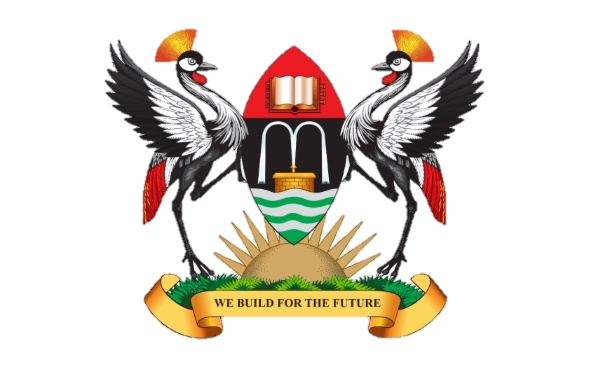
**MAKERERE**  **UNIVERSITY**

**MICROBIOME METATRANSCRIPTOMICS OF THE BLACK SOLDIER FLY LARVAE GUT FOR IDENTIFICATION AND FUNCTIONAL CHARACTERIZATION OF LIGNOCELLULOSIC BIOMASS-DEGRADING MICROBES**

**BY**

**KARIUKI GATHIRWA ERIC (MSC. BIOINFORMATICS)**

**2019/HD07/30034U**

**A DISSERTATION SUBMITTED TO THE DIRECTORATE OF RESEARCH AND GRADUATE TRAINING IN PARTIAL FULFILMENT FOR THE AWARD OF MASTER OF SCIENCE IN BIOINFORMATICS OF MAKERERE UNIVERSITY**

**DECEMBER 2021**

**DECLARATION**

I hereby declare that this submission is my original work and that, to the best of my knowledge and belief, it contains neither resources previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree of the university or other institutes, except for where due acknowledgment has been made in the text.

**Name**: Kariuki Gathirwa Eric **Reg no**: 2019/HD07/30034U

Signature: Date:

I confirm that this project proposal has been submitted with my supervisors’ approval:

1. **Dr. Gerald Mboowa (Ph.D.)**

Department of Immunology, Makerere University

P.O. Box 7062 Kampala, Uganda

Signature: Date:

1. **Dr. Juan C. Paredes (Ph.D.)**

International Center for Insect Physiology and Ecology *(icipe),* Duduville campus, Nairobi.

P.O. Box 30772-00100 Nairobi, Kenya

Signature: Date:

1. **Dr. Caleb Kibet (Ph.D.)**

International Center for Insect Physiology and Ecology *(icipe),* Duduville campus, Nairobi.

P.O. Box 30772-00100 Nairobi, Kenya

Signature: Date:

1. **Dr. Chrysantus Tanga (Ph.D.)**

International Center for Insect Physiology and Ecology *(icipe),* Duduville campus, Nairobi.

P.O. Box 30772-00100 Nairobi, Kenya

Signature: Date:

1. **Dr. Daniel Masiga (Ph.D.)**

International Center for Insect Physiology and Ecology *(icipe),* Duduville campus, Nairobi.

P.O. Box 30772-00100 Nairobi, Kenya

Signature: Date:

**DEDICATION**

This work is dedicated to my late uncle Erastus Githua. Rest in peace uncle. You are a huge part of who I am today.

I would also like to dedicate this work to my brother Brian Muriga. May it quench your curious soul.

**ACKNOWLEDGMENTS**

I wish to express my sincere gratitude to the East African Network for Bioinformatics Training (EANBiT) for offering me the opportunity to further my learning. I would like to thank my supervisors for making the pursuit and completion of this work possible, the International Center for Insect Physiology and Ecology *(icipe),* and its entire fraternity for according me with decent work facilities and a conducive working environment. Special thanks to Dr. Tanga, for his generosity and belief in this work, and Dr. Juan and Dr. Kibet for their valuable ideas and attention to detail. I would also like to acknowledge the Makerere University fraternity for the skills I have gained over my term as an MSc. Bioinformatics student.

I would like to acknowledge my parents, for their prayers and unwavering support even when quitting felt like an option. Their enthusiasm and daily phone calls made this work worth pursuing. I would also like to thank, Sarah Mueni for her prayers, encouragement, and support, Eric Njiraini for standardizing and improving the R analysis scripts, Ibra Lujumba for offering a listening ear and the extra pair of eyes when I needed them, Ian Ochieng’, Stephen Kiniti, and Meshack Mutua for their compassion, Ruth Nanjala for aiding me with GPU resources, Johannes Horstmann for his compliments and believing in my work, my mentees Oscar Mwaura and John Njogu for believing in my research and their great effort in developing the 16S rRNA analysis pipeline, and finally, myself for showing up and making big sacrifices even when the work hurt more than it progressed. I am most grateful to God for the gift of good health and a sober mind when undertaking this work.

**Table of Contents**

[LIST OF ILLUSTRATIONS x](#_Toc92192634)

[LIST OF TABLES xii](#_Toc92192635)

[ABBREVIATIONS xiii](#_Toc92192636)

[ABSTRACT xvi](#_Toc92192637)

[1.0 CHAPTER ONE: INTRODUCTION 1](#_Toc92192638)

[1.1 Background Information 1](#_Toc92192639)

[1.2 Problem Statement 2](#_Toc92192640)

[1.3 Hypothesis 3](#_Toc92192641)

[1.4 General Objective 3](#_Toc92192642)

[To identify and functionally characterize lignocellulosic biomass-degrading microbes and enzymes from the BSF larval gut microbiome. 3](#_Toc92192643)

[1.5 Specific Objectives 3](#_Toc92192644)

[1.6 Significance 4](#_Toc92192645)

[1.7 Justification 4](#_Toc92192646)

[1.8 Conceptual Framework 6](#_Toc92192647)

[2.0 CHAPTER 2: LITERATURE REVIEW 7](#_Toc92192648)

[2.1 Introduction 7](#_Toc92192649)

[2.2 Aims of Feedstock Pretreatment 8](#_Toc92192650)

[2.3 Current Biomass Pretreatment and Conversion Technologies 9](#_Toc92192651)

[2.4 The Black Soldier Fly 10](#_Toc92192652)

[2.5 Diet Selection and Breeding Conditions for the BSF Larvae 11](#_Toc92192653)

[2.6 The Black Soldier Fly larval microbiome 13](#_Toc92192654)

[2.7 RNA-sequencing and Metatranscriptomic Analysis 16](#_Toc92192655)

[2.7.1 The Advancement of Microbiome Research 16](#_Toc92192656)

[2.7.2 Long-read Sequencing 17](#_Toc92192657)

[2.7.3 The ONT MinION Sequencing Platform 18](#_Toc92192658)

[2.7.4 Multiplexing Sequencing Approach 19](#_Toc92192659)

[2.7.5 Ribodepletion in Metatranscriptomic Analysis 19](#_Toc92192660)

[2.7.6 Error Correction 20](#_Toc92192661)

[2.8 Taxonomic Validation using 16S rRNA Sequences 20](#_Toc92192662)

[2.9 Differential Expression of Genes (DEGs) 21](#_Toc92192663)

[2.10 Carbohydrate-Active Enzymes (CAZymes) 22](#_Toc92192664)

[3.0 CHAPTER THREE: MATERIALS AND METHODS 24](#_Toc92192665)

[3.1 Research Design and Approach 24](#_Toc92192666)

[3.2 Ethical Considerations 24](#_Toc92192667)

[3.3 Growth conditions and colony maintenance 24](#_Toc92192668)

[3.4 Sampling 25](#_Toc92192669)

[3.5 RNA isolation 25](#_Toc92192670)

[3.6 cDNA Synthesis and mRNA Enrichment 26](#_Toc92192671)

[3.7 Sequencing 27](#_Toc92192672)

[3.8 Data Analysis 28](#_Toc92192673)

[3.8.1 Physicochemical parameter collection 28](#_Toc92192674)

[3.8.2 Feed composition analysis 28](#_Toc92192675)

[3.8.3 Quality statistics of basecalled reads with pycoQC 28](#_Toc92192676)

[3.8.4 Trimming, orienting, and defusing cDNA reads with Pychopper 28](#_Toc92192677)

[3.8.5 Adapter trimming with Porechop 29](#_Toc92192678)

[3.8.6 De novo clustering of cDNA isoforms with IsONclust 29](#_Toc92192679)

[3.8.7 Error correction of clustered cDNA reads with IsONcorrect 30](#_Toc92192680)

[3.8.8 Ribodepletion with SortMeRNA 30](#_Toc92192681)

[3.8.9 Taxonomic Validation using Filtered rRNA reads 31](#_Toc92192682)

[3.8.10 Alignment with Minimap2 32](#_Toc92192683)

[3.8.11 Alignment statistics with samtools 33](#_Toc92192684)

[3.8.12 Obtaining raw read counts 34](#_Toc92192685)

[3.8.13 Annotation of unmapped reads with DIAMOND 34](#_Toc92192686)

[3.8.13 Aggregation of annotated reads 35](#_Toc92192687)

[3.8.14 Statistical analysis and visualization 36](#_Toc92192688)

[3.8.15 Annotation of CAZymes with the Hotpep module of dbCAN 37](#_Toc92192689)

[3.8.16 Identification of species of interest with Krona enzyme-specific multi-layered pie-charts 37](#_Toc92192690)

[3.8.17 Screening for PULs from lignocellulolytic CAZyme families 38](#_Toc92192691)

[4.0 CHAPTER 4: RESULTS 39](#_Toc92192692)

[4.1 Growth parameters collection 39](#_Toc92192693)

[4.2 Feed composition analysis 42](#_Toc92192694)

[4.3 Gel electrophoresis 44](#_Toc92192695)

[4.4 Quality statistics of basecalled reads with pycoQC 44](#_Toc92192696)

[4.5 Trimming, orienting, and defusing cDNA reads with Pychopper 47](#_Toc92192697)

[4.6 rRNA filtering with SortMeRNA 49](#_Toc92192698)

[4.7 Error correction of IsONclust clustered cDNA reads with IsONcorrect 50](#_Toc92192699)

[4.8 Raw read counts 52](#_Toc92192700)

[4.9 Annotation and aggregation of with DIAMOND 53](#_Toc92192701)

[4.10 Statistical analysis and visualization 58](#_Toc92192702)

[4.10.1 Relative activity of microorganisms in the metatranscriptomes 58](#_Toc92192703)

[4.10.2 Taxonomic validation of metatranscriptome sequences using 16S rRNA reads 60](#_Toc92192704)

[4.10.3 Metatranscriptome diversity statistics 62](#_Toc92192705)

[4.10.4 DESeq2 organism heatmap 63](#_Toc92192706)

[4.10.5 Organism PCA plot 64](#_Toc92192707)

[4.10.7 Relative functional activity in the metatranscriptomes 67](#_Toc92192708)

[4.10.8 DESeq2 function heatmap 69](#_Toc92192709)

[4.10.9 CAZymes annotation with dbCAN2 Hotpep module 70](#_Toc92192710)

[4.10.10 CAZy family GH43 subfamily 16 (GH43\_16) from the BSG sample 72](#_Toc92192711)

[4.10.11 CAZy family GH51 subfamily 2 (GH51\_2) from the WH sample 73](#_Toc92192712)

[4.10.12 Screening for polysaccharide utilization loci (PULs) 75](#_Toc92192713)

[5.0 CHAPTER 5: DISCUSSION 80](#_Toc92192714)

[6.0 CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS 86](#_Toc92192715)

[6.1 Conclusion 86](#_Toc92192716)

[6.2 Recommendations 87](#_Toc92192717)

[REFERENCES 88](#_Toc92192718)

[APPENDICES 104](#_Toc92192719)

[Project Budget 104](#_Toc92192720)

[Time plan 106](#_Toc92192721)

[Parameter Collection Template 107](#_Toc92192722)

[SBS-REC Approval Letter 109](#_Toc92192723)

[NACOSTI Research Permit 111](#_Toc92192724)

# LIST OF ILLUSTRATIONS

[Figure 1 : An illustration of the conceptual framework 6](file:///A:\Thesis\Eric_Gathirwa_Kariuki_MSC.Dissertation_JP_CK.docx#_Toc92192527)

[Figure 2:The structure of lignocellulose (Baruah et al., 2018) 8](#_Toc92192528)

[Figure 3: The life cycle of the Black soldier fly (De Smet et al., 2018) 10](#_Toc92192529)

[Figure 4: Substrate reduction index 39](#_Toc92192530)

[Figure 5: Mean weight of larvae 40](#_Toc92192531)

[Figure 6: Substrate temperature levels 40](#_Toc92192532)

[Figure 7: Substrate temperature levels 41](#_Toc92192533)

[Figure 8:Mean larval lengths 41](#_Toc92192534)

[Figure 9: Pupation rates observed per dietary substrate 42](#_Toc92192535)

[Figure 10: Comparison of macronutrients and micronutrients in the dietary substrates 43](#_Toc92192536)

[Figure 11: Agarose gel electrophoresis 44](#_Toc92192537)

[Figure 12: Sequencing quality statistics with pycoQC 46](#_Toc92192538)

[Figure 13: Demultiplexing with Guppy 46](#_Toc92192539)

[Figure 14: Classification and orientation of cDNA reads with Pychopper 47](#_Toc92192540)

[Figure 15: Ribodepletion statistics using MultiQC (v1.11) 50](#_Toc92192541)

[Figure 16: Differences in mapping percentages and mapping quality of reads mapped to the BSF reference between the uncorrected and corrected reads 52](#_Toc92192542)

[Figure 17: Stacked bar plots showing microbial activity in each metatranscriptome 59](#_Toc92192543)

[Figure 18: Taxonomic profiling using 16S rRNA sequences filtered at the ribodepletion step 60](#_Toc92192544)

[Figure 19: Order Bacteroidales taxonomic profiling using 16S rRNA sequences 61](#_Toc92192545)

[Figure 20:Phylogenetic tree of 16S rRNA samples from order Bacteroidales 62](#_Toc92192546)

[Figure 21: Alpha diversity evaluated using the Shannon-Wiener index (H) 63](#_Toc92192547)

[Figure 22: DESeq2 distance heatmap for metatranscriptome function profiles 64](#_Toc92192548)

[Figure 23: Organism PCA plot 65](#_Toc92192549)

[Figure 24: Alpha diversity analysis using 16S rRNA sequences 66](#_Toc92192550)

[Figure 25: Beta diversity analysis using 16S rRNA sequences 67](#_Toc92192551)

[Figure 26: Dodged bar plots of active SEED subsystems level 1 functions against the average levels in the pooled experimental metatranscriptomes compared to the pooled control metatranscriptome (CF) 69](#_Toc92192552)

[Figure 27: DESeq2 distance heatmap for metatranscriptome function profiles 69](#_Toc92192553)

[Figure 28: GH43\_16 Krona multi-layered taxonomic pie-chart 73](#_Toc92192554)

[Figure 29: GH51 Krona multi-layered taxonomic pie-chart 74](#_Toc92192555)

# LIST OF TABLES

[Table 4. 1: ICP-MS composition analysis 42](#_Toc89593846)

[Table 4. 2: Sequencing summary statistics (sequencing run 1) 44](#_Toc89593847)

[Table 4. 3: Sequencing summary statistics (sequencing run 2) 44](#_Toc89593848)

[Table 4.4: Classification and orientation of sequence reads 47](#_Toc89593849)

[Table 4.5: Percentage of rRNA reads per sample 49](#_Toc89593850)

[Table 4. 6: Evaluation of the read error-correction efficacy 50](#_Toc89593851)

[Table 4. 7:Total number of unmapped reads per metatranscriptome 53](#_Toc89593852)

[Table 4. 8: Most abundant organisms per dietary substrate 55](#_Toc89593853)

[Table 4. 9: CAZyme identification and annotation using the dbCAN2 Hotpep module 71](#_Toc89593854)

[Table 4. 10: Screening for polysaccharide utilization loci (PULs) 77](#_Toc89593855)

# ABBREVIATIONS

**BSF** Black Soldier Fly

**RNA** Ribonucleic Acid

**RNA-Seq** RNA Sequencing

**rRNA** Ribosomal RNA

**tRNA** Transfer RNA

**ITS** Internal Transcribed Spacer

**DNA** Deoxyribonucleic Acid

**PCR** Polymerase Chain Reaction

**cDNA** Complementary DNA

**OTU** Operational Taxonomic Unit

**AFEX** AmmoniaFiber Explosion

**ONT** Oxford Nanopore Technologies

**PacBio** Pacific Biosciences

**DEG** Differentially expressed genes

**DEA** Differential Expression Analysis

**WRI** Waste Reduction Index

**SGS** Second Generation Sequencing

**TGS** Third Generation Sequencing

**OLC** Open Layout Consensus

**ASIC** Application-specific Integrated circuit

**SSD** Solid-state drive

**KEGG** Kyoto Encyclopedia of Genes and Genomes

**KGML** KEGG Markup Language

**DEPC** Diethylpyrocarbonate

**FDR** False Discovery Rate

**BLAST** Basic Local Alignment Search Tool

**BLAT** BLAST-like alignment tool

**HPC** High-Performance Computing

**RCT** Randomized Control Trial

**BSG** Brewers’ Spent Grain

**CF** Processed Chicken Feed (Grower’s mash)

**CM** Chicken Manure

**WH** water hyacinth

**FM** Feed Mix

***Icipe*** International Center for Insect Physiology and Ecology

**INSEFF** Insects for Food and Feed program

**SBS-REC** School of Biomedical Sciences Research and Ethics Committee

**NACOSTI** National Commission for Science Technology and Innovation

**ICP-MS** Inductive Coupled Plasma – Mass Spectrometry

**SRI** Substrate reduction index

**VNP** VN primers

**SSP** Strand-switching Primers

**SQB** sequencing buffer

**RAP** Rapid adapters

**LB** loading beads

**TBE** Tris/Borate/EDTA buffer

**NEB** New England Biolabs

**GPU** Graphical Processing Unit

**CPU** Central Processing Unit

**HMM** Hidden Markov Models

**SAM** Sequence Alignment Map file

**BAM** Binary SAM file

**Hotpep** Homology to Peptide

**CAZyme** Carbohydrate-active Enzyme

**CAZy** Carbohydrate-active Enzyme family (CAZy family)

**PUL** Polysaccharide Utilization Loci

**GT** Glycosyltransferases

**GH** Glycoside Hydrolase

**CE** Carbohydrate Esterases

**PL** Polysaccharide Lyases

**abf** α-L-arabinofuranosidase

**abfB** α-L-arabinofuranosidase B

**abf2**  α-L-arabinofuranosidase 2

# ABSTRACT

Second-generation biofuel production has emerged as a sustainable and alternative energy option to the fast depleting, ecologically unfriendly petroleum-based fuels. Second-generation biofuels are produced from lignocellulosic industrial and agricultural wastes, energy crops, and crop residues, and therefore, are not a threat to food security. Despite second-generation biofuels being a promising technology, obstacles in their production prevent the bioconversion process from attaining optimal performance under minimal capital investment. The Black soldier fly (BSF) (*Hermetia illucens*) larvae have stood out as a key source of alternative protein in the animal feeds industry, the production of chitin and essential oils, and as a useful tool in the valorization of organic biomass and other biodegradable wastes, mainly attributed to their potent larval gut microbiome. The BSF larval gut microbiome is an active area of study due to the rapid growth and broad degradation capabilities of the larvae, and their non-competence as a vector for any known human diseases. This presents the BSF larvae as a potential source of lignocellulolytic microorganisms and enzymes capable of the breakdown of recalcitrant organic biomass and potential applications in the second-generation biofuel industry. Members of the genera *Dysgonomonas, Bacteriodes*, and *Actinomycetes* were identified in high abundance from preliminary BSF microbiota studies. These genera have shown lignocellulolytic abilities and thus have the potential to be implemented in the production of different value-added products such as biofuels and animal feeds.

To investigate the effects of dietary intervention and the presence of microorganisms and enzymes associated with lignocellulolytic activities in the BSF larval microbiome, a randomized control trial design was employed in this study where BSF larvae were bred under different diets, selected based on their increasing lignin contents i.e., processed chicken feed (CF), chicken manure (CM), brewers’ spent grain (BSG), water hyacinth (WH) and feed mix (FM) that contained all diets mixed in the ratio 1:4. mRNA libraries from 16 samples (n=16) were prepared and RNA-Sequencing was conducted using the PCR-cDNA approach with the MinION sequencing platform. The gut microbiome and functional profiles were compared between the samples while mainly screening for the lignocellulolytic microbial genera reported from preliminary studies.

Metatranscriptome samples from diets BSG and WH possessed in high abundance two of the three main genera identified in preliminary 16S BSF microbiota studies and hypothesized to be involved in lignocellulolytic functions - *Bacteroides* and *Dysgonomonas*. Additionally, lignocellulolytic CAZyme families GH43\_16 and GH51 containing enzyme α-L-arabinofuranosidase (abf, EC3.2.1.55), known for degrading arabinoxylan and arabinogalactan hemicellulose fractions, were identified in the BSG and WH metatranscriptome samples, respectively. Polysaccharide Utilization Loci (PUL) screening further revealed PUL0013 and PUL0395 gene clusters that encode for the hemicellulolytic enzyme abfB in CAZy family GH51. We conclude that the BSF larvae gut microbiome profiles were significantly altered with dietary intervention and that the BSF gut microbiome could not only be used as a source of lignocellulolytic microbes and CAZymes, but could also guide the identification of degradative pathways involved that could be applied in, but not limited to, enzyme hydrolysis in second-generation biofuel production.

# 1.0 CHAPTER ONE: INTRODUCTION

## 1.1 B**ackground Information**

There has been a recent need for the development, commercialization, and use of eco-friendly renewable fuels as cleaner alternatives to the rapidly depleting petroleum-based (fossil) fuels. The dynamic has further shifted from just renewable biofuels to cellulosic and lignocellulosic (second-generation) biofuels (Limayem & Ricke, 2012). As opposed to the first-generation biofuels produced from food-grade products such as corn and sugarcane, second-generation biofuels are produced from energy crops, industrial and agricultural wastes, and crop residues, with no impact on food prices, closing the debate on the competition with food products for biomass (Robak & Balcerek, 2018).

However, there are inherent challenges associated with handling cellulosic and lignocellulosic biomass. Additional pretreatment and hydrolysis steps increase production costs and have led to the closure and repurposing of many second-generation biofuel production plants (Nuelle, 2019); the actual concern is how to make them economically and commercially feasible. Recent studies have shown that the costs associated with the production of second-generation biofuels are 2-3 times more than petroleum-based fuels on an energy equivalent basis with about 30% of the total costs associated with biomass acquisition and handling, 40-49% associated with enzymes costs, and about 15% incurred in fermentation costs (Balan, 2014).

The Black Soldier Fly (*Hermetia illucens*; BSF) is a useful tool in the valorization of organic biomass and other biodegradable wastes, and its larval gut microbiome, which is thought to be involved in the bioconversion, has emerged as an active area of study (Samayoa et al., 2016b). This can be attributed to the broad substrate degradation capabilities and fast growth rates of the larvae (Klammsteiner et al., 2020; Nguyen et al., 2015; Sheppard et al., 2002). Furthermore, some studies have shown that dietary interventions alter the gut microbiome in BSF larvae presenting the possibility of inducing the establishment of desired microbial communities through changes in dietary regimens (Bruno et al., 2019; Klammsteiner et al., 2020; Rothe & Blaut, 2013).

Members of the genera *Dysgonomonas, Bacteriodes*, and *Actinomycetes* were identified in high abundance in BSF microbiota studies (Tanga et al., 2021; Jeon et al., 2011; Jiang et al., 2019). This study focused on these microorganisms because of their lignocellulolytic capabilities, and thus, the potential to identify the associated metabolic pathways and complex PULs to be applied in the production of different value-added products such as second-generation biofuels.

By sequencing the metatranscriptome, *i.e.* mRNA of the BSF larvae microbiome, we used the data to study the microbial and functional shifts under different dietary regimens and their possible contribution to lignocellulolytic activity (F. Li et al., 2019). By implementing the metatranscriptomics pipeline, the active microbial species were identified by annotation against the Refseq bacterial database and the SEED subsystems hierarchical database, and the respective functions performed by these microbes are inferred (Westreich et al., 2018).

Further annotation of metatranscriptomic sequences against Carbohydrate-Active Enzymes (CAZymes) aided in the identification of enzyme families involved in the breakdown of complex carbohydrates. Additionally, the organisms involved in the production of these enzymes were identified by taxonomically assigning the signature proteins to species where the respective CAZyme modules have been previously identified (Lombard et al., 2014). It is also possible to identify Polysaccharide Utilization Loci (PULs), which are gene clusters from groups of organisms that work together to degrade these complex carbohydrates. These resources provide a clearer understanding of what microorganisms present in the study metatranscriptomes are directly involved and capable of producing the identified lignocellulolytic enzymes (Terrapon et al., 2015). Analyzing the pathways the identified enzymes are involved in, provides a mechanistic understanding of specific genes that are more expressed than expected by chance (Reimand et al., 2017). The viable bacterial species identified from taxonomic analysis together with the enzymes identified from pathway analysis steps could be prospectively tested on bioreactors in a future study to ascertain their efficacy in increasing biofuel conversion rates through lignocellulosic biomass degradation and hydrolysis of complex sugars into fermentable sugars respectively.

For this study, we carefully selected four diets for the BSF larvae which included: chicken feed, chicken manure, brewers' spent grain, and water hyacinth for their varying lignin contents, to identify which of these diets induced microbial lignocellulolytic activity in the larvae.

## 1.2 Problem Statement

The costs incurred in pretreatment and hydrolysis of raw materials during the production of second-generation (cellulosic) biofuels are not economically viable and are not incurred in the first-generation (starch-based) biofuel production. There is hence an urgent need to find inexpensive and time-efficient ways to pretreat and hydrolyze lignocellulosic biomass to produce second-generation biofuels (Antunes et al., 2019). Pretreatment is required to break down the lignin structure and to disrupt cellulose’s crystalline structure so that the acids or enzymes can easily access and hydrolyze the cellulose. Hydrolysis requires enzymes to break down the cellulose and hemicellulose into fermentable sugars that can finally be fermented by the appropriate microbial enzymes. These processes take place in varying physicochemical conditions such as adverse temperatures, pH and ionic concentration changes, high alcohol concentration, among others, that microbes must withstand (Amin et al., 2017).

In this study, a metatranscriptomics analysis pipeline was formulated to: (i) identify and functionally characterize organisms that play a direct role in the effective degradation of lignocellulosic biomass present in the BSF larvae gut microbiota; (ii) improve the accuracy and throughput of Oxford Nanopore long-cDNA reads through error correction and optimization methods; (iii) identify lignocellulolytic CAZymes present in the BSF larvae metatranscriptomes and their associated PULs; (iv) further understand the microbial breakdown of recalcitrant biopolymers and propose approaches to overcome the inherent challenges in lignocellulosic biomass hydrolysis.

The findings of this study provided novel insights into the BSF larvae microbiome as a source of microorganisms and enzymes that could be applied in, but not limited to, enzyme hydrolysis in second-generation biofuel production.

## 1.3 Hypothesis

Implementing a metatranscriptomic analysis pipeline can aid in the identification and characterization of unculturable microbes responsible for lignocellulosic biomass degradation in the gut of the BSF larvae.

## 1.4 General Objective

## To identify and functionally characterize lignocellulosic biomass-degrading microbes and enzymes from the BSF larval gut microbiome.

## 1.5 Specific Objectives

1. To determine whether dietary intervention significantly alters the BSF larvae gut microbiome composition profiles.
2. To improve the accuracy and throughput of Oxford Nanopore long-cDNA reads through error correction and statistical optimization.
3. Design and implement a metatranscriptomic analysis pipeline on ONT RNA-sequencing data, obtained from the BSF larvae gut, to acquire species level functions associated with the genera of interest from the preliminary 16S rRNA data.

## 1.6 Significance

The use of second-generation (cellulosic) biofuels has for long enough been deemed as a solution to the ecologically unfriendly fossil fuels and food insecurity caused by the production of first-generation biofuels. The success of this study will enable us to identify and characterize microbes that can effectively degrade lignocellulosic biomass and enzymes that can hydrolyze cellulose and hemicellulose biomass fractions into simple sugars that can easily be fermented. These can be applied in large-scale industrial processes such as biofuel generation, pharmaceuticals, food industries, and the manufacture of other useful bioproducts. Ultimately, the implementation of the findings and recommendations from this study will potentially ease the production of eco-friendly second-generation biofuels cheaper, easier, and more time-efficient.

## 1.7 Justification

This study utilized RNA-Seq data generated using the MinION MK1B platform from Oxford Nanopore Technologies (ONT), to study the BSF larvae gut metatranscriptomes under different dietary regimens. The resultant sequencing reads possess a relatively high accuracy that can be further improved using statistical correction algorithms (Sahlin et al., 2021). This sequencing platform is considered cost-effective and scalable for various sequencing needs (Zhang et al., 2020). The sequencing libraries were prepared using the SQK-PCB109 PCR-cDNA barcoding kit from ONT. This sequencing approach allows for multiplex sequencing by assigning each sample a unique set of barcode sequences with the possibility of sequencing 12 samples in one sequencing run. RNA-Sequencing approaches possess a higher resolution than 16S rRNA amplicon sequencing; and thus, facilitate better analysis of bacterial communities present in the different metatranscriptomes studied and their respective functional profiles. Using this approach, we were able to observe variations in the organism and functional profiles that arose with dietary intervention. The genera *Dysgonomonas* and *Bacteroides* were identified in high abundance in the highly lignocellulosic diets, BSG and WH.  *Dysgonomonas, Bacteriodes*, and *Actinomycetes* were reported in high abundance in previous studies and have been portrayed to possess lignocellulolytic capabilities (Tanga et al., 2021; Jeon et al., 2011; Jiang et al., 2019). Also identified from these highly lignocellulosic diets, were CAZyme classes involved in xylan degradation. These enzymes, found in this study, could be potentially exploited in commercial biofuel production – to hydrolyze hemicellulosic fractions of lignocellulosic feedstock into fermentable sugars to increase the biomass to biofuel conversion rates.

## 1.8 Conceptual Framework

Figure 1 : An illustration of the conceptual framework

**Outcomes**

* Altered gut microbiomes under different dietary conditions
* Different microbial and functional profiles
* Identification of microorganisms and enzymes actively involved in lignocellulolytic functions.
* enriched metabolic pathways

**Applications**

* Identified and functionally characterized lignocellulolytic organisms and enzymes can be applied in, but not limited to, cellulosic biofuel production.

**Component**

* Gut Microbiome
* Microbial species
* Dietary intervention
* Gene expression

**Attribute**

* Diversity Analyses
* Composition
* Altered microbial and functional profiles
* Enrichment of metabolic pathways

**Mechanisms**

* BSF Breeding under different dietary conditions
* RNA Extraction
* ONT PCR-cDNA synthesis
* ONT MinION sequencing
* Quality analyses
* Error correction of full-length cDNA reads
* Implement Metatranscriptomic analysis pipeline:
* rRNA removal
* Annotation and analyses against reference databases
* Generate abundance counts
* Differential expression Analysis (DEA)
* Pathway enrichment Analysis
* Interpret results, publish and disseminate research findings

## 2.0 CHAPTER 2: LITERATURE REVIEW

### 2.1 Introduction

Second-generation biofuels are a viable alternative to petroleum-based fossil fuels. They are produced from lignocellulosic biomass that normally comes from non-food products. This has anticipated benefits such as easing the evident burden on food security caused by the production of first-generation biofuels that are sourced from food crops (Lynd, 2017). There has been an exponential increase in world biofuel production which increased five-fold from about 20 billion liters/year in 2001 to 100 billion liters/year in 2011, and the numbers keep rising with the incessant appeals to stakeholders in the energy industry to embrace more climate-friendly and sustainable fuel options. The sharp rise in demand to produce biofuels has been identified by researchers and industry players as a key factor leading to the increase in food prices.The sharpest increase in the production of biofuel happened in 2007/2008, in tandem with a sharp rise in the prices of food commodities. A substitution effect also arises, at a consumption level as well as at a production level causing a ripple price increase to other crops (MS Swaminathan, Maryam Rahmanian, John Wilkinson, 2013). On the other hand, plant biomass accounts for about 10% of the global primary energy and is projected to rise to about 25% in 2050 (Dale et al., 2014). Amongst the several types of plant-derived biomass, cellulosic feedstocks are believed to possess the highest potential for mitigating climate change and are available at a lesser cost per unit of energy (e.g., per megajoule) than fossil fuels (*AR5 Climate Change 2014: Mitigation of Climate Change — IPCC*, n.d.).

Governments saw in cellulosic biofuels the potential to contribute massively to economic development and employment especially in less developed areas while cementing their energy security (Lynd, 2017). This raised the hopes and expectations of stakeholders in the energy industry. However, by 2016, the production capacity for liquid biofuels derived from cellulosic biomass was at 0.7 billion liters for cellulosic ethanol and 4.4 billion liters for thermochemically derived renewable diesel and jet fuel, against projections of 17 billion liters globally Overview for Renewable Fuel Standard, EPA, 2017). Many of these biofuel companies have ceased operations while those that are still existent are trading well below their initial public offering price and primarily turning their focus to higher-value products other than fuels (Lynd, 2017). It is universally recognized that the main challenge to the cost-effective production of second-generation biofuels is the inherent challenge that exists in pretreatment and converting lignocellulosic biomass into more fermentable intermediates. This is translated into high pretreatment and enzyme hydrolysis process costs for biological processing and the cost of gasification or pyrolysis in the thermochemical process (Canabarro et al., 2013). These render the entire process economically unfeasible. The removal of lignin, a highly recalcitrant polymer, and improving the low yields of sugars from other plant polysaccharides e.g., hemicellulose and cyclodextrins, must be effectively conducted and the technology improved to make cellulosic biofuels sustainable and economical to produce (Himmel et al., 2007).

### 2.2 Aims of Feedstock Pretreatment

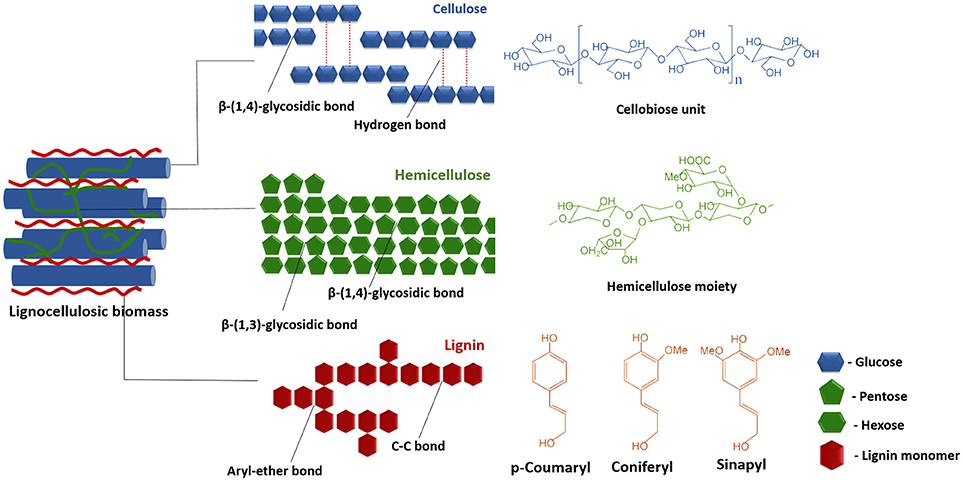


Figure 2:The structure of lignocellulose (Baruah et al., 2018)

The main goals of successful pretreatment of biomass are the removal or disintegration of the lignin polymer, solubilizing hemicellulose into simple for easier fermentation, decreasing the cellulose’s crystallinity, and increasing the permeability of lignocellulosic feedstock. Therefore, for a pretreatment strategy to be effective, it must adhere to the following requirements: be cost-effective, improve the overall bioconversion of complex sugars into simple sugars by hydrolysis, avoid the loss or degradation of sugars in the process, and evade the build-up or formation of inhibitory byproducts that might hinder subsequent pretreatment and fermentation steps (P. Kumar et al., 2009).

### 2.3 Current Biomass Pretreatment and Conversion Technologies

The biorefinery comprises four major sections: harvesting and handling of feedstock, pretreatment (thermal or chemical), enzymatic hydrolysis, and fermentation. The resistance of lignocellulosic biomass to biological or physicochemical breakdown is known as recalcitrance. This is attributed to several factors such as the crystalline structure of cellulose and the extent of lignification characterized by the structural complexity and heterogeneity of the cell-wall components. These factors must be overcome for the successful utilization of these feedstocks (Guerriero et al., 2016). Most existing biomass conversion structures rely on both chemical and enzymatic treatments (Srivastava et al., 2020). Pretreatment is normally done using acids, alkalis, heat treatments, steam explosion, or a blend of the methods to decrease the recalcitrance of the lignocellulosic biomass by depolymerizing lignin and solubilizing cellulose and hemicellulose. The cellulose and hemicellulose end up being converted to oligosaccharides or monosaccharides that can be easily enzymatically hydrolyzed and fermented. Thermochemical biomass pretreatment methods include pretreatment with dilute sulphuric acid at 1400 – 2000C, which has for a long time been regarded as the standard reference technology that yields biomass with enzymatic digestibilities that are generally acceptable (Himmel et al., 2007). Physicochemical pretreatment methods include steam explosion at 1600C-2600C and pressures of <50atm, and Ammonia Fiber Explosion (AFEX) using liquid ammonia at high temperature and pressure for some time before suddenly dropping the pressure levels (P. Kumar et al., 2009). Milling is a physical pretreatment method that involves breaking down the feedstock into small fractions to increase the surface area to volume ratio of the feedstocks while reducing the crystallinity of cellulose and its polymerization. Microwave radiation pretreatment uses the magnetic field components of microwaves to speed up the biological, and physicochemical processes due to the heat and the extensive collisions brought about by the vibration of ion movements and polar molecules (Nomanbhay et al., 2013). Alkali pretreatment e.g., on the other hand, involves the addition of bases to the feedstock, leading to the swelling of the internal surface and the disentanglement of lignin. It is, however, a suitable method for biomass with low lignin content (Badiei et al., 2014).

The main disadvantage of using physical pretreatment methods e.g. steam explosion is the high energy consumption which significantly increases the overall costs. Chemical pretreatment methods, on the other hand, face challenges such as toxic byproducts produced by the chemicals, and the high costs of organic solvents. Biological methods provide a cheaper, less toxic alternative but face major disadvantages such as low efficiency when used in isolation and longer turnaround times. They involve using potent microorganisms such as fungi and bacteria which secrete lignocellulolytic enzymes (Baruah et al., 2018). In light of these reasons, studying the lignocellulolytic efficiency of entire microbial communities in their natural environments e.g. gut microbiota of potent waste-degrading BSF larvae using techniques such as metatranscriptomics can provide better outcomes for biological methods by identifying Polysaccharide Utilization Loci (PULs) where various microorganisms work together to break down complex polysaccharides.

### 2.4 The Black Soldier Fly

Natively from South America, the Black Soldier Fly has found its way into other tropical and subtropical regions of the world and can be found in all continents excluding the Arctic and Antarctic regions. The adult flies are large and can grow up to 2cm long. The larvae normally develop through 6 larval instars and can attain lengths of 1.8-2cm (Čičková et al., 2015). Adult flies are not very mobile and spend most of the time resting on vegetation. The flies do not eat as adults and therefore, their larvae are known to consume diets rich in fats and proteins to be able to complete their development and survive their adulthood long enough to mate and lay eggs (Nguyen et al., 2015).

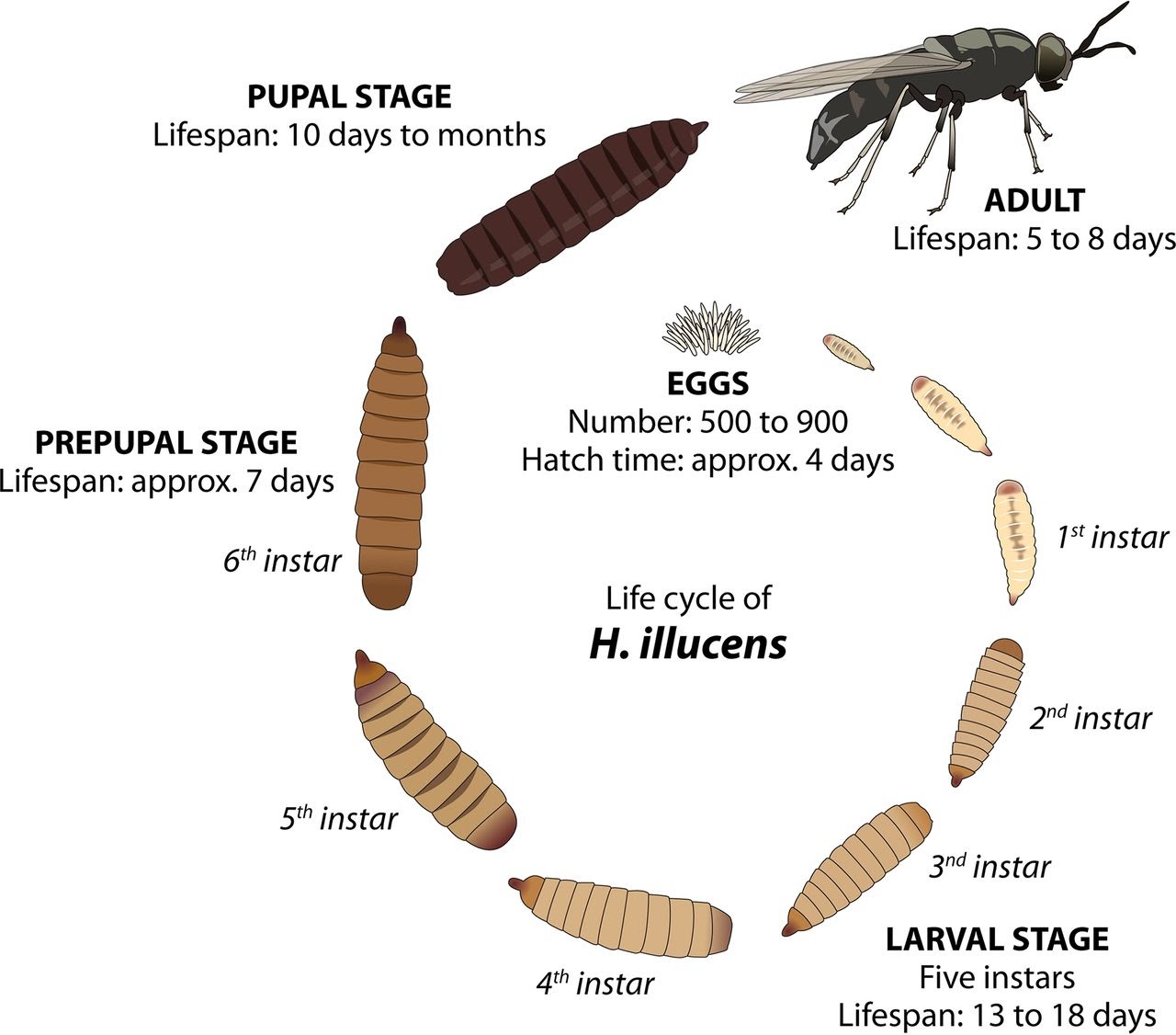


Figure 3: The life cycle of the Black soldier fly (De Smet et al., 2018)

The use of larvae for processing organic wastes was recommended about a century ago. The Black Soldier Fly and the house fly have been extensively studied for this reason. Temperature and pH are some of the common physiological parameters observed and analyzed during BSF larvae breeding. Degradation under suitable dietary conditions changes the pH from neutral to alkaline, the moisture decreases, and the temperature rises. A study by (Čičková et al., 2015) found that some microbial pathogen loads present in the organic substrate tend to be substantially reduced upon degradation, but further treatment may be required to eliminate all pathogens. The rapid growth rate, its broad degradation capabilities make the larvae an ideal candidate for large-scale solid waste degradation as well as other industrial applications such as converting low-value wastes to high-value protein and oil for the manufacture of alternative animal feeds (Kim et al., 2020). The adult fly has not been identified as a vector or pest for any known diseases (Klammsteiner et al., 2020; Samayoa et al., 2016a). However, it is still unclear whether opportunistic pathogens do proliferate in the presence of BSF larvae posing further health and environmental risks (Khamis et al., 2020).

### 2.5 Diet Selection and Breeding Conditions for the BSF Larvae

The choice of feedstock has been found to significantly impact the duration of the BSF larval phase (Klammsteiner et al., 2020). Processed feedstocks such as processed chicken feed, or manure are degraded faster compared to lignocellulosic biomass containing lignin, hemicellulose, and cellulose fractions. If different diets are used in experimental breeding, the moisture content, as well as other physiological conditions e.g. the photoperiod and the texture of the diets, should be standardized in each.

To investigate the impact of different diets on the BSF larvae on microbial diversity, we selected the diets selected based on varying lignocellulose content. A control diet should be selected based on its ability to provide a steady and balanced supply of nutrients (Klammsteiner et al., 2020), and in our study, it had to be well-digestible by the BSF larvae. Processed animal feeds are normally preferred as control diets. Processed chicken feed (Grower’s mash) was selected as the control for this study. However, this diet is expensive and is not among the low-value waste materials which are degraded and transformed by BSF larvae into high-value commodities such as fertilizer. On the other hand, the other dietary substrates selected for this study possessed varying lignocellulose content.

Chicken manure (CM) has been extensively used for BSF larvae breeding as it is readily available and possesses high nutritive value. This diet has reported a high turnover of nutrients and minerals compared with other common diets such as kitchen waste and brewers’ spent grain (Shumo et al., 2019).

The highly lignocellulosic feedstocks - Brewer’s spent grain (BSG) and Water hyacinth (WH), contained high proportions of lignin, hemicellulose, and cellulose. Due to lower amounts of available carbon, the breakdown of such biomass is more complex and is performed by specialized groups of bacterial and fungal communities. Therefore, feedstocks with higher lignocellulosic fractions are expected to be broken down less efficiently by BSF larvae as compared to non-lignocellulosic feedstocks (Lalander et al., 2019). However, the breakdown of these lignocellulosic dietary substrates by BSF larvae provided insights into their efficiency in breaking down complex polysaccharides such as cellulose and hemicellulose.

BSG has been widely used in BSF rearing studies and has been recommended for its sustainability footprint as it is a by-product of the alcohol brewing process. It has been regarded as a good substrate, especially for BSF larvae reared for use in animal feeds. As compared to many other feedstocks, BSF larvae fed on this feedstock possessed higher fat and nutrient content (Chia et al., 2020). This is normally characteristic of larvae fed on feedstocks they can effectively break down.

Water hyacinth (Eichhornia crassipes) is a rapid-growing prolific water weed that has choked freshwater bodies all over the world and has emerged as a potential feedstock for second-generation biofuel production due to its low lignin levels and high availability. The plant can reach growth rates as high as 60-100 tons/ha/yr and does not compete with food crops for arable land. Additionally, the waterweed can reach these growth rates without human interference (Mishima et al., 2008). Successful degradation of this plant by the BSF larvae would prove potentially useful in the identification and characterization of microbial communities present in their guts. Further analysis of metabolic pathways would be useful in potentially identifying pathways and novel enzymes that can be utilized in the hydrolysis of lignocellulosic biomass into simple sugars for more efficient fermentation.

Choosing an appropriate sampling strategy for the BSF larval samples is crucial for a study (Setia, 2016). Samples can either be individually sampled or pooled. The main factors that guide the selection of a sampling strategy include the expected analytical variance and statistical power, the costs for sample collection, and sample preparation and analysis. The number of pools selected should be able to achieve high-cost efficiency and good statistical power (Bignert et al., 2014). Pooling also improves the statistical significance of the samples and minimizes the amount of information that is lost below a detection threshold (Hoang et al., 2017). For this study, 3 samples were sequenced for each dietary substrate, except for the control substrate where 4 samples were sequenced. For CAZyme family analysis and screening for PULs, all individual samples belonging to a dietary substrate were pooled.

## 2.6 The Black Soldier Fly larval microbiome

The microbiome was described by Nobel laureate Joshua Lederberg as the combination of commensal, symbiotic, and pathogenic microorganisms that colonize the human body (Lederberg & McCray, 2001). Bacteria are a key component of organic waste decomposition and insects such as BSF employ these resources. Some insects largely depend on bacterial mutualism to survive and reproduce effectively (Werren et al., 1995). Bacteria that have been isolated from BSF larvae have in many instances been used as probiotics in enhancing waste reduction and are gradually finding their way in other applications such as the production of protein and oil for alternative animal feeds, chitin production, among other uses (Antonov et al., 2019). However, it has been reported that the microbial diversity based on diet and geographic distribution of BSF also determines the core components of the microbiome (Khamis et al., 2020; Tanga et al., 2021; Wynants et al., 2019). Although the diet fed to the BSF larvae is assumed to be a key source of gut bacteria, other interrelating biotic and abiotic factors also contribute to the microbiota composition (Barragan-Fonseca et al., 2017). The presence of some species of bacteria such as *Providencia* can be attributed to vertical transmission as they were reported across all stages of the BSF life cycle (De Smet et al., 2018).

Despite the great industrial potential possessed by the BSF larvae, the biochemical and molecular potential of microbial hydrolytic enzymes present in the larval gut are not properly understood as many gut microbes are unculturable. Advanced metagenomic approaches such as metatranscriptomic analyses have come in handy by providing potent tools for novel gene surveillance without necessarily cultivating gut microbes (C. M. Lee et al., 2018).

The BSF larval gut is divided into three sections: the anterior, the middle, and the posterior section. A study by (Bruno et al., 2019) showed that changing the diet results in dysbiosis limited to the anterior gut region which also portrays the highest diversity and that the feeding activity of the larvae does significantly alter the microbiota of the substrate. The diversity gradually decreases through the midgut to the posterior region as the bacterial load increases. However, when the substrate contains more protein, as compared to other macronutrients, dysbiosis was found to favor the midgut region which has been attributed to different structural and functional domains present in this region (Bruno et al., 2019). Therefore, based on different studies, by placing the larvae on different diets, one can expect to observe altered taxonomic profiles.

Findings from different 16S rRNA gene studies have shown some significant differences in the gut microbial composition of BSF larvae. This has been mainly attributed to the different dietary substrates used in each study, and environmental factors credited to the various study sites. Other factors such as the sections of the 16S rRNA gene studied and the choice of analysis tools contribute to changes in abundance and diversity, especially for lower taxonomic classes (Jeong et al., 2021), but to a lower extent (Chung et al., 2020). The different dietary substrates ingested by the larvae are capable of shaping the gut microbial load and diversity profiles (Bruno et al., 2019). As potent as the BSF larvae are in degrading organic compounds, their waste reduction capabilities show a notable reduction when the substrates in question contain recalcitrant compounds such as cellulose and lignin, and in turn, portray altered gut microbial diversity profiles as compared to those observed in processed substrates. This has been revealed by diversity analysis using various alpha diversity indices with very small changes being detected at a beta-diversity level (Klammsteiner et al., 2020; Tanga et al., 2021; Khamis et al., 2020).

From previous studies, several bacterial conserved Operational Taxonomic Units (OTUs) have been found in larval guts exposed to various diets, laying assertions of an extant core microbiome. Nevertheless, microbial acquisition of the BSF larval gut microbiota from the surroundings, diets, and other factors contributes significantly to the microbial composition (De Smet et al., 2018). The core microbial community has been reported in about 80% of the gut samples. From multiple studies, *Actinomyces, Dysgonomonas, Enterococcus, Bacteroides* genera have been reported to constitute this core microbiome (Khamis et al., 2020; Klammsteiner et al., 2020; Tanga et al., 2021).

*Actinomyces* sp. is a commensal, Gram-positive, filamentous facultative anaerobe that can degrade a wide array of organic materials such as chitin and lignin. It has also been used in the production of a variety of antibiotics while inhibiting the growth of other organisms in the gut, normally a benefit to the BSF larval gut (Wang et al., 2014).

*Dysgomonas* sp. is known for playing a key role in the degradation of recalcitrant and complex polysaccharides and has also been positively correlated with genes involved in carbohydrate, sulfate, and nitrogen metabolism (Shelomi et al., 2020). From a recent study, the metagenomic analysis of the gut of the BSF larvae, it was found out that the origin of a novel α-galactosidase gene that makes it possible to break down α-galactoses that are abundant in most non-digestible plant carbohydrates was traced back to a specific *Dysgonomonas* strain (C. M. Lee et al., 2018).

*Enterococcus* sp. is a Gram-positive, facultative anaerobe that contributes to gut health by availing important nutrients for the host organism (Klammsteiner et al., 2020), and in some insects such as the greater wax moth (*Galleria mellonella*), these species have been identified to provide immunity-related antimicrobial peptides (Krams et al., 2017).

*Bacteroides* sp. is a Gram-negative, obligate anaerobic bacteria. These species are non-endospore-forming bacilli and are either motile or nonmotile, reliant on the species(*Brock Biology of Microorganisms*, 2005). These bacteria make up a substantive portion of the gut microbiota especially in mammals and are involved in the fundamental role of breaking complex molecules into simpler molecules (Wexler, 2007).

In some microbiomes such as the human gut microbiome, PULs (Polysaccharide Utilization Loci) from the class have been identified in bacteria of phylum *Bacteroidetes*. These are genomic regions that form sets of gene clusters that work together to degrade the various carbohydrates ranging from the simple monosaccharides to the polysaccharides e.g. starch, glycogen, and the more complex glycans commonly regarded as dietary fiber (Terrapon et al., 2015). These genomic regions encode all the requisite machinery for the transport, binding, and depolymerization of complex glycan structures. Large and diverse CAZyme (carbohydrate-active enzymes) repertoires have been identified in species with large genomes from genus *Bacteroides* such as *Bacteroides ovatus* with polysaccharide lyases (PL) from ~60 distinct families and over 320 glycoside hydrolases (GH) (Seshadri et al., 2018). By identifying the PULs in the BSF larval gut samples, gene clusters responsible for lignocellulolytic activity can be identified.

## 2.7 RNA-sequencing and Metatranscriptomic Analysis

### 2.7.1 The Advancement of Microbiome Research

The microbiome is defined as a complex of symbiotic, commensal, and pathogenic microorganisms that inhabit a particular environment normally consisting of fungal, bacterial, and viral communities. However, the most predominant studies today are more inclined towards the bacterial components of these communities (Lederberg & McCray, 2001). Earlier microbiome research was majorly characterized by the usage of DNA sequencing-based 16S rDNA or shotgun metagenome sequencing to allow for the exposition of genome structure and microbial composition. With technological advances in RNA-sequencing, the scientific community has been granted the ability to gain further insight into the genes that are actively expressed in intricate bacterial communities. This has enabled the exposition of the functional changes that direct the roles of the microbiome under different circumstances, its host interactions, and changes in functions that accompany significant changes in an organism such as the onset of a disease, prolonged dietary change, or exposure to xenobiotics (Bashiardes et al., 2016). Metatranscriptomics allows for the characterization of the functional microbiome providing information on what functions the microbes present are carrying out and has the potential to better associate microbes with the host performances as compared to classical metagenomics (F. Li et al., 2019). It also allows researchers to discriminate between resident, active, gut microbes, or transient microbes that come with the diet and travel across the gut without colonization (Bost et al., 2018).

Metagenomics catalogs the genes present in a sample using DNA sequences while metatranscriptomics studies the RNA transcripts, and it gives detailed insights into the mechanistic understanding of the inter-community relationships between the microbial community and its host (Martinez et al., 2016). Classical microbial profiling methods are normally reliant on targeted sequencing of the highly conserved barcode gene regions i.e. 16S region for bacteria and the ITS (Internal Transcribed Spacer) region for fungi, followed by analyzing the microbial diversity, relative microbial abundances, and in some cases, the prediction of the microbial community’s functional capabilities (Langille et al., 2013). Shotgun sequencing is a more thorough and effective technique that aids in the identification of bacterial genes and their potential functions in a community. This is because it can decode the entire genetic material in a community at a better resolution. However, the analysis from both methods is limited to the study of the microbial composition and diversity characteristics of the microbiome of interest without elucidating the active functions of these microbial communities (Chung et al., 2020).

Succinctly, functional profiling with metatranscriptomics can tell what organisms are present in a microbial community and what functions they are performing. Functional microbiome RNA-sequencing data used for metatranscriptomics focuses on the active microbes in a microbiome of interest and is hence more reliable as it filters out the noise e.g. biases brought about by the use of DNA methods such as PCR and chimera biases, and not being able to differentiate live from dead microbiota (Qian et al., 2020). A substantial number of genes that might not be portrayed at a metagenomic level might be noticeably active at the metatranscriptomic level attributable to the metatranscriptomics’ propensity to focus on the active members of the microbiome (Shakya et al., 2019). Therefore, metagenomic analysis alone may underestimate or overestimate the functional importance of the transcribed genes in the microbiome while metatranscriptomics can be able to detect more lowly expressed genes due to its ability to discriminate active from inactive microbiota in a microbial community (Chung et al., 2020).

### 2.7.2 Long-read Sequencing

Third generation sequencing (TGS) technologies such as Oxford Nanopore Technology (ONT) and Pacific Biosciences (PacBio) have enhanced the progress of life science research by generating longer reads than SGS platforms. This has come as a response to demands by the scientific community for technologies that operate at faster speeds, more portable, and can generate longer reads (Lu et al., 2016). However, they are laden with high error rates normally attributed to the low signal-to-noise ratio; and therefore, a large error rate during base-calling which constitutes nearly double the magnitude of second-generation sequencing technologies (Laver et al., 2015).

Longer reads have allowed for sequencing of full genomes for smaller organisms, e.g. microbial genomes, and have allowed for sequencing through repetitive regions, detect mutations associated with diseases, and allow for the identification of gene isoforms and reliable discovery of novel genes with transcriptome sequencing (Fu et al., 2019; Rhoads & Au, 2015). Recent studies have also reported that long-read sequencing data exhibits lesser GC biases normally witnessed in second-generation sequencing platforms as they do not rely on PCR (Aird et al., 2012). To accurately measure gene expression occurring in a microbiome, it is important to obtain sufficient sequencing depth as well as use adequately long reads as they are less likely to be erroneously aligned to references.

### 2.7.3 The ONT MinION Sequencing Platform

The MinION is a portable device that weighs about 90g. It contains a flowcell at its core that bears about 2048 nanopores, each of which is individually addressable. Using an Application-specific Integrated circuit (ASIC), the nanopores are controlled in groups of 512. Adapters are first ligated to the nucleotide strands to facilitate strand capture and enzyme loading at the 5’ end of one strand. As the nucleotide strand passes through the pore, the sensor detects the changing ionic current due to the shifting nucleotide sequences present in the strand. For DNA, this is done for the template and complementary strand to produce a high-quality 2-D read (Jain et al., 2016). The MinION can be plugged into a computer directly using a standard USB3 port on a computer with a simple configuration and low hardware requirements i.e. 16GB cache and 1TB solid-state drive (SSD). With these specifications, the computer should be able to run MinKNOW, a proprietary software by ONT that carries out core tasks such as data acquisition; real-time analysis and feedback; sample identification and tracking. This ensures that the platform operates correctly in sample processing (Lu et al., 2016).

The sequences are interpreted computationally as 3-6 nucleotide k-words (kmers) with graphical methods i.e. De Bruijn graphs (Henson et al., 2012). While earlier assemblers used the Open Layout Consensus (OLC) paradigm e.g., Sangar sequencing, modern assemblers prefer graphical methods as they are faster. However, OLC-based algorithms perform better for long reads but with higher error rates. This is because longer reads allow for longer repeats and when read and overlap lengths are long enough, steps involving pre-masking and recovery of repeats can be omitted in the OLC algorithm However, for the De-Bruijn graph approach, most software accept k-mer sizes ranging from 31bp - 127bp, making the algorithm inefficient for assembling longer reads (Z. Li et al., 2012). The global error rate of MinION sequencing from ONT is about 11% and is caused majorly by inserted and deleted bases. At 40× sequencing coverage, MinION data can generate around one false insertion and substitution every 10–50 kb, and a false deletion at every 1000 bp, making the use of the technology still challenging in applications such as variant discovery. However, long reads present promising options for the discovery and identification of structural variants (Jiang et al., 2020). However, ONT’s MinION data is showing a lot of potential in detecting genomic regions at better specificity and sensitivity than other third-generation sequencing technologies. (Magi et al., 2017).

### 2.7.4 Multiplexing Sequencing Approach

A multiplex sequencing approach that involves barcoding multiple samples into one single sequencing run was recently integrated into the long line of ONT sequencing kits. The method combines multiplex PCR directly with barcoding of samples and high-throughput sequencing (Stiller et al., 2009). This has reduced the sequencing costs per sample significantly enabling the sequencing of multiple samples in a single flow cell at a fraction of the previous costs. However, this has been done at the expense of a reduced number of reads per sample. Therefore, a trade-off must be arrived at to sequence the maximum number of samples without immensely reducing the number of reads per sample to the point of losing statistical power.

The SQK-PCB109 kit by ONT allows multiplexing of up to 12 samples each with unique barcode primers in one sequencing flow cell. The kit possesses a time-efficient robust protocol that entails an RT-PCR step for reverse-transcription and strand switching, followed by selecting for full-length PCR transcripts, and ultimately assigning barcodes to the different samples that are to be pooled together in one sequencing run. The preceding kits such as the SQK-PCS109 incorporate an RT-PCR step for cDNA synthesis but do not entail the multiplexing of different samples in one sequencing run. The RNA-Seq PCR line of kits from ONT are regarded as ideal when the starting material (DNA/RNA) is of low concentration, but they are prone to amplification biases. However, a recent study by (Grünberger et al., 2021) reported that cDNA kits from ONT offer improved yield and accuracy without bias in quantification compared to direct RNA sequencing kits such as the SQK-RNA002 kit.

## 2.7.5 Ribodepletion in Metatranscriptomic Analysis

Most metatranscriptomic analysis pipelines necessitate the removal of ribosomal RNA (rRNA) by incorporating a ribodepletion step to increase the relative amount of mRNA reads. This step ensures that functional inferences are derived from mRNA sequences that encode various proteins. For samples that are not ribodepleted before sequencing, most of the RNA is normally rRNA (Westreich et al., 2018). For this reason, it is recommended that most of the rRNA be removed using laboratory mRNA enrichment methods, while residual rRNA can be removed using in-silico approaches. However, some bacteria genera have shown resistance to the commercially recommended rRNA depletion kits causing an overall bias towards the rRNA removal of some bacterial phyla over others as was reported by (Westreich et al., 2016).

### 2.7.6 Error Correction

Error correction methods fall into two broad categories: self-correction and hybrid correction. The self-correction strategy corrects reads that are error-prone by generating a consensus from a constructed arrangement of long-reads. In some correction software, self-correction involves correcting the erroneous regions using a consensus of overlapping long-reads. Others use De-Bruijn-based graphs with increasing kmer lengths. However, self-correction is only feasible for long-reads with high coverage. The self-correction approach is employed in tools such as PBcR (Berlin et al., 2015). The hybrid-correction methods, on the other hand, makes use of highly accurate and cost-effective short-reads. The methods can salvage more erroneous long-reads, especially those that have low coverage. Error correction tools like LoRDEC (Salmela & Rivals, 2014), proovread (Hackl et al., 2014), Jabba (Miclotte et al., 2016), and LSC (Au et al., 2012) are hybrid methods. Hybrid correction methods fall into three categories based on algorithm design: graph-based, alignment-based, and dual-based (graph/alignment based). Graph-based methods construct graphs i.e. De-Bruijn graphs, then search for matched/shared paths to the long-reads for correction. Alignment-based techniques map short-read sequences to error-prone long-reads and generate a consensus. The strategies are assessed and evaluated based on the following aspects: accuracy, sensitivity, alignment rate, runtime, memory usage, and output rate. Most importantly, the effects of the correction strategies on downstream analysis are also evaluated (Fu et al., 2019).

### **2.8 Taxonomic Validation using 16S rRNA Sequence**s

In metatranscriptomics experiments, 16S rRNA sequences have been used as additional validation for taxonomic assignment and phylogeny or as a preliminary step before meta-omics projects (Seshadri et al., 2018). Combining 16S rRNA and mRNA (metatranscriptomic) data can provide additional perspectives of the various factors involved in the manifestation of the study condition (Martinez et al., 2016). Novel sequence reads that might otherwise be rendered as unclassified under metatranscriptomic analyses might be broadly classified using the 16S rRNA analysis. This can be attributed to the widely available resources that are well-curated, and well-established pipelines for 16S analyses. By combining both metatranscriptomic and 16S rRNA analyses, the robustness of the findings obtained can be well ascertained (Peters et al., 2019).

### 2.9 Differential Expression of Genes (DEGs)

RNA-sequencing has become the idealized technique for the transcriptome-wide analysis of gene expression. However, there exist unique challenges in estimating expression from short sequence reads as the accurate alignment of reads in the occurrence of sequencing errors, measurement bias depending on the method of library preparation, and the intricacy in estimating the expression of distinct mRNA transcripts that share exons. Two or more conditions are compared to find the rate of change in genes expressed, also referred to as upregulation or downregulation. From the comparison, markers or functional changes of the parameter under investigation are inferred, thereby providing a reference estimate of underlying biological facts (Williams et al., 2017). Testing for differential expression at each gene is considered as one experiment; testing across thousands of genes requires corrections for multiple comparisons such as False Discovery Rate (FDR) or Bonferroni correction (McDonald, 2014).

The study and exploitation of uncultivated microbial communities to identify and isolate effective enzymes is still an active research area. Prokaryotic organisms possess high functional and metabolic diversity and are a rich source of genetic diversity. Metatranscriptomic approaches have been employed to elucidate novel pathways in different bacteria and archaea (Madhavan et al., 2017). There are several tools initially designed for single transcriptome analyses that can be applied for differential expression of genes in metatranscriptomics. The input required is mainly the abundance data per transcript or gene in every sample. Abundance data is acquired by either mapping the transcripts or gene to a reference, a reference gene set or de novo assembly (Shakya et al., 2019). To further identify upregulated or downregulated genes, R packages such as DESeq2 (Love et al., 2014), EdgeR (Robinson et al., 2010), and Limma (Ritchie et al., 2015) are mainly used. Other tools like Generally Applicable Gene-Set/Pathway Analysis (GAGE) can be utilized in the identification of enriched pathways in the study treatments (Luo et al., 2009). A major challenge that arises during metatranscriptomic differential expression analysis, is shared genes among closely related organisms and this can lead in false assessment of expression profiles. Normalization methods by basis of taxonomic composition has been proposed by (Klingenberg & Meinicke, 2017) though biases can be introduced by the taxonomic databases used (Shakya et al., 2019).

### 2.10 Carbohydrate-Active Enzymes (CAZymes)

CAZymes comprise all enzymes that are involved in the biosynthesis, modification, and modification of carbohydrates and their derivatives intracellularly and extracellularly. They are regarded as the most important enzymes in agricultural and bioenergy industries (Huang et al., 2018). The enzyme group is key towards understanding core biological activities such as protein glycosylation, metabolism of sugars, biosynthesis of compounds, and degradation. These enzymes have found great use in biotechnology within sectors including but not limited to bioenergy, textile, and other bio-based industries (Contesini et al., 2021).

CAZymes are divided into several classes based on their predominant functions; glycosyltransferases (GT) that are involved in carbohydrate assembly, glycoside hydrolases (GH), carbohydrate esterases (CE), and polysaccharide lyases (PL) that are involved in carbohydrate breakdown (Lombard et al., 2014). The classification of CAZymes into families is assigned based on significant amino-acid similarity with an extant biochemically characterized group. However, some groups that display little similarity with the classified enzyme groups are omitted, yet they may possess other functions. However, some marginal sequence hits are domiciled in the ‘unclassified’ section awaiting biochemical characterization (Lombard et al., 2014).

A database <http://www.cazy.org/> containing the unique classification and curation of these enzymes was created in 1991 and to date comprises more than 300 families of >400 unique protein signatures (Busk et al., 2017). Bioprocesses have shown high dependence on polysaccharide-degrading CAZymes for the deconstruction of lignocellulosic fractions of plant biomass during their transformation into value-added products such as biofuels and animal feeds. CAZymes are popular in biofuel production due to their ability to break down recalcitrant polysaccharides e.g.cellulose and hemicellulose into fermentable sugars such as glucose and xylose. The identification of lignocellulose-metabolizing microbial communities, coupled with the curation and accurate prediction of CAZyme functions has become a fundamental step in industrial bioprocesses. This has been pushed further to improve the catalytic ability of such microbial communities isolated from metatranscriptomes and metagenomes of interest. Microbial sequencing projects unveil a significantly diverse pool of CAZyme encoding genes, some of whose functions remain unknown. Transcriptomic and metatranscriptomic analyses have emerged as ways of acquiring relevant insights on the enzymatic machinery utilized by microorganisms to break down recalcitrant plant biomass. Such analyses reveal sets of expressed genes and their encoding transcription factors, as well as the responsible enzymes. The enzymes involved in these degradative pathways are mainly CAZymes (Contesini et al., 2021). CAZymes used in biofuel production are mainly obtained from genetically engineered organism strains or commercially produced (Tingley et al., 2021).

In the past one and a half decades, studies have shown that different CAZymes work together to effusively break down recalcitrant carbohydrates and that the genes that encode these proteins have a tendency of forming gene clusters that are physically linked (Ausland et al., 2021). These gene clusters have been referred to as PULSs and have been cited with concerted efforts that orchestrate the breakdown of complex polysaccharides (Terrapon et al., 2018).

# 3.0 CHAPTER THREE: MATERIALS AND METHODS

### 3.1 Research Design and Approach

The study was a randomized control trial (RCT) where BSF larvae were selected then assigned to different dietary groups (Chalmers et al., 1981). There were four dietary groups, three experimental diets, and one control diet. The experimental diets were brewers’ spent grain (BSG), chicken manure (CM), and water hyacinth (WH), while the processed chicken feed (CF) was used as the control diet. An additional fifth diet consisting of the 4 diets mixed in equal proportions (1:4) feed mix (FM) was also included as the starting substrate and to serve as an additional control. The WH dietary substrate contained water hyacinth leaves and stems mixed in equal proportions. The larvae were randomly sampled from each dietary group during the different larval phases and their entire guts were extracted for isolation and sequencing of RNA by the ONT MinION long-read sequencing platform using the cDNA approach. The data generated was used to perform metatranscriptomics analysis and functional annotation analysis.

### 3.2 Ethical Considerations

This work was presented to the Makerere University College of Health Sciences School of Biomedical Sciences Research and Ethics Committee (SBS-REC) and was approved on 18th October 2021 for one year. A research permit was also obtained from the National Commission for Science Technology and Innovation (NACOSTI) in Kenya as a mandatory research requirement.

### 3.3 Growth conditions and colony maintenance

The study was conducted at the International Center for Insect Physiology and Ecology (*icipe*), Duduville campus, Kasarani, Nairobi, Kenya, where the BSF larvae were bred molecular work conducted, and sequencing conducted. The Black Soldier Flies were sourced from the *icipe* insectary where they are bred under the Insects for Food and Feed (INSEFF) program.

BSF larvae were reared in clean plastic crates of dimensions 1m by 0.5m. The larvae were bred on four different dietary substrates selected based on varying lignocellulosic/fiber content: chicken manure (CM), brewers’ spent grain (BSG), water hyacinth (WH), and processed chicken feed (CF) that was used as the control diet. An additional fifth diet consisting of the 4 diets mixed in equal proportions (1:4) feed mix (FM) was used as the starting substrate from Day 0- Day 9 after the BSF eggs were collected, and was used as an additional control. From Day 10, the larvae were moved to their respective diets (BSG, CF, CM, WH) while the FM diet was maintained. The substrate was replenished every 5 days until the experiment was terminated (Heussler et al., 2019). The feeding of the larvae was terminated once 25% pupation was achieved for each dietary substrate. Samples were collected weekly for entire gut dissection and stored in RNALaterTM at -80˚C. Each dietary substrate was composed of >200 pooled larvae kept at 27˚C, 60% - 70% humidity, and a photoperiod of 12h:12h in a greenhouse structure. All the feedstocks were milled and the moisture standardized by oven drying at 105˚C until they achieved uniform dryness. Distilled water was then added to each dietary substrate in the ratio of 3:2. The gender of the larvae was ignored, while other parameters such as the length, color, and stage of growth were collected during sampling to guide further analyses.

### 3.4 Sampling

After every 3 days from day 10, the BSF larvae in each diet were randomly sampled and the length, color, substrate reduction index, the pupation span and rate per diet, and the pH and temperature of the frass waste of the BSF larvae diets were recorded. The pH is informative of the physicochemical properties of the BSF larvae and their environment and has been found to increase with the increase in larval activity. The pupation span helped us understand how each diet impacted the successful pupation of the BSF larvae, while the pupation rate is informative of the survival of the larvae in each of the given diets (Meneguz et al., 2018).

Taking measurements for these parameters was done to provide grounds for comparison, while the different dietary substrates selected were hypothesized to adjust the gut microbiota of the BSF larvae to metabolize the various substrates under different conditions such as pH, ionic concentrations, and alcohol concentrations that are likely to be present in biofuel production plants. Additionally, we further quantified how the diet affects microbial profiles in the gut and the corresponding CAZymes they express.

### 3.5 RNA isolation

BSF larvae gut samples stored in RNALaterTM at -80˚C were randomly selected to include all five dietary substrates. The gut samples were thawed on ice for 1 hour then RNA was extracted using the Bioline ISOLATE II RNA Mini Kit from Meridian Bioscience as per the manufacturer’s specifications. Lysis buffer (400µl) and 2-Mercaptoethanol (3.5µl) were added to each of the gut samples and homogenized for 5 minutes using the Qiagen Tissue Lyser in 1.5ml microfuge tubes. This was done to inactivate RNases and prevent the degradation of released RNA. The RNA binding conditions were adjusted by adding 70% ethanol (350µl) in each tube followed by brief vortexing (2 × 5s). The binding of RNA was done using a spin column provided by the manufacturer. After binding, desalting was performed in a silica membrane using the membrane desalting buffer by centrifuging at 11,000 × g for 1 minute. DNase1 was then used to selectively remove genomic DNA. The RNA, bound to a silica membrane was then subjected to 2 wash steps using wash buffers provided by the manufacturer before eluting using RNAse free water (*ISOLATE II RNA Mini Kit | Bioline | Meridian Bioscience*, n.d.).

### 3.6 cDNA Synthesis and mRNA Enrichment

The cDNA libraries were prepared using the PCR-cDNA approach using the Barcoding kit (SQK-PCB109) kit from ONT according to the manufacturer’s instructions. Total RNA was used as the template from 16 samples. Library preparation involved mRNA selection using VN “anchor” primers (VNP) and reverse transcription and strand-switching step using strand-switching primers (SSP) for 90 min at 420C, followed by heat inactivation for 5 minutes at 850C. This was followed by selecting for full-length cDNA transcripts by PCR using the following conditions: initial denaturation for 30 seconds at 950C, followed by 17-18 cycles of denaturation for 15 seconds at 950C, annealing at 620C for 15 seconds, and extension at 650C for 120-150 seconds. The final extension was carried out for 6 minutes at 650C.

The SQK-PCB109 kit allows multiplexing up to 12 samples in one sequencing run using unique barcodes provided with the kit. Barcodes were assigned to the 16 samples in two sequencing runs. For the first sequencing run, 5 samples were multiplexed and assigned barcodes 07-11, while 11 samples were multiplexed for the second sequencing run that contained all barcodes apart from barcode 05. To remove any unamplified primers and single strands, 1µl of Exonuclease I was added to each sample. Gel electrophoresis was carried out to analyze the integrity and size of the resultant cDNA libraries. The 1.5% agarose gel was run in Tris/Borate/EDTA (TBE) buffer at 70 volts for 1 hour with 3.5µl of each sample, 2µl of loading dye, and 1.5µl of 1Kb ladder.

The cDNA libraries were subsequently cleaned using the magnetic Agencourt AMPure XP beads and pelleted on a magnetic stand. This was followed by 2 washes using 70% ethanol diluted using Nuclease-free water. The ethanol was removed and discarded, then the beads were allowed to dry off any residual ethanol, before being eluted in 12µl Elution Buffer (EB) provided with the kit. The amplified DNA (1µl) was then analyzed for quantity using the NanoDrop 2000. Guided by their concentrations, all the amplified barcoded cDNA samples totaling 100 fmol were pooled together in a 1.5ml Eppendorf DNA LoBind tube to a final volume of 11 µl in Elution Buffer. This was done for both sequencing runs. Before sequencing, 1µl of sequencing adapters (RAP) were added to the amplified cDNA libraries before mixing by pipetting.

### 3.7 Sequencing

The MinION MK1B platform was used for both sequencing runs. However, two types of flow cells were used: R9.4.1 and R10.4. Better throughput was achieved with version R9.4.1 than version R10.4, as the latter is recommended for sequencing homopolymer regions and is regarded as ideal for use with the newer sequencing chemistries being developed by ONT for experiments where high consensus accuracy is required (~99%) (Oxford Nanopore Technologies, 2020).

The flow cells require priming before the cDNA libraries can be loaded into the flow cell sequencing port (SpotON), and this was done using the Flow cell Priming Kit EXP-FLP002 from ONT while adhering to the manufacturer’s specifications. 37.5 µl of sequencing buffer (SQB), 25.5 µl of loading beads (LB), and 12 µl of the cDNA library were properly mixed in a LoBind 1.5 ml Eppendorf tube and loaded onto the SpotON sample sequencing port. The MinKNOW software (version 21.06.0) was used for data acquisition, real-time analysis and feedback, and sample tracking and identification, while the Guppy (v5.0.11) was used to perform basecalling after sequencing using the High-Accuracy Model (HAC) model (Lu et al., 2016).

Basecalling was run on an NVIDIA Tesla V100 Graphical Processing Unit (GPU) with the Guppy software (v5.0.11) using following command line parameters: --recursive --config dna\_r9.4.1\_450bps --barcode\_kits “SQK-PCB109” --detect\_mid\_strand\_barcodes --min\_score\_mid\_barcodes 60 --gpu\_runners\_per\_device 16 --compress\_fastq --device cuda:0. The --detect\_mid\_strand\_barcodes option was used to classify more reads whose barcode sequences are located within the sequences with an identity threshold of 60 that might have otherwise been considered as chimeric reads. The configuration file was modified and the quality score of the reads was set at the minimum threshold of 7. Every read that failed to attain this threshold was discarded as a failed read.

### 3.8 Data Analysis

### 3.8.1 Physicochemical parameter collection

During BSF larval breeding, various physicochemical parameters were collected and recorded to provide an in-depth understanding of the state, larval activity, and performance of each dietary group. The parameters were collected using the parameter collection template (**Appendices - page 106**) until 25% pupation was attained in each dietary substrate. These parameters were analyzed and visualized using Microsoft Excel 2016.

### 3.8.2 Feed composition analysis

To acquire a complete understanding of the material composition of the dietary substrates used to feed the BSFL, a complete feed composition analysis was carried out using the wet chemistry Inductive Coupled Plasma – Mass Spectrometry technique (Wetchem, ICP-MS). The ICP-MS is regarded as one of the most versatile tools used in bio-analytical and environmental applications for the quantification of elements, by utilizing noble gases at high temperatures to fragment the samples to be analyzed into detectable ionized molecules that are then used as surrogates to detect complex molecules e.g. proteins (Prange & Pröfrock, 2012).

### 3.8.3 Quality statistics of basecalled reads with pycoQC

Once basecalling was complete, the sequencing\_summary.txt file generated by Guppy (v5.0.11) was used to obtain sequencing statistics such as length distribution of the reads, pass reads, read quality, and run duration using the PycoQC tool using the following command:

pycoQC –f $sequencing\_summary.txt -o $RESULTS/PyqoQc\_results.html .

PycoQC generates interactive plots and control metrics from basecalled reads sequenced on ONT platforms using summary files generated from basecalling software such as Guppy and Albacore (Leger & Leonardi, 2019).

### 3.8.4 Trimming, orienting, and defusing cDNA reads with Pychopper

Pychopper (v2.2.0) is a bioinformatics tool written in the python language that is used to identify, orient, trim and defuse cDNA reads generated from ONT platforms. The tool accomplishes these tasks by identifying the alignment hits of the primers used across the length of basecalled sequences. The tool possesses two inbuilt backends, *nhmmscan* which is set as the default and is pre-trained using Hidden Markov Models (HMM), and the *edlib* backend which uses a mixture of local and global alignment strategies to detect primer hits within the reads. This tool also employs dynamic programming algorithms to defuse and rescue merged reads. To prevent assigning such rescued reads twice, once a hit is identified as a rescued read, the subsequent segment must be excluded since it has been used up in the preceding segment (*nanoporetech/pychopper: A tool to identify, orient, trim and rescue full-length cDNA reads*, 2018).

The *edlib* backend produced better classification and better overall performance than the default *nhmmscan* backend. Pychopper was run using 32 CPU threads, -m option specified the edlib backend –r option specified the names of the output pdf reports, the –u option specified the names of the unclassified fastq files, -w specified the rescued fastq files, while the full\_length\_output.fastq was the full length fastq containing the rescued using the command below:

cdna\_classifier.py -t 32 -m edlib -r $RESULTS/report.pdf -u $unclassified.fastq -w $rescued.fastq $output.fastq $full\_length\_output.fq

### 3.8.5 Adapter trimming with Porechop

Porechop (v0.2.4) is a software tool written in C++ used to find and remove sequencing adapters from ONT reads. Adapters located at the end of the reads are trimmed off while those reads that have adapters located in the middle are treated as chimeric reads and are split into different reads (Wick et al., 2017). Porechop was run with default parameters using the command shown below:

porechop -i $ full\_length\_output.fq -o $adapter\_free.fastq --threads 32

### 3.8.6 De novo clustering of cDNA isoforms with IsONclust

IsONClust is a bioinformatics tool written in the Python language that is used to cluster long reads from PacBio and ONT platforms into gene clusters where every cluster represents a set of reads that came from a particular gene. The output is a tab-separated file with every read allotted to a cluster-ID, and a sorted fastq file containing merged reads that have been clustered. The tab-separated file contains rows equivalent to the number of reads in descending order with the largest cluster occupying the first position and the smallest cluster occupying the last position. This same order is adhered to in the sorted fastq file (Sahlin & Medvedev, 2020).

IsONclust (v0.0.6.1) was run using 32 CPU threads and used to cluster our adapter free reads using the following command:

isONclust --ont --t 32 --fastq $mergfqs/$adapter\_free.fastq --outfolder $isonclust/$clustered.fastq

To obtain separate fastq clusters to be used during the subsequent correction step, the following command was run using the default parameters:

isONclust write\_fastq --clusters $isonclust/$clustered.fastq/final\_clusters.tsv --fastq $mergfqs/$adapter\_free.fastq --outfolder $isonclust/fastq\_clusters --N 1

### 3.8.7 Error correction of clustered cDNA reads with IsONcorrect

Despite ONT’s capability to sequence transcripts end-to-end, the technology is still associated with high error rates limiting its uses to reference-based analyses. IsONcorrect jointly leverages the shared regions from the cDNA isoforms clustered de novo using isONclust, allowing it to perform error correction even at low sequencing depths to achieve consensus accuracies of about 99% (Sahlin et al., 2021).

IsONcorrect (v0.0.8) was run using 32 CPU threads with default parameters using the following command:

run\_isoncorrect --t 32 --fastq\_folder $clusfqs --outfolder $correction/

The output of the above command is one file for each cluster with headers identical to the original reads.

### 3.8.8 Ribodepletion with SortMeRNA

Ribodepletion refers to ribosomal depletion, a key step in transcriptomics analysis that enables the efficient detection and removal of highly abundant rRNA species. Most of the rRNA was removed during the library preparation step, but residual rRNA species have been known to cause misclassification in metatranscriptomics during annotation steps (Martinez et al., 2016).

SortMeRNA is a bioinformatics tool that is designed to filter out rRNA fragments from metatranscriptomic datasets, by aligning the sequences to various rRNA databases. Non-coding RNA such as rRNA makes up to 90% of total RNA. Therefore, it is necessary to eliminate these non-coding RNA species before sequencing to maximize the number of coding RNA required for expression analyses (Kopylova et al., 2012).

SortMeRNA (v4.3.3) was run using 32 CPU threads for both corrected and uncorrected reads using the command below:

*for r in `cat $sample\_list`*

*do*

*sortmerna -ref $silva-bac-16s-database-id85.fasta \*

*-ref $silva-arc-16s-database-id95.fasta \*

*-ref $silva-euk-28s-id98.fasta \*

*-ref $silva-euk-18s-id95.fasta \*

*-ref $silva-arc-23s-id98.fasta \*

*-ref $rfam-5.8s-database-id98.fasta \*

*-reads ${r}[corrected/uncorrected]\_reads.fastq \*

*--aligned ${r}\_rna[.fq] --other ${r}\_clean[.fq] --threads 32 \*

*-sam -fastx -blast 1 -num\_alignments 1 -v*

*done*

The –sam parameter was specified to return the alignments in the .sam file format while the --fastx format specified that the output files be converted to fastq format. The --blast 1 option was specified to output a single alignment in Blast-like format, while the -num\_alignments 1 option specified only the best alignment as output. The –v option was invoked to verbose the index building step (Kopylova et al., 2012).

### 3.8.9 Taxonomic Validation using Filtered rRNA reads

16S rRNA reads filtered out from the study metatranscriptomes during ribodepletion with SortMeRNA were used for further taxonomic validation of bacterial sequences. From the previous step, an assortment of rRNA reads (both eukaryotic and prokaryotic) was obtained. Additional filtering was done using SortMeRNA to retain only the bacterial rRNA sequences (16S rRNA) using the Silva-16S-database at 95 percent identity (Quast et al., 2013). A custom pipeline was designed to perform 16S rRNA analysis that involved: (i) merging the sequence files from the same metatranscriptome, (ii) quality filtering and sequence dereplication using Vsearch (v2.16.0) (Rognes et al., 2016) to remove spurious and duplicated sequences using *vsearch -fastq\_filter* *$file.fq* and *vsearch --derep\_fulllength $filtered\_file.fa* commands, (iii) getting rid of chimeric sequences and generating amplicon sequence variant (ASV) files using Usearch (v11.0) (Edgar & Bateman, 2010) to distinguish sequences that differ at minimal base positions that would have otherwise been lost by performing classical operational taxonomic unit (OTU) clustering. This was achieved using the *usearch -unoise3 unique\_seqs.fa -zotus ASVs.fa* command. (iv) An ASV counts table was generated using the command *vsearch -usearch\_global $filtered\_file.fa --db ASVs.fa --id 0.85 --otutabout $ASV\_counts.txt* set at 85% identity (specified by the --id 0.85 flag). (v) Taxonomic assignment using the Ribosome Database Project (RDP) classifier for 16S rRNA sequences (Cole et al., 2014).

The Phyloseq package (v1.24.2) (McMurdie & Holmes, 2013) in Bioconductor (Love et al., 2014) was used in 16S rRNA statistical analysis using in-house R Software (v4.0.2) scripts.

### 3.8.10 Alignment with Minimap2

Minimap2 has emerged as a multipurpose sequence alignment program that aligns DNA, cDNA, or mRNA sequences against various reference databases. The program has especially shown superior performance with various long-read datasets from ONT and PacBio platforms. Another major advantage of the program is the fast runtimes compared with other mapping programs. The program reports ≥30 times faster speeds than other long-read aligners and is 3-4 times faster than other short-read alignment programs. Minimap2 is also easy to use, allowing users to index the reference genomes/transcriptomes and perform alignment in one command (H. Li, 2018).

Minimap2 was used to align the reads to the BSF reference genome downloaded from National Center for Biotechnology Information (NCBI) genomes database: <https://ftp.ncbi.nlm.nih.gov/genomes/refseq/invertebrate/Hermetia_illucens/all_assembly_versions/GCF_905115235.1_iHerIll2.2.curated.20191125/>. The reference genome was downloaded alongside the annotation and transcriptome information. The following commands were used to perform indexing and alignment for both corrected and uncorrected reads:

*minimap2 -t 16 -G 500k -k 13 -w 5 -ax splice $genome \ $[corrected/uncorrected]\_ribodepleted.fastq > $[corrected/uncorrected].sam*

To improve the accuracy over the default parameters, non-default parameters were invoked. The aligner was run using 16 CPU threads (-t 16), the -ax splice parameter was used to specify the nature of the reads, which were ONT 2D cDNA reads. The –k option that specifies the k-mer size was set at 13 from the default 15, while –w specified the minimizer window size was set at 5 from the default 10. The –G option that specifies the maximum intron length was set at 500,000 from a default 200,000 to enable the identification of mapped isoforms longer than 200,000 bases. The use of the above non-default parameters was reported by (Sahlin & Mäkinen, 2021) to improve the number of correct alignments over the default parameters.

### 3.8.11 Alignment statistics with samtools

SAMtools software is used to manipulate and parse alignments in the SAM/BAM file formats. It can perform functions such as converting files from various alignment formats, merging and sorting alignments, generating per-position information, and mapping statistics such as coverage, depth, mapping quality, percentage of mapping among others (H. Li et al., 2009).

SAMtools (v1.12) was used to obtain mapping statistics for corrected and uncorrected reads. The .sam files generated from alignment with Minimap2 were converted to .bam files, before being sorted and indexed. This was followed by running the $*samtools flagstat* command to obtain the percentage of reads mapped, and $*samtools coverage* to obtain the coverage, depth, and mapping quality statistics. Finally, mapped and unmapped reads were obtained separately, sorted, indexed, and converted into fastq files for subsequent steps. This was done using the following commands:

*# Generating .bam files...*

samtools view -b $[corrected/uncorrected].sam > $[corrected/uncorrected].bam

*# Sorting the .bam file...*

samtools sort $[corrected/uncorrected].bam -o $[corrected/uncorrected]\_sorted.bam

*# Indexing the sorted bam file...*

samtools index $[corrected/uncorrected]\_sorted.bam

*# Running flagstat*...

samtools flagstat $[corrected/uncorrected]\_sorted.bam >> $[corrected/uncorrected].tsv

# *Obtain coverage statistics...*

samtools coverage $[corrected/uncorrected]sorted.bam -o $[corrected/uncorrected]\_coverage.out

The mapping statistics generated above namely coverage, mean depth, and percentage of reads mapped to the reference were used as the metrics for evaluating the efficacy of the clustering and error correction steps in improving the accuracy and throughput of the sequencing data. The paired t-test (α=0.05) was used to check whether the differences between these two groups were statistically significant. The reads that showed better alignment scores were used for the subsequent commands to generate the mapped and unmapped .fastq files from the sorted and indexed .bam files as shown in the commands below:

# *Getting unmapped reads in .bam file...*

samtools view -b -f 4 $sorted.bam > $unmapped.bam

# *Getting mapped reads in .bam file...*

samtools view -b -F 4 $sorted.bam > $mapped.bam

# *Indexing the unmapped reads…*

samtools index $unmapped.bam

# *Indexing the mapped reads…*

samtools index $mapped.bam

# *Sorting the unmapped bam files...*

samtools sort $unmapped.bam -o $unm\_sorted.bam

# *Sorting the mapped bam files...*

samtools sort $mapped.bam -o $map\_sorted.bam

# *Converting the unmapped reads to .fastq files…*

samtools bam2fq $unm\_sorted.bam > $unmapped.fastq

# *Converting the unmapped reads to .fastq files…*

samtools bam2fq $map\_sorted.bam > $mapped.fastq

### 3.8.12 Obtaining raw read counts

To compute diversity statistics and differential expression statistics with DESeq2 (Love et al., 2014) in R software, raw read counts had to be computed from each sample using a Python script (raw\_read\_counter.py) adopted from the SAMSA2 analysis pipeline (Westreich et al., 2018) by specifying the unmapped .fastq files as input, and the tab-separated .txt file containing two columns - the sample, and the counts recorded as output using the following command:

python $raw\_read\_counter.py –I $Unmapped\_corrected.fastq -O $ rawcounts.txt

### 3.8.13 Annotation of unmapped reads with DIAMOND

The SAMSA2 metatranscriptomics analysis (Westreich et al., 2018) pipeline was adopted for this step. The pipeline uses the DIAMOND program (Buchfink et al., 2021) for annotation of metatranscriptomes, a BLAST-like aligner that is designed to deal with the large number of reads generated from metatranscriptomics experiments at 10,000 times the speed of classical BLAST. DIAMOND is capable of performing annotation against multiple reference databases. The NCBI bacterial RefSeq database (Tatusova et al., 2014) was used to generate organism and functional annotations, while the SEED subsystems database was selected to generate annotations for functional activities in a hierarchical format (Overbeek et al., 2014).

The databases had to be indexed into DIAMOND format using the following command:

diamond makedb --in $diamond \_database --db $database\_basename

The --db flag was used to indicate the basename of the indexed files. The commands used to perform the annotation step and create a results table for the aggregation of organism and functional counts were as follows:

1. $diamond blastx --db $diamond\_database -q $diamond\_indexed\_input\_sequences -a $output --sensitive -t ./ -k 1
2. $diamond view --daa $RefSeq.daa -o $output -f tab

From (i) above, the blastx option indicates DIAMOND performed a translated nucleotide search (BLASTx), while the --db option specified the reference database to be used. The -q option was used to specify the query name of the diamond indexed input sequences, while -a was used to indicate the name of the output file in DIAMOND format. As DIAMOND ran, temporary files and directories were generated, and the -t flag was used to specify their location. To increase the consensus accuracy of the annotation step, the --sensitive option was specified. For command (ii), the --daa option was used to specify the output files in .daa format, while -o was used to specify the output file name in a tab-separated format (specified by the -f option).

### 3.8.13 Aggregation of annotated reads

After annotation, DIAMOND generated results with each sequence from the metatranscriptome that had a corresponding hit from the reference database occupying a single line in the annotated output file. Aggregation scripts from the SAMSA2 pipeline (Westreich et al., 2018) written in Python language were adopted to aggregate the annotated output files into summarized tables. These scripts generate two files (for organism and functional annotations) containing three columns from the annotation output from the previous step. The first column contained the percentage of each entry compared to the total reads, the second column contained the respective read counts, and the third column contained the annotation (organism or function). This was achieved using the following commands:

1. python $standardized\_DIAMOND\_analysis\_counter.py -I $anotation\_file –D $RefSeq\_db -O
2. python $standardized\_DIAMOND\_analysis\_counter.py -I $anotation\_file –D $RefSeq\_db –F

The -I option referred to the annotation file from the previous step that is used as the input, -D referred to the database used. The -O option used in command (i.) performed aggregation of all reads by organism while –F in command (ii.) aggregated all reads by function.

For hierarchical annotations using the SEED subsystems database, a separate Python script (DIAMOND\_subsystems\_analysis\_counter.py) from the SAMSA2 pipeline (Westreich et al., 2018) was used to generate outputs containing information from the different hierarchy levels for later use in R software for statistical analysis and visualization. The output was further compressed using the DIAMOND\_subsystems\_reducer.py to get the results into a more summarized format for statistical analysis by removing redundant annotations. The following commands were used for this step:

1. python $DIAMOND\_subsystems\_analysis\_counter.py -I $file -D $Subsys\_db –O $file.hierarchy -P $file.receipt
2. python $DIAMOND\_subsystems\_reducer.py -I $file.hierarchy

The output was a tab-separated summarized file with non-redundant entries used in R software to generate SEED subsystems dodged bar plots.

### 3.8.14 Statistical analysis and visualization

Statistical analysis was performed using R software (V4.0.2). The R scripts used were adopted from the SAMSA2 pipeline, except the ones to generate SEED subsystems dodged bar plots which were custom-made scripts. Some of the scripts were also modified to improve their functionality to achieve the desired output. The scripts were utilized to normalize raw read counts, calculate Shannon and Simpson diversity statistics, generate stacked taxonomic bar plots, principal component analysis (PCA), and calculate differential expression statistics for the various organisms and functions from DIAMOND annotation results. These scripts were used to perform within-group and between-group comparisons between control and experimental samples while generating the pertinent visualizations.

### 3.8.15 Annotation of CAZymes with the Hotpep module of dbCAN

The Homology to Peptide (Hotpep) module (Busk et al., 2017) of dbCAN (v2.0.11) (Huang et al., 2018) was used to annotate highly active CAZymes in the different metatranscriptomes. The first step involved downloading the sequences of the five enzyme classes present in the CAZymes database: Auxiliary activities (AA), Carbohydrate esterases (CE), Glycoside hydrolases (GH), Glycosyltransferases (GT), and Polysaccharide lyases (PL). The Hotpep module performs protein annotations by matching conserved (signature) peptides to the protein sequences of interest. For a sequence to be assigned to a particular CAZyme family, it was required to contain at least three conserved peptides belonging to that CAZy group (Busk et al., 2017).

To run this step, all the samples from one condition were merged and converted into fasta format. The dbCAN software and the CAZymes database were downloaded and compiled using the instructions on dbCAN’s Github repository (<https://github.com/linnabrown/run_dbcan>), then the following command was used to perform CAZymes annotation on each of the metatranscriptome fasta files:

run\_dbcan.py $file.fasta meta --db\_dir $dbase --dbCANFile dbCAN-HMMdb-V8.txt -t hotpep --hotpep\_cpu 20 --out\_dir $dbcan2\_results

The --db\_dir option specified the location of the indexed database to be used for annotation. dbCAN software primarily uses HMMs. The HMM database was specified using the --dbCANFile option. The -t option specified the use of the Hotpep module in dbCAN, while the --hotpep\_cpu option specified the number of CPU cores that the Hotpep module used.

### 3.8.16 Identification of species of interest with Krona enzyme-specific multi-layered pie-charts

After annotation of CAZymes with the Hotpep module of dbCAN 2, the CAZy database ([www.cazy.org](http://www.cazy.org)) was used to validate the annotations, and to identify the species of interest using the Krona multi-layered taxonomic pie-charts which show the number of CAZy modules that have been identified in each species. This enabled us to navigate across the different levels of taxonomy and identify the species abundant in the study metatranscriptomes that might have contributed to the expression of the various CAZymes. The polysaccharide-degrading enzymes expressed in each metatranscriptome together with the four most abundant bacterial genera and their subspecies (abundance > 0.1%) were selected for this step. Krona charts allow for interactive identification of species and easy taxonomic navigation by zooming in and out to attain the desired taxonomic resolution (Ondov et al., 2011).

### 3.8.17 Screening for PULs from lignocellulolytic CAZyme families

To further understand whether the highly abundant bacteria identified from our metatranscriptomes were organized as PUL gene clusters responsible for producing lignocellulolytic CAZymes, the dbCAN-PUL BLASTx resource (<https://bcb.unl.edu/dbCAN_PUL/blast>) (Ausland et al., 2021) was used to screen for PULs in the identified CAZy families in comparison with the control diet metatranscriptome sequences.

# 4.0 CHAPTER 4: RESULTS

## 4.1 Growth parameters collection

To understand how the BSF larvae responded to various dietary substrates, various physicochemical parameters were collected over the course of the experiment. Figure 4-9, below, show how the different parameters varied in the dietary substrates used for BSF larvae breeding, from the egg collection date (Day 0) to the termination of the experiment upon attainment of 25% pupation (from day 39 to 68). A standard substrate FM (Feed Mix), that contained each of the four diets in equal proportion was used as the hatching substrate. On Day 9, the larvae were moved from the hatching substrate to their respective dietary substrates. Parameter collection began on Day 14 and ended on Day 68 when all BSF larvae from the various dietary substrates had attained at least 25% pupation.

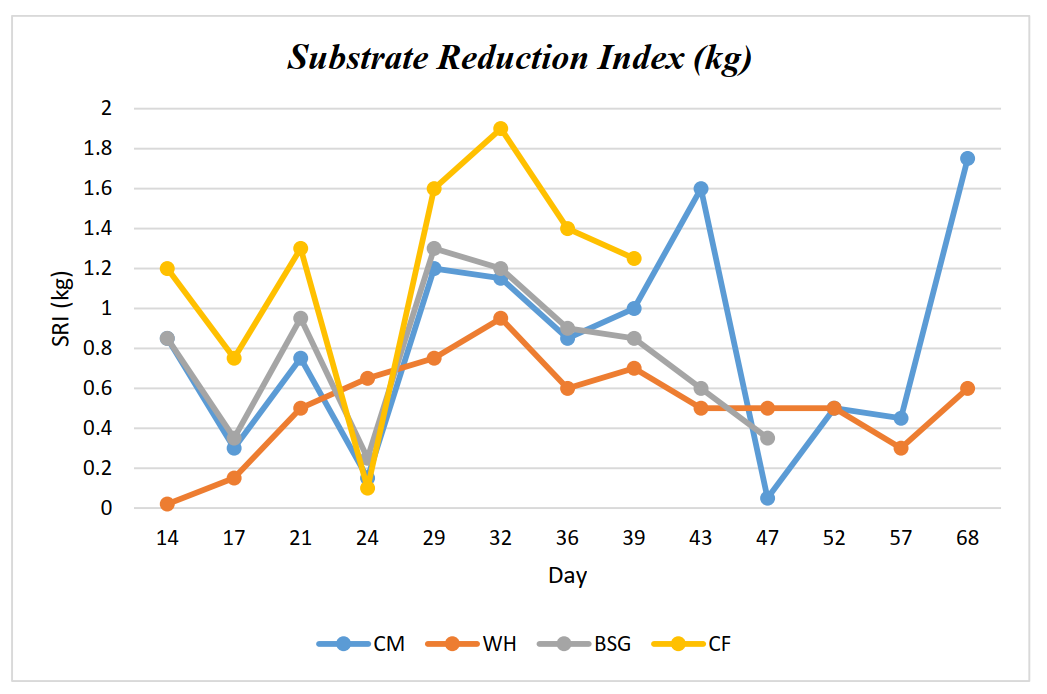


Figure 4: Substrate reduction index

The Substrate Reduction Index (SRI) in kg was used to understand the ease of digestibility of the various dietary substrates used to feed each group of larvae. The highest SRI was recorded for the CF dietary substrate (peaked at ~1.9 kg) while the lowest SRI was recorded for the lignocellulose-rich WH dietary substrate (peaked at ~0.95 kg).

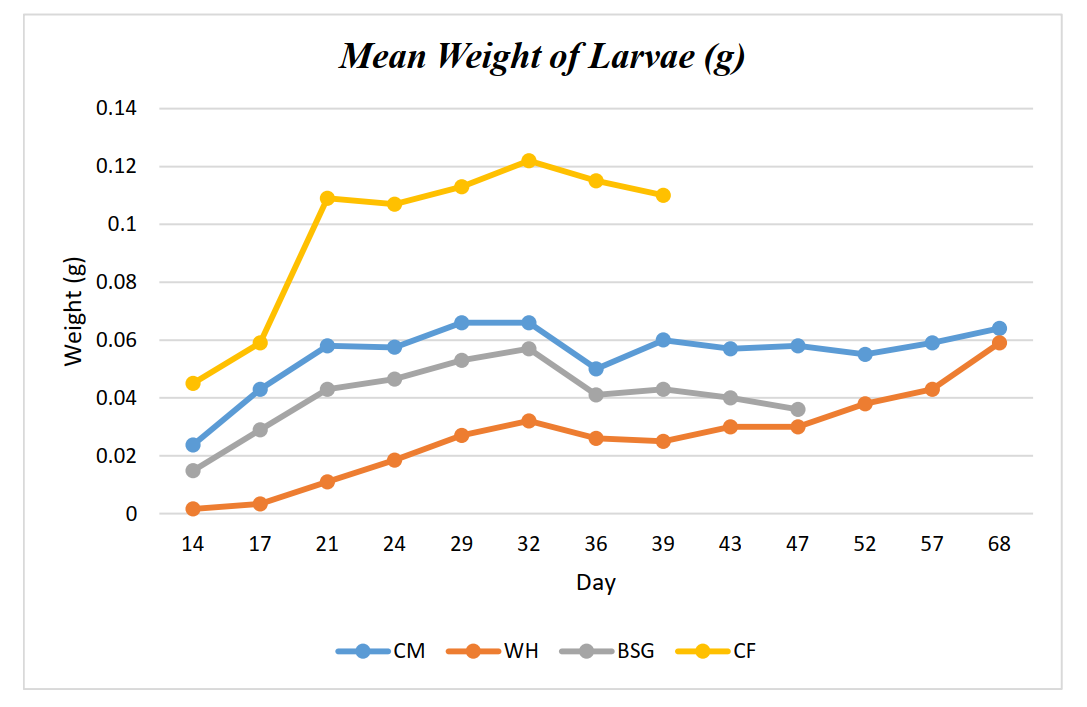


Figure 5: Mean weight of larvae

The ease in the digestion of the dietary substrates used to feed the BSF larvae was directly proportional to the BSF larval weight gain, with the easily digestible control diet, CF, recording the highest mean weight per larva while the lignocellulose-rich WH and BSG dietary substrates recorded the lowest mean weight per larva as shown in **Figure 5**.

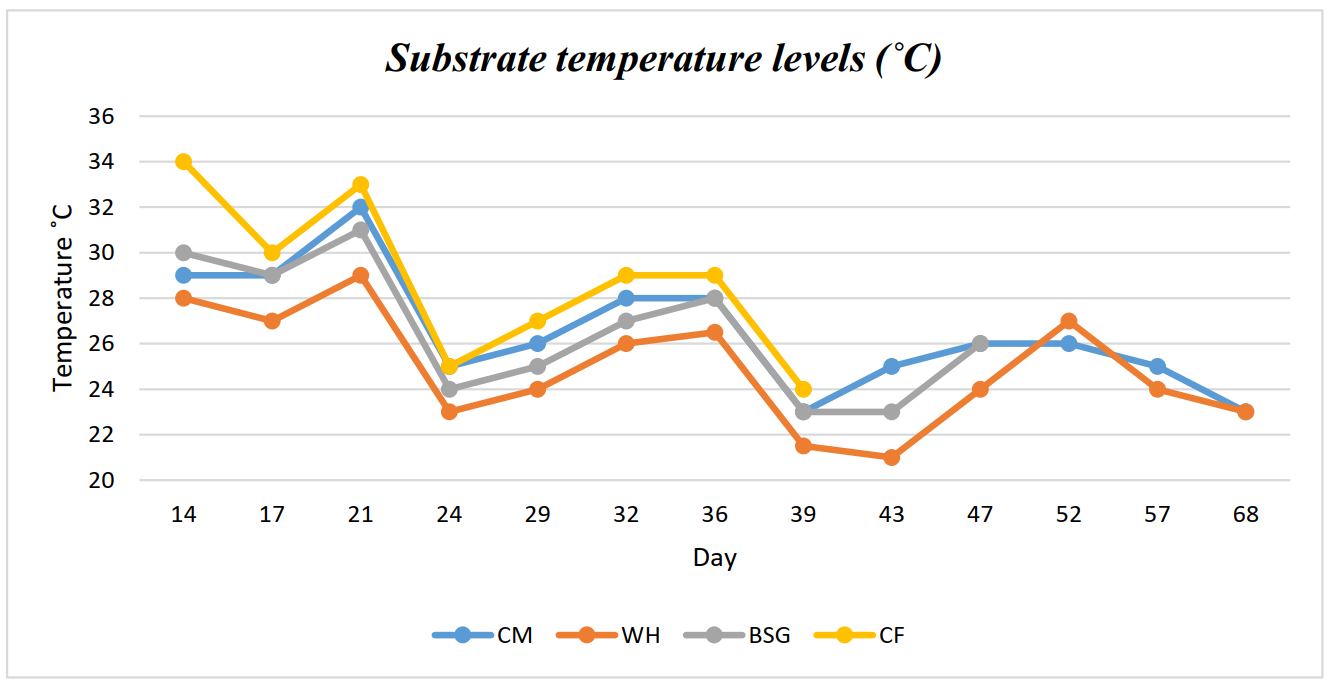


Figure 6: Substrate temperature levels

**Figure 6** above shows the temperature levels recorded for each dietary substrate throughout the larval phase. Higher temperatures were indicative of more feeding activity. The highest temperatures were recorded in the CF dietary substrate while the lowest temperatures were recorded in diet WH.

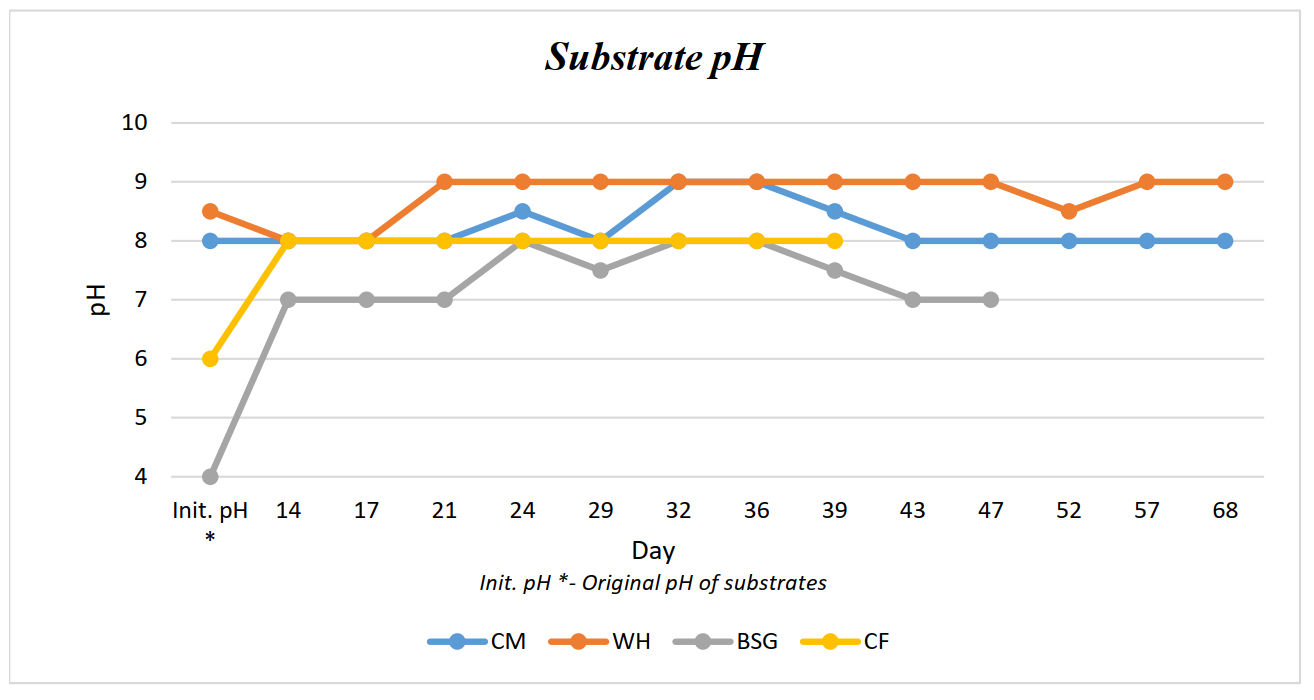


Figure 7: Substrate temperature levels

**Figure 7** above shows the pH levels recorded in the substrate in comparison with the original pH levels of each dietary substrate. The WH dietary substrate had the highest initial pH (8.5). The highest changes in pH in comparison to the initial pH were recorded in the BSG dietary substrate.

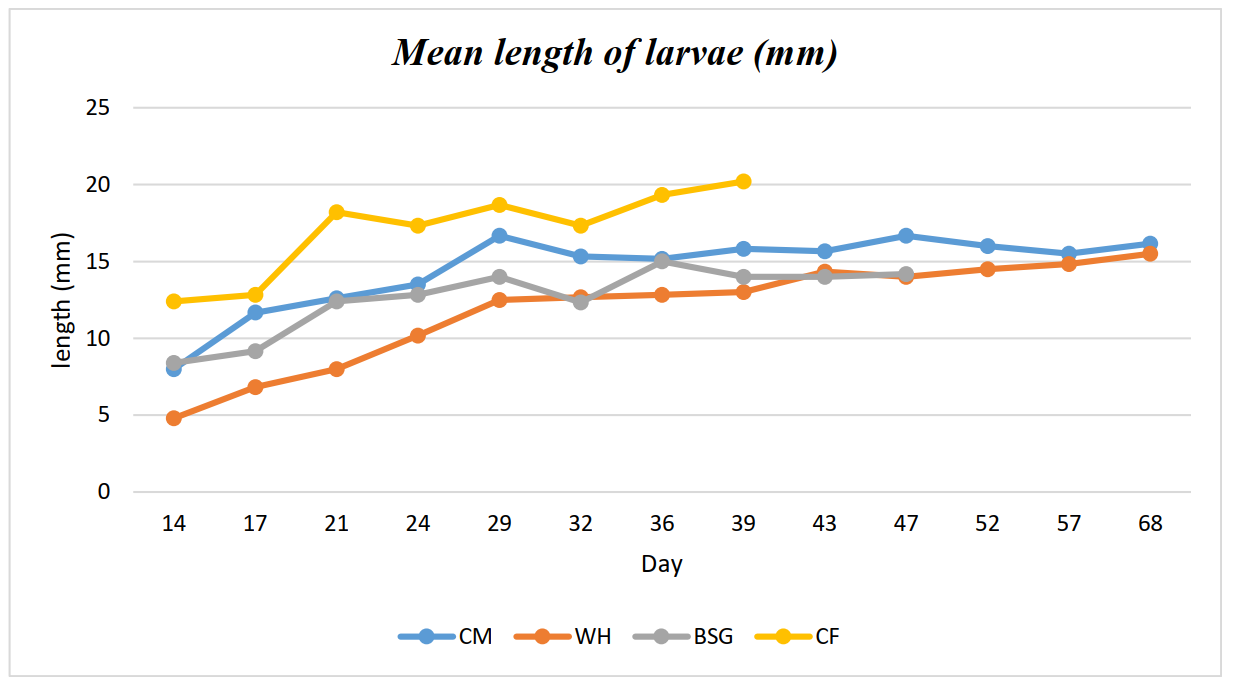


Figure 8:Mean larval lengths

From **Figure 8** above, the longest mean lengths for the BSF larvae were observed in the CF dietary substrate (the average of the mean lengths ~17.04 mm), while the lowest mean lengths were observed in the WH (the average of the mean lengths ~11.84 mm).

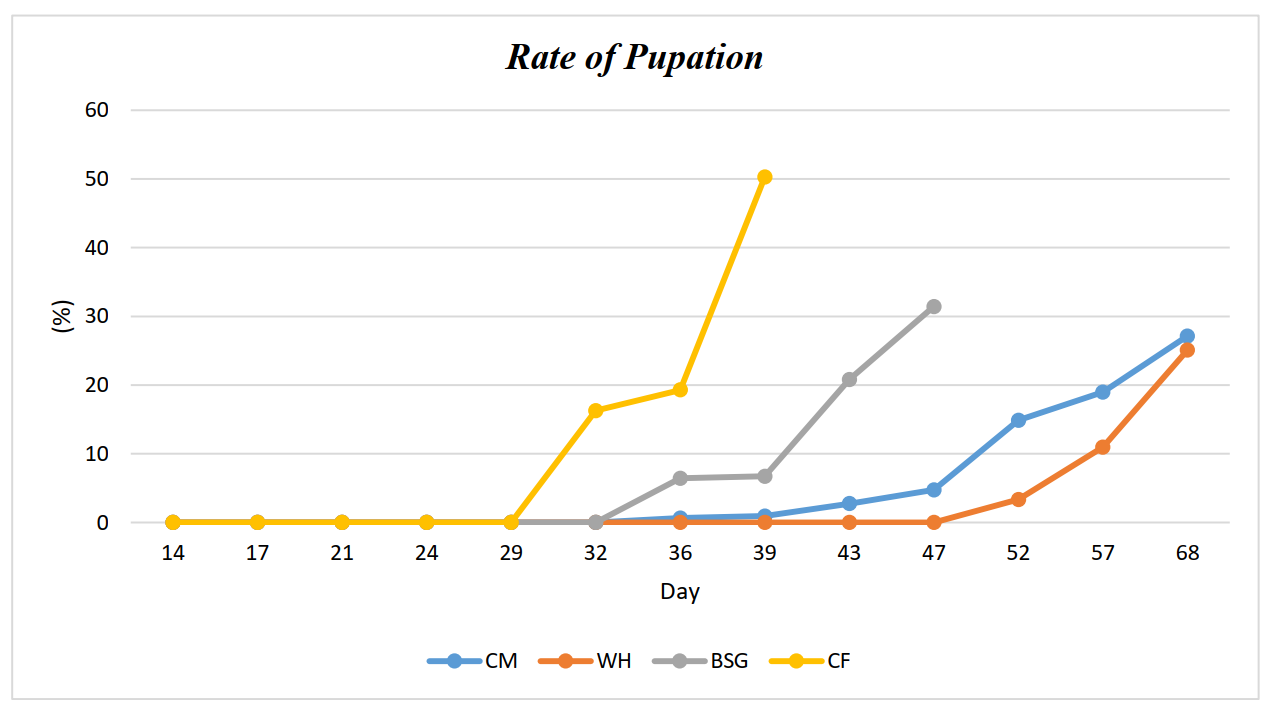


Figure 9: Pupation rates observed per dietary substrate

Like other parameters such as mean length and weight of larvae, **Figure 9** shows that faster pupation rates were recorded in the easily digestible CF diet, for which the pupation threshold for termination of the experiment (25%) was attained on Day 39. The slowest pupation rates were recorded for WH and CM diets, where the pupation thresholds for termination of the experiment were attained on day 68.

## 4.2 Feed composition analysis

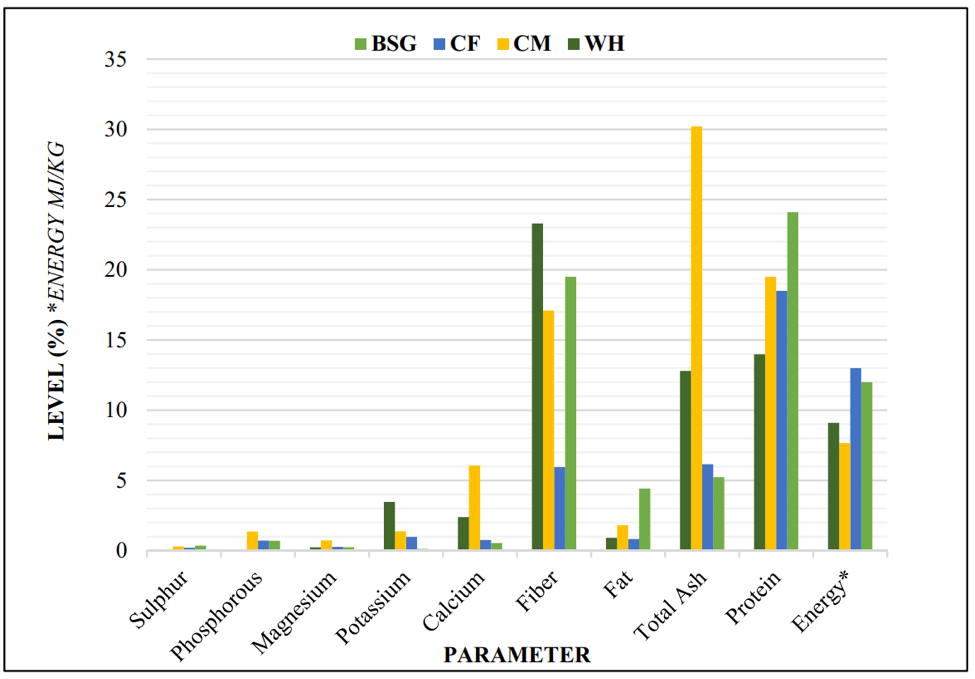
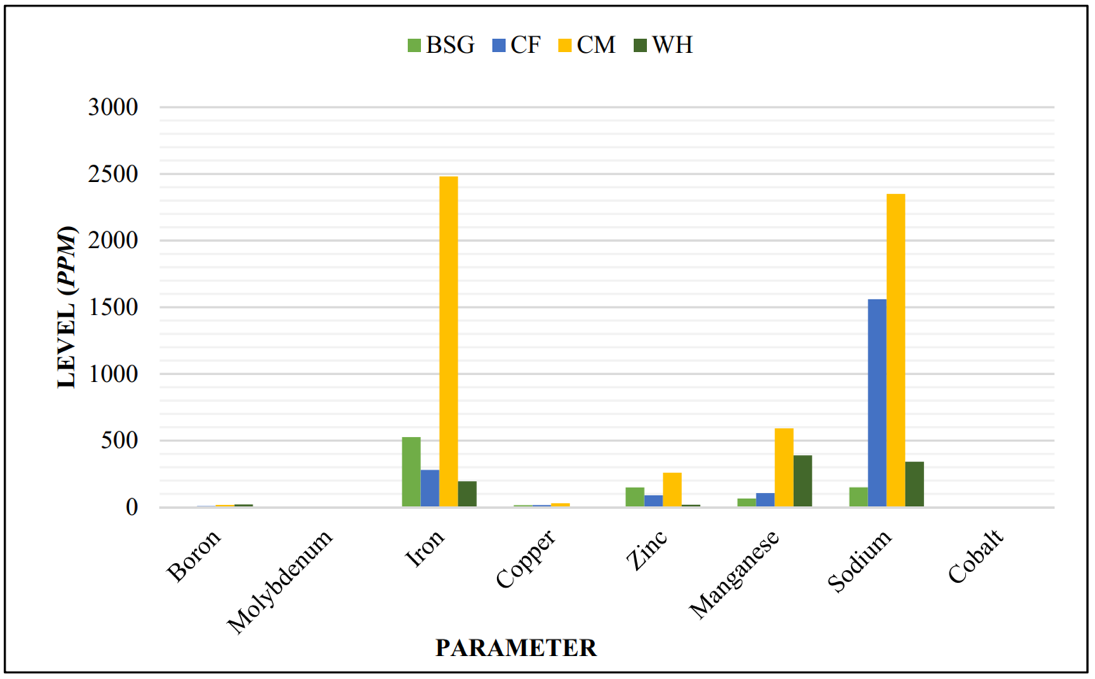
To further understand the components of each dietary substrate, feed composition analysis was conducted on the dried substrates using the Wetchem ICP-MS (Inductively coupled plasma -mass spectrometry) technique (H. C. Lee, 2018), and the concentrations of various macronutrients and micronutrients were obtained and are summarized in **Figure 10**.

Table 4. 1: ICP-MS composition analysis

The table below shows a summary of the results obtained from Wetchem, ICP-MS analyses carried out on the various dietary substrates used to feed the BSFL.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Parameter*** | ***Method*** | ***Unit*** | ***Result*** | | | |
|  |  |  | **BSG** | **CF** | **CM** | **WH** |
| *Energy* | Calculated | MJ/Kg | 12 | 13 | 7.66 | 9.1 |
| *Protein* | ISO 5983-2 | % | 24.1 | 18.5 | 19.5 | 13.98 |
| *Total Ash* | ISO 5984 | % | 5.24 | 6.16 | 30.2 | 12.8 |
| *Fat* | Gafta 3 | % | 4.43 | 0.83 | 1.82 | 0.915 |
| *Fiber* | ISO 6865 | % | 19.5 | 5.95 | 17.1 | 23.3 |
| *Dry Matter @103C Animal Feed* | ISO6496 | % | 91.4 | 89 | 90.7 | 89.3 |
| *Calcium* | CN-TM-P01 | % | 0.54 | 0.76 | 6.06 | 2.395 |
| *Potassium* | CN-TM-P01 | % | 0.12 | 0.98 | 1.38 | 3.475 |
| *Magnesium* | CN-TM-P01 | % | 0.25 | 0.27 | 0.73 | 0.24 |
| *Phosphorous* | CN-TM-P01 | % | 0.71 | 0.72 | 1.36 | 0.094 |
| *Sulphur* | CN-TM-P01 | % | 0.36 | 0.21 | 0.3 | 0.093 |
| *Boron* | CN-TM-P01 | ppm | 2.88 | 8.96 | 17.4 | 20.8 |
| *Molybdenum* | CN-TM-P01 | ppm | 1.65 | 1.1 | 2.57 | 2.66 |
| *Iron* | CN-TM-P01 | ppm | 526 | 280 | 2480 | 194 |
| *Copper* | CN-TM-P01 | ppm | 15.8 | 16.2 | 29.4 | 2.27 |
| *Zinc* | CN-TM-P01 | ppm | 149 | 89.5 | 259 | 18.6 |
| *Manganese* | CN-TM-P01 | ppm | 65.8 | 106 | 591 | 388.5 |
| *Sodium* | CN-TM-P01 | ppm | 150 | 1560 | 2350 | 341.75 |
| *Cobalt* | CN-TM-P01 | ppm | 0.17 | 0.63 | 1.59 | 0.115 |

The highest energy content levels (13 MJ/Kg) were recorded in the CF diet while the CM diet recorded the lowest (7.66 MJ/Kg). The lignocellulose-rich diets WH and BSG recorded the highest fiber contents at 23.3% and 19.5% respectively while the CF diet recorded the lowest fiber levels at 5.95%.

**a.**

**b.**

Figure 10: Comparison of macronutrients and micronutrients in the dietary substrates

**a.** shows the levels of macronutrients recorded while **b.** shows the levels of micronutrients recorded in each of the dietary substrates analyzed using Wetchem ICP-MS analyses.

The CF dietary substrate recorded the highest energy content while the CM recorded the lowest. CM also recorded the highest levels of calcium, total ash, iron, sodium, and considerable amounts of most macro and micronutrients compared to the other diets, while the BSG diet recorded the highest fat and protein levels. The WH diet recorded the highest fiber content while the CF diet recorded the lowest.

## 4.3 Gel electrophoresis

**b.**

**a.**

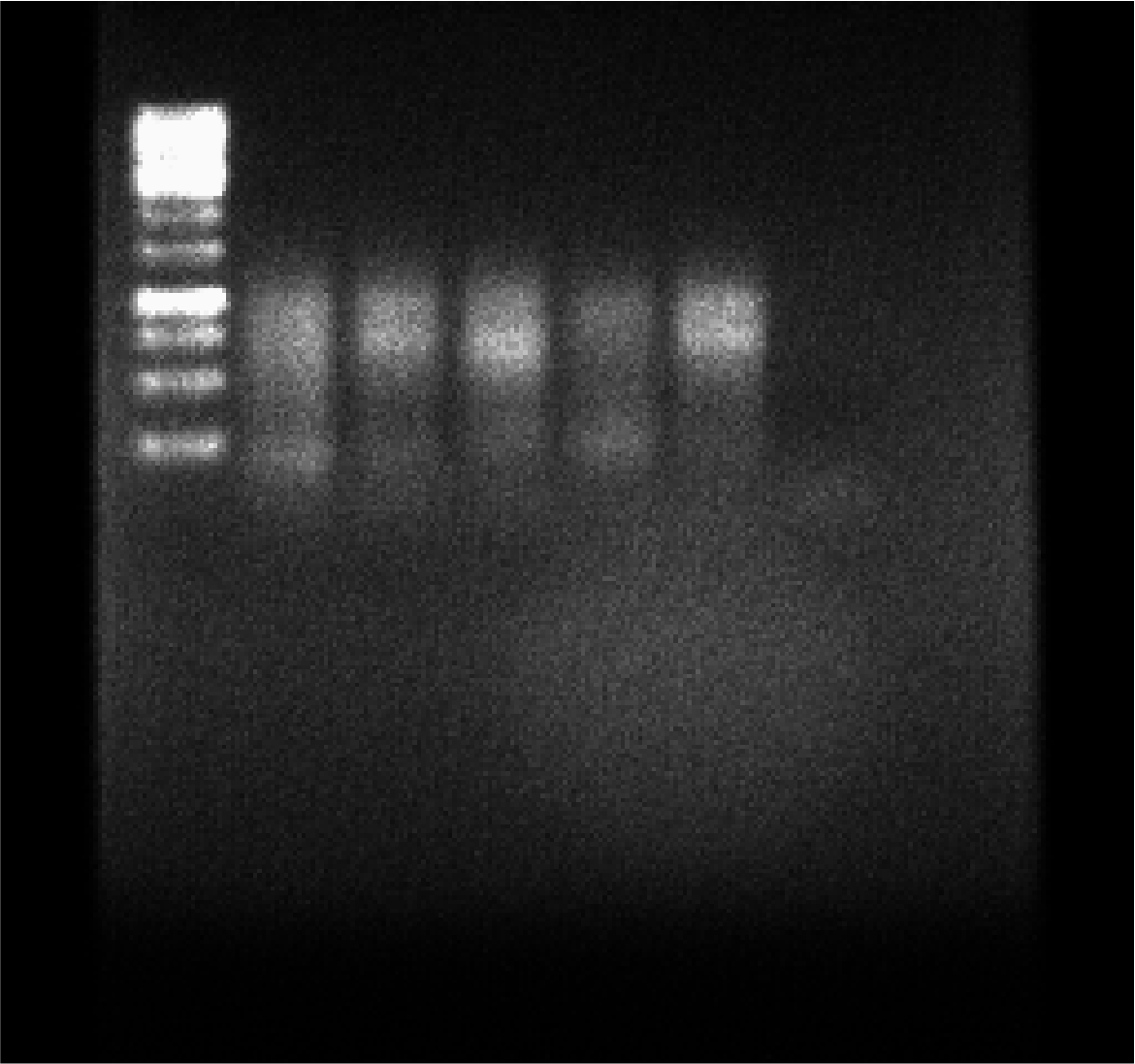
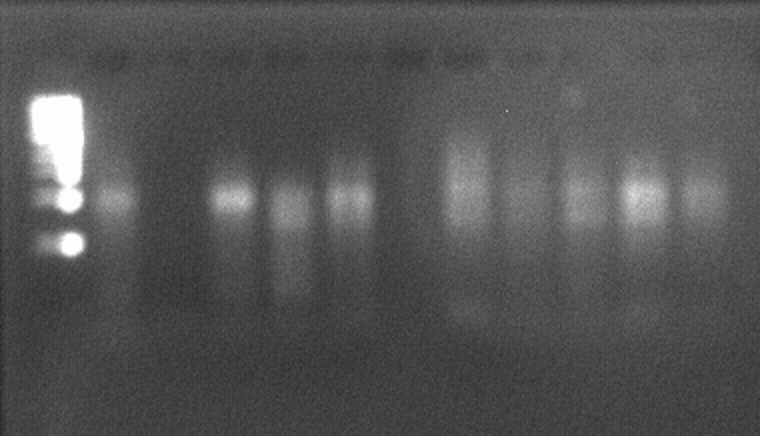
 

Figure 11: Agarose gel electrophoresis

Agarose gel electrophoresis (1.5%) was performed to show the integrity and size of the cDNA libraries from the mRNA library enrichment step using the PCR-cDNA barcoding kit (SQK-PCB109) from ONT. **a.** shows cDNA smears from sequencing run 1 while **b.** shows cDNA smears from sequencing run 2. The cDNA strands were roughly 0.5-2 kb long as observed in the above gel images using the 1kb NEB (New England Biolabs) ladder.

## 4.4 Quality statistics of basecalled reads with pycoQC

Using sequencing\_summary.txt files generated by the Guppy (v5.0.11) basecaller, the quality statistics of the basecalled reads from the two sequencing runs were obtained using pycoQC (Leger & Leonardi, 2019). Tables **4.2** and **4.3** below show a summary of these statistics.

Table 4. 2: Sequencing summary statistics (sequencing run 1)

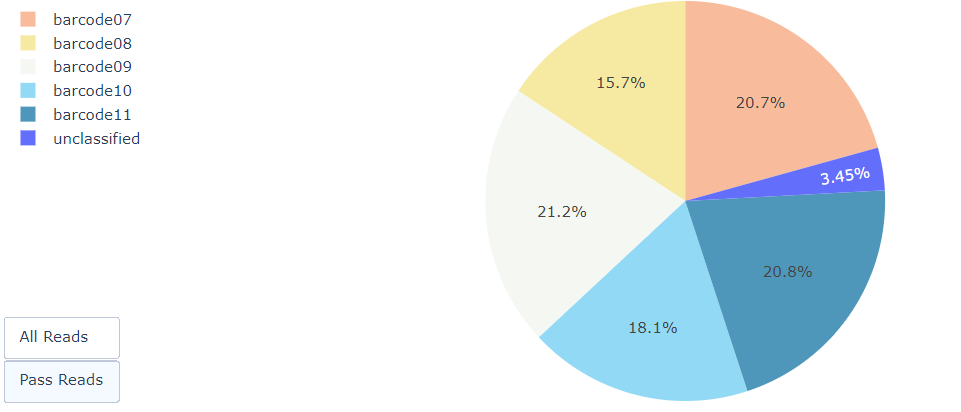
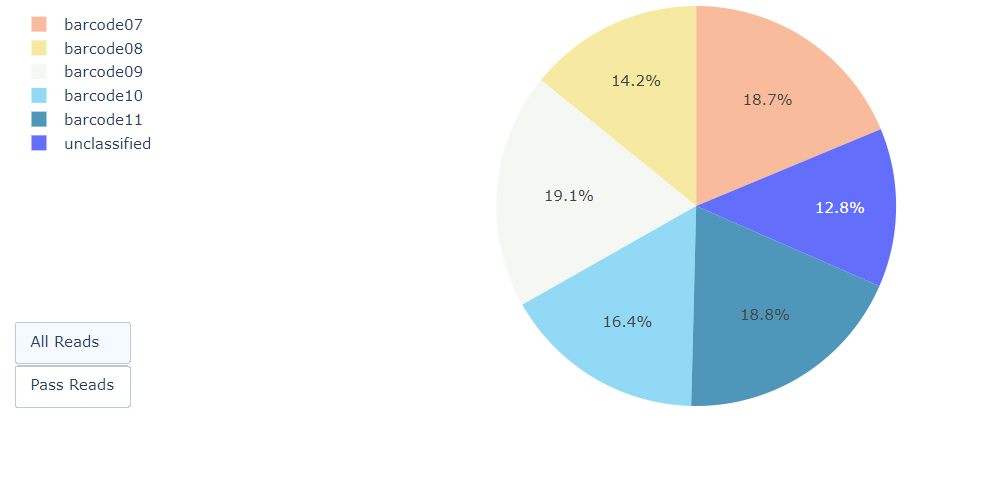
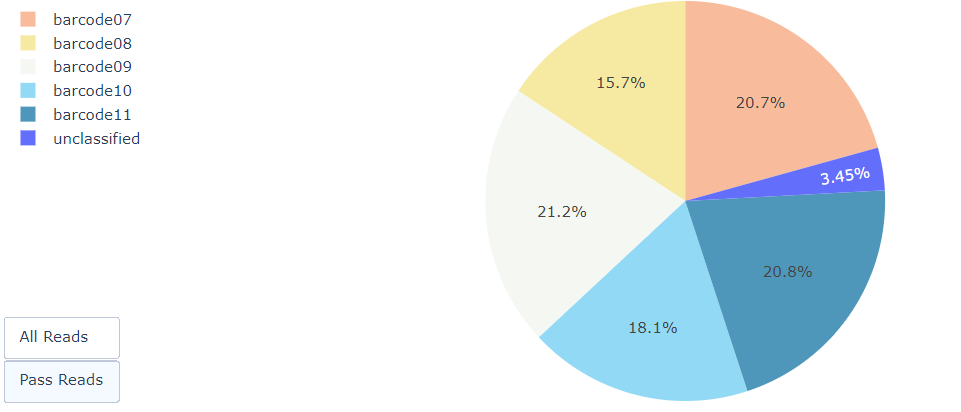
Table 4.2 shows summary statistics of the total reads, passed reads, the median read length, the active flow cell channels, and the run duration from the first sequencing run.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Barcode*** | ***Sample ID*** | ***Total reads*** | ***Pass reads*** | ***Median length (bp)*** | ***N50 Length*** | ***Median read quality*** | ***Active flow cell Channels*** | ***Run Duration (h)*** |
| *barcode07\_1* | CF1 | 1,100,574 | 1,060,876 | 542 | 802 | 10.57 | 459 | 15.87 |
| *barcode08\_1* | FM1 | 831,907 | 805,845 | 620 | 728 | 10.62 | 459 | 15.87 |
| *barcode09\_1* | BSG1 | 1,120,069 | 1,086,833 | 541 | 653 | 10.66 | 461 | 15.87 |
| *barcode10\_1* | CM1 | 963,542 | 928,013 | 390 | 761 | 10.50 | 460 | 15.87 |
| *barcode11\_1* | WH1 | 1,107,121 | 1,066,337 | 641 | 712 | 10.84 | 462 | 15.87 |
| *unclassified\_run1* |  | 754,426 | 176,677 | 508 | 756 | 8.32 | 465 | 15.87 |

Table 4. 3: Sequencing summary statistics (sequencing run 2)

Table 4.3 shows summary statistics of the total reads, passed reads, the median read length, the active flow cell channels, and the run duration from the second sequencing run.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Barcode*** | ***Sample ID*** | ***Total reads*** | ***Pass reads*** | ***Median length (bp)*** | ***N50 Length*** | ***Median Read Quality*** | ***Active flow cell Channels*** | ***Run Duration (h)*** |
| *barcode01* | CM2 | 202,066 | 189,220 | 683 | 763 | 10.33 | 230 | 148.71 |
| *barcode02* | CF2 | 66,298 | 61,563 | 610 | 828 | 10.20 | 225 | 148.71 |
| *barcode03* | FM2 | 114,106 | 107,090 | 682 | 854 | 10.27 | 223 | 148.71 |
| *barcode04* | WH2 | 80,959 | 75,004 | 664 | 779 | 10.18 | 223 | 148.71 |
| *barcode06* | BSG2 | 212,881 | 199,615 | 684 | 887 | 10.21 | 231 | 148.71 |
| *barcode07* | CF3 | 198,416 | 186,509 | 674 | 871 | 10.27 | 223 | 148.71 |
| *barcode08* | CM3 | 293,572 | 276,123 | 748 | 884 | 10.17 | 228 | 148.71 |
| *barcode09* | FM3 | 344,353 | 321,919 | 455 | 683 | 9.92 | 229 | 148.71 |
| *barcode10* | WH3 | 203,430 | 190,193 | 700 | 866 | 10.35 | 230 | 148.71 |
| *barcode11* | BSG3 | 123,542 | 114,942 | 779 | 914 | 10.36 | 225 | 148.70 |
| *barcode12* | CF4 | 172,584 | 161,575 | 672 | 869 | 10.26 | 228 | 148.71 |
| *unclassified\_run2* |  | 1,830,961 | 92,060 | 608 | 858 | 8.02 | 502 | 181.28 |

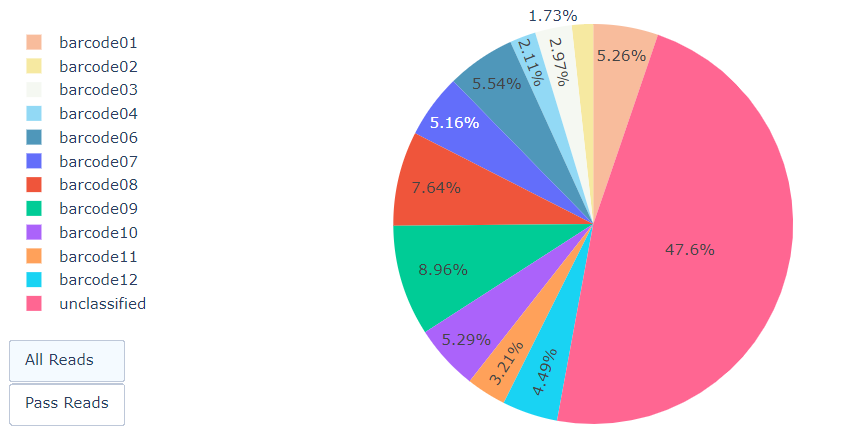
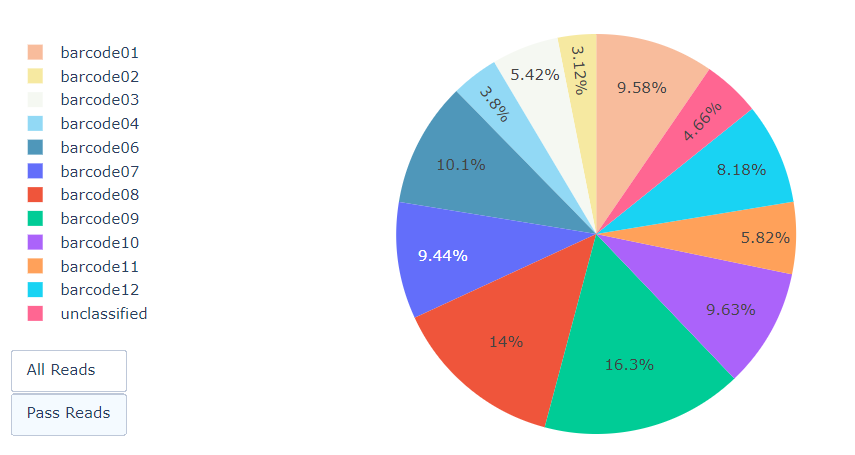
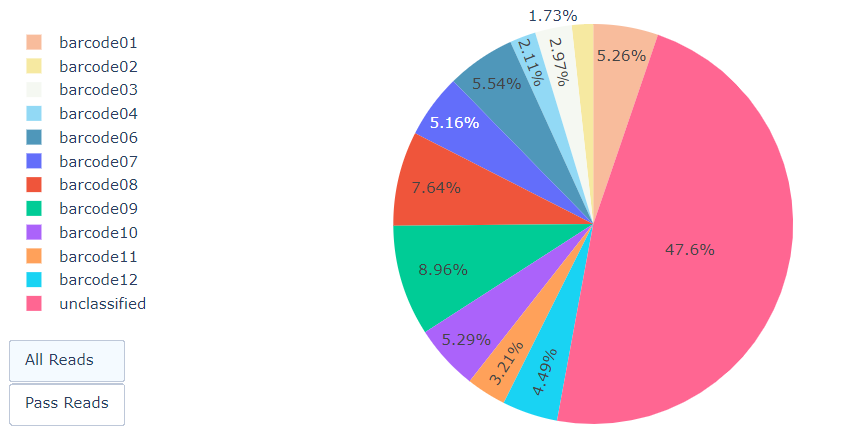


**a.**

**b.**

Figure 12: Sequencing quality statistics with pycoQC

Pie charts showing the distribution of barcodes from sequencing run 1 after basecalling and demultiplexing with Guppy (v5.0.11). **a.** shows the distribution of barcodes in all the basecalled reads while **b.** shows the distribution of barcodes in the reads with a quality score >7 (passed reads).



**a.**

**b.**

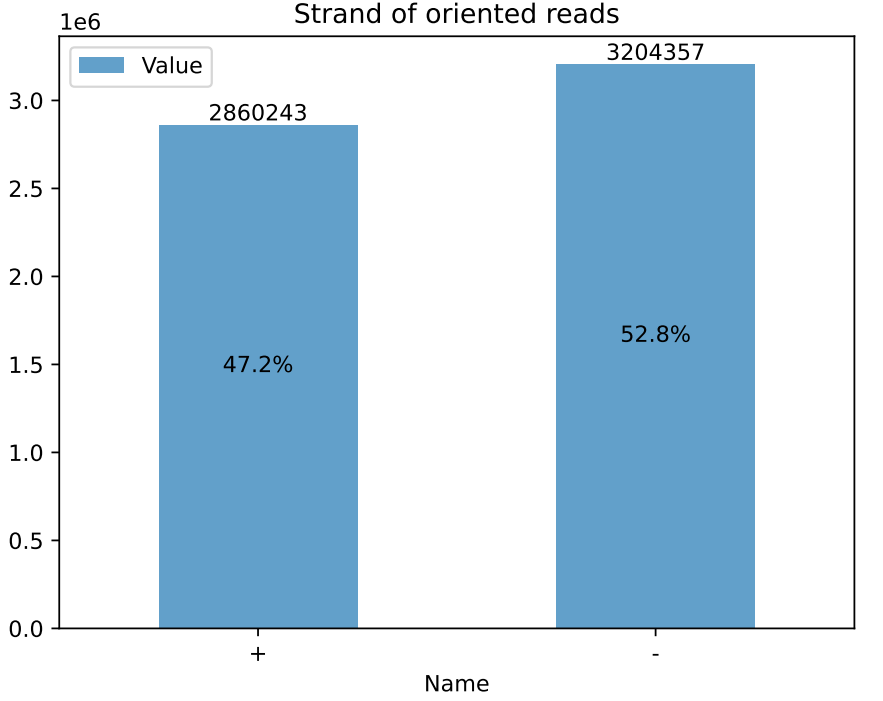
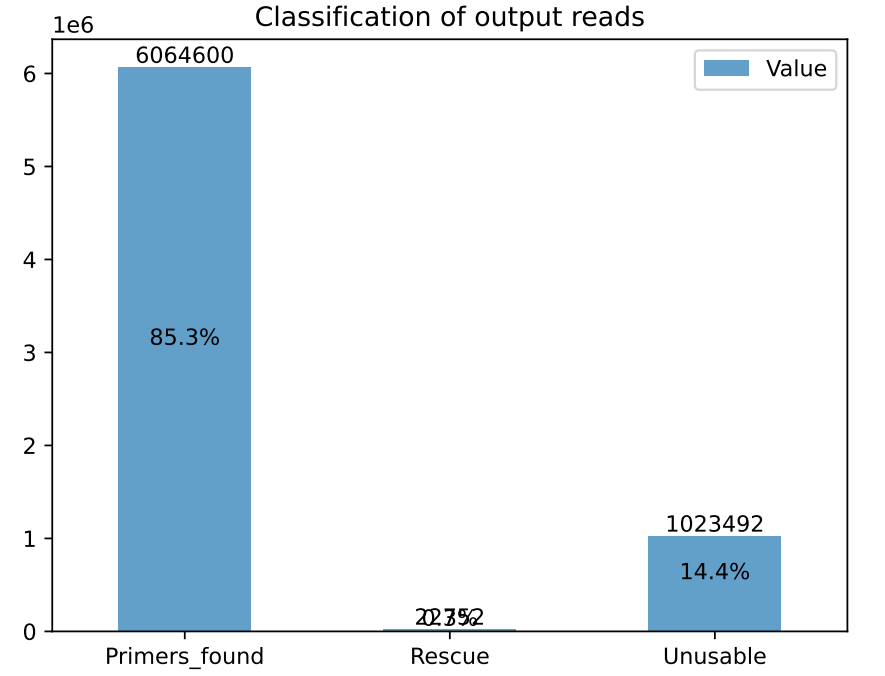
Figure 13: Demultiplexing with Guppy

Pie charts showing the distribution of barcodes from sequencing run 2 after basecalling and demultiplexing with Guppy (v5.0.11). **a.** shows the distribution of barcodes in all the basecalled reads while **b.** shows the distribution of barcodes in the reads with a quality score >7 (passed reads).

Unclassified reads are regarded as the reads that were not assigned to any of the barcodes provided in the library preparation kit (Leidenfrost et al., 2020). Apart from being unassigned to the provided barcodes, reads can be termed as unclassified if they are fused and possess more than one barcode. Such reads can be ‘rescued’ by using algorithms that separate such fused reads (*Nanoporetech/Pychopper: A tool to identify, orient, trim and rescue full-length cDNA reads*, 2018). From all the reads sequenced, sequencing run 1 recorded 12.8% of unclassified reads while sequencing run 2 recorded 47.6% of unclassified reads. From the reads that passed the quality cut-off, sequencing run 1 recorded 3.45% of unclassified reads while sequencing run 2 recorded 4.66% of unclassified reads. This was attributed to the different flow cell chemistries used in the two sequencing runs. Flow cell R10.4 was used for sequencing run 2 while flow cell R9.4.1 was used for sequencing run 1. The flowcell R9.4.1 flow cell is well optimized for use with the MinION MK1B platform and basecalling with Guppy hence its better performance. The R10 sequencing chemistry is well optimized for sequencing homopolymer regions recording up to 99.5% accuracy (*R10.3: the newest nanopore for high accuracy nanopore sequencing*, 2020). However, the computing infrastructure has not been fully optimized for this flow cell type and is set to be wholly integrated into subsequent releases of ONT sequencing platforms and software. The best performance for the R10.4 flow cell is observed by tagging the templates being sequenced with Unique Molecular Identifiers (UMIs) for high consensus accuracy (Karst et al., 2021).

## 4.5 Trimming, orienting, and defusing cDNA reads with Pychopper

After running Pychopper (v2.2.0) to trim barcode primers, orient + and – strands, and defuse merged reads, about 0.3% of the total reads were rescued, while 85.3% were classified and 14.4% were unclassified. The results are further summarized in **Figure 14** below.



**a..**

**b.**

**22752**

**0.3%**

Figure 14: Classification and orientation of cDNA reads with Pychopper

**a.** shows a summary of the read classification statistics obtained from Pychopper (v2.2.0) for both sequencing runs. More than 85% of the total reads were classified against the primers provided by the library preparation kit. Also, in **b.** 47.2% of the reads were found to be positively (+) oriented strands while 52.8% were negatively (-) oriented strands. The ideal orientation ratio is often (+):(-) 1:1 (*nanoporetech/pychopper: A tool to identify, orient, trim and rescue full-length cDNA reads*, 2018).

Table 4.4: Classification and orientation of sequence reads

Table 4.4 below presents the classification of reads into their respective barcodes and the orientation statistics for each of the sequenced samples.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Barcode*** | ***Sample ID*** | ***Classified reads***  ***(%)*** | ***Unclassified reads***  ***(%)*** | ***Rescued reads***  ***(%)*** | ***Orientation (%)*** | |
| **+** | **-** |
| *barcode07\_1* | CF1 | 903,473 (85.2) | 153,373 (14.8) | 60 (0) | 389,347 (43.1) | 514,126 (56.9) |
| *barcode08\_1* | FM1 | 691,444 (85.8) | 114,370 (14.2) | 62 (0) | 317,178 (45.9) | 374,266 (54.1) |
| *barcode09\_1* | BSG1 | 953,056 (87.7) | 133,750 (12.3) | 54 (0) | 427421 (45.9) | 515,635 (54.1) |
| *barcode10\_1* | CM1 | 803,584 (86.6) | 124,406 (13.4) | 46 (0) | 368,384 (45.8) | 435,200 (54.2) |
| *barcode11\_1* | WH1 | 884,170 (82.9) | 182,151 (17.1) | 32 (0) | 365,074 (41.3) | 519,096 (58.7) |
| *barcode01* | CM2 | 170,876 (90.3) | 18,343 (9.7) | 2 (0) | 100,894 (59) | 69,982 (41) |
| *barcode02* | CF2 | 56,024 (91) | 5,536 (9) | 6 (0) | 28,455 (50.8) | 27569 (49.2) |
| *barcode03* | FM2 | 96,446 (90.1) | 10,643 (9.9) | 2 (0) | 49,873 (51.7) | 46,573 (48.3) |
| *barcode04* | WH2 | 67,176 (89.6) | 7,822 (10.4) | 12 (0) | 33,347 (49.6) | 33,829 (50.4) |
| *barcode06* | BSG2 | 181,499 (90.9) | 18,109 (9.1) | 14 (0) | 102,190 (56.3) | 79,309 (43.7) |
| *barcode07* | CF3 | 170,007 (91.2) | 16,499 (8.8) | 6 (0) | 92,728 (54.5) | 77,279 (45.5) |
| *barcode08* | CM3 | 251,968 (91.2) | 24,140 (8.7) | 30 (0.1) | 133,242 (52.9) | 118,726 (47.1) |
| *barcode09* | FM3 | 296,918 (92.2) | 24,991 (7.8) | 20 (0) | 170,296 (57.4) | 126,622 (42.6) |
| *barcode10* | WH3 | 170,605 (89.7) | 19,584 (10.3) | 8 (0) | 85,804 (50.3) | 84,801 (49.7) |
| *barcode11* | BSG3 | 101,893 (88.6) | 13,077 (11.4) | 4 (0) | 51,286 (50.3) | 50,577 (49.7) |
| *barcode12* | CF4 | 144,949 (89.7) | 16,625 (10.3) | 2 (0) | 75,176 (51.9) | 69,773 (58.7) |

**Key**

|  |
| --- |
|  |
|  |

Sequencing run 1

Sequencing run 2

## 4.6 rRNA filtering with SortMeRNA

The SortMeRNA program (v4.3.3) (Kopylova et al., 2012) used to filter, map, and perform OTU picking in metagenomic and metatranscriptomic sequences, was used to perform ribodepletion to filter out rRNA sequences. The statistics from this step are summarized in **Table 4.5** below and were visualized using MultiQC (v1.11) (Ewels et al., 2016) shown in **Figure 15**.

Table 4.5: Percentage of rRNA reads per sample

**Table 4.5** shows the rRNA statistics from the ribodepletion step with SortMeRNA. Residual rRNA from the metatranscriptomes ranged from 25.6% - 44%.

|  |  |  |
| --- | --- | --- |
| ***Percentage of rRNA reads per sample*** | | |
| ***Barcode*** | **Sample\_ID** | **% rRNA** |
| ***barcode07\_1*** | **CF1** | 30.40% |
| ***barcode08\_1*** | **FM1** | 28.60% |
| ***barcode09\_1*** | **BSG1** | 25.60% |
| ***barcode10\_1*** | **CM1** | 30.30% |
| ***barcode11\_1*** | **WH1** | 30.90% |
| ***barcode01*** | **CM2** | 32.70% |
| ***barcode02*** | **CF2** | 44.00% |
| ***barcode03*** | **FM2** | 40.30% |
| ***barcode04*** | **WH2** | 43.40% |
| ***barcode06*** | **BSG2** | 27.90% |
| ***barcode07*** | **CF3** | 39.10% |
| ***barcode08*** | **CM3** | 32.10% |
| ***barcode09*** | **FM3** | 36.10% |
| ***barcode10*** | **WH3** | 37.20% |
| ***barcode11*** | **BSG3** | 35.80% |
| ***barcode12*** | **CF4** | 36.20% |

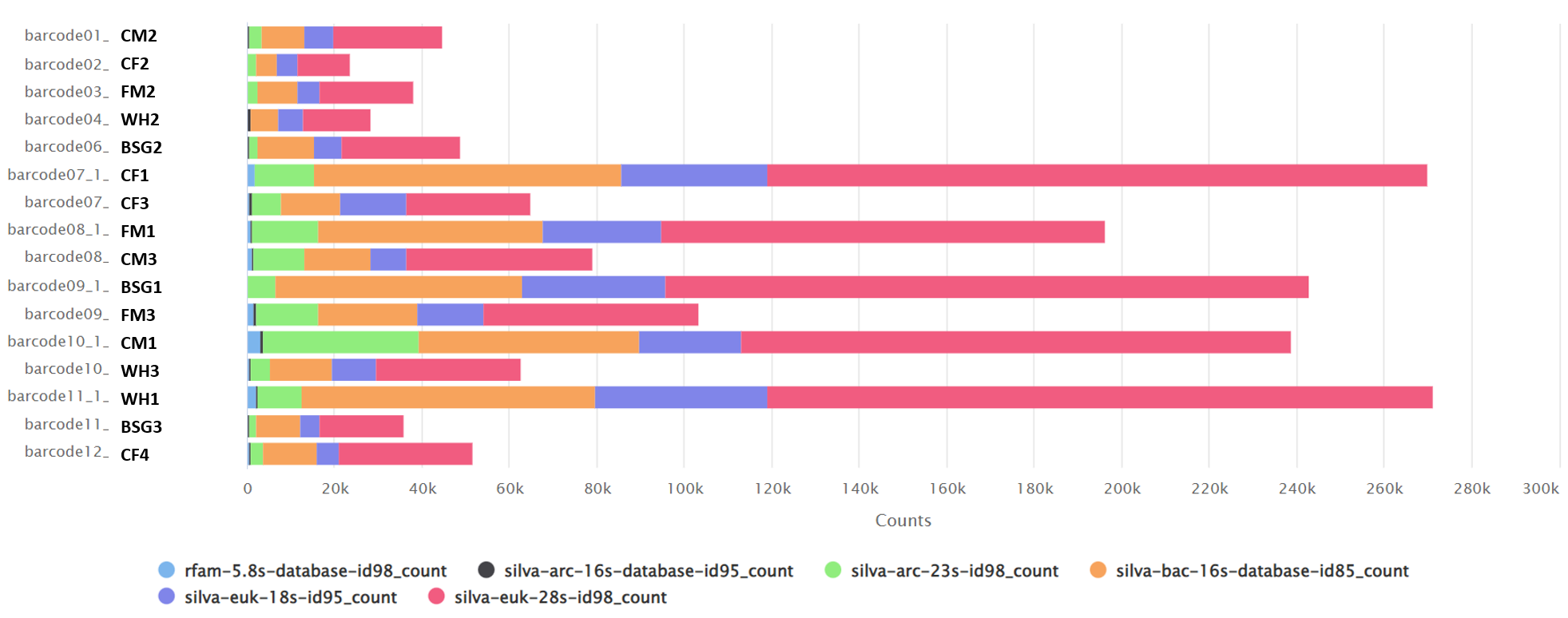


Figure 15: Ribodepletion statistics using MultiQC (v1.11)

**Figure 15** above shows the number and type of rRNA sequence reads filtered from each sample. The highest numbers of rRNA reads were 28S rRNA reads from the silva-euk-28s-id98\_count database while the lowest type were the 16S rRNA reads from the silva-arc-16s-database-id98\_count database.

## 4.7 Error correction of IsONclust clustered cDNA reads with IsONcorrect

From the IsONclust greedy clustering algorithm, longer sequences with better quality scores were prioritized and ordered before the shorter reads with lower scores, and one consensus read was maintained as a representative of each cluster (Sahlin & Medvedev, 2020). From error correction with IsONcorrect - evaluated using reads that mapped to the BSF reference genome for both the clustered, corrected reads and unclustered, uncorrected reads, we found that there were differences in mean coverage, mean depth, mapping quality, and mapping percentages with significant differences being observed in mapping percentages using the paired t-test as shown in **Table 4.6**, and **Figure 16**.

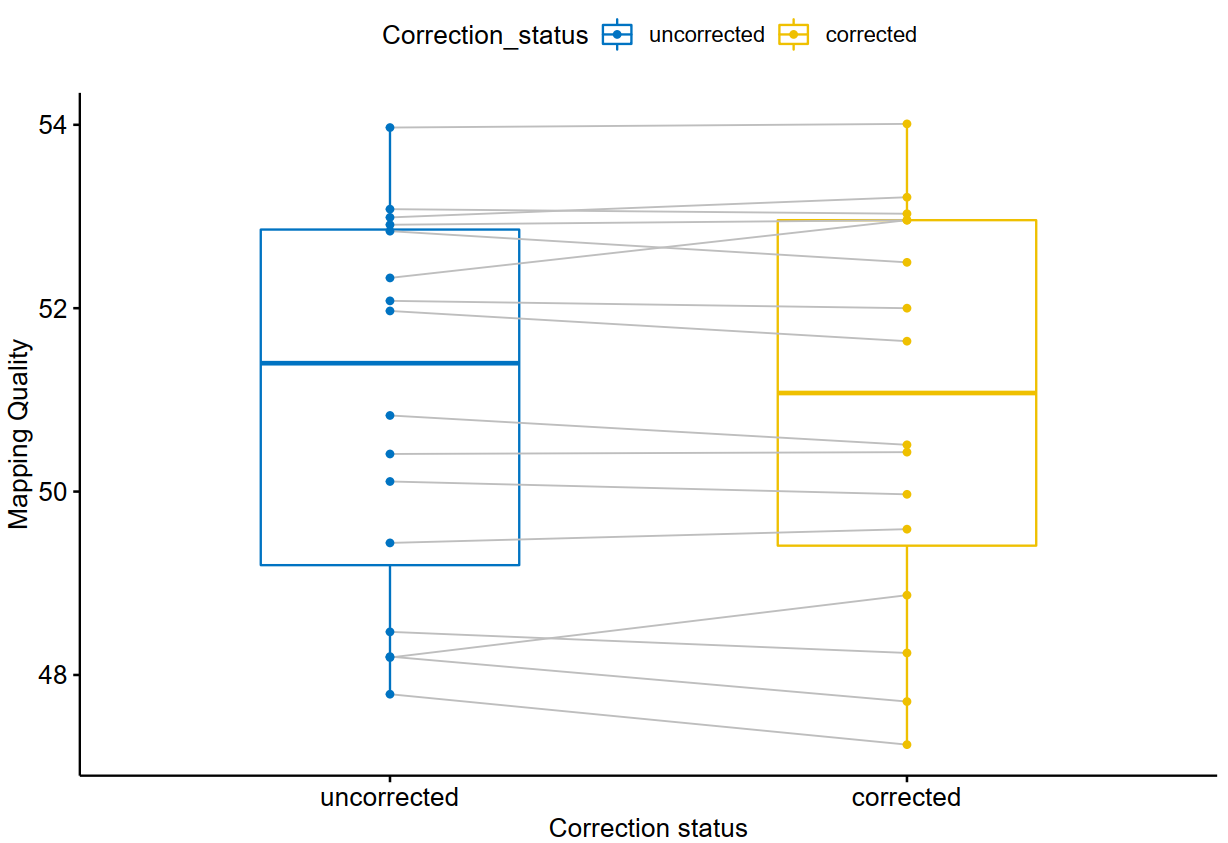
Table 4. 6: Evaluation of the read error-correction efficacy

The table below shows the comparison between reads subjected to de novo clustering with isONclust (Sahlin & Medvedev, 2020) and correction with isONcorrect (Sahlin et al., 2021) and uncorrected reads that were neither clustered nor corrected. There was a slight variation in the mean coverage, mean depth, and mapping quality between the corrected and uncorrected reads. However, all the corrected reads recorded higher (%) of reads that mapped to the BSF genome as compared to the uncorrected reads.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Barcode\_ID*** | ***Sample\_ID*** | ***Mean coverage*** | | ***Mean depth*** | | ***Mapping quality*** | | ***(%) of mapped reads*** | |
|  |  | *cor.* | *uncor.* | *cor.* | *uncor.* | *cor.* | *uncor.* | *cor.* | *uncor.* |
| ***barcode01*** | CF1 | 0.262 | 0.264 | 0.043 | 0.052 | 52.96 | 52.33 | 94.55 | 93.66 |
| ***barcode02*** | FM1 | 0.373 | 0.373 | 0.015 | 0.015 | 52.5 | 52.84 | 90.93 | 86.75 |
| ***barcode03*** | BSG1 | 0.268 | 0.269 | 0.025 | 0.025 | 50.43 | 50.41 | 92.03 | 89.01 |
| ***barcode04*** | CM1 | 0.3 | 0.3 | 0.018 | 0.018 | 54.01 | 53.97 | 95.45 | 92.24 |
| ***barcode06*** | WH1 | 0.559 | 0.561 | 0.058 | 0.057 | 51.64 | 51.97 | 89.62 | 85.84 |
| ***barcode07*** | CM2 | 0.389 | 0.389 | 0.047 | 0.047 | 53.21 | 52.99 | 89.35 | 86.26 |
| ***barcode08*** | CF2 | 0.351 | 0.353 | 0.078 | 0.077 | 50.51 | 50.83 | 92.2 | 88.4 |
| ***barcode09*** | FM2 | 0.312 | 0.314 | 0.056 | 0.055 | 47.24 | 47.79 | 82.5 | 76.16 |
| ***barcode10*** | WH2 | 0.398 | 0.4 | 0.052 | 0.052 | 52 | 52.08 | 94.76 | 92.48 |
| ***barcode11*** | BSG2 | 0.43 | 0.43 | 0.035 | 0.034 | 53.03 | 53.08 | 95.09 | 92.75 |
| ***barcode12*** | CF3 | 0.406 | 0.407 | 0.0402 | 0.04 | 52.96 | 52.91 | 91.73 | 88.7 |
| ***barcode07\_1*** | CM3 | 0.717 | 0.717 | 0.203 | 0.199 | 49.97 | 50.11 | 84.86 | 80.88 |
| ***barcode08\_1*** | FM3 | 0.672 | 0.675 | 0.174 | 0.171 | 48.24 | 48.47 | 89.57 | 86.29 |
| ***barcode09\_1*** | WH3 | 1.075 | 1.077 | 0.229 | 0.224 | 48.87 | 48.19 | 89.15 | 85.18 |
| ***barcode10\_1*** | BSG3 | 1.09 | 1.094 | 0.162 | 0.158 | 47.71 | 48.2 | 83.3 | 78.06 |
| ***barcode11\_1*** | CF4 | 0.572 | 0.574 | 0.223 | 0.22 | 49.59 | 49.44 | 92.42 | 90.17 |

*\*cor.*- corrected

*\*uncor*.-uncorrected

**b.**

**a.**

Figure 16: Differences in mapping percentages and mapping quality of reads mapped to the BSF reference between the uncorrected and corrected reads

From **Figure 16a.** above, the mapping percentages were higher for the corrected reads as compared to the uncorrected reads after analysis with the paired t-test. The observed p-value 2×10e8 (α value - 0.05) showed that there was a statistically significant difference between the mapping percentages of corrected and uncorrected sequence reads. From **Figure 16b.** the observed mapping quality of reads aligned to the reference pre-correction was not statistically significant compared to the mapping quality observed post-correction. The p-value (0.6) was higher than the α value (0.05).

## 4.8 Raw read counts

The total number of corrected reads from each sample that failed to map to the host genome (unmapped reads) during the alignment step, were required to perform differential statistical analyses in R software. These statistics are shown in (**Table 4.7)**.

Table 4. 7:Total number of unmapped reads per metatranscriptome

From **Table 4.7** below, sample CF1 recorded the highest number of unmapped reads (149,255) while sample WH2 recorded the lowest number of unmapped reads (2,716). The unmapped reads ranged from 4.04% (WH2) to 18.53% (FM3) of the total classified reads.

|  |  |  |
| --- | --- | --- |
| ***Sample\_ID*** | ***Total classified reads*** | ***No. of unmapped reads (%)*** |
| *CF1* | 903,473 | 149,255 (16.52) |
| *CF2* | 56,024 | 3,747 (6.69) |
| *CF3* | 170,007 | 18,073 (10.63) |
| *CF4* | 144,949 | 12,021 (8.29) |
| *BSG1* | 953,056 | 105,398 (11.06) |
| *BSG2* | 181,499 | 21,247 (11.71) |
| *BSG3* | 101,893 | 5,567 (5.46) |
| *CM1* | 803,584 | 142,086 (17.68) |
| *CM2* | 170,876 | 8,746 (5.12) |
| *CM3* | 251,968 | 25,897 (10.28) |
| *FM1* | 691,444 | 92,612 (13.39) |
| *FM2* | 96,446 | 7,765 (8.05) |
| *FM3* | 296,918 | 55,030 (18.53) |
| *WH1* | 884,170 | 87,292 (9.87) |
| *WH2* | 67,176 | 2,716 (4.04) |
| *WH3* | 170,605 | 9,921 (5.82) |

The subsequent annotation steps were performed on these unmapped reads, while statistical analyses steps in R required these read counts to perform library normalization and differential statistics between the sample metatranscriptomes.

## 4.9 Annotation and aggregation of with DIAMOND

The DIAMOND aligner (Buchfink et al., 2021) was used to perform both organism and functional annotation with the RefSeq database (Tatusova et al., 2014) and the SEED subsystems hierarchical database (Overbeek et al., 2014) for the different sample metatranscriptomes using the DIAMOND BLASTx module. The results of the BLASTx search were tabular-separated (.tsv) format files composed of 12 fields containing information such as *Query accession, Target accession, Sequence identity, Length, E-value*, among others. This file format corresponds to the format generated by the BLAST command line module using the option -outfmt 6 (Buchfink et al., 2021).

Once aggregation was performed on each of the matched sequences that had corresponding hits in the reference databases, the output files were tab-separated (.tsv) files containing 3 columns namely percentages of each entry compared to the total reads, the respective read counts, and the respective annotation (organism or function).

To get a clearer picture of the most abundant samples in every dietary substrate, all the samples belonging to one dietary substrate were pooled together, and the 4 most abundant genera and their species in each dietary substrate exceeding an abundance of >0.1% were selected and are shown in **Table 4.8** below.

Table 4. 8: Most abundant organisms per dietary substrate

This table shows a summary of the 4 most abundant genera and their species exceeding 0.1% abundance in each sample. In the control sample, CF, *Sphingobacterium sp., Flavobacterium sp., Pedobacter sp.,* and *Myroides sp.* were the most abundant genera. In sample BSG, *Sphingobacterium sp., Bacteroides sp., Dysgonomonas sp.,* and *Prevotella sp*. were most abundant. In sample CM, *Sphingobacterium sp., Clostridium sp., Bacillus sp.,* and *Chryseobacterium sp.* were found to be the most abundant genera. In sample FM, *Sphingobacterium sp., Pedobacter sp., Flavobacterium sp.,* and *Chryseobacterium sp.* were found to be the most abundant genera while in the WH sample, *Gilliamella sp., Dysgonomonas sp., Bacteroides sp.,* and *Parabacteroides sp*. were most abundant.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Diet*** | ***Organism (>0.1% Abundance)*** | ***Raw counts*** | ***(%)*** |
| ***BSG*** | *Sphingobacterium sp.* | 723 | 4.529 |
|  | *Bacteroides sp.* | 283 | 1.773 |
|  | *Dysgonomonas sp.* | 242 | 1.516 |
|  | *Prevotella sp.* | 220 | 1.377 |
|  | *Sphingobacterium spiritivorum* | 160 | 1.002 |
|  | *Dysgonomonas gadei* | 144 | 0.901 |
|  | *Dysgonomonas mossii* | 136 | 0.851 |
|  | *Bacteroidales* | 125 | 0.782 |
|  | *Sphingobacterium nematocida* | 98 | 0.613 |
|  | *Sphingobacteriaceae bacterium* | 97 | 0.607 |
|  | *Sphingobacterium deserti* | 93 | 0.583 |
|  | *Sphingobacterium thalpophilum* | 91 | 0.57 |
|  | *Sphingobacterium psychroaquaticum* | 90 | 0.564 |
|  | *Sphingobacterium paucimobilis* | 88 | 0.551 |
|  | *Dysgonomonas capnocytophagoides* | 82 | 0.514 |
|  | *Dysgonomonas macrotermitis* | 79 | 0.495 |
|  | *Bacteroides fragilis* | 56 | 0.351 |
|  | *Bacteroidales bacterium* | 44 | 0.275 |
|  | *Bacteroides thetaiotaomicron* | 41 | 0.257 |
|  | *Prevotella ruminicola* | 25 | 0.157 |
|  | *Bacteroides coprocola* | 22 | 0.138 |
|  | *Bacteroides cellulosilyticus* | 21 | 0.132 |
|  | *Bacteroides coprophilus* | 21 | 0.132 |
|  | *Bacteroides luti* | 21 | 0.132 |
|  | *Bacteroides clarus* | 20 | 0.125 |
|  | *Bacteroides pyogenes* | 19 | 0.12 |
|  | *Bacteroides finegoldii* | 19 | 0.12 |
|  | *Bacteroides xylanisolvens* | 18 | 0.112 |
|  | *Bacteroides uniformis* | 17 | 0.106 |
|  | *Bacteroides coprosuis* | 17 | 0.106 |
|  | *Bacteroides propionicifaciens* | 17 | 0.106 |
|  | *Bacteroides vulgatus* | 16 | 0.1 |
|  | *Bacteroides eggerthii* | 16 | 0.1 |
| ***CF*** | *Sphingobacterium sp.* | 13250 | 10.14 |
|  | *Flavobacterium sp.* | 3695 | 2.828 |
|  | *Pedobacter sp.* | 2758 | 2.111 |
|  | *Sphingobacterium spiritivorum* | 2348 | 1.797 |
|  | *Myroides sp.* | 2094 | 1.603 |
|  | *Sphingobacterium deserti* | 1653 | 1.265 |
|  | *Myroides oratus* | 1644 | 1.259 |
|  | *Sphingobacterium psychroaquaticum* | 1560 | 1.194 |
|  | *Sphingobacterium nematocida* | 1558 | 1.193 |
|  | *Sphingobacteriaceae bacterium* | 1521 | 1.164 |
|  | *Sphingobacterium paucimobilis* | 1447 | 1.108 |
|  | *Sphingobacterium thalpophilum* | 1423 | 1.089 |
|  | *Myroides odoratimimus* | 896 | 0.686 |
|  | *Myroides injenensis* | 837 | 0.641 |
|  | *Flavobacterium columnare* | 835 | 0.639 |
|  | *Pedobacter glucosidilyticus* | 655 | 0.501 |
|  | *Myroides marinus* | 492 | 0.377 |
|  | *Flavobacterium suncheonense* | 391 | 0.299 |
|  | *Myroides profundi* | 389 | 0.298 |
|  | *Pedobacter luteus* | 372 | 0.285 |
|  | *Pedobacter oryzae* | 366 | 0.28 |
|  | *Flavobacteria bacterium* | 340 | 0.26 |
|  | *Flavobacterium beibuense* | 337 | 0.258 |
|  | *Pedobacter arcticus* | 315 | 0.241 |
|  | *Flavobacterium enshiense* | 306 | 0.234 |
|  | *Flavobacterium haoranii* | 292 | 0.224 |
|  | *Pedobacter steynii* | 283 | 0.217 |
|  | *Flavobacterium terrae* | 282 | 0.216 |
|  | *Flavobacterium indicum* | 272 | 0.208 |
|  | *Flavobacterium cauense* | 258 | 0.197 |
|  | *Flavobacterium fontis* | 257 | 0.197 |
|  | *Flavobacterium saliperosum* | 250 | 0.191 |
|  | *Flavobacterium filum* | 236 | 0.181 |
|  | *Flavobacterium cucumis* | 235 | 0.18 |
|  | *Flavobacterium psychrophilum* | 234 | 0.179 |
|  | *Pedobacter africanus* | 229 | 0.175 |
|  | *Flavobacterium sasangense* | 229 | 0.175 |
|  | *Flavobacterium akiainvivens* | 229 | 0.175 |
|  | *Flavobacterium limnosediminis* | 229 | 0.175 |
|  | *Flavobacterium subsaxonicum* | 220 | 0.168 |
|  | *Pedobacter antarcticus* | 217 | 0.166 |
|  | *Flavobacterium rivuli* | 205 | 0.16 |
|  | *Flavobacterium johnsoniae* | 202 | 0.155 |
|  | *Pedobacter nyackensis* | 200 | 0.153 |
|  | *Flavobacterium gelidilacus* | 195 | 0.149 |
|  | *Pedobacter cryoconitis* | 189 | 0.145 |
|  | *Flavobacterium frigoris* | 188 | 0.145 |
|  | *Pedobacter panaciterrae* | 181 | 0.139 |
|  | *Pedobacter heparinus* | 178 | 0.136 |
|  | *Flavobacterium daejeonense* | 167 | 0.128 |
|  | *Flavobacterium branchiophilum* | 162 | 0.124 |
|  | *Flavobacterium soli* | 159 | 0.122 |
|  | *Pedobacter jeongneungensis* | 158 | 0.121 |
|  | *Pedobacter kyungheensis* | 155 | 0.118 |
|  | *Flavobacterium aquatile* | 150 | 0.115 |
|  | *Pedobacter lusitanus* | 145 | 0.111 |
|  | *Pedobacter caeni* | 145 | 0.111 |
|  | *Flavobacterium antarcticum* | 145 | 0.111 |
|  | *Pedobacter agri* | 140 | 0.107 |
|  | *Flavobacterium flevense* | 137 | 0.105 |
|  | *Flavobacterium succinicans* | 136 | 0.104 |
|  | *Flavobacterium fluvii* | 136 | 0.104 |
| ***CM*** | *Sphingobacterium sp.* | 3182 | 2.936 |
|  | *Clostridium sp.* | 1385 | 1.278 |
|  | *Bacillus sp.* | 1353 | 1.249 |
|  | *Chryseobacterium sp.* | 1277 | 1.179 |
|  | *Sphingobacterium spiritivorum* | 572 | 0.528 |
|  | *Clostridium botulinum* | 474 | 0.438 |
|  | *Clostridioides difficile* | 458 | 0.423 |
|  | *Sphingobacterium nematocida* | 387 | 0.358 |
|  | *Sphingobacterium paucimobilis* | 383 | 0.353 |
|  | *Bacillus cereus* | 377 | 0.348 |
|  | *Sphingobacterium psychroaquaticum* | 376 | 0.347 |
|  | *Sphingobacterium deserti* | 353 | 0.326 |
|  | *Sphingobacterium thalpophilum* | 324 | 0.299 |
|  | *Clostridium ultunense* | 169 | 0.156 |
|  | *Clostridium purinilyticum* | 157 | 0.145 |
|  | *Clostridium perfringens* | 141 | 0.13 |
|  | *Clostridium formicaceticum* | 120 | 0.111 |
|  | *Bacillus megaterium* | 111 | 0.102 |
| ***FM*** | *Sphingobacterium sp.* | 7242 | 7.946 |
|  | *Flavobacterium sp.* | 2019 | 2.281 |
|  | *Pedobacter sp.* | 1712 | 1.94 |
|  | *Chryseobacterium sp.* | 1029 | 1.166 |
|  | *Sphingobacterium deserti* | 876 | 0.993 |
|  | *Sphingobacterium psychroaquaticum* | 858 | 0.973 |
|  | *Sphingobacteriaceae bacterium* | 807 | 0.915 |
|  | *Sphingobacterium nematocida* | 794 | 0.9 |
|  | *Sphingobacterium paucimobilis* | 778 | 0.882 |
|  | *Sphingobacterium thalpophilum* | 774 | 0.878 |
|  | *Flavobacterium columnare* | 550 | 0.624 |
|  | *Pedobacter glucosidilyticus* | 513 | 0.582 |
|  | *Pedobacter luteus* | 352 | 0.399 |
|  | *Pedobacter oryzae* | 343 | 0.389 |
|  | *Pedobacter arcticus* | 338 | 0.383 |
|  | *Flavobacterium haoranii* | 189 | 0.214 |
|  | *Flavobacterium suncheonense* | 178 | 0.202 |
|  | *Flavobacterium terrae* | 171 | 0.194 |
|  | *Flavobacterium indicum* | 166 | 0.188 |
|  | *Flavobacterium psychrophilum* | 156 | 0.177 |
|  | *Flavobacterium filum* | 150 | 0.17 |
|  | *Flavobacterium beibuense* | 145 | 0.164 |
|  | *Pedobacter africanus* | 144 | 0.163 |
|  | *Pedobacter nyackensis* | 130 | 0.147 |
|  | *Flavobacterium enshiense* | 130 | 0.147 |
|  | *Flavobacterium cucumis* | 130 | 0.147 |
|  | *Flavobacterium akiainvivens* | 125 | 0.141 |
|  | *Flavobacterium cauense* | 117 | 0.132 |
|  | *Flavobacterium rivuli* | 114 | 0.129 |
|  | *Flavobacterium sasangense* | 113 | 0.128 |
|  | *Parabacteroides merdae* | 113 | 0.128 |
|  | *Flavobacterium subsaxonicum* | 110 | 0.125 |
|  | *Pedobacter steynii* | 109 | 0.124 |
|  | *Flavobacterium johnsoniae* | 108 | 0.122 |
|  | *Flavobacterium branchiophilum* | 103 | 0.117 |
|  | *Flavobacterium saliperosum* | 98 | 0.111 |
|  | *Pedobacter heparinus* | 98 | 0.111 |
|  | *Flavobacterium frigoris* | 97 | 0.11 |
|  | *Pedobacter antarcticus* | 96 | 0.109 |
|  | *Flavobacterium daejeonense* | 94 | 0.107 |
|  | *Pedobacter panaciterrae* | 91 | 0.103 |
|  | *Flavobacterium flevense* | 90 | 0.102 |
|  | *Flavobacterium aquatile* | 90 | 0.102 |
|  | *Flavobacterium limnosediminis* | 89 | 0.101 |
|  | *Pedobacter cryoconitis* | 89 | 0.101 |
| ***WH*** | *Gilliamella apicola* | 831 | 2.797 |
|  | *Parabacteroides sp.* | 580 | 1.952 |
|  | *Bacteroides sp.* | 491 | 1.649 |
|  | *Dysgonomonas sp.* | 433 | 1.457 |
|  | *Dysgonomonas mossii* | 267 | 0.899 |
|  | *Dysgonomonas gadei* | 244 | 0.821 |
|  | *Gilliamella sp.* | 205 | 0.69 |
|  | *Dysgonomonas capnocytophagoides* | 180 | 0.606 |
|  | *Dysgonomonas macrotermitis* | 170 | 0.572 |
|  | *Parabacteroides distasonis* | 142 | 0.478 |
|  | *Bacteroides fragilis* | 96 | 0.323 |
|  | *Parabacteroides merdae* | 91 | 0.306 |
|  | *Parabacteroides chartae* | 78 | 0.263 |
|  | *Bacteroides thetaiotaomicron* | 67 | 0.226 |
|  | *Parabacteroides johnsonii* | 64 | 0.215 |
|  | *Parabacteroides goldsteinii* | 62 | 0.209 |
|  | *Bacteroides uniformis* | 50 | 0.168 |
|  | *Bacteroides finegoldii* | 41 | 0.138 |
|  | *Bacteroides pyogenes* | 41 | 0.138 |
|  | *Bacteroides massiliensis* | 39 | 0.131 |
|  | *Bacteroides ovatus* | 39 | 0.131 |
|  | *Gilliamella mensalis* | 33 | 0.111 |
|  | *Bacteroides vulgatus* | 31 | 0.104 |

## 4.10 Statistical analysis and visualization

### 4.10.1 Relative activity of microorganisms in the metatranscriptomes

The relative activity of the organisms present in the metatranscriptomes was illustrated using stacked bar plots.

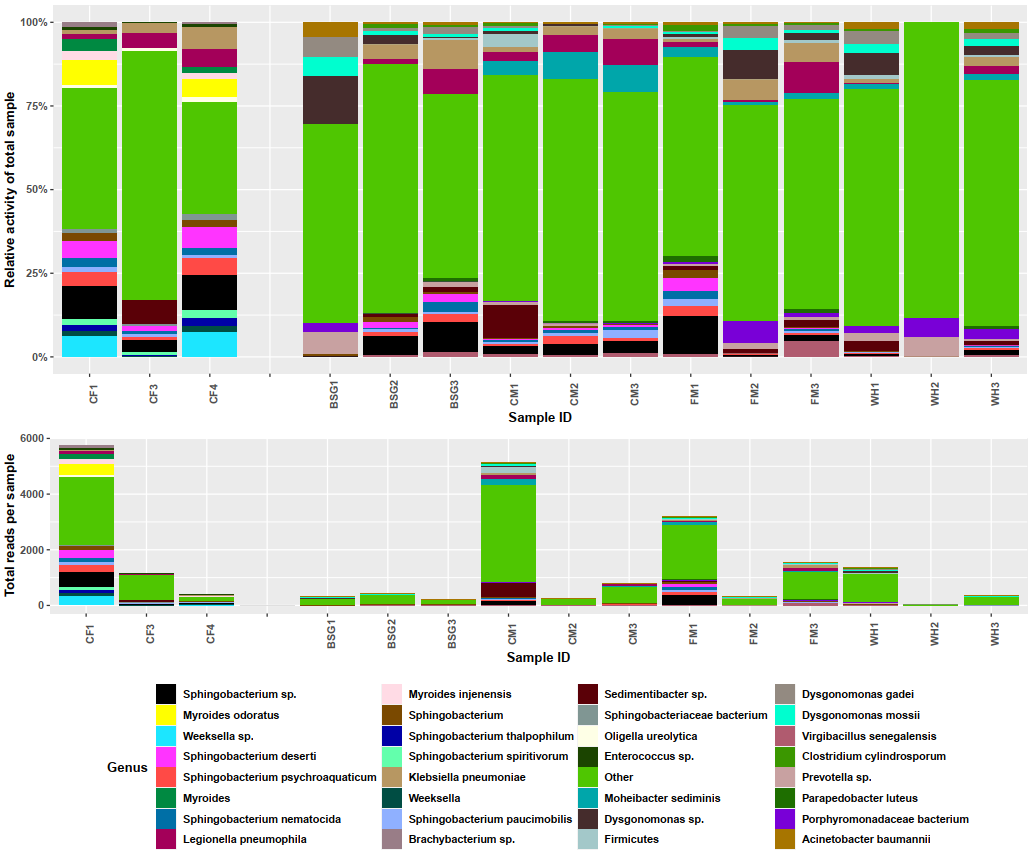


Figure 17: Stacked bar plots showing microbial activity in each metatranscriptome

The Stacked bar plots show the relative activity of various microorganisms in the various study metatranscriptomes and the total reads present in each metatranscriptome.The suffixes after each diet’s name were indicative of the replicate ID while the blank space after CF4 demarcates the control and experimental samples.The total number of reads in each metatranscriptome varied significantly. For this reason, library normalization was performed using the DESeq2 package in R software, and reads from the same dietary substrate were pooled for hierarchical functional analysis and CAZyme annotation.

### 4.10.2 Taxonomic validation of metatranscriptome sequences using 16S rRNA reads



Figure 18: Taxonomic profiling using 16S rRNA sequences filtered at the ribodepletion step

To provide additional validation of the metatranscriptomic sequences, the 16S rRNA reads filtered out during the ribodepletion step were used to perform taxonomic profiling. The taxonomic bar plot (**Figure 18**) shows classification at the Order level after generating ASV files from the 16S Usearch-Vsearch analysis pipeline. The order Bacteroidales which showed high abundance from the 16S rRNA taxonomic profiling was further subsetted to attain resolution at the genus level for further comparison with the most active genera from the metatranscriptomic sequences.

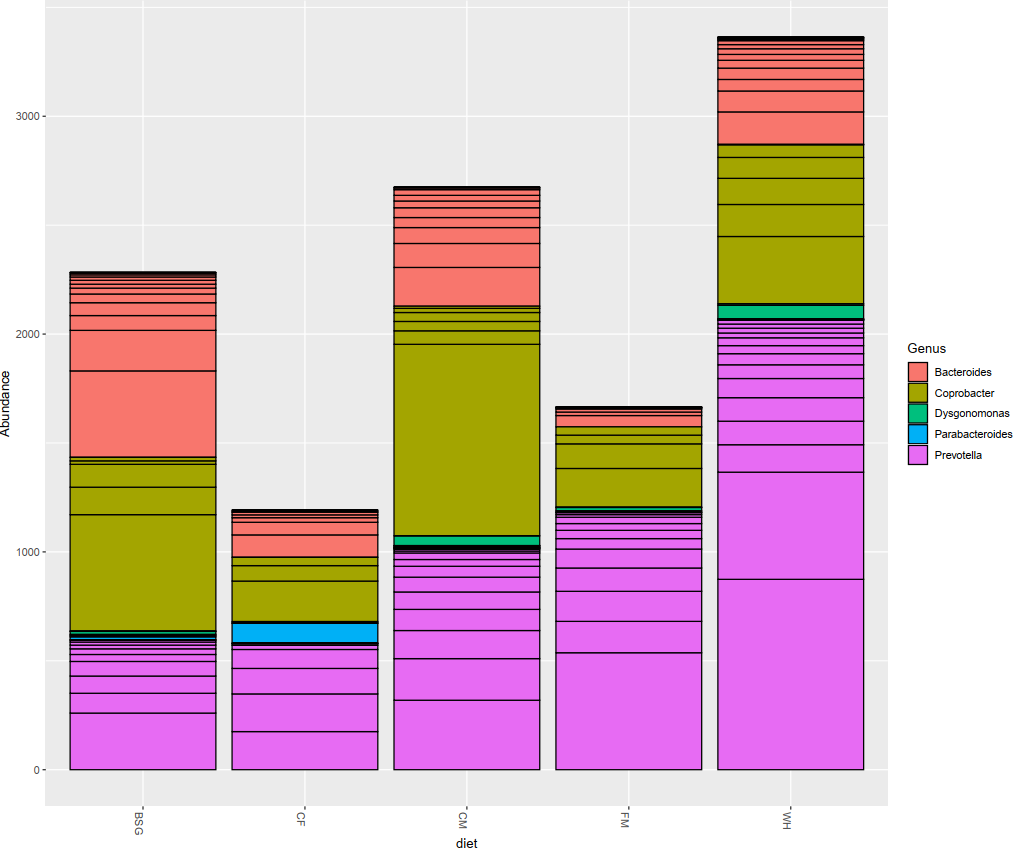


Figure 19: Order Bacteroidales taxonomic profiling using 16S rRNA sequences

From **Figure 19**, a further subset of the order Bacteroidales revealed 5 dominant genera, 3 (*Bacteroides, Coprobacter, and Prevotella*) of which were common in all the 5 study diets. Individual lines within each color represent the number of distinct ASVs observed in each genus. This was further elaborated using the phylogenetic tree (**Figure 20**).

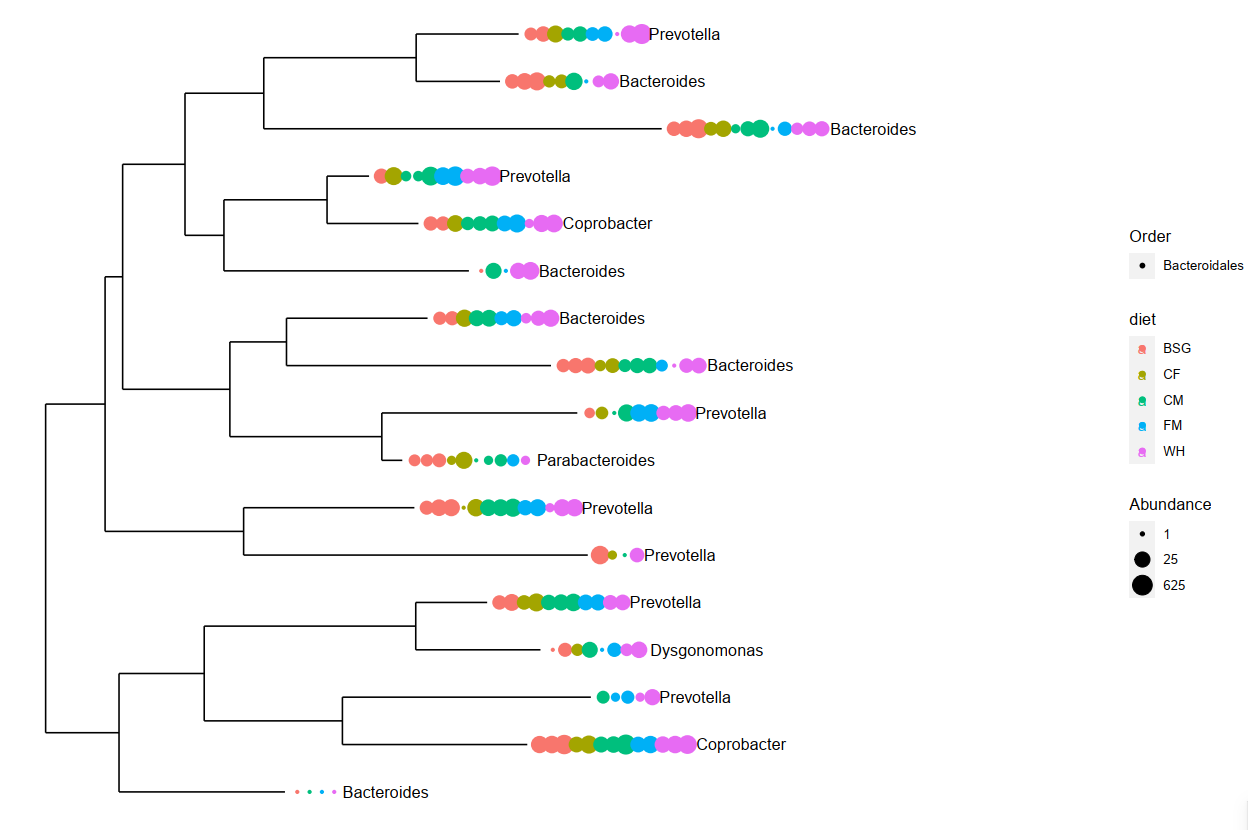


Figure 20:Phylogenetic tree of 16S rRNA samples from order Bacteroidales

The phylogenetic tree (**Figure 20**) further showed the taxonomic relationship between the 5 abundant genera observed during 16S rRNA analysis in order Bacteroidales in the different dietary samples. The color of the dot represents the dietary sample, the number of dots in each clade represents the number of individual ASVs observed in each taxon, and the size of the dot indicates the abundance of individual ASVs in each taxon. There was a notable abundance of genus *Coprobacter* across all dietary substrates, a genus that was not observed in abundance from the Metatranscriptomics analysis.

### 4.10.3 Metatranscriptome diversity statistics

After the metatranscriptome libraries were normalized and differential statistics computed using the DESeq2 package in R software, we found that higher median diversity was observed for experimental samples in all except the BSG metatranscriptome samples using the Shannon-Wiener (H) diversity index as shown in **Figure 21**. The Shannon-Wiener index (H) is a widely accepted alpha diversity index, as it accounts for entropy in representative samples or an eco-system (Jost, 2006).

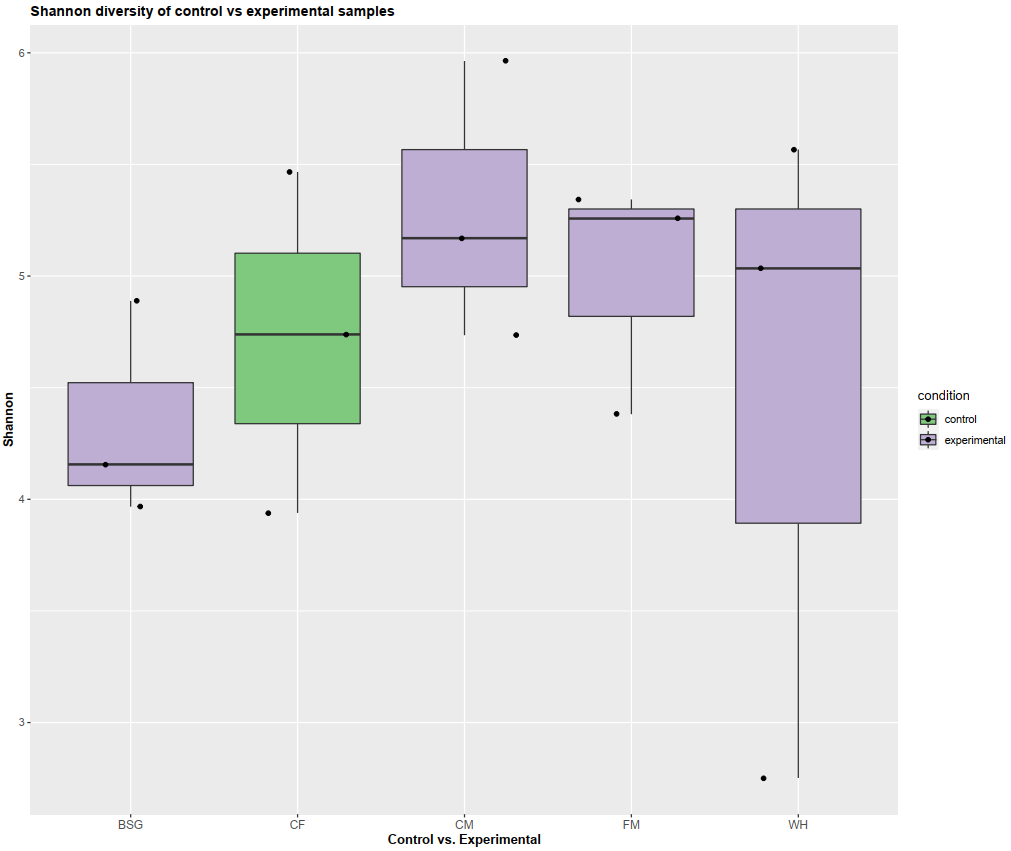


Figure 21: Alpha diversity evaluated using the Shannon-Wiener index (H)

Higher Shannon diversity (H) in the experimental metatranscriptomes was observed in all except the BSG metatranscriptome in comparison to the control metatranscriptomes. The mean Shannon diversity (H) recorded for control metatranscriptomes was 3.96 while the mean diversity of the experimental samples was 4.77.

### 4.10.4 DESeq2 organism heatmap

The differential organism statistics in the study metatranscriptomes were calculated using DESeq2, and a distance heatmap showing the organism profiles in each metatranscriptome was generated.

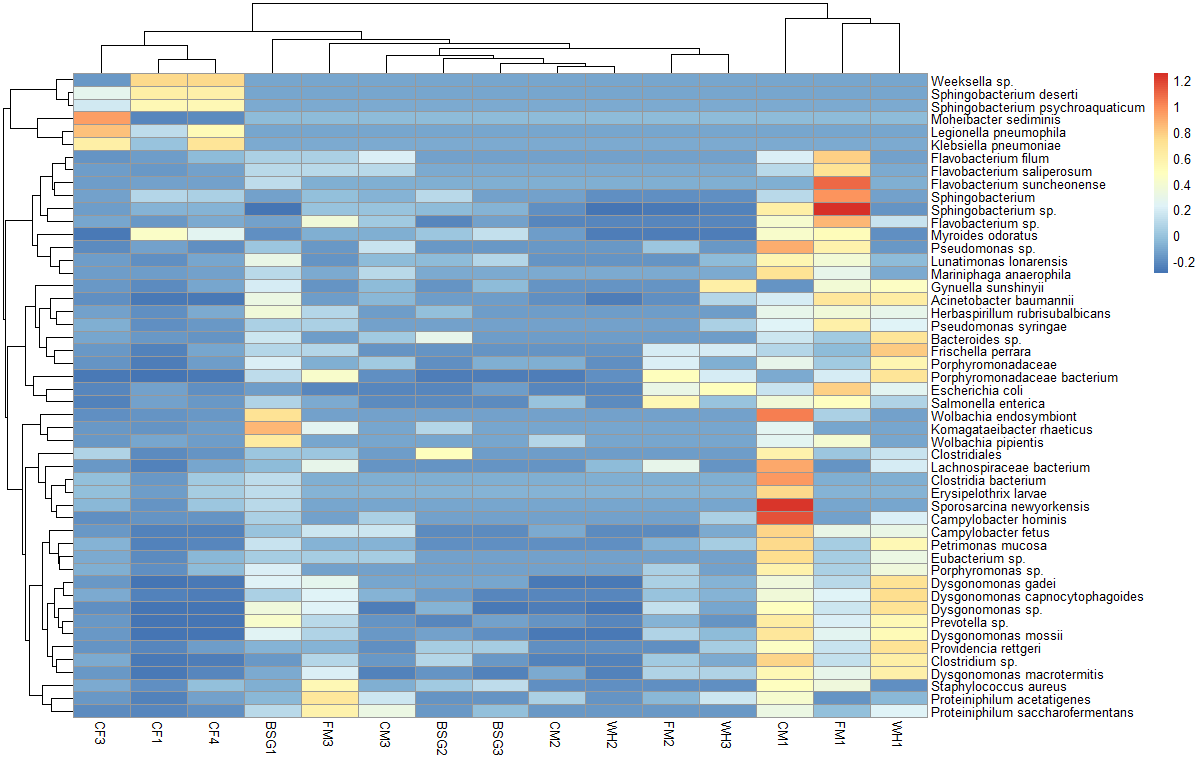


Figure 22: DESeq2 distance heatmap for metatranscriptome function profiles

From the organism heatmap, the highest dissimilarity as illustrated by the Euclidean distance dendrogram was observed between the CF metatranscriptome samples (controls) and the WH metatranscriptome samples (highly lignocellulosic).

### 4.10.5 Organism PCA plot

To further demonstrate how the organism profiles in the study metatranscriptomes varied from each other, a Principal Component Analysis (PCA) plot was generated with ellipses included illustrating the variances between each dietary group.

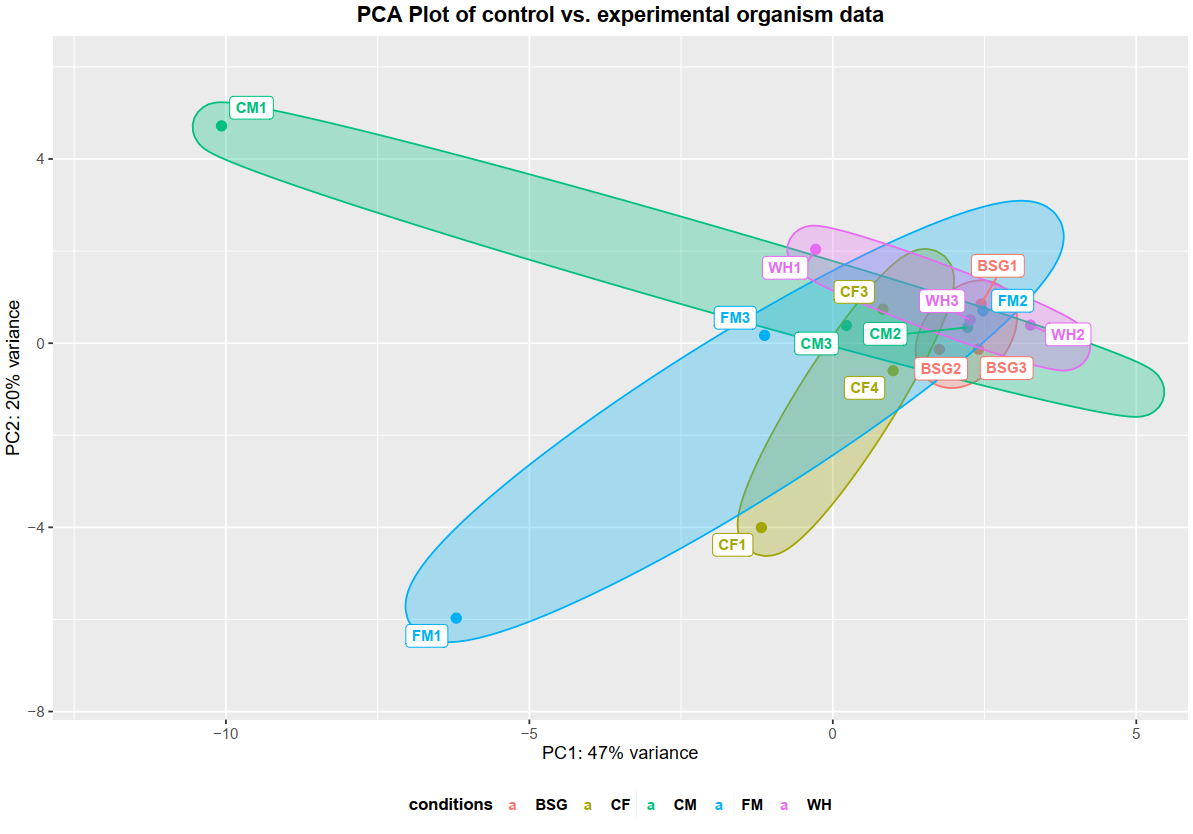


Figure 23: Organism PCA plot

The two main principal components (PC1 and PC2) accounted for 67% of the total variance. The lignocellulose-rich diets (BSG and WH) appeared to cluster closely on the right side of the plot.

**4.10.6 Comparison of diversity statistics using16S rRNA sequences**

16S rRNA sequences filtered out at the ribodepletion step were further used to compare the metatranscriptome diversity statistics using the Phyloseq package (McMurdie & Holmes, 2013) in R software’s Bioconductor library (v3.11) (Love et al., 2014). Observed and Shannon Alpha diversity statistics and Bray Curtis beta diversity statistics were computed and are shown in **Figure 24**. The species Observed index is equivalent to the richness measurement and does not provide insights into the evenness element while the Shannon index equally weighs the evenness and richness (Wagner et al., 2018).

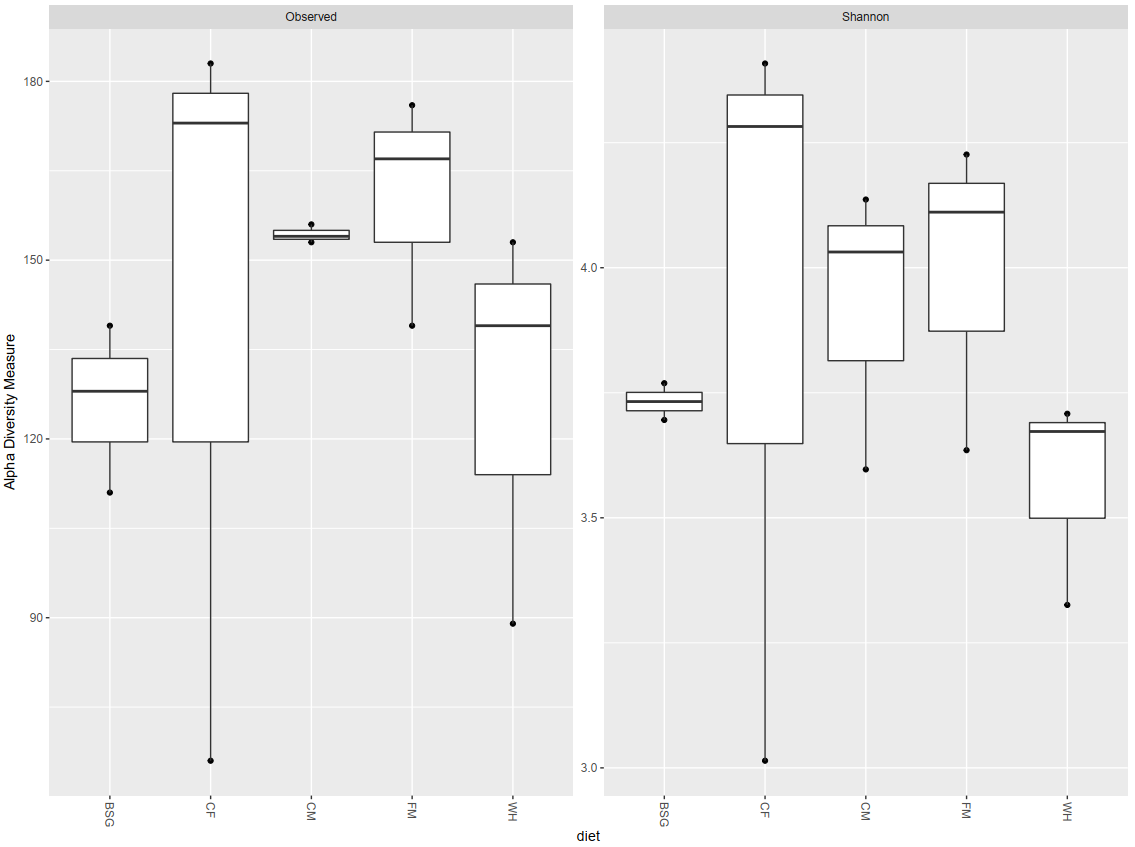


Figure 24: Alpha diversity analysis using 16S rRNA sequences

The highest observed diversity was recorded in the CF diet samples indicative of higher richness in the CF metatranscriptomes while the lowest observed diversity (lowest richness) was recorded in the BSG diet samples. However, the lowest Shannon diversity was observed in the WH diet samples compared to the highest Shannon diversity which was recorded in the CF diet samples. This was indicative that the highest richness and evenness were observed in the control CF diet samples compared to the experimental samples, where the lowest richness and evenness were observed in the WH diet samples.

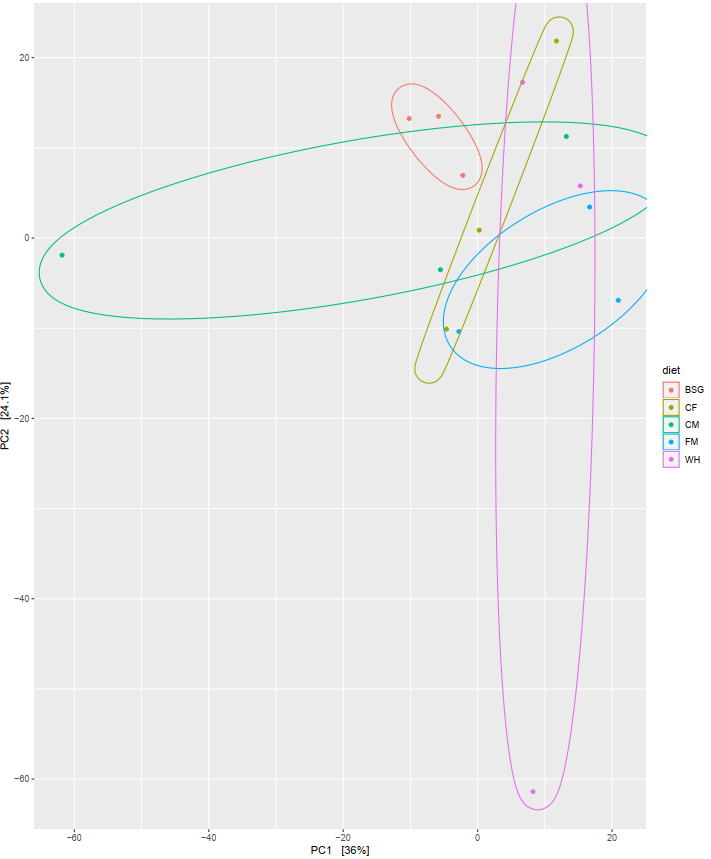
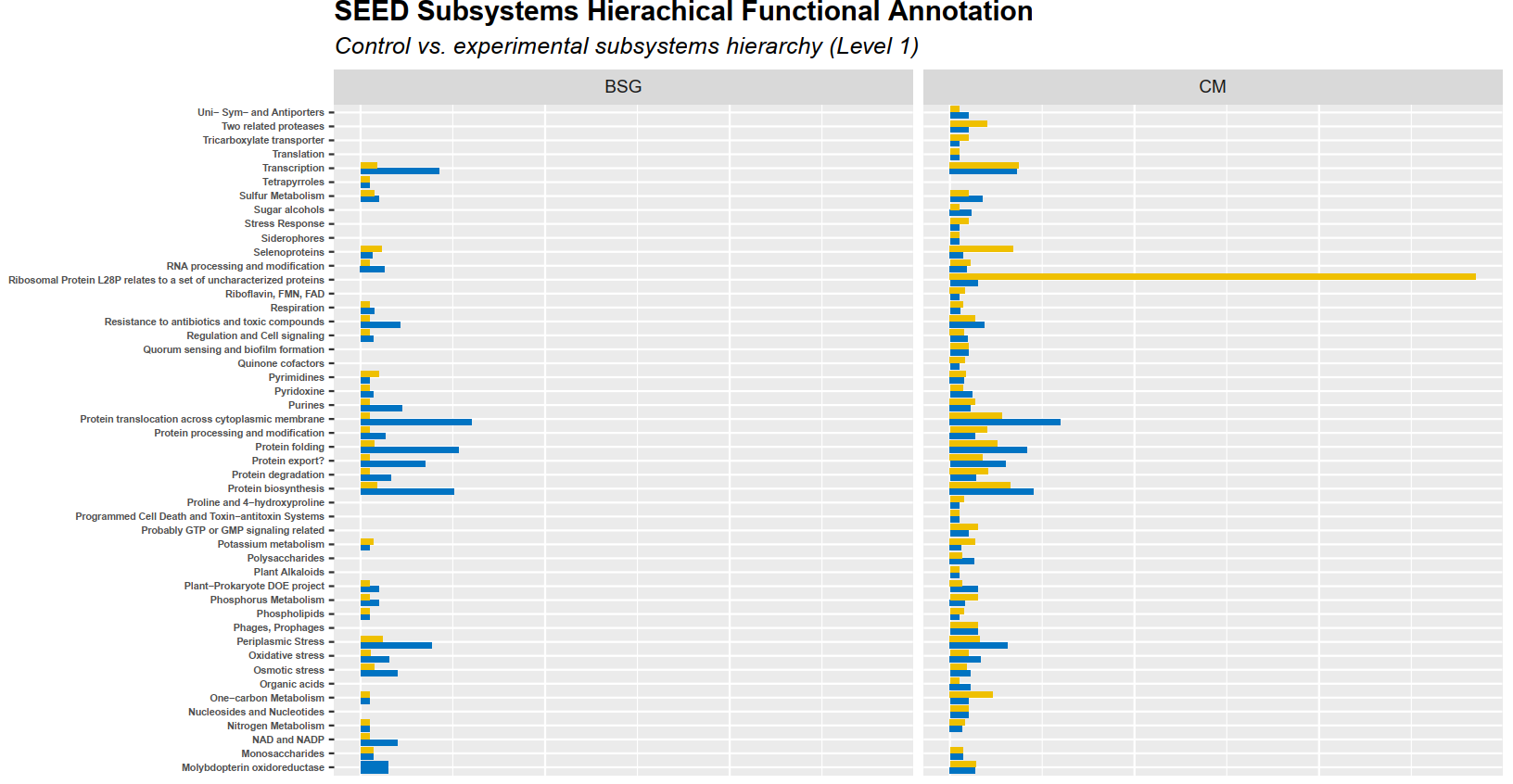


Figure 25: Beta diversity analysis using 16S rRNA sequences

From the Bray-Curtis PCA plot recommended for dissimilarity ordination on organism abundance data (Høj et al., 2018) we found that the two main principal components PC1 and PC2 accounted for 60.1% of the total variance with no distinct inter-dietary sample variance patterns observed.

### 4.10.7 Relative functional activity in the metatranscriptomes

Since metatranscriptomics involves understanding which organisms are performing what functions in a microbial community, the highly active functions were studied by pooling all experimental metatranscriptomes belonging to one diet and comparing them against the pooled control samples (CF). These were represented using the dodged bar plots generated from DIAMOND annotation against the SEED subsystems hierarchical database (**Figure 26**). Functional analysis on the individual metatranscriptome samples (unpooled) was also performed and is represented by the function heatmap (**Figure 27**) generated by DIAMOND annotation against the RefSeq database.



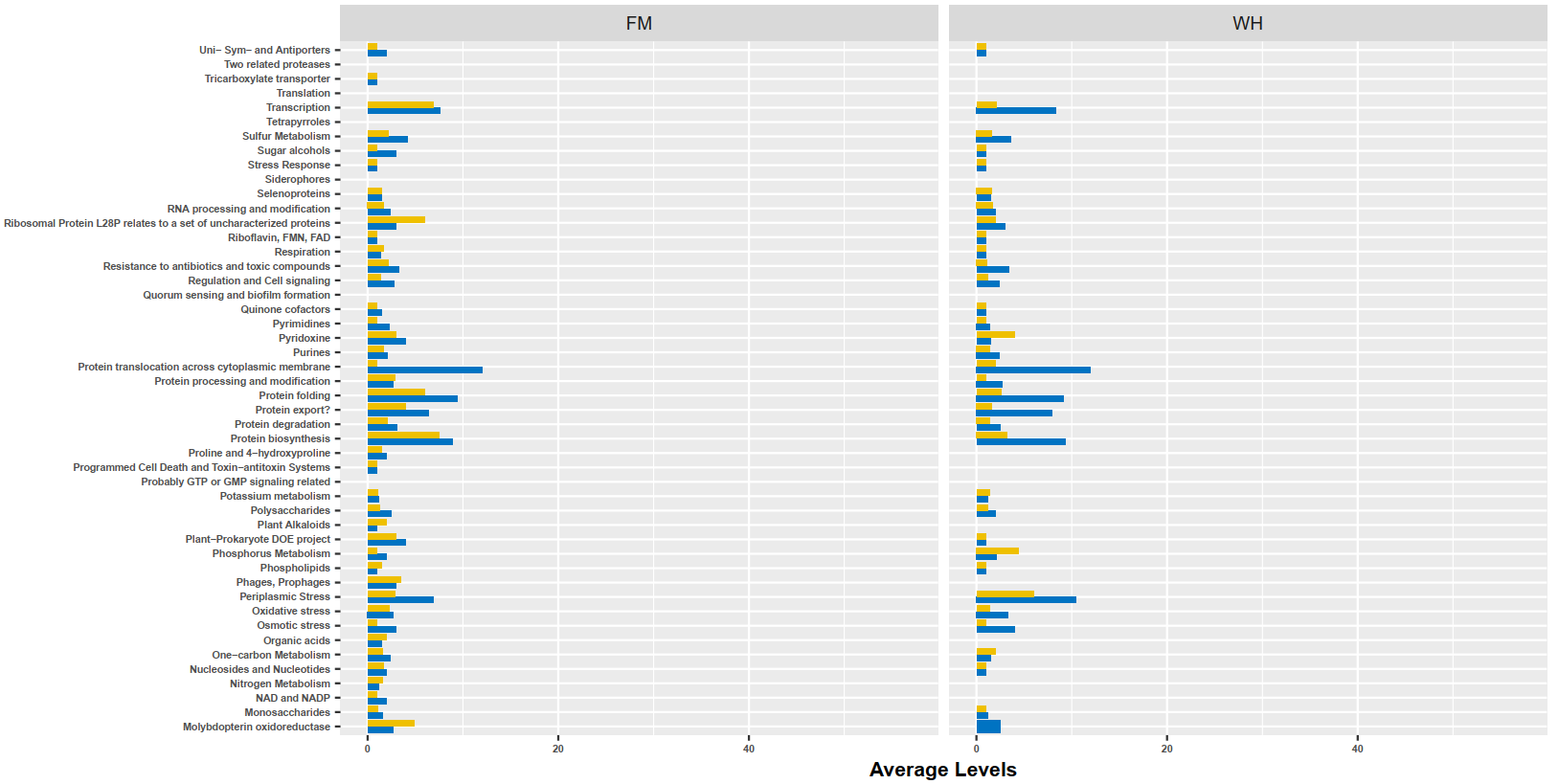
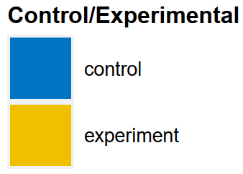
 

Figure 26: Dodged bar plots of active SEED subsystems level 1 functions against the average levels in the pooled experimental metatranscriptomes compared to the pooled control metatranscriptome (CF)

In all the metatranscriptomes, most functions were observed to be more abundant in the pooled control metatranscriptome compared to the experimental metatranscriptomes with a few exceptions. In the BSG pooled metatranscriptome, the average levels of pyrimidines, potassium metabolism, and selenoproteins observed were higher than in the control metatranscriptome. Transcription, ribosomal protein L28P, selenoproteins, one-carbon metabolism, and RNA processing modification functions were found to be more enriched in the pooled CM metatranscriptome than in the pooled control metatranscriptome, CF. In the pooled FM metatranscriptome, molybdopterin oxidoreductase, phages and prophages, organic acids, and nitrogen metabolism were observed in more abundance than in the pooled control metatranscriptome. In the WH pooled metatranscriptome pyridoxine and phosphorous metabolism average levels were observed to be more than in the pooled control metatranscriptome.

### 4.10.8 DESeq2 function heatmap

The DESeq2 function heatmap with Euclidean distances was generated to compare the functional profiles between individual metatranscriptomes.

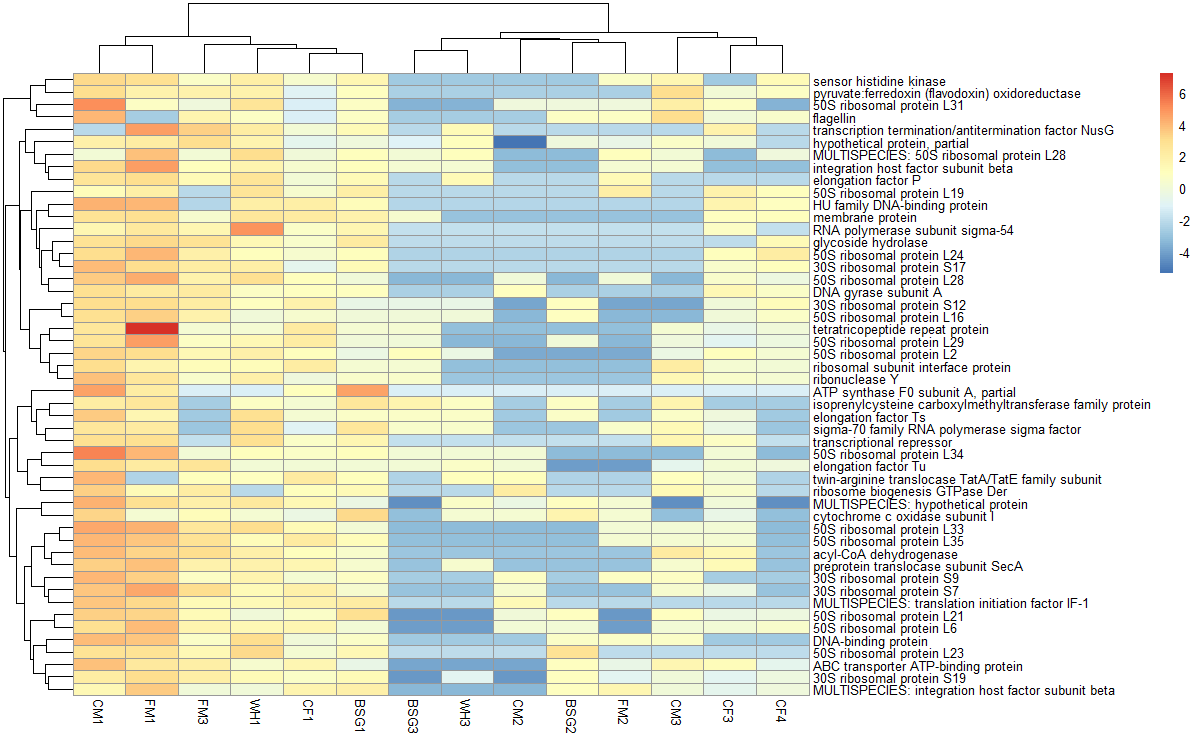


Figure 27: DESeq2 distance heatmap for metatranscriptome function profiles

From the function heatmap incorporating the 50 most active functions based on annotation with the bacterial RefSeq database, the highest dissimilarity as illustrated by the Euclidean distance dendrogram was observed between the CF4 metatranscriptome (control) and the CM1 metatranscriptome. Two metatranscriptomes (CF2 and WH2) with very few active functions recorded were not included in this heat map.

### 4.10.9 CAZymes annotation with dbCAN2 Hotpep module

Study metatranscriptomes from the same diet were pooled together before this step and the CAZymes were annotated with the Hotpep module that performs annotations by matching the conserved peptides to the protein sequences of interest. The results are summarized in **Table 4.8**.

Table 4. 9: CAZyme identification and annotation using the dbCAN2 Hotpep module

The FM metatranscriptome recorded the highest number of CAZymes (11), while the CM metatranscriptome recorded the lowest (1). Metatranscriptomes BSG, control CF, and WH recorded 3,5, and 2 CAZyme groups respectively. For some of the CAZyme groups, the Enzyme Commission (EC) number which uniquely identifies an enzyme based on the reaction it catalyzes was included (Webb E. C., 1993).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Dietary Substrate*** | ***CAZyme Family*** | ***Sub-family*** | ***Signature peptides (Hotpep)*** | ***EC Number*** |
| ***BSG*** | **GH43** | **16** | **ESRSVT, SRSVTS, EVNRRW, VNRRWT, NRRWTE, RRWTEG** | **3.2.1.37, 3.2.1.55** (Mewis et al., 2016) |
|  | GT2 | 46 | YILWLD, FAAARN, ILWLDA, LDADDI, WLDADD, LWLDAD, EYILWL |  |
|  | GT51 | 2 | GSTITQ, STITQQ, GGSTIT, QGGSTI, ITQQLA, TITQQL | 2.4.1.129:58 |
| ***CF*** | CE11 | 1 | LDAIGD, MLDAIG, DEFVRH, EFVRHK, HKMLDA, DAIGDL, KMLDAI, MRDIEY, DIEYLQ, RDIEYL, IEYLQS, GFMRDI, FGFMRD, FMRDIE | 3.5.1.108:260 |
|  |  | 3 | IEIDGP, IDGPEV, PILDGS, GPEVPI, PEVPIL, VPILDG, DGPEVP, EVPILD, EIDGPE |  |
|  |  | 6 | VPDPEN, PGDTLI,GDTLIF, QTGGIL, AQTGGI, MAQTGG, AMAQTG, EAMAQT, PDPENY |  |
|  | GH19 | 4 | GTNGLA, TNGLAD, NGGTNG, LADRQA, GGTNGL, NGLADR, GLADRQ, INGGTN |  |
|  | GH95 | 1 | PANLQG, QPANLQ, YTININ, KYTINI, FGRYLL, LQGIWN, QFGRYL, NYWPAE, ANLQGI, TININT, GRYLLI, SKYTIN, NLQGIW |  |
|  | GT26 | 29 | IDELPQ, MSLVGP, SLVGPR, DMSLVG, GDMSLV, LVGPRP |  |
|  | CBM13 | 55 | NLQGIW,PANLQG,GRYLLI,QPANLQ,LQGIWN,ANLQGI |  |
| ***CM*** | GH13 | 27 | QIFPDR, IFPDRF, FYQIFP, VFYQIF, AVFYQI, YQIFPD | 3.2.1.54:41 |
| ***FM*** | CE7 | 2 | FTDAVR, RVFTDA, RRVFTD, TDAVRA, DAVRAV, VFTDAV | 3.1.1.72:35 |
|  | CE9 | 1 | TAPAGA, VTDATA, TDATAP, LVTDAT, ATAPAG, DATAPA | 3.5.1.25:306 |
|  | GH0 | 99 | GCGGRG, VGCGGR, RGTGAA, GGRGTG, GRGTGA, CGGRGT, LVGCGG |  |
|  | GH28 | 11 | GHGGFV, TRGRGG, KSTRGR, FKSTRG, STRGRG, HGGFVV, GGFVVG |  |
|  | GH31 | 13 | AFAGQQ,TRSAFA,KRVFIL,GQQRYG,VFILTR,AGQQRY,RVFILT,RSAFAG,SDKRVF,SAFAGQ,ILTRSA,FAGQQR,LTRSAF,FILTRS |  |
|  | GH33 | 43 | TEAEWE, AEWEYA, PTEAEW, LPTEAE, RLPTEA, EAEWEY |  |
|  | GT2 | 55 | NEKENI, DGSPDG, TYNEKE, IPTYNE, DDGSPD, YNEKEN, SPDGTA, PTYNEK, IIPTYN, LVIIPT, GSPDGT, VIIPTY | 2.4.1.83:53 |
|  | GT28 | 13 | VNPLLA, LAGGGT, PLLATA, TAGHVN, AGGGTA, HVNPLL, GTAGHV, AGHVNP, GGTAGH, LLAGGG, GGGTAG, GHVNPL |  |
|  | GT35 | 1 | GTGNMK, GTLDGA, ASGTGN, LTIGTL, EASGTG, SGTGNM, TIGTLD, GALTIG, ALTIGT, NGALTI, IGTLDG | 2.4.1.1:1499 |
|  | CBM50 | 9 | KLHFEI,VKLHFE,LHFEIR,TGTDRV,TDRVKL,HFEIRR,DRVKLH,GTDRVK,RVKLHF |  |
|  | CBM6 | 53 | TEAEWE, AEWEYA, PTEAEW, EWEYAA, RLPTEA, LPTEAE, EAEWEY |  |
| ***WH*** | GH0 | 16 | GRGTGA, LIGCGG, GCGGRG, ALIGCG, GGRGTG, IGCGGR, CGGRGT |  |
|  | **GH51** | **2** | **GNENWG, WGCGGN, ENWGCG, NENWGC, GCGGNM, NWGCGG** | **3.2.1.55:729** |

Family GH43 in the BSG metatranscriptome and family GH51 in the WH metatranscriptome were identified from the CAZy database (www.cazy.org) as possessing lignocellulolytic activity. The Krona taxonomic display multi-layered pie-charts (Ondov et al., 2011) from the CAZy database provided further information on which bacterial organisms these enzyme families have been identified. The organisms present in the taxonomic multi-layered pie-charts were then validated against the most abundant organisms present in the respective metatranscriptomes where these lignocellulolytic enzyme families were identified.

### 4.10.10 CAZy family GH43 subfamily 16 (GH43\_16) from the BSG sample

The GH43 enzyme family contains 37 subfamilies whose conserved sequence residues correspond to the results of structural studies and biochemical assays. Enzymes identified from GH43\_16 include β-glycosidase (EC 3.2.1.-), β-D-xylosidase (EC 3.2.1.37), and α-L-arabinofuranosidase (EC 3.2.1.55) (Mewis et al., 2016).

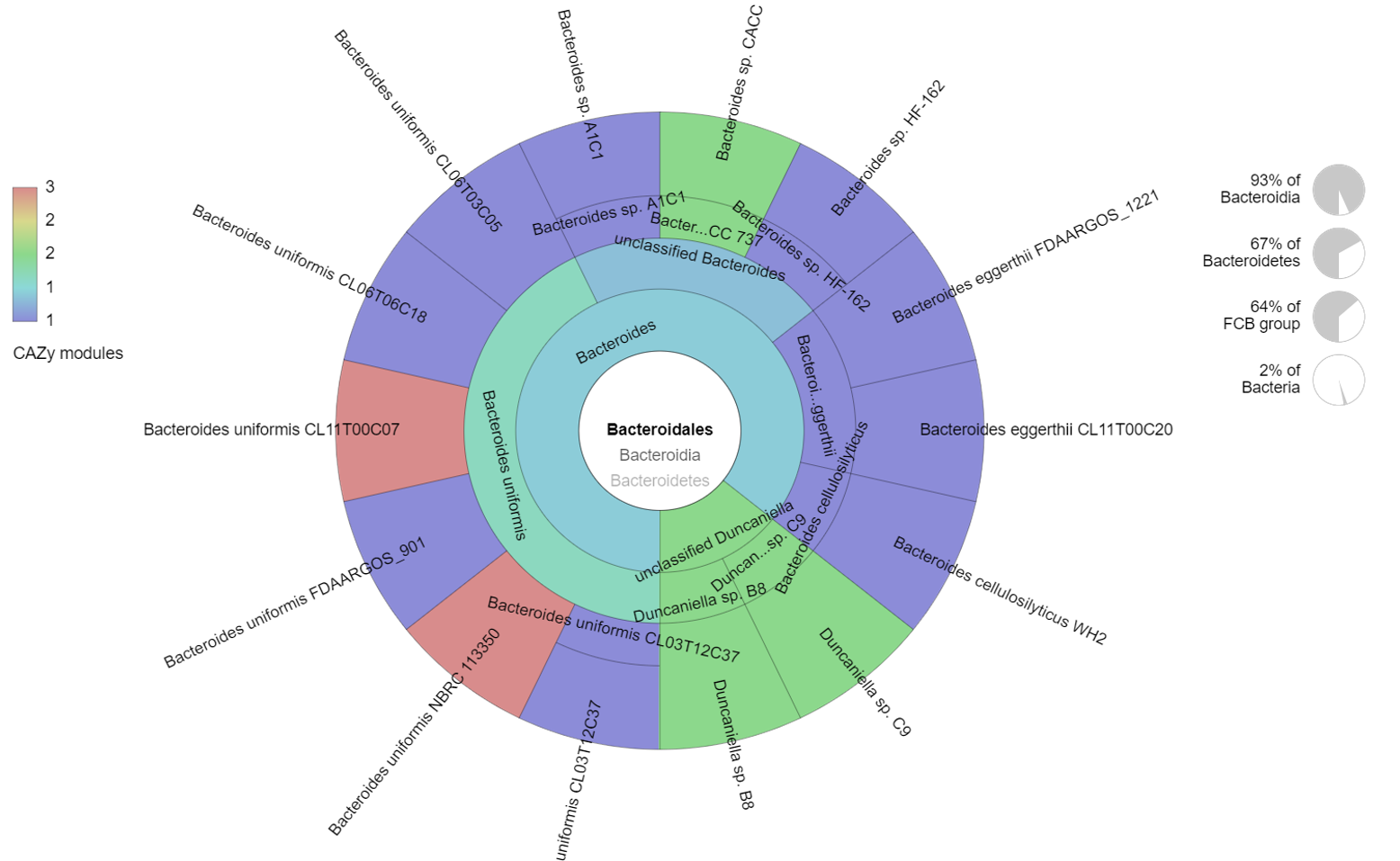


Figure 28: GH43\_16 Krona multi-layered taxonomic pie-chart

The Krona chart (**Figure 28**) above shows CAZy modules (from the CAZy database) that have been identified in bacterial taxonomic order *Bacteroidales* where 3 of the 4 most abundant genera (*Bacteroides, Dysgonomonas,* and *Prevotella*) in the BSG metatranscriptome are classified. From this chart, most of the GH43\_16 CAZy modules have been identified in the *Bacteroides uniformis* species, *Bacteroides sp.*, *Bacteroides Eggerthii* species, and *Bacteroides cellulosilyticus* species. From **Table 4.8**, these bacteria had a collective abundance of 2.111%. From the CAZy database, there were no GH43\_16 CAZy modules identified from genera *Sphingobacterium, Dysgonomonas,* and *Prevotella* genera, which were the most abundant in the BSG metatranscriptome.

### 4.10.11 CAZy family GH51 subfamily 2 (GH51\_2) from the WH sample

From the Hotpep module of dbCAN2, subfamily 2 of the GH51 enzyme family was identified, but the subfamily annotation of this enzyme family was not represented in the CAZy database. Therefore, the generalized GH51 Krona multi-layered taxonomic pie-chart was used to validate the abundant organisms in the WH metatranscriptome. Like the GH43 enzyme family, the GH51 enzyme family contains α-L-arabinofuranosidases (abf) (EC 3.2.1.55) to release the arabinofuranosyl (araf) side-chains during arabinoxylan hydrolysis (Dos Santos et al., 2018).

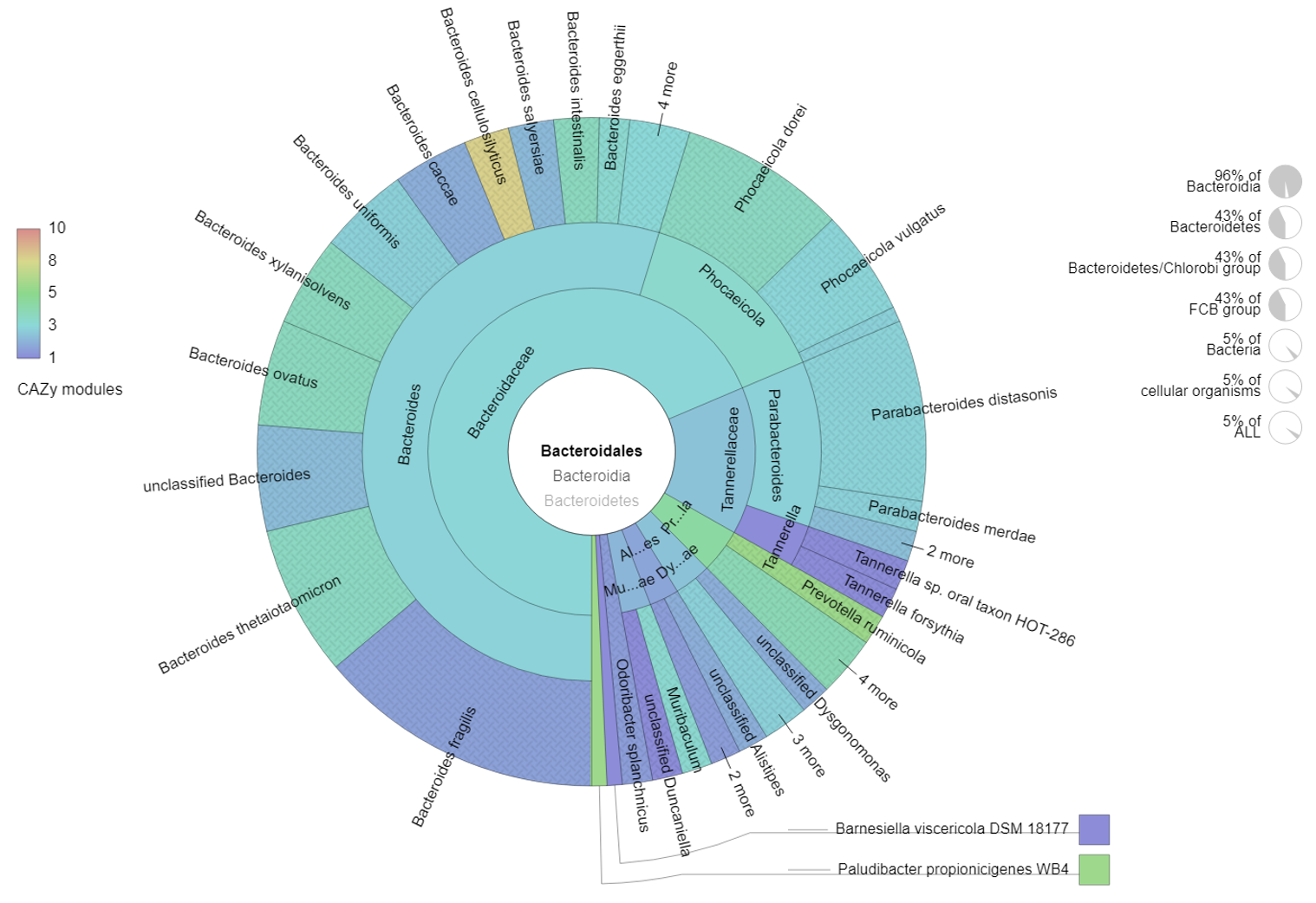


Figure 29: GH51 Krona multi-layered taxonomic pie-chart

**Figure 29** above shows CAZy modules (from the CAZy database) that have been identified in bacterial taxonomic order *Bacteroidales,* where 3 of the 4 most abundant genera (*Dysgonomonas, Bacteroides,* and *Parabacteroides*) in the WH metatranscriptome are classified. From this figure, most of the GH51 CAZy modules have been identified in the *Bacteroides cellulosilyticus, Prevotella ruminicola,* and *Plaudibacter propionicigenes* species. From the WH metatranscriptome,these species did not attain the >0.1% threshold abundance as indicated in **Table 4.8**. However, bacterial species *Parabacteroides diastonis*, *Bacteroides fragilis*, *Parabacteoides merdae*, *Bacteroides thetaiotamicron, Bacteroides uniformis,* and *Bacteroides* *ovatus* attained the >0.1 abundance threshold (**Table 4.8**), and have been identified in **Figure 29** to possess CAZy modules of the GH51 enzyme family. These bacteria had a collective abundance of 1.632% in the WH metatranscriptome. The *Dysgonomonadaceae* family has also been identified to possess CAZy modules from the GH51 family. However, from **Figure 29** above, none had been classified into any of the known *Dysgonomonas* species. From **Table 4.8**, these bacteria had a collective abundance of 2.111%. From the CAZy database, no CAZy modules from the GH51 family have been identified from genus *Gilliamella,*  which was among the most abundant in the WH metatranscriptome.

### 4.10.12 Screening for polysaccharide utilization loci (PULs)

To screen for PULs, the metatranscriptome sequences from the highly lignocellulosic diets BSG and WH that revealed lignocellulolytic CAZy families were screened against the dbCAN-PUL database (Ausland et al., 2021) and compared with the PULs identified in the control metatranscriptome, CF. From the 15 PULs with the highest percentage identity to these metatranscriptomes, keen interest was taken in PULs that had CAZy hits that corresponded to the CAZy families identified using the dbCAN2 Hotpep module (**Table 4.9)**.

Table 4. 10: Screening for polysaccharide utilization loci (PULs)

PULs screening using dbCAN-PUL in the control sample sequences (CF) and the highly lignocellulosic metatranscriptomes, BSG and WH.

Table 4.10a CF Metatranscriptome

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Query ID*** | ***PUL ID*** | ***Hit name*** | ***Hit Locus*** | ***Hit Protein ID*** | ***Hit Type*** | ***Identity (%)*** | ***E-value*** | ***Annotation*** |
| *107:375|+* | PUL0553 | N.A | *Bacteroides xylanisolvens XB1A* | CBK67945.1 | CAZyme | 76.19 | 9.85E-07 | GH97 |
| *104:193|+* | PUL0328 | N.A | *Gramella flava JLT2011* | APU67024.1 | CAZyme | 76 | 3.82E-09 | GH2 |
| *113:249|-* | PUL0557 | N.A | *Bacteroides thetaiotaomicron/818* | AAO77739.1 | CAZyme | 75 | 5.01E-07 | GH125 |
| *113:249|-* | PUL0375 | N.A | *Bacteroides thetaiotaomicron/818* | AAO77739.1 | CAZyme | 75 | 5.01E-07 | GH125 |
| *113:249|-* | PUL0161 | N.A | *Bacteroides thetaiotaomicron/818* | WP\_011108410.1 | CAZyme | 75 | 5.01E-07 | GH125 |
| *113:249|-* | PUL0378 | N.A | *Bacteroides thetaiotaomicron VPI-5482* | AAO78886.1 | CAZyme | 75 | 5.92E-07 | GH125 |
| *113:249|-* | PUL0163 | N.A | *Bacteroides thetaiotaomicron/818* | WP\_011108966.1 | CAZyme | 75 | 5.92E-07 | GH125 |
| ***110:395|+*** | **PUL0558** | **N.A** | ***Bacteroides thetaiotaomicron VPI-5482*** | **AAO76117.1** | **CAZyme** | **74.627** | **3.07E-29** | **GH95** |
| ***110:395|+*** | **PUL0530** | **N.A** | ***Bacteroides thetaiotaomicron VPI-5482*** | **AAO76117.1** | **CAZyme** | **74.627** | **3.07E-29** | **GH95** |
| ***110:395|+*** | **PUL0189** | **N.A** | ***Bacteroides xylanisolvens XB1A*** | **CBK65850.1** | **CAZyme** | **74.627** | **6.17E-29** | **GH95** |
| *104:193|+* | PUL0476 | N.A | *Flavobacterium johnsoniae/986* | WP\_012024095.1 | CAZyme | 74.074 | 4.47E-09 | GH2 |
| ***110:395|+*** | **PUL0044** | **N.A** | ***Bacteroides ovatus ATCC 8483*** | **EDO10805.1** | **CAZyme** | **73.134** | **2.57E-28** | **GH95** |
| ***110:395|+*** | **PUL0553** | **N.A** | ***Bacteroides xylanisolvens XB1A*** | **CBK67959.1** | **CAZyme** | **73.134** | **3.65E-28** | **GH95** |
| ***110:395|+*** | **PUL0342** | **N.A** | ***Prevotella ruminicola/839*** | **WP\_013064566.1** | **CAZyme** | **72.131** | **2.10E-26** | **GH95** |
| *108:382|-* | PUL0016 | N.A | *Lactococcus lactis/1358* | WP\_011834483.1 | CAZyme | 70.833 | 2.55E-08 | GH1 |

Table 4.10b: BSG metatranscriptome

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Query ID*** | ***PUL ID*** | ***Hit name*** | ***Hit Locus*** | ***Hit Protein ID*** | ***Hit Type*** | ***Identity (%)*** | ***E-value*** | ***Annotation*** |
| *108:282|-* | PUL0382 | N.A | *Bacteroides thetaiotaomicron VPI-5482* | AAO79348.1 | CAZyme | 70.732 | 3.01E-18 | GH109 |
| *111:299|+* | PUL0325 | N.A | *Bacteroides thetaiotaomicron VPI-5482* | AAO78420.1 | CAZyme | 68 | 6.87E-16 | GH3 |
| *111:299|+* | PUL0526 | N.A | *Bacteroides ovatus/28116* | WP\_004322518.1 | CAZyme | 56.923 | 2.83E-15 | GH3 |
| *111:299|+* | PUL0067 | N.A | *Dysgonomonas gadei ATCC BAA-286* | EGJ99270.1 | CAZyme | 54 | 9.58E-11 | GH3 |
| *116:443|-* | PUL0395 | abfB | N.A | ACE73681.1 | CAZyme | 50 | 2.53E-13 | GH51 |
| *116:443|-* | PUL0013 | abfB | N.A | ACE73681.1 | CAZyme | 50 | 2.53E-13 | GH51 |
| *116:443|-* | PUL0218 | N.A | *termite gut metagenome* | CCO20976.1 | CAZyme | 50 | 5.14E-11 | GH51 |
| *116:443|-* | PUL0532 | abf2 | *Bacteroides cellulosilyticus* | ALJ58197.1 | CAZyme | 47.368 | 3.63E-09 | GH51 |
| *116:443|-* | PUL0218 | N.A | *termite gut metagenome* | CCO20976.1 | CAZyme | 44.737 | 5.14E-11 | GH51 |
| *116:443|-* | PUL0190 | N.A | *Bacteroides xylanisolvens XB1A* | CBK66679.1 | CAZyme | 43.182 | 3.27E-09 | GH51 |
| *116:443|-* | PUL0302 | N.A | *Bacteroides thetaiotaomicron VPI-5482* | AAO75455.1 | CAZyme | 42.105 | 3.69E-09 | GH51 |
| *116:443|-* | PUL0395 | abfB | N.A | ACE73681.1 | CAZyme | 41.509 | 2.53E-13 | GH51 |
| *116:443|-* | PUL0013 | abfB | N.A | ACE73681.1 | CAZyme | 41.509 | 2.53E-13 | GH51 |
| *116:443|-* | PUL0309 | N.A | *Caldanaerobius polysaccharolyticus/44256* | WP\_026486272.1 | CAZyme | 39.623 | 2.75E-07 | GH51 |
| *116:443|-* | PUL0302 | N.A | *Bacteroides thetaiotaomicron VPI-5482* | AAO75455.1 | CAZyme | 38.636 | 3.69E-09 | GH51 |

Table 4.10c: WH metatranscriptome

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Query ID*** | ***PUL ID*** | ***Hit name*** | ***Hit Locus*** | ***Hit Protein ID*** | ***Hit Type*** | ***Identity (%)*** | ***E-value*** | ***Annotation*** |
| ***118:230|+*** | **PUL0395** | **abfB** | **N.A** | **ACE73681.1** | **CAZyme** | **92.308** | **1.11E-08** | **GH51** |
| ***118:230|+*** | **PUL0013** | **abfB** | **N.A** | **ACE73681.1** | **CAZyme** | **92.308** | **1.11E-08** | **GH51** |
| ***118:230|+*** | **PUL0395** | **abfB** | **N.A** | **ACE73681.1** | **CAZyme** | **88.889** | **1.11E-08** | **GH51** |
| ***118:230|+*** | **PUL0013** | **abfB** | **N.A** | **ACE73681.1** | **CAZyme** | **88.889** | **1.11E-08** | **GH51** |
| ***118:230|+*** | **PUL0309** | **N.A** | ***Caldanaerobius polysaccharolyticus/44256*** | **WP\_026486272.1** | **CAZyme** | **83.333** | **1.62E-07** | **GH51** |
| *106:209|+* | PUL0482 | N.A | *Flavobacterium johnsoniae/986* | WP\_012026054.1 | CAZyme | 77.419 | 6.18E-12 | GH78|CBM67 |
| *101:188|-* | PUL0436 | N.A | *Chitinophaga pinensis/79329* | WP\_012792687.1 | CAZyme | 77.419 | 8.92E-09 | GH65 |
| ***118:230|+*** | **PUL0309** | **N.A** | ***Caldanaerobius polysaccharolyticus/44256*** | **WP\_026486272.1** | **CAZyme** | **76.923** | **1.62E-07** | **GH51** |
| *108:415|-* | PUL0472 | N.A | *Bacteroides thetaiotaomicron VPI-5482* | AAO79197.1 | CAZyme | 75 | 7.90E-08 | GH92 |
| *108:415|-* | PUL0380 | N.A | *Bacteroides thetaiotaomicron VPI-5482* | AAO79095.1 | CAZyme | 75 | 2.06E-07 | GH92 |
| *108:415|-* | PUL0353 | N.A | *Bacteroides thetaiotaomicron VPI-5482* | AAO79095.1 | CAZyme | 75 | 2.06E-07 | GH92 |
| *108:415|-* | PUL0426 | N.A | *Bacteroides ovatus ATCC 8483* | EDO08952.1 | CAZyme | 64.706 | 2.98E-08 | GH92 |
| *107:391|+* | PUL0526 | N.A | *Bacteroides ovatus/28116* | WP\_004322518.1 | CAZyme | 63.636 | 5.76E-14 | GH3 |
| *111:254|-* | PUL0564 | N.A | *Bacteroides thetaiotaomicron VPI-5482* | AAO79262.1 | CAZyme | 63.333 | 4.98E-10 | GH27|CBM35 |
| *109:324|-* | PUL0505 | N.A | *Bacteroides fragilis NCTC 9343* | CAH08299.1 | CAZyme | 57.143 | 3.42E-07 | GT2 |

**Key**

N.A – Not assigned

abfB - α-L-arabinofuranosidase B

abf2 - α-L-arabinofuranosidase 2

**Bold** – CAZy family identified by the dbCAN2 Hotpep module (**Table 4.9**)

From **Table 4.10a**, multiple PUL hits for CAZy class GH95 that was identified in the control dietary transcriptome using the dbCAN2 Hotpep module were detected. These hits were PUL0558 and PUL0530 from locus *Bacteroides thetaiotaomicron VPI-5482,* PUL0189, and PUL0553 from *Bacteroides xylanisolvens XB1A,* PUL0044 from *Bacteroides ovatus ATCC 8483,* andPUL 0342 from *Prevotella ruminicola/839.* However, none of these organisms belonged to the 4 most abundant genera identified from the CF metatranscriptome (**Table 4.8**). However, *Bacteroides* and *Prevotella* genera were detected in high abundance from the 16S rRNA taxonomic analysis of the order Bacteroidales for the CF diet (**Figure 19,20**). From **Table 4.10b**, none of the PUL hits detected in the highly lignocellulosic diet, BSG, belonged to any of the CAZy families identified using the dbCAN2 Hotpep module in **Table 4.10**. However, PUL0395 and PUL0013 identified as belonging to GH51 produce the same family of enzymes (α-L-arabinofuranosidases - EC3.2.1.55) as GH43 subfamily 16 identified in the BSG metatranscriptome. These 2 PUL hits encoding the abfB enzyme, were however not assigned to any organism. From **Table 4.10c**, the 3 top PUL hits (PUL0395, PUL0013, and PUL0309) that were detected in the highly lignocellulosic WH sample belonged to the GH51 CAZy family. This was concordant with the results obtained using the dbCAN2 Hotpep module in **Table 4.9.** However, only PUL0309 was assigned to an organism - *Caldanaerobius polysaccharolyticus.* From the dbCAN-PUL results, PUL0395 and PUL0013 identified in the BSG and WH metatranscriptomes encoding the abfB enzyme were not assigned to any organism (**Table 4.10 b,c**).

# 5.0 CHAPTER 5: DISCUSSION

Production of second-generation biofuels has been continually proposed as a viable and sustainable alternative to the fast-depleting, ecologically unfriendly petroleum-based fuels. These fuels, mainly produced from lignocellulosic agricultural and industrial wastes, present a sustainable model of waste management without posing a threat to food security as they are produced from non-food-grade materials. However, despite the immense ecological benefits they present, inherent challenges in their production process hinder their commercial feasibility and integration into everyday use as they sustain high production costs, which are not incurred in the production of first-generation biofuels. These challenges are attributed to the pretreatment step required to disintegrate lignin and disrupt the crystalline structure of complex sugars present in lignocellulosic feedstock i.e. cellulose and hemicellulose, and the hydrolysis step required to break down these complex sugars into fermentable sugars that can be fermented into bioethanol fuel and other value-added products. There is hence a need for inexpensive and time-efficient methods to pretreat and hydrolyze lignocellulosic biomass to improve the economic feasibility of second-generation biofuels. Microbiome studies have continued to enrich our knowledge of complex microbial communities and reveal microorganisms capable of producing enzymes that can effectively degrade recalcitrant polymers and hydrolyze complex polysaccharides to readily utilizable forms.

This study aimed at identifying and functionally characterizing lignocellulosic-biomass degrading microorganisms and enzymes from the BSF larval gut microbiome as well as to formulate and implement a metatranscriptomics analysis pipeline to examine the effects dietary intervention had on the microbiome profiles of the BSF larvae.

Based on the ICP-MS analysis results in **Table 4.1**, two diets BSG (19.5%) and WH (23.3%) were found to possess the highest level of crude fiber, the dietary fraction that is associated with poorly digestible, recalcitrant components such as lignin, cellulose, and hemicellulose (Van Soest, 1966). The highest SRI (**Figure 4**), mean weight of larvae (**Figure 5**), mean larval lengths (**Figure 8**), and the fastest rates of pupation (**Figure 9**), were recorded in the control diet, CF, while the lowest were recorded in the WH dietary substrate. These were attributed to the high digestibility and low crude fiber content of the CF dietary substrate.

This study revealed that dietary intervention using substrates of varying lignocellulosic content had a significant effect on the microorganism profiles and in turn the functional profiles of the BSF larvae. Two of three genera that were hypothesized from preliminary BSF larvae microbiota studies (*Bacteroides* and *Dysgonomonas*)(Tanga et al., 2021; Jeon et al., 2011; Jiang et al., 2019) to be directly involved in the degradation of lignocellulosic substrates were found to be highly abundant in the two highly lignocellulosic metatranscriptomes, BSG and WH but not in the other metatranscriptomes (**Table 4.7**). These two genera belong to the taxonomic order Bacteroidales where about 40% of the PULs present in the PULDB database have been identified, and half of these PULs belong to the *Bacteroides* genus (Terrapon et al., 2018). From the functional analysis performed using the SEED Subsystems Hierarchical database, dodged bar plots in **Figure 26** show that most of the functions were highly enriched in the CF metatranscriptomes with a few exceptions, indicative of the high metabolic activity in the control diet. This could be attributed to the ease of digestibility of this diet.

Notably, also dominant in 4 of the 5 metatranscriptomes (BSG, CF, CM, and FM) were organisms from the genera *Sphingobacterium*. Sphingobacteria have been previously found in the guts of insects that feed on wood substrates (Schloss et al., 2006) and have been found to possess xylanases directly involved in hemicellulose degradation (Zhou et al., 2009). From PULDB (Terrapon et al., 2018), PUL hits associated with CAZy GH51 have been identified in some *Sphingobacterium sp.,* but after screening with dbCAN-PUL (Ausland et al., 2021), none of the PULs recorded in the WH sample were associated with bacteria from this genera. However, none of the PULs identified from the dbCAN-PUL and PULDB (Terrapon et al., 2018) databases for CAZy GH43 subfamily 16 found in the BSG sample were associated with *Sphingobacterium sp.* (**Table 4.9**). This could be because *Sphingobacterium sp.* identified from these metatranscriptomes were not organized into gene clusters responsible for the breakdown of complex polysaccharides, or because they were involved in executing other functions in these metatranscriptomes. Since the PUL resources are recent, another probable reason may be that PULs from Sphingobacteria are yet to be identified and curated. Sphingobacteria have been identified from some BSF studies to constitute the core gut microbiota (Yang et al., 2021) (Zhineng et al., 2021), (Tegtmeier et al., 2021), but in none of these studies were they been identified as the most dominant species.

The α-L-arabinofuranosidases (EC3.2.1.55) are exo-enzymes that cleave terminal α-(1,2), α-(1,3), or α-(1,5) linked L-arabinofuranosyl residues from hemicelluloses or oligosaccharides containing arabinose to liberate arabinofuranosides of arabinoxylan, arabinogalactan, and arabinan (McKee et al., 2012). In the CAZy database, these enzymes are classified into families GH2, GH3, GH43, GH51, GH54, and GH62 (Wilkens et al., 2017). From our samples, two of these GH classes (GH43 and GH51) were identified from the highly lignocellulosic BSG and WH metatranscriptomes respectively using the dbCAN2 Hotpep module (Busk et al., 2017). GH43 are more specific arabinofuranosidases and act on terminal α-1,5-linked arabinofuranosides while GH 51 cleave both α-1,2 and α-1,3 arabinofuranosyl moieties from xylans and arabinan (Margolles & De los Reyes-Gavilán, 2003). Complete degradation of hemicellulosic fractions of lignocellulosic biomass can therefore be achieved by the joint activity of these enzyme families. A study conducted by (Mewis et al., 2016) that characterized the GH43 enzymes into subfamilies, found carbohydrate-binding module 6 (CBM6) with an established function of binding to β-1,4-xylan and amorphous cellulose, to be highly prevalent in the GH43\_16 subfamily which was identified in the BSG metatranscriptome (**Table 4. 9**). This subfamily was also found to be multimodular, sharing some protein modules with other GH43 subfamilies potentially broadening its substrate specificity. CAZy class GH0 which was identified in the WH metatranscriptome is annotated as “Not classified” in the CAZy database. This CAZy class possesses enzymes with CAZyme functionality based on significant amino-acid similarity, but await biochemical characterization since they might display characteristics of multiple CAZy families (Lombard et al., 2014). Therefore, novel CAZy families could be identified from our data upon further biochemical characterization assays.

Arabinofuranosidases have been previously associated with genus *Bifidobacterium*, where they play crucial physiological roles in immunomodulation, colon cancer prevention, among other functions (Tannock, 1999). PULs PUL0395 and PUL0013 were identified in both the BSG and WH metatranscriptomes but were not associated with any organism clusters in their respective metatranscriptomes (**Table 4.10 b,c**). These PULs have been previously isolated from the L-arabinan utilization system of the *Geobacillus stearothermophilus* bacterium (Shulami et al., 2011). However, this does not rule out the presence of PULs responsible for the breakdown of complex polysaccharides and the production of the identified arabinofuranosidases (GH43 and GH51) in our metatranscriptome sequences, since the dbCAN-PUL database used, only focuses on CAZyme-containing PULs annotated from extant literature, and therefore novel or unannotated PULs from our data may have likely been missing from this resource (Ausland et al., 2021).

From taxonomic validation using 16S rRNA sequences, a subset of order Bacteroidales revealed genera *Coprobacter* and *Prevotella* as among the five most abundant genera in all dietary substrates (**Figure 18**). However, in metatranscriptomic analysis, genus *Prevotella* was only found to be among the most abundant genera in the highly lignocellulosic BSG diet, while genus *Coprobacter* was not identified among the four most abundant genera in any of the metatranscriptomes. Nonetheless, genera *Bacteroides* and *Dysgonomonas* that were abundant in the highly lignocellulosic BSG and WH metatranscriptomes, were also found to be among the most abundant in their respective 16S rRNA sequences. These additional bacterial genera identified from the 16S rRNA analysis were possibly representative of the inactive gut microbiota. These bacteria may have not been identified with metatranscriptomic analysis whose focus is on microbes with active functions in the microbial community. Other possible reasons behind this discrepancy may have arisen from the misclassification of short-length sequences that are highly similar, or due to sequencing and basecalling errors. Moreover, the 16S rRNA sequences were filtered before the error correction step, and no specific variable region of the 16S rRNA gene was being targeted by these filtered sequences (Jeong et al., 2021). Additionally, the 85% identity threshold set while filtering for the 16S rRNA sequences to rescue more sequences may have been too lenient and could have been a likely source of taxonomic misclassification. Taxonomic classification with 16S rRNA sequences has been reported to be more accurate with a high abundance of sequence reads and a high identity threshold (~99%) (Kaehler et al., 2019).

Ribodepletion is a highly essential step in metatranscriptomics analysis that was performed to remove the highly abundant rRNA reads from our data which constitute up to 90% of all sequences yet they do not contribute to key downstream transcriptomic analysis steps such as pathway characterization or differential expression studies (Shakya et al., 2019). A physical ribodepletion step, normally using ribodepletion kits, is highly encouraged before metatranscriptomics sequencing to remove rRNA sequences and ensure that higher numbers of mRNA reads are sequenced. Ribodepletion kits vary greatly in their efficiency. The best performing of these kits reduce up to ~80% of rRNA reads, and an additional in-silico ribodepletion step has to be incorporated (Westreich et al., 2018). Another inherent challenge encountered with most ribodepletion kits is that they are optimized towards specific commonly studied hosts i.e. mammalian, bovine, and murine species but very few of such resources exist for invertebrate species (N. Kumar et al., 2016). Despite ribodepletion resulting in the loss of some indicators on organism abundance within study metatranscriptomes, mRNA sequences can be used as a substitute for examining the overall organism prevalence within a microbial community. Additionally, rRNA reads have also been associated with higher incidences of false positives during organism annotation making mRNA sequences a more comprehensive measure of abundance, diversity, and prevalent functions (Westreich et al., 2016).

The SQK-PCB109 kit from ONT used during library preparation entailed a step that annealed most of the mRNA removing >50% of rRNA reads from all samples. From the in-silico ribodepletion step, the residual rRNA reads constituted 25.6 - 44% of the total reads (**Table 4.5**). Further obtaining non-host mRNA reads from the sequenced samples presented another challenge, since the vast majority of the mRNA reads sequenced were of host origin. Host mRNA reads were removed by mapping to the reference genome using the Minimap2 splice aware aligner, followed by obtaining the unmapped reads using Samtools. The mapped reads were used to evaluate the efficiency of the error correction step. Ultimately, the number of unmapped reads required for the organism and functional annotation had reduced significantly in comparison to the originally classified reads (**Table 4.7**). A study by (Penaranda & Hung, 2019) shows that bacterial cells can produce up to 100 times less total mRNA compared to their eukaryotic host. Therefore, in light of this finding, it would be highly efficient to employ sequencing approaches that generate high amounts of sequence data, or better still, sequence one sample multiple times to maximize the amount of bacterial mRNA generated.

Error correction had a significant impact on the accuracy and throughput of the reads as shown in **Figure 16.** The percentage of reads that mapped to the BSF reference was higher for the corrected reads across all the samples. This was validated using the paired t-test where the observed p-value, 2×10e8 was lower than the α value (0.05) indicative of statistical significance. However, other metrics like the coverage, mean depth, and mapping quality varied between the corrected and uncorrected reads (**Table 4.6**). This could be attributed to the IsONcorrect algorithm that is designed to handle highly variable read coverage and exon variation within reads even for low abundant transcripts. After de novo isoform clustering with isONclust (Sahlin & Medvedev, 2020), isONcorrect leverages the shared regions between transcripts by jointly using all the isoforms from a gene, enabling the correction of genes at low sequencing depths (Sahlin et al., 2021). This means that reads that could have been otherwise discarded are corrected and retained based on their common regions, regardless of their mean depths.

# 6.0 CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS

## 6.1 Conclusion

From this study, a metatranscriptomics analysis pipeline was designed for ONT-cDNA long-read sequences and implemented on reads sequenced from the gut of the BSF larvae with the ONT MinION MK1B platform. Dietary intervention using substrates of varying lignocellulose content was found to induce notable shifts in the microorganism profiles and ultimately, the CAZyme profiles.

Two of the three genera hypothesized to possess lignocellulolytic abilities, *Dysgonomonas* and *Bacteroides* were found to be among the most dominant in the two highly lignocellulosic diets, brewer’s spent grain (BSG) and water hyacinth (WH) from the metatranscriptomic analysis. After subsetting order Bacteroidales where most of the abundant genera from the metatranscriptomic analysis were classified, 16S rRNA analysis using ribodepleted sequences further revealed genera *Bacteroides* and *Dysgonomonas* were among the most abundant species in the highly lignocellulosic diets BSG and WH. Genus *Prevotella* which was identified from metatranscriptomic analysis as among the most abundant in the BSG diet was also identified as among the most abundant from the 16S rRNA analysis. However, genera *Coprobacter* and *Prevotella* which were identified as the most abundant from the 16S analysis were not among the highly abundant genera in the metatranscriptomic analysis. Since metatranscriptomic analysis only focused on microbial species with active functions in the microbiome, these genera likely represented the inactive gut microbiota.

Two families of α-L-arabinofuranosidases (EC3.2.1.55), GH43 and GH51 were identified from the highly lignocellulosic BSG and WH metatranscriptomes. The joint catalytic action of these two enzyme families is capable of the complete hydrolysis of hemicellulosic fractions of lignocellulosic biomass, rendering them highly useful in biotechnological applications. Therefore, the BSF gut microbiome could be used as a source of novel microbes that produce enzymes with lignocellulolytic activity that could be applied in, but not limited to, enzyme hydrolysis in second-generation biofuel production. Additionally, an error correction strategy that was adopted to improve the accuracy of ONT cDNA reads was found to have a significant impact on the overall mapping of reads to the BSF reference genome (p= 2×10e8) using a paired t-test.

## 6.2 Recommendations

From the findings of this study, it is recommended that;

1. A transcriptomics study focusing on the host transcriptome reads which are not required for metatranscriptomic analysis but account for most of the sequenced mRNA should be carried out to understand the effects of dietary intervention on host gene expression
2. Bioinformatics resources that are designed to study CAZy families should be further annotated since most of the existing resources only rely on CAZy families cited and annotated in literature
3. Metatranscriptomic analyses should be adopted as the first line of screening for active microbes and enzymes present in complex microbial communities before culturing, biochemical characterization assays, and metabolic engineering of strains and enzymes used in various biotechnological applications, since metagenomics and targeted sequencing approaches e.g. 16S rRNA (demonstrated in this study), also reveal inactive microbes in microbial communities
4. ‘Broad-spectrum’ enrichment kits capable of efficient bacterial mRNA enrichment from various hosts should be designed to reduce non-specific, untargeted sequencing of unwanted mRNA which is costly and time-consuming
5. Further microbiological and biochemical assays should be carried out to isolate and characterize the candidate microorganisms and arabinofuranosidase enzymes identified from the BSF larvae gut microbiome, and ultimately test their lignocellulolytic performance in various bioprocesses.

# REFERENCES

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., Gnirke, A., Oyola, S. O., Otto, T. D., Gu, Y., Maslen, G., Manske, M., Campino, S., Turner, D. J., MacInnis, B., Kwiatkowski, D. P., Swerdlow, H. P., … Jaffe, D. B. (2012). Characterizing and measuring bias in sequence data. *Genome Biology*, *02*(5), 1. http://genomebiology.com/2013/14/5/R51%5Cnhttp://www.biomedcentral.com/1471-2164/13/1%5Cnhttp://genomebiology.com/2011/12/2/R18

Amin, F. R., Khalid, H., Zhang, H., Rahman, S., Zhang, R., Liu, G., & Chen, C. (2017). Pretreatment methods of lignocellulosic biomass for anaerobic digestion. In *AMB Express* (Vol. 7, Issue 1, p. 72). Springer Verlag. https://doi.org/10.1186/s13568-017-0375-4

Antonov, A., Ivanov, G., Pastukhova, N., & Bovykina, G. (2019). Production of chitin from dead Hermetia Illucens. *IOP Conference Series: Earth and Environmental Science*, *315*(4). https://doi.org/10.1088/1755-1315/315/4/042003

Antunes, F. A. F., Chandel, A. K., Terán-Hilares, R., Ingle, A. P., Rai, M., dos Santos Milessi, T. S., da Silva, S. S., & dos Santos, J. C. (2019). Overcoming challenges in lignocellulosic biomass pretreatment for second-generation (2G) sugar production: emerging role of nano, biotechnological and promising approaches. *3 Biotech*, *9*(6). https://doi.org/10.1007/s13205-019-1761-1

*AR5 Climate Change 2014: Mitigation of Climate Change — IPCC*. (n.d.). Retrieved November 4, 2020, from https://www.ipcc.ch/report/ar5/wg3/

Au, K. F., Underwood, J. G., Lee, L., & Wong, W. H. (2012). Improving PacBio long read accuracy by short read alignment. *PloS One*, *7*(10). https://doi.org/10.1371/JOURNAL.PONE.0046679

Ausland, C., Zheng, J., Yi, H., Yang, B., Li, T., Feng, X., Zheng, B., & Yin, Y. (2021). dbCAN-PUL: A database of experimentally characterized CAZyme gene clusters and their substrates. *Nucleic Acids Research*, *49*(D1), D523–D528. https://doi.org/10.1093/nar/gkaa742

Badiei, M., Asim, N., Jahim, J. M., & Sopian, K. (2014). Comparison of Chemical Pretreatment Methods for Cellulosic Biomass. *APCBEE Procedia*, *9*(Icbee 2013), 170–174. https://doi.org/10.1016/j.apcbee.2014.01.030

Balan, V. (2014). Current Challenges in Commercially Producing Biofuels from Lignocellulosic Biomass. *ISRN Biotechnology*, *2014*(i), 1–31. https://doi.org/10.1155/2014/463074

Barragan-Fonseca, K. B., Dicke, M., & van Loon, J. J. A. (2017). Nutritional value of the black soldier fly (Hermetia illucens L.) and its suitability as animal feed - a review. *Journal of Insects as Food and Feed*, *3*(2), 105–120. https://doi.org/10.3920/JIFF2016.0055

Baruah, J., Nath, B. K., Sharma, R., Kumar, S., Deka, R. C., Baruah, D. C., & Kalita, E. (2018). Recent trends in the pretreatment of lignocellulosic biomass for value-added products. *Frontiers in Energy Research*, *6*(DEC), 141. https://doi.org/10.3389/FENRG.2018.00141/BIBTEX

Bashiardes, S., Zilberman-Schapira, G., & Elinav, E. (2016). Use of metatranscriptomics in microbiome research. *Bioinformatics and Biology Insights*, *10*, 19–25. https://doi.org/10.4137/BBI.S34610

Berlin, K., Koren, S., Chin, C. S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, *33*(6), 623–630. https://doi.org/10.1038/NBT.3238

Bignert, A., Eriksson, U., Nyberg, E., Miller, A., & Danielsson, S. (2014). Consequences of using pooled versus individual samples for designing environmental monitoring sampling strategies. *Chemosphere*, *94*, 177–182. https://doi.org/10.1016/j.chemosphere.2013.09.096

Bost, A., Franzenburg, S., Adair, K. L., Martinson, V. G., Loeb, G., & Douglas, A. E. (2018). How gut transcriptional function of Drosophila melanogaster varies with the presence and composition of the gut microbiota. *Molecular Ecology*, *27*(8), 1848–1859. https://doi.org/10.1111/MEC.14413

*Brock Biology of Microorganisms* (11th ed.). (2005). Prentice Hall.

Bruno, D., Bonelli, M., De Filippis, F., Di Lelio, I., Tettamanti, G., Casartelli, M., Ercolini, D., & Caccia, S. (2019). The intestinal microbiota of Hermetia illucens larvae is affected by diet and shows a diverse composition in the different midgut regions. *Applied and Environmental Microbiology*, *85*(2). https://doi.org/10.1128/AEM.01864-18

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods 2021 18:4*, *18*(4), 366–368. https://doi.org/10.1038/s41592-021-01101-x

Busk, P. K., Pilgaard, B., Lezyk, M. J., Meyer, A. S., & Lange, L. (2017). Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. *BMC Bioinformatics*, *18*(1), 1–9. https://doi.org/10.1186/S12859-017-1625-9/TABLES/4

Canabarro, N., Soares, J. F., Anchieta, C. G., Kelling, C. S., & Mazutti, M. A. (2013). Thermochemical processes for biofuels production from biomass. *Sustainable Chemical Processes 2013 1:1*, *1*(1), 1–10. https://doi.org/10.1186/2043-7129-1-22

Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials*, *2*(1), 31–49. https://doi.org/10.1016/0197-2456(81)90056-8

Chia, S. Y., Tanga, C. M., Osuga, I. M., Cheseto, X., Ekesi, S., Dicke, M., & van Loon, J. J. A. (2020). Nutritional composition of black soldier fly larvae feeding on agro-industrial by-products. *Entomologia Experimentalis et Applicata*, *168*(6–7), 472–481. https://doi.org/10.1111/eea.12940

Chung, Y. W., Gwak, H. J., Moon, S., Rho, M., & Ryu, J. H. (2020). Functional dynamics of bacterial species in the mouse gut microbiome revealed by metagenomic and metatranscriptomic analyses. *PLoS ONE*, *15*(1), 1–19. https://doi.org/10.1371/journal.pone.0227886

Čičková, H., Newton, G. L., Lacy, R. C., & Kozánek, M. (2015). The use of fly larvae for organic waste treatment. *Waste Management*, *35*, 68–80. https://doi.org/10.1016/j.wasman.2014.09.026

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., & Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(Database issue), D633. https://doi.org/10.1093/NAR/GKT1244

Contesini, F. J., Frandsen, R. J. N., & Damasio, A. (2021). Editorial: CAZymes in Biorefinery: From Genes to Application. *Frontiers in Bioengineering and Biotechnology*, *9*, 86. https://doi.org/10.3389/FBIOE.2021.622817/BIBTEX

Dale, B. E., Anderson, J. E., Brown, R. C., Csonka, S., Dale, V. H., Herwick, G., Jackson, R. D., Jordan, N., Kaffka, S., Kline, K. L., Lynd, L. R., Malmstrom, C., Ong, R. G., Richard, T. L., Taylor, C., & Wang, M. Q. (2014). Take a closer look: Biofuels can support environmental, economic and social goals. In *Environmental Science and Technology* (Vol. 48, Issue 13, pp. 7200–7203). American Chemical Society. https://doi.org/10.1021/es5025433

De Smet, J., Wynants, E., Cos, P., & Van Campenhout, L. (2018). Microbial community dynamics during rearing of black soldier fly larvae (Hermetia illucens) and impact on exploitation potential. In *Applied and Environmental Microbiology* (Vol. 84, Issue 9). American Society for Microbiology. https://doi.org/10.1128/AEM.02722-17

Dos Santos, C. R., De Giuseppe, P. O., De Souza, F. H. M., Zanphorlin, L. M., Domingues, M. N., Pirolla, R. A. S., Honorato, R. V., Tonoli, C. C. C., De Morais, M. A. B., De Matos Martins, V. P., Fonseca, L. M., Büchli, F., De Oliveira, P. S. L., Gozzo, F. C., & Murakami, M. T. (2018). The mechanism by which a distinguishing arabinofuranosidase can cope with internal di-substitutions in arabinoxylans. *Biotechnology for Biofuels*, *11*(1), 1–19. https://doi.org/10.1186/S13068-018-1212-Y/TABLES/2

Edgar, R. C., & Bateman, A. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/BIOINFORMATICS/BTQ461

EPA. (n.d.). *Overview for Renewable Fuel Standard*. Retrieved December 9, 2021, from https://www.epa.gov/renewable-fuel-standard-program/overview-renewable-fuel-standard

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/BIOINFORMATICS/BTW354

Fu, S., Wang, A., & Au, K. F. (2019). A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biology*, *20*(1), 1–17. https://doi.org/10.1186/s13059-018-1605-z

Grünberger, F., Ferreira-Cerca, S., & Grohmann, D. (2021). Nanopore sequencing of RNA and cDNA molecules expands the transcriptomic toolbox in prokaryotes. *BioRxiv*, 2021.06.14.448286. https://doi.org/10.1101/2021.06.14.448286

Guerriero, G., Hausman, J. F., Strauss, J., Ertan, H., & Siddiqui, K. S. (2016). Lignocellulosic biomass: Biosynthesis, degradation, and industrial utilization. *Engineering in Life Sciences*, *16*(1), 1–16. https://doi.org/10.1002/ELSC.201400196

Hackl, T., Hedrich, R., Schultz, J., & Förster, F. (2014). proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics (Oxford, England)*, *30*(21), 3004–3011. https://doi.org/10.1093/BIOINFORMATICS/BTU392

Henson, J., Tischler, G., & Ning, Z. (2012). Next-generation sequencing and large genome assemblies. *Pharmacogenomics*, *13*(8), 901–915. https://doi.org/10.2217/pgs.12.72

Heussler, C. D., Walter, A., Oberkofler, H., Insam, H., Arthofer, W., Schlick-Steiner, B. C., & Steiner, F. M. (2019). Correction: Influence of three artificial light sources on oviposition and half-life of the Black Soldier Fly, Hermetia illucens (Diptera: Stratiomyidae): Improving small-scale indoor rearing (PLoS ONE (2018) 13:5 (e0197896) DOI: 10.1371/journal.pone.0197. *PLoS ONE*, *14*(12), 40–45. https://doi.org/10.1371/journal.pone.0226670

Himmel, M. E., Ding, S. Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W., & Foust, T. D. (2007). Biomass recalcitrance: Engineering plants and enzymes for biofuels production. In *Science* (Vol. 315, Issue 5813, pp. 804–807). Science. https://doi.org/10.1126/science.1137016

Hoang, N. V., Furtado, A., Mason, P. J., Marquardt, A., Kasirajan, L., Thirugnanasambandam, P. P., Botha, F. C., & Henry, R. J. (2017). A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. *BMC Genomics*, *18*(1), 1–22. https://doi.org/10.1186/s12864-017-3757-8

Høj, L., Levy, N., Baillie, B. K., Clode, P. L., Strohmaier, R. C., Siboni, N., Webster, N. S., Uthicke, S., & Bourne, D. G. (2018). Crown-of-thorns sea star Acanthaster cf. solaris has tissue-characteristic microbiomes with potential roles in health and reproduction. *Applied and Environmental Microbiology*, *84*(13), 181–199. https://doi.org/10.1128/AEM.00181-18/SUPPL\_FILE/ZAM013188596S1.PDF

Huang, L., Zhang, H., Wu, P., Entwistle, S., Li, X., Yohe, T., Yi, H., Yang, Z., & Yin, Y. (2018). dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. *Nucleic Acids Research*, *46*(D1), D516–D521. https://doi.org/10.1093/NAR/GKX894

*ISOLATE II RNA Mini Kit | Bioline | Meridian Bioscience*. (n.d.). Retrieved November 14, 2020, from https://www.bioline.com/isolate-ii-rna-mini-kit.html

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 1–11. https://doi.org/10.1186/s13059-016-1103-0

Jeong, J., Yun, K., Mun, S., Chung, W. H., Choi, S. Y., Nam, Y. do, Lim, M. Y., Hong, C. P., Park, C. H., Ahn, Y., & Han, K. (2021). The effect of taxonomic classification by full-length 16S rRNA sequencing with a synthetic long-read technology. *Scientific Reports 2021 11:1*, *11*(1), 1–12. https://doi.org/10.1038/s41598-020-80826-9

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., & Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology*, *21*(1), 1–24. https://doi.org/10.1186/S13059-020-02107-Y/FIGURES/5

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*(2), 363–375. https://doi.org/10.1111/J.2006.0030-1299.14714.X

Kaehler, B. D., Bokulich, N. A., McDonald, D., Knight, R., Caporaso, J. G., & Huttley, G. A. (2019). Species abundance information improves sequence taxonomy classification accuracy. *Nature Communications 2019 10:1*, *10*(1), 1–10. https://doi.org/10.1038/s41467-019-12669-6

Karst, S. M., Ziels, R. M., Kirkegaard, R. H., Sørensen, E. A., McDonald, D., Zhu, Q., Knight, R., & Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nature Methods 2021 18:2*, *18*(2), 165–169. https://doi.org/10.1038/s41592-020-01041-y

Khamis, F. M., Ombura, F. L. O., Akutse, K. S., Subramanian, S., Mohamed, S. A., Fiaboe, K. K. M., Saijuntha, W., Van Loon, J. J. A., Dicke, M., Dubois, T., Ekesi, S., & Tanga, C. M. (2020). Insights in the Global Genetics and Gut Microbiome of Black Soldier Fly, Hermetia illucens: Implications for Animal Feed Safety Control. *Frontiers in Microbiology*, *11*, 1538. https://doi.org/10.3389/fmicb.2020.01538

Kim, Y. B., Kim, D. H., Jeong, S. B., Lee, J. W., Kim, T. H., Lee, H. G., & Lee, K. W. (2020). Black soldier fly larvae oil as an alternative fat source in broiler nutrition. *Poultry Science*, *99*(6), 3133–3143. https://doi.org/10.1016/J.PSJ.2020.01.018

Klammsteiner, T., Walter, A., Bogataj, T., Heussler, C. D., Stres, B., Steiner, F. M., Schlick-Steiner, B. C., Arthofer, W., & Insam, H. (2020). The Core Gut Microbiome of Black Soldier Fly (Hermetia illucens) Larvae Raised on Low-Bioburden Diets. *Frontiers in Microbiology*, *11*(May), 1–14. https://doi.org/10.3389/fmicb.2020.00993

Klingenberg, H., & Meinicke, P. (2017). How to normalize metatranscriptomic count data for differential expression analysis. *PeerJ*, *5*(10). https://doi.org/10.7717/PEERJ.3859

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217. https://doi.org/10.1093/BIOINFORMATICS/BTS611

Krams, I. A., Kecko, S., Jõers, P., Trakimas, G., Elferts, D., Krams, R., Luoto, S., Rantala, M. J., Inashkina, I., Gudrā, D., Fridmanis, D., Contreras-Garduño, J., Grantina-Ievina, L., & Krama, T. (2017). Microbiome symbionts and diet diversity incur costs on the immune system of insect larvae. *Journal of Experimental Biology*, *220*(22), 4204–4212. https://doi.org/10.1242/JEB.169227/262697/AM/MICROBIOME-SYMBIONTS-AND-DIET-DIVERSITY-INCUR

Kumar, N., Lin, M., Zhao, X., Ott, S., Santana-Cruz, I., Daugherty, S., Rikihisa, Y., Sadzewicz, L., Tallon, L. J., Fraser, C. M., & Dunning Hotopp, J. C. (2016). Efficient Enrichment of Bacterial mRNA from Host-Bacteria Total RNA Samples. *Scientific Reports 2016 6:1*, *6*(1), 1–10. https://doi.org/10.1038/srep34850

Kumar, P., Barrett, D. M., Delwiche, M. J., & Stroeve, P. (2009). Methods for pretreatment of lignocellulosic biomass for efficient hydrolysis and biofuel production. *Industrial and Engineering Chemistry Research*, *48*(8), 3713–3729. https://doi.org/10.1021/ie801542g

Lalander, C., Diener, S., Zurbrügg, C., & Vinnerås, B. (2019). Effects of feedstock on larval development and process efficiency in waste treatment with black soldier fly (Hermetia illucens). *Journal of Cleaner Production*, *208*, 211–219. https://doi.org/10.1016/j.jclepro.2018.10.017

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology 2013 31:9*, *31*(9), 814–821. https://doi.org/10.1038/nbt.2676

Laver, T., Harrison, J., O’Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, *3*, 1–8. https://doi.org/10.1016/j.bdq.2015.02.001

Lederberg, B. J., & McCray, A. T. (2001). ’ Ome Sweet ’ Omics-- A Genealogical Treasury of Words. *The Scientist*, *15*(7), 8.

LEDERBERG, J., & MCCRAY, A. T. (2001). `Ome Sweet `Omics--A Genealogical Treasury of Words. *The Scientist*, *15*(7), 8–8. https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=08903670&v=2.1&it=r&id=GALE%7CA73535513&sid=googleScholar&linkaccess=fulltext

Lee, C. M., Kim, S. Y., Song, J., Lee, Y. S., Sim, J. S., & Hahn, B. S. (2018). Isolation and characterization of a halotolerant and protease-resistant α-galactosidase from the gut metagenome of Hermetia illucens. *Journal of Biotechnology*, *279*, 47–54. https://doi.org/10.1016/j.jbiotec.2018.05.003

Lee, H. C. (2018). Review of inductively coupled plasmas: Nano-applications and bistable hysteresis physics. *Applied Physics Reviews*, *5*(1), 011108. https://doi.org/10.1063/1.5012001

Leger, A., & Leonardi, T. (2019). pycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open Source Software*, *4*(34), 1236. https://doi.org/10.21105/JOSS.01236

Leidenfrost, R. M., Pöther, D.-C., Jäckel, U., & Wünschiers, R. (2020). Benchmarking the MinION: Evaluating long reads for microbial profiling. *Scientific Reports*, *10*(1). https://doi.org/10.1038/S41598-020-61989-X

Li, F., Hitch, T. C. A., Chen, Y., Creevey, C. J., & Guan, L. L. (2019). Comparative metagenomic and metatranscriptomic analyses reveal the breed effect on the rumen microbiome and its associations with feed efficiency in beef cattle 06 Biological Sciences 0604 Genetics 06 Biological Sciences 0605 Microbiology. *Microbiome*, *7*(1), 6. https://doi.org/10.1186/s40168-019-0618-5

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/BIOINFORMATICS/BTY191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/BIOINFORMATICS/BTP352

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., & Fan, W. (2012). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, *11*(1), 25–37. https://doi.org/10.1093/BFGP/ELR035

Limayem, A., & Ricke, S. C. (2012). Lignocellulosic biomass for bioethanol production: Current perspectives, potential issues and future prospects. *Progress in Energy and Combustion Science*, *38*(4), 449–467. https://doi.org/10.1016/j.pecs.2012.03.002

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, *42*(Database issue), D490. https://doi.org/10.1093/NAR/GKT1178

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology 2014 15:12*, *15*(12), 1–21. https://doi.org/10.1186/S13059-014-0550-8

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics and Bioinformatics*, *14*(5), 265–279. https://doi.org/10.1016/j.gpb.2016.05.004

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., & Woolf, P. J. (2009). GAGE: Generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, *10*(1), 1–17. https://doi.org/10.1186/1471-2105-10-161/FIGURES/5

Lynd, L. R. (2017). The grand challenge of cellulosic biofuels. In *Nature Biotechnology* (Vol. 35, Issue 10, pp. 912–915). Nature Publishing Group. https://doi.org/10.1038/nbt.3976

Madhavan, A., Sindhu, R., Parameswaran, B., Sukumaran, R. K., & Pandey, A. (2017). Metagenome Analysis: a Powerful Tool for Enzyme Bioprospecting. *Applied Biochemistry and Biotechnology*, *183*(2), 636–651. https://doi.org/10.1007/s12010-017-2568-3

Magi, A., Giusti, B., & Tattini, L. (2017). Characterization of MinION nanopore data for resequencing analyses. *Briefings in Bioinformatics*, *18*(6), 940–953. https://doi.org/10.1093/bib/bbw077

Margolles, A., & De los Reyes-Gavilán, C. G. (2003). Purification and Functional Characterization of a Novel α-l-Arabinofuranosidase from Bifidobacterium longum B667. *Applied and Environmental Microbiology*, *69*(9), 5096. https://doi.org/10.1128/AEM.69.9.5096-5103.2003

Martinez, X., Pozuelo, M., Pascal, V., Campos, D., Gut, I., Gut, M., Azpiroz, F., Guarner, F., & Manichanh, C. (2016). MetaTrans: An open-source pipeline for metatranscriptomics. *Scientific Reports*, *6*, 1–12. https://doi.org/10.1038/srep26447

McDonald, J. H. (2014). Handbook of Biological Statistics - Paired t–test. *Sparky House Publishing*, 180–185. http://www.biostathandbook.com/pairedttest.html

McKee, L. S., Peña, M. J., Rogowski, A., Jackson, A., Lewis, R. J., York, W. S., Krogh, K. B. R. M., Viksø-Nielsen, A., Skjøt, M., Gilbert, H. J., & Marles-Wright, J. (2012). Introducing endo-xylanase activity into an exo-acting arabinofuranosidase that targets side chains. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(17), 6537–6542. https://doi.org/10.1073/PNAS.1117686109/-/DCSUPPLEMENTAL

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE*, *8*(4), e61217. https://doi.org/10.1371/JOURNAL.PONE.0061217

Meneguz, M., Gasco, L., & Tomberlin, J. K. (2018). Impact of pH and feeding system on black soldier fly (Hermetia illucens, L; Diptera: Stratiomyidae) larval development. *PLOS ONE*, *13*(8), e0202591. https://doi.org/10.1371/JOURNAL.PONE.0202591

Mewis, K., Lenfant, N., Lombard, V., & Henrissat, B. (2016). Dividing the large glycoside hydrolase family 43 into subfamilies: A motivation for detailed enzyme characterization. *Applied and Environmental Microbiology*, *82*(6), 1686–1692. https://doi.org/10.1128/AEM.03453-15/SUPPL\_FILE/ZAM999116989SO1.PDF

Miclotte, G., Heydari, M., Demeester, P., Rombauts, S., Van de Peer, Y., Audenaert, P., & Fostier, J. (2016). Jabba: Hybrid error correction for long sequencing reads. *Algorithms for Molecular Biology*, *11*(1), 1–12. https://doi.org/10.1186/S13015-016-0075-7/FIGURES/5

Mishima, D., Kuniki, M., Sei, K., Soda, S., Ike, M., & Fujita, M. (2008). Ethanol production from candidate energy crops: Water hyacinth (Eichhornia crassipes) and water lettuce (Pistia stratiotes L.). *Bioresource Technology*, *99*(7), 2495–2500. https://doi.org/10.1016/J.BIORTECH.2007.04.056

MS Swaminathan, Maryam Rahmanian, John Wilkinson, V. G. (2013). HPLE, 2013 Biofuels and food. In *Biofuels and Food Security* (Issue 5).

*nanoporetech/pychopper: A tool to identify, orient, trim and rescue full length cDNA reads*. (n.d.). Retrieved October 25, 2021, from https://github.com/nanoporetech/pychopper

Nguyen, T. T. X., Tomberlin, J. K., & Vanlaerhoven, S. (2015). Ability of Black Soldier Fly (Diptera: Stratiomyidae) Larvae to Recycle Food Waste. *Environmental Entomology*, *44*(2), 406–410. https://doi.org/10.1093/ee/nvv002

Nomanbhay, S. M., Hussain, R., & Palanisamy, K. (2013). Microwave-Assisted Alkaline Pretreatment and Microwave Assisted Enzymatic Saccharification of Oil Palm Empty Fruit Bunch Fiber for Enhanced Fermentable Sugar Yield. *Journal of Sustainable Bioenergy Systems*, *03*(01), 7–17. https://doi.org/10.4236/jsbs.2013.31002

Nuelle, B. (2019). *Cellulosic ethanol struggles to climb commercialization ladder| Agri-Pulse Communications, Inc.* Agri-Pulse. https://www.agri-pulse.com/articles/12894-cellulosic-ethanol-struggles-to-climb-commercialization-ladder

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-385

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., & Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, *42*(Database issue), D206. https://doi.org/10.1093/NAR/GKT1226

Oxford Nanopore Technologies, O. (n.d.). *Flow Cell (R9.4.1)*. 2020. Retrieved October 22, 2021, from https://store.nanoporetech.com/flow-cell-r9-4-1.html

Penaranda, C., & Hung, D. T. (2019). Single-Cell RNA Sequencing to Understand Host–Pathogen Interactions. *ACS Infectious Diseases*, *5*(3), 336–344. https://doi.org/10.1021/ACSINFECDIS.8B00369

Peters, B. A., Wilson, M., Moran, U., Pavlick, A., Izsak, A., Wechter, T., Weber, J. S., Osman, I., & Ahn, J. (2019). Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients. *Genome Medicine*, *11*(1), 1–14. https://doi.org/10.1186/S13073-019-0672-4/FIGURES/5

Prange, A., & Pröfrock, D. (2012). Inductively Coupled Plasma-Mass Spectrometry (ICP-MS) for Quantitative Analysis in Environmental and Life Sciences: A Review of Challenges, Solutions, and Trends. *Applied Spectroscopy, Vol. 66, Issue 8, Pp. 843-868*, *66*(8), 843–868. https://doi.org/10.1366/12-06681

Qian, X. B., Chen, T., Xu, Y. P., Chen, L., Sun, F. X., Lu, M. P., & Liu, Y. X. (2020). A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. *Chinese Medical Journal*, *133*(15), 1844–1855. https://doi.org/10.1097/CM9.0000000000000871

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590. https://doi.org/10.1093/NAR/GKS1219

*R10.3: the newest nanopore for high accuracy nanopore sequencing – now available in store*. (n.d.). Retrieved October 25, 2021, from https://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., Merico, D., & Bader, G. (2017). Pathway enrichment analysis of - omics data. *BioRxiv*, 232835. https://doi.org/10.1101/232835

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47. https://doi.org/10.1093/NAR/GKV007

Robak, K., & Balcerek, M. (2018). Review of second generation bioethanol production from residual biomass. In *Food Technology and Biotechnology* (Vol. 56, Issue 2, pp. 174–187). University of Zagreb. https://doi.org/10.17113/ftb.56.02.18.5428

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139. https://doi.org/10.1093/BIOINFORMATICS/BTP616

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*(10). https://doi.org/10.7717/PEERJ.2584

Rothe, M., & Blaut, M. (2013). Evolution of the gut microbiota and the influence of diet. *Beneficial Microbes*, *4*(1), 31–37. https://doi.org/10.3920/BM2012.0029

Sahlin, K., & Mäkinen, V. (2021). Accurate spliced alignment of long RNA sequencing reads. *Bioinformatics*. https://doi.org/10.1093/BIOINFORMATICS/BTAB540

Sahlin, K., & Medvedev, P. (2020). De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm. *Https://Home.Liebertpub.Com/Cmb*, *27*(4), 472–484. https://doi.org/10.1089/CMB.2019.0299

Sahlin, K., Sipos, B., James, P. L., & Medvedev, P. (2021). Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nature Communications*, *12*(1). https://doi.org/10.1038/S41467-020-20340-8

Salmela, L., & Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics (Oxford, England)*, *30*(24), 3506–3514. https://doi.org/10.1093/BIOINFORMATICS/BTU538

Samayoa, A. C., Chen, W. T., & Hwang, S. Y. (2016a). Survival and development of Hermetia illucens (Diptera: Stratiomyidae): A biodegradation agent of organic waste. *Journal of Economic Entomology*, *109*(6), 2580–2585. https://doi.org/10.1093/jee/tow201

Samayoa, A. C., Chen, W. T., & Hwang, S. Y. (2016b). Survival and development of Hermetia illucens (Diptera: Stratiomyidae): A biodegradation agent of organic waste. *Journal of Economic Entomology*, *109*(6), 2580–2585. https://doi.org/10.1093/jee/tow201

Schloss, P. D., Delalibera, I., Handelsman, J., & Raffa, K. F. (2006). Bacteria associated with the guts of two wood-boring beetles: Anoplophora glabripennis and Saperda vestita (Cerambycidae). *Environmental Entomology*, *35*(3), 625–629. https://doi.org/10.1603/0046-225X-35.3.625

Seshadri, R., Leahy, S. C., Attwood, G. T., Teh, K. H., Lambie, S. C., Cookson, A. L., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Hadjithomas, M., Varghese, N. J., Paez-Espino, D., Perry, R., Henderson, G., Creevey, C. J., Terrapon, N., Lapebie, P., Drula, E., Lombard, V., Rubin, E., … Cerón Cucchi, M. (2018). Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nature Biotechnology*, *36*(4), 359–367. https://doi.org/10.1038/nbt.4110

Setia, M. (2016). Methodology Series Module 5: Sampling Strategies. *Indian Journal of Dermatology*, *61*(5), 505. https://doi.org/10.4103/0019-5154.190118

Shakya, M., Lo, C. C., & Chain, P. S. G. (2019). Advances and challenges in metatranscriptomic analysis. *Frontiers in Genetics*, *10*(SEP), 904. https://doi.org/10.3389/FGENE.2019.00904/BIBTEX

Shelomi, M., Wu, M. K., Chen, S. M., Huang, J. J., & Burke, C. G. (2020). Microbes Associated With Black Soldier Fly (Diptera: Stratiomiidae) Degradation of Food Waste. *Environmental Entomology*, *49*(2), 405–411. https://doi.org/10.1093/ee/nvz164

Sheppard, D. C., Tomberlin, J. K., Joyce, J. A., Kiser, B. C., & Sumner, S. M. (2002). Rearing Methods for the Black Soldier Fly (Diptera: Stratiomyidae): Table 1. *Journal of Medical Entomology*, *39*(4), 695–698. https://doi.org/10.1603/0022-2585-39.4.695

Shulami, S., Raz-Pasteur, A., Tabachnikov, O., Gilead-Gropper, S., Shner, I., & Shoham, Y. (2011). The L-Arabinan utilization system of Geobacillus stearothermophilus. *Journal of Bacteriology*, *193*(11), 2838–2850. https://doi.org/10.1128/JB.00222-11

Shumo, M., Osuga, I. M., Khamis, F. M., Tanga, C. M., Fiaboe, K. K. M., Subramanian, S., Ekesi, S., van Huis, A., & Borgemeister, C. (2019). The nutritive value of black soldier fly larvae reared on common organic waste streams in Kenya. *Scientific Reports 2019 9:1*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-46603-z

Srivastava, N., Mishra, P. K., & Upadhyay, S. N. (2020). Significance of lignocellulosic biomass waste in the biofuel production process. *Industrial Enzymes for Biofuels Production*, 1–18. https://doi.org/10.1016/B978-0-12-821010-9.00001-2

Stiller, M., Knapp, M., Stenzel, U., Hofreiter, M., & Meyer, M. (2009). Direct multiplex sequencing (DMPS)—a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. *Genome Research*, *19*(10), 1843. https://doi.org/10.1101/GR.095760.109

Tanga, C. M., Waweru, J. W., Tola, Y. H., Onyoni, A. A., Khamis, F. M., Ekesi, S., & Paredes, J. C. (2021). Organic Waste Substrates Induce Important Shifts in Gut Microbiota of Black Soldier Fly (Hermetia illucens L.): Coexistence of Conserved, Variable, and Potential Pathogenic Microbes. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.635881

Tannock, G. W. (1999). Analysis of the intestinal microflora: a renaissance. *Antonie van Leeuwenhoek 1999 76:1*, *76*(1), 265–278. https://doi.org/10.1023/A:1002038308506

Tatusova, T., Ciufo, S., Fedorov, B., O’Neill, K., & Tolstoy, I. (2014). RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Research*, *42*(Database issue). https://doi.org/10.1093/NAR/GKT1274

Tegtmeier, D., Hurka, S., Mihajlovic, S., Bodenschatz, M., Schlimbach, S., & Vilcinskas, A. (2021). Culture-independent and culture-dependent characterization of the black soldier fly gut microbiome reveals a large proportion of culturable bacteria with potential for industrial applications. *Microorganisms*, *9*(8). https://doi.org/10.3390/microorganisms9081642

Terrapon, N., Lombard, V., Drula, É., Lapébie, P., Al-Masaudi, S., Gilbert, H. J., & Henrissat, B. (2018). PULDB: the expanded database of Polysaccharide Utilization Loci. *Nucleic Acids Research*, *46*(D1), D677–D683. https://doi.org/10.1093/NAR/GKX1022

Terrapon, N., Lombard, V., Gilbert, H. J., & Henrissat, B. (2015). Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics*, *31*(5), 647–655. https://doi.org/10.1093/BIOINFORMATICS/BTU716

Tingley, J. P., Low, K. E., Xing, X., & Abbott, D. W. (2021). Combined whole cell wall analysis and streamlined in silico carbohydrate-active enzyme discovery to improve biocatalytic conversion of agricultural crop residues. *Biotechnology for Biofuels*, *14*(1), 1–19. https://doi.org/10.1186/S13068-020-01869-8/FIGURES/5

Van Soest, P. J. (1966). Nonnutritive Residues: A System of Analysis for the Replacement of Crude Fiber. *Journal of AOAC INTERNATIONAL*, *49*(3), 546–551. https://doi.org/10.1093/JAOAC/49.3.546

Wagner, B. D., Grunwald, G. K., Zerbe, G. O., Mikulich-Gilbertson, S. K., Robertson, C. E., Zemanick, E. T., & Harris, J. K. (2018). On the use of diversity measures in longitudinal sequencing studies of microbial communities. *Frontiers in Microbiology*, *9*(MAY), 1037. https://doi.org/10.3389/FMICB.2018.01037/BIBTEX

Wang, C., Guo, X., Deng, H., Dong, D., Tu, Q., & Wu, W. (2014). New insights into the structure and dynamics of actinomycetal community during manure composting. *Applied Microbiology and Biotechnology*, *98*(7), 3327–3337. https://doi.org/10.1007/s00253-013-5424-6

Webb E. C. (1993). Enzyme nomenclature: Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Pp 862. Academic Press, San Diego. 1992 ISBN 0-12-227165-3. *Biochemical Education*, *21*(2), 102–102. https://doi.org/10.1016/0307-4412(93)90058-8

Werren, J. H., Zhang, W., & Guo, L. R. (1995). Evolution and phylogeny of *Wolbachia* : reproductive parasites of arthropods. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *261*(1360), 55–63. https://doi.org/10.1098/rspb.1995.0117

Westreich, S. T., Korf, I., Mills, D. A., & Lemay, D. G. (2016). SAMSA: a comprehensive metatranscriptome analysis pipeline. *BMC Bioinformatics*, *17*(1). https://doi.org/10.1186/S12859-016-1270-8

Westreich, S. T., Treiber, M. L., Mills, D. A., Korf, I., & Lemay, D. G. (2018). SAMSA2: a standalone metatranscriptome analysis pipeline. *BMC Bioinformatics 2018 19:1*, *19*(1), 1–11. https://doi.org/10.1186/S12859-018-2189-Z

Wexler, H. M. (2007). Bacteroides: The good, the bad, and the nitty-gritty. In *Clinical Microbiology Reviews* (Vol. 20, Issue 4, pp. 593–621). https://doi.org/10.1128/CMR.00008-07

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics*, *3*(10), e000132. https://doi.org/10.1099/MGEN.0.000132

Wilkens, C., Andersen, S., Dumon, C., Berrin, J. G., & Svensson, B. (2017). GH62 arabinofuranosidases: Structure, function and applications. *Biotechnology Advances*, *35*(6), 792–804. https://doi.org/10.1016/J.BIOTECHADV.2017.06.005

Williams, C. R., Baccarella, A., Parrish, J. Z., & Kim, C. C. (2017). Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*, *18*(1), 38. https://doi.org/10.1186/s12859-016-1457-z

Wynants, E., Frooninckx, L., Crauwels, S., Verreth, C., De Smet, J., Sandrock, C., Wohlfahrt, J., Van Schelt, J., Depraetere, S., Lievens, B., Van Miert, S., Claes, J., & Van Campenhout, L. (2019). Assessing the Microbiota of Black Soldier Fly Larvae (Hermetia illucens) Reared on Organic Waste Streams on Four Different Locations at Laboratory and Large Scale. *Microbial Ecology*, *77*(4), 913–930. https://doi.org/10.1007/s00248-018-1286-x

Yang, F., Tomberlin, J. K., & Jordan, H. R. (2021). Starvation Alters Gut Microbiome in Black Soldier Fly (Diptera: Stratiomyidae) Larvae. *Frontiers in Microbiology*, *12*, 160. https://doi.org/10.3389/FMICB.2021.601253/BIBTEX

Zhang, P., Ganesamoorthy, D., Nguyen, S. H., Au, R., Coin, L. J., & Tey, S. K. (2020). Nanopore sequencing as a scalable, cost-effective platform for analyzing polyclonal vector integration sites following clinical T cell therapy. *Journal for Immunotherapy of Cancer*, *8*(1). https://doi.org/10.1136/JITC-2019-000299

Zhineng, Y., Ying, M., Bingjie, T., Rouxian, Z., & Qiang, Z. (2021). Intestinal microbiota and functional characteristics of black soldier fly larvae (Hermetia illucens). *Annals of Microbiology*, *71*(1), 1–9. https://doi.org/10.1186/S13213-021-01626-8/FIGURES/4

Zhou, J., Huang, H., Meng, K., Shi, P., Wang, Y., Luo, H., Yang, P., Bai, Y., Zhou, Z., & Yao, B. (2009). Molecular and biochemical characterization of a novel xylanase from the symbiotic Sphingobacterium sp. TN19. *Applied Microbiology and Biotechnology*, *85*(2), 323–333. https://doi.org/10.1007/S00253-009-2081-X

# APPENDICES

## Project Budget

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item** | **Item Description** | **Value (USD)** | **Qty.** | **Total**  **(USD)** |
| Sample collection and experimental design | * Rearing and dietary intervention of the BSF larvae * Obtaining samples from the gut microbiome | 450 |  | 450 |
| ***RNA Sequencing*** | | | | |
| RNA isolation kit | For RNA isolation from the BSF larvae gut | 430 | 1 | 430 |
| MinION starter pack | A portable real-time device from ONT for long-read sequencing  **Contents**  1x MinION Sequencing Device  1x Starter Pack Flow Cell  1x Included kits  1x Flow Cell Wash Kit  1x Control Expansion  Voucher for Flow Cell | 1,050 | 1 | 1,050 |
| PCR-cDNA Barcoding kit | * Allows for Multiplex sequencing * Enables multiplexing of 12 samples | 650 | 1 | 650 |
| Extra Flow Cell | For additional sequencing runs | 900 | 1 | 900 |
| Extra Expansion Kit | For additional sequencing runs | 300 | 1 | 300 |
| Flow cell priming kit | For additional sequencing runs | 40 |  | 40 |
| ThermoFisher MICROBEnrich kit | Enrichment of microbial RNA from host RNA | 550 | 1 | 550 |
| Single Place Magnetic stand (1.5ml) | Used in place of centrifugation. Uses magnetic capture beads. | 95 |  | 95 |
| Agencourt AMPure XP Beads | Used together with a magnetic stand for magnetic separation. | 450 |  | 450 |
| Shipping | From the UK to Kenya with Carramore International Ltd. | 790 |  | 790 |
| Training costs | Free training from an ONT representative | 0 |  | 0 |
| ***Data Handling Management, and Analysis*** | | | | |
| Data management and downstream analysis | * Data transfer and storage * High-Performance Computing (HPC) facilities (For outsourced facilities only) * Programs and software used will be mainly open-source. However, commercial software e.g., long-read sequence analysis and pathway analysis software will be required. * Data Analysis, statistical inferencing | 1,200 |  | 1,200 |
| ***Consumables*** | | | | |
| 10 mM dNTP solution | For PCR reactions | 2 | 65 | 130 |
| LongAmp Taq 2X Master Mix | For PCR reactions | 2 | 138 | 276 |
| Maxima H Minus Reverse Transcriptase (200U/µl) with 5x RT Buffer | Reverse transcription and strand-switching | 2 | 75 | 150 |
| RNaseOUT, 40 U/μl | Reverse transcription and strand-switching | 2 | 190 | 380 |
| Exonuclease I | Removal of single-stranded primers in PCR reactions before sequencing | 2 | 128 | 356 |
| Thin-walled, frosted lid, RNase-free PCR tubes (0.2 mL) (1000 units) | PCR Reactions | 1 | 150 | 150 |
| Gloves | Medium Sized powder-free gloves | 8 | 5 | 40 |
| Pipette tips | DNA-free 20 µl Pipette tips 1000 units | 2 | 1 | 2 |
|  | DNA-free 100 µl Pipette tips 1000 units | 4 | 1 | 4 |
|  | DNA-free 200 µl Pipette tips 1000 units | 2.5 | 2 | 5 |
|  | DNA-free 1000 µl Pipette tips 1000 units | 3 | 1 | 3 |
|  | DEPC Water | 75 | 1 | 75 |
| Microcentrifuge tubes | Eppendorf LoBind tubes (1.5ml) | 16 | 5 | 80 |
| **Grand Total** |  |  |  | **8556** |

## Time plan

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Year** | **2020** | **2021** | | | |
| **Month** | **Q4** | **Q1** | **Q2** | **Q3** | **Q4** |
| **Activity** |  |  |  |  |  |
| Writing and approval of project concept |  |  |  |  |  |
| Proposal writing |  |  |  |  |  |
| Fieldwork, breeding, and dietary Intervention of the BSF larvae |  |  |  |  |  |
| RNA extraction and Sequencing using ONT Minion Platform |  |  |  |  |  |
| Optimization and Error Correction of ONT Data |  |  |  |  |  |
| Data Analysis & implementation of metatranscriptomic pipeline |  |  |  |  |  |
| Identification and functional characterization of lignocellulolytic enzymes |  |  |  |  |  |
| Thesis Writing |  |  |  |  |  |
| Graduation | ***2022 (Tentative)*** | | | | |

## Parameter Collection Template

**BSFL Breeding**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day |  |  |  |  |  |  |  |  |  |  |  |
| Weight (gms) |  |  |  |  |  |  |  |  |  |  |  |

1. Weight of the larvae

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day |  | |  | |  | |  | |  | |  | |  | |  | |  | |
|  | Pre-feed | Post- feed | Pre-feed | Post-feed | Pre-feed | Post-feed | Pre-feed | Post-feed | Pre-feed | Post-feed | Pre-feed | Post-feed | Pre-feed | Post-feed | Pre-feed | Post-feed | Pre-feed | Post-feed |
| Weight (gms) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. Substrate Reduction Index
2. Temperature

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day |  |  |  |  |  |  |  |  |  |  |  |
| Temp 0C |  |  |  |  |  |  |  |  |  |  |  |

1. pH

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day |  |  |  |  |  |  |  |  |  |  |  |
| pH |  |  |  |  |  |  |  |  |  |  |  |

1. Length

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day |  |  |  |  |  |  |  |  |  |  |  |
| Length (mm) |  |  |  |  |  |  |  |  |  |  |  |

1. Color of larvae

\**photo must be against a white background*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day |  |  |  |  |  |  |  |  |  |  |  |
| Photo number |  |  |  |  |  |  |  |  |  |  |  |

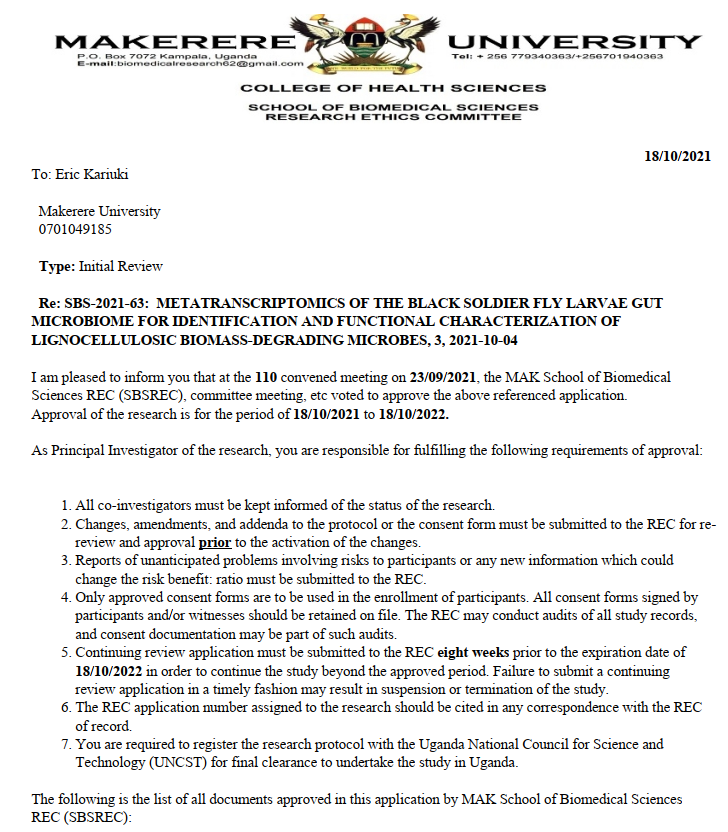
1. Rate of pupation

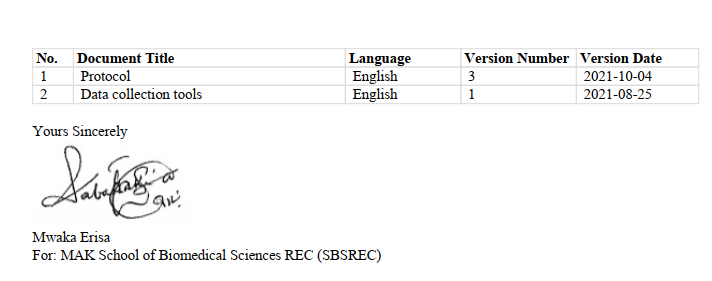
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day |  |  |  |  |  |  |  |  |  |  |  |
| Number of pupa/quadrat |  |  |  |  |  |  |  |  |  |  |  |

Compiled by: ……………………………………………

Sign: ……………………………………..

## SBS-REC Approval Letter





## NACOSTI Research Permit

