemical Engineering/Biotechnology* 108 (April): 95–120. https://doi.org/10.1007/10_2007_066.

Mewis, Keith, Nicolas Lenfant, Vincent Lombard, and Bernard Henrissat. 2016. "Dividing the Large Glycoside Hydrolase Family 43 into Subfamilies: A Motivation for Detailed Enzyme Characterization." *Applied and Environmental Microbiology* 82 (6): 1686–92. https://doi.org/10.1128/aem.03453-15.

Mhuantong, Wuttichai, Varodom Charoensawan, Pattanop Kanokratana, Sithichoke Tangphatsornruang, and Verawat Champreda. 2015. “Comparative Analysis of Sugarcane Bagasse Metagenome Reveals Unique and Conserved Biomass-Degrading Enzymes among Lignocellulolytic Microbial Communities.” *Biotechnology for Biofuels* 8 (1): 16. https://doi.org/10.1186/s13068-015-0200-8.

Michlmayr, Herbert, Johannes Hell, Cindy Lorenz, Stefan Böhmdorfer, Thomas Rosenau, and Wolfgang Kneifel. 2013. “Arabinoxylan Oligosaccharide Hydrolysis by Family 43 and 51 Glycosidases from Lactobacillus Brevis DSM 20054.” *Applied and Environmental Microbiology* 79 (21): 6747–54. https://doi.org/10.1128/aem.02130-13.

Millati, Ria, Siti Syamsiah, Claes Niklasson, Muhammad Nur Cahyanto, Knut Lundquist, and Mohammad J Taherzadeh. 2011. “Biological Pretreatment: Review.” *BioResources* 6 (4): 5224–59.

Miller, Gail Lorenz. 1959. “Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar.” *Analytical Chemistry* 31 (3): 426–28. https://doi.org/10.1021/ac60147a030.

Montaña, Sabrina, Sareda T.J. Schramm, German Matías Traglia, Kevin Chiem, Gisela Parmeciano Di Noto, Marisa Almuzara, Claudia Barberis, et al. 2016. “The Genetic Analysis of an Acinetobacter Johnsonii Clinical Strain Evidenced the Presence of Horizontal Genetic Transfer.” *PLoS ONE* 11 (8). https://doi.org/10.1371/journal.pone.0161528.

Moran, Nancy A., Howard Ochman, and Tobin J. Hammer. 2019. “Evolutionary and Ecological Consequences of Gut Microbial Communities.” *Annual Review of Ecology, Evolution, and Systematics* 50: 451–75. https://doi.org/10.1146/annurev-ecolsys-110617-062453.

Moreira, Leonardo Marmo, Fernando de Paula Leone, Ricardo Augusto Mendonça Vieira, and José Carlos Pereira. 2013. “A New Approach about the Digestion of Fibers by Ruminants.” Revista Brasileira de Saude e Producao Animal. 2013. https://doi.org/10.1590/S1519-99402013000200008.

Morita, Masahiko, Nikhil S. Malvankar, Ashley E. Franks, Zarath M. Summers, Ludovic Giloteaux, Amelia E. Rotaru, Camelia Rotaru, and Derek R. Lovleya. 2011. “Potential for Direct Interspecies Electron Transfer in Methanogenic Wastewater Digester Aggregates.” *MBio* 2 (4). https://doi.org/10.1128/mBio.00159-11.

Mosier, Annika C., Nicholas B. Justice, Benjamin P. Bowen, Richard Baran, Brian C. Thomas, Trent R. Northen, and Jillian F. Banfield. 2013. “Metabolites Associated with Adaptation of Microorganisms to an Acidophilic, Metal-Rich Environment Identified by Stable-Isotope-Enabled Metabolomics.” *MBio* 4 (2). https://doi.org/10.1128/mBio.00484-12.

Moss, Eli L., Dylan G. Maghini, and Ami S. Bhatt. 2020. “Complete, Closed Bacterial Genomes from Microbiomes Using Nanopore Sequencing.” *Nature Biotechnology* 38 (6): 701–7. https://doi.org/10.1038/s41587-020-0422-6.

Mullany, Peter. 2014. “Functional Metagenomics for the Investigation of Antibiotic Resistance.” *Virulence* 5 (3): 443–47. https://doi.org/10.4161/viru.28196.

Nagy, Tibor, Kaveh Emami, Carlos M.G.A. G A Fontes, Luis M.A. A Ferreira, David R. Humphry, and Harry J. Gilbert. 2002. “The Membrane-Bound α-Glucuronidase from Pseudomonas Cellulosa Hydrolyzes 4-O-Methyl-D-Glucuronoxylooligosaccharides but Not 4-O-Methyl-D-Glucuronoxylan.” *Journal of Bacteriology* 184 (17): 4925–29. https://doi.org/10.1128/JB.184.17.4925-4929.2002.

Naughton, Lynn M., Stefano Romano, Fergal O’Gara, and Alan D.W. Dobson. 2017. “Identification of Secondary Metabolite Gene Clusters in the Pseudovibrio Genus Reveals Encouraging Biosynthetic Potential toward the Production of Novel Bioactive Compounds.” *Frontiers in Microbiology* 8 (AUG). https://doi.org/10.3389/fmicb.2017.01494.

Navarro, David, Marie Couturier, Gabriela G.D. da Silva, Jean Guy Berrin, Xavier Rouau, Marcel Asther, and Christophe Bignon. 2010. “Automated Assay for Screening the Enzymatic Release of Reducing Sugars from Micronized Biomass.” *Microbial Cell Factories* 9 (1): 58. https://doi.org/10.1186/1475-2859-9-58.

Neddersen, Mara, and Skander Elleuche. 2015. “Fast and Reliable Production, Purification and Characterization of Heat-Stable, Bifunctional Enzyme Chimeras.” *AMB Express* 5 (1): 1–12. https://doi.org/10.1186/s13568-015-0122-7.

Nelson, Karen E., Stephen H. Zinder, Ioana Hance, Patrick Burr, David Odongo, Delia Wasawo, Agnes Odenyo, and Richard Bishop. 2003. “Phylogenetic Analysis of the Microbial Populations in the Wild Herbivore Gastrointestinal Tract: Insights into an Unexplored Niche.” *Environmental Microbiology* 5 (11): 1212–20. https://doi.org/10.1046/j.1462-2920.2003.00526.x.

Nesme, Joseph, Sébastien Cécillon, Tom O. Delmont, Jean Michel Monier, Timothy M. Vogel, and Pascal Simonet. 2014. “Large-Scale Metagenomic-Based Study of Antibiotic Resistance in the Environment.” *Current Biology* 24 (10): 1096–1100. https://doi.org/10.1016/j.cub.2014.03.036.

Ni, Jinfeng, and Gaku Tokuda. 2013. “Lignocellulose-Degrading Enzymes from Termites and Their Symbiotic Microbiota.” *Biotechnology Advances* 31 (6): 838–50. https://doi.org/10.1016/j.biotechadv.2013.04.005.

Nikaido, Hiroshi. 2010. “Structure and Mechanism of RND-Type Multidrug Efflux Pumps.” *Advances*

in Enzymology and Related Areas of Molecular Biology. Vol. 77 1. https://doi.org/10.1002/9780470920541.ch1.

Nota, Benjamin, Martijn J.T.N. Timmermans, Oscar Franken, Kora Montagne-Wajer, Janine Mariën, Muriel E. De Boer, Tjalf E. De Boer, Bauke Ylstra, Nico M. Van Straalen, and Dick Roelofs. 2008. "Gene Expression Analysis of Collembola in Cadmium Containing Soil." *Environmental Science and Technology*. https://doi.org/10.1021/es801472r.

Numan, Mondher Th, and Narayan B. Bhosle. 2006. "α-L-Arabinofuranosidases: The Potential Applications in Biotechnology." *Journal of Industrial Microbiology and Biotechnology* 33 (4): 247–60. https://doi.org/10.1007/s10295-005-0072-1.

Nurizzo, Didier, Tibor Nagy, Harry J. Gilbert, and Gideon J. Davies. 2002. "The Structural Basis for Catalysis and Specificity of the Pseudomonas Cellulosa α-Glucuronidase, GlcA67A." *Structure*. Vol. 10. https://doi.org/10.1016/S0969-2126(02)00742-6.

Nurizzo, Didier, Johan P. Turkenburg, Simon J. Charnock, Shirley M. Roberts, Eleanor J. Dodson, Vincent A. McKie, Edward J. Taylor, Harry J. Gilbert, and Gideon J. Davies. 2002. "Cellvibrio Japonicus α-l-Arabinanase 43a Has a Novel Five-Blade β-Propeller Fold." *Nature Structural Biology* 9 (9): 665–68. https://doi.org/10.1038/nsb835.

Ofori-Boateng, Cynthia, and Keat Teong Lee. 2013. "Sustainable Utilization of Oil Palm Wastes for Bioactive Phytochemicals for the Benefit of the Oil Palm and Nutraceutical Industries." *Phytochemistry Reviews* 12 (1): 173–90. https://doi.org/10.1007/s11101-013-9270-z.

Oliphant, Kaitlyn, and Emma Allen-Vercoe. 2019. "Macronutrient Metabolism by the Human Gut Microbiome: Major Fermentation by-Products and Their Impact on Host Health." *Microbiome* 7 (1): 1–15. https://doi.org/10.1186/s40168-019-0704-8.

Olofsson, Kim, Magnus Bertilsson, and Gunnar Lidén. 2008. "A Short Review on SSF - An Interesting Process Option for Ethanol Production from Lignocellulosic Feedstocks." *Biotechnology for Biofuels* 1: 1–14. https://doi.org/10.1186/1754-6834-1-7.

Omoboye, Olumide Owolabi, Niels Geudens, Matthieu Duban, Mickaël Chevalier, Christophe Flahaut, José C. Martins, Valérie Leclère, Feyisara Eyiwumi Oni, and Monica Höfte. 2019. "Pseudomonas Sp. COW3 Produces New Bananamide-Type Cyclic Lipopeptides with Antimicrobial Activity against Pythium Myriotylum and Pyricularia Oryzae." *Molecules* 24 (22). https://doi.org/10.3390/molecules24224170.

Omoboye, Olumide Owolabi, Feyisara Eyiwumi Oni, Humaira Batool, Henok Zimene Yimer, René De Mot, and Monica Höfte. 2019. "Pseudomonas Cyclic Lipopeptides Suppress the Rice Blast

Fungus Magnaporthe Oryzae by Induced Resistance and Direct Antagonism." *Frontiers in Plant Science* 10 (July): 901. https://doi.org/10.3389/fpls.2019.00901.

Owji, Hajar, Navid Nezafat, Manica Negahdaripour, Ali Hajiebrahimi, and Younes Ghasemi. 2018. "A Comprehensive Review of Signal Peptides: Structure, Roles, and Applications." *European Journal of Cell Biology*. Elsevier GmbH. https://doi.org/10.1016/j.ejcb.2018.06.003.

Page, Malcom G.P. 2019. "The Role of Iron and Siderophores in Infection, and the Development of Siderophore Antibiotics." *Clinical Infectious Diseases* 69 (Suppl 7): S529–37. https://doi.org/10.1093/cid/ciz825.

Parisutham, Vinuselvi, Tae Hyun Kim, and Sung Kuk Lee. 2014. "Feasibilities of Consolidated Bioprocessing Microbes: From Pretreatment to Biofuel Production." *Bioresource Technology*. Elsevier Ltd. https://doi.org/10.1016/j.biortech.2014.03.114.

Pason, Patthra, Kanok Wongratpanya, Thidarat Nimchua, Somphit Sornyotha, Siriluck Imjongjairak, Khanok Ratanakhanokchai, Chakrit Tachaapaikoon, Paripok Phitsuwan, and Rattiya Waeonukul. 2015. "Multifunctional Properties of Glycoside Hydrolase Family 43 from Paenibacillus Curdlanolyticus Strain B-6 Including Exo-β-Xylosidase, Endo-Xylanase, and α-L-Arabinofuranosidase Activities." *BioResources* 10 (2): 2492–2505. https://doi.org/10.15376/biores.10.2.2492-2505.

Pass, Daniel Antony, Andrew John Morgan, Daniel S. Read, Dawn Field, Andrew J. Weightman, and Peter Kille. 2015. "The Effect of Anthropogenic Arsenic Contamination on the Earthworm Microbiome." *Environmental Microbiology* 17 (6): 1884–96. https://doi.org/10.1111/1462-2920.12712.

Passerini, Delphine, Michèle Coddeville, Pascal Le Bourgeois, Pascal Loubière, Paul Ritzenthaler, Catherine Fontagné-Faucher, Marie-Line Daveran-Mingot, and Muriel Cocaign-Bousquet. 2013. "The Carbohydrate Metabolism Signature of Lactococcus Lactis Strain A12 Reveals Its Sourdough Ecosystem Origin." *Applied and Environmental Microbiology* 79 (19): 5844–52. https://doi.org/10.1128/aem.01560-13.

Pauchet, Yannick, and David G. Heckel. 2013. "The Genome of the Mustard Leaf Beetle Encodes Two Active Xylanases Originally Acquired from Bacteria through Horizontal Gene Transfer." *Proceedings of the Royal Society B: Biological Sciences* 280 (1763). https://doi.org/10.1098/rspb.2013.1021.

Paula, Débora P., Benjamin Linard, Alex Crampton-Platt, Amrita Srivathsan, Martijn J.T.N. Timmermans, Edison R. Sujii, Carmen S.S. Pires, Lucas M. Souza, David A. Andow, and Alfried P. Vogler. 2016. "Uncovering Trophic Interactions in Arthropod Predators through DNA

Shotgun-Sequencing of Gut Contents." *PLoS ONE* 11 (9). https://doi.org/10.1371/journal.pone.0161841.

Pearman, William S., Nikki E. Freed, and Olin K. Silander. 2019. "The Advantages and Disadvantages of Short- And Long-Read Metagenomics to Infer Bacterial and Eukaryotic Community Composition." *BioRxiv*. bioRxiv. https://doi.org/10.1101/650788.

Pereira, Sara B., Marina Santos, José P. Leite, Carlos Flores, Carina Eisfeld, Zsófia Büttel, Rita Mota, et al. 2018. "The Role of the Tyrosine Kinase Wzc (Sll0923) and the Phosphatase Wzb (Slr0328) in the Production of Extracellular Polymeric Substances (EPS) by Synechocystis PCC 6803." *MicrobiologyOpen*, no. June 2018: 1–15. https://doi.org/10.1002/mbo3.753.

Peterson, Brittany F., and Michael E. Scharf. 2016a. "Lower Termite Associations with Microbes: Synergy, Protection, and Interplay." *Frontiers in Microbiology* 7 (APR). https://doi.org/10.3389/fmicb.2016.00422.

———. 2016b. "Metatranscriptome Analysis Reveals Bacterial Symbiont Contributions to Lower Termite Physiology and Potential Immune Functions." *BMC Genomics* 17 (1): 1–12. https://doi.org/10.1186/s12864-016-3126-z.

Piddock, Laura J.V. 2006a. "Clinically Relevant Chromosomally Encoded Multidrug Resistance Efflux Pumps in Bacteria." *Clinical Microbiology Reviews* 19 (2): 382–402. https://doi.org/10.1128/CMR.19.2.382-402.2006.

———. 2006b. "Multidrug-Resistance Efflux Pumps - Not Just for Resistance." *Nature Reviews Microbiology* 4 (8): 629–36. https://doi.org/10.1038/nrmicro1464.

Piñeiro, Valeria, Joaquín Arias, Jochen Dürr, Pablo Elverdin, Ana María Ibáñez, Alison Kinengyere, Cristian Morales Opazo, et al. 2020. "A Scoping Review on Incentives for Adoption of Sustainable Agricultural Practices and Their Outcomes." *Nature Sustainability* 3 (10): 809–20. https://doi.org/10.1038/s41893-020-00617-y.

Powell, Sean, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Michael Kuhn, Jean Muller, Roland Arnold, et al. 2012. "EggNOG v3.0: Orthologous Groups Covering 1133 Organisms at 41 Different Taxonomic Ranges." *Nucleic Acids Research* 40 (D1). https://doi.org/10.1093/nar/gkr1060.

Puniya, Anil Kumar, Rameshwar Singh, and Devki Nandan Kamra. 2015. *Rumen Microbiology: From Evolution to Revolution*. *Rumen Microbiology: From Evolution to Revolution*. https://doi.org/10.1007/978-81-322-2401-3.

Qin, Zhiwei, Rebecca Devine, Thomas J Booth, Elliot H E Farrar, Matthew N Grayson, Matthew I

Hutchings, and Barrie Wilkinson. 2020. “Formicamycin Biosynthesis Involves a Unique Reductive Ring Contraction †.” https://doi.org/10.1039/d0sc01712d.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. 2005. “InterProScan: Protein Domains Identifier.” *Nucleic Acids Research* 33 (SUPPL. 2). https://doi.org/10.1093/nar/gki442.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

R Core Team. 2018. “A Language and Environment for Statistical Computing.” *R Foundation for Statistical Computing* 2: https://www.R-project.org. http://www.r-project.org.

Rappé, Michael S., and Stephen J. Giovannoni. 2003. “The Uncultured Microbial Majority.” *Annual Review of Microbiology* 57: 369–94. https://doi.org/10.1146/annurev.micro.57.030502.090759.

Rasekh, Behnam, Khosro Khajeh, Bijan Ranjbar, Nasrin Mollania, Banafsheh Almasinia, and Hassan Tirandaz. 2014. “Protein Engineering of Laccase to Enhance Its Activity and Stability in the Presence of Organic Solvents.” *Engineering in Life Sciences* 14 (4): 442–48. https://doi.org/10.1002/elsc.201300042.

Ravindra, Khaiwal, Tanbir Singh, and Suman Mor. 2019. “Emissions of Air Pollutants from Primary Crop Residue Burning in India and Their Mitigation Strategies for Cleaner Emissions.” *Journal of Cleaner Production* 208 (January): 261–73. https://doi.org/10.1016/j.jclepro.2018.10.031.

Ravindran, Rajeev, and Amit Kumar Jaiswal. 2016. “A Comprehensive Review on Pre-Treatment Strategy for Lignocellulosic Food Industry Waste: Challenges and Opportunities.” *Bioresource Technology* 199: 92–102. https://doi.org/10.1016/j.biortech.2015.07.106.

Reid, Anne N, and Chris Whitfield. 2005. “Functional Analysis of Conserved Gene Products Involved in Assembly of Escherichia Coli Capsules and Exopolysaccharides: Evidence for Molecular Recognition between Wza and Wzc for Colanic Acid Biosynthesis.” *Journal of Bacteriology* 187 (15): 5470–81. https://doi.org/10.1128/JB.187.15.5470-5481.2005.

Rhee, Mun Su, Neha Sawhney, Young Sik Kim, Hyun Jee Rhee, Jason C. Hurlbert, Franz J. St. John, Guang Nong, John D. Rice, and James F. Preston. 2017. “GH115 α-Glucuronidase and GH11 Xylanase from Paenibacillus Sp. JDR-2: Potential Roles in Processing Glucuronoxylans.” *Applied Microbiology and Biotechnology* 101 (4): 1465–76. https://doi.org/10.1007/s00253-016-7899-4.

Ricard, Guénola, Neil R. McEwan, Bas E. Dutilh, Jean Pierre Jouany, Didier Macheboeuf, Makoto

Mitsumori, Freda M. McIntosh, et al. 2006. “Horizontal Gene Transfer from Bacteria to Rumen Ciliates Indicates Adaptation to Their Anaerobic, Carbohydrates-Rich Environment.” *BMC Genomics* 7 (1): 22. https://doi.org/10.1186/1471-2164-7-22.

Richardson, Leif L., Lynn S. Adler, Anne S. Leonard, Jonathan Andicoechea, Karly H. Regan, Winston E. Anthony, Jessamyn S. Manson, and Rebecca E. Irwin. 2015. “Secondary Metabolites in Floral Nectar Reduce Parasite Infections in Bumblebees.” *Proceedings of the Royal Society B: Biological Sciences* 282 (1803). https://doi.org/10.1098/rspb.2014.2471.

Riesenfeld, Christian S., Robert M. Goodman, and Jo Handelsman. 2004. “Uncultured Soil Bacteria Are a Reservoir of New Antibiotic Resistance Genes.” *Environmental Microbiology* 6 (9): 981–89. https://doi.org/10.1111/j.1462-2920.2004.00664.x.

Riley, David R., Karsten B. Sieber, Kelly M. Robinson, James Robert White, Ashwinkumar Ganesan, Syrus Nourbakhsh, and Julie C. Dunning Hotopp. 2013. “Bacteria-Human Somatic Cell Lateral Gene Transfer Is Enriched in Cancer Samples.” *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1003107.

Rinninella, Emanuele, Pauline Raoul, Marco Cintoni, Francesco Franceschi, Giacinto Abele Donato Miggiano, Antonio Gasbarrini, and Maria Cristina Mele. 2019. “What Is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases.” *Microorganisms* 7 (1): 14. https://doi.org/10.3390/microorganisms7010014.

Roberts, Ian S, and Chris Whitfield. 1999. “Structure , Assembly and Regulation of Expression of Capsules in Escherichia Coli.” *Molecular Microbiology* 31 (5): 1307–19.

Roelofs, Dick, Martijn J.T.N. Timmermans, Paul Hensbergen, Hans Van Leeuwen, Jessica Koopman, Anna Faddeeva, Wouter Suring, et al. 2013. “A Functional Isopenicillin N Synthase in an Animal Genome.” *Mol. Biol. Evol.* 30 (3): 541–48. https://doi.org/10.1093/molbev/mss269.

Rogers, M. John, Eric Cundliffe, and Thomas F. McCutchan. 1998. “The Antibiotic Micrococcin Is a Potent Inhibitor of Growth and Protein Synthesis in the Malaria Parasite.” *Antimicrobial Agents and Chemotherapy* 42 (3): 715–16. https://doi.org/10.1128/aac.42.3.715.

Rogowski, Artur, Arnaud Baslé, Cristiane S. Farinas, Alexandra Solovyova, Jennifer C. Mortimer, Paul Dupree, Harry J. Gilbert, and David N. Bolam. 2014. “Evidence That GH115 α-Glucuronidase Activity, Which Is Required to Degrade Plant Biomass, Is Dependent on Conformational Flexibility.” *Journal of Biological Chemistry* 289 (1): 53–64. https://doi.org/10.1074/jbc.M113.525295.

Rowland, Ian, Glenn Gibson, Almut Heinken, Karen Scott, Jonathan Swann, Ines Thiele, and Kieran

Tuohy. 2018. “Gut Microbiota Functions: Metabolism of Nutrients and Other Food Components.” *European Journal of Nutrition*. Dr. Dietrich Steinkopff Verlag GmbH and Co. KG. https://doi.org/10.1007/s00394-017-1445-8.

Ruile, Peter, Christoph Winterhalter, and Wolfgang Liebl. 1997. “Isolation and Analysis of a Gene Encoding α-Glucuronidase, an Enzyme with a Novel Primary Structure Involved in the Breakdown of Xylan.” *Molecular Microbiology* 23 (2): 267–79. https://doi.org/10.1046/j.1365-2958.1997.2011568.x.

Sagné, C, M F Isambert, J P Henry, and B Gasnier. 1996. “SDS-Resistant Aggregation of Membrane Proteins: Application to the Purification of the Vesicular Monoamine Transporter.” *The Biochemical Journal* 316 (Pt 3: 825–31. http://www.ncbi.nlm.nih.gov/pubmed/8670158%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1217424.

Saini, Jitendra Kumar, Reetu Saini, and Lakshmi Tewari. 2015. “Lignocellulosic Agriculture Wastes as Biomass Feedstocks for Second-Generation Bioethanol Production: Concepts and Recent Developments.” *3 Biotech*. Springer Verlag. https://doi.org/10.1007/s13205-014-0246-5.

Sangwan, Naseer, Fangfang Xia, and Jack A. Gilbert. 2016. “Recovering Complete and Draft Population Genomes from Metagenome Datasets.” *Microbiome* 4. https://doi.org/10.1186/s40168-016-0154-5.

Schalk, Isabelle J., and Laurent Guillon. 2013. “Pyoverdine Biosynthesis and Secretion in Pseudomonas Aeruginosa: Implications for Metal Homeostasis.” *Environmental Microbiology* 15 (6): 1661–73. https://doi.org/10.1111/1462-2920.12013.

Schloss, Patrick D., and Jo Handelsman. 2005. “Metagenomics for Studying Unculturable Microorganisms: Cutting the Gordian Knot.” *Genome Biology*. BioMed Central. https://doi.org/10.1186/gb-2005-6-8-229.

Schmid, Jochen, Volker Sieber, and Bernd Rehm. 2015. “Bacterial Exopolysaccharides: Biosynthesis Pathways and Engineering Strategies.” *Frontiers in Microbiology* 6 (MAY): 1–24. https://doi.org/10.3389/fmicb.2015.00496.

Scholl, Elizabeth H., Jeffrey L. Thorne, James P. McCarter, and David Mck Bird. 2003. “Horizontally Transferred Genes in Plant-Parasitic Nematodes: A High-Throughput Genomic Approach.” *Genome Biology* 4 (6): R39. https://doi.org/10.1186/gb-2003-4-6-r39.

Schröder, Gunnar, Ralf Schuelein, Maxime Quebatte, and Christoph Dehio. 2011. “Conjugative DNA Transfer into Human Cells by the VirB/VirD4 Type IV Secretion System of the Bacterial

Pathogen Bartonella Henselae." *Proceedings of the National Academy of Sciences of the United States of America* 108 (35): 14643–48. https://doi.org/10.1073/pnas.1019074108.

Schutyser, Wouter, Tom Renders, Gil Van den Bossche, Sander Van den Bosch, Steven-Friso Koelewijn, Thijs Ennaert, and Bert F. Sels. 2017. "Catalysis in Lignocellulosic Biorefineries: The Case of Lignin Conversion." In *Nanotechnology in Catalysis*, 537–84. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA. https://doi.org/10.1002/9783527699827.ch23.

Scully, Erin D., Scott M. Geib, Kelli Hoover, Ming Tien, Susannah G. Tringe, Kerrie W. Barry, Tijana Glavina del Rio, Mansi Chovatia, Joshua R. Herr, and John E. Carlson. 2013. "Metagenomic Profiling Reveals Lignocellulose Degrading System in a Microbial Community Associated with a Wood-Feeding Beetle." *PLoS ONE* 8 (9). https://doi.org/10.1371/journal.pone.0073827.

Sebastian, Raveendar, Jae-Young Kim, Tae-Hun Kim, and Kyung-Tai Lee. 2013. "Metagenomics: A Promising Approach to Assess Enzymes Biocatalyst for Biofuel Production." *Asian Journal of Biotechnology* 5 (2): 33–50. https://doi.org/10.3923/ajbkr.2013.33.50.

Septiningrum, Krisna, Hiroshi Ohi, Rattiya Waeonukul, Patthra Pason, Chakrit Tachaapaikoon, Khanok Ratanakhanokchai, Junjarus Sermsathanaswadi, Lan Deng, Panida Prawitwong, and Akihiko Kosugi. 2015. "The GH67 α-Glucuronidase of Paenibacillus Curdlanolyticus B-6 Removes Hexenuronic Acid Groups and Facilitates Biodegradation of the Model Xylooligosaccharide Hexenuronosyl Xylotriose." *Enzyme and Microbial Technology* 71: 28–35. https://doi.org/10.1016/j.enzmictec.2015.01.006.

Shannon, A. L., G. Attwood, D. H. Hopcroft, and J. T. Christeller. 2001. "Characterization of Lactic Acid Bacteria in the Larval Midgut of the Keratinophagous Lepidopteran, Hofmannophila Pseudospretella." *Letters in Applied Microbiology* 32 (1): 36–41. https://doi.org/10.1046/j.1472-765X.2001.00854.x.

Shelomi, Matan, Irnayuli R. Sitepu, Kyria L. Boundy-Mills, and Lynn S. Kimsey. 2015. "Review of the Gross Anatomy and Microbiology of the Phasmatodea Digestive Tract." *Journal of Orthoptera Research* 24 (1): 29–40. https://doi.org/10.1665/034.024.0105.

Shiner, Erin K., Kendra P. Rumbaugh, and Simon C. Williams. 2005. "Interkingdom Signaling: Deciphering the Language of Acyl Homoserine Lactones." *FEMS Microbiology Reviews* 29 (5): 935–47. https://doi.org/10.1016/j.femsre.2005.03.001.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. "Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7 (1): 539. https://doi.org/10.1038/msb.2011.75.

Siezen, Roland J., Jumamurat R. Bayjanov, Giovanna E. Felis, Marijke R. van der Sijde, Marjo Starrenburg, Douwe Molenaar, Michiel Wels, Sacha A. F. T. van Hijum, and Johan E. T. van Hylckama Vlieg. 2011. "Genome-Scale Diversity and Niche Adaptation Analysis of Lactococcus Lactis by Comparative Genome Hybridization Using Multi-Strain Arrays." *Microbial Biotechnology* 4 (3): 383–402. https://doi.org/10.1111/j.1751-7915.2011.00247.x.

Singh, Pranveer, Likhesh Sharma, S. Rajendra Kulothungan, Bharat V. Adkar, Ravindra Singh Prajapati, P. Shaik Syed Ali, Beena Krishnan, and Raghavan Varadarajan. 2013. "Effect of Signal Peptide on Stability and Folding of Escherichia Coli Thioredoxin." *PLoS ONE* 8 (5). https://doi.org/10.1371/journal.pone.0063442.

Smith, Chad C., Robert B. Srygley, Frank Healy, Karthikeyan Swaminath, and Ulrich G. Mueller. 2017. "Spatial Structure of the Mormon Cricket Gut Microbiome and Its Predicted Contribution to Nutrition and Immune Function." *Frontiers in Microbiology* 8 (MAY). https://doi.org/10.3389/fmicb.2017.00801.

Smith, Matthew A., Andrea Rentmeister, Christopher D. Snow, Timothy Wu, Mary F. Farrow, Florence Mingardon, and Frances H. Arnold. 2012. "A Diverse Set of Family 48 Bacterial Glycoside Hydrolase Cellulases Created by Structure-Guided Recombination." *FEBS Journal* 279 (24): 4453–65. https://doi.org/10.1111/febs.12032.

Sohn, Jang Il, and Jin Wu Nam. 2018. "The Present and Future of de Novo Whole-Genome Assembly." *Briefings in Bioinformatics* 19 (1): 23–40. https://doi.org/10.1093/bib/bbw096.

Soucy, Shannon M., Jinling Huang, and Johann Peter Gogarten. 2015. "Horizontal Gene Transfer: Building the Web of Life." *Nature Reviews Genetics* 16 (8): 472–82. https://doi.org/10.1038/nrg3962.

Souza, Wagner Rodrigo de. 2013. "Microbial Degradation of Lignocellulosic Biomass." In *Sustainable Degradation of Lignocellulosic Biomass - Techniques, Applications and Commercialization*. https://doi.org/10.5772/54325.

Sowani, Harshada, Mohan Kulkarni, and Smita Zinjarde. 2018. "An Insight into the Ecology, Diversity and Adaptations of Gordonia Species." *Critical Reviews in Microbiology* 44 (4): 393–413. https://doi.org/10.1080/1040841X.2017.1418286.

Spriestersbach, Anne, Jan Kubicek, Frank Schäfer, Helena Block, and Barbara Maertens. 2015. "Purification of His-Tagged Proteins." In *Methods in Enzymology*, 559:1–15. Academic Press Inc. https://doi.org/10.1016/bs.mie.2014.11.003.

Stark, Lucy, Tina Giersch, and Röbbe Wünschiers. 2014. "Efficiency of RNA Extraction from Selected

Bacteria in the Context of Biogas Production and Metatranscriptomics." *Anaerobe* 29 (October): 85–90. https://doi.org/10.1016/j.anaerobe.2013.09.007.

Stevenson, Gordon, Kanella Andrianopoulos, Matthew Hobbs, and Peter R Reeves. 2006. "Organization of the Escherichia Coli K-12 Gene Cluster Responsible for Production of the Extracellular Polysaccharide Colanic Acid Downloaded from Http://Jb.Asm.Org/ on February 24 , 2015 by Tamil Nadu Veterinary & Animal Science University." *Journal of Bacteriology* 178 (16): 4885–93. http://jb.asm.org/.

Stewart, Christopher J., Nicholas D. Embleton, Emma C.L. Marrs, Daniel P. Smith, Tatiana Fofanova, Andrew Nelson, Tom Skeath, et al. 2017. "Longitudinal Development of the Gut Microbiome and Metabolome in Preterm Neonates with Late Onset Sepsis and Healthy Controls." *Microbiome* 5 (1): 75. https://doi.org/10.1186/s40168-017-0295-1.

Streit, Wolfgang R., and Ruth A. Schmitz. 2004. "Metagenomics - The Key to the Uncultured Microbes." *Current Opinion in Microbiology* 7 (5): 492–98. https://doi.org/10.1016/j.mib.2004.08.002.

Su, Lijuan, Lele Yang, Shi Huang, Yan Li, Xiaoquan Su, Fengqin Wang, Cunpei Bo, En Tao Wang, and Andong Song. 2017. "Variation in the Gut Microbiota of Termites (Tsaitermes Ampliceps) Against Different Diets." *Applied Biochemistry and Biotechnology* 181 (1): 32–47. https://doi.org/10.1007/s12010-016-2197-2.

Suen, Garret, Jarrod J. Scott, Frank O. Aylward, Sandra M. Adams, Susannah G. Tringe, Adrián A. Pinto-Tomás, Clifton E. Foster, et al. 2010. "An Insect Herbivore Microbiome with High Plant Biomass-Degrading Capacity." *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1001129.

Sundset, Monica A., Kirsti E. Præsteng, Isaac K.O. Cann, Svein D. Mathiesen, and Roderick I. MacKie. 2007. "Novel Rumen Bacterial Diversity in Two Geographically Separated Sub-Species of Reindeer." *Microbial Ecology* 54 (3): 424–38. https://doi.org/10.1007/s00248-007-9254-x.

Suresh, Cuddapah, Ahmed Abu Rus'd, Motomitsu Kitaoka, and Kiyoshi Hayashi. 2002. "Evidence That the Putative α-Glucosidase of Thermotoga Maritima MSB8 Is a PNP α-D-Glucuronopyranoside Hydrolyzing α-Glucuronidase." *FEBS Letters* 517 (1–3): 159–62. https://doi.org/10.1016/S0014-5793(02)02611-X.

Suring, Wouter, Janine Mariën, Rhody Broekman, Nico M. Van Straalen, and Dick Roelofs. 2016. "Biochemical Pathways Supporting Beta-Lactam Biosynthesis in the Springtail Folsomia Candida." *Biology Open* 5 (12): 1784–89. https://doi.org/10.1242/bio.019620.

Suring, Wouter, Karen Meusemann, Alexander Blanke, Janine Mariën, Tim Schol, Valeria

Agamennone, Anna Faddeeva-Vakhrusheva, et al. 2017. “Evolutionary Ecology of Beta-Lactam Gene Clusters in Animals.” *Molecular Ecology* 26 (12): 3217–29. https://doi.org/10.1111/mec.14109.

Susmel, P., and B. Stefanon. 1993. “Aspects of Lignin Degradation by Rumen Microorganisms.” *Journal of Biotechnology* 30 (1): 141–48. https://doi.org/10.1016/0168-1656(93)90035-L.

Sützl, Leander, Christophe V.F.P. Laurent, Annabelle T. Abrera, Georg Schütz, Roland Ludwig, and Dietmar Haltrich. 2018. “Multiplicity of Enzymatic Functions in the CAZy AA3 Family.” *Applied Microbiology and Biotechnology* 102 (6): 2477–92. https://doi.org/10.1007/s00253-018-8784-0.

Sweeney, Matt D., and Feng Xu. 2012. “Biomass Converting Enzymes as Industrial Biocatalysts for Fuels and Chemicals: Recent Developments.” *Catalysts* 2 (2): 244–63. https://doi.org/10.3390/catal2020244.

Tai, Vera, Erick R. James, Christine A. Nalep, Rudolf H. Scheffrahn, Steve J. Perlman, and Patrick J. Keelinga. 2015. “The Role of Host Phylogeny Varies in Shaping Microbial Diversity in the Hindguts of Lower Termites.” Edited by C. R. Lovell. *Applied and Environmental Microbiology* 81 (3): 1059–70. https://doi.org/10.1128/AEM.02945-14.

Tajima, K., R. I. Aminov, T. Nagamine, H. Matsui, M. Nakamura, and Y. Benno. 2001. “Diet-Dependent Shifts in the Bacterial Population of the Rumen Revealed with Real-Time PCR.” *Applied and Environmental Microbiology* 67 (6): 2766–74. https://doi.org/10.1128/AEM.67.6.2766-2774.2001.

Tauzin, Alexandra S., Elisabeth Laville, Yao Xiao, Sébastien Nouaille, Pascal Le Bourgeois, Stéphanie Heux, Jean Charles Portais, et al. 2016. “Functional Characterization of a Gene Locus from an Uncultured Gut Bacteroides Conferring Xylo-Oligosaccharides Utilization to Escherichia Coli.” *Molecular Microbiology* 102 (4): 579–92. https://doi.org/10.1111/mmi.13480.

Taylor, Martin J., Hassan A. Alabdrabalameer, and Vasiliki Skoulou. 2019. “Choosing Physical, Physicochemical and Chemical Methods of Pre-Treating Lignocellulosic Wastes to Repurpose into Solid Fuels.” *Sustainability (Switzerland)* 11 (13). https://doi.org/10.3390/su11133604.

Taylor, Maureen E., and Kurt Drickamer. 2014. “Convergent and Divergent Mechanisms of Sugar Recognition across Kingdoms.” *Current Opinion in Structural Biology*. Elsevier Ltd. https://doi.org/10.1016/j.sbi.2014.07.003.

Tessler, Michael, Johannes S. Neumann, Ebrahim Afshinnekoo, Michael Pineda, Rebecca Hersch, Luiz Felipe M. Velho, Bianca T. Segovia, et al. 2017. “Large-Scale Differences in Microbial

Biodiversity Discovery between 16S Amplicon and Shotgun Sequencing." *Scientific Reports* 7 (1). https://doi.org/10.1038/s41598-017-06665-3.

Thakur, Abhijeet, Kedar Sharma, and Arun Goyal. 2019. "α-l-Arabinofuranosidase: A Potential Enzyme for the Food Industry." In , 229–44. https://doi.org/10.1007/978-981-13-3263-0_12.

Thimm, Torsten, Andrea Hoffmann, Heinz Borkott, Jean Charles Munch, and Christoph C. Tebbe. 1998. "The Gut of the Soil Microarthropod Folsomia Candida (Collembola) Is a Frequently Changeable but Selective Habitat and a Vector for Microorganisms." *Appl. Environ. Microbiol.* 64 (7): 2660–69. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC106441/pdf/am002660.pdf.

Thursby, Elizabeth, and Nathalie Juge. 2017. "Introduction to the Human Gut Microbiota." *Biochemical Journal*. Portland Press Ltd. https://doi.org/10.1042/BCJ20160510.

Tian, Shijing, Muhammad Ali, Li Xie, and Lin Li. 2016. "Genome-Sequence Analysis of Acinetobacter Johnsonii MB44 Reveals Potential Nematode-Virulent Factors." *SpringerPlus*. https://doi.org/10.1186/s40064-016-2668-5.

Till, M., D. Goldstone, G. Card, G. T. Attwood, C. D. Moon, and V. L. Arcus. 2014. " Structural Analysis of the GH43 Enzyme Xsa43E from Butyrivibrio Proteoclasticus ." *Acta Crystallographica Section F Structural Biology Communications* 70 (9): 1193–98. https://doi.org/10.1107/s2053230x14014745.

Tiwari, Pragya, B. N. Misra, and Neelam S. Sangwan. 2013. "β -Glucosidases from the Fungus Trichoderma: An Efficient Cellulase Machinery in Biotechnological Applications." *BioMed Research International* 2013. https://doi.org/10.1155/2013/203735.

Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. "MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling." *Nature Methods*. Nature Publishing Group. https://doi.org/10.1038/nmeth.3589.

Tsirigotaki, Alexandra, Jozefien De Geyter, Nikolina Šoštarić, Anastassios Economou, and Spyridoula Karamanou. 2017. "Protein Export through the Bacterial Sec Pathway." *Nature Reviews Microbiology* 15 (1): 21–36. https://doi.org/10.1038/nrmicro.2016.161.

Um, Soohyun, Antoine Fraimout, Panagiotis Sapountzis, Dong Chan Oh, and Michael Poulsen. 2013. "The Fungus-Growing Termite Macrotermes Natalensis Harbors Bacillaene-Producing Bacillus Sp. That Inhibit Potentially Antagonistic Fungi." *Scientific Reports* 3 (1): 1–7. https://doi.org/10.1038/srep03250.

Valdes, Ana M., Jens Walter, Eran Segal, and Tim D. Spector. 2018. "Role of the Gut Microbiota in

Nutrition and Health." *BMJ* 361: 36–44. https://doi.org/10.1136/bmj.k2179.

Valvano, Miguel A, Sarah E Furlong, and Kinnari B Patel. 2011. *Bacterial Lipopolysaccharides*. https://doi.org/10.1007/978-3-7091-0733-1.

Vandermarliere, Elien, Tine M. Bourgois, Martyn D. Winn, Steven Van Campenhout, Guido Volckaert, Jan A. Delcour, Sergei V. Strelkov, Anja Rabijns, and Christophe M. Courtin. 2009. "Structural Analysis of a Glycoside Hydrolase Family 43 Arabinoxylan Arabinofuranohydrolase in Complex with Xylotetraose Reveals a Different Binding Mechanism Compared with Other Members of the Same Family." *Biochemical Journal* 418 (1): 39–47. https://doi.org/10.1042/BJ20081256.

Várnai, Anikó, Laura Huikko, Jaakko Pere, Matti Siika-aho, and Liisa Viikari. 2011. "Synergistic Action of Xylanase and Mannanase Improves the Total Hydrolysis of Softwood." *Bioresource Technology* 102 (19): 9096–9104. https://doi.org/10.1016/j.biortech.2011.06.059.

Vásquez, Alejandra, Eva Forsgren, Ingemar Fries, Robert J. Paxton, Emilie Flaberg, Laszlo Szekely, and Tobias C. Olofsson. 2012. "Symbionts as Major Modulators of Insect Health: Lactic Acid Bacteria and Honeybees." *PLoS ONE* 7 (3). https://doi.org/10.1371/journal.pone.0033188.

Vinés, Enrique D., Cristina L. Marolda, Aran Balachandran, and Miguel A. Valvano. 2005. "Defective O-Antigen Polymerization in TolA and Pal Mutants of Escherichia Coli in Response to Extracytoplasmic Stress." *Journal of Bacteriology* 187 (10): 3359–68. https://doi.org/10.1128/JB.187.10.3359-3368.2005.

Wade, W. 2002. "Unculturable Bacteria - The Uncharacterized Organisms That Cause Oral Infections." In *Journal of the Royal Society of Medicine*, 95:81–83. Royal Society of Medicine Press. https://doi.org/10.1258/jrsm.95.2.81.

Wang, Guozeng, Huiying Luo, Kun Meng, Yaru Wang, Huoqing Huang, Pengjun Shi, Xia Pan, et al. 2011. "High Genetic Diversity and Different Distributions of Glycosyl Hydrolase Family 10 and 11 Xylanases in the Goat Rumen." *PLoS ONE* 6 (2). https://doi.org/10.1371/journal.pone.0016731.

Wang, Guozeng, Huiying Luo, Yaru Wang, Huoqing Huang, Pengjun Shi, Peilong Yang, Kun Meng, Yingguo Bai, and Bin Yao. 2011. "A Novel Cold-Active Xylanase Gene from the Environmental DNA of Goat Rumen Contents: Direct Cloning, Expression and Enzyme Characterization." *Bioresource Technology* 102 (3): 3330–36. https://doi.org/10.1016/j.biortech.2010.11.004.

Wang, Lin, Yu Feng, Jianqing Tian, Meichun Xiang, Jingzu Sun, Jianqing Ding, Wen Bing Yin, Marc Stadler, Yongsheng Che, and Xingzhong Liu. 2015. "Farming of a Defensive Fungal Mutualist

by an Attelabid Weevil." *ISME Journal* 9 (8): 1793–1801. https://doi.org/10.1038/ismej.2014.263.

Wang, Lingling, Ayat Hatem, Umit V. Catalyurek, Mark Morrison, and Zhongtang Yu. 2013. "Metagenomic Insights into the Carbohydrate-Active Enzymes Carried by the Microorganisms Adhering to Solid Digesta in the Rumen of Cows." Edited by Hauke Smidt. *PLoS ONE* 8 (11): e78507. https://doi.org/10.1371/journal.pone.0078507.

Wang, Xue, Haibo Zhou, Hanna Chen, Xiaoshu Jing, Wentao Zheng, Ruijuan Li, Tao Sun, et al. 2018. "Discovery of Recombinases Enables Genome Mining of Cryptic Biosynthetic Gene Clusters in Burkholderiales Species." *Proceedings of the National Academy of Sciences of the United States of America* 115 (18): E4255–63. https://doi.org/10.1073/pnas.1720941115.

Wang, Yaru, Huiying Luo, Wenxia Yang, Peilong Yang, Pengjun Shi, Bin Yao, Huoqing Huang, Kun Meng, and Yingguo Bai. 2015. "A Novel Bifunctional GH51 Exo-α-l-Arabinofuranosidase/Endo-Xylanase from Alicyclobacillus Sp. A4 with Significant Biomass-Degrading Capacity." *Biotechnology for Biofuels* 8 (1): 197. https://doi.org/10.1186/s13068-015-0366-0.

Wang, Yong, Jiao Mei Huang, Ying Li Zhou, Alexandre Almeida, Robert D. Finn, Antoine Danchin, and Li Sheng He. 2020. "Phylogenomics of Expanding Uncultured Environmental Tenericutes Provides Insights into Their Pathogenicity and Evolutionary Relationship with Bacilli." *BMC Genomics* 21 (1): 408. https://doi.org/10.1186/s12864-020-06807-4.

Warnecke, Falk, Peter Luginbühl, Natalia Ivanova, Majid Ghassemian, Toby H. Richardson, Justin T. Stege, Michelle Cayouette, et al. 2007. "Metagenomic and Functional Analysis of Hindgut Microbiota of a Wood-Feeding Higher Termite." *Nature* 450 (7169): 560–65. https://doi.org/10.1038/nature06269.

Waterhouse, Andrew, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T. Heer, et al. 2018. "SWISS-MODEL: Homology Modelling of Protein Structures and Complexes." *Nucleic Acids Research* 46 (W1): W296–303. https://doi.org/10.1093/nar/gky427.

Watford, Shelby, and Steven J. Warrington. 2018. *Bacterial DNA Mutations*. *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/pubmed/29083710.

Weber, Tilmann, Kai Blin, Srikanth Duddela, Daniel Krug, Hyun Uk Kim, Robert Bruccoleri, Sang Yup Lee, et al. 2015. "AntiSMASH 3.0-A Comprehensive Resource for the Genome Mining of Biosynthetic Gene Clusters." *Nucleic Acids Research* 43 (W1): W237–43. https://doi.org/10.1093/nar/gkv437.

Whitney, J. C., and P. L. Howell. 2013. “Synthase-Dependent Exopolysaccharide Secretion in Gram-Negative Bacteria.” *Trends in Microbiology* 21 (2): 63–72. https://doi.org/10.1016/j.tim.2012.10.001.

Wi, Seung Gon, In Seong Choi, Kyoung Hyoun Kim, Ho Myeong Kim, and Hyeun Jong Bae. 2013. “Bioethanol Production from Rice Straw by Popping Pretreatment.” *Biotechnology for Biofuels* 6 (1): 166. https://doi.org/10.1186/1754-6834-6-166.

Wickham, Hadley. 2009. *Ggplot2*. *Ggplot2*. Springer New York. https://doi.org/10.1007/978-0-387-98141-3.

Wilkins, Laetitia G.E., Matthieu Leray, Aaron O’Dea, Benedict Yuen, Raquel S. Peixoto, Tiago J. Pereira, Holly M. Bik, et al. 2019. “Host-Associated Microbiomes Drive Structure and Function of Marine Ecosystems.” *PLoS Biology* 17 (11): e3000533. https://doi.org/10.1371/journal.pbio.3000533.

Wilmanski, Tomasz, Christian Diener, Noa Rappaport, Sushmita Patwardhan, Jack Wiedrick, Jodi Lapidus, John C. Earls, et al. 2020. “Gut Microbiome Pattern Reflects Healthy Aging and Predicts Extended Survival in Humans.” *BioRxiv* 3 (2): 274–86. https://doi.org/10.1101/2020.02.26.966747.

Wilson, David B. 2008. “Three Microbial Strategies for Plant Cell Wall Degradation.” *Annals of the New York Academy of Sciences* 1125: 289–97. https://doi.org/10.1196/annals.1419.026.

Wong, Chun Nin Adam, Patrick Ng, and Angela E. Douglas. 2011. “Low-Diversity Bacterial Community in the Gut of the Fruitfly Drosophila Melanogaster.” *Environmental Microbiology* 13 (7): 1889–1900. https://doi.org/10.1111/j.1462-2920.2011.02511.x.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. “Improved Metagenomic Analysis with Kraken 2.” *BioRxiv* 20 (1): 1–13. https://doi.org/10.1101/762302.

Wood, Derrick E, and Steven L Salzberg. 2014. “Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments.” *Genome Biology* 15 (3): R46. https://doi.org/10.1186/gb-2014-15-3-r46.

Xiao, Chaowen, and Charles T. Anderson. 2013. “Roles of Pectin in Biomass Yield and Processing for Biofuels.” *Frontiers in Plant Science*. Frontiers Research Foundation. https://doi.org/10.3389/fpls.2013.00067.

Xiaohui Zhang, Melissa R Jacob, R Ranga Rao, Yan-Hong Wang, Ameeta K Agarwal, David J Newman, Ikhlas A Khan, Alice M Clark, and Xing-Cong Li. 2012. “Antifungal Cyclic Peptides from the Marine Sponge Microscleroderma Herdmani.” *Research and Reports in Medicinal*

Chemistry 2 (May): 7. https://doi.org/10.2147/rrmc.s30895.

Xie, Shangxian, Ryan Syrenne, Su Sun, and Joshua S. Yuan. 2014. "Exploration of Natural Biomass Utilization Systems (NBUS) for Advanced Biofuel-from Systems Biology to Synthetic Design." *Current Opinion in Biotechnology*. Elsevier Ltd. https://doi.org/10.1016/j.copbio.2014.02.007.

Xu, Jing, Han Zhang, Jinfang Zheng, Philippe Dovoedo, and Yanbin Yin. 2020. "ECAMI: Simultaneous Classification and Motif Identification for Enzyme Annotation." Edited by Jinbo Xu. *Bioinformatics* 36 (7): 2068–75. https://doi.org/10.1093/bioinformatics/btz908.

Xu, Zhongli, Kathrin Jakobi, Katrin Welzel, and Christian Hertweck. 2005. "Biosynthesis of the Antitumor Agent Chartreusin Involves the Oxidative Rearrangement of an Anthracyclic Polyketide." *Chemistry and Biology* 12 (5): 579–88. https://doi.org/10.1016/j.chembiol.2005.04.017.

Yadav, Meera, and H. S. Yadav. 2015. "Applications of Ligninolytic Enzymes to Pollutants, Wastewater, Dyes, Soil, Coal, Paper and Polymers." *Environmental Chemistry Letters*. Springer Verlag. https://doi.org/10.1007/s10311-015-0516-4.

Yang, Shih-Chun, Chih-Hung Lin, Calvin T. Sung, and Jia-You Fang. 2014. "Antibacterial Activities of Bacteriocins: Application in Foods and Pharmaceuticals." *Frontiers in Microbiology* 5 (MAY): 241. https://doi.org/10.3389/fmicb.2014.00241.

Ye, Xiaokun, Zhen Zhang, Yuancai Chen, Jiaqi Cheng, Zhenghua Tang, and Yongyou Hu. 2016. "Physico-Chemical Pretreatment Technologies of Bioconversion Efficiency of Paulownia Tomentosa (Thunb.) Steud." *Industrial Crops and Products* 87 (September): 280–86. https://doi.org/10.1016/j.indcrop.2016.04.045.

Yin, Yanbin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu. 2012. "DbCAN: A Web Resource for Automated Carbohydrate-Active Enzyme Annotation." *Nucleic Acids Research* 40 (W1). https://doi.org/10.1093/nar/gks479.

Yoav, Shahar, Johanna Stern, Orly Salama-Alber, Felix Frolow, Michael Anbar, Alon Karpol, Yitzhak Hadar, Ely Morag, and Edward A. Bayer. 2019. "Directed Evolution of Clostridium Thermocellum β-Glucosidase a towards Enhanced Thermostability." *International Journal of Molecular Sciences* 20 (19): 4701. https://doi.org/10.3390/ijms20194701.

Yonehara, Hiroshi, Haruo Seto, Shojiro Aizawa, Tetsuro Hidaka, Akira Shimazu, and Noboru Ōtake. 1968. "The Detoxin Complex, Selective Antagonists of Blasticidin S." *Journal of Antibiotics*. J Antibiot (Tokyo). https://doi.org/10.7164/antibiotics.21.369.

Yoon, Sang Hwal, Tae Seok Moon, Pooya Iranpour, Amanda M Lanza, and Kristala Jones Prather.

2009. “Cloning and Characterization of Uronate Dehydrogenases from Two Pseudomonads and Agrobacterium Tumefaciens Strain C58.” *Journal of Bacteriology* 191 (5): 1565–73. https://doi.org/10.1128/JB.00586-08.

Yoshida, Yuki, Georgios Koutsovoulos, Dominik R. Laetsch, Lewis Stevens, Sujai Kumar, Daiki D. Horikawa, Kyoko Ishino, et al. 2017. *Comparative Genomics of the Tardigrades Hypsibius Dujardini and Ramazzottius Varieornatus*. *BioRxiv*. Vol. 15. Public Library of Science. https://doi.org/10.1101/112664.

You, Minsheng, Zhen Yue, Weiyi He, Xinhua Yang, Guang Yang, Miao Xie, Dongliang Zhan, et al. 2013. “A Heterozygous Moth Genome Provides Insights into Herbivory and Detoxification.” *Nature Genetics* 45 (2): 220–25. https://doi.org/10.1038/ng.2524.

Yun, Ji Hyun, Seong Woon Roh, Tae Woong Whon, Mi Ja Jung, Min Soo Kim, Doo Sang Park, Changmann Yoon, et al. 2014. “Insect Gut Bacterial Diversity Determined by Environmental Habitat, Diet, Developmental Stage, and Phylogeny of Host.” *Applied and Environmental Microbiology* 80 (17): 5254–64. https://doi.org/10.1128/AEM.01226-14.

Zaide, Galia, Dalia Shallom, Smadar Shulami, Gennady Zolotnitsky, Gali Golan, Timor Baasov, Gil Shoham, and Yuval Shoham. 2001. “Biochemical Characterization and Identification of Catalytic Residues in α-Glucuronidase from Bacillus Stearothermophilus T-6.” *European Journal of Biochemistry* 268 (10): 3006–16. https://doi.org/10.1046/j.1432-1327.2001.02193.x.

Zhang, Han, Tanner Yohe, Le Huang, Sarah Entwistle, Peizhi Wu, Zhenglu Yang, Peter K. Busk, Ying Xu, and Yanbin Yin. 2018. “DbCAN2: A Meta Server for Automated Carbohydrate-Active Enzyme Annotation.” *Nucleic Acids Research* 46 (W1): W95–101. https://doi.org/10.1093/nar/gky418.

Zhao, Zhongtao, Huiquan Liu, Chenfang Wang, and Jin Rong Xu. 2013. “Comparative Analysis of Fungal Genomes Reveals Different Plant Cell Wall Degrading Capacity in Fungi.” *BMC Genomics* 14 (1): 274. https://doi.org/10.1186/1471-2164-14-274.

Zheng, Yu, Ayana Saitou, Chiung Mei Wang, Atsushi Toyoda, Yohei Minakuchi, Yuji Sekiguchi, Kenji Ueda, et al. 2019. “Genome Features and Secondary Metabolites Biosynthetic Potential of the Class Ktedonobacteria.” *Frontiers in Microbiology* 10 (APR): 893. https://doi.org/10.3389/fmicb.2019.00893.

Zhou, Tao, Yemin Xue, Fengjiao Ren, and Yuanyuan Dong. 2018. “Antioxidant Activity of Xylooligosaccharides Prepared from Thermotoga Maritima Using Recombinant Enzyme Cocktail of β-Xylanase and α-Glucuronidase.” *Journal of Carbohydrate Chemistry* 37 (4): 210–24. https://doi.org/10.1080/07328303.2018.1455843.

Zhu, Ning, Jinshui Yang, Lei Ji, Jiawen Liu, Yi Yang, and Hongli Yuan. 2016. “Metagenomic and Metaproteomic Analyses of a Corn Stover-Adapted Microbial Consortium EMSD5 Reveal Its Taxonomic and Enzymatic Basis for Degrading Lignocellulose.” *Biotechnology for Biofuels* 9 (1). https://doi.org/10.1186/s13068-016-0658-z.

Zhu, Wenhan, Alexandre Lomsadze, and Mark Borodovsky. 2010. “Ab Initio Gene Identification in Metagenomic Sequences.” *Nucleic Acids Research* 38 (12): e132. https://doi.org/10.1093/nar/gkq275.

Zhu, Zhi, Suqin Hang, Shengyong Mao, and Weiyun Zhu. 2014. “Diversity of Butyrivibrio Group Bacteria in the Rumen of Goats and Its Response to the Supplementation of Garlic Oil.” *Asian-Australasian Journal of Animal Sciences* 27 (2): 179–86. https://doi.org/10.5713/ajas.2013.13373.

Summary

ISOLATION AND CHARACTERIZATION OF NOVEL ENZYMATIC ACTIVITIES FROM GUT METAGENOMES TO SUPPORT LIGNOCELLULOSE BREAKDOWN

The agricultural sector produces a large amount of organic waste as by-products (crop remains, foliage, seed pods, straw, etc.). Currently, these materials are not properly treated and their uncontrolled disposal can lead to many problems. In many countries crop remains are burned on the field, sometimes causing severe air pollution as well as damage health. This contributes to climate change and could impact the climate further to the point of irreversible damage. To realize a sustainable agriculture, organic wastes should not be disposed or burned but used as a cheap source for biomaterials. In line with the philosophy of the bio-based economy, agricultural waste is recycled and used as raw material in the chemical industry, replacing fossil fuels. However, the process of converting agricultural waste into useful products is not efficient and can be improved further for optimization.

One of the dominant components of agricultural waste is lignocellulose, a complex biomaterial that is difficult to handle. To break down this complex structure, large amounts of energy or chemicals for treatment are required. This thesis aims to explore novel natural biological catalysts that can help to degrade lignocellulose and deliver useful compounds for the bio-based chemical industry.

To discover such novel catalysts, I looked at the digestive systems of a number of different animals: goats, springtails, isopods and termites. Animal guts are mini ecosystems that contain many unknown interesting bacteria. These microorganisms are adapted to the host and might have interesting properties that can be explored. By looking at these organisms and their catalysts it is possible to identify and mine novel genes that can be used to breakdown biomass. This process creates substrates that can be used for many other procedures.

In this thesis my main focus was on enzymes that can break down carbohydrates. The various bonds in complex carbohydrate molecules are cleaved by different enzymes. Every bacterium has a suit of carbohydrate-active enzymes, called CAZymes. Using metagenomics and

bioinformatic tools, I explored the genomes of microbial communities in search of novel CAZymes. Unlike traditional culturing methods, metagenomics is aimed at the whole genome of the communities involved, that is, all bacteria jointly. In addition, I also investigated genes encoding antibiotic resistance, and the production of secondary metabolites since these two gene categories greatly contribute to the survival of bacteria in complex microbial communities. This provided a better understanding about the bacterial contribution to the host and within the bacterial community.

In Chapter 2 a metagenomic approach was applied to identify the bacterial species and their gene complements in the guts of the Ninh Binh's mountain goats. These goats from Vietnam feed on grass as well as woody plants and so have a large number of carbohydrate degrading enzymes. In our survey, we identified 821 carbohydrate esterases and polysaccharide lyases, 816 cellulases and 2,252 hemicelullases. A promising protein with a carbohydrate-binding domain was recombinantly expressed in *Escherichia coli* and its catalytic activity studied. The protein accelerated the action of a commercial cellulase in the degradation of paper.

In Chapter 3, we looked at the functional potential of the microbiome associated with the springtail *Folsomia candida*. Springtails (Collembola) are invertebrates that feed on dead organic material and fungi and so are expected to harbor a specialized microbial community to aid in digestion. Using a bioinformatics approach, we focused on carbohydrate metabolism functions, as well as antibiotic biosynthesis gene clusters and secondary metabolites. In the microbiome, we found several genes with strong homology to genes in the *F. candida* genome, that were previously identified as genes resulting from horizontal gene transfer. These genes are part of the soil-adaptive repertoire as they help the springtails to degrade recalcitrant compounds and defense against pathogens. The microbiome constitutes an important source of new functions for the springtail.

In Chapter 4, the gut microbiomes of the three invertebrates: springtails, isopods and termites were compared. While springtails are mostly fungivorous, isopods are detritivores and termites can degrade woody materials. The analysis revealed an enormous diversity of gut bacteria. Interestingly, the core enzymes for breaking down carbohydrate are similar in the three hosts. In addition to the core complement, each species had 10-30 CAZy families specific to that species. The diversity of organic matter breakdown potential is much greater than commonly assumed in ecological studies.

Chapter 5 looks at two interesting hemicellulase genes, an α-L-arabinofuranosidase and an α-glucuronidase, identified in the previous chapter. These genes were isolated, cloned and expressed in a recombinant system. Their activity as a function of the substrate concentration was characterized and the Michaelis-Menten parameters estimated. The data showed how efficient bioinformatic tools can be used to identify novel and active genes.

In Chapter 6 we compared the α-L-arabinofuranosidase from the springtail with a similar gene found in the termite. This gene was predicted to be a functional novel horizontal gene, which was transferred from long time ago. Since its uptake in a eukaryotic genome, the gene was modified to fit the new host, as it contained a 19 amino acid signal peptide, normally only found in eukaryotes, which targets the gene product for excretion. This illustrates how the springtail has adapted to the soil environment by recruiting and modifying genes from its microbiome.

The work on this thesis shows the possibilities of using bioinformatic tools to investigate the microbiome communities and mine for interesting enzymes from metagenomes. The bacterial community appears to be very diverse and different between hosts. However, at the enzymatic level there is a core group of carbohydrate enzymes and antibiotic resistances. Some of the studied enzymes show the potential for bio-applications. The method could be further tested on different animal metagenomes.

With the expansion of sequencing and bioinformatic tools as well as advances in computing and machine learning, it is possible to use and understand more about the natural environment. These tools could help to mine enzymes with interesting properties. Together with enzyme characterization, the bioinformatic tools could be improved further. When the whole process is streamlined and part of a pipeline, a large number of enzymes can be identified and tested to find optimal conditions for their action. These enzymes can be combined together to create cocktails, which can efficiently breakdown biomass. Since the enzymes are natural and the products are used as substrates, little energy and resources are wasted. Beneficial enzymes and bacteria are also preserved and promoted. By creating a recycling plant, agricultural waste can turn into new substrates and products, which in turn can help to improve the environment.

Samenvatting

ISOLATIE EN KARAKTERISERING VAN NIEUWE ENZYMATISCHE ACTIVITEITEN IN DARM-METAGENOMEN TEN BEHOEVE VAN DE AFBRAAK VAN LIGNOCELLULOSE

De landbouwsector produceert als nevenproduct een grote hoeveelheid organisch afval (overblijfselen van gewassen, loof, zaadhuiden, stro, enz.). Op dit moment wordt dit materiaal niet correct behandeld; deze ongecontroleerde afvalbehandeling kan milieuproblemen veroorzaken. In veel landen worden de organische restanten verbrand op het veld en veroorzaken luchtverontreiniging en bedreigen de volksgezondheid. Dit draagt ook bij aan klimaatverandering en kan het klimaat zelfs brengen naar het punt van een onomkeerbare verandering. Voor een duurzame landbouw moet het afval niet weggegooid of verbrand worden, maar gebruikt als een goedkope bron van biomaterialen. In lijn met het uitgangspunt van de bio-gebaseerde economie, moet organisch afval hergebruikt worden als uitgangsmateriaal in de chemische industrie, ter vervanging van fossiele brandstoffen. Echter, het proces om landbouw-afval om te zetten in waardevolle producten is niet erg efficiënt en moet verder geoptimaliseerd worden.

Een van de dominante bestanddelen van landbouw-afval is lignocellulose, een complexe biologische materie waar niet goed mee te werken is. Om deze complexe structuur af te breken zijn behandelingen nodig die veel energie vergen of veel chemicaliën. Dit proefschrift stelt zich tot doel om nieuwe biologische katalysatoren te verkennen die kunnen helpen om lignocellulose af te breken en waardevolle bestanddelen voor de bio-gebaseerde chemische industrie te leveren.

Om zulke nieuwe katalysatoren te ontdekken heb ik gekeken naar de verteringsstelsels van een aantal verschillende dieren: geiten, springstaarten, isopoden en termieten. De darm van een dier is een mini-ecosysteem dat vele onbekende bacteriën bevat. Door te kijken naar deze organismen en hun enzymen is het mogelijk om nieuwe genen te identificeren die gebruikt kunnen worden om biomassa af te breken. Dit proces levert substraten die gebruikt kunnen worden voor vele andere procedures.

Mijn belangrijkste doel in dit proefschrift was om enzymen te vinden die koolhydraten kunnen afbreken. De verschillende chemische bindingen in complexe koolhydraat-moleculen worden door verschillende enzymen gesplitst. Elke bacterie heeft een verzameling van koolhydraat-actieve enzymen, genoemd CAZymes. Met behulp van metagenomica en bioïnformatische technieken heb ik de genomen van microbiële gemeenschappen onderzocht, op zoek naar nieuwe CAZymes. In tegenstelling tot traditionele microbiële kweekmethodes is metagenomica gericht op het hele genoom van de betreffende levensgemeenschap, dat wil zeggen, alle bacteriën gezamenlijk. Bovendien heb ik ook genen onderzocht die coderen voor resistentie tegen antibiotica en genen die betrokken zijn bij de productie van secundaire metabolieten, aangezien deze twee gencategorieën belangrijk bijdragen aan de overleving van bacteriën in complexe microbiële gemeenschappen. Hiermee werd een beter begrip verkregen van de bijdrage van bacteriën aan de gastheer en aan de bacteriële gemeenschap.

In Hoofdstuk 2 heb ik een metagenomica-benadering toegepast om de bacteriële soorten en hun verzameling genen te identificeren in de darm van Ninh-Binh-berggeiten. Deze Vietnamese dieren eten gras en houtige planten en hebben daarom een groot aantal koolhydraat-afbrekende enzymen. In ons genoomonderzoek identificeerden we 821 koolhydraat-esterases en polysaccharide-lyases, 816 cellulases and 2,252 hemicellulases. Een veelbelovend koolhydraat-bindend eiwit werd recombinant tot expressie gebracht in *Escherichia coli* en de katalytische activiteit werd bestudeerd. Het eiwit versnelde de werking van een commercieel cellulase bij de afbraak van papier.

In Hoofdstuk 3 keken we naar het functionele potentieel van het microbioom geassocieerd met de springstaart *Folsomia candida*. Springstaarten (Collembola) zijn evertebraten die zich voeden met dood organisch materiaal en schimmels en daarom naar verwachting beschikken over een microbiële gemeenschap die de spijsvertering ondersteunt. Met gebruikmaking van de bioïnformatica richtten we onze aandacht op functies in het koolhydraatmetabolisme en ook op genclusters betrokken bij de synthese van antibiotica en secundaire metabolieten. In het microbioom vonden we verschillende genen die sterk homoloog waren met genen in het genoom van *F. candida* die eerder gekwalificeerd waren als horizontaal overgedragen. Deze genen zijn onderdeel van het bodem-adaptieve repertoire, het gencomplement dat de springstaarten helpt om recalcitrante verbindingen af te breken en zich te verdedigen tegen

pathogenen. Het microbioom vertegenwoordigt een belangrijke bron van nieuwe functies voor de springstaart.

In Hoofdstuk 4 werden de darm-microbiomen van drie evertebraten vergeleken: springstaarten, isopoden en termieten. Terwijl springstaarten hoofdzakelijk fungivoor zijn, zijn isopoden detritivoor en termieten kunnen houtig materiaal afbreken. De analyse legde een enorme diversiteit aan darmbacteriën bloot. Interessant was dat de kernenzymen voor de afbraak van koolhydraten van de drie gastheren vergelijkbaar waren. Maar bovenop de kerngenen had elke soort nog 10-30 CAZy-families die specifiek waren voor die soort. De diversiteit van het potentieel om organische stof af te breken is veel groter dan wat gebruikelijk wordt aangenomen in ecologische studies.

Hoofdstuk 5 richtte zich op twee interessante hemicellulase-genen, een alfa-L-arabinofuranosidase en een alfa-glucuronidase, die in het voorgaande hoofdstuk werden geïdentificeerd. Deze genen werden geïsoleerd, gekloneerd en tot expressie gebracht in een recombinant-DNA-systeem. Hun activiteit als functie van de substraatconcentratie werd gekarakteriseerd en de Michaelis-Menten-parameters werden geschat. De gegevens laten zien hoe efficiënt bioïnformatische technieken kunnen zijn bij de identificatie van nieuwe werkzame genen.

In Hoofdstuk 6 vergeleken we het alfa-L-arabinofuranosidase van de springstaart met een soortgelijk gen dat aangetroffen werd in de termiet. Dit gen was eerder aangewezen als een nieuw functioneel gen dat lang geleden in het genoom terecht is gekomen door horizontale genoverdracht. Sinds de opname in een eukaryoot genoom is het gen gemodificeerd en aangepast aan de nieuwe gastheer, namelijk, het bleek een signaalpeptide te bevatten van 19 aminozuren dat normaal gesproken alleen bij eukaryoten voorkomt en dat het genproduct adresseert voor excretie. Dit laat zien hoe de springstaart aangepast is aan het bodemmilieu door uit zijn microbioom genen te recruteren en te modificeren.

Het onderzoek van dit proefschrift illustreert de mogelijkheden om met bioïnformatische technieken microbioomgemeenschappen te onderzoeken en na te speuren op interessante enzymen in het metagenoom. De bacteriële gemeenschap blijkt zeer divers te zijn en te verschillen tussen gastheren. Echter op het niveau van de enzymen is er een gedeelde kerngroep van koolhydraat-geassocieerde enzymen en genen betrokken bij resistentie tegen

antibiotica. Sommige van deze enzymen zijn potentieel biotechnologisch toepasbaar. De bioïnformatische aanpak zou verder getest kunnen worden op andere dierlijke metagenomen.

Met de verdere ontwikkeling van methodes voor sequentiebepaling en bioïnformatische analyse, gekoppeld aan geavanceerde rekenmethodes en machinaal leren, is het mogelijk om het natuurlijk milieu beter te begrijpen. Deze technieken zouden ons kunnen helpen om meer enzymen met interessante eigenschappen op te sporen. Naast de verdere karakterisering van de enzymen kunnen ook de bioïnformatische technieken verbeterd worden. Als het hele proces opgenomen wordt in een gestroomlijnde geautomatiseerde productielijn, kan een groot aantal enzymen ontdekt worden en getest om de optimale condities te vinden voor hun werking. Deze enzymen kunnen gecombineerd worden om mengsels te maken die efficiënt biomassa kunnen afbreken. Aangezien enzymen een natuurlijke oorsprong hebben en de producten direct gebruikt worden als uitgangsmaterialen, gaan er nauwelijks energie en hulpbronnen verloren. Gunstige enzymen en bacteriën worden behouden en gestimuleerd. Door bio-fabrieken op te richten kan landbouw-afval omgezet worden in nieuwe substraten en producten, wat vervolgens kan helpen om het milieu te verbeteren.

Acknowledgement

The most valuable thing I have learned during my PhD is teamwork. This work would not be possible without the help and contribution of so many people, which ranged from scientific guidance, practical support to simple friendship. Due to the sheer amount of the contributors, influencers and helpers during my PhD, it is not possible for me to list all down below. Please know that even if your name does not appear below, I am wholeheartly thankful for all your help.

Foremost, I would like to thank my promoters Nico, Hai, my supervisor Dick and Bram, for giving me the opportunity to do this PhD. Thank you for helping and guiding me all steps of the way. I am grateful for your support, guidance, patience and believing in me during these years. This PhD has been a great learning experience and I am happy to be a part of it.

Dick, I remember meeting you at the BE-BASIC meeting in Vietnam. My first impression was how knowledgeable you are at sequencing. I am so happy and thankful that you let me be a part of your group. Your valuable feedback and encouragement motivated me to excel in completing this interesting and yet challenging research topic. Every time we have issues/tasks/problems, you would always have a solution at hand. You are knowledgeable, funny, enthusiastic and have a passion for music and running. It is truly amazing to see how well you balance work and leisure. Thank you for being there for me.

Nico, to me, you are an exceptional promoter. I feel very fortunate to have you. I remember walking into your office and be amazed by the number of books and PhD students you have over the years. Your knowledge/experience is vast, and it was always fun to talk and discuss matters with you. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I benefited greatly from your wealth of knowledge and meticulous editing. Your passion and perseverance helped me to be here. You were always available for me through thick and thin. I am so happy that this thesis will be a part of your bookshelf. Thank you so much from the bottom of my heart.

I am thankful for my second promoter, Hai. You took me into your lab after I finished with my master degree. You taught me how to do better research and present my work. I also enjoyed all the lessons you taught at different birthday parties and social gathering events.

Your wisdom and attention to detail are amazing. It was under your vision and guidance that I started to focus and expand on bioinformatics. Thank you very much with all my heart.

To Bram, thank you for the collaboration with BE-Basic. Without your bio-based economy vision of the future and this programme, I would not be able to be part of this unforgettable journey.

Thanks to Huyen from the Institute of Technology in Vietnam for teaching, helping and inspiring me throughout these years. Your works are splendid.

I am grateful for the collaborators Valeria, Wouter, Anna and Peter. This work would not be possible without the groundwork from Anna, Valeria and Wouter, who worked on the springtail. Similarly, the research of Thao, Huyen, Quynh Giang, who worked on the termite from the Institute of Technology. Also, thanks to Khoa, Tung Lam, Nguyen, Lam, Huong for helping me on the field trip and preparing the samples of the goat.

To Peter, thank you for sparing your time and energy in helping and teaching me with coding. You opened my eyes with another dimension that I fell in love with instantly. You opened up other possibilities for me with bioinformatics. I will remember the day you helped me with my computer issue while we are countries apart.

A part of my thesis was done at the 02 building. Thanks to Peter van Ulsen from Department of Molecular Microbiology and Rob from Department of Molecular Cell Biology for helping me with setting up and run experiments. You both made the experiences exciting. I enjoyed experimenting, talking and discussing scientific as well as life experiences with you both. Thanks to Wilbert for letting me do my experiments in your lab.

Special thank goes to Renée, Wendy, Claudia, Anouk and Jacintha. Without your help, my wife and kid would not be a part of my daily life here in the Netherlands.

Also, I am thankful to technicians at Animal Ecology, especially Janine, Riet and Rudo, for their assistance and advice. Thanks Gregory from Molecular Microbiology for helping me with the experiments.

During my PhD time at VU and IBT, I had the opportunity to be a supervisor, and I supervised Mohammed and Ly. Thank you both for helping me with my experiments and challenging me to become a better supervisor to other students.

While completing my PhD programme, I also came across many other PhD students and postdocs. I have a great memory with Tjalf, Monica, Astra, Simon, Dre, Oscar, Jeroen, Mark, Claudia, Estfania, and Trang Phan. Thank you for making the department a friendly work environment. I enjoyed your company at lunch, where we had many fun stories and discussions.

To my fellow BE-BASIC PhD students, Hoang Ha, Dat, Long, Anh and Lan Anh, thank you for being part of this journey.

Finally, my sincere thanks go to my family. My parents and in-law family helped my wife and me through the difficult time. Without your help and sacrifice, I would not be here today. Big thanks to my little sister. She gathered all of her high computer skills for assisting me with resizing fonts and formatting this thesis.

A personal thank you to a special person of mine, my wife, Trang, and Khue my adorable daughter for your love, kind understanding and unconditional support. Trang, you look after Khue while studying so I can have more time to complete my own study. Both of you are also an important part of my journey. I am also immensely thankful that finally, you are both here with me.

.

About the Author

Giang Le was born on the 6th October 1986 in Hanoi, Vietnam. He completed his honour bachelor degree in Genetics at Glasgow University from 2006 until 2010. Giang carried on further into his master degree in Medical Genetics at Glasgow University. Upon completion, he returned to Vietnam and worked at the Institute of Biotechnology in Hanoi. While working there, Giang gained more knowledge and experience in molecular genetics using bioinformatics tools to mine for active enzymes. In 2015, he successfully applied for a sandwhich PhD position between Vrije Universiteit Amsterdam, the Netherlands and the Institue of Biotechnology, Vietnam. During his PhD programme, Giang found his passion in bioinformatics, which he used to understand the microbial composition as well as mining for functional enzymes. As of March 2020, he is working full-time at the Maastricht University Medical Center (MUMC+) as a bioinformatician. His current project is to create pipelines to analyse 16S amplicon, outbreak typing and identifying COVID-19 variants.

List of publications

Ngoc Giang Le, Valeria Agamennone, Nico M. van Straalen, Abraham Brouwer, and Dick Roelofs. 2019. "Antimicrobial Activity and Carbohydrate Metabolism in the Bacterial Metagenome of the Soil-Living Invertebrate Folsomia Candida." Scientific Reports 9 (1): 7308. https://doi.org/10.1038/s41598-019-43828-w.

Faddeeva-Vakhrusheva, Anna, Ken Kraaijeveld, Martijn F.L. Derks, Seyed Yahya Anvar, Valeria Agamennone, Wouter Suring, Andries A. Kampfraath, Jacintha Ellers, **Ngoc Giang Le**, et al. 2017. "Coping with Living in the Soil: The Genome of the Parthenogenetic Springtail Folsomia Candida." BMC Genomics 18 (1). https://doi.org/10.1186/s12864-017-3852-x.

Thi Huyen Do, **Ngoc Giang Le**, Trong Khoa Dao, Thi Mai Phuong Nguyen, Tung Lam Le, Han Ly Luu, Khanh Hoang Viet Nguyen, et al. 2018. "Metagenomic Insights into Lignocellulose-Degrading Genes through Illumina Based de Novo Sequencing of the Microbiome in Vietnamese Native Goats' Rumen." Journal of General and Applied Microbiology 64 (3): 108–16. https://doi.org/10.2323/jgam.2017.08.004.

Thi Huyen Do, Trong Khoa Dao, Khanh Hoang Viet Nguyen, **Ngoc Giang Le**, Thi Mai Phuong Nguyen, Tung Lam Le, Thu Nguyet Phung, Nico M. van Straalen, Dick Roelofs, and Nam Hai Truong. 2018. "Metagenomic Analysis of Bacterial Community Structure and Diversity of Lignocellulolytic Bacteria in Vietnamese Native Goat Rumen." Asian-Australasian Journal of Animal Sciences 31 (5): 738–47. https://doi.org/10.5713/ajas.17.0174.

Other publication

Nguyen, Thi Quy, Thu Huong Duong, Thi Ngoc Ha Dang, **Ngoc Giang Le**, Quynh Giang Le, Thi Huyen Do, Van Do Nguyen, Thi Thu Hong Le, and Nam Hai Truong. 2018. "Enhanced Soluble Expression and Effective Purification of Recombinant Human Interleukin-11 by SUMO Fusion in Escherichia Coli." Indian Journal of Biotechnology. Vol. 17. http://nopr.niscair.res.in/bitstream/123456789/45403/1/IJBT 17%284%29 579-585.pdf.

Ngoc Giang Le, Le Thi Hong Minh, Vu Thi Quyen, Nguyen Mai Anh, Nguyen Thi Kim Cuc, and Vu Thi Thu Huyen. 2021. “Genome Mining of a Marine-Derived Streptomyces Sp. PDH23 Isolated from Sponge in Da Nang Sea for Secondary Metabolite Gene Clusters.” Vietnam Journal of Biotechnology 18 (4): 709–21. https://doi.org/10.15625/1811-4989/18/4/14970.