Identification and characterisation of lignin degrading bacteria and enzymes from larval guts of the African Palm Weevil (*Rhynchophorus phoenicis*).

Jessica Lapshak Lenka
School of Science, Engineering and Environment
University of Salford

A thesis submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy (PhD).

September 2022

Table of Content

List of Figures

List of Tables

Dedication

This thesis is dedicated to the Almighty, one true God and to my children Rotkang and Ritji Lenka.

Acknowledgement

Firstly, I would like to express my deepest gratitude to Almighty God who has led me to this knowledge seeking path and blessed me continuously with the opportunities to progress and excel in this journey, especially in this particular period of PhD studies.

I would like to thank my supervisor, Dr Natalie Ferry for giving me the opportunity to undertake my PhD at the University of Salford. I am especially thankful for her extremely patient supervision, continuous encouragement, and generous support throughout my project. Without her inspiration, understanding and timely support, I could never have learnt the appropriate way to carrying out this research systematically.

My appreciation also goes to my co-supervisor Dr Rhoderick Elder, and other academic and technician staff at Salford University including Dr Rachael Antwis, Prof. Ian Goodhead, Prof. Chloe James, Dr Tony Bodell, Dr Muna Abubaker and Dr Lee Haman for their support with sequencing, and helpful assistance during my period of laboratory experiments. I remain very thankful to Shweta, Poppy, and Masood for offering support with bioinformatic analysis of my sequencing data.

Many thanks to all of my team members especially Stephen and Henry, I appreciate your constant help, support, scientific discussions, and friendship that made my study period more meaningful and enjoyable.

I would also want to express my unreserved gratitude to all my family members (mum and sisters) and friends too numerous to mention who helped and supported me in various ways especially with caring for my children while I battle with the rigours of this studies, may God bless you all and meet you at the points of your needs.

Finally, I would forever be thankful to my sponsors, the Petroleum Technology Development Fund (PTDF) for giving me the opportunity to study my PhD in the UK where I have developed valuable skills. I have also encountered and networked with distinguished scientists and academics and learnt within an enabling environment that offers so much support. Also, I thank the British Federation of Women Graduates (BFWG) for offering me a grant to support my cost of living during my extended period of study after my scholarship elapsed, I don’t know how I could have coped without it.

List of abbreviations

AA	Auxiliary activities
ABTS	2,2 -azino-bis (3-ethylbenozthiazoline-6-sulfonic acid
AFEX	Ammonia fibre explosion
APW	African palm weevil
ASVs	Amplicon sequence variants
BBSRC	Biotechnology and Biological Sciences Research Council
Bio-PMs	Bio-Platform Molecules
BLAST	Basic local alignment search tool
CAZy	Carbohydrate active enzymes
CBMs	Carbohydrate binding modules
CD	Conserved domain
CE	Carbohydrate esterases
CO_2	Carbon dioxide
COG	Clusters of orthologous groups
DNA	Deoxyribonucleic acid
DyP	Dye decolourising peroxidase
EC number	Enzyme Commission number
EU	European Union
FDCA	Furan dicarboxylic acid
FOLy	Fungal oxidative lignin enzymes
GH	Glycoside hydrolases
GHG	Greenhouse gas

GT	Glycosyltransferases
H_2O_2	Hydrogen peroxide
HMF	Hydroxymethyl furfural
HRP	Horseradish peroxidase
IEA	International energy agency
IEX	Ion exchange chromatography
IMAC	Immobilised metal affinity chromatography
KL	Kraft lignin
LB	Luria-Bertani
LB Net	Lignocellulosic biorefinery network
LCB	lignocellulosic biomass
LDAs	Lignin degrading accessory enzymes
LiP	Lignin peroxidase
LMEs	Lignin modifying enzymes
LPMOs	Lytic polysaccharide monooxygenases
MnP	Manganese peroxidase
NAP buffer	Nucleic acid preservation buffer
NGS	Next generation sequencing
ORFs	Open reading frames
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PLs	polysaccharide lyases
pI	Isoelectric point
RB19	Reactive Blue 19
rpm	Rotations per minute
rRNA	ribosomal ribonucleic acid

SDS-PAGE	Sodium Dodecyl Sulphate – Poly Acrylamide Gel Electrophoresis
USDOE	United States Department of Energy
VP	Versatile peroxidase
WGA	Whole genome amplification
WGS	Whole genome sequencing

Abstract

As developing economies continue to grow, the world's demand for energy which currently stands at 84MB (million barrels oil) per day is projected to rise to 116 MB per day by the year 2030. The need to meet this continuous rise in demand for energy while lowering the emission of CO_2 and other greenhouse gases has necessitated a shift in focus from the exploitation of fossil fuels which are limited, to more renewable and environmentally safe biological resources such as lignocellulosic biomass, the main structural components of plant cell walls. Lignocellulose, however, is resistant to degradation, thus there is a high cost and energy requirement associated with its pre-treatment in order to access the lignin bound polysaccharides for subsequent hydrolysis, fermentation and conversion to biofuels and biomaterials.

Xylophagous (wood-feeding) insects such as the African palm weevil (*Rhynchophorus phoenicis*) have developed the ability to effectively utilize lignocellulosic substrates as an energy source due to the synergistic association with their gut microbes. This makes them viable resources to explore for novel lignocellulose degrading enzymes. Metagenomics allows access to the entire microbiome present in a particular environment and has been adopted in recent studies, rather than culture-based methods, thereby allowing for discovery of novel genes and enzymes from both culturable and non-culturable microbes.

In this study, we carried out taxonomic profiling of the bacteria in the gut of the African palm weevil's larvae using 16S rRNA gene amplicon sequencing with particular interest in identifying lignin degrading bacteria. We also performed functional metagenomics analysis from whole metagenome sequencing data derived from whole gut metagenomic DNA of APW larvae to identify genes and by extension, enzymes that can deconstruct lignocellulose and degrade its lignin component. The predominant bacterial genera found across all gut segments were *Enterococcus, Lactococcus*, *Shimwellia, Lelliotia, Klebsiella* and *Enterobacter*, with the foregut having the most diverse and abundant lignin degraders mostly from the *Proteobacteria* phylum. One thousand, one hundred and forty-one (1,141) annotated genes identified from the *R. phoenicis* larval gut bacterial metagenome aligned with genes encoding CAZymes and 249 of these belonged to the "Auxiliary Activities" class which harbours the suite of

genes implicated to play different roles in lignin deconstruction. Three genes of putative lignin degraders were successfully amplified by PCR, one of the three amplified genes B-38773 (encoding a putative deferrochelatase/ peroxidase of approximately 46kDa in size that has a conserved domain match to the dye decolourising peroxidase superfamily) was produced by heterologous expression and was found to exhibit activity on the peroxidase substrate ABTS, and the anthraquinone dye RB19 but no activity was observed with kraft lignin. Specific activity of 12.9Umg^{-1} at optimum temperature of 40°C and pH of 4 were recorded when B-38773 enzyme was assayed against ABTS as a substrate. Kinetic parameters: V_{max}, $K_{m,}$ $K_{cat,}$ and catalytic efficiency were determined to be 3.68 µMol/min, 1.089mM, 540.9S^{-1} and 4.96 X 10^5 M^{-1}S^{-1} respectively.

This study elucidates the lignocellulose degrading potential of the gut community associated with the African palm weevil (APW) by robustly defining the bacterial community structure of the APW gut. Also, massive data from the metagenomic library generated will serve as a storehouse from where genes with various potential functions identified by the inhabitant gut bacteria can be harvested to contribute to areas of biotechnological relevance for industrial applications.

Chapter 1: Biological Lignin valorisation: Challenges, prospects, and potential applications for biofuel and bioproducts generation from lignocellulose.

1.1 Fossil fuels: Overview, challenges, and the need for alternatives

Fossil fuels are hydrocarbons (organic compounds) which are found in soil and sediment, rocks, and sea, and they exist in three forms: coal, oil, and natural gas (Stephenson, 2018). The carbon source for the formation of these fossil fuels emanates from the degradation of photosynthesizing plants, algae and planktons which trap atmospheric CO_2 to produce high energy carbon compounds. When terrestrial plants die and decay, their carbon rich content is transferred to soil (in the case of coal) while the remnants of land plants and aquatic organisms (mostly microscopic organisms such as phytoplankton, algae, and bacteria) in lakes and sea supply the organic matter in the case of oil and natural gas. This happens through a series of biological, physical, and chemical alterations resulting in the various fossil fuels. Fossil fuels are therefore being made continually through the geological carbon cycle by sequestration of carbon deep in the earth. It is however, a very slow process which typically takes several million years (Stephenson, 2018).

Since the industrial revolution in the early 1800s, the world has relied on fossil fuels as the main source for generating energy and chemicals (Takkellapati *et al.,* 2018; Rana and Rana, 2017). According to Abas *et al.,* 2015, the peaks, declines and depletions of fossil fuels depend on their proven reserve, exploration, and consumption rates (Table 1.1).

Table 1.1 Global reserve, exploration, and consumption rates of fossil fuels (Abas *et al.,* 2015)

Fossil fuel Type	Oil ($x10^9$ Barrels)	Gas ($x10^{12}$ Cubic feet)	Coal ($x10^9$ Tons)
Total Reserve	1688	6558	891
Consumption (per day)	0.092	0.329	7.89
Rate of increase in reserves (Per annum)	0.6	0.4	19.2
Rate of increase in consumption (Per annum)	0.0014	0.0045	0.0031

Although the numbers in Table 1.1 indicate a steady increase in fossil fuels with no threat of immediate depletion, fossil fuels are still considered finite and non-sustainable in the long run as 86% of global energy demands and 96% chemicals are from fossils against 13.6% from renewable and alternative sources (Abas *et al.,* 2015; Stephenson 2018, Weiss *et al.,* 2020). From an environmental perspective, the mining, exploration and burning of fossil fuels to produce energy required for industrialization and modern living for an ever-increasing population (approximately 200,000 persons per day) returns stored-up carbon into the atmosphere. About $39.5x10^9$ tons of CO_2 is released from the over 12×10^9 tons of oil equivalent global annual energy demand, and this is projected to increase to 75×10^9 tons of CO_2 as energy needs in the future rise to $24\text{-}25 \times 10^9$ tons of oil equivalent (Abas *et al.,* 2015). CO_2 and other greenhouse gases such as nitrous oxide (N_2O), water vapour, methane (CH_4), chlorofluorocarbons (CFCs) and ozone (O_3) contribute to green-house effect by absorbing heat and reradiating it, resulting in increased global temperatures (global warming) (Stephenson 2018; Ayadi *et al.,* 2016). Global warming has adverse effects on agricultural crops and has resulted in other phenomena such as faster melting of glaciers and rising ocean levels, acid rain, excessive rain, floods and droughts, hurricanes, increased occurrences of heat and cold waves, and ultimately the extinction of fauna and flora (Suranovic, 2013; Howe and Leiserowitz, 2013; Mwangi *et al.*, 2015; Saini *et al.,* 2018).

Beside the menace of green-house gas emission and global warming, over reliance on fossil fuels can cause damage to the environment in the form of oil spills that pollute the sea and threaten aquatic life, the production of non-biodegradable products that

are non-recyclable and not readily degraded which end up littering land and water bodies depleting the environment and affecting quality of life. Environmentalists have advised that the use of fossil fuels be avoided or kept at the bare minimum because of their harmful consequences on nature as there is presently no known chemical process that can effectively get rid of the large amounts of CO_2 increasingly being released into the environment (Abas *et al.,* 2015). GHG emissions are also responsible for a variety of health and respiratory problems such as pneumonia, bronchitis, eye irritation, sneezing and coughing etc (Mofijur *et al.,* 2015).

Alternative and sustainable energy sources are therefore being sought to replace fossils and thereby counter the adverse effects associated with their utilization. Reviews on solar, wind, hydrogen, bioenergy, artificial photosynthesis, and fusion technologies show that other natural and artificial sources can more safely meet the world's energy demands (Dresselhaus and Thomas, 2001; Jia *et al.,* 2018; Didane *et al.,* 2016; Bartels *et al.,* 2010; Faunce *et al.,* 2013; Michaelides, 2012). Therefore, the move for a swift change in energy source from fossils to alternative and renewable resources which are environmentally friendly and economically viable must be supported by the global community (Cherubini 2010; Abas *et al.,* 2015).

1.2 Lignocellulosic biomass as alternative fossil fuel replacements

To reduce the dependency on fossil fuels due to its adverse effect on climate change, non-renewable and unsustainable nature, there are efforts worldwide geared towards shifting focus to other energy sources which are renewable, sustainable, and environmentally safe.

Plant biomass is a non-pollutant and abundant resource which has the capacity to produce bioenergy and biomaterials and can potentially replace products made from fossils as it is an equally carbon-rich resource (Takkellapati *et al.,* 2018, Cherubini 2010; Menon and Rao, 2012; Gupta and Verma, 2015). Lignocellulosic plant biomass also has the added advantage of mitigating global warming by contributing a near net zero CO_2 (IEA bioenergy Task 42 report; Rana and Rana, 2017). The CO_2 released when biofuels are burnt for energy is taken up through the process of photosynthesis to produce more plants that can be used again as feedstock, while material products

can be biodegraded, and their carbon content recycled (Naik *et al.,* 2010, Saini *et al.,* 2018). The drive for the transition from fossil dependent to bio-based economies stems from the need to reduce overdependence on fossil-based products, encourage diversification of energy sources, and climate change concerns (IEA bioenergy Task 42 report). Efforts to decelerate CO_2 and greenhouse gas emissions and restrict global warming to below the proposed limits of 2°C above pre-industrial levels continues to gain massive commitments from countries in the form of national policies and legislation, and globally as evidenced in several discourses, resolutions and ratifications from the Kyoto Protocol in 1997 to the Paris Agreement in 2015 (COP21) and the recently concluded climate change conference (COP26) held in Glasgow in 2021 (Suranovic, 2013; Gupta and Verma, 2015; Silva *et al.,* 2018).

Recently, there is high interest in lignocellulosic plant biomass (the woody component of plants) derived mostly from agricultural, forest, and municipal waste, due to their considerable potential as sustainable alternatives to petroleum-based resources as feedstock in the production of biofuels and biomaterials (Naik *et al.,* 2010; Gupta and Verma, 2015; Kumar *et al.,* 2017). This resource has however not been fully exploited due to the inherent challenge in its degradation attributed by its highly complex and heterogeneous structure (Xie *et al.,* 2014). Lignin is the aromatic component of lignocellulose from which a myriad of industrial chemicals and materials can be produced. Lignin is highly hydrophobic, and it binds tightly to cellulose and hemicellulose rendering them inaccessible to hydrolytic enzymes. These phenomena contribute to the recalcitrance of lignocellulose and hence frustrates its bioconversion. To achieve maximum and sustainable benefit from utilizing lignocellulosic biomass in place of fossil fuels, the hurdle of recalcitrance must be surpassed by employing strategies that can achieve its degradation in a manner that is cost effective and friendly to the environment (Xie *et al.,* 2014; Bundhoo, 2018; Bundhoo and Mohee, 2018).

1.3 Overview of lignocellulose structure

The cell wall of plants is made up of approximately 90% lignocellulose dry weight and is the most abundant source of organic carbon on earth (Gilbert, 2010; Prasad *et al.,* 2018). Lignocellulose serves a structural role in plants, provides strength to help them

withstand harsh environmental conditions and protects them against herbivores and pathogenic micro-organisms (Scheller and Ulvskov, 2010). Lignocellulose consists primarily of three main components: cellulose (35-50%), hemicellulose (15-35%), and lignin (10-30%) (Bugg *et al.,* 2020; Ni and Tokuda, 2013; Chen, 2014; Do *et al.,* 2014; Arumugam and Mahalingam, 2015). These components are tightly bound to each other via non-covalent forces as well as by covalent cross-links, resulting in a composite material which is resistant to degradation because of the encasement of crystalline cellulose by the lignin–hemicellulose matrix (Figure 1.1) (Barakat *et al.,* 2013**;** Mathews *et al.,* 2016). The mix ratio of these different components, their physical and chemical structure and interconnectivity varies in different species, tissues, cell types and maturity stage of plants (Barakat *et al.,* 2013; Isikgor and Becer, 2015; Le, 2021). Pectin and other nitrogenous based compounds are also found in plant cell walls (Gilbert 2010; Chen, 2014).

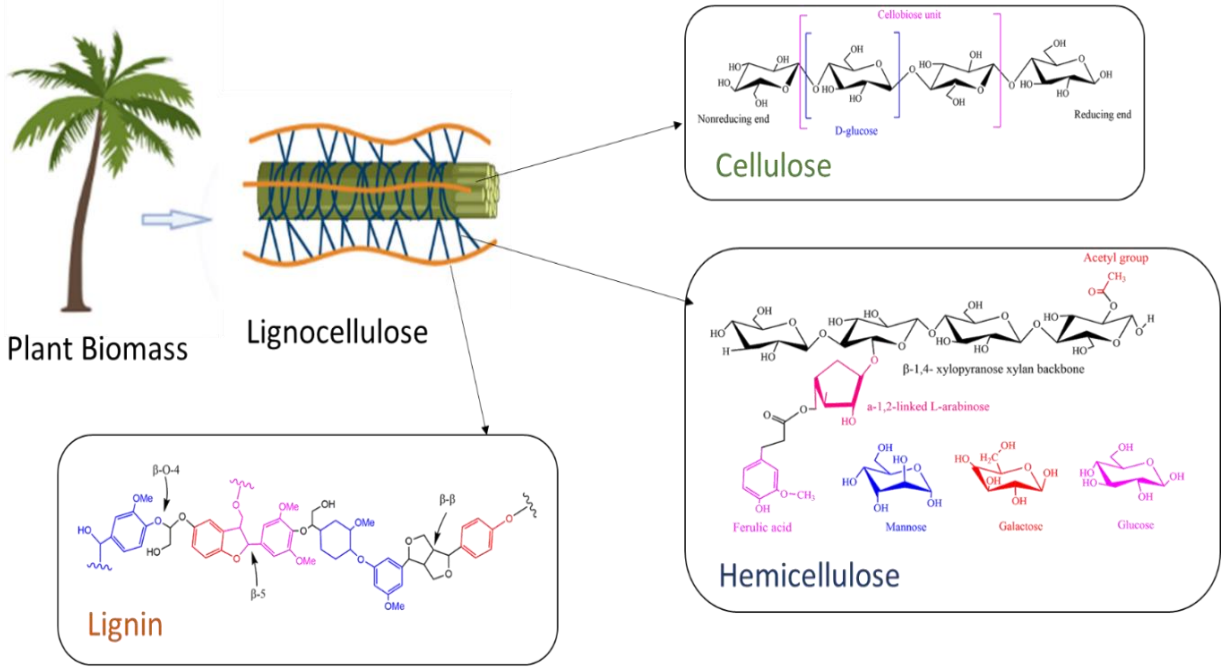


Figure 1.1 Lignocellulose Structure
The major components of lignocellulose (cellulose, hemicellulose, and lignin) and their molecular arrangement in lignocellulosic plant biomass.

1.3.1 Cellulose

Cellulose is a carbohydrate homopolymer composed of D-glucose units linked by β-1, 4- glycoside bonds in a linear, repeating disaccharide unit called cellobiose. It is the most abundant polysaccharide component of lignocellulose and makes up about 35-50% of its dry weight (except in cotton bolls which is almost 100% cellulose), hence making it the most abundant renewable organic polymer in nature, and its metabolism is an important part of the carbon cycle (Chen, 2014; Isikgor and Becer, 2015; Mussatto and Dragone, 2016). Due to the presence of inter and intra molecular hydrogen bonding and Van der Waals interactions, the linked glucose molecules can aggregate into highly ordered fibrillar arrangement of about 500nm diameter formed from several entwined microfibrils of 10-25nm diameter. Consequently, cellulose fibrils have high tensile strength and are stronger than a steel wire of equal thickness (Chen, 2014). The fibrils are further entwined, forming a network which forms the basic framework of the cell wall. Cellulose fibrils have regions of high order (crystalline regions) and regions of less order (amorphous regions) with no sharp demarcation between the two regions. Within the crystalline regions, the individual microfibrils are so tightly packed making the molecule impermeable to water and less accessible to cellulases hence the crystalline regions of cellulose are more resistant to biodegradation than the amorphous regions (Sun and Zhou, 2011; Menon and Rao, 2012).

Unlike cellulose, starch is a non-linear molecule as it is composed of α-1, 4 linkages, with α-1, 6 bonds occurring at branch points. Hence, starch tends to form helical structures in the solid state and in solution. The occurrence of branching and helix formation combine to make it difficult for starch to aggregate because the molecules cannot stack together. These molecular differences result in starch being soluble whilst cellulose is not and being soluble makes starch much easier to degrade compared to cellulose hence its exploitation in the production of first-generation biofuels (Ahmad, 2010; Bugg *et al.,* 2011, Sun and Zhou, 2011)**.**

To achieve complete breakdown of cellulose into the simple sugar glucose, the following three enzymes act cooperatively in this order: endo-β-1, 4-glucanases (EC 3.2.1.4) hydrolyze cellulose chains in a random, non-processive manner creating new ends, exo-β-1, 4 glucanases, e.g cellobiohydrolases (EC 3.2.1.91) or cellodextrinases

(EC 3.2.1.74) depolymerize cellulose chains from their reducing or non-reducing ends in a processive or ordered manner releasing cellobiose units, and β-glucosidases (EC 3.2.1.21) break down the glycosidic bonds between the cellobiose units to release free glucose molecules (Ezeilo *et al.,* 2017; Gilbert, 2010; Willis *et al.,* 2010, Chang *et al.,* 2012). The glucose molecules released following the breakdown of cellulose can then undergo fermentation into ethanol (biofuel) which can further be reduced to ethane (Feedstock chemical for production of other industrially relevant chemicals) (Ahmad, 2010; Barakat *et al.,* 2013; Joynson *et al.,* 2014).

1.3.2 Hemicellulose

The second most abundant polysaccharide component of lignocellulose is hemicellulose (15-35%). Hemicellulose is made up of several heteropolymers including xylan, galactomannan, glucuronoxylan, arabinoxylan, glucomannan and xyloglucan (Le, 2021). The heteropolymers of hemicellulose are composed of different pentoses (5- carbon monosaccharides) such as xylose and arabinose, and hexoses (6-carbon monosaccharide units) such as mannose, rhamnose, glucose, and galactose, as well as uronic acids, with the backbone sugars in a β-linkage. The backbone sugars are decorated with other different kinds of sugars and acetyl groups in a random manner; thus, hemicelluloses are highly branched and non-crystalline polymers with a much lower degree of polymerisation compared to cellulose (Takkellapati *et al.,* 2018; Sun and Zhou, 2011; Isikgor and Becer, 2015). Hemicelluloses differ from one species of plant to another in the different combination of pentoses and hexoses that are found in the heteropolymers that make up the backbone of each one and from the different side chains that could be found attached to each. Also, different species of plants and cell types have varying types of subunits in their overall structure, with different degrees of abundance. For example, the predominant hemicelluloses in hardwoods are glucuronoxylans while in softwoods, glucomannans predominate (Takkellapati *et al.,* 2018, Scheller and Ulvskov, 2010).

Hemicellulose contributes to the strengthening of plant cell walls due to its cross-linking ability with cellulose microfibrils and lignin, forming a complex network of bonds and therefore further block the access of enzymes to the glucose molecules stored up in

cellulose and increasing the insolubility and resistance of the cell wall components to biodegradation (Isikgor and Becer, 2015; Chen, 2014). The conversion of biomass to biofuels requires the removal and hydrolysis of these complex heteropolysaccharides. Thermal and chemical pre-treatments result in the degradation of hemicelluloses into furfurals and hydroxymethyl furfurals and formic acids which have been reported as fermentation inhibitors. This adds to the cost and complexity of downstream processing, decreasing the sugar yield obtainable for bioethanol production (Rana and Rana, 2017; Kane and French, 2018).

1.3.3 Lignin

Lignin, unlike cellulose and hemicellulose is an aromatic, non-polysaccharide component of lignocellulose, and the second most abundant natural organic polymer next to cellulose comprising 20-30% dry weight of plant biomass (Sun and Zhou, 2011; de Gonzalo *et al.,* 2016)**.** Lignin is a heterogenous polymer of assembled phenylpropanoid subunits: guaiacyl (G), p-hydroxyphenyl (H) and syringyl (S) which are produced from three oxidatively coupled hydroxycinnamyl monomers (monolignols): coniferyl, *p*-coumaryl and sinapyl alcohols respectively to give the structure of lignin (Figure 1.2). The three phenyl propanoid oligomers differ in their degree of methoxy group substitutions and are held together by several C-O (occurring as the majority) and C-C bonds such as the β-0-4 (aryglycerol-β-aryl ether), α-O-4 (non-cyclic benzyl aryl ether), 4-O-5 (Bi-phenyl ether), β-β (Resinol), 5-5 (Biphenyl), β-5 (phenylcoumaran), β-1(1,2-Diarylpropane) to form different dimeric structures (Pollegioni *et al.,* 2015; Olsson, 2016; Datta *et al.,* 2017; Santos *et al.,* 2013; Chen and Wan 2017; Rana and Rana, 2017). These dimer units are polymerised into lignin polymers with their specific amounts and arrangements being different in different plant species thereby resulting in distinct compositions and structures of lignin (Fisher and Fong, 2014; Shewa *et al.,* 2016). For example, in softwoods, lignins are comprised of mainly guaiacyl and sometimes p-hydroxyphenyl subunits and make up 20-30%, in hardwoods, guaiacyl and syringyl subunits predominate and they make up 15-25%, while in grasses, all the three subunits (guaiacyl, p-hydroxyphenyl and syringyl) are found making up 10-15% of total biomass (Sun and Zhou, 2011; Lazaridis *et al.,* 2018).

The type of subunits found in each species imparts on the reactivity of the lignin in these species due to the variety of functional groups present (Arumugam *et al.,* 2014). The lignin structures generally found in hardwoods are rich in methoxy groups linked predominantly by the biphenyl linkage with less possibility of branching which translates into the great degree of rigidity seen in hardwoods. On the other hand, where all three lignin subunits occur, there will be fewer methoxy groups, increased branching, less rigidity, ultimately resulting in a more flexible wood (Ahmad, 2010)**.**

The enormous variety of structures and bonds in lignin which make it difficult to degrade is believed to serve as a powerful defence mechanism against pathogen attack. The aromatic rings linked by alkyl ethers are a chemically resistant motif that grants lignin a high degree of stability to many chemical methods of breakdown. Lignin crosslinks with hemicellulose and cellulose after its synthesis to form a complex matrix (Fisher and Fong, 2014) and hence lignin is the principal reason that lignocellulose is difficult to deconstruct. Although lignin cannot be broken down to simple sugars and converted to ethanol via fermentation like cellulose and hemicellulose, lignin is an abundant source of aromatic units which can be used as a renewable feedstock to produce other high-value chemicals and products in a biorefinery (Cherubini, 2010; Olsson, 2016).

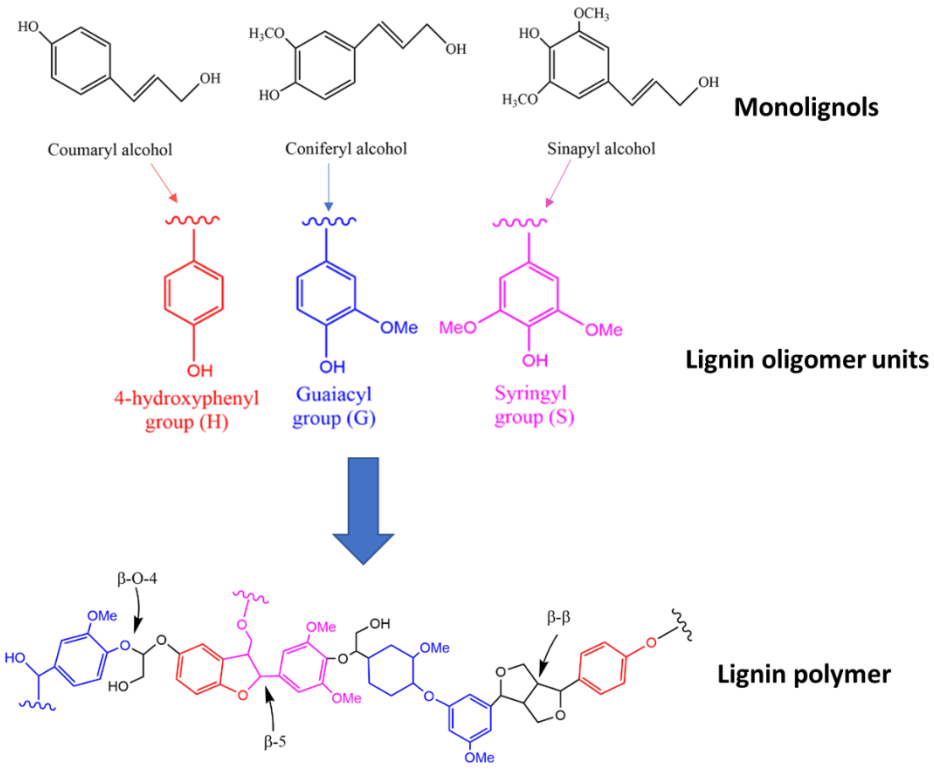


Figure 1.2 Lignin structure
Structures of the monolignols and their corresponding phenyl propanoid monomeric units, and a representation of how these subunits are heterogeneously arranged, held by different linkage types forming the lignin polymer

Lignocellulose being the most abundant organic resource on earth is a form of sustainable biomass with high promise for application in the production of renewable fuels and chemicals. Cellulose and hemicellulose are easily degraded both enzymatically and chemically and hence are readily accessible for application in industrial processes. However, lignin is much more resistant to degradation, and this limits the full exploitation of the potential range of products that can be produced from lignocellulose as a result of inadequate knowledge and technologies required for its efficient deconstruction (Lambertz *et al.,* 2016). The principal challenge with lignocellulosic biomass utilisation is therefore overcoming its resistance (which is mainly imparted by the presence of lignin, as well as other related obstacles) by employing different pre-treatment processes to solubilise and remove lignin (Olsson, 2016; Silva *et al.,* 2018). Biotechnological processes that exploit lignin degrading microbes and their enzymes can contribute to more efficient and environmentally

sound use of renewable lignocellulosic feedstocks for sustainable production of materials, chemicals, biofuels, and energy (Ruiz-Dueñas and Martínez, 2009).

1.4 Biomass Pre-treatment

The recalcitrance of lignocellulosic biomass to degradation is because of its inherent properties such as the crystalline nature of cellulose, the heterogenous structures of hemicellulose and lignin, and the composite macromolecular assemblage of its components whereby they form a tight knit association, further cemented by lignin (Barakat *et al.,* 2013). The ether and C–C linkages that are found in lignin are not susceptible to hydrolytic breakdown, hence ultimately making lignin/ lignocellulose difficult to degrade. This recalcitrant lignin also crosslinks with hemicellulose and shields cellulose filaments, providing a physical barrier and hindering access to enzymes required for its deconstruction (Bugg *et al.,* 2011a)**.** Therefore, there is a requirement for lignin removal as a pre-requisite step in biorefining plant biomass (Narra *et al.,* 2015; Kumar *et al.,* 2017)**.**

Before lignocellulosic biomass can be converted to useful products in a biorefinery, it needs to undergo pre-treatment. The purpose of pre-treatment is to increase the access of hydrolytic enzymes to cellulose for rapid and high yielding hydrolysis. However, the development of efficient pre-treatment technologies remains one of the main bottlenecks in the bioconversion of lignocellulose on a commercial scale. Many processes for pre-treatment are in use with continuous improvement required as there is no one method that can be applied to all feedstock types and by-products. The pre-treatment process is crucial for efficient bio refining as it significantly affects other processes downstream and impacts on overall cost and product yield. Therefore, a pre-treatment method will be deemed viable for large-scale industrial application if it can be applied to a variety of starting materials, has low negative impact on the environment, requires minimum further treatment for other downstream processes and will ultimately result in maximum product yield with minimum overall cost arising from both pre- and post- pre-treatment activities (Kurian *et al.,* 2013; Kumar *et al.,* 2017; Silva *et al.,* 2018; Mussatto and Dragone, 2016).

Pre-treatment methods alter the binding and interactions between the different components of lignocellulose, disrupting the physical cell wall barrier and causing modifications to its structure, enabling enzyme access to the polysaccharides thus enhancing digestibility at an accelerated rate (Barakat *et al.,* 2013; Isikgor and Becer, 2015,). This involves the removal and breakdown of lignin and hemicellulose, decreasing the crystallinity and particle size of cellulose, making it more porous with increased surface area for enzymes to act on and hence resulting in increased recovery of sugars from cellulose (Gupta and Verma 2015; Cheng and Brewer, 2017, Silva *et al.,* 2018, Mussatto and Dragone, 2016).

Methods of pre-treatment can be considered under the following categories: physical/ mechanical, physico-chemical, chemical, and biological (Cheng and Brewer, 2017, Rana and Rana, 2017; Le, 2021)**.**

1.4.1 Physical/ mechanical methods

These methods are employed to cut down the size and break open the usually large and bulky starting materials by way of grinding, milling, chipping, or even using high energy radiation (Gupta and Verma 2015; Kumar *et al.,* 2017). This results in a reduction in the crystalline form of the cellulose and increased surface area leading to a consequent increase in efficiency and improvement in enzymatic hydrolysis of the lignocellulosic materials (Kumar *et al.,* 2017). However, large energy input is demanded to run the machinery for these processes. Depending on the nature of the biomass, the associated cost can outweigh the benefits and make the overall process not economically attractive especially for hardwoods (Menon and Rao, 2012). Pyrolysis, which uses high temperatures to increase the chemical breakdown of cellulose can be used as an alternative to grinding and milling. However, at the very high temperatures (400–600 °C) which is the temperature range required for the lignin fraction to begin to solubilize and degrade, phenolic breakdown products that are toxic to the microbes required for fermentation stages are produced (Lazaridis *et al.,* 2018; Gupta and Lee, 2010). Some common phenolic breakdown products, mostly volatile oils such as 4-ethyl-2-methoxyphenol, vanillin, 2-methoxyphenol, 2-methoxy-4-vinylphenol, acetaldehyde, and 2-methoxy-4-methylphenol can interact with

hemicellulose to form precipitable compounds known as lignin carbohydrate complexes (LCCs) thereby complicating downstream processing if not removed (Usino *et al.,* 2020; Ansari *et al.,* 2019).

1.4.2 Physico-chemical methods

Physico-chemical pre-treatments combine both physical and chemical processes. Examples are steam explosion (SE), SO_2 or CO_2 catalysed steam explosion, ammonia fibre explosion (AFEX), liquid hot water and microwave-chemical pre-treatment methods among others (Menon and Rao, 2012).

Steam explosion is one of the early physico-chemical methods developed. Here, steam at very high pressure is used to raise the temperature of the feedstock to over 240°C for an appropriate length of time that is determined by the moisture content of the biomass. Steam explosion was developed by consolidating on the effect of steam on biomass, but this time, high pressure steam is applied to chipped biomass and then rapidly lowered. This causes water within the biomass to expand rapidly increasing its porosity (Kurian *et al.,* 2013). While this method uses less energy, it has the disadvantage of resulting in the loss of the hemicellulose component of the biomass, incomplete lignin separation and production of toxic compounds that can inhibit the action of microbes during the later stage of fermentation (Kumar *et al.,* 2017).

The principle behind AFEX is same as that of SE, but ammonia rather than steam is used at temperatures of about 90°C (far less than that required for SE). AFEX pre-treatment yields better depolymerisation of biomass by efficient removal of acetyl groups on hemicellulose, cleavage of lignin carbohydrate linkages and C-O-C bonds within lignin polymers, all resulting in a disruption of the structure, increased surface area for enzyme action and de-crystallization of cellulose. AFEX is a less complex process compared to SE, the ammonia used is recoverable and can be reused and does not complicate subsequent downstream processes, however, it is not efficient when using biomass that is rich in lignin such as woods and nuts, and it is not cost effective as both the cost of ammonia and the cost involved in its recovery are very high (Kumar *et al.,* 2017; Menon and Rao, 2012).

1.4.3 Chemical methods

In terms of energy conservation, chemical methods of pre-treatment have been the most successful and the most used so far. Chemicals such as alkalis, inorganic and organic acids, organic solvents, and ionic liquids (IL) have all been used to pre-treat biomass (Kurian *et al.,* 2013).

Acid pre-treatment uses concentrated or dilute acids to deconstruct the composite structure of lignocellulose. Dilute H_2SO_4 is most frequently used commercially because it is amenable to different biomass types (Kumar *et al.,* 2017). Hydrochloric acid, phosphoric acid, and nitric acid have also been studied (Menon and Rao, 2012; Gupta and Verma, 2015). Acid hydrolysis uses low temperatures, hence less energy and is used to efficiently remove hemicelluloses and separate out the fractions of lignocellulose, but slurries arising from chemical pre-treatment must be neutralized prior to downstream processing (Saha *et al.,* 2005).

Alkaline pre-treatment uses bases such as NaOH, Ca $(OH)_2$, KOH, hydrazine, NH_4OH, NH_3 etc (Kumar *et al.,* 2017). Alkaline pre-treatment of biomass results in the breakdown of ester and glycosidic side chain linkages thus altering the structure of lignin. Alkaline pre-treatment also results in partial solubilization of hemicellulose, swelling and partial de-crystallization of cellulose, hence making it available to hydrolytic enzymes. The best studied alkali is sodium hydroxide (Menon and Rao, 2012). Different alkalis are more effective for different feedstocks, for example, NaOH works best for substrates rich in lignin content such as hardwoods while NH_4OH is optimal for low lignin containing substrates (Gupta and Lee, 2010)**.** To obtain pure cellulose using chemical pre-treatment, acid treatment, which is more effective in removing hemicellulose, should be used followed by alkalis which are effective in removing lignin (Menon and Rao, 2012).

Ionic liquids (ILs) or “green solvents” as they are often called are potent chemicals composed of paired ions (large cations and small anions) that can replace organic solvents in several chemical processes due to their characteristic low vapour pressure, stability at high temperatures, low melting points and high polarity (Socha *et al.,* 2014; Kumar *et al.,* 2017). ILs prepared from imidazole and pyridine cations such as 1,3-dimethylimidazolium methylsulfate ([mmim][MeSO4]), 1-hexyl-3-methylimidazolium

trifluoromethanesulfonate ([hmim][CF3SO3]), 1-butyl-2,3-dimethylimidazolium tetrafluoroborate ([bm2im][BF4]) and 1-butyl-4-methylpyridinium hexafluorophosphate ([bmpy][PF6]) have been studied and used to dissolve lignin in biomass pre-treatment. These ILs can be recovered and reused, act on a wide variety of biomass substrates and do not form toxic by-products, but they do not efficiently breakdown lignocellulose, they inactivate cellulases during hydrolysis and are very costly (Pu *et al.,* 2007).

Organic solvents such as ethanol, methanol, ethylene glycol, glycerol, ethers, phenols, and ketones have also been employed in the delignification of lignocellulose either in the presence or absence of a catalyst in a process called organosolv pre-treatment (OP). It is suitable for removing lignin from biomass that has a high lignin content, but it is expensive (Kumar *et al.,* 2017; Kurian *et al.,* 2013).

1.4.4 Biological methods

Unlike other pre-treatment methods, biological methods do not require the use of sophisticated or expensive equipment, a high input of energy nor the application of harsh chemicals. It is less harmful to the environment and relatively cost effective because it exploits the natural abilities of micro-organisms to degrade lignocellulose (Riyadi *et al.,* 2020; Gupta and Verma, 2015; Kumar *et al.,* 2017). Here, microorganisms such as the brown, white and soft rot fungi and several bacterial species that secrete ligninolytic enzymes (laccases and peroxidases) are used for degradation of lignin and hemicelluloses in the lignocellulosic biomass (Sindhu *et al.,* 2016; Vasco-Correa *et al.,* 2016; Tsegaye *et al*., 2019).

Although each of the pre-treatment methods discussed above have their advantages and disadvantages for certain feedstocks and circumstances, biological pre-treatment seems promising compared to other conventional methods as it is considered relatively inexpensive, uses highly active, specific and stable enzymes that are biodegradable, the by-products produced normally do not inhibit subsequent hydrolysis steps and it requires minimal energy input as the pre-treatment is performed at near ambient conditions of temperature and pressure thereby having minimal negative impact on the environment (Gupta and Verma, 2015, Silva *et al.,* 2018; Riyadi *et al.,* 2020). However, to achieve optimum sugar yield, and increase overall efficiency of pre-treatment

methods using biological enzymes, certain conditions such as length of incubation and increased efficiency of microbial enzymes used still need to be improved upon to make it of higher and comparable advantage, suitable enough to replace other conventional pre-treatment methods for industrial scale application (Sindhu *et al.,* 2016; de Gonzalo *et al.,* 2016)**.**

Extensively characterised microorganisms predominantly of fungal and bacterial origin have been exploited and a significant number of ligninolytic enzymes have been identified, purified, and characterised from their genomes. Examples of fungal enzymes include lignin and manganese peroxidases from *Phlebia radiata* (Vares *et al.,* 1995) and *Phanerochaete chrysosporium (Tien and Kirk, 1983)*; Laccases from *Trametes versicolor* (Casland and Jonsson, 1999), DyP-type peroxidases from *Irpex lacteus* (Qin *et al.,* 2018) and *Pleurotus sapidus* (Lauber *et al.,* 2017). Bacterial laccases have also been reportedly produced by both Gram-positive and negative bacterial genera including *Bacillus, Geobacillus, Streptomyces, Rhodococcus, Staphylococcus, Azospirillum, Lysinibacillus, Pseudomonas, Enterobacter, Delftia, Proteobacterium, Alteromonas* and *Aquisalibacillus* (Chauhan *et al.,* 2017). A variety of DyP-type peroxidases have been identified and characterised from *Rhodococcus jostii* (Ahmad *et al.,* 2011), *Pseudomonas fluorescens* (Rahmanpour and Bugg, 2015; Loncar *et al.,* 2019), *Thermobifida fusca* (Rahmanpour and Bugg, 2015; van Bloois *et al.,* 2010), *Bacillus subtilis* (Min *et al.,* 2015), *Rhodococcus sp.* T1 (Sahinkaya *et al.,* 2019), *Thermomonospora curvata* (Chen *et al.,* 2015). (See comprehensive list of bacteria implicated in lignin degradation on table 1.4)

1.5 Bio-based products obtained from bio-refining lignocellulosic biomass

The awareness that plant biomass is capable of potentially replacing a large fraction of fossil-based resources as an alternative industrial feedstock material is on the increase globally and this has resulted in the development of the concept of a biorefinery. This addresses the main disadvantages of using fossil fuels (finite supply and environmental hazard) in the production of both energy and non-energy products (Cherubini, 2010).

According to the various definitions, a biorefinery is a facility with a similar concept but alternative to a fossil-based refinery where biomass is processed in a sustainable and

cost-effective manner into a wide spectrum of marketable products with the aim of progressively replacing petroleum refinery products (IEA bioenergy-Task42 report; Ahmad, 2010; Ragauskas *et al.,* 2014; Silva *et al.,* 2018).

Broadly, two categories of products can be obtained from a biorefinery system: Energy products (these are fuels used for their energy content, electricity, and heat generation or for transportation) and material products (products required for their physical or chemical attributes but not for energy generation) (Cherubini, 2010; Chukwuma *et al.,* 2021).

1.5.1 Energy products (Biofuels)

Biofuels are renewable sources of energy produced from natural biological resources such as photosynthesizing microorganisms and higher plants that capture energy from the sun. Biofuels can be in liquid, solid or gaseous forms and the major ones include bioethanol, biodiesel, biobutanol, bio-oil (liquid biofuels), biogas, biomethane, bio-syngas, bio-hydrogen (Gaseous fuels), charcoal and lignin pellets (solid biofuels) (Gupta and Verma, 2015; Majidian *et al.,* 2018; Bundhoo, 2018). From the above-mentioned fuels, the production and utilization of bioethanol and biodiesel have received the most global attention with respect to development of established technologies and utilization as alternative fuels. Bioethanol is produced mainly from plants with a high composition of polysaccharides (starch, cellulose, and other sugars) which can be hydrolysed and fermented to ethanol while biodiesel is produced from fats and oils (lipids) by transesterification of fatty acids to produce fatty acid methyl esters (FAME) (Sarma *et al.,* 2014). Bioethanol and biodiesel can be used as substitutes for gasoline and diesel respectively either alone (with requirement for minor engine modifications) or blended with fossil-based fuels (USDOE Biomass Multi-Year Program, 2008; Rana and Rana, 2017).

While the United States Department of Energy (USDOE) set a target to reduce gasoline use by 20% of the figure in 2007 and to reduce crude oil demand by 30% in 2030 by replacing with biofuels predominantly, the EU has also mandated that biofuels should account for 10% of transportation fuels by 2020 and 25% by 2030 in all member countries. The UK transport sector on 15th April 2018 made new changes to the

renewable transport fuel obligation (RTFO) which is meant to achieve a doubling in the use of renewable fuels within the next 15 years by compelling transportation fuel suppliers to increase the biofuel volume ratio from 4.75% in 2018 to 9.75% in 2020, and then to 12.4% by 2032 (https://www.gov.uk/government/news/new-regulations-to-double-the-use-of-sustainable-renewable-fuels-by-2020). These policies and legislations are directed towards lowering the pressure and over dependence on fossil fuels and their hazardous consequence of CO_2 emission and climate change on the environment in alignment with the objectives of the United Nations Framework Convention on Climate Change (UNFCC) to replace fossil fuels with biofuels (Saini *et al.,* 2018).

1.5.1.1 Classification of Biofuels

There are mainly four categories to classify biofuels based on the biological feedstock used for their production: first, second, third and fourth generation biofuels (Abdullah *et al.,* 2019). The development of first-generation biofuels began by exploiting agricultural crops such as sugarcane, corn, beet, rice, wheat to produce bioethanol; plant oils such as sunflower and rapeseed oil to produce biodiesel (Figure1.3) (Ahmad, 2010; Aro, 2016), and starch-derived biogas, biomethanol and bioethers were produced from food crops (Cherubini, 2010)**.** For bioethanol production, the carbohydrates are hydrolysed into simple sugars, and the simple sugars (hydrolysate) undergo fermentation to produce the alcohol ethanol, which is then distilled. Plant and vegetable oils are subjected to transesterification reactions where fatty acid methyl esters are produced along with glycerine as the by-product in biodiesel production (Escobar *et al.*, 2009).

Billions of litres of biofuels are being produced annually in different parts of the world (Naik *et al.,* 2010). The agricultural specialisation in each location determines the available raw material and hence the predominant type of biofuel produced. In North America and Brazil, corn and sugarcane are the main crops from which bioethanol is made while in Western Europe, the dominant biofuel is biodiesel from recycled cooking oil (Havlík *et al.,* 2011). In the early part of the year 2000, there was a massive upsurge in production of bioethanol from crop plants with continuous increase until 2013 (Gupta

and Verma, 2015). These increases reflect the relatively low cost and ease of production in the use of food crops as raw materials to make first generation biofuels as the need for costly pre-treatment does not arise (Sarma *et al.,* 2014). Unfortunately, however, as increase in land utilization for cultivation of energy crops to serve as biofuel raw materials began to translate into increases in prices of food for human and animal consumption, especially in developing countries, public opinion and acceptability of the technology decreased as it sparked social, environmental, economic, and moral concerns and debates (the "so called" food vs fuel controversy) (Naik *et al.,* 2010; Mohr and Raman, 2013). In the period between 2005 and 2007 when food prices spiked, several reports have attributed the spike to the use of agricultural products, in particular maize, wheat, and vegetable oil, as feedstock for biofuel production at the time as nearly 60% increase in global consumption of cereals and vegetable oils was due to the production of the biofuels: bioethanol and biodiesel (Tangermann, 2008). A report by Agence France-presse quoted the chairman and chief executive of Nestlé, Peter Brabeck-Letmathe saying "If as predicted we look to use biofuels to satisfy twenty percent of the growing demand for oil products, there will be nothing left to eat. To grant enormous subsidies for biofuel production is morally unacceptable and irresponsible" (Tenenbaum, 2008). A confidential report based on detailed analysis of the food vs fuel crisis put together by an internationally respected economist at the world bank obtained and reported by The Guardian paper estimated a 75% rise in food prices triggered by government policies that clamoured for increased biofuel production as against the US government's claim of only a 3% rise (World bank, 2008). While those who disprove of the use of agricultural crops for fuel production have continued to raise such concerns citing hike in food prices among other reasons as mentioned above, governments and some other interested stakeholders have contradicted those claims by attributing such spikes to other factors such as price of fertilizer, weather, and climatic factors, increase in global population that translates into increased demand for food etc. (Popp *et al.,* 2014; Ajanovic, 2011). Again, the "green" potential of these biofuels is being questioned because the farming of energy crops to achieve mass production is intense and requires large input of energy. Of great concern too is the release of N_2O gas (a more dangerous greenhouse gas compared to CO_2) into the environment from

excessive application of nitrogen fertilizers to the fields where these energy crops are farmed (Kurian *et al.,* 2013). Also, high overall cost resulting from high energy input requirement to produce biodiesel from oilseed is prohibitive (Ghosh and Ghose, 2003; Hirani *et al.*, 2018), therefore, large government subsidies are required to make biodiesel at prices that can compete with those of petroleum-based fuels (Joynson *et al.,* 2014). These situations have motivated the increased research on the use of non-edible plant biomass as fuel feedstock.

The future of biomass derived fuels is now focussed on the use of plant materials that are not fit for consumption, such as woody agricultural waste and forestry residues or co-products such as wheat straw (lignocellulosic materials). These are called second generation biofuels (Figure 1.3) (Avanthi and Banerjee, 2016) and over the years, there has been a progression from “first generation” to “second generation” biofuels, which differ essentially in the feedstock raw materials used to produce these fuels. Lignocellulosic materials are the plant materials that are not suitable for consumption by humans. They are renewable as 1×10^{10} metric tonnes are produced worldwide via photosynthesis each year. These non-edible lignocellulosics include both grazable and non-grazable plant materials both rich in holocellulosic content (60–70% w/w) that can serve as raw material for bioethanol production (Da Silva *et al.,* 2013). Second generation biofuel production from lignocellulosics therefore has the advantages of addressing the food vs fuel controversy, being economically sustainable and environmentally safe and as well ensuring energy security, thanks to the co-production of both bioenergy and high value chemicals in a biorefinery (Cherubini, 2010).

There are two main situations that pose a great challenge to the use of these lignocellulosics to produce fuels and other biochemicals. The first has to do with the unavailability of a large quantity of one feedstock type to meet the requirement of a large-scale biorefinery. Secondly, the presence of lignin in plant biomass which is largely responsible for its resistance to degradation (Narra *et al.,* 2015). This, therefore, requires the biomass to undergo pre-treatment to separate the lignin portion of lignocellulose, making the carbohydrates available for enzymatic hydrolysis and fermentation to bioethanol (Ahmad, 2010; Bugg *et al.,* 2011a).

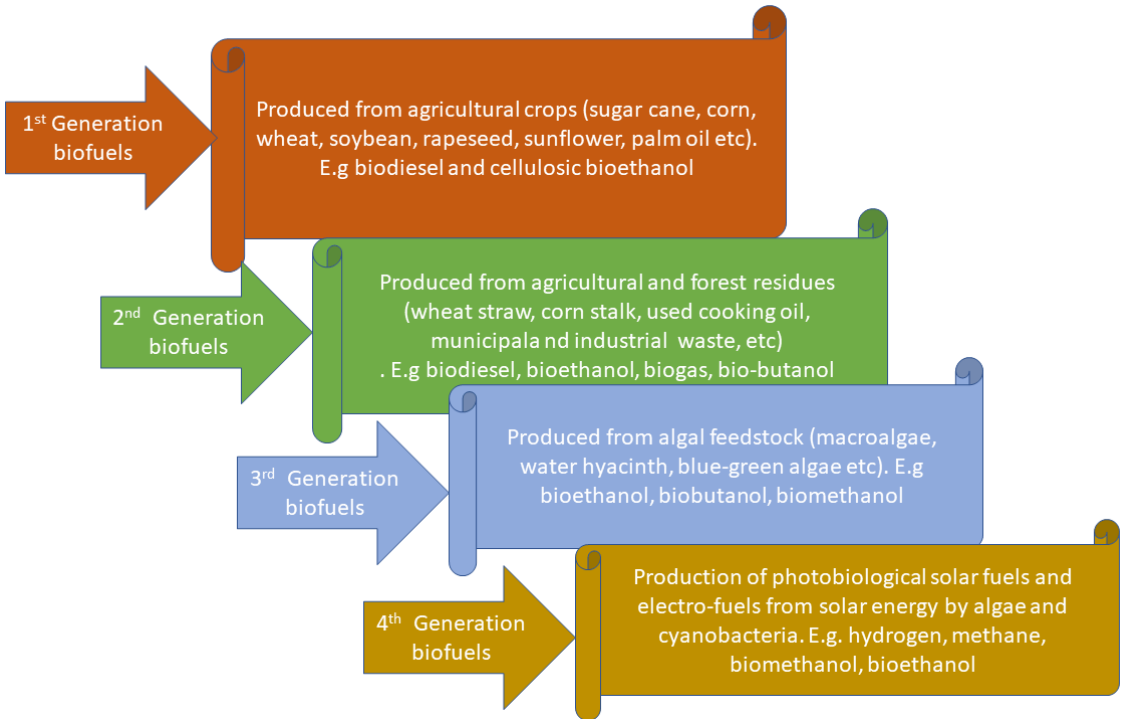


Figure 1.3 Classes of Biofuels
Different generations of biofuels and their respective biomass sources. (Modified from Kumari and Singh, 2018)

Carbohydrates and oils produced by algae via photosynthesis using CO_2, water and sunlight serve as feedstock for producing a class of biofuels designated "third generation" biofuels (Figure 1.3) (Behera *et al.*, 2015). With over 40,000 identified algal species existing in both fresh and sea water, high growth rates and ability to tolerate a wide range of environmental conditions, these micro machines can serve as renewable and sustainable biomass feedstock for energy and chemicals production. This method of biofuel production has received a huge amount of effort and research interest by way of identifying new organisms (increasing biodiversity) or optimising the engineering of the production process (Aro, 2016). Considering that the algae are non-terrestrial organisms made up of very soft homogenous tissues devoid of lignocellulose which serve as biomass entirely unlike plants which have roots, stems and leaves, and also because algae can be cultivated in both sea and brackish water not suitable for agricultural purposes all year round, this approach to biofuel production surpasses the ethical concerns of diverting agricultural resources such as land and water for crop production, and pre-treatment challenges associated with first- and second-generation

biofuels (Hays and Ducat, 2015, Naik *et al.,* 2010, Olguín, 2012). However, challenges relating to suitability of algal strains with high yields, optimising culture techniques and conditions, strain modification and difficulty of cell recovery still constrain the rapid development of this technology (Gareet *et al.,* 2010; Hays and Ducat, 2015).

The technology described as “fourth generation” for biofuel production also takes advantage of the synthetic capability of genetically modified cyanobacteria, macro- and microalgae which serve as the biomass source for biofuel production (Figure 1.3) (Godbole *et al.,* 2021; Abdullah *et al.,* 2019). Some common genetic modifications that could be done to improve algal biomass yield and increase their biofuel production capacity include increasing their photosynthesizing efficiency by enhancing light penetration and reducing photoinhibition. Other genetic modification to the algae could be in the form of metabolic engineering targeting genes for the direct synthesis of biofuels, improved nutrient use and hydrogen production, enhanced cell disruption and bio flocculation, improved lipid, and carbohydrate synthesis (Radakovits *et al.,* 2010). Biofuels from genetically modified algae are a great alternative to fossil fuels, but the potential environmental and health-related risks from the cultivation, processing and disposal of genetically engineered organisms are of great concern (Abdullah *et al.,* 2019).

Of all the classes of biofuels discussed above, only first and second generations have been commercialised. The third and fourth generations are still at basic stages of research and development with challenges such as high costs, insufficient biomass production, environment and health hazards posed by the release of toxic algal strains standing in the way of their commercial application in spite of their numerous advantages mentioned above (Godbole *et al.,* 2021; Aro, 2016).

1.5.2 Material products (Biochemicals / Biomaterials)

Aside from the production of energy, other high-value products (chemicals and materials) can be generated from plant biomass in a biorefinery. Following effective hydrolysis of the polysaccharide components in lignocellulose, five-carbon (xylose, arabinose) and six-carbon (mannose, galactose, rhamnose) membered monosaccharides can be produced (Takkellapati *et al.,* 2018). These simple sugars

can then be converted into several building block molecules; collectively called the Bio-Platform Molecules (Bio-PMs), that can potentially be used as substrates in the production of an array of value-added chemicals via fermentations or chemical synthesis (Cherubini, 2010).

The USDOE in 2004, compiled a list of the most promising Bio-PMs to include succinic acid, fumaric acid, malic acid, glycerol, 3-hydroxypropanoic acid, L-aspartic acid, 3-hydroxybutyrolactone, xylitol/arabinitol, L-glutamic acid, itaconic acid, levulinic acid, 2, 5- furan-dicarboxylic acid (2,5-FDCA), glucaric acid, and sorbitol (Table 1.2). Ethanol and lactic acid were not included in the list as they were considered to have been considerably researched and their use already at an advanced stage (Cherubini, 2010; Isikgor and Becer, 2015, Takkellapati *et al.,* 2018). The USDOE list was updated in 2010 to include ethanol, lactic acid, furfural, hydroxymethyl furfural, and isoprene. Except for glycerol and isoprene, all the above listed chemicals can be derived from the carbohydrate components of biomass (Takkellapati *et al.,* 2018).

Table 1.2 Bio-Platform Molecules from biomass sugars.
(Extracted from Cherubini *et al.,* 2010; Ahmad, 2010; Takkellapati *et al.,* 2018; Mussatto and Dragone 2016; Martin-Dominguez *et al.,* 2018; West, 2017; Weiss *et al.,* 2020)

S/No.	Bio-platform chemical	Function/ uses
1.	Ethanol	Used as biofuel (bioethanol), for production of ethylene, propylene and butanediene (building blocks for polymer synthesis), PVC from ethylene. Can also be converted into acetaldehyde and acetic acid (chemicals)
2.	Succinic acid	Production of completely bio-renewable polyester copolymers of succinic acid and 1,4-butanediol such as polyethylene succinate (PES), polypropylene succinate (PPS), polybutylene succinate (PBS).

3.	Fumaric acid	As an acidulant, preservative and flavouring agent in food and feed industry and an acid sizing agent in pulp and paper industry. Precursor for production of other acids, like L-aspartic and L-malic acid that are also used in beverages, health drinks, and cosmetics. Production of unsaturated polyester resins (UPEs), a family of polymers with applications as coatings, insulating materials, drug delivery systems and biomedical applications etc.
4.	L-Malic acid	Industrial applications in foods and pharmaceuticals as an acidulant and flavour enhancer. Other commercial applications are in metal cleaning, textile finishing, pharmaceuticals, and agriculture.
5.	2,5-Furan dicarboxylic acid (2, 5- FDCA)	Many potential applications in polyesters, polyamides, and plasticizers E.g., production of polyethylene furanoates (PEF), a stronger alternative to PET used to make plastic bottles and food packagings.
6.	3-hydroxypropanoic acid/ aldehyde	Can be converted to acrolein, acrylonitrile (for synthesis of various polymers), acrylic acid, acrylic acid esters, 1,3-propanediol, malonic acid and 3-hydroxypropionic esters. Used to make polyacrylamide which is used in various applications such as in water treatment, paper manufacture, mining, oil recovery, absorbents and as electrophoresis gels.
7.	Sorbitol	Used as a sugar substitute in food, beverages, drugs, cosmetics. Can be converted to glycerol, propyleneglycol, ethylene glycol, ethanol, and methanol. These lower alcohols can then be further converted to biodegradable polymers in biocomposites and biomedicines. e.g poly (isosorbide carbonate), a promising alternative to the petroleum-based Bisphenol A (BPA) polycarbonate.
8.	Levulinic acid	Potential for substituting petroleum-based chemicals. E.g., levulinic acid derived diphenolic acid (DPA) can serve as a substitute for bisphenol-A (BPA) in food containers and consumer products. Can also be converted into various higher value-added products such as levulinic acid esters, 5-aminolevulinic acid, valeric acid, γ-valerolactone, and 2-methyltetrahydrofuran. Acts as a building block in many applications such as pharmaceuticals, plasticizers, fragrances, and cosmetics.

9.	L- Aspartic acid	Serves as salts for chelating agents. Used to produce fumaric/maleic acid, 2-amino-1,4-butanediol, β-alanine, aspartic anhydride, amino-γ-butyrolactone, water soluble biodegradable polymers such as Polyaspartic acids (PASA) and Thermal polyaspartate (TPA) for production of performance chemicals, diapers and agricultural chemicals
10.	Glucaric acid	Prevents deposits of limescale and dirts on fabric or dishes hence can be used as a green replacement for phosphate-based detergents.
11.	L-Glutamic acid	Monomers for polyesters and polyamides such as dimethyl glutarate and poly-γ-glutamic acid (γ-PGA). γ-PGA is water-soluble, edible, biodegradable, biocompatible and non-toxic for humans and the environment. Hence, γ-PGA and its derivatives have applications in food, cosmetics, medicine, and water treatment industries.
12.	Itaconic acid	A replacement for petroleum based acrylic acid. Used to make absorbent materials for nappies and resins used in high performance marine and automotive components. As a copolymer with acrylic acid and in styrene-butadiene systems
13.	3-Hydroxybutyrolactone (3HBL)	Intermediate for high value pharma compounds e.g the statin class of cholesterol-reducing drugs such as Crestor and Lipitor, as well as the antibiotic Zyvox, and the antihyperlipidemic medication Zetia. Other pharmaceuticals derived from 3HBL include HIV inhibitors and the nutritional supplement L-carnitine.
14.	Xylitol	The metabolism of xylitol is not dependent on insulin; thus, it is an ideal sugar substitute for people with diabetes as it is 20% sweeter than sucrose, but without 40% of the calories.
15.	Furfural	Used in production of furfuryl alcohol and can be converted to succinic and levulinic acids. Used extensively in plastics, pharma and agro chemical industries, and as adhesives and flavour enhancers
16.	Hydroxymethyl furfural	Can be converted into 2,5-furandicarboxylic acid (FDCA), 2,5-bis(hydroxymethyl)furan, and potential biofuels 2,5-dimethylfuran, 5-ethoxymethylfurfural, ethyl levulinate, and γ-valerolactone. Also converted to 1,6-hexanediol (1,6-HDO), used in the preparation of polycarbonatediols for production of polyurethanes for use in coatings, elastomers,

		and adhesives.1,6-HDO can be converted to 1,6-heaxanediamine and ε-caprolactone, which are used in the synthesis of various polymers
17.	Lactic acid	Can undergo various conversions to produce propylene glycol, propylene oxide, acrylic acid, and esters, and polylactic acid used to make biodegradable plastics
18.	Glycerol	Hydrogenation to produce ethylene glycol, propylene glycol, and acetol. Production of 1,2-propanediol, 1,3-propanediol, glycerol carbonate used in synthesis of gycidol and in polymers, coatings, adhesives, and lubricants. Production of other chemicals such as epichlorohydrin, glyceraldehyde, dihydroxyacetone, glyceric acid,1-butanol, 2,3-butanediol, ethanol, lactic acid, succinic acid, propionic acid, and dihydroxyacetone
19.	Isoprene	Can be converted into the polyisoprene polymer, which is used in a variety of products such as footwear, mechanical instruments, medical appliances, sporting goods, and rubber tyres.

These Bio-PMs possess several functional groups in their structures available for numerous potentially possible reactions. Therefore, contrary to the situation in oil-based chemical industries whereby desired functionalities are being added, the Bio-PMs will already possess most of the desired functionality or pre-functionality and this translates into a greater potential for the production of a wider variety of products from bio-based industries than from oil-based ones (Isikgor and Becer, 2015). Again, Bio-PMs are richer in oxygen content compared to their oil-derived counterparts such as ethylene, benzene, etc. This implies that their chemistries and reactions will mostly be reductions which are “greener chemistries” compared to the harsh and environmentally unfriendly oxidation chemistries that are seen with fossil-based molecules (Cherubini, 2010).

Although the lignin component of lignocellulose is difficult to breakdown, it is non-toxic, versatile, and highly available directly from plants or as by-products from industrial conversions of lignocellulose. A wide range of products, particularly aromatic chemicals can be produced from lignin in a biorefinery (Isikgor and Becer, 2015). Considering that the β-aryl ether and biphenyl linkages are the predominant linkage types found in lignin,

industrially relevant chemicals, and molecules of high value such as vanillin, benzoic acid, cinnamic acid, vinyl guaiacol, adipic acid, optically active lignans, ferulic acid, dimers of monolignols and p-coumaric acid, etc can be generated from the β-aryl ether and biphenyl catabolic pathways (Bugg *et al.,* 2011b; Mathews *et al.,* 2016; Lambertz *et al.,* 2016). Also, through the β-ketoadipate pathway, aromatic units derived from the degradation of lignin can be used to synthesize high molecular weight compounds such as lipids and polyhydroxyalkanoates (PHAs) (Figure 1.4) (Chen and wan, 2017; IEA bioenergy Task 42 report).

The Lignocellulose Biorefinery Network (LBNet), which has now been integrated into the Biomass Biorefinery Network (BBNet), a BBSRC (Biotechnology and Biological Sciences Research Council) funded network in the UK, identified top 10 lignocellulose-derived chemicals; Lactic acid, 2,5- Furandicarboxylic acid (FDCA), levoglucosenone, 5 hydroxymethyl furfural (HMF), muconic acid, Itaconic acid, 1,3 butanediol, glucaric acid, levulinic acid, n-butanol) as high value chemicals which can replace petrochemicals to produce a wide range of products creating a more sustainable bio-economy that can generate billions of pounds. From a report following its 3rd international conference in 2018, government and businesses were urged to focus attention and invest in these specific biochemicals to grow the bioeconomy and maintain the UK's position as a world leader in biochemicals production (LBNet and BBSRC- NIBB, 2019; Barret, 2018).

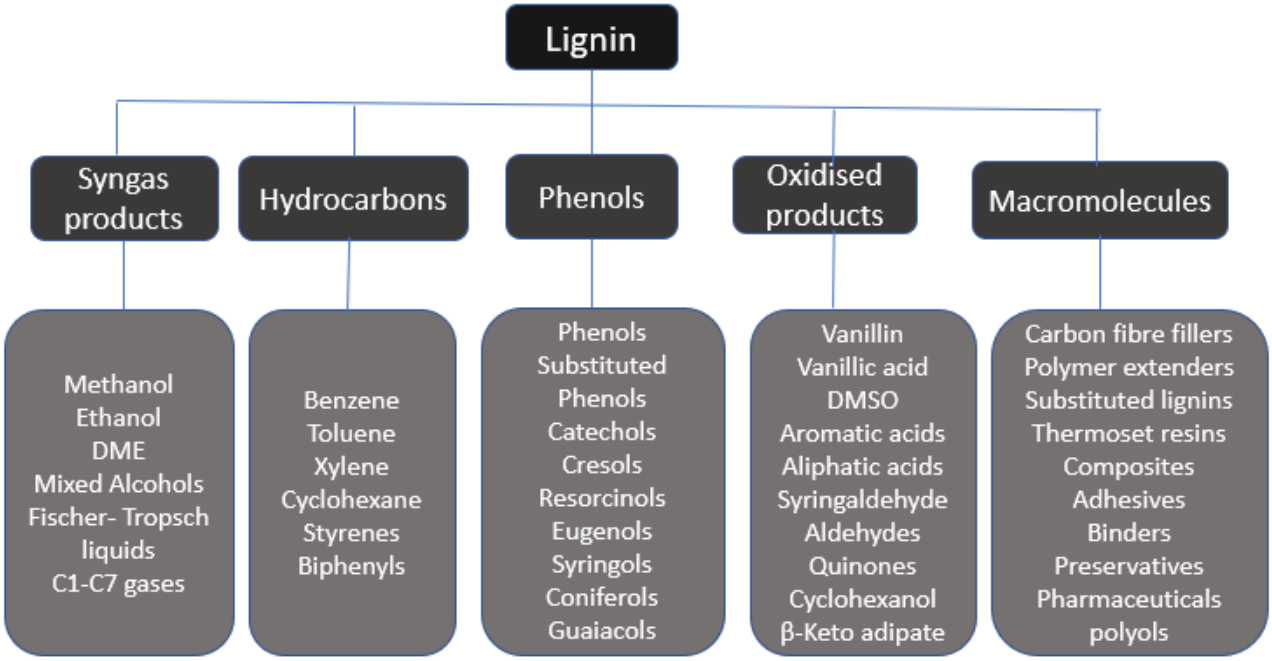


Figure 1.4 Potential products from lignin valorisation
(Modified from IEA bioenergy-task 42 report)

Currently, the chemical industry in the UK employs over 105,000 people and generates about £9 billion gross added value each year (E4tech (UK) Ltd, 2015). Over 25 companies and about 10 universities are actively developing and producing some of these bio-based products, translating potentials into commercial realities. While some are already in the market, a lot more are still in the pipeline being developed.

Table 1.3 Some lignocellulose based chemical industries and collaborating universities producing bio-based products
(Silva *et al.,* 2018; Cherubini 2010; Takkellapati *et al.,* 2018, IEA bioenergy Task 42 report; E4tech (UK) Ltd, 2015)

S/No.	Bio-platform chemical	Company/ collaborating UK university
1.	Ethanol	GranBio, Chempolis biorefinery, Cometha projects, Ensus, Biogasol, Instilbio, Corbion etc
2.	Succinic acid	BIoAmber, MyrIant, BASF, Purac, Reverdia (DSM/ Roquette), PTT Chem/ Mitsubishi CC University of Huddersfield
3.	Fumaric acid	DSM and Myriant,
4.	L-Malic acid	Novozymes

5.	2,5-Furan dicarboxylic acid (2, 5- FDCA)	Avantium University of Huddersfield, Imperial College London, Aston University, University of Liverpool, The University of York, The University of Manchester, and Biome Bioplastics
6.	3-hydroxypropanoic acid/ aldehyde	Cargill
7.	Sorbitol	Roquette, Archer Daniels Midland (ADM)
8.	Levulinic acid	Malne BioProducts, Avantium, SegetIs, Circa group, Biofine, DSM, Segetis, GF Biochemicals. Aston university
9.	Glucaric acid	Rivertop renewables and Cargill, Rennovia, Johnson Matthey
10.	L-Glutamic acid	Global Biotech, Melhua, Fufeng, Juhua
11.	Itaconic acid	Itaconix, Qingdao Kehal Biochemistry Co. The University of York and The University of Nottingham
12.	Xylitol	Danisco/Lenzing, Xylitol Canada, S2G BioChem.
13.	Hydroxymethyl furfural (HMF)	Imperial College and at The University of Liverpool
14.	Lactic acid	Purac, NatureWorks,Galactic, Henan Jindan, BBCA, Cellulac, Plaxica/ Imperial college London, Rebio/CPI UK/ University of Bath
15.	Isoprene	Goodyear/Genencor, GlycosBio, Amyris/ DuPont/ Michelin, Lanza Tech Aemetis.
16.	Levoglucosenone	Circa Sustainable Chemicals UK. University of York, University of Huddersfield

Natureworks remains the leading producer of lactic acid in the UK after they acquired Plaxica's D-lactic acid process technology, and Cellulac are producing pure lactic acid from lactose whey (Takkellapati *et al.,* 2018; IEA bioenergy task42 report). FDCA development is one of the UK's areas of research strength with Biome bioplastics, the University of Liverpool, the University of York, and the University of Manchester actively engaged (Isikgor and Becer, 2015). Other areas of strength where the UK is well positioned include muconic acid development (E4tech (UK) Ltd, 2015). The UK companies, Green biologics and Solvert are seeking to commercialise production of bio-based butanol (IEA bioenergy task 42 report). Due to the well-established, and synthetic biology capabilities of UK universities, development of itaconic acid and its

polymer derivatives are carried out at Itaconix, the University of York and the University of Nottingham. As of now, only Circa sustainable chemicals company is producing levoglucosenone commercially (Takkellapati *et al.,* 2018). At Imperial College London and the University of Liverpool, research is ongoing on development of HMF technologies, while Aston university is researching levulinic acid. American companies such as Goodyear, Michelin, DuPont, Amyris, GlycosBios, and Aemetis are working on development of bio-based and fermentative isoprene, however, there has not been any known isoprene activities in the UK (E4tech (UK) Ltd, 2015).

Overall, there is a range of promising bio-based chemicals with good market opportunities as a result of improved functionality and greater sustainability. The development of these bio-based chemicals and their derivatives is still at a stage where it is possible to innovate and compete, and the UK has promising strengths. The economic value of the markets that could be accessed by these bio-based chemicals is very large. There is therefore a strong rationale for investing in this area, though investments should follow more careful and detailed assessments of the technical and economic prospects of the specific bio-based chemical production pathways (IEA bioenergy Task 42 report: E4tech (UK) Ltd, 2015).

From a technical standpoint, majority of industrial chemicals and materials currently derived from fossil fuel resources could be replaced by their lignocellulosic biomass derived counterparts thereby offering significant solutions to the problems associated with utilization of fossil-based products (Mathews *et al.,* 2016). Some of the most important bio-based products from biomass, some of which are already commercially available include fuels, chemicals (fine chemicals, building blocks chemicals, bulk chemicals, solvents and sorbents), organic acids (succinic, lactic, itaconic and other sugar derivatives), polymers and resins (starch-based plastics, phenol resins, furan resins), biomaterials (wood panels, pulp, paper, cellulose, carbon fibre,), food and animal feed, detergents and cleaning agents, pharmaceuticals, fertilizers, paints and coatings, etc. (Table 1.3 and Figure 1.4) (IEA bioenergy- task42 report, Cherubini, 2010; Bugg and Rhamanpour, 2015; Ragauskas *et al.,* 2014; Mussatto and Dragone, 2016).

International organisations such as the International Energy Agency (IEA), Organisation for Economic Co-operation and Development (OECD), World Economic Forum (WEF) as well as national and regional governments have continued to emphasize the need for a transition from our current fossil-dependent economy to a more sustainable bioeconomy (Takkellapati *et al.,* 2018). The obvious reasons driving the need for this transition include the necessity to develop a global economy that is environmentally, socially and economically more sustainable, the reduction in GHG emissions, and to minimise global over-reliance on fossil resources (Takkellapati *et al.,* 2018, Weiss *et al.,* 2020) especially now that prices of oil is on the rise and supply is limited, the population is fast growing, and there is increased public awareness about detrimental consequences of fossil based products resulting in increased demand for environmentally friendly products. All these have now opened new windows of opportunities for bio-based chemicals and polymers with much investment interest from industry. However, some of the constraints hampering the transition to a bio-based economy have to do with current high cost of producing biobased products which far exceeds that for petrochemicals production arising from high cost of obtaining lignocellulosic feedstock, biorefineries operating at below maximum capacity and mostly generating only a single product thereby not exploiting the full energy potential of lignocellulose (Silva *et al.,* 2018). Therefore, to maximise the economic sustainability of a biorefinery, cost effective methods of production of a wide spectrum of marketable products must be achieved, aiming for net zero waste. Biobased products must also be proven to perform better or at least, as good as their petrochemical equivalents by being effective and having lower environmental impact (IEA-bioenergy task 42).

1.6 Microbial degradation of lignocellulose/lignin

In industrial bio-refining processes, a large amount of the costs and pollution produced are derived from the process of delignification during biomass pre-treatment (Tsegaye *et al*., 2019; Cragg *et al.,* 2015; Angzzas *et al.,* 2016). Physico-chemical pre-treatment methods are costly and contribute to the pollution and degradation of the environment and may also alter the structure of lignocellulosics thereby decreasing the yield of fermentable sugar or leading to the generation of by-products that inhibit enzymatic

hydrolysis and fermentation, hence biological pre-treatment methods that would eliminate or substantially replace thermochemical pre-treatment steps altogether are desired (Fisher and Fong, 2014).

In nature, microorganisms have acquired several sophisticated enzymatic strategies to decompose and utilize the various components of lignocellulosic biomass, and for survival in different environments (Janusz *et al.,* 2017, Chen and Wan, 2017). Many new strains of ligninolytic microbes have commonly been isolated from compost, sludge, forest soil and animal guts to facilitate the bioconversion of lignocellulose into useful chemicals (Chen and Wan, 2017). The primary organisms identified as lignin degraders include the white rot basidiomycetes and some ligninolytic bacteria. They are known to secrete ligninolytic enzymes such as laccases and peroxidases (de Gonzalo *et al.,* 2016; Cragg *et al.,* 2015).

1.6.1 Lignin degradation by fungi

Fungi are the organisms more predominantly studied that are capable of degrading lignin in both hard and soft woods. Macro fungi belong to two classes: brown-rot and white-rot, depending on their pattern of decay. The mechanism of modifying and degrading lignin has been studied extensively in basidiomycetes, with more attention focussed on the white-rot fungi than brown-rot fungi (Ahmad, 2010). These studies have led to the discovery of several enzymes, and elucidation of mechanisms of action involved in lignin depolymerisation (Arumugam *et al.,* 2014; Bugg *et al.,* 2011a)**.** White-rot fungi produce heme-containing peroxidases such as lignin peroxidases (LiP), manganese peroxidases (MnP), versatile peroxidases (VP), dye decolorizing peroxidases (DyP) and laccases as well (Lambertz *et al.,* 2016). These enzymes are usually secreted extracellularly by the fungi to assist in the degradation of lignin, however, there is not a single fungal species that can produce all these enzymes. For example, lignin and manganese peroxidases are produced by *Phanerochaete chrysosporium* but this fungus cannot produce versatile peroxidase or laccases (Ahmad, 2010).

Degradation of lignin by fungi is an oxidative and non-specific process that brings about a decrease in the methoxy, phenoxy and aliphatic content of lignin, cleaves aromatic

rings and forms new carbonyl groups by the action of these enzymes with a consequent depolymerisation of the lignin molecule and carbon dioxide production (Arumugam *et al.,* 2014). The best studied and most extensively characterised fungus that has been used as a model species for the study of fungal lignin degradation is *Phanerochaete chrysosporium.* However, other species such as *Pleurotus ostreatus*, *Coriolus versicolor*, *Cyathus stercoreus* and *Ceriporiopsis subvermispora* have also been studied (Fisher and Fong, 2014, Lai *et al.,* 2017). Lignin degrading enzymes have also been produced by *Bjerkandera adusta, Tinea versicolor*, *Irplex lacteus, Dichomitus squalens, Thanatephorus cucumeris, Trametes versicolor, Phlebia tremellosa, and Pinus radiata* so far (Cagide and Castro-Sowinski, 2020; Sahadevan *et al.,* 2016, Tsegaye *et al*., 2019).

1.6.2 Lignin degradation by bacteria

Research on bacteria capable of lignin degradation have only recently gained much attention as most studies of microbial lignin degradation have centred on fungi (Bugg *et al.,* 2011a). The ability to directly degrade and modify lignin has been shown in several bacterial phyla such as Proteobacteria, some Firmicutes and, Actinobacteria, the majority of which were obtained from the guts of ruminants, termites, and other wood-feeding insects (Bugg *et al.,* 2011b; Huang *et al.,* 2012; Arumugam *et al.,* 2014; Bugg and Rahmanpour, 2015; Kassim *et al.,* 2016; Janusz *et al.,* 2017).

So far, *Streptomyces viridosporus* is the organism most studied with respect to lignin degradation by bacteria (Fisher and Fong, 2014; Bugg *et al.,* 2011a). They have been found to secrete peroxidases that can digest the β-aryl ether bonds of lignin to release low molecular weight phenols. *Thermobifida fusca* is a thermophilic bacterium that breaks down cellulose. It has also been shown to partially degrade lignin in lignocellulose from pulp. *T. fusca* also forms acid precipitable polymeric lignin (APPL) at higher levels compared to *S. viridosporus*. Ahmad, (2010) identified *Pseudomonas putida* and *Rhodococcus jostii* as effective lignin degrading bacteria from their research that employed novel spectrophotometric assays for determination of lignin breakdown ability by different organisms (Ahmad, 2010). Angzzas *et al.,* 2016 identified four major species of bacteria from the genera of *Klebsiella, Enterobacter, Serratia* and

Pseudomonas when they cultured the gut microbiome of *Rhynchophorus ferrugineus* larvae on minimal salt media enriched with 0.5% lignin as sole carbon source (Angzzas *et al.,* 2016). Other bacteria identified and shown to possess lignin degrading or modifying ability from several other research outcomes have been compiled and are presented in Table 1.4 below.

Table 1.4 Lignin degrading bacterial genera identified from literature reports

Phyla	Genus	Reference(s)
Proteobacteria		
	Pseudomonas	Bugg and Rhamanpour, 2015; Ahmad, 2010; Janusz *et al.,* 2017; de Gonzalo *et al.,* 2016; Chen and Wan, 2017; Li et al., 2009; Gong *et al.,* 2017, Bugg *et al.,* 2020, Chauhan 2020; Beckham *et al.,* 2016; Kumar and Chandra, 2020.
	Xanthomonas	Ahmad, 2010; Bugg *et al.,* 2011a, Ceballos *et al.,* 2017
	Acinetobacter	Ahmad, 2010; Fisher and Fong, 2014; Bugg *et al.,* 2011a; Chauhan, 2020.
	Variovorax	Janusz *et al.,* 2017; Bugg *et al.,* 2011b; Woo *et al.,* 2017
	Aeromonas	Bugg *et al.,* 2011a; Ahmad, 2010
	Klebsiella	Angzzas *et al.,* 2016; Janusz *et al.,* 2017; Kameshwar and Qin, 2017b; Bugg *et al.,* 2011a; Chauhan 2020; Beckham *et al.,* 2016
	Enterobacter	Angzzas *et al.,* 2016; Janusz *et al.,* 2017; Bugg *et al.,* 2011a; Chauhan 2020; Bugg *et al.,* 2020, Kumar and Chandra, 2020; Deangelis *et al.,* 2013
	Serratia	Angzzas *et al.,* 2016; Bugg *et al.,* 2011a; Chauhan 2020; Kumar and Chandra, 2020
	Sphingobium	Janusz *et al.,* Chen and Wan, 2017; de Gonzalo *et al.,* 2016; Bugg and Rhamanpour, 2015; Xie et al., 2014; Pollegioni *et al.,* 2014; Kameshwar and Qin, 2017b
	Sphingomonas	Bugg *et al.,* 2011a; Janusz *et al.,* 2017; Fisher and Fong, 2014; Bugg *et al.,* 2020; Cragg et al., 2015

	Ochrobactrum	Taylor *et al.,* 2012; Janusz *et al.,* 2017; Kumar and Chandra, 2020, Bugg *et al.,* 2011b, Bugg *et al.*, 2020
	Pantoea	de Gonzalo *et al.,* 2016, Kumar and Chandra, 2020
	Comamonas	Bugg *et al.,* 2011b; Chauhan, 2020, Chen et al., 2012
	Escherichia	Ceballos *et al.,* 2017; Janusz *et al.,* 2017; de Gonzalo *et al.,* 2016, Kumar and Chandra, 2020; Cagide and Castro-sowinski, 2020
	Rhizobium	Kameshwar and Qin, 2017b
	Raoultella	Kameshwar and Qin, 2017b; Chauhan 2020
	Brucella	Ceballos *et al.,* 2017; Bugg *et al.,* 2011a; Janusz *et al.,* 2017
	Citrobacter	Chauhan, 2020; Ceballos *et al.,* 2017; Bugg *et al.,* 2011b; Asina *et al.,* 2016
	Pandoraea	Kumar *et al.,* 2018; Chauhan, 2020; Chen and Wan, 2017; Bugg *et al.,* 2011b; Asina *et al.,* 2016
	Burkholderia	Bugg *et al.,* 2011b; Kameshwar and Qin, 2017b, Cragg *et al.,* 2015; Ceballos *et al.,* 2017
	Novosphingobium	Chen and Wan, 2017; Bugg *et al.,* 2011a; Kameshwar and Qin, 2017b; Bugg *et al.*, 2020; Asina *et al.,* 2016
	Shigella	Chauhan, 2020
	Delftia	Chauhan, 2020
Actinobacteria		
	Actinomadura	Ahmad, 2010
	Atrobacter	Ahmad, 2010; Kameshwar and Qin, 2017b; Xie *et al.,* 2014; Li *et al.,* 2009; Bugg *et al.,* 2011b
	Corynebacterium	Ahmad *et al.,* 2011; Li *et al.,* 2009; Bugg *et al.,* 2011a
	Mycobacterium	Ahmad *et al.,* 2011; Bugg *et al.,* 2011a; Ceballos *et al.,* 2017; Bugg *et al.,* 2011b
	Rhodococcus	Fisher and Fong 2014; Bugg and Rhamanpour, 2015; Ahmad, 2010; Janusz *et al.,* 2017; de Gonzalo *et al.,* 2016; Chen and Wan, 2017; Li *et al.,* 2009; Bugg *et al.,* 2020, Chauhan 2020; Cragg *et al.,* 2015; Pollegioni *et al.,* 2014

	Nocardia	Ahmad, 2010; Bugg *et al.,* 2011a; Li *et al.,* 2009; Woo *et al.,* 2017; Xie *et al.,* 2014; Lai *et al.,* 2016; Chauhan 2020
	Streptomyces	Brown and Chang, 2014; Kameshwar and Qin, 2017b; Fisher and Fong 2014; Bugg and Rhamanpour, 2015; Ahmad, 2010; Janusz *et al.,* 2017; de Gonzalo *et al.,* 2016; Chen and Wan, 2017; Li *et al.,* 2009; Bugg *et al.,* 2020, Chauhan 2020; Cragg *et al.,* 2015; Pollegioni *et al.,* 2014; Asina *et al.,* 2016
	Saccharomonospora	Ahmad, 2010; Bugg *et al.,* 2020
	Thermomonospora	Chen *et al.,* 2015; Blooise *et al.,* 2010; Ahmad, 2010; Bugg *et al.,* 2020; de Gonzalo *et al.,* 2016;
	Microbacterium	Taylor *et al.,* 2012, Bugg *et al.,* 2011b; Ceballos *et al.,* 2017; Kameshwar and Qin, 2017b
	Amycolatopsis	Cragg *et al.,* 2015; Bugg and Rhamanpour, 2015; Kameshwar and Qin, 2017b; de Gonzalo *et al.,* 2016; Bugg *et al.,* 2020; Brown and Chang, 2014; Pollegioni *et al.,* 2014, Beckham *et al.,* 2016
	Alcaligenes	Li *et al.,* 2009; Ceballos *et al.,* 2017
	Rubrobacter	Ceballos *et al.,* 2017
	Leucobacter	Chauhan 2020
	Azotobacter	Kumar and Chandra 2020, Chauhan 2020
Firmicutes		
	Paenibacillus	Ahmad *et al.,* 2011; Woo *et al.,* 2017, Kumar and Chandra, 2020
	Bacillus	Ahmad *et al.,* 2011; Woo *et al.,* 2017; de Gonzalo et al., 2016; Chen and Wan, 2017; Bugg et al., 2011a; Gong et al., 2017; Xie et al., 2014; Lai *et al.,* 2016; Pollegioni *et al.,* 2014; Kameshwar and Qin, 2017b; Min et al., 2015; Chauhan 2020; Asina *et al.,* 2016
	Aneurinibacillus	Ahmad *et al.,* 2011; Chen and Wan, 2017; Bugg *et al.,* 2011a, Kumar and Chandra, 2020
Bacteroidetes		

	Bacteroides	Cragg *et al.,* 2015; Bugg and Rhamanpour, 2015
	Vogesella	Woo *et al.,* 2017
	Cupriavidus	Xie *et al.,* 2014; Kameshwar and Qin, 2017b

All these research findings point to the fact that the role bacteria play in lignin deconstruction is more significant than it was previously assumed. However, despite the growing evidence from decades of studies that have identified several lignin degrading bacteria, a lot still needs to be done on the enzymology of this process as only a few enzymes of bacterial origin have been produced, purified, and conclusively reported to depolymerise lignin (Ahmad, 2010; de Gonzalo *et al.,* 2016).

1.6.3 Why interest in bacterial rather than fungal lignin degradation?

Despite the available information and evidence about the ability of fungi to secrete enzymes that can degrade lignin, these enzymes have not been developed for large-scale, industrial application due to several challenges. Most of the fungal enzymes are unable to efficiently decompose lignin at the temperature and pH extremes, and low oxygen or anaerobic conditions which are characteristic of industrial processes as they are more effective at lower pH (4–7) and temperatures (Yang *et al.* 2011; Mathews *et al.,* 2016). However, there are certain bacteria (extremophiles) that can resist extreme conditions of temperature and pH and hence produce enzymes that can function under conditions that are comparable to those obtainable in industrial plants. Also, in general, the practicality of fungal protein expression and its genetic manipulation is quite challenging contrary to bacteria which are more amenable to genetic manipulation due to their ability to adapt to different environments, and being biochemically versatile (Li *et al.,* 2009; Ahmad 2010; Bugg *et al.,* 2011b; Bugg and Rhamanpour 2015). Fungi take a longer time to grow (they require a minimum of 2–4 weeks) hence the rate of enzyme production will be equally slow unlike with bacteria which are easy to culture, often exhibit rapid growth rates and can allow for higher recombinant production of enzymes in hosts such as *Escherichia coli* (Sahadevan *et al.,* 2016). Bacteria can withstand environmental stress better as they are biochemically versatile, with the

ability to adapt to changes in temperature, salinity, pH, and oxygen availability (Chukwuma *et al.,* 2021)

The desire to overcome these challenges and achieve enzymatic degradation of lignocellulose has spurred research interests in diverse fields such as molecular genetics, enzyme engineering, metabolic engineering, and other allied fields (Huang *et al.,* 2012; Bugg and Rahmanpour, 2015).

1.7 Lignin Degrading Enzymes

Until recently, most research efforts on the biodegradation of lignocellulose have focussed on the breakdown of the polysaccharide components which brought about the discovery of several cellulases and hemicellulases with limited attention being paid to lignin as it was considered a low value product because of its resistance to degradation (Pollegioni *et al.,* 2015).

The increased acknowledgement that the controlled and selective hydrolysis of the C-O and C-C bonds of lignin could produce a series of monomeric and aromatic molecules has seen the upsurge of research efforts and interest in lignin valorisation strategies. These molecules could represent sources of renewable aromatic chemicals in a biorefinery. Detailed studies surrounding the enzymology of ligninolytic enzymes is being intensified and several classes of enzymes potentially possessing ligninolytic activity have been identified from lignin-degrading fungi and bacteria (Fisher and Fong, 2014).

Biological degradation of lignin could be described as an “enzymatic combustion” involving several enzymes that have high-redox-potentials, that exploit the oxidising capacity of enzymatically generated hydrogen peroxide or molecular oxygen to oxidise aromatic units (Fisher and Fong, 2014, Ruiz-Duenas and Martinez, 2009). It is more of an aerobic oxidation process rather than hydrolysis as seen with cellulose and hemicellulose (Ruiz-Duenas and Martinez, 2009).

The oxidative enzymes involved in aerobic degradation of lignin can be classified into two main categories; peroxidases and phenol oxidases (laccases and polyphenol oxidases) (Li *et al.,* 2009; Sun and Zhou, 2011). However, lignin degrading enzymes can generally belong to one of two groups: Lignin-modifying enzymes-LME classified

as lignin peroxidase-LiP, Manganese peroxidase-MnP, Versatile peroxidase-VP, the recently classified DyP-type peroxidase and laccases, or lignin-degrading auxiliary enzymes -LDA such as quinone reductases (QR), glyoxal oxidase (GOx), glucose oxidase (GO), pyranose 2-oxidase (POx), aryl alcohol oxidases (AAO), aryl-alcohol dehydrogenases (AAD), cellobiose dehydrogenase (CBD), etc (Xie *et al.,* 2014). The classification as LME or LDA is based on the degree to which these enzymes are involved in lignin degradation. While LMEs are directly involved with the breaking of bonds to deconstruct lignin, LDA enzymes are not capable of independently degrading lignin, but they perform accessory functions required to bring about complete lignin degradation (Chen and Wan, 2017; Janusz *et al.,* 2017).

Chloroperoxidases (CPO, EC 1.11.1.10) and aromatic peroxygenases (APO, EC 1.11.2.1) are heme-thiolate haloperoxidases (HTPs) which are a recent addition of a group of enzymes that share catalytic properties with at least three other groups of heme-containing oxidoreductases and have been suggested to have a role in lignin degradation. These HTP enzymes have neither been classified as LME nor LDA group members but since there is no experimental evidence that they can degrade lignin alone, it is suggested that they be considered as LDA enzymes (Janusz *et al.,* 2017).

1.7.1 Lignin Modifying Enzymes (LMEs)

Lignin modifying enzymes are either peroxidases or laccases

1.7.1.1 Peroxidases

Peroxidases are heme-containing oxidoreductases that catalyse several oxidative and hydroxylation reactions of a wide range of substrates including phenols, aromatic amines, and other compounds such as alkyl peroxides and aromatic per acids, using hydrogen peroxide (H_2O_2) as the electron acceptor. The strategy adopted by ligninolytic peroxidases is based on nonspecific one electron oxidation of the benzenic rings in the different lignin substrates in synergy with oxidases that generate hydrogen peroxide. Ligninolytic peroxidases share similar folding structure to cytochrome c peroxidases (CCP) by having 10 helices making up about half the total percentage of the molecule. These peroxidases possess two domains between which a single heme group tethered

by a histidine residue is sandwiched (Ahmad, 2010) and hence they have unusually high redox potentials due to the heme pocket architecture that enables oxidation of non-phenolic aromatic groups (Ruiz-Duenas and Martinez, 2009). They are able to catalyse the cleavage of α, β and β-ether bonds (including β-O-4 linkages) leading to the efficient degradation of lignin into mono-aromatic structures, which has been demonstrated using lignin model compounds (Schoenherr *et al.,* 2018). Peroxidases are ubiquitous, found in both micro and macro-organisms (fungi, bacteria, plants, and animals) and most of them have a common catalytic cycle in which one oxidation and two reduction steps are involved as summarised below (Li *et al.,* 2009).

1. Native peroxidase (Fe^{3+}) + H_2O_2……. Compound I + H_2O
2. Compound I + AH_2……. Compound II + AH*
3. Compound II + AH_2……. Native peroxidase (Fe^{3+}) +AH* + H_2O

In reaction 1, the ferriheme prosthetic group of the native peroxidase (ferriperoxidase) undergoes a two-step oxidation by H_2O_2 or organic hydroperoxides forming an intermediate that consist of an oxo-ferryl iron and a porphyrin cation radical with an oxidation state of +5 (compound-I). In reaction 2, non-phenolic aromatic substrates donate an electron to compound-I reducing it to form compound-II which has an oxidation state of +4. In the third step of the cycle, compound-II receives another electron from the substrates, undergoes a further reduction which returns the enzyme complex back to its native state (oxidation state +3) ready for another round of the reaction cycle (Figure 1.5). Different oxidation products are formed during this reaction cycle depending on the nature of the substrates involved (Li *et al.,* 2009, Fisher and Fong, 2014).

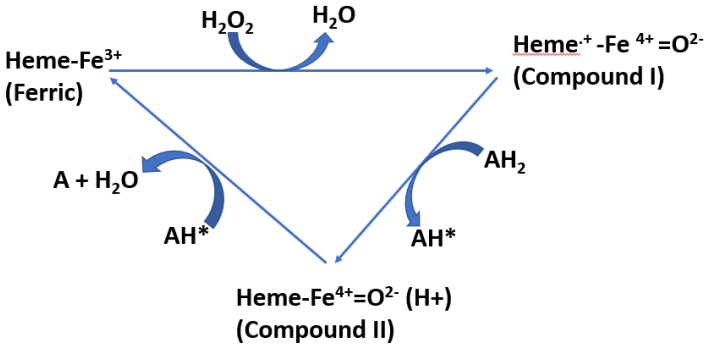


Figure 1.5 Catalytic cycle of peroxidases

1.7.1.1.1 Lignin Peroxidase

Lignin peroxidase (LiP, EC 1.11.1.14), formerly called ligninase, is a glycosylated, heme-containing protein with an iron protoporphyrin prosthetic group that requires hydrogen peroxide to catalyse the oxidation of non-phenolic lignin units and deconstruct the recalcitrant aromatic compounds. LiP is produced by most white rot fungi, and it was first discovered in *P. chrysosporium* (Bugg *et al.,* 2011a; Fisher and Fong, 2014) having a globular structure composed of 8 major and 8 minor *α*-helices arranged into two domains. The domains form an active center cavity composed of a heme-chelating single ferric ion (Choinowski *et al.* 1999). The LiP contains two glycosylation sites, two Ca^{2+} binding sites and four disulfide bridges, all stabilizing the three-dimensional structure of this enzyme. Depending on its degree of glycosylation, the molecular mass of lignin peroxidases could range between 35 to 48 kDa and it has a pI between 3.1 and 4.7 (Janusz *et al.,* 2017). LiP exhibits the same catalytic mechanism characteristic of all peroxidases (Figure 1.6).

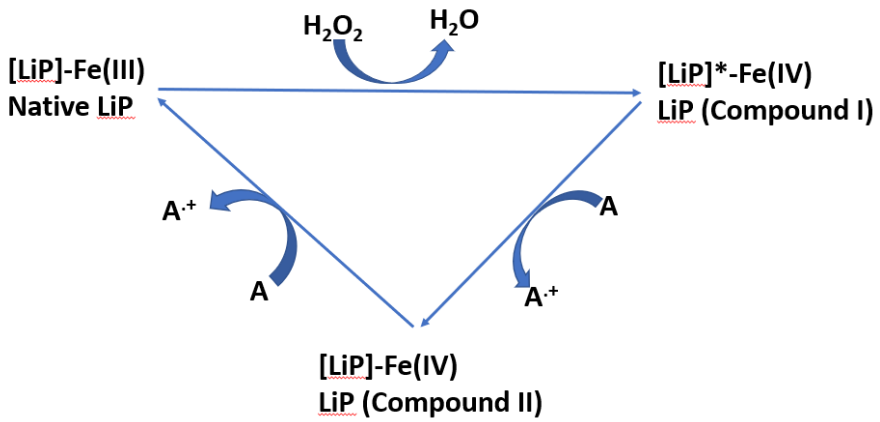


Figure 1.6 Catalytic cycle of Lignin peroxidase

Lignin peroxidase has a high redox potential (around 1.2 V at pH 3), therefore, it can catalyse the oxidation of an extremely broad range of substrates such as lignin monomers, dimers, and trimers, polycyclic aromatic compounds in lipid peroxidative pathways as well as molecules that are unrelated to lignin (Li *et al.,* 2009; Janusz *et al.,* 2017). The radicals produced during catalysis (compounds I and II) cause breakdown of sidechains, cleavage of various bonds and opening of aromatic rings bringing about the degradation of the lignin polymer (Datta *et al.,* 2017).

1.7.1.1.2. Manganese Peroxidase

Another important enzyme belonging to the LME class is manganese peroxidase (MnP, EC 1.11.1.13). It was also first detected in *P. chrysosporium* several decades ago and exists in several isoforms (Janusz *et al.,* 2017).

MnP, as a heme-containing peroxidase shares a similar catalytic cycle to LiP except that MnP utilizes Mn^{2+} as reducing substrate (electron donor), which is oxidized to Mn^{3+} by hydrogen peroxide (Figure 1.9). The reactive Mn^{3+} diffuses out of the enzyme's active site and is chelated with dicarboxylic acids, such as malate, oxalate, fumarate, glyoxylate and malonate to enhance its stability. The chelated Mn^{3+} acts as a small molecular weight, diffusible redox mediator, and is capable of oxidizing phenolic substrates and targets (but cannot oxidize non-phenolic compounds) that are far from the active site of the enzyme via hydrogen or one electron abstraction (Fisher and Fong, 2014; Chen and Wan, 2017). Compounds I and compound II can undergo

conversion by the addition of various phenolic substances which serve as electron donors, but the rate of this conversion occurs very slowly. However, compound II of MnP is not efficiently converted to its native form by phenolic compounds as Mn^{2+} is required to serve as a redox coupler for the catalytic cycle to be completed (Ahmad, 2010; Janusz *et al.,* 2017).

The molecular structure of MnP also consists of a heme group sandwiched between two *α*-helical domains just as LiP. MnP has five disulphide bridges and two Ca^{2+} ions, which maintain the structure of the active enzyme (Sutherland *et al.* 1997). The Mn (II)-binding site consists of two glutamate and one aspartate *γ* -carboxylic groups and is located close to the porphyrin macrocycle (Wong, 2009). The molecular mass of MnPs ranges from 38 to 62.5 kDa, which include 4%– 18% glycans, and their pI ranges from 2.9 to 7.1 (Sigoillot *et al.* 1997, Janusz *et al.,* 2017).

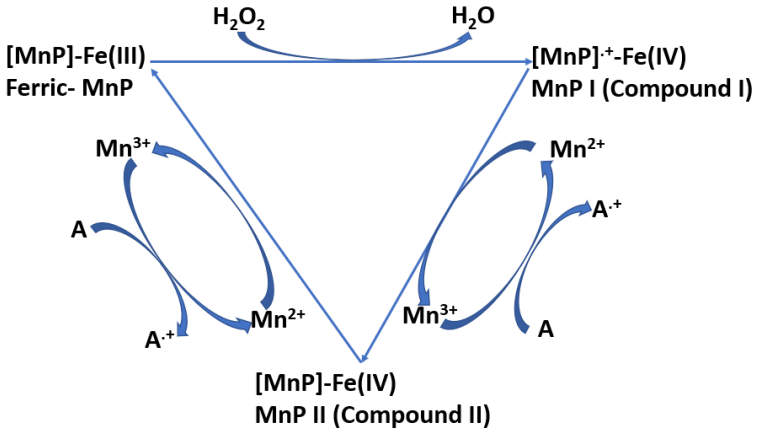


Figure 1.7 Catalytic mechanism of manganese peroxidase

1.7.1.1.2 Versatile Peroxidase

Versatile peroxidases (VP, EC 1.11.1.16) as have been fittingly named, are unique non-specific enzymes that combine the molecular architecture and mechanism of catalysis of LiP and MnP. They possess several binding sites that can cleave high redox potential non-phenolic compounds and lower redox potential aromatic compounds and amines such as veratryl alcohol, methoxybenzenes, azo dyes and Mn^{2+} in the absence of mediators (Garcia-Ruiz *et al.* 2012, Datta *et al.,* 2017).

Structurally, VP has 11–12 helices, 4 disulphide bridges, 2 structural Ca^{2+} sites, a heme pocket and a Mn^{2+}-binding site like that of MnP (Perez-Boada *et al.* 2005). VPs are secreted as isoenzymes, and they have molecular mass ranging between 40 and 45 kDa with a pI ranging between 3.4 and 3.9 (Perez-Boada *et al.* 2005). VP shares a basic catalytic mechanism that resembles those of the other peroxidases already discussed which include the formation of compounds I and II, but the ability of VPs to utilize a wider variety of potential substrates makes it more complex (Sigoillot *et al.* 1997, Janusz *et al.,* 2017; Chen and Wan, 2017).

1.7.1.1.3 Dye-decolorizing peroxidases

Dye-decolorizing peroxidases (DyP, EC 1.11.1.19) are a class of heme-peroxidases most recently discovered. They are not similar in structure or sequence to other known peroxidases, but they also use hydrogen peroxide as electron acceptor (Adamo *et al.,* 2022; Chen and Wan, 2017). Structurally, they possess two domains that contain *α*-helices and anti-parallel *β*-sheets with a heme cofactor located at the cavity between the two domains (Janusz *et al.,* 2017). Based on phylogenetic analysis of genomic sequences, DyPs can be classified into four types (A, B, C, and D). The A and C type DyPs are predominantly bacterial enzymes while type D is mostly clustered to fungal species. In addition to lignin and other typical peroxidase substrates, DyPs can also oxidize non-phenolic methoxylated aromatics, and high redox synthetic dyes such as anthraquinone and azo dyes and that's how they came to be named "dye-decolorizing" (Datta *et al.,* 2017, Chen and Wan, 2017, Schoenerr *et al.,* 2018). DyPs may be bi-functional enzymes as it has been suggested that they may have hydrolase or oxygenase activity aside their typical peroxidase activity and they are active at low pH (3-4) (Adamo *et al.,* 2022). Their physiological roles remain yet unclear though there is growing evidence that some bacterial variants of this are effective lignin degraders and/or involved in oxidative stress defence mechanisms (Janusz *et al.,* 2017).

1.7.1.2 Phenol Oxidases (PO)

Phenol oxidases (PO) are secreted mainly by microbes. They use molecular oxygen as final electron acceptor to catalyse the oxidation and depolymerisation of lignin and

other complex aromatic compounds into more readily available substrates (Kersten and Cullen, 2014). POs have been reported to be involved in biodegradation and detoxification of some aromatic pollutants and hence have been applied for bioremediation of polluted water and soil. Phenol oxidases can be classed as laccases or polyphenol oxidases depending on the substrates they specifically act on (Li *et al.,* 2009).

1.7.1.2.1. Laccases

Laccases (benzenediol: oxygen oxidoreductase EC 1.10.3.2) belong to the group of phenol oxidases. They are an important group of enzymes that contribute to the ligninolytic cocktail of enzymes produced by wood-degrading microbes. Laccases are abundant, and they can be found in plants, fungi, bacteria, and insects, but were first isolated by Yoshida in 1883 from the Japanese lacquer tree, *Rhus venicifera.* In 1896, Bertrand and Laborde showed its presence in fungi for the first time (de Gonzalo *et al.,* 2016).

They are multicopper-containing oxidoreductase enzymes. They possess four copper atoms located in three distinct binding sites within their active centre which are critical for their ability to oxidise a large variety of organic (Particularly phenols) and inorganic substances to their corresponding radical species, with the concomitant reduction of atmospheric oxygen which serves as an electron acceptor, to water (Li *et al.,* 2009; Fisher and Fong, 2014; Datta *et al.,* 2017). Structural and comparative studies have identified conserved regions in which histidine residues can bind four copper atoms located at two main sites (T1 and T2/T3) that are involved in catalytic activity. Electrons captured by the T1 site are transferred via T2/T3 to the product, leading to product oxidation and reduction of oxygen to water (Beloqui *et al.,* 2006). Laccases can also oxidize Mn^{2+} to Mn^{3+} and organometallic compounds. Generally, there is wide variation in the structure, molecular weight, and oligomeric state of laccases (de Gonzalo *et al.,* 2016; Janusz *et al.,* 2017)

The high molecular weight (MW 70000) and low-redox potential (0.5–0.8 V) of laccase makes it impossible to penetrate deep into wood and hinders it from oxidizing non-phenolic lignin units which have high-redox potentials (>1.5 V) on their own (Galli and

Gentili, 2004). Hence, to degrade and depolymerise lignin, and other aromatic phenols, amines, and even non-phenolic aromatic targets, they mostly require mediators to act as shuttles that convey electrons between their active site and the substrate though direct oxidation of substrates without the need for mediators is also possible (Figure 1.8). These small molecular weight mediators serve to help in overcoming the problem of substrate's inaccessibility to the enzyme's active site due to the large size of the enzyme. (Li *et al.,* 2009; Galli and Gentili, 2004). The mediators are low molecular weight compounds that can access the enzyme's active site easily where they are oxidized into more stable, high redox intermediates. These oxidized intermediates then move away from the active site to oxidize the complex substrates, and subsequently returns to its original state. The electrons taken by Laccases are finally transferred back to oxygen to form hydrogen peroxide (Datta *et al.,* 2017). Several organic and inorganic compounds, such as thiol and phenol aromatic derivatives, N-hydroxy compounds and ferrocyanide, have been reported as effective mediators for the above-mentioned purposes e.g ABTS (2,2 -azino-bis (3-ethylbenozthiazoline-6-sulfonic acid) and N-hydroxybenzotriazole (HBT) and acetosyringone. (Reid 1995; Li *et al.,* 2009; de Gonzalo *et al.,* 2016,). In the presence of redox mediators, laccases can even catalyze the breakdown of non-phenolic lignin structures, including the cleavage of β-*O*-4 linkages (Schoenherr *et al.,* 2018). They are useful enzymes for a variety of applications, including decolorization of different types of recalcitrant dyes, bioremediation of soils and water, kraft pulp biobleaching, and in other biotechnological applications (Beloqui *et al.,* 2006)

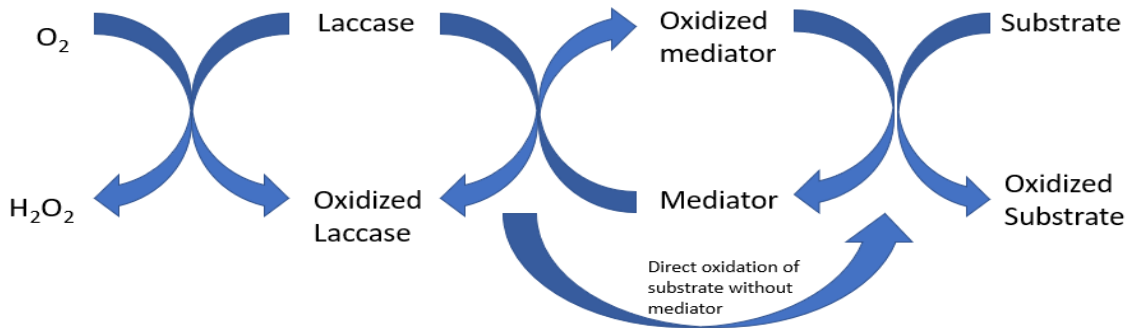


Figure 1.8 Illustration of the catalytic cycle of laccases

Whilst laccases have been shown to be involved in lignin breakdown reactions, they can also catalyse lignin polymerisation reactions. Plant laccases with 20-50% glycosylations are more involved with lignin polymerisation while fungal laccases with 5-25% glycosylation are more involved in depolymerisation (Ahmad, 2010).

1.7.1.2.2. Polyphenol oxidases

Polyphenol oxidases or tyrosinases (PPO), are members of the “phenol oxidase” class which possess a binuclear copper centre. They are able to oxidize diphenols to their corresponding quinones. This group of enzymes do not directly participate in lignin degradation but are classed as lignin degrading accessory enzymes (Li *et al.,* 2009).

1.7.2 Lignin Degrading Accessory (LDA) enzymes

Aside these peroxidases and laccases that react directly to bring about lignin modification, there are other auxiliary enzymes that have been discovered to play significant accessory roles in the complete biodegradation of lignin. They include antioxidant enzymes such as alcohol dehydrogenases, catalases, cytochrome P450, aryl-alcohol or veratryl alcohol oxidases, superoxide dismutase, and glyoxal oxidases that produce hydrogen peroxide required by the peroxidases, oxidoreductases such as dioxygenases, quinone oxidoreductases and cellobiose dehydrogenases that reduce the radical methoxy-groups of lignin-derived compounds (Sun and Zhou, 2011; Fisher and Fong, 2014; Janusz *et al.,* 2017). Also, the glutathione-dependent–etherases, which are multi-enzyme systems that can catalyse the reductive cleavage of ether bonds in lignin-related compounds. This system is composed of three separate proteins; LigD (a Cα-dehydrogenase), LigF (a β-etherase) and LigG (a glutathione lyase) (de Gonzalo *et al.,* 2016). Other enzymes which are viewed as secondary degradative enzymes such as protocatechuate-3,4 dioxygenase cannot degrade lignin on their own but have been reported to offer a great synergistic advantage when in combination with other enzymes (Li *et al.,* 2009; Kumar and Chandra, 2020).

1.8 The CAZy and FOLy databases and lignin degrading enzymes

The biosynthesis, assembly, modification, and catabolism of carbohydrate polymers and glycoconjugates (carbohydrate binding modules-CBMs) are carried out by a diverse group of enzymes called the Carbohydrate-Active enzymes, or 'CAZymes' (Levasseur *et al.,* 2013; Gilbert, 2010). The deconstruction of lignocellulose in plant cell walls which is composed mainly of polysaccharides benefit from members of this group. CAZymes are classified based on their amino acid sequence similarities into families maintained and continuously updated with genomic information from Genbank in an online platform called 'CAZy database' (www.cazy.org). Members of the same family display a common fold, with their catalytic apparatus and mechanism similarly conserved (Busk *et al.,* 2017; Levasseur *et al.,* 2013). The CAZy database currently incorporates more than 400,000 unique sequences classified in more than 300 families subdivided into the following classes with the side listed characteristics:

1. Glycoside Hydrolases (GH): Hydrolysis and/or rearrangement of glycosidic bonds
2. Glycosyl Transferases (GT): Formation of glycosidic bonds
3. Polysaccharide Lyases (PL): Non-hydrolytic cleavage of glycosidic bonds
4. Auxiliary Activities (AA): Redox enzymes that act in conjunction with CAZymes.
5. Carbohydrate Esterases (CE): Hydrolysis of carbohydrate esters
6. Carbohydrate-Binding Modules (CBM): Adhesion to carbohydrates

(Gilbert, 2010; Kameshwar and Qin, 2017a; Levasseur *et al.,* 2013; Kunath *et al.,* 2017).

This database does not only provide a means for rationalising enzymatic action on glycosidic bonds but has also been applied to describe and elucidate major aspects of the carbohydrate metabolism of fungi and other organisms with important consequences for both fundamental and applied knowledge. E.g, it can serve as a specialized and comprehensive database for annotating genes in a metagenomic project dedicated to the identification of novel CAZymes (Kunath *et al.,* 2017).

Although lignin is not a polysaccharide component of lignocellulose, its co-occurrence and interconnectivity with the polysaccharides within the plant cell wall superstructure suggests there might be no strict boundaries in the overall deconstruction of each component and there could exist some sort of cooperation amongst the enzymes that

act on each component. Lignin fragments act in concert with other families of CAZymes (CBM 33 and GH61 which are lytic polysaccharide monooxygenases-LPMOs) to efficiently breakdown plant cell wall. Hence, the family of lignin degrading enzymes was expanded to include the LPMOs, and reclassified into a new and more suitable class, the "Auxiliary Activities" (Levasseur *et al.*, 2013; Lombard *et al.,* 2014).

The AA class currently encompass redox enzymes and are divided into 16 families: 9 being ligninolytic enzymes and 7 LPMOs with 18740 classified modules and 28 non-classified modules as at present (www.cazy.org/Auxiliary-Activities.html; accessed August 2021). However, AA enzymes involved in lignin valorisation/depolymerisation function as highly reactive and non-specific free radicals which cleave carbon-carbon and ether inter-unit bonds and are found mainly in AA1- AA10 families presented in Table 1.5.

Other families (AA11- AA17) are monooxygenases but active in the catabolism of other compounds such as chitin, starch, and xylan not lignin.

Table 1.5 Auxilliary Activities enzyme families/ subfamilies involved in Lignin degradation
(Adapted from Levasseur *et al.,* 2013 and www.cazy.org/Auxiliary-Activities.html)

(Sub)Families	Known activities	EC number	Number of AA
AA1	Multicopper oxidase		4552
AA1_1	Laccase	EC 1.10.3.2	
AA1_2	Ferroxidase	EC 1.10.3.2	
AA1_3	Laccase-like multicopper oxidase	EC 1.10.3.2	
AA2	Class II peroxidase		721
	Manganese peroxidase	EC 1.11.1.13	
	Lignin peroxidase	EC 1.11.1.14	
	Versatile peroxidase	EC 1.11.1.16	
AA3	GMC oxidoreductase		2,154
AA3_1	Cellobiose dehydrogenase	EC 1.1.99.18	
AA3_2	Aryl alcohol oxidase/ glucose oxidase	EC 1.1.3.7/ 1.1.3.4	
AA3_3	Alcohol oxidase	EC 1.1.3.3	
AA3_4	Pyranose oxidase	EC 1.1.3.10	
AA4	Vanillyl alcohol oxidase	EC 1.1.3.38	62
AA5	Copper radical oxidase		751
AA5_1	Glyoxal oxidase	EC 1.1.3.-	
AA5_2	Galactose oxidase	EC 1.1.3.9	
AA6	1,4-Benzoquinone reductase	EC 1.6.5.6	748
AA7	Glucooligosaccharide oxidase	EC 1.1.3.-	93
AA8	Iron reductase domain		177
AA9	Lytic polysaccharide monooxygenase (GH61)	EC 1.-.-.-	871
AA10	Lytic polysaccharide monooxygenase (GH61)	EC 1.-.-.-	7,424

Unfortunately, A systematic and integrated database that specifically classifies lignin degrading enzymes has not been fully developed like in the case of the CAZymes and this limits the knowledge and information available on lignin degrading enzymes, and consequently constrains potential biotechnological applications (Levasseur *et al.,* 2008; Kameshwar and Qin, 2017a). However, Levasseur *et al.,* made a significant effort to develop a database for the enumeration and classification of enzymes involved

in the breakdown of lignin which they called the Fungal Oxidative Lignin enzymes (FOLy) database. The database is structured similarly to the CAZy database as it classifies the enzymes and related sequences according to their sequence similarity and structure corresponding to those in several public databases such as GenBank and the Protein Data Bank (PDB), and its module-by-module description (Levasseur *et al.,* 2008). They classified 379 full length and 601 partial sequences into 10 families consisting of 3 Lignin Oxidases (LOs) and 7 Lignin Degrading Auxiliary (LDA) enzymes based on their direct and indirect involvement in lignin degradation respectively. Lignin oxidases are designated numerically by order of creation as LO1 (Laccases), LO2 (Lignin peroxidases, Manganese peroxidases, Chloroperoxidases), and LO3 (Cellobiose dehydrogenase). The Lignin Degrading Auxiliary Enzymes (LDA) families are also numerically designated as follows: LDA1 (Aryl alcohol oxidase); LDA2 (Vanillyl alcohol oxidase); LDA3 (Glyoxal Oxidase); LDA4 (Pyranose Oxidase); LDA5 (Galactose Oxidase); LDA6 (Glucose Oxidase); and LDA7 (Benzoquinone reductase) (Kameshwar and Qin, 2017a; Levasseur *et al.,* 2008).

1.9 Overview of microbiome studies and metagenomics

Microorganisms are ubiquitous and exist in well-structured communities referred to as microbiomes. The study of microbiomes is very crucial for an overall insight into the different microorganisms that exist in nature, their interactions with each other, their immediate environment, and the entire biosphere as they are a critical component of these environments, providing essential ecosystem services. In studying host-associated microbial communities, an understanding of the factors that influence the composition, stability, and dynamics of the associated microbes and how these factors impact the phenotype, ecology and evolution of the host is paramount (McDonald *et al.,* 2010). The interactions between microorganisms and their hosts are complex and dynamic resulting in fluctuations in diversity and compositional abundance of resident microorganism in response to factors such as host genotype, developmental stage, diet, spatial and temporal dynamics etc (Antwis *et al.,* 2017). For the purpose of mining novel functionalities, an understanding of the structure and function of microbiomes offers great promise for biotechnological exploitation and applications in agriculture,

manufacturing, health, and the environment (Quince *et al.,* 2017). Traditionally, the study of microorganisms has been performed using culture dependent methods. With current estimates suggesting that of the approximately 4–6 x 10^{30} bacteria that may inhabit the earth, 99% are not cultivable, the implication is that cultivation-dependent approaches of studying microbiomes is limiting as access to an enormous amount of information within the genomes of uncultured microorganisms is not possible (Madhavan *et al.,* 2017; Sleator *et al.,* 2008; Simon and Daniel, 2011; Batista-Garcia *et al.,* 2016). "Unculturable" as used in this context does not mean that the microorganisms can never be cultured, it only indicates that the current techniques for culturing microorganisms in the laboratory are not right to support the growth of such microbes at the moment due to a lack of adequate information about their biology such as lack of necessary symbionts, nutrients, or surfaces, excess inhibitory compounds, incorrect combinations of temperature, pressure, or atmospheric gas composition, accumulation of toxic waste products from their own metabolism, and intrinsically slow growth rate or rapid dispersion from colonies (Rabelo-Fernandez *et al.,* 2018; Handelsman, 2004; Stewart, 2012). Inability to culture in the lab has limited our ability to understand microbial ecosystems and has slowed down our effort to discover and utilize new and beneficial functionalities from microorganisms e.g., enzymes for biotechnological applications, bioremediation improvement processes, discovery of biomarkers for disease diagnosis and therapeutic targets etc. Hence, the need for non-culture dependent methods of studying microbiomes (Schmeisser *et al.,* 2007; Knight *et al.,* 2012; Kunath *et al.,* 2017).

Depending on the research question, hypothesis, sample type, budget and a host of other considerations, microbiome studies can aim to focus on the structure (composition) or the function of the microbial community (structure-based surveys or function-based surveys), or both (Knight *et al.,* 2018; Madhavan *et al.,* 2017).

Structure based surveys are concerned with finding out what the composition of the microbial community looks like, that is, what types of microorganisms are present in the community and in what abundance. These kinds of experiments take advantage of the availability of universally conserved genes with enough variability that can be used to distinguish different taxa within a population following sequencing and analysis. The

commonly exploited genes are the 16S rRNA gene, 18S/ 28S rRNA gene and ITS (Internal Transcribed Spacer) as markers for identification of bacteria, eukaryotes, and fungal taxa respectively. Well known housekeeping genes e.g RecA, rpoD or genes performing specific functions such as nitrate reductase genes, sulphate reductase genes needed for specific functions necessary for the survival of the microbes can also be used as markers for taxonomic profiling. Knowing the microbes within a community could also serve to give insight into the possible biological function of such a community e.g., the presence of cyanobacteria could suggest that the community is photosynthetic (Sharpton, 2014). On the other hand, function-based surveys are particular about figuring out what the community does rather than who's in the community. The most common methods are metagenomics, metatranscriptomics and metaproteomics. Most experiments combine both approaches to obtain information on both the organisms present in a community and the roles they perform (Sleator *et al.,* 2008; Schmeisser *et al.,* 2007, Kunath *et al.,* 2017; Streit and Daniel, 2010; Simon and Daniel, 2011).

Metagenomics, also called community genomics, environmental genomics, or population genomics has emerged as a powerful centrepiece among the methods designed to gain access to the physiology and genetics of uncultured microorganisms in their environments. In theory, a metagenome is the total isolated DNA of all microorganisms that inhabit a particular environment. Metagenomes can be quite large and diverse, having several hundreds to thousands of distinct species and genomes depending on the sampled environment (Streit and Daniel, 2010).

By way of definition, metagenomics is the culture-independent analysis of the entire microbial genomes (both culturable and non-culturable microorganisms) present in a particular environment whereby nucleic acids are directly isolated from samples and analysed (Handelsman, 2004; Thomas *et al.,* 2012; Sleator *et al.,* 2008). This discipline has evolved as a new field of research, and it has seen a significant level of development with great advancement in technologies. Metagenomics has transformed our capacity to investigate complex microbial communities as it helps us to correlate the phylogenetic and functional attributes of a community to the physical, chemical, and biological manifestations that uniquely identifies that community (Thomas *et al.,* 2012; Knight *et al.,* 2018).

Initially, metagenomics began with the cloning of environmental DNA, followed by screening for expressed functions. This has now been complemented or in most cases, replaced by direct random shotgun sequencing of environmental DNA (Thomas *et al.,* 2012). Metagenomics can be used to explore the composition of microbial communities and their ecological patterns of diversity to have a robust understanding of the inter relationships that exist between the entire inhabitants that make up the biosphere. It is also a useful tool for identifying and bioprospecting novel enzymes from natural environments such as soil, marine water and the gastrointestinal tracts of vertebrates and invertebrates driven by the increasing biotechnological demands for these enzymes and biomolecules (Sleator *et al.,* 2008; Bugg *et al.,* 2011a, Yu *et al.,* 2018, Joynson *et al.,* 2014; Simon and Daniel 2011; Harnpicharnchai *et al.,* 2007). These enzymes have applications in several industries where they facilitate reactions that are difficult or expensive when using chemical catalysts, to increase the diversity of products, improve efficiencies, reduce costs and energy needs, and to reduce the burden of industrial processes on the environment (Ewuim *et al.,* 2011; Ferrer *et al.,* 2009).

Metagenomics based experiments can be approached in two ways: function-driven and sequence-driven metagenomics. While sequence-driven techniques are mostly used for biotechnological studies (bioprospecting enzymes and proteins with desired abilities), function-driven techniques are more powerful with respect to the identification of gene functions and their role in complex microbial communities (Schmeisser *et al.,* 2007; Rosnow *et al.,*2017). Most experiments use one or a combination of these approaches.

1.9.1 Function-driven approach to microbiome studies

In a function-driven approach to studying microbiomes, metagenomics libraries for a particular phenotypic characteristic expressed from the genes that code for the protein responsible for the function or activity of interest are screened and analysed, e.g., enzymatic activity, production of antibiotics, salt tolerance, etc. The phylogenetic source of the gene from the cloned DNA can then be identified by monitoring evidence or rate of enzyme activity or product formation (Dinsdale *et al.,* 2008; Streit and Daniel

2010; Rosnow *et al.,* 2017; Kunath *et al*., 2017). This strategy can potentially identify novel functional enzymes and /or biomolecules as screening is not dependent on sequence information (Reid 1995; Bugg *et al.,* 2011a; Joynson *et al.,* 2017; Yu *et al.,* 2018). Three different strategies can be employed in carrying out function-driven microbiome studies: phenotypical detection of the desired activity, heterologous complementation of host strains or mutants, and induced gene expression. In most cases, phenotypical detection which are the more commonly used strategies, use dyes and insoluble or chromophore-bearing derivatives of substrates incorporated into growth medium or reaction mixtures and then the clones or reaction is monitored for metabolic capability (Simon and Daniel, 2011).

Many successful function-based microbiome studies which require efficient systems for screening and expression have resulted in the identification of entirely new families of proteins including amino acylated antibiotics, turbomycin A and B, lipases, marine chitinases and some membrane bound proteins (Streit and Daniel 2010; Batista-Garcia *et al.,* 2016).

1.9.2 Sequence-driven approach to microbiome studies

Techniques with a sequence-driven approach are used mainly in functional metagenomic surveys and for bioprospecting novel proteins. Targeting of specific classes of enzymes can be directly incorporated into PCR-based analyses of metagenomic DNA, with the help of highly conserved domains present in those classes of enzymes. In these strategies, primer design represents the critical step and can introduce intrinsic bias through a marked influence on the types and relative novelty of genes that may be amplified (Batista-Garcia *et al.,* 2016). DNA probes or primers derived from conserved regions of genes or protein families with already known sequences and roles in metabolic pathways are designed in such a manner that the primers bind to the conserved regions of the genes within the metagenomic DNA following PCR amplification. The respective amplified PCR fragment is either directly sequenced and analysed or cloned into a vector followed by sequence analysis of selected clones compared against known databases (Streit and Daniel 2010, Batista-Garcia *et al.,* 2016). Therefore, only new variants of proteins with known functional

classifications can be identified but no entirely new genes can be detected (Schmeisser *et al.,* 2007). Successful application of this strategy has resulted in the identification of genes that code for novel enzymes such as chitinases, nitrite reductases, dioxygenases, glycerol dehydratases, dimethyl sulfoniopropionate-degrading enzymes, lignocellulose degrading enzymes etc (Madhavan *et al.,* 2017; Streit and Daniel, 2010; Simon and Daniel, 2011).

Microbial communities can be analysed by two different strategies in a sequence-driven approach. These are the marker gene/ targeted sequencing and whole (meta)genome shotgun sequencing, depending on the research goal.

1.9.2.1 Marker gene/ targeted sequencing.

In marker gene sequencing, primers are specifically designed to target a specific region of a gene of interest. These genes will typically be abundantly present and code for a protein performing an essential function across species but also have enough variability to distinguish between species. The genes contain variable and conserved regions. Primers are attached to the conserved regions for PCR amplification while the highly variable regions can be used for detailed phylogenetic identification and classification of microbes to specific genera or species (Santo Domingo 1998; Langille *et al.*, 2013; Jovel *et al.,* 2016; Segata *et al.,* 2012, Bragg and Tyson, 2014). Marker gene amplification and sequencing using unique markers such as the sequence of the 16S rRNA gene for bacteria and archaea, 18S rRNA gene for eukaryotes and internal transcribed spacer (ITS) for fungi are the principal measures of phylogenetic relatedness and thus of biodiversity. These methods are fast, cost effective and well developed for obtaining a low-resolution information about a microbial community (Knight *et al.,* 2017; Ferrer *et al.,* 2009; Batista-Garcia *et al.,* 2016; Alcon-Giner *et al.,* 2017).

Systematic characterization of diversity will not only provide new understanding of how microorganisms in different habitats interact with respect to carrying out metabolic activities and other functions individually or collectively but can also predict the functional composition of metagenomes from marker gene data (using the bioinformatic software PICRUSt; Phylogenetic Investigation of Communities by

Reconstruction of Unobserved States) (Langille *et al.,* 2013; Escobar-Zepeda *et al.,* 2015; Ceballos *et al.,* 2017; Mohammed *et al.,* 2018; Chew *et al.,* 2018; Knight *et al.,* 2018). Therefore, marker gene sequencing can provide new ideas and insights that may help the cultivation of microorganisms that are currently unculturable (Ferrer *et al.,* 2009).

16S rRNA gene sequencing is the most commonly performed marker gene analysis and has been employed to assess microbial diversity in environments such as soil, water, guts of animals etc. The prokaryotic 16S rRNA gene is about 1,500 bp long made up of nine variable regions interspersed between conserved regions. Other conserved genes, such as recA, *rpoB*, *gyrB*, *fusA* and radA, genes encoding heat shock protein 70, elongation factor Tu, or elongation factor G, have also been employed as markers for phylogenetic analyses (Simon and Daniel 2010; Scheller and Ulvskov, 2010; Bragg and Tyson, 2014).

Extensive sequencing of the ribosomal RNA gene has led to the generation of several large databases containing reference sequences which serve as very important resources for comparing sequences for rRNA gene-based classification of microorganisms. Some examples of these databases include the ribosomal database project (RDP) II (Cole *et al.,* 2003), SILVA (Yilmaz *et al.*, 2014), Greengenes (DeSantis *et al.,* 2006) and a manually curated database called EzTaxon-e (Kim *et al.,* 2012) among others. Typically, the 16S rRNA genes are PCR amplified and cloned prior to sequencing and therefore some inherent disadvantages such as PCR bias, low species level resolution, or the varying number of gene copies between taxa are limitations of this approach (Simon and Daniel 2009; Venter *et al.* 2004).

Although the marker gene sequencing analysis methods provide great information on taxonomic and phylogenetic composition of microbial communities, they have limited resolution, therefore, the shotgun metagenomic sequencing of entire communities has become a viable alternative thanks to the decreasing costs of sequencing protocols. This approach is applicable to samples of uncultured microbiota and avoids some of the limitations of marker gene sequencing.

1.9.2.2 Whole metagenome shotgun sequencing.

Whole metagenome sequencing involves sequencing fragments of all microbial genomes within a sample without any prior cloning or amplification of a specific genomic locus. Shotgun metagenomics can theoretically capture all DNA from all the genomes present in a sample, collectively referred to as the metagenome, fragment the DNA into small pieces which can be sequenced independently to generate reads that cut across both taxonomically informative genomic loci such as the 16S rRNA gene and genes that code for other biological functions (Sharpton, 2014). This technique can therefore be used to profile taxonomic composition and functional potential of microbial communities and to recover whole genome sequences by reconstructing large fragments (contigs) or even complete genomes from organisms in a community for further analysis (Sleator *et al.,* 2008; Kunath *et al.,* 2017). Therefore, this sequencing approach provides access to the entire gene composition of microbial communities and thus gives a much broader description and taxonomic resolution as well as detailed genomic information about a community than marker gene surveys, which are often based only on the diversity of one gene (Bragg and Tyson, 2014). Also, because this approach is cloning and amplification-independent, the biases that could result from cloning and amplification are avoided (Quince *et al.,* 2017, Escobar-Zepeda *et al.,* 2015; Sharpton 2014). The target of a function driven approach to metagenomics is the identification of metabolic activity as it can reveal novel enzymes, but it cannot directly reveal the microbial origin or genetic context of the enzymes without further analysis being performed. Hence, if isolation of the identified enzyme is desired, a sequence-driven approach must be employed to enable PCR amplification and recombinant production of the target enzyme of interest from sequence data (Handelsman, 2004; Schmeisser *et al.,* 2007; Sleator *et al.,* 2008; Ferrer *et al.,* 2009).

1.9.3 Shotgun Metagenomics as a tool for bioprospecting enzymes from microbial environments

Microorganisms play a vital role in biogeochemical cycles and are an attractive source of novel biomolecules with biotechnological applications (Kunath *et al.,* 2017). They can grow in a vast range of environments, from alkaline lakes to hydrothermal vents,

indicating that they produce active and stable enzymes that help them thrive under these extreme conditions. Hence, proteins from microbial sources are more stable because they can retain their activities even under less-than-ideal conditions compared to their counterparts obtained from plant or animal sources (Simon and Daniel, 2011). This explains why most enzymes for commercial application are being sourced from microorganisms to replace chemical catalysts in industrial processes (Acharya and Chaudhary, 2012; Madhavan *et al.,* 2017; Steele *et al.,* 2009).

The advent of next generation sequencing (NGS) techniques has increased the ease and reduced the cost of studying microbial communities using metagenomics (Yu *et al.,* 2018), and in over 15 years since it evolved, metagenomics has enabled large-scale investigations of complex microbiomes including those of soil, sea, human, ruminant, gastropods, crustacean and other insect guts (Quince *et al.,* 2017; Segata *et al.,* 2012, Madhavan *et al.,* 2017).

Typically, experiments that aim to bio prospect for enzymes and other biomolecules from microbial environments employ the shotgun sequencing metagenomics approach. This process typically involves the steps outlined below but lots of alterations and adjustments can be done in line with what the specific experiment is designed to achieve.

The very first and most critical decision in carrying out a shotgun metagenomic sequencing experiment involves the choice of sample collection, preservation and preparation as all other downstream processes will depend on this. It is crucial to ensure the most appropriate environments and methods are chosen that will give the best representative sample and outcome in line with the aim of the experiment. All information and details of the collection process must be recorded in the experimental metadata in line with checklists provided by the Minimum Information about any (x) Sequence (MIxS) for a metagenome sequence (MIMs), and marker sequence (MIMARKS) recommended by the genomic sequence consortium (GSC) (Yilmaz *et al.,* 2011; Field *et al.,* 2011). These standards should be followed to account for variability and to ensure that experiments are reproducible and hence verifiable (Knight *et al.,* 2018; Thomas *et al.,* 2012, Quince *et al.,* 2017). Isolation of pure, high molecular weight and representative nucleic acid target (DNA in the case of metagenomics and RNA in

the case of metatranscriptomics) and preparation of sequencing libraries from the collected samples are the next vital steps which are equally important in achieving quality reads from the various next generation sequencing platforms (Head *et al.,* 2014; Madhavan *et al.,* 2017).

NGS technologies such as the 454/Roche and the Illumina/Solexa systems, and more recently, the PacBio and Nanopore technologies have been the most extensively applied methods for metagenomic sequencing, shifting away from sanger sequencing technology (Thomas *et al.,* 2012; Head *et al.,* 2014; Quinn *et al.,* 2018). The Illumina platform is widely available, costs drastically reduced, has high accuracy (0.1-1% typical error rates), very high output (can generate up to 1.5Tb data per run), requires only very small amount of DNA input for library preparation (0.001-1ng),

and can sequence up to 96 or 384 samples in one run by multiplexing using dual indexing barcodes. The above reasons have made the Illumina technology an increasingly popular choice for metagenomics research, although SOLiD, ion torrent and other fast emerging sequencing technologies being developed might be useful or even out-perform the currently used technologies in the future (Thomas *et al.,* 2012; Quince *et al.,* 2017; Bragg and Tyson 2014).

The short-read fragments generated from shotgun sequencing may need to be assembled into longer contigs (contiguous sequence) by identifying overlapping reads from the same genome. Metagenomic data can be assembled with reference to an existing database (reference-based) or without a reference database (*de novo*). Reference based assembly relies on the availability of related reference genomes while *de novo* assembly requires large computational resources and is very difficult to perform but overall, the aim of the research will determine what assembly method will be carried out. There are several challenges associated with metagenomic assembly, first, it is difficult to assemble genomes of less abundant taxa within a community, then production of chimera's and thirdly, assembly can be computationally intensive. A wide range of assembly tools such as Metavelvet, SOAP, Meta-IDBA, IDBA-UD, MetaSPAdes, MEGAHIT etc based on the de Bruijn graphs have been developed and are in use for assembly of metagenomic data (Knight *et al.,* 2018; Thomas *et al.,* 2012; Quince *et al.,* 2017; Bragg and Tyson 2014; Sharpton, 2014).

Raw or assembled sequences are sorted into groups that represent an individual genome or genomes with closely related taxonomic origins with the aid of several algorithms in a process referred to as Binning. Binning algorithms take advantage of shared features or characteristics across genomes such as similarity in conserved nucleotide compositions (composition-based binning or unsupervised method) e.g., distribution of abundant k-mers or a certain GC composition to group sequences into taxonomic groups. Binning can also be done by comparing the similarity of the gene encoded by the DNA fragment with known genes from reference databases (similarity based binning or supervised method) (Quince *et al.,* 2017; Kunath *et al.,* 2017). Phylopythia, S-GSOM, PCAHIER and TACAO are examples of composition-based algorithms while IMG/M, MG-RAST, MEGAN, CARMA, SOrt-ITEMS, and MetaPhyler are examples of similarity-based algorithms. PhymmBL and MetaCluster consider both composition and similarity (Thomas *et al.,* 2012; Sharpton, 2014). Composition based methods are more in use as against the similarity-based methods because most microbial species have not been sequenced, so a large fraction of reconstructed genomic fragments cannot be mapped to reference genomes. Although binning may be conducted on assembled or unassembled data, most binning algorithms have shown that the longer the sequences, the better the binning accuracy (Bragg and Tyson 2014; Sharpton, 2014).

In bioprospecting novel functionalities from metagenomic data, sequences (may be assembled or not) are subjected to sequence-based homology searches against databases of proteins or conserved domains. The functional annotation of metagenomic sequence data is comprised of two steps; gene prediction where sequences that represent a gene are identified and labelled as genomic elements; and functional annotation, where putative functions are assigned to identified genes by comparing the coding sequence to a database of genes, proteins, protein families, or metabolic pathways for which some functional information is known. Some tools designed for prediction of gene coding sequences from metagenomic data include MetaGeneMark, MetaGene/ MetaGeneAnnotator, FragGeneScan, and Orphelia (Bragg and Tyson, 2014, Sharpton 2014). Functional annotation is computationally challenging and currently, it is estimated that only 20 to 50% of metagenomic

sequences can be annotated which means the function or importance of many identified genes remain unknown (Sharpton, 2014). This is because, annotation is not done *de novo*, but by mapping to gene or protein databases that are already in existence hence previously uncharacterised proteins cannot be identified. Many reference databases are available to give functional context to metagenomic datasets and they come in two varieties: sequence and hidden Markov model-HMM based databases; examples include SEED, Kyoto Encyclopedia of Genes and Genomes (KEGG), EggNOG, Clusters of Orthologous Groups (COG), MetaCyc, Pfam (protein families database), and TIGRFAM (Thomas *et al.,* 2012; Kunath *et al.,* 2017; Simon and Daniel, 2011). However, recent versions of MG-RAST and IMG/M can merge and visualise interpretations of all database searches as no single reference database covers all biological functions (Thomas *et al.,* 2012).

The entire metabolic functional profiling of a microbiome can be performed from whole metagenome shotgun sequencing data, but also, a complementary approach is to characterize specific functions of interest from the metagenome. For example, identification of genes coding for specific enzymes like the β-lactamases, sulfur oxygenase reductases, chitinases, alcohol oxidoreductases, monooxygenases, diol dehydratases, carboxypeptidases, and antibiotic resistance genes (Ferrer *et al.,* 2009). *Ad hoc* methods and manually curated databases of genes encoding these specific functions (e.g CAZy, CAT, and dbCAN databases for automated CAZy annotation, Antibiotic resistance databases- ARDB and Resfams, FOLy- database of fungal oxidative lignin enzymes, MetaBioMe, etc) are increasingly being developed and have been crucial to the success of this approach (Escobar-Zepeda *et al.,* 2015, Quince *et al.,* 2017; Sharpton, 2014, Levasseur *et al.,* 2008).

Strategies and tools employed for the statistical analysis of higher organisms from years of research in quantitative ecology can be adapted and applied in analysing metagenomic sequencing projects in microbial ecology. Statistical tools such as the Primer-E package allow for a range of multivariate analysis including generation of MDS (multidimensional scaling) plots, ANOSIM (analysis of similarities), analysis of variance (ANOVA), SIMPER which identifies species or functions that contribute to differences between two samples. Some statistical tools have been incorporated into

web-based pipelines e.g Metastats, MG-RAST, and FunctionalizeR (Escobar-Zepeda *et al.,* 2015, Quince *et al.,* 2017; Sharpton 2014).

Notwithstanding the great potential of the shotgun metagenomic approach of studying miocrobiomes, some drawbacks such as high cost associated with sample preparation, sequencing, and the requirement for extensive data analysis (alignments and annotations) still exist. It is also, not possible to discover truly novel enzymes and biomolecules as de novo annotation of functional genes cannot be done (Baldrian and Lopez-Mondejar, 2014; Escobar-Zepeda *et al.,* 2015; Knight *et al.,* 2018).

Metagenomics can be complemented with metatranscriptomic or metaproteomic approaches that use RNA and proteins respectively to describe expressed activities. However, metagenomics on its own, can provide information about potentially new biocatalysts, genomic interrelationships between function and phylogeny, and the structural and functional evolutionary profile of a community of unculturable microorganisms (Thomas *et al.,* 2012).

In an ideal situation, it is expected that any comprehensive metagenomics study should use all the above sequencing methods in analysing each sample. Realistically though, there could be constraints in terms of sufficiency of sample material or funding, and in other cases, the sample may not be amenable to one of the sequencing methods. It is therefore imperative that researchers consider and adopt the method of sequencing that will most effectively answer their specific research question. If no funding challenges exist, then it is recommended as common practise to perform marker gene sequencing first to obtain a basic clue of the microbial composition of the community and then move on to whole metagenome sequencing and /or metatranscriptomics/ metaproteomics depending on the focus of the research. If it is not possible to adopt all methods, whole genome metagenomics alone is recommended against just marker gene sequencing (Quince *et al.,* 2017, Knight *et al.,* 2018).

In our study, we adopted a two-tiered approach in which we first performed a 16S rRNA sequencing of the APW gut to determine the bacterial community structure and search for the presence/ abundance of lignin degrading bacteria in different gut segments of the larvae being the potential sources of our enzymes of interest (structural survey); then we used whole metagenome sequencing and functional annotation to identify

putative genes that code for lignin degrading enzymes expectedly retrieved from the identified bacteria within the gut metagenome (functional survey), using sequence based approaches (targeted/ marker gene and whole metagenome shotgun sequencing) in both survey methods.

1.9.4 Review of selected metagenomics-based microbiome studies

The Genomes OnLine Database (GOLD) is an online resource where information on genome and metagenome related sequencing projects spanning a wide spectrum of environments, with their associated metadata from all around the world are kept and continuously monitored centrally (Pagani *et al.,* 2011; Bragg and Tyson, 2014; Mukherjee *et al.,* 2019). The database has experienced steady increase since its launch in 1997 with the microbial genome projects being responsible for most of that increase (Pagani *et al.,* 2011). In 2005, the database contained 1575 sequencing projects and in 2015, barely 10 years later, a massive 70,000 projects were recorded (Reddy *et al.,* 2015; Vestergaard *et al.,* 2017). As of 2021, 416,202 sequencing projects, and 322,792 analysis projects have been recorded from 49,674 studies that have involved about 412,323 organisms (https://gold.jgi.doe.gov).

In the nearly two decades since it was first used, metagenomics has enabled large-scale investigations of complex microbiomes and enriched our understanding of uncultured communities (Sharpton, 2014).

Initial research on mining genes from microbial communities focused on external and extreme environments (Simon and Daniel, 2009) such as the sea, acid mine drainage, hot springs, and landfill sites. Currently, genes and biomolecules especially those involved with lignocellulose breakdown and metabolism have been mined from diverse internal environments of vertebrate and invertebrate guts as well as from the genomes of individual microorganisms.

Presented in table 1.6 are selected examples of metagenomic projects and the significant discoveries from each of them.

Table 1.6 Some discoveries from selected metagenomics-based projects

Study Environment	Discovery	Reference
External Environments		
Landfill	• 8,371 CAZymes belonging to 244 families. Fibrobacter cellulase system and polysaccharide utilization locus (PUL) in landfill sites	Ransom-Jones et al., 2017
	• Functional genes and enzymes related to production of 7 valuable products from the microbial community of the activated sludge from a landfill leachate treatment plant	Yasuda et al., 2021
	• Various genes related to degradation of xenobiotic compounds in Landfill Lysimeter Soil of Ghazipur Landfill Site	Gupta et al., 2017
Sugarcane bagasse	• 1,774 biomass degrading genes including cellulases, glycoside hydrolases and auxiliary activity proteins	Mhuantong et al., 2015
	• 34 putative carbohydrate-active enzymes belonging to 17 glycosyl hydrolase (GH) families.	Colombo et al., 2016
	• A variety of putative genes encoding GH families and production of recombinant GH9 endoglucanase (Cel9) and GH11 endo-xylanase (Xyn11)	Kanokratana et al., 2015
Anaerobic poplar	• 28,793 carbohydrate active enzymes identified from 230 candidate genes were identified by homology with CAZy or FOLy genes from 230 gene families, with the 22 most dominant gene families containing 19,510 candidate genes (67%).	Van der Lelie et al., 2012

Hot springs	• Taq DNA polymerase from Thermus aquaticus, purified and isolated from hot springs. • Genes potentially involved in nitrogen and sulfur cycling. • genes associated with sulfur, nitrogen, and methane metabolism, and genes of many novel carbohydrate-transforming enzymes	Chien et al., 1976 Jiménez et al., 2012 Sharma et al., 2020
Sea	• Identified 1800 genomic species, 1.2 million previously unknown genes, including 148 previously unknown bacterial phylotypes and more than 782 new rhodopsin-like photoreceptors.	Venter et al., 2004
Acid mine drainage	• Identified an abundance of genes that function in detoxifying cells of toxic elements	Tyson et al. 2004
Sludge	• Identified 253 thermo-stable genes as putatively carbohydrate-active, dominated by GH9 and CBM3	Xia et al., 2013
Vertebrates		
Human guts	• Single study that yielded the largest database of 16Sr RNA sequences having 11,831 bacterial and 1524 archeal sequences • Seventy-three CAZymes from 35 different families were discovered for catabolising dietary fibre	Sleator et al., 2008 Tasse et al., 2010
Giant panda	• Genes encoding cellulase, β-glucosidase, xylan 1,4-β-xylosidase, and endo-1,4-β-xylanase. • Antibiotic resistance genes and biosynthesis of antibiotics	Zhu et al., 2011 Mustafa et al., 2021
Buffalos	• Identified potential 2614 contigs encoding biomass degrading enzymes, 1943 GH contigs, 23	Singh et al., 2014

	CBM contigs, 373 glycosyl transferase contigs, 259 carbohydrate esterases contigs and 16 polysaccharide lyases 16 contigs	
Goat	• 821 ORFs encoding carbohydrate esterases (CEs) and polysaccharide lyases (PLs) serving for lignocellulose pre-treatment, 816 ORFs encoding 11 glycoside hydrolase families (GHs) of cellulases, and 2252 ORFs encoding 22 GHs of hemicellulases, were mined	Do et al., 2018
Wallabies	• multigene polysaccharide utilization loci-like systems coupled with genes encoding β-1,4-endoglucanases, and β-1,4-endoxylanases were identified	Pope et al., 2010
Cow rumen	• 12 esterases, 9 endo-b-1,4-glucanases and 1 cyclodextrinase were identified in the library and characterized • Identified 27,755 putative carbohydrate-active genes and expressed 90 proteins with 57% of the expressed proteins demonstrating catalytic activity against cellulosic substrates	Ferrer et al., 2005 Hess et al., 2011
Rats	• Found 587 carbohydrate-active enzyme genes belonging to different families, including 7 carbohydrate esterase families and 21 glycoside hydrolase families	Bai et al., 2021
Invertebrates		
Snails	• Identified over 2700 genes of CAZymes and also genes that could facilitate the detoxification of xenobiotics and synthesis of essential amino acids and vitamins. • System functional analysis showed that 35.00% of genes belong to transposable elements, 10.00% of genes	Cardoso et al., 2012 Rabelo-Fernandez et al., 2018

	belong to clustering-based subsystems, 4.00% of genes belong to the production of cofactors and secondary metabolites, and 2.00% resistance to antibiotics and toxic compounds	
Slugs	• Identified over 3,383 carbohydrate active enzymes (CAZymes) including multiple enzymes associated with lignin degradation	Joynson et al., 2017;
Termites	• Several bacterial genes encoding cellulose and xylan degrading enzymes • Two cellulases and 12 xylanases • Cellulases and associated genes	Warnecke et al., 2007 Nimchua et al., 2012 Watanabe and Tokuda, 2010
Beetles	• Identified pectinase and Cellulase, in the Digestive System of the Red Palm Weevil (Rhynchophorus ferrugineus) • Genes encoding enzymes were identified in the A. glabripennis gut metagenome that could have key roles in woody tissue digestion including candidate lignin degrading genes (laccases, dye-decolorizing peroxidases, novel peroxidases and β-etherases), 36 families of glycoside hydrolases (such as cellulases and xylanases), and genes that could facilitate nutrient recovery, essential nutrient synthesis, and detoxification. • Identified a novel cellulose Bh-EGaseI belonging to the glycoside hydrolase family45(gh45-1) obtained from the beetle Batocera horsfieldi	Vatanparast et al., 2014 Scully et al., 2013 Mei et al., 2016 Bozorov et al., 2019

	• Identified bacterial isolates exhibiting cellulolytic, xylanase, glucanase, cellobiose and lignin peroxidase activity • Identified genes that encode enzymes involved in lignocellulose degradation (such as peroxidases, alpha-L-fucosidases, beta-xylosidases, beta-mannosidases, endoglucanases, beta-glucosidases and others, and nitrogen fixation (nitrogenases).	Mohammed et al., 2018,

1.10 Insect gut microbiota as sources of lignocellulose degrading enzymes

Insects are a very successful group of organisms both in terms of diversity and their remarkable adaptability for survival in different ecological niches with an estimated 6-10 million species existing on earth (Prasad *et al.,* 2018). The insect gut microbial diversity represents a large source of unexplored microbes that participate in various activities from utilization of different organic polymers, nitrogen fixation, methanogenesis, pesticide degradation, pheromone production to pathogen prevention (Engel and Moran, 2013; Prasad *et al.,* 2018). It is estimated that the gut of insects houses ten times more microbial cells than total insect cells and a hundred times more microbial genes than animal genes (Lluch *et al.,* 2015).

There is increased evidence that suggests the genomes of insects lack many of the catabolic enzymes necessary to fully digest the vegetal meals they consume to extract energy from it and there is sufficient evidence that points to the microbiota in their gut as being the facilitators responsible for this ability (Engel and Moran, 2013; Ben-Yosef *et al.,* 2014; Sugio *et al.,* 2015; Xu *et al.,* 2016; Muhammad *et al.,* 2017).

Microorganisms colonize the gut of insects through food intake, and they perform significant roles in digestion and metabolism as well as other beneficial roles for their hosts (Rajagopal, 2009; Engel and Moran, 2013) contributing to the diversity and

evolutionary success of insects. Wood feeding insects and other herbivores use the enzymes produced by their gut associated microbes to facilitate the digestion of woody tissue by liberating carbohydrates from plant tissues. The enzymes produced contribute greatly to digestion of lignocellulose in many insect classes where microbial fermentation products have been detected in the gut (Scully *et al.,* 2013; Jia *et al.,* 2013; Hansen and Moran, 2014).

The sequencing of the complete genome of the fruit-fly Drosophila melanogaster pioneered an era of insect genomics (Sleator *et al.,* 2008). Since then, there has been increased interest in bioprospecting lignocellulose degrading enzymes from the guts of wood feeding insects due to the natural ability of these insects to utilize wood as a nutrient source. Initially, research in this field has focussed on termites (Isoptera). A review by Ni *et al.,* (2013) of metagenomic and metatranscriptomic based research has shed more light on our understanding of the diversity of lignocellulolytic enzymes within the termite gut. Their findings indicate that there has been consistent increased research interest in the biology of termite enzymes in the last 10 years which resulted in the termites being dubbed as the most successful plant decomposers (da Costa *et al.,* 2018). Recently, the gut microbiomes of insects from other insect orders such as Coleoptera: sheet winged insects, e.g., Beetles (Egert *et al.,* 2003;Tagliavia *et al.,* 2014; Franzini *et al.,* 2016; Luo *et al.,* 2019; Mohammed *et al.,* 2018), Lepidoptera: paired winged insects, e.g., Butterflies and moths (Paniagua Voirol *et al.,* 2018; Gomes *et al.,* 2020; Gong et al., 2020), Diptera: e.g., Flies and mosquitoes (Sontowski *et al.,* 2020; Coastworth *et al.,* 2018; Park *et al.,* 2019; Dada *et al.,* 2021), Hymenoptera: e.g., Bees, Ants, and wasps (Quinn, 2017; Ramalho *et al.,* 2017; Romero *et al.,* 2019), Isoptera: e.g., termites (Ni and Tokuda, 2013; Scully *et al.,* 2013; da Costa *et al.,* 2017; Ali *et al.,* 2019), Orthoptera: e.g., crickets and grasshoppers (Zheng *et al.,* 2021; McClenaghan *et al.,* 2015, Santo Domingo *et al.,* 1998), Hemiptera: Plant sap feeders e.g., aphids, stinkbugs, psyllids etc. (Shan *et al.,* 2021; Overholt *et al.,* 2015; Lin *et al.,* 2019), Blattaria: e.g., cockroaches (Guzman and Vilcinskas, 2020), Mantodea: e.g., praying mantids (Tinker and Ottesen, 2018). Other orders such as Neuroptera, Siphonoptera, Archaeognatha, Dermaptera, Ephemeroptera, Mecoptera, Megaloptera, Odonata, Phasmatodea, Plecoptera, Thysanoptera, Thysanura, and Trichoptera have

also been studied in large scale (Wheeler *et al.,* 2001; Jones *et al.,* 2013; Yun *et al.,* 2014).

Although most of previous reports suggest that many wood-feeding insects overcome the recalcitrant lignin barrier by feeding on predegraded wood or through exosymbiotic relationships with wood-degrading fungi, there are species of insects that feed on the inner wood of living, healthy trees. How these insects are able to circumvent the lignin barrier and gain access to the polymer carbohydrates still begs for explanations. Essentially, in all animals, the digestive tract is the hub of microbial communities associated with that organism (Rajagopal, 2009). Therefore, it is reasonable to suggest that the gut microbial communities of such insects that naturally feed on healthy living trees possess the ability to produce enzymes that can perform lignin breakdown to avail these insects' access to their food source. Hence, the guts of wood-feeding insects represent unique environments where novel lignin degrading enzymes and proteins could be mined to improve the efficiency of industrial biomass pre-treatment processes, detaching lignin from the polysaccharides, and facilitating access to the fermentable sugars in cellulose and hemicellulose (Scully *et al.,* 2013; Fisher and Fong, Rasiravuthanahalli *et al.,* 2017; Muhammad *et al.,* 2017).

Insects have a complete digestive system just like vertebrates (tube from the mouth to the anus). The insect digestive system has three major regions, foregut, midgut, and hindgut though bearing diversified modifications to show adaptation to specific environmental situations and feeding habits, therefore, different gut compartments house specific microorganisms. Several biological and ecological factors influence the composition of microorganisms within different gut compartments, these include developmental stage of insects, social behaviour, morphological variations, and physiochemical conditions in the lumen of each gut compartment (temperature, pH, and oxygen availability) and majorly, host diet and metabolism (Engel and Moran, 2013; Chapman *et al.,* 2013; Prasad *et al.,* 2018; Pal and Karmakar, 2018).

1.11 The African Palm Weevil (*Rhynchophorus phoenicis*)

The African Palm Weevil *(Rhynchophorus phoenicis)* belongs to the curculionidae family of beettles (Coleoptera). It is an important pest affecting mostly oil palm trees in

Nigeria, Cameroon, and other subtropical African countries where it is found. Other host plants of this insect include sugar cane, coconut, and raffia palm as well as the sago palm (Omotoso and Adedire, 2011; Mba *et al.*, 2017). The life cycle of *R. phoenicis,* from egg stage to the newborn adult, takes place within the trunk of host trees (Thomas and Dimkpa, 2016; Montagna *et al.,* 2015). The adults oviposit in palm trees that are wounded or dying, larval development occurs in a week. The young larvae begin to bore tunnels to make their way into the inner part of the trunk and they develop into adult larvae in about 4 weeks (Muafor *et al.*, 2015; Montagna *et al.,* 2015). The larval stage lasts about 2 months after which it enters the pupal stage and metamorphoses into an adult in about 25 days bringing the total life cycle to approximately 3 months (Chung, 2012). The larval stage of development is the most destructive stage of this insect (Figure 1.9). At this stage, the larvae of *R. phoenicis* can burrow and create cavities of more than a metre deep whilst feeding on the trunk of the palm trees without any physical symptoms appearing on the palm tree which may eventually lead to the death of the tree after three to four months of infestation (Bamidele *et al.,* 2013; Harris *et al.,* 2015, Angzzas *et al.,* 2016; Omotoso 2013).

Figure 1.9 Larvae of the African Palm Weevil (*Rhynchophorus phoenicis*)

The major host tree of the APW is the common African oil palm (*Elaeis guineensis*) which has a stout trunk and stands erect, attaining a height of 30m when fully grown. In Nigeria, the oil palms are abundant in the Niger Delta area of southern Nigeria due to the prevailing favourable climatic conditions, which include a humid rainy environment with an annual rainfall exceeding 2000– 3000mm evenly distributed throughout the year and total relative humidity of over 80% with mean monthly

temperatures ranging between 24 - 34°C (NIFOR, 2015; Thomas and Dimkpa, 2016) and thereby having large oil palm plantations as well as the trees growing in the forests which are consequently subject to high infestation by APW.

R. phoenicis, is consumed in some local communities as food (Bamidele *et al.,* 2013). It is popularly known as “Edible worm”, it is eaten in many parts of West Africa and has been reported to be highly nutritious and a rich source of dietary lipids and proteins therefore, studies on the nutritional and medicinal value of *R. phoenicis* are the most reported (Ekpo and Onigbinde, 2004; Womeni *et al.*, 2012; Banjo *et al.,* 2006; Koffi *et al.,* 2017; Mba *et al.,* 2017).

Table 1.7 Lignin content of major host plants of *R. phoenicis*

Host Species	proportion of lignin (%)	References
Phoenix dactylifera	17-27	Al-Zuhair et al., 2015; Ammar et al., 2014, Nasser et al., 2016
Elaeis guineensis	22-52	Saka et al., 2008
Cocos nucifera	25-52	Khalil et al., 2007; Bensah et al., 2015
Raphia spp.	21-24	Israel et al., 2008; Fadele et al., 2017
Saccharum officinarum	18-29	Yao et al., 2015; Ameh et al., 2016

The plants *R. phoenicis* attack have high lignin content (Table 1.7). The predominant lignin moieties found in palm species as is characteristic of angiosperms (Lu *et al.,* 2015) and monocotyledons (Ross and Maza, 2011) are of the guaiacyl (G) and syringyl (S) types (Lu et al., 2015; Saka *et al.,* 2018; Sim *et al.,* 2015) and hence are categorised as hardwoods. Hardwoods comprising predominantly of S type lignin generally contain less lignin (20-25%) than softwoods (25-35%) which have more of the G type lignin known to be more resistant to degradation (Al-Zuhair *et al.,* 2015; Hatakka and Hammel, 2011). Despite this lignin content, the APW is able to overcome the lignin barrier as it excavates and burrows deeply into the interior of the trunk of healthy trees

to feed on the sap leading to their eventual destruction. This indicates that they could have mechanisms that enable them degrade lignin (Geib *et al.,* 2008), hence, the APW larva is a good candidate organism whose gut metagenome should be investigated for novel lignin degrading enzyme. To the best of our knowledge, there has not been any comprehensive investigation or exploration of the gut microbiota of *Rhynchophorus phoenicis* for lignocellulose and lignin degrading enzymes.

1.12 Hypothesis, aim and objectives of study

Hypothesis: Wood feeding arthropods are a potential reservoir of novel lignin degrading bacteria and enzymes

Aim: The Identification and characterization of novel enzyme(s) capable of lignin degradation from the gut of the African Palm Weevil (APW) larvae.

Objective 1: Field Collection and preservation of APW larvae

Objective 2: Bacterial DNA extraction and sequencing for taxonomic profiling and identification of lignin degrading bacteria in different gut compartments of APW larvae (16S rRNA gene sequencing and analysis)

Objective 3: Extraction of whole gut metagenomic DNA from APW larvae and shotgun sequencing

Objective 4: Pre-processing and functional annotation of whole gut metagenome shotgun sequencing data for identification of putative genes with lignin degrading role.

Objective 5: Amplification, Cloning and heterologous expression of selected putative gene sequences in a microbial host and purification of recombinant protein(s)

Objective 6: Evaluation of biochemical activity and enzymatic characteristics of expressed protein(s).

Chapter 2: Taxonomic profiling and identification of lignin degrading bacteria in different gut segments of APW larvae using 16S rRNA gene sequencing.

2.1 Abstract

The microbiota within the guts of insects plays beneficial roles for their hosts thereby contributing to its host's sustenance and survival. One such role is in the regulation of host's metabolism through efficient digestion of ingested food to extract maximum energy. Prior to field collection of our study insect, we compared the DNA preservation efficiencies of three readily available solutions in our laboratory (70% ethanol; E-70, 95% ethanol; E-95, and Nucleic Acid Preservation buffer; NAP buffer) on the wax moth (*Galleria mellonella*) larvae to guide our choice of sample preservation method. Samples preserved in NAP buffer yielded DNA of higher molecular weight and integrity compared to the ethanol preserved samples after a 4-week period. Therefore, the NAP buffer was chosen to preserve the APW larvae we collected from field in Nigeria, to the laboratory in the UK, before DNA extraction. In this study, bacterial metagenomic DNA was extracted from foregut, midgut, and hindgut of NAP preserved larvae of the APW, the V3-V4 hypervariable region of the 16S rRNA gene was amplified and sequenced using the Illumina Miseq platform. The data generated was analysed and taxonomically classified to identify the different bacterial phylotypes present within the gut community cummulatively, and per gut segment. We also determined the presence, diversity and abundance of bacteria associated with lignin degradation within each larval gut compartment as a basis for possible discovery of novel lignin degrading enzymes of bacterial origin and in order to suggest the gut segment(s) where lignin degradation occurs. All sequences were classified and belonged to the bacterial domain. *Firmicutes* (63.7%), and *Proteobacteria* (33.2%) were the most dominant phyla within the gut, followed distantly by *Bacteroidetes* (1.9%) and *Actinobacteria* (1.0%) while *Campylobacteria*, *Desulfovibrio*, and *Verrucomicrobiota* each had very low abundances below 0.1% of the total abundance of taxa recorded. *Enterococcus, Lactococcus, Shimwellia, Lelliotia, Klebsiella* and *Enterobacter* constituted the most abundant genera found across all gut segments. The foregut and midgut had lots of similar genera while the hindgut appeared to be more unique. Overall, 12.3% of total

gut bacteria comprising 16 genera are potential lignin degraders found predominantly in the *Proteobacteria* phylum (91.4%), then moderately in *Actinobacteria* (5.9%) and *Bacteroidetes* (2.7%). The most abundant ligninolytic genera were *Klebsiella* (55.13%), *Enterobacter* (25.58%), *Citrobacter* (5.36%), *Corynebacterium* (4.36%), *Serratia* (3.66%), *Bacteroides* (2.68%), and *Leucobacter* (1.33%) found in different amounts in different gut compartments. The others are *Acinetobacter*, *Ochrobactrum*, *Sphingobium, Microbacterium*, *Novosphingobium*, *Thermomonas*, *Delftia*, and *Pseudomonas* each having total abundance of less than 1%. The foregut had the most diverse and highest abundance of lignin degrading phylotypes and we present reasons that point to the foregut as the location for the depolymerisation of lignin in the APW larval gut.

This study served to justify and rationalise the proposal for further exploration of the APW gut environment for bacterial lignin degrading enzymes using whole metagenome sequencing techniques and functional annotation carried out in the succeeding chapters of this research project.

2.2 Introduction

Beneficial associations between insects and their gut microbial inhabitants especially with respect to host's nutrition can be exploited for biotechnological applications (Harrison *et al.,* 2021; Rajagopal, 2009; Chukwuma *et al.,* 2021).

It has been long known that wood feeding insects are able to digest and utilise plant biomass by the synergistic association they enjoy with the micro-organisms that inhabit their gut (Ali *et al.,* 2019; Chew *et al.,* 2018; Scully *et al.,* 2013; Chauhan, 2020; Kougias *et al.,* 2018). Recently, much attention has been given to understanding the composition of the inhabitant microbes and how they are naturally adapted to facilitate these bioconversion processes (Prasad *et al.,* 2018; Ransom-Jones *et al.,* 2017). Molecular techniques such as PCR and high throughput sequencing have facilitated the studies of microbial communities without depending on the ability to culture individual members of the community as the optimum conditions for growing different species of microbes vary or are yet undetermined for most species (Lazarevic *et al.,* 2016; Stewart, 2012). Structural survey methods of studying microbiomes aim to

identify the taxonomic profiles of the study environments with respect to the types of micro-organisms present (diversity), and their amounts (abundance or richness), from which functional capability can be predicted if desired (Kunath *et al.,* 2017; Knight *et al.,* 2018). Microbiome composition using next generation sequencing can be determined using a whole metagenome sequencing approach or by targeting regions with variability present in all species which can be used to identify species origins called amplicon sequencing or metaprofiling (Escobar-Zepeda *et al.,* 2015). While whole metagenome sequencing can provide a greater breadth of information without the restrictions of sequencing only a single gene, amplicon-based microbiome studies can give more sequencing depth with the same amount of sequencing power and avoids sequencing host DNA instead of microbial DNA (Jovel *et al.,* 2016; Ranjan *et al.,* 2016; Segata *et al.,* 2012). Amplicon sequencing also has reduced computational requirements compared to whole metagenome sequencing, but is limited to the target organisms, excluding non-conforming targets or other interesting organisms in the analysed communities. Both of these techniques rely on accurate and accessible databases of known microbial sequences to compare sequencing reads for phylogenetic classification (Ranjan *et al.,* 2016).

When designing any sequencing-based analysis of microbial communities associated with a host organism where field collection of the host organisms is involved and immediate extraction of DNA is not possible, it is crucial to consider the method of sample storage and DNA isolation to ensure that the DNA is of high quality, appreciable quantity and captures all microbes within the community of interest for a holistic and reliable representation of all taxa and function (Knight *et al.,* 2018; Hammer *et al.,* 2015; Moreau *et al.,* 2013). Poor sample preservation methods if employed, affects the quality and quantity of DNA recovered which in turn affect successful amplification and community structure, thereby resulting in biased inferences (Quince *et al.,* 2017, Hammer *et al.,* 2015). Cryopreservation of samples could be considered the best method of sample preservation for genetic and expression studies where immediate DNA isolation is not possible, but this method cannot always be used for field sampling-based studies as it is also not feasible to freeze samples immediately during collection in the field (Camacho-Sanchez *et al.,* 2013). Alternatively, alcohol and salt-based

solutions such as ethanol, RNAlater, dimethyl sulfoxide (DMSO), cetrimonium bromide (CTAB), nucleic acid preservation buffer (NAP), propylene glycol, etc, that can preserve nucleic acids integrity at ambient temperature are used in such situations and have been tested in previous studies (Campbell *et al.,* 2004; Moreau *et al.,* 2013; Sanders *et al.,* 2014).

Despite the abundance of information on effects of sample preservation methods for metagenomics studies, methods that have been validated for certain sample types cannot be assumed to be optimal for all samples and sampling environments (Hammer *et al.,* 2015). Hence, careful preliminary work to optimize conditions for specific sample types is often necessary (Quince *et al.,* 2017).

Both culture dependent and culture independent approaches have been used in structural surveys of insect gut environments for identification of general or specific microbial forms, e.g., bacterial, fungal, or archaeal communities, or the identification of microbial communities associated with specific functions, e.g., nutrient utilisation or antibiotic resistance etc (Yun *et al.,* 2014). However, culture independent approaches based on high throughput sequencing technologies, consider the full complement of microbial genomes within an environment. These approaches give a more robust and comprehensive representation of microbial populations unlike the culture dependent methods which can only account for populations of microbes that are culturable under the given experimental conditions that are never a perfect simulation of the real environments, thereby producing incomplete, biased, and skewed information (Yun *et al.,* 2014, Egert *et al.,* 2003).

In analysing composition of bacterial communities, 16S rRNA gene amplicon sequencing is the most widely used method (Yang *et al.,* 2016; Thompson *et al.,* 2017; Joynson *et al.,* 2017; Reich *et al.,* 2018; Li *et al.,* 2018; Ransom-Jones *et al.,* 2017; Rajagopal, 2009; Rasiravuthanahalli *et al.,* 2017; Edwards *et al.,* 2010; Muhammad *et al.,* 2017; Sharpton, 2014; Langille *et al.,* 2013). The gene codes part of the small ribosomal subunit of all known archaea and bacteria. It is made up of nine highly conserved regions to which primers can be designed and annealed, and also nine hypervariable regions (V1-V9) which can be used to identify different phylogenetic characteristics of different bacterial taxa and hence used to cluster them based on

phylogenetic affinities as closely related species have similar sequences in each variable region (Woese, 1987; Yang *et al.,* 2016). The most used sequencing platforms such as Illumina Miseq and Roche 454 are designed to reliably and efficiently sequence short DNA fragments generating paired end reads by designing primers matching the conserved regions of targeted hypervariable regions on either side (Garcia-Lopez *et al.,* 2020). Massive sequencing of the approximately 1500bp full-length amplicons of the gene is not feasible with the current short-read high-throughput sequencing technologies, though theoretically, this should yield the best taxonomic resolution (Yang *et al.,* 2016; Alcon-Giner *et al.,* 2017). The choice of hypervariable region(s) to be targeted must be considered carefully to determine the optimum region that can provide the most representative taxonomic profile for the relevant organisms being investigated as different regions have different sensitivities and can significantly introduce bias (Garcia-Lopez *et al.,* 2020; Baker *et al.,* 2003; Klindworth *et al.,* 2013; Walker *et al.,* 2015; Walters *et al.,* 2015). The V3-V4 hypervariable region of the 16S rRNA gene has been targeted in many published sequencing studies of phylogenetic and taxonomic classification of insect gut microbiomes (Ben Guerrero *et al.,* 2016; Lazarevic *et al.,* 2016; Garcia-Lopez 2020, Lluch *et al.,* 2015; Chew *et al.,* 2018).

There is a plethora of studies which have investigated insect gut bacterial compositions using the 16S rRNA gene amplicon sequencing technique and have mostly identified *Proteobacteria*, *Firmicutes*, *Actinobacteria* and *Bacteroidetes* as the predominant bacterial phyla in insect guts amongst many other species and environment specific findings (Prasad *et al.,* 2018; Scully *et al.,* 2013; Do *et al.,* 2014; Santo-Domingo 1998; Ali *et al.,* 2019; Bozorov *et al.,* 2018). Some of these studies have also pointed out the fact that the gut microbiome of insects are non-static and are influenced by factors such as environment (Yun *et al.,* 2014), host phylogeny (Franzini *et al.,* 2016, Mohammed *et al., 2018*; Jones *et al.,* 2013), developmental stage and season (Valzano *et al.,* 2012; Jia *et al.,* 2013), nutrition and diet (Montagna *et al.,* 2015; Muhammad *et al.,* 2017; Ben Guerrero *et al.,* 2016; Berasategui *et al.,* 2017), gut physiology and conditions with respect to pH, temperature, and oxygen availability (Egert *et al.,* 2003; Chew *et al.,* 2018). Regardless, there are core members of the community that are only mildly influenced by such factors that may persist thereby defining the most fundamental

functions performed by the microbiome (Pal and Karmakar, 2018; Reich *et al.,* 2018; Franzini *et al.,* 2016; Ben Guerrero *et al.,* 2016).

Despite the increase in studies of gut microbial communities, studies about how these communities are organized within each gut compartment using culture independent methods are not readily available, as most gut bacterial diversity studies have been about the whole gut communities or are taxa specific. This presents a need for broader and systematic identification of the diversity in each segment of the gut of these insects, in order to provide a wider description of the microbial community and relate the contribution of members of the community in each gut segment to overall host's metabolism, adaptability and survival (Engel and Moran, 2013; Poelchau *et al.,* 2016; Santo-Domingo, 1998).

Industrial scale bioprocessing of lignocellulosic biomass as viable substitutes to fossil-based sources is plagued by lack of efficient pre-treatment and lignin valorisation strategies that align with the global outcry for green and sustainable processes to minimise environmental damage and their climate change consequences. In biorefineries, substituting currently used chemical and thermophysical methods of biomass pre-treatment with biological enzyme-based methods will go a long way in alleviating cost and slowing down climate change. In view of this, researchers have prioritised the exploration of natural biomass utilising systems in a bid to maximize chances of isolating the most efficient candidate enzymes (Olsson, 2016; Brown and Chang, 2014).

Our study insect, being pests of palm trees, live their entire life cycle within the trunk of palm trees, feeding on the palm tissue which have been reported to have high lignin content (Table 1.7). Consequently, it is expected that their guts should harbour an abundance of ligninolytic bacteria that hitherto produce lignin degrading enzymes that facilitate their natural ability to digest their lignocellulosic diet. Accordingly, these gut environments constitute reservoirs for novel lignocellulose degrading enzymes that could be explored for increased efficiency of industrial biomass bioconversion processes into energy and material products.

2.3 Methods

2.3.1 Field collection and preservation of *R. phoenicis* larvae

In order to assess what method of sample preservation to adopt during sample collection as the field is far from the laboratory where the research is being conducted, a preliminary experiment to compare DNA preservation efficiencies of three readily available and comparatively cheap solvents (70% ethanol, 95% ethanol and the NAP buffer pH 5.2) on larvae of the wax moth (*Galleria mellonella*) was undertaken.

2.3.1.1 Preliminary assessment of DNA preservation efficiencies of ethanol versus NAP buffer on larval guts of *Galleria mellonella*

Wax worm (*Galleria mellonella*) larvae were purchased from a pet store in Eccles, UK (Fig. 2.1) and the experiment was set up as shown on table 2.1.

Table 2.1 Experimental set up for determination of DNA preservation efficiencies of different solvents

Sample	Preservative	Conditions	Duration of storage
Six whole larvae	Cryopreservation- (PC)	Storage at -80^{o}C	4 weeks
Six whole larvae	95% Ethanol- (E-95)	Storage at room temperature	4 weeks
Six whole larvae	70% Ethanol- (E-70)	Storage at room temperature	4 weeks
Six whole larvae	Nucleic acid preservation buffer (NAP)	Storage at room temperature	4 weeks
None	1 X PBS only (EC)	Room temperature	4 weeks

Figure 2.1 Larvae of the wax moth (*Galleria mellonella*)

The experimental set up was made up of 6 whole larvae per sample group for the positive control (PC) which were stored directly at -80^{o}C, 30ml of 95% ethanol (E-95), 30ml of 70% ethanol (E-70), and 30ml of nucleic acid preservation buffer (NAP). An extraction control (EC) was also set up containing no larvae but only 30ml of 1XPBS. These samples were kept for 4 weeks at room temperature (with the exception of the cryopreserved samples) after which the larvae were taken out of the respective solutions, placed in petri dishes to dry, dissected, whole guts removed and the 6 larvae in each solvent were treated as one sample from which DNA was extracted. The positive control samples stored at -80^{o}C were also thawed, dissected and bacterial DNA extracted from whole guts of the 6 cryopreserved larvae as explained below.

All gut pieces from each sample above were aseptically cut into smaller pieces under a biosafety hood and quickly transferred to sterile falcon tubes, 1.5 ml of 1x phosphate buffered saline (1x PBS) was added to each tube (for PC and solvent preserved samples) and partially homogenised using a sterile rod to release the microbial cells. DNA extraction protocol was applied for the extraction control sample with only PBS and no larvae stored at room temperature for 4 weeks) to account for any possible contamination in the extraction process as well as for effect of reagents used.

The homogenate was centrifuged at 3000 x g for 3 minutes to separate gut tissue from the released microbial cells in the supernatant. The supernatants from the centrifugation step for each sample was transferred into 2 ml centrifuge tubes and

again centrifuged at 3000 x g for 1 minute to ensure all gut tissue particles have been removed (Santo Domingo *et al.,* 1998).

Bacterial DNA was extracted from the resultant supernatant in duplicates using QIAamp DNA microbiome kit from Qiagen, UK according to the manufacturer's instructions presented in the scheme below (Figure 2.2). Twenty-five microlitres (25 µl) of pure bacterial DNA was eluted from each of the QIAamp mini columns into 1.5ml Eppendorf tubes.

This kit differentially lyses and degrades the contaminating host nucleic acids (by incubation with benzonase) based on the physiological difference in cellular architecture between bacteria and host cells. During this host cell lysis step, the bacterial cells remain intact, and they are subsequently lysed in a bead beating process to ensure efficient lysis of both gram-negative and gram-positive bacteria. Two wash steps are employed to eliminate contaminants and ensure elution of bacteria enriched DNA (QIAamp DNA microbiome handbook, www.qiagen.com).

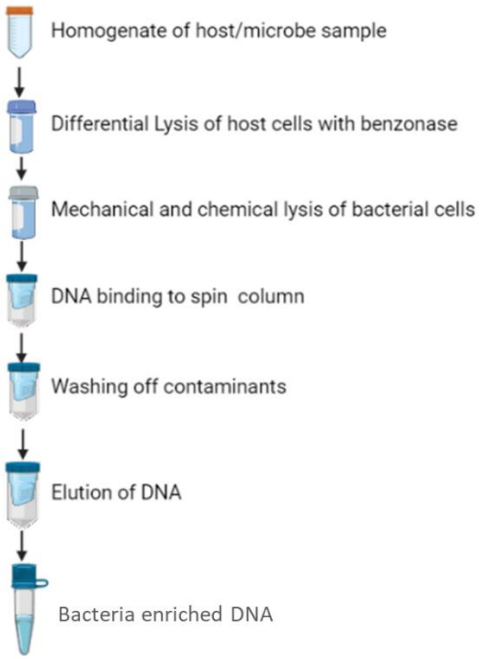


Figure 2.2 Bacterial DNA extraction.
Schematic representation of the QIAamp DNA microbiome kit procedure for extraction of bacteria enriched DNA from mixed host/microbe samples (Adapted from QIAamp DNA microbiome kit manual). Created with BioRender.com

DNA sample concentration and purity were determined using a Nanodrop spectrophotometer (Nanodrop 2000, Thermo Fisher scientific, UK). The DNA samples were run on a 1% agarose gel alongside a 1Kb DNA ladder (Bioline, UK) to confirm the successful extraction of high quality and intact DNA and to compare the integrity of each solvent preserved sample to the cryopreserved sample (positive control) (Santo Domingo *et al.,* 1998).

2.3.1.2 Collection of APW larvae from field

Actively feeding larvae of the African Palm Weevil (*R. phoenicis*) were collected from different freshly felled palm tree trunks (Figure 2.4A and 2.4B) at the Ejekimomi forest reserves of Amukpe village in Sapele, Delta state, Nigeria (5°52'29.9"N 5°42'14.3"E) in May 2018 with the help of locals recruited for this purpose.

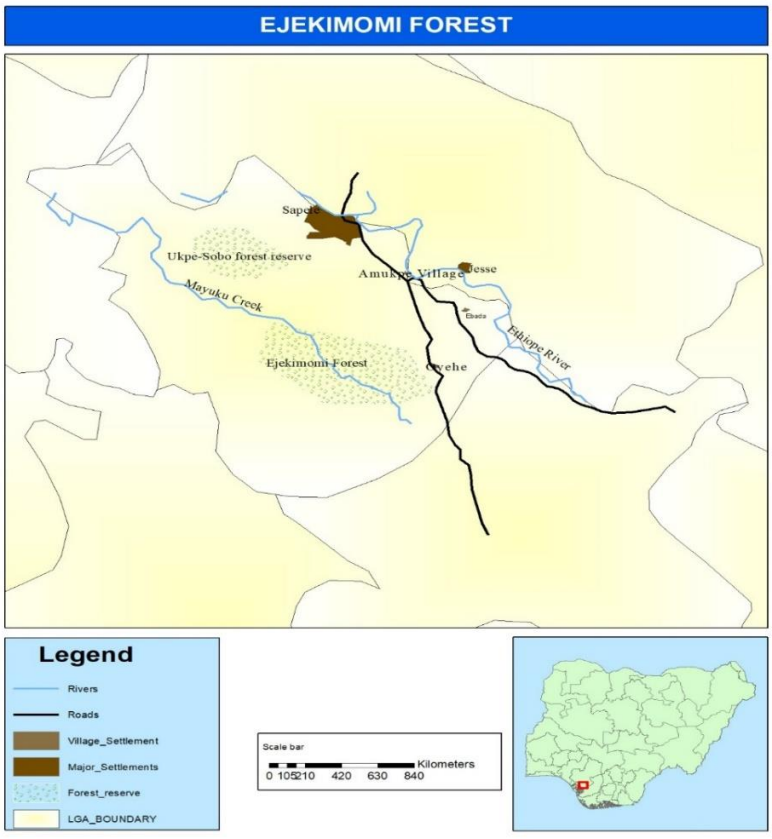


Figure 2.3 Geographical location of the sampling site
A map of Ejekimomi forest reserve area in Amukpe village of Sapele town in Delta state Nigeria where African palm weevil larvae were collected.

The larvae were identified by their morphological characteristics as *R. phoenicis* by Dr. Manasseh Manyi, an entomologist with the Federal university of Agriculture Makurdi, Benue state, Nigeria. They were transported to the Biochemistry Laboratory of the University of Jos, Plateau state, alive in open plastic containers to which chopped bits of the palm tree trunk were added to keep the larvae feeding continuously. The larvae were washed in sterile water to remove dirt and chopped tree particles, surface sterilisation (Figure 2.4C) was done using 70% ethanol and 10% bleach and a second rinse in distilled water (Hammer *et al.,* 2015, Mohammed *et al.,* 2018). Larvae were packaged in sterile containers in the laboratory prepared and sterilised NAP buffer (Figure 2.4D) (Camacho-Sanchez *et al.,* 2013), and subsequently, safely transported to the United Kingdom. The samples were stored at 4^{o}C until dissection and DNA extraction.

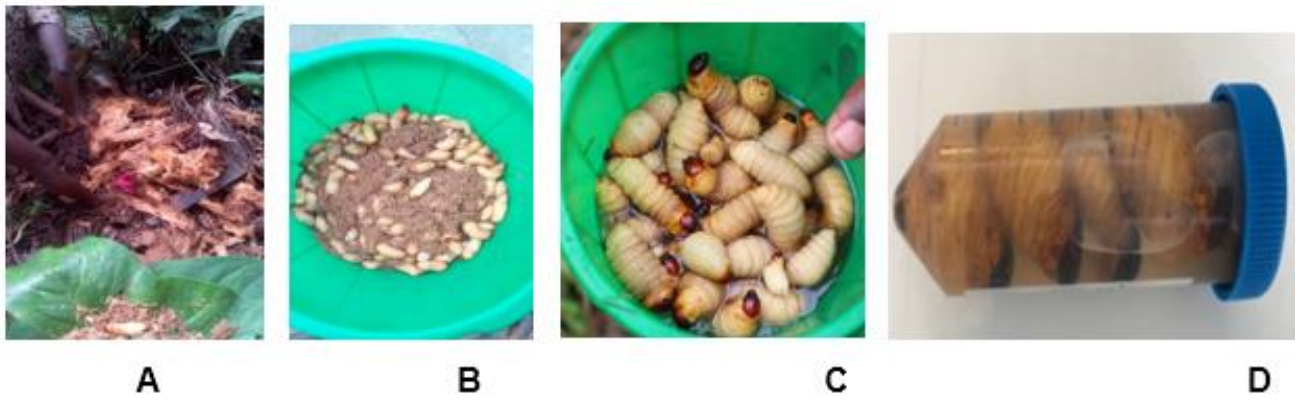


Figure 2.4 Pictures from larvae collection and preservation
A: Felling and hacking down of palm tree to access the APW larvae; **B**: Harvested larvae in a plastic bowl with chopped parts from palm tree trunk; **C**: Surface sterilization of APW larvae; **D**: Sterilized larvae packaged in NAP buffer

2.3.2 Ethics statement

Ethical clearance is not necessary for work carried out on insects (Franzini *et al.,* 2016). Also, *Rhynchophorus phoenicis* has not been listed as protected or endangered species in any national or regional laws. However, ethical approval was obtained to undertake this research due to the Nagoya protocol which emphasizes the need for agreement and benefit sharing when accessing genetic materials from a different country as enshrined in the provisions of the biodiversity convention (Ajai, 1997;

Temitope, 2012). The sample collection was done in open and not protected forests with the agreement and support of the local community.

2.3.3 Dissection and bacterial DNA extraction from larval guts of APW for 16S rRNA gene amplicon sequencing

Stored larvae were removed from the NAP buffer and allowed to dry in a petri dish. Using sterile scalpel and forceps, the larvae were cut open from the mouth to the end of the abdomen and the whole gut aseptically removed. Whole guts were further separated into ‘Foregut’ (F), ‘Midgut’ (M) and ‘Hindgut’ (H) guided by description of each gut segment in Omotoso, 2013. Tissues from the same gut segment from 10 larvae were transferred to separate petri dishes to avoid mixing of gut juices. Gut samples were cut into small pieces, homogenised, centrifuged, supernatant collected, and bacterial DNA extracted from the supernatant from each gut section in duplicates using QIAamp DNA microbiome kit in a similar process described in section 2.3.1.1.

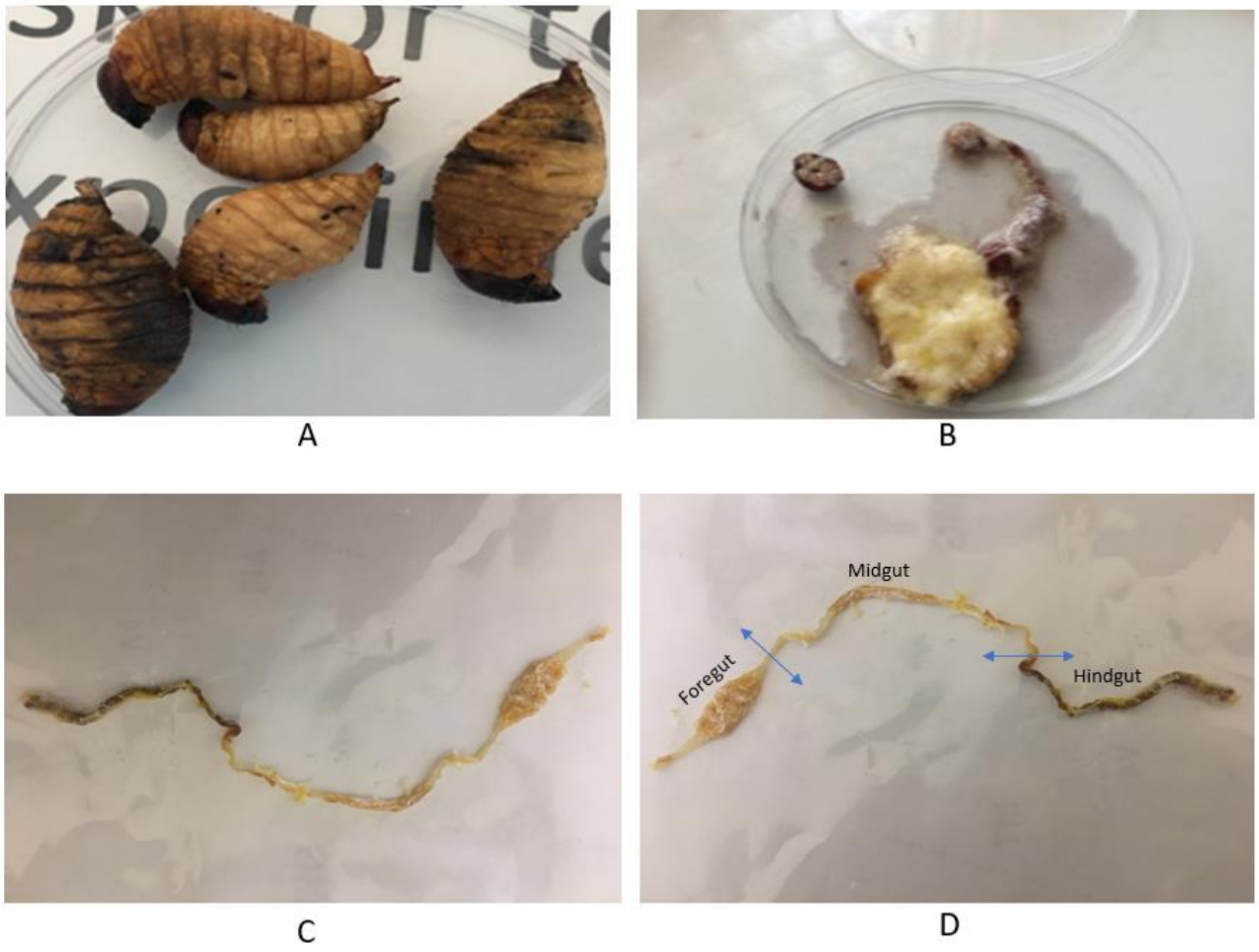


Figure 2.5 Photographs of larval dissection and gut segmentation
A: Preserved larvae on a petri dish prior to dissection; **B:** Dissected larva; **C:** Whole gut of APW larva; **D**: Different segments of the APW larval gut (Foregut, Midgut and Hindgut)

Twenty-five microlitre (25 µl) of DNA was eluted from each QIAamp spin column into 1.5 ml eppendorf tubes. The concentration and purity of the eluted DNA samples were measured using a Nanodrop spectrophotometer (Thermo Fisher, UK). Aliquots from the samples were used as templates for amplification of the 16S rRNA gene and the remaining DNA samples were preserved at -20°C till further required for use. Negative extraction control in which no DNA was added, and the ZymoBiomics microbial community DNA standard (ZYMO research, USA) having a well-defined composition ideal for validation of microbiomics workflows which served as positive control were prepared and processed alongside the gut segment samples.

2.3.4 PCR Amplification and sequencing of 16S rRNA gene

To confirm successful extraction of bacterial DNA from the different gut segments of the APW larvae, the 16S rRNA gene was amplified in a PCR reaction using the Weisberg universal primers as an initial quality check. The primer sequence, composition of the reaction mixture and thermocycling conditions are shown in Tables 2.2, 2.3 and 2.4 respectively.

Table 2.2 Sequence of Weisberg universal primers used for 16S rRNA gene amplification

Name	Forward/Reverse	Sequence
Weisberg bacterial universal primers (*Weisburg et al., 1991*)	Forward	5' AGA GTT TGA TCC TGG CTC- 3'
	Reverse	5' AGF GCT ACC TTG TTA CGA-3'

Table 2.3 Reaction mixture for PCR amplification of 16S rRNA gene

Component	Amount (µL)
2x My Taq red master mix	12.5
Forward Primer (10µM)	1
Reverse Primer (10µM)	1
Template DNA	Variable
Nuclease-free water	Up to 25
Total volume	25

Table 2.4 Thermocycling conditions for 16S rRNA gene amplification reaction using the Weisberg primers

Stage	Temperature(°C)	Time (seconds)	Cycles
Initial denaturation	94	300	1
Denaturation	94	30	30
Annealing	55	45	
Elongation	72	60	
Final extension	72	420	1
Hold	4		∞

Bacterial genomic DNA samples in duplicates per gut sample: Foregut-1(F1), Foregut-2(F2), Midgut-1(M1), Midgut-2 (M2), Hindgut-1(H1), Hindgut-2(H2), Mock microbial community DNA standard (S) and Negative extraction control (C) were sent to Macrogen, Inc. (NGS), Seoul, Republic of Korea, for library preparation and sequencing. The V3-V4 region of the 16S rRNA gene was amplified, libraries prepared, quality validated, and then sequenced on Illumina Miseq 2 X 300bp platform following the protocols in the Illumina 16S metagenomic sequencing library preparation guide (Part #15044223 Rev. B) briefly explained below.

The DNA samples submitted to Macrogen Inc were re-quantified by picogreen using the victor 3 fluorometry method and the purity checked ($A_{260/280}$) on a nanodrop spectrophotometer to ensure they were of good enough quality for library preparation (between 1.8- 2.0). Sequencing libraries were prepared using the Herculase II Fusion DNA Polymerase Nextera XT Index Kit V2 (Illumina). The V3-V4 hypervariable region of the 16S rRNA gene was amplified from the quality assessed template DNA using the set of primers shown in Table 2.5 which were designed to have overhang sequences to allow for addition of Illumina and dual multiplexing index adapters. The PCR product was cleaned up and Illumina compatible sequencing adapters and index adapters were ligated to the amplicon (~460bp) in a limited PCR cycle.

Table 2.5 Primer set used to amplify the V3-V4 region of the 16S rRNA gene

Primer Name	Forward/Reverse	Sequence
V3-V4 region specific primers (337F/ 805R)	Forward- 337F	5’ GACTCCTACGGGAGGCWGCAG -3’
	Reverse- 805R	5’ GACTACCAGGGTATCTAATCC -3’

The quality of the prepared libraries was validated on an Agilent Technologies 2100 Bioanalyzer, using a DNA 1000 chip. Libraries that passed quality control were normalized, pooled, and sequenced on Illumina MiSeq reagent kit V3 (2 x 300bp paired end reads) platform. The sequencing data received from Macrogen were already demultiplexed, paired end fastq.gz files untrimmed for primers and sequencing adapters but had the PhiX sequencing control reads removed.

2.3.5 Data processing and analysis

The data file containing forward and reverse reads for each sample was imported into R-studio software version 4.1.0 (R core team 2020) and was processed following guides from the DADA2 pipeline tutorial 1.16 with slight modifications to suit our reads and desired outcome employing DADA2 package version 1.20.0 (Callahan *et al.,* 2016).

Data pre-processing involved quality profiling, trimming, and filtering of raw data to eliminate read duplicates, low quality reads, adapter, and barcode sequences, etc followed by the generation of an amplicon sequence variant (ASV) table from where taxonomy was assigned as detailed below.

Reads were sorted into “forward” and “reverse” reads, quality assessed by plotting to visualize the sequence quality profiles. Based on the examination of visualised quality plots, the forward and reverse reads were quality trimmed by truncating the reads at positions 290 and 230 (truncLen=c (290,230)) cutting off the last 10 and 70 nucleotides respectively. Reads with unknown bases were discarded and a maximum expected error threshold was set at 2 for forward reads and 5 for reverse reads (MaxEE=c (2,5)) and any reads left were truncated at quality score 2 (truncQ=2). The “trimLeft” function was used to trim off the first 17 and 21 reads from the forward and reverse reads respectively which correspond to primer sequences employed in amplification of the V3-V4 region of the 16S rRNA gene prior to sequencing. The learnErrors function of DADA2 was used to estimate the dataset errors for the filtered and trimmed forward and reverse reads for error correction. Reads that passed quality processing were denoised, paired reads merged, and amplicon sequence variants (ASVs) with corresponding frequencies for each sample were generated (Callahan *et al.,* 2016). Chimeric sequences were identified and removed using removeBimeraDenovo with the “consensus” method whereby each ASV is independently checked and removed if it could be exactly reconstructed by combining segments from two or more “parent” sequences (Faith *et al.,* 2013). After the quality assessment and control, merging and chimera removal process, a total of 658,757 reads out of the initial 1,054,375 raw reads at the start of the analysis were used for taxonomic assignment. Taxonomy was assigned to each ASV using the AssignTaxonomy function which employs a naïve

Bayesian classifier method to compare 8 nucleotide segments of each ASV to a database of known sequences, or a “training set”, and scores ASVs based on their degree of likeness to known taxa, assigning each ASV to the sequence with the highest score (Wang *et al.,* 2007). We assigned taxonomy to genus level only using the Silva_nr99_v138 training set database for bacterial 16S rRNA as the reference database because taxonomic assignments at species level do not yield satisfactory resolution with amplicon sequencing in most cases (Callahan, 2018).

The taxonomy assigned ASVs were transferred to phyloseq package version 1.36.0. Sample sheet was imported, and the sample identities (ID) were merged with the metadata on the sample sheet to make a phyloseq object. Each sample was identified as a true sample (standard, foregut, midgut, and hindgut samples) or a negative (control sample). This phyloseq object was analysed for contaminating taxa from external sources (not sample cross-contamination) using the automated prevalence-based strategy in the decontam package version 1.12.0 (Davis *et al.,* 2018). The strategy relies on the assumption that, as any identified taxa are only a sample of the total present in the sample, contaminating taxa are more likely to be present in negative samples than true samples (Davis *et al.,* 2018). ASVs detected in the control sample were manually analysed to look out for presence and abundance of non-expectant taxa to consolidate on the output from the automatic decontamination. No taxa were identified as external contaminants, so no taxa were filtered out as contaminants. ASVs that correspond to sequences identified as mitochondria and chloroplast sequences were removed and all ASVs that were identified to belong to the same genus were merged.

To assess the accuracy of the sequencing and taxonomic identification procedure, a separate phyloseq object was created containing only the mock community sample, which was analysed by examining ASV counts and comparing their observed relative proportions to the expected theoretical proportions of species declared in the ZymoBIOMICS microbial community DNA standard product literature. This information was represented as a bar chart using Microsoft excel.

Bacterial genera with <0.1% cumulative abundances (total abundance from all gut segments) were filtered out on Microsoft excel and only those with ≥0.1% cumulative

percentage abundances were presented in a Microsoft word table. The most abundant bacterial genera identified per gut segment (≥1.0% for meaningful comparison) were also presented on a bar chart plotted in Microsoft excel. A Venn diagram was created using Microsoft PowerPoint to show taxa shared between the gut segments.

Analyses of bacterial community diversities were performed using the vegan package version 2.5-7 (Oksanen *et al.,* 2019) and data were visualised using ggplot2 version 3.3.5 (Wickham, 2016) on R software. Beta diversity which compares the dissimilarity in features between different samples was calculated using the Bray-Curtis dissimilarity method presented graphically by a non-metric multi-dimensional scaling (NMDS) plot. Diversity within individual samples (Alpha diversity) was estimated by determining the Shannon diversity indices for each gut segment sample which represents the abundance, richness, and evenness of species, and was visualised as box plots.

All lignin degrading bacterial genera identified within our samples were identified manually from the list cross-referenced against literature sources in Table 1.4, and their percentage abundances per gut segment were plotted as histograms on R studio.

2.4 Results

2.4.1 Assessment of DNA preservation efficiencies of ethanol versus NAP buffer on larval guts of *Galleria mellonella*

Bacterial DNA of the wax moth larvae (*Galleria mellonella*) was extracted after preservation in different solvents (70% Ethanol, 95% Ethanol and NAP buffer) for 4 weeks, using the QIAamp DNA microbiome kit as previously described. The concentration and purity measured by a nanodrop spectrophotometer, and the integrity of extracted DNA assessed by agarose gel electrophoresis are shown in Table 2.6 and Figure 2.6 respectively.

Table 2.6 Concentrations and purity of DNA samples extracted from different solution preserved larval guts of the wax moth (*Galleria mellonella*)

Sample	Conc. (ng/ µl)	Purity (260/280)
Positive control (PC)	175.4	1.9
70% ethanol (E-70)	8.6	2.1
95% ethanol (E-95)	45.0	1.88
NAP buffer (NAP)	93.9	1.83
Extraction control (EC)	0.2	1.68

Samples stored at -80°C (PC) had the highest DNA concentration and good purity as expected being the positive control (Glasel, 1995; Hassan and Cheong, 2015). Of the three preservative solutions compared, NAP preserved samples had the best and most closely comparable values to the Positive Control samples followed by 95% ethanol, and 70% ethanol preserved samples. Extraction Control had only a negligible amount of DNA evidencing very minimal contamination during DNA extraction.

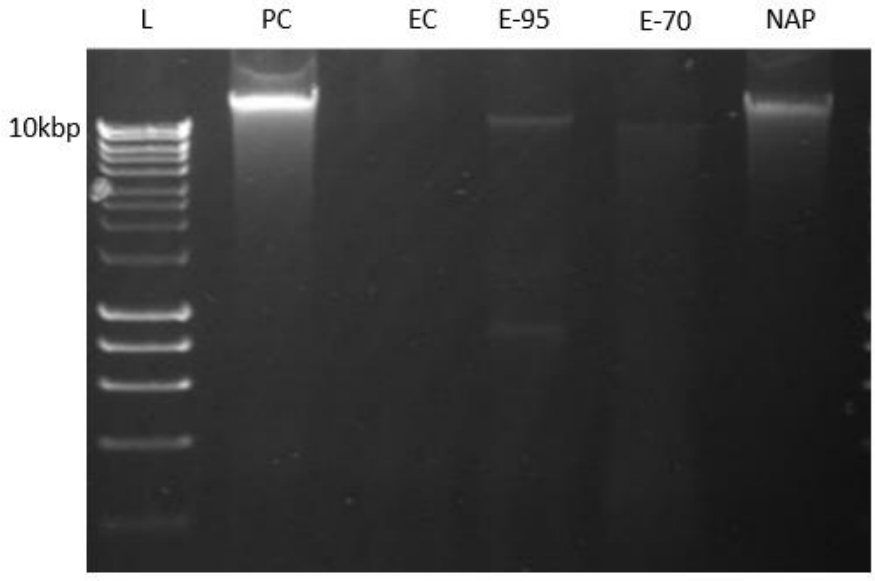


Figure 2.6 Gel electrophoresis image of DNA samples from the wax moth larvae

In lane L, 2µl of 1kb DNA ladder (L) was loaded as DNA marker. Lanes PC, EC, E-95, E-70, and NAP, each contain 2µl of the Positive Control sample (PC), Extraction Control (EC), sample preserved in 95% Ethanol (E-95), 70% Ethanol (E-70), and NAP buffer pH 5.2 (NAP) respectively for 4-weeks.

From the gel image, a band size (≥10 kbp) comparable to the positive control was observed with NAP and 95% ethanol preserved samples. The intensity of the band for NAP preserved samples was closest to that of the positive control, a very faint band was seen with 95% ethanol, no clear band (smear) was seen for 70% ethanol preserved samples and there was no visible band on the gel in the extraction control loaded lane. This shows the DNA from the NAP-preserved gut samples having higher concentration and integrity compared to that from the ethanol preserved samples.

2.4.2 16S rRNA gene amplification

Following DNA extraction, PCR was performed in order to amplify the 16S rRNA marker gene as validation of presence of bacterial DNA. The image on Figure 2.7 confirms the successful amplification of the gene in each sample with the observation of a band size at ~1500bp mark which is the expected size of the amplification product.

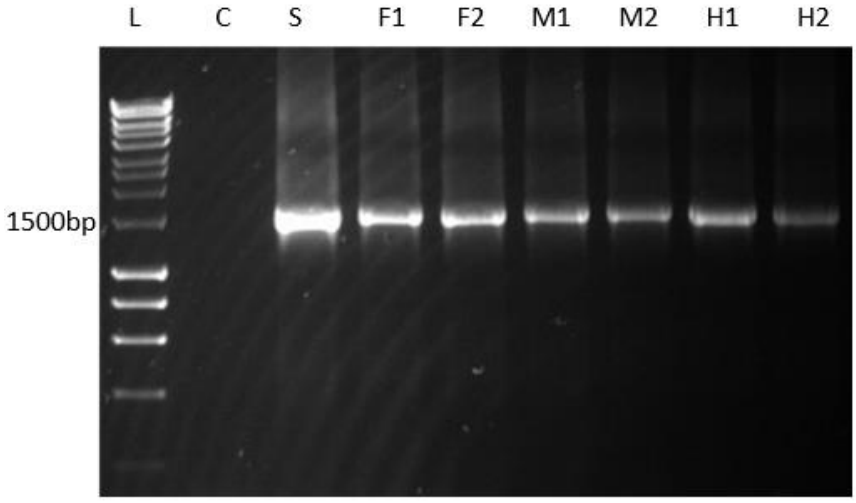


Figure 2.7 Agarose gel electrophoresis image of amplification of the 16S rRNA gene

Lane L: 2µl of 1kb DNA hyperladder, Lanes C - H2: 2µl of DNA negative extraction control (C), 5µl each of Mock microbial community standard (100ng/µl) (S), Foregut 1 (F1), Foregut 2 (F2), Midgut 1 (M1), Midgut 2 (M2), Hindgut 1 (H1) and Hindgut 2 (H2) PCR products respectively.

The gel image in figure 2.7 shows successful amplification of the full length 16S rRNA gene (~1500bp) in sufficient quantity as indicated by the intensity of the bands from all samples except the Negative extraction control.

2.4.3 DNA sample quality control

The figure below shows the concentrations, volumes and thus total amounts of each sample as quantified by fluorometry using picogreen method prior to library preparation given the importance of accurate DNA template quantification for next generation sequencing. All samples passed quality control.

Table 2.7 DNA quality of all samples prior to library preparation

S/No.	Sample Name	Conc. (ng/µl)	Purity (260/280)	Final Volume (µl)	Total Amount (µg)	Result
1	Control	0.351	1.76	10	0.004	Pass
2	Standard	0.978	1.85	5	0.005	Pass
3	Foregut1	22.251	1.82	15	0.334	Pass
4	Foregut2	13.273	1.79	15	0.199	Pass
5	Midgut1	41.106	1.88	15	0.617	Pass
6	Midgut2	14.209	1.86	15	0.213	Pass
7	Hindgut1	5.678	1.82	15	0.085	Pass
8	Hindgut2	6.636	1.89	15	0.100	Pass

2.4.4 Sequencing library quality control

All prepared libraries passed quality control except the “extraction control” sample which was very low in concentration. Nevertheless, all libraries were normalised, pooled, and sequenced. The concentration and size of each library is shown in Table 2.8.

Table 2.8 Quality control data of prepared libraries prior to sequencing

S/No.	Library name	Conc. (ng/µl)	Conc. (nM)	Size (bp)	Result
1	Control	1.81	12.80	218	Low
2	Standard	5.41	41.65	200	Pass
3	Foregut-1	2.59	18.20	219	Pass
4	Foregut-2	3.79	27.52	212	Pass
5	Midgut-1	4.04	29.15	213	Pass
6	Midgut-2	7.04	51.10	212	Pass
7	Hindgut-1	2.03	14.84	210	Pass
8	Hindgut-2	4.10	30.01	210	Pass

2.4.5 Summary of Raw amplicon sequence data statistics

A summary of the raw data generated and submitted by Macrogen Inc following the sequencing of the libraries on a 2 x 300bp Illumina platform is presented below (Table 2.9).

Table 2.9 Raw amplicon sequencing data statistics

S/No.	Sample ID	Total reads bases (bp)	Total reads	GC (%)	AT (%)	Q20 (%)	Q30 (%)
1	Control	11,468,100	38,100	54.235	45.77	91.722	83.379
2	Standard	112,710,052	374,452	53.888	46.11	91.686	83.206
3	Foregut1	165,148,466	548,666	54.300	45.70	91.372	82.707
4	Foregut2	156,689,764	520,564	54.982	45.02	91.244	82.423
5	Midgut1	145,899,516	484,716	54.812	45.19	91.098	82.325
6	Midgut2	136,404,772	453,172	54.341	45.66	91.190	82.516
7	Hindgut1	119,979,202	398,602	54.719	45.28	91.797	83.499
8	Hindgut2	119,740,208	397,808	54.729	45.27	91.430	82.780

The statistics shown above indicate that the sequencing run was a success with each sample recording high total number of paired end reads (except the negative extraction control sample due to low concentration earlier observed with the template DNA and library). Average GC content was 54%. Approximately 92% of total reads sequenced had Phred quality scores of 20 (reasonably good quality data), while ~83% had quality score of 30 (very good quality data) (Andrew, 2010; https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf).

2.4.6 Analysis of Negative control sample (Decontamination)

The duplicate negative control samples “Control” which underwent all amplification, library preparation, sequencing and bioinformatic analysis steps as the true samples were analysed for external contaminants using the decontam package in R. The output returned a “False” result with respect to the assumption that contaminating taxa are more likely to be present in the negative “control” sample compared to true samples. This therefore means that the “contaminant” taxa identified in the control samples are more present in the true samples than in the control. The negative control contained 42 ASVs which were all present in the true samples and had a total abundance corresponding to just about 1.7% of the total taxa abundance found in the true samples. Only *Enterococcus*, *Lactococcus*, *Acinetobacter*, and *Bacteroides* were present at >0.1% each. All the other bacteria each had much more lower values (< 0.1%). Notwithstanding, these taxa were not removed from the true samples as contaminants as they are expected in the true samples and their abundances in the control sample are far lower than what was observed for each of these taxa in the true samples.

2.4.7 Analysis of mock microbial community standard

The mock microbial community DNA standard made up of eight bacterial strains with various theoretical compositions for each strain was sequenced and analysed alongside the other larval gut, and negative extraction control samples to serve as a positive control for ascertaining the fidelity of the 16S rRNA gene amplicon sequencing process, and the performance of the data analysis pipeline employed. Compositions of the bacteria observed in the sequenced “standard” sample compared to the expected theoretical values are shown below (Figure 2.9).

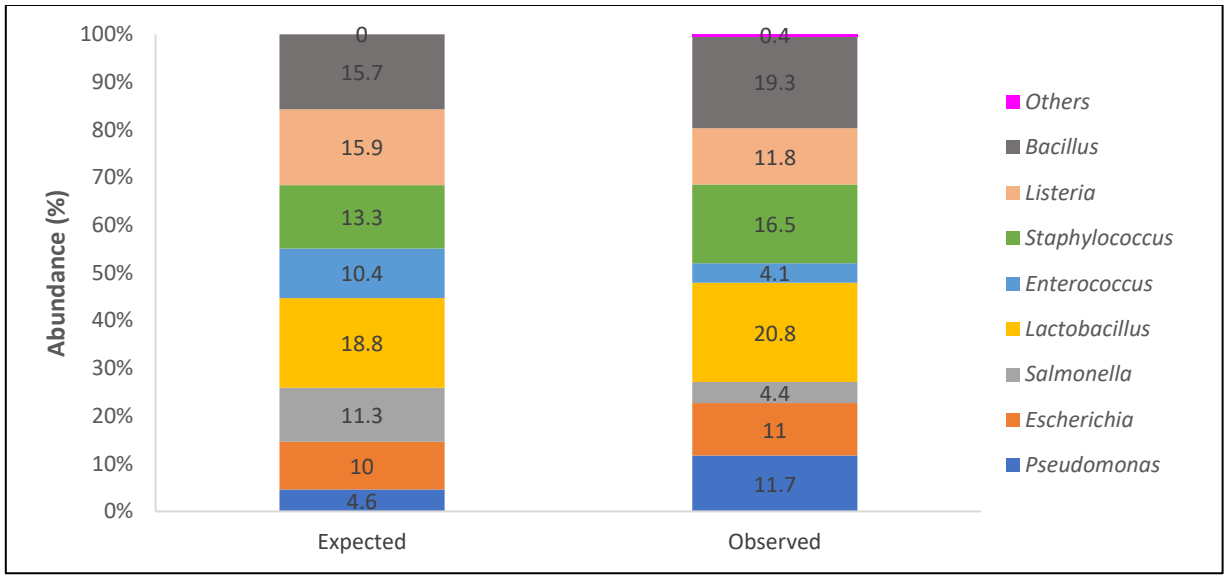


Figure 2.8 Observed versus expected bacterial taxa in the mock microbial DNA community standard.

Stacked bar chart showing percentage abundances of bacterial strains expected within the mock microbial DNA community standard from theoretical data and actual observed bacterial genera identified following sequencing and analysis of the positive control "standard" sample.

All bacterial strains in the mock community standard (expected) were present in the analysed sample (observed) with only few additional strains seen. All strains were abundant in slightly higher or lower percentages but within close range compared to expected values. This is a good indication that the sequencing and analyses progressed successfully without much bias.

2.4.8 Taxonomic profile of APW larval gut bacteria

2.4.8.1 Total bacterial diversity in the APW gut showing percentage abundances

Following taxonomic assignment of identified ASVs generated from the processed sequencing data of the APW larval gut segments only (excluding control and standard), we present a list of bacterial taxa with ≥0.1% cumulative abundances in Table 2.10.

Table 2.10 Taxonomic classification of total bacterial genera identified within the gut of APW larvae.

K, P, C and **G** indicate the "Kingdom", "Phylum", "Class", and "Genus" taxonomic levels.

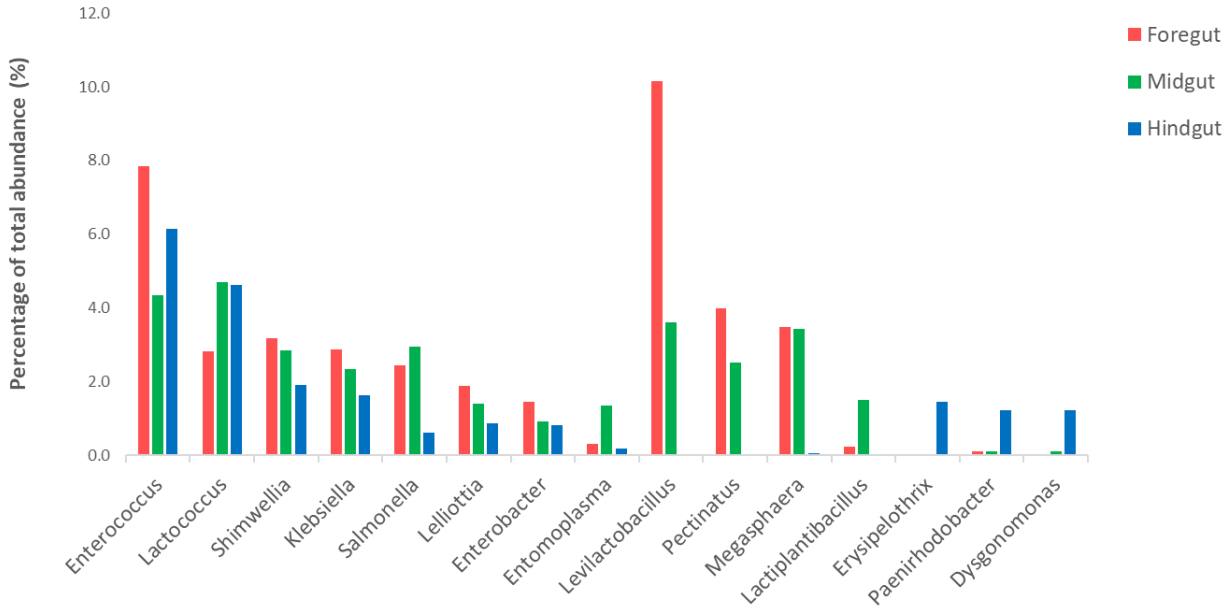
Taxonomic classification	**Percentage Abundance (%)**
K_Bacteria	100
P_Firmicutes	63.7
C_Bacilli	
G_Enterococcus	18.2
G_Lactococcus	12.0
G_Levilactobacillus	13.7
G_Entomoplasma	1.8
G_Lactiplantibacillus	1.7
G_ Erysipelothrix	1.5
G_Paucilactobacillus	0.5
G_Ligilactobacillus	0.3
G_Secundilactobacillus	0.3
G_Leuconostoc	0.2
G_Liquorilactobacillus	0.1
C_Negativicutes	
G_Megasphaera	6.9
G_Pectinatus	6.5
P_Proteobacteria	32.9
C_Gammaproteobacteria	
G_Shimwellia	7.8
G_Salmonella	5.9
G_Lelliottia	4.1
G_Klebsiella	6.8
G_Enterobacter	3.2
G_Cronobacter	0.8
G_Serratia	0.4
G_Pseudocitrobacter	0.4
G_Morganella	0.4
G_Citrobacter	0.7
G_Leminorella	0.1
G_Providencia	0.1
G_Acinetobacter	0.1
G_Mangrovibacter	0.1
G_Yokenella	0.1

C_Alphaproteobacteria	
G_Paenirhodobacter	1.4
G_Gemmobacter	0.3
G_Paracoccus	0.1
G_Rhizobium	0.1
P_Bacteroidetes	1.8
C_ Bacteroidia	
G_Bacteroides	0.3
G_Dysgonomonas	1.3
G_Sphingobacterium	0.1
G_Moheibacter	0.1
P_Actinobacteriodota	0.8
C_Actinobacteria	
G_Corynebacterium	0.5
G_Leucobacter	0.2
C_Coriobacteriia	
G_Atopobium	0.1
Total abundance (0.1% and above)	99.3
Others (< 0.1% Abundance)	0.7

All taxa identified (100%) belonged to the kingdom Bacteria. In all, 165 genera spanning 7 phyla (*Firmicutes, Proteobacteria, Actinobacteria, Bacteroidetes, Campylobacteria, Desulfovibrio,* and *Verrucomicrobiota*) were identified. The dominant phyla with individual genera having a sequence abudance of 0.1% and above were *Firmicutes* (63.7% of total ASV abundance), *Proteobacteria* (32.9%), *Bacteroidetes* (1.8%), and *Actinobacteria* (0.8%) abundance consisting of 40 genera. The remaining 125 genera each had less than 0.1% abundance and constituted only 0.7% of total bacterial abundance within the gut. *Enterococcus, Levilactococcus, Lactococcus, Shimwellia, Megasphaera, Klebsiella, Pectinatus, Salmonella, Lelliotia* and *Enterobacter* was the most dominant genus, listed in decreasing order of abundance. The raw ASV table generated can be seen in appendix.

2.4.8.2 Genus level bacterial diversity and percentage abundances in different gut segments of APW larvae

A.



B.

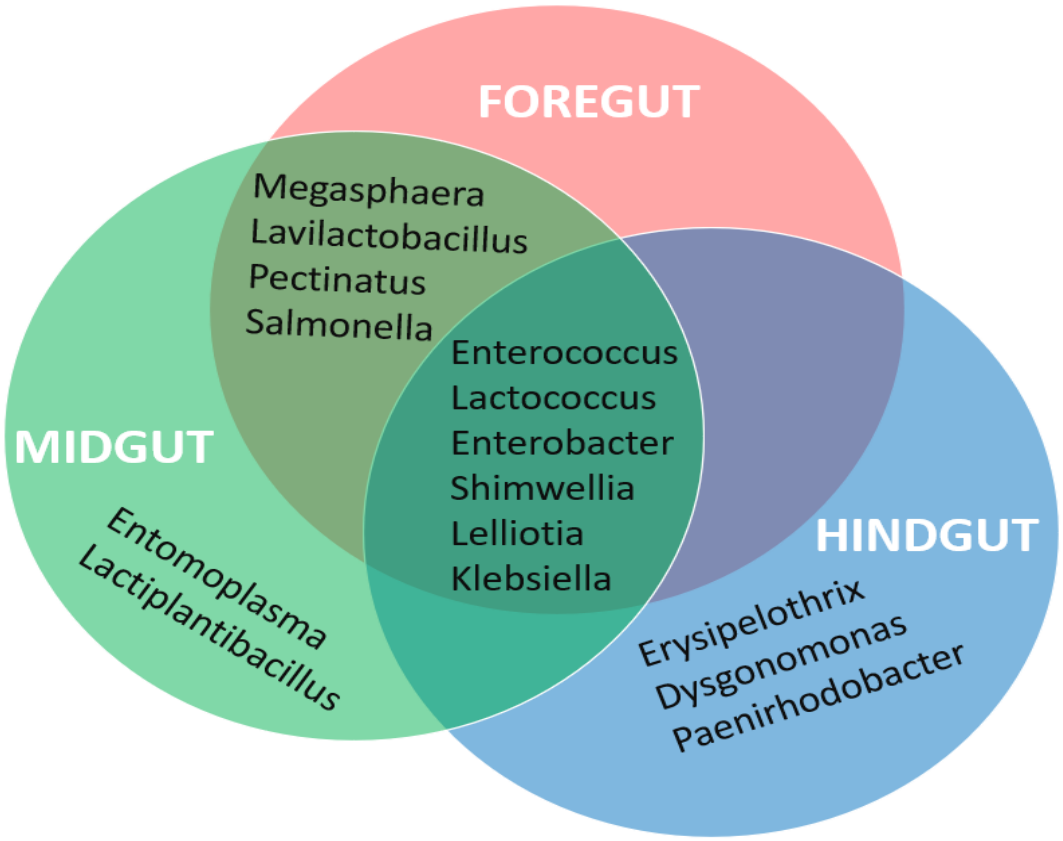


Figure 2.9 Genus-level bacterial diversity and percentage abundances (≥1.0%) in the gut segments of APW larvae.
(A). Bar charts of the dominant bacterial genera identified in the different gut segments of the APW larvae (1% abundance and above).
(B). Venn diagram showing interrelatedness of identified taxa present in the different gut segments of APW larvae
Note: Only the most abundant genera with abundance ≥1% were shown on figure 2.10 for ease of visualisation. However, Table 2.9 shows all bacteria with abundance ≥0.1% mentioned in the description below.

All sequences were classified and belonged to the bacteria domain. *Enterococcus, Lactococcus, Shimwellia, Lelliotia, Klebsiella*, *Enterobacter, Psuedocitrobacter*, *Salmonella*, *Entomoplasma* and *Cronobacter* were the genera found across all gut segments which were deemed to represent the core bacterial microbiota. The foregut and midgut shared lots of similar genera which were completely absent in the hindgut (*Megasphaera, pectinatus, levilactobacillus, Paucilactobacillus, Secundilactobacillus, Leuconostoc, and Atopobium).*The hindgut appeared to be unique containing *Erysipelothrix*, *Morganella, Gemmobacter, Paracoccus, Providencia, Leminoria,*

Yokenella, and *Rhizobium* (0.1%) exclusively, sharing only *Dysgonomonas* and *Bacteroides* (both of *Bacteroidetes* phylum) with the foregut, and *Ligilactobacillus, serratia, Corynebacterium* and *Leucobacter* with the hindgut. Members of the *Bacteriodetes* phylum were absent in the Foregut at >0.1% abundance and a low abundance of *Acinetobacter* and *Liquorilactobacillus* were detected exclusively in the Foregut and Midguts respectively. Compared to the foregut, the midgut had more bacteria in common with the Hindgut though most were present at very low abundances (<1.0%).

2.4.9 Analysis of Beta diversity by Bray-Curtis dissimilarity method.

The distance or dissimilarity between identified bacterial communities from each gut segment was calculated using the Bray-Curtis dissimilarity method, which is based on phylotype abundances, and is shown in multidimensional space on a non-metric multidimensional scaling (NMDS) plot (Figure 2.10)

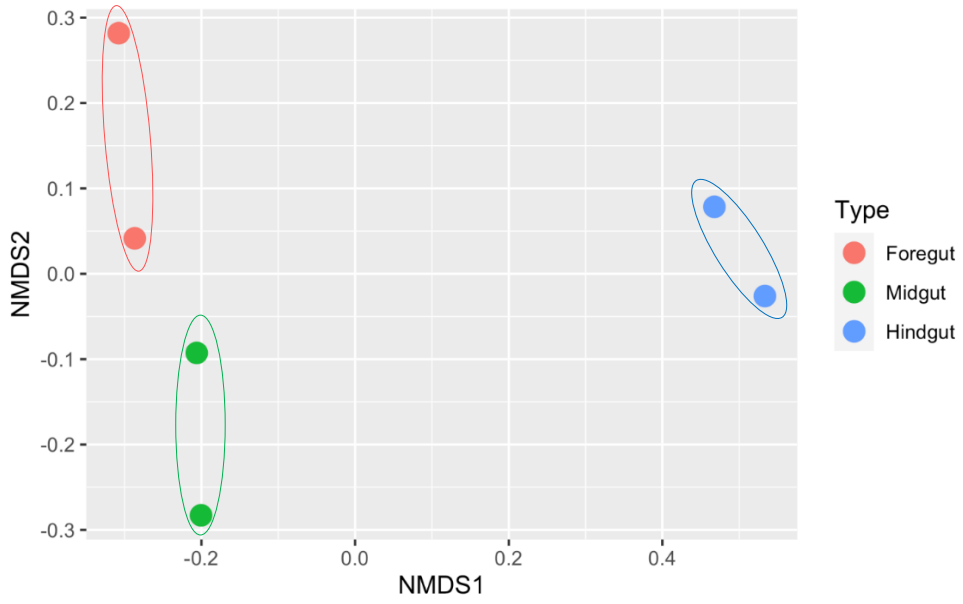


Figure 2.10 NMDS plot

Plot derived from Bray-curtis dissimilarity estimation of bacteria in the different gut communities (Foregut, Midgut and Hindgut) of the African palm weevil larvae. Each point on the graph represents the position of a particular sample in multidimensional space and each gut segment is represented by a different colour. The distance between points represents the difference between each sample. The closer the samples on the graph, the more their similarity.

The duplicate representatives from the Hindgut were more clearly separated from those of midgut and foregut which were more closely ordinated to each other. However, points representing the same gut segment are closer to each other and clearly separated from those representing other gut segments. This shows that the midgut and foregut microbial communities are more similar to each other while the hindgut community is distinctively different.

2.4.10 Alpha diversity estimation by Shannon diversity indices and plot

The species diversity was estimated using the Shannon diversity indices which considers both the abundance and evenness of species present in the different gut segments. Figure 2.11 shows the indices for each gut segment.

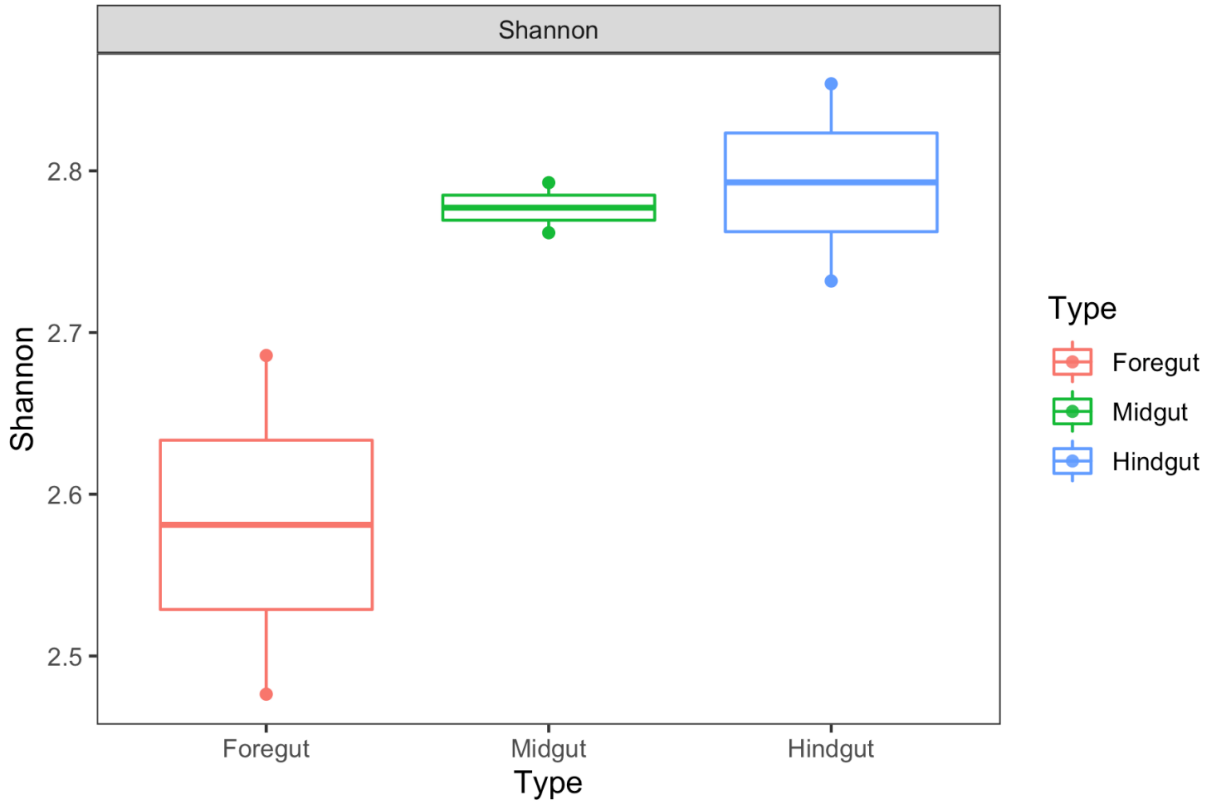


Figure 2.11 Boxplot of Shannon index in different gut segments.
The boxes denote interquartile ranges (IQR) between the first and third quartiles (Q1-Q3) and the horizontal line inside the boxes defines the median. The whiskers which extend from Q1 and Q3 represent the lowest and highest points within 1.5-fold IQR respectively.

The hindgut had the highest Shannon diversity index followed by the midgut and foregut had the least. Species within the midgut community are more evenly distributed as seen by the narrow size of the interquartile range (width of the box), then the hindgut, and the foregut had the most uneven distribution of species.

2.4.11 Diversity and abundance of all identified lignin degrading bacteria in the different gut segments of APW larvae

A total of 16 bacterial genera reported to have lignin degrading ability from several literatures were identified across the different gut segments of the African palm weevil microbiome. They represented a total of 12.3% of all identified genera within the gut. 91.4% of the lignin degraders were from the *Proteobacteria* phylum while only 5.9% and 2.7% were from the *Actinobacteria* and *Bacteroidetes* phyla respectively. *Klebsiella* (55.1%), *Enterobacter* (25.3%), *Citrobacter* (5.4%), *Corynebacterium* (4.4%), *Serratia* (3.7%), *Bacteroides* (2.7%), and *Leucobacter* (1.3%) were the most dominant lignin degrading genera in the gut cumulatively in the listed order. Figure 2.12 shows the different genera identified in each gut segment and their percentage abundances.

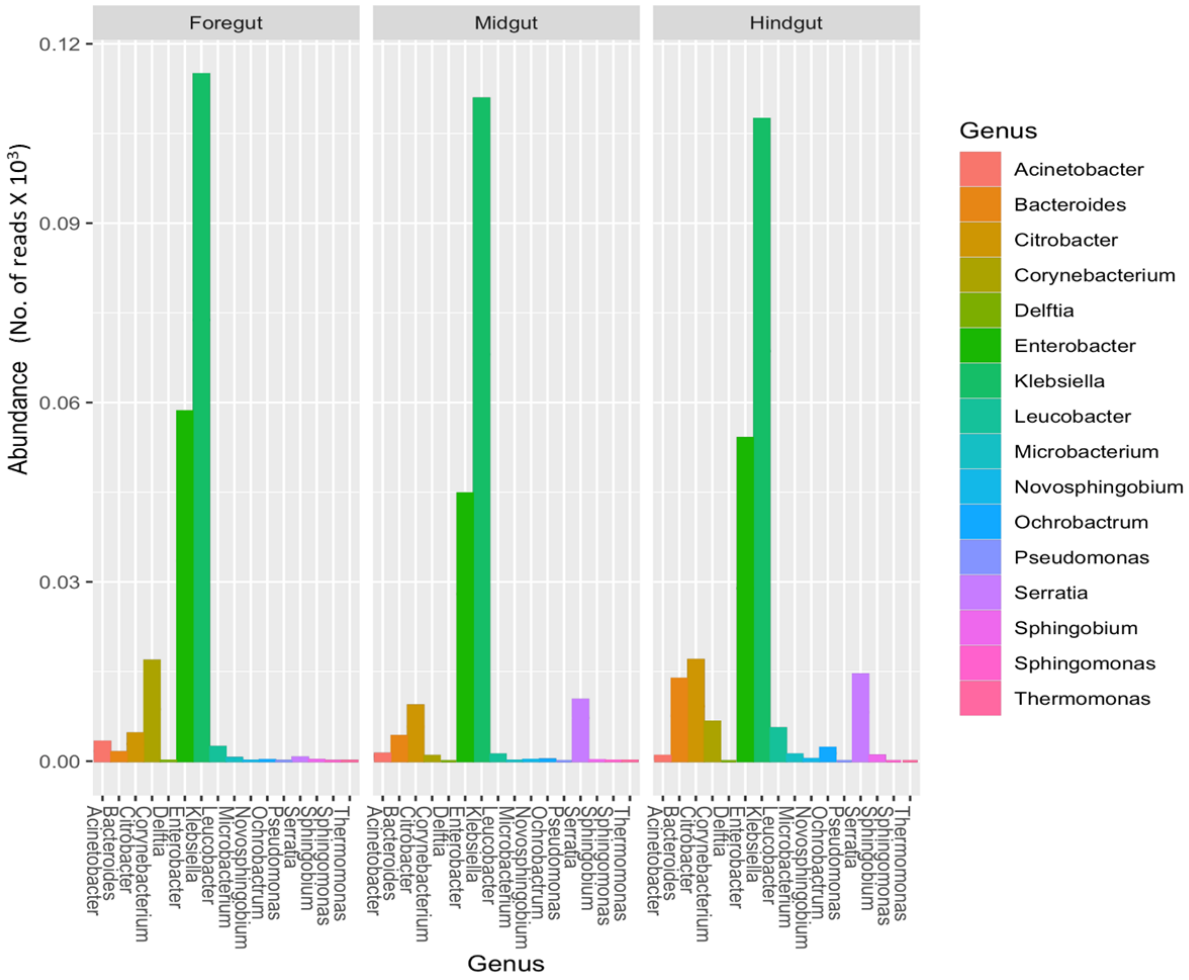


Figure 2.12 Percentage abundances of lignin degrading bacterial genera in different gut segments.
Histogram plots of percentage abundances of individual lignin degrading bacteria identified in the Foregut, Midgut and Hindgut of the APW larvae plotted using R software.

The foregut had all the 16 identified lignin degrading genera adding up to a total percentage abundance of 41.3%. The midgut followed closely having 14 of the identified genera (only *Pseudomonas* and *Delftia* absent) and a total abundance of 31.4%. The Hindgut had the least number of identified bacterial taxa (only 12 genera, with *Pseudomonas*, *Delftia*, *Thermomonas* and *Sphingomonas* absent) and least total abundance of 27.3%. To facilitate comparison, we calculated the percentage abundance of each bacterial genus in each gut segment as a fraction of the total abundance of lignin degrading taxa identified. The three gut segments shared twelve similar taxa in varying abundances with the foregut having the highest abundance of

each taxon in most cases. The percentages of the most abundant bacteria as detected in the foregut, midgut, and hindguts respectively include *Klebsiella* (23.1%, 18.7%, and 13.2%), *Enterobacter* (11.6%, 7.44% and 6.52%), *Corynebacteria* (3.4%, 0.15%, 0.8%), *Citrobacter* (1.42%, 1.9%, 2.03%), *Serratia* (0.12%, 1.78% and 1.75%), *Leucobacter* (0.48%, 0.18% and 0.67%), and *Bacteroides* (0.31%, 0.72, 1.66%). *Novosphingobium, Microbacterium, Ochrobactrum* and *Sphingobium* were also present in all gut segments but in far less amounts. Overall, the foregut had the most diverse and abundant lignin degrading genera followed by the midgut and the hindgut had the least.

2.5 Discussion

The quest to find enzymes capable of biological degradation of lignin as an alternative to chemical and physical methods of lignocellulose breakdown has resulted in research efforts geared towards bioprospecting these enzymes from environments where lignin-degradation is known to occur naturally such as in the guts of wood feeding insects (Fisher and Fong, 2014; Ali *et al.,* 2019). Recently, research on mining the microbiota of insects for genes that code for enzymes and bioactive compounds has greatly increased and are mostly being carried out via metagenomics (Steele *et al.,* 2009; Hammer *et al.,* 2015; Harnpicharnchai *et al.,* 2007; Quince *et al.,* 2017).

Considering the distance between the locations of the field (Nigeria) and the laboratory where this PhD research is to be carried out (United Kingdom), a preliminary assessment of sample preservation efficiency between ethanol (70% and 95%) and NAP buffer with respect to DNA quantity and quality after 4 weeks of storage was carried out. These solvents have been shown to have comparable efficiency according to Hammer *et al.,* 2015, who after studying the effect of six different methods of DNA preservation (Ethanol and NAP buffers inclusive) on four different insect species, reported that, method of preservation had no significant effect on the microbial composition of insect microbiota as differences observed were largely due to the differences in insect species. However, Quince *et al.,* 2017 recommend that preliminary assessment to obtain optimum conditions for specific samples be performed as

methods that have been validated for certain sample types cannot be assumed to be optimal for all samples and sampling environments (Quince *et al.,* 2017). Bacterial DNA extraction (devoid of host DNA) was performed from larval guts of the wax moth in order to have a close enough representation of what we can expect from our samples (Gut of the APW larva) as both are larval stages of insects from closely related orders. From our results, DNA samples of whole gut from larvae preserved in NAP buffer had the closest concentration of DNA and band size appearance to the positive control. We found the concentration of DNA to be best in the NAP buffer preserved guts compared to the ethanol preserved samples tested (Figure 2.6).

Ideally, the positive control should be the bacterial DNA extracted from freshly caught larvae, but cryopreservation maintains the gut microbiota composition and DNA integrity of samples close enough to what is obtainable with fresh samples, so samples stored at -80°C were used as positive control (Camacho-Sanchez *et al.,* 2013; Knight *et al.,* 2018). Also, the possibility of transporting live insects from the field (Nigeria) to the lab (UK) is realistically not achievable due to logistics and biodiversity regulatory constraints. The EU has adopted several strategies to protect its biodiversity as part of her “EU vision and 2020 mission on biodiversity” agreement reached in March 2010, one of which is to combat the pollution of her biodiversity by invasive alien species (DEFRA, 2011).

Although ethanol preservation is relatively cheap and used frequently in studies of insect microbiota, evidence from the shotgun sequencing of DNA from preservative ethanol suggests that DNA may be released from stored samples into the ethanol solution (Linard *et al.,* 2016) thereby possibly explaining the low concentration observed with the ethanol preserved samples as some of the DNA must have leaked out of the samples into the ethanol. Ethanol also has the disadvantage of being highly flammable which makes it unsuitable for transporting specimens by plane (Moreau *et al.,* 2013) further presenting the NAP buffer as a more convenient and safe option considering that our samples were transported on commercial flights (IATA, 2019). To add further support to our findings, Camacho-Sanchez *et al*., 2013 reported from their studies that NAP buffer preserved DNA quality and quantity better than freezing and 95% ethanol in rat brain and tail tissues following storage for both 7 weeks and 10

months. This can be explained by the fact that ethylene diamine tetra-acetic acid (EDTA), a chelator of divalent ions required by nucleases, is a constituent of the NAP buffer (Camacho-Sanchez *et al.,* 2013). Its presence therefore helps to protect DNA from degradation by inhibiting the action of nucleases during extraction process. The protective role of EDTA on DNA has also been reported by Kilpatrick (2002) where they showed that addition of EDTA to ethanol prevented DNA degradation, they also posited that salt-based buffers that contain EDTA such as DMSO and Longmire could preserve DNA at room temperature for at least 2 years or even longer (Kilpatrick, 2002).

In order to assess if it is justifiable to invest time and resources into mining the complete metagenome of the African palm weevil's gut for lignin degrading enzymes of bacterial origin, we decided to investigate the gut for the presence and abundance of lignin degrading bacteria using 16S rRNA amplicon sequencing technique which is a fast, simple and cost effective method that can provide a low-level identification of bacterial taxa adequate as a preliminary investigation of bacterial community structure (Jovel *et al.,* 2016; Alcon-Giner *et al.,* 2017).

APW larvae were collected in Nigeria, preserved in NAP buffer, and transported to the UK based on the outcome of our preliminary experiment with the wax moth (Figures 2.3 and 2.4). We extracted bacterial DNA using the QIAamp DNA microbiome kit (Qiagen, UK. Cat. No. 51704) from the NAP preserved APW larvae since our research interest on lignin degradation was specifically targeted to bacteria and their enzymes. The assignment of all sequences to the bacteria kingdom without any to archaea following 16S rRNA gene amplicon sequencing validates the effectiveness of the extraction kit and reconfirms that indeed only bacterial DNA was extracted.

All other data quality parameters such as total reads per sample, average GC content, quality of base calling denoted by phred scores were within expected ranges (Table 2.9 and Figure 2.8) attesting to the quality of the sequencing run. However, data quality trimming and filtering of poor and uninformative sequences prior to analysis is highly recommended to ensure that only the highest possible quality data is used in order to achieve best results (Jovel *et al.,* 2016).

It is not just enough to have good quality data, it is important to assess that the sequencing did not introduce bias that will misrepresent the true composition of the

microbial community after analysis, hence the need for controls and standards (Lazarevic *et al.,* 2016; Reich *et al.,* 2018). The negative control sample had only bacteria which were also present in the true samples and are bacteria that have been shown to be associated with insect guts (Egert *et al.,* 2003; Janusz *et al.,* 2017; Ceballos *et al.,* 2017). This implies that there were no external or unexpected contamination by any foreign or exogenous bacteria. Reich *et al.,* and Lazarevic *et al.,* recommend that microbial taxa found in the control sample that correspond to genuine or biologically expected microbiota of interest should not be removed from true samples except where they occur in higher relative abundances compared to the samples (Reich *et al.,* 2018; Lazarevic *et al.,* 2016). Hence, the true samples from our results were not decontaminated and no taxa were filtered out as contaminants. The presence of the identified bacteria in the control sample can be explained as a consequence of mild cross-contamination during extraction and library preparation, or from reagents and chemicals used which were the same for all samples.

Analysis of the mock microbial community DNA standard also found all the bacterial strains as expected with only a few additional strains in very low amounts of about 0.4% (Figure 2.9). Although the expected percentage abundances for the bacterial components in the community standard were slightly over-represented as with *Bacillus, Staphylococcus, Lactobacillus, Escherichia,* and *Pseudomonas* or under- represented as with *Salmonella, Enterococcus,* and *Listeria,* overall, the discrepancies are minimal and validate our sequencing and bioinformatic analysis pipeline. The identified discrepancies could be as a result of primer and hypervariable region choice, PCR conditions, library preparation, sequencing, and data pre-processing, and several other variables known to introduce bias in 16S rRNA sequencing which cannot be completely eliminated but can only be minimised (Lluch *et al.,* 2015, Jovel *et al.,* 2016).

Microbiome studies of host-associated gut communities have identified four bacterial phyla (*Proteobacteria, Actinobacteria, Firmicutes* and *Bacteriodetes*) predominantly colonising the gut of insects and most animals (Table 2.10). Other phyla include *Acidobacteria, Verrucomicrobia, Spirochaetes, Tenericutes* etc (Liu *et al.,* 2021; Batista-Garcia *et al.,* 2016; Coleman *et al.,* 2012; Engel and Moran, 2013; Franzini *et al.,* 2016; Huang *et al.,* 2012, Fisher and Fong, 2014). Several factors such as diet and

nutrition, host taxonomy, developmental stage and habitat, seasons, gut morphology and physicochemical conditions etc have been shown to affect the structure of the microbiota in most insect guts and these findings have reported host phylogeny as being the most influential factor with diet contributing significantly especially in lignocellulose feeding insects (Colman *et al.,* 2012; Yun *et al.,* 2014; Chew *et al.,* 2018; Franzinni *et al.,* 2016; Huang *et al.,* 2012; Jia *et al.,* 2013; Tsegaye *et al.,* 2019). Our results agree with the above findings from previous studies of insect gut associated bacterial communities with the detection of the four mentioned phyla being predominant in the APW gut. *Firmicutes* was most abundant (63.7%) followed by *Proteobacteria* (32.9%), then *Bacteroidetes* (1.8%) and *Actinobacteria* (0.8%)

A large scale cross taxon analysis of insect-associated bacterial diversity which investigated the bacterial communities associated with 137 insect specimens representing 39 species of insects using 16S rRNA gene sequencing, reported that on average, most insect bacterial communities were not diverse containing less than 8 phylotypes and being mostly dominated by a single phylotype in which in most of the insects they sampled, the dominant phyla were *Proteobacteria* or *Firmicutes* (Jones *et al.,* 2013). Another large-scale deep sequencing effort by Yun *et al.,* of 305 individual insects belonging to 218 species reported that the gut of insects harbours a diverse collection of bacteria (Yun *et al.,* 2014). These two submissions may seem to be at par with each other. However, Jones *et al*., had excluded phylotypes with less than 1% of the bacterial community in each sample which must have eliminated a large number of taxa with low abundance thereby presenting a community with low diversity. We detected an abundance of diverse bacterial taxa from our data agreeing with Yun *et al.,* but after we used a threshold of 0.1% of total counts per taxa to define abundance (Table 2.10), only a handful (40 out of 165 identified taxa) passed the abundance threshold and was used for further analyses and gut microbiota description, resulting in less diversity and predominance of the *Firmicutes* phylum agreeing with the report by Jones *et al.* The situation explained above calls for caution when comparing findings across different studies as subtle variations in methods and analyses parameters if not carefully considered may lead to wrong conclusions (Knight *et al.,* 2018; Thomas *et al.,* 2012; Quince *et al.,* 2017).

At the genus level, the most dominant individual genera identified were mostly aerobes and facultative anaerobes from the *Firmicutes* (*Enterococcus, Levilactococcus, Lactococcus, Megasphaera, Pectinatus*), and *Proteobacteria* phyla (*Shimwellia, Klebsiella, Salmonella, Lelliotia* and *Enterobacter*). Investigations to detect the effect of different developmental stages on the gut microbiota of the red palm weevil (*Rhynchophorus ferrugineus*), a sister species to the APW (*Rhynchophorus phoenicis*), using non-culture dependent 16S rRNA amplicon sequencing of the V4 hypervariable region also detected similar bacterial taxa including *Enterobacter, Citrobacter, Serratia, Klebsiella, Lactococcus, Entomoplasma, Erysipelothrix*, etc though in varying abundances from our results (Muhammad *et al.,* 2017) which is supported by Jones et al., who posited that greater similarity in bacterial community structure exists among closely related insect species than in less-related species (Jones *et al.,* 2016). The red palm weevil gut has also been reported to have a stable gut microbiota across all developmental stages with differences owing more to nutrition than host taxonomy (Muhammad *et al.,* 2017). The detection of similar abundant taxa or what could be called a 'core microbiome' from our results studying the larval stage and those of studies in larval, pupal, and adult stages of *Rhynchophorus* species are in tandem with this report. There have been several other studies into the microbiota of the red palm weevil being the most investigated species of the *Rhynchophorus* weevils, but the sequencing methods, experimental design and parameters used, and focus of these studies may not allow for an accurate comparison of total gut bacterial profile with our results (Liu *et al.,* 2021, Tagliavia *et al.,* 2014; Montagna *et al.,* 2015; Jia *et al.,* 2013; Angzzass *et al.,* 2016). To the best of our knowledge, this is the first attempt at profiling the microbiome of *Rhynchophorus phoenicis* and there is no published record of gut microbiota studies of other *Rhynchophorus* relatives such as *R. cruentatus*, *R. Palmarum*, and *R. vulneratus.*

All the gut segments of the APW larvae shared an appreciable number of core taxa while the foregut and midgut particularly had more taxa in common hence exhibiting greater similarity in community structure compared to the hindgut which was more compositionally unique (Fig 2.10B). The Bray-Curtis dissimilarity estimation of beta diversity visually represented by the NMDS plot (Figure 2.11) showed each community

to be distinct from the others, and also confirmed the uniqueness of the hindgut community, and the closer relationship between the foregut and midgut communities, judging from the clear separation of hindgut duplicate points being the most distant while duplicate points representing the foregut and midgut are in closer proximity. This same trend was noticed in Figure 2.10. When comparing the diversity between communities, both the number of taxa (richness) and the shortest difference between the observed and hypothetical distribution of each taxon (evenness) should be considered (Escobar-Zepeda *et al.,* 2015). Alpha diversity estimation of the different gut segments using the Shannon diversity indices visualised by a box plot (Figure 2.11) shows that the hindgut harboured more diverse bacterial taxa followed by the midgut and then the Hindgut (the higher the Shannon index the higher the diversity). However, taxa distribution within the foregut was more even than in the hindgut with the most uneven distribution found in the midgut (the wider the size of the box, i.e, the interquartile range, the less even the distribution) (Li *et al.,* 2017, Li *et al.,* 2018). These observed differences support the submission made from other researchers that the difference in morphology (shape, size), and physico chemical conditions (oxygen, temperature, pH, mineral elements) within each gut compartment affects the structure of the microbial community that exists in it (Engel and Moran, 2013; Valzano *et al.,* 2012; Egert *et al.,* 2003; Chew *et al.,* 2018).

In the absence of an existing database of lignin-degrading bacteria to our knowledge, we manually compiled information from many pieces of literature stemming from research where bacteria have been implicated or tentatively confirmed to be associated with the decomposition of any part of the lignin molecule (Table 1.4) and we used that list as a reference document to label bacteria identified within our insect's taxonomic profile as lignin degraders. The lignin degraders constituted 12.3% of the total taxa identified within the larval gut and were drawn from the phyla *Proteobacteria* predominantly (91.4%), *Actinobacteria* (5.9%) and *Bacteroidetes* (2.7%). Although the total gut is dominated by members of the *Firmicutes* phylum, no lignin degraders were identified from this phylum. This may be due to only few members of the *Firmicutes* phylum being reported in the literature we gathered as being lignin degraders and none of those were present within the gut of our study insect from our results. However, the

absence of members of *Firmicutes* is not a misnomer or a cause for serious concern as only the *Proteobacteria* phylum has been consistently reported in all previous research we have accessed on best characterised lignin degrading bacteria. *Firmicutes* along with other phyla such as *Bacteriodetes,* and *Actinobacteria* have been frequently mentioned alongside *Proteobacteria* but not in all cases (See detailed list of references in Table 1.4).

The distribution of the 16 bacterial genera (*Klebsiella, Enterobacter, Citrobacter, Corynebacterium, Serratia, Bacteroides, Leucobacter, Acinetobacter, Ochrobactrum, Microbacterium, Sphingobium, Novosphingobium, Thermomonas, Sphingomonas, Delftia,* and *Pseudomonas*) across the different gut segments and their total abundance per gut is presented in Figure 2.13.

The physical and chemical characteristics of the major components of lignocellulose and the physico chemical conditions such as pH and oxygen availability within the insect's gut are the major determining factors responsible for the distribution pattern of the lignocellulose degrading machinery in insects (Sun and Zhou, 2011; Yun *et al.,* 2014). A detailed morphological and histological description of the APW digestive tract was reported to consist of a foregut, midgut, and hindgut. The foregut is the largest gut segment made up of the buccal cavity, oesophagus, crop and proventriculus which are all adapted for intake, mechanical grinding, storage, and onward passage of food to the midgut. The midgut and hindgut are structurally and functionally adapted for digestion of food, assimilation of nutrients and excretion of wastes (Omotoso, 2013). Same alimentary tract structure was reported for *R. ferrugineus* by Harris *et al* in a similar study of the morphology and histology of the RPW larval gut (Harris *et al.,* 2015) Different segments of the gut have unique characteristics which make them susceptible to colonisation by different types of bacteria (Engel and Moran, 2013). The microbiome within a gut compartment is affected by morphology which varies as insects metamorphose from one developmental stage to the other in most insect orders. The size and shape of the gut additionally influences the availability of oxygen due to the partial pressure of oxygen from the external environment, which in turn determines the metabolism of the inhabitant bacteria (Yun *et al.,* 2014). For effective utilisation of lignocellulose by wood feeding insects for energy, depolymerisation of lignin must

occur first in order to grant access to hydrolytic enzymes to release the stored-up energy in the carbohydrate polymers; cellulose and hemicellulose (Sun and Zhou, 2011; Kumar *et al.,* 2017, Silva *et al.,* 2018). Lignin degradation is an aerobic oxidation process requiring oxidative enzymes such as peroxidases, oxidases, and laccases hence it is believed that these reactions are most likely to occur in the foregut being the anterior part of the gut closest to the external environment where oxygen supply is highest (Chew *et al.,* 2018; Sun and Zhou). In contrast, the midgut and hindgut have been reported as the sites for cellulose and hemicellulose degradation (Egert *et al.,* 2003; Sun and Zhou, 2011; Chew *et al.,* 2018; Yun *et al.,* 2014). These processes which are fermentative, occur by anaerobic hydrolysis thus, it is reasonable to expect the degradation machinery to be domiciled in the interior, anaerobic compartments of the gut that are farther away from oxygen supply.

The presence almost exclusively of aerobic and facultatively anaerobic bacteria (except Bacteroides alone which is anaerobic) within the gut of APW and the specific abundance distribution of lignin degrading bacteria within each gut compartment demonstrates the adaptability of the APW larvae to digesting its diet and suggests where lignin degradation most likely occurs (Mohammed and Alyamani, 2008). Olsson has reported that the gut of mammals houses more obligate and facultative anaerobes while insect guts have a prevalence of aerobes and facultative anaerobes and a large variety of lignin-associated enzymes (Olsson, 2014)

From our results, the foregut of APW larvae possessed the most diverse and highest percentage abundance of lignin-degrading phylotypes compared to the midgut and hindgut. The presence of the proventriculus as part of the foregut of the APW's digestive tract shows their adaptation to their food source (lignocellulosic palm tissues) and explains their ability to offer some sort of mechanical pre-treatment to the lignin in their diet as a first step towards extracting energy from the polysaccharides which occur in the later parts of the gut following a logical order. A similar investigation of bacterial community structure in the foregut, midgut, and hindgut of the wood-feeding termite *bulbitermes sp.* by Chew *et al.,* suggested that lignin degradation was most probably held in the foregut due to the significantly higher relative abundance of the lignin-degrading bacteria, *Actinomycetales* in the foregut compared to the other segments.

They further justified their assertion following a predictive functional profiling where they found energy and co-factors metabolism predominantly occurring in the hindgut whereas oxidative xenobiotics degradation reactions (which are related to lignin degradation reactions) occurred mostly in the foregut (Chew *et al.,* 2018). Overall, our results, supported by the works of Chew *et al.,* and the several other pieces of literature cited above seem to rationalize the foregut of the APW larvae as being the site for lignin degradation prior to cellulose and hemicellulose degradation in the other gut compartments.

2.6 Conclusion

In this chapter, we present evidence to support the use of NAP buffer as a cost-effective and convenient preservative for field collection of insects for metagenomic studies as against the use of ethanol. We also recommend the use of QIAamp microbiome Kit for the selective extraction of pure bacterial DNA devoid of host DNA from host-associated microbiomes. Most importantly, our study which represent the first known metaprofiling effort of the bacteria colonizing the gut of the African palm weevil, *R. phoenicis* to date, has revealed great similarity in bacterial community structure with those identified in most insects, and specifically with the bacterial microbiota of the phylogenetically related red palm weevil, *Rhynchophorus ferrugineus*. The presence of an appreciable number of lignin-degrading bacteria within the larval gut suggest an immense potential for the discovery of lignin-degrading genes and enzymes. Furthermore, Lignin degradation in the African palm weevil is believed to be domiciled in its foregut due to the presence of a proventriculus that serves to mechanically decrease the structural complexity of lignocellulose as a first step towards degradation, and the greatest abundance of mostly aerobic and facultatively anaerobic bacteria capable of oxidatively decomposing lignin predominating the foregut. Our findings point towards the gut of the African palm weevil being a reservoir that harbours a consortium of bacteria capable of lignin degradation/modification from which lignin-degrading genes and enzymes can be harvested. We, therefore, have a good reason and justification to employ the more expensive, time and resources-consuming shotgun metagenomic

sequencing and functional annotation methods to bio prospect bacterial lignin degrading enzymes from the African palm weevil's gut microbiota.

Chapter 3: Functional metagenomic studies of APW larval gut metagenome in search of lignin degrading genes

3.1 Abstract

Following the identification of an interesting, unique, and diverse community of bacteria associated with lignin breakdown present in the gut of the African palm weevil using 16S rRNA sequencing, a full functional metagenome analysis was conducted. Total bacterial DNA from whole guts of APW larvae were prepared as sequencing libraries and sequencing of the whole gut bacterial metagenome was conducted using Illumina-based next generation sequencing (NGS) technology. Bioinformatic analysis including data quality control, metagenome assembly, taxonomic profiling, open reading frame (ORF) prediction and functional annotation of predicted ORFs produced an annotation data output file with gene sequences assigned putative functions. In total, 2.89 Gbp of data was generated and analysed, 60,615 ORFs were identified and annotated from 2.71 Gbp of data after quality control. Putative functions were assigned to 15,892 genes whereas a vast majority (43,847) were designated “hypothetical” genes. Genes related to lignocellulose degradation, particularly lignin degradation were identified using the CAZy database as a reference. We found 249 genes potentially encoding lignin degrading enzymes (members of the Auxiliary Activities class of the CAZy database) from AA1, AA2, AA3, AA5, AA7, AA9 and AA10 families. Genes corresponding to members of other classes of enzymes in the CAZy database were also detected in substantial amounts (Glycoside hydrolases- GH: 191; Glycosyl transferases- GT: 276; Polysaccharide lyases- PL: 32; and Carbohydrate esterases- CE: 423) suggesting the presence of a microbial consortium within the gut of APW capable of decomposing all components of lignocellulose and thus facilitating its ability to extract energy from its palm tissue diet. To validate the above findings from sequencing and bioinformatics analysis, we selected 3 genes predicted to be lignin degraders (2 deferrochelatase/ peroxidases; belonging to the DyP-type peroxidase superfamily and 1 polyphenol oxidase) for PCR amplification from whole genome amplified stocks of the original DNA that was sequenced. All selected genes were successfully amplified yielding products of the expected sizes as predicted validating the computationally generated gene

prediction and thus indicating that these genes are naturally present in the APW gut microbiota. The thousands of hypothetical genes reported offer an inexhaustible resource that could be studied in the future for in-depth characterisation of the functions contributed by the APW gut bacteria, as well as the recombinant production of enzymes for biotechnological applications.

3.2 Introduction

In order to reduce the cost, energy demand and environmental burden of industrial processes currently employed in the manufacture of a wide and diverse range of products required to meet the consumer needs of an ever increasing global population, it is imperative to exploit biological systems for biocatalysts (enzymes) to facilitate biochemical transformation methods that can replace fossil-based products in agricultural, pulp and paper, pharmaceutical, and chemical industries (Arnau *et al.,* 2020; Gurung *et al.,* 2013; Ekas *et al.,* 2019; Choi *et al.,* 2015). The majority of commonly applied industrial biocatalysts are sourced from micro-organisms (Antwis *et al.,* 2020; Ferrer *et al.,* 2009; Simon and Daniel, 2009). However, about 99% of microbes are not amenable to lab culture thereby limiting the genetic space we could sample to bio prospect for novel catalysts. Thus, the contribution of microbial biochemical diversity to problems such as lignocellulose degradation cannot be reasonably exploited by employing culture-dependent methods. These methods involve cloning fragments of DNA libraries into bacterial artificial chromosomes (BACs) or fosmids, and functional activity identification on an agar plate or liquid culture assays is limited to only a few out of thousands of cultivated clones (Schmeisser *et al.,* 2007; Joynson, 2015; Bodor *et al.,* 2020; Batista-Garcia *et al.,* 2016; Stewart, 2012). Metagenomics has allowed researchers to access the full genetic potential of both culturable and non-culturable organisms present in a defined community (Quince *et al.,* 2017; Knight *et al.,* 2018; Thomas *et al.,* 2012). The advent and rapid development of high throughput sequencing based metagenomics in concert with the fast-developing field of bioinformatics and computational analysis makes it possible to extract meaningful information about the taxonomic, functional, and evolutionary aspects of microbial communities from their biochemical repertoire (Knight *et al.,* 2018; McDonald

et al., 2010; Simon and Daniel, 2011; Cragg *et al.,* 2015). Genetic constituents of complex microbial communities from sequence data can be compared against the rapidly increasing, publicly available databases that contain thousands of known gene functions that code for the corresponding enzymes of interest (Madhavan *et al.,* 2017; Rooks *et al.,* 2012; Bragg and Tyson 2014; Sharpton, 2014).

Studies using NGS-based metagenomics have facilitated the discovery of lignocellulolytic genes and enzymes from a wide range of environments (Table 1.6) as well as other biomolecules evidenced by the increasing number of such research output recorded in the Genomes OnLine Database- GOLD and other publicly relevant databases (Mukherjee *et al.,* 2019; Pagani *et al.,* 2012).

These methods have been specifically applied in bioprospecting lignocellulose degrading enzymes for biomass biorefining. Environments with extreme conditions of temperature and pH such as landfill sites (Ransom-Jones *et al.,* 2017), sugarcane bagasse (Mhuantong *et al.,* 2015), and more predominantly the guts of wood feeding organisms have been the most common targets (Watanabe and Tokuda, 2010; Huang *et al.,* 2012). It is expected that enzymes from these understudied environments may be optimally active and stable at wider ranges of these physical conditions which are more compatible with most industrial processes and hence more suitable for industrial scale applications (Simon and Daniel, 2011; Steele *et al.,* 2009). Gut microbial communities of wood-feeding insects in particular, have the capacity to produce enzymes that facilitate the degradation of lignocellulosic material thereby constituting unique ecosystems that may serve as store houses of novel proteins and enzymes that could be exploited to enhance the efficiency of industrial biomass pre-treatment processes, decoupling lignin from wood polysaccharides and facilitating access to fermentable sugars in cellulose and hemicellulose in line with the drive for replacing fossil based products with more sustainable and environmentally friendly products (Watanabe and Tokuda, 2010; Huang *et al.,* 2012).

While appreciable success has been recorded with the discovery of cellulose and hemicellulose degrading genes/ enzymes from both fungal and bacterial sources (Arnau *et al.,* 2020; Joynson *et al.,* 2014; Willis *et al.,* 2010; da Costa *et al.,* 2018; Gong *et al.,* 2017; Liu *et al.,* 2018; Bhalla *et al.,* 2014; Ransom-Jones *et al.,* 2017; Edwards

et al., 2010), the majority of identified lignin degraders are of fungal origin, not well suited for industrial applications due to inability to adapt to extreme conditions and genetic manipulations. Efficient lignin degrading enzymes of bacterial origin remain elusive for industrial application, although metagenomic investigations of the guts of slugs, long horned beetle, gribbles, sugar cane bagasse, and more has identified genes with lignin degrading potentials (See Table 1.6), only a few of these genes have been recombinantly produced and characterised to any significant extent (Ahmad, 2010; Chen and Li, 2016). In this section of our research, the lignin/lignocellulose degrading potential of the notorious pest of palm trees, *R. phoenicis*, is investigated via whole shotgun sequencing and functional metagenomic analysis, followed by PCR amplification of selected genes of interest to validate the sequencing and bioinformatic analyses as a first step towards cloning and recombinant expression of the protein products of those genes.

3.3 Methods

R. phoenicis larvae collected from Nigeria and preserved in NAP buffer were dissected, and metagenomic DNA was extracted from whole gut tracts using the QIAamp DNA microbiome kit (Qiagen, UK). To fully describe the lignin degrading potential of *R. phoenicis* larval gut microbiota, shotgun sequencing of the bacterial metagenomic DNA was performed using the V2 chemistry on Illumina Miseq. Raw reads generated from shotgun sequencing were quality assessed, quality trimmed and assembled into contigs from which gene prediction and functional annotation was conducted. Whole genome amplification (WGA) was performed to increase the available library DNA pool and three selected genes of interest from the functional annotation file generated were targeted for PCR amplification using the whole genome amplified DNA as template to validate the sequencing and bioinformatic processes so far.

3.3.1 Dissection of APW larvae and bacterial DNA extraction

Whole gut tracts of *R. phoenicis* larvae were removed following dissection as previously described in section 2.3.3. Five guts were each placed in 3 different tubes labelled MG-1, MG-2, and MG-3 and bacterial metagenomic DNA was extracted from the samples

in each tube using QIAamp microbiome kit as already described in section 2.3.1.1. An extraction control (1.5 ml 1x PBS) with no gut sample; MG-C was also prepared to account for any possible contamination in the extraction process as well as for effect of reagents used.

Equal amounts of the DNA samples (MG-1, MG-2, MG-3) were additionally pooled together to make one composite sample (MG-P) with increased volume and encompassing more guts (15 whole guts in all) to ensure the microbiota of all possible bacteria are represented. DNA quality indicators, which include DNA integrity (quantity and quality) and purity (Lucena-Aquilar, 2016) were checked (for both replicate and pooled samples) using a nanodrop spectrophotometer (Nanodrop 2000, Thermo Fisher scientific, UK) and a Qubit fluorimeter (Qubit 3, Invitrogen, UK). The size and quality of the extraction control (MG-C), and the pooled metagenomic DNA samples (MG-P) were checked by agarose gel electrophoresis run on a 1% agarose gel (Bioline, UK) supplemented with SYBR™ Safe DNA Gel Stain (Fisher Scientific, UK), alongside a 1Kb DNA ladder (Bioline, UK). Gel images were captured using the UV transillumination feature of the G: Box (Syngene) to confirm the successful extraction of high integrity and intact DNA (Lee *et al.,* 2012). The purity of the DNA sample, which is crucial for shotgun library preparation, was determined spectrophotometrically from the absorbance ratios $A_{260}/_{280}$ (indicator of protein contamination) and $A_{260}/_{230}$ (indicator of organic solvent residues). Values in the range between 1.8 - 2.0 and 2.0 - 2.2 for $A_{260}/_{280}$ and $A_{260}/_{280}$ respectively, generally indicate DNA of high purity (Liu *et al.,* 2009; Glasel, 1995; Gallagher, 1998). Samples were stored at -20°C till future use.

3.3.2 Whole gut metagenomic library preparation and shotgun sequencing

The Nextera DNA library prep kit from Illumina (Illumina Inc., UK) was used to prepare indexed, paired end libraries from the bacterial DNA extracted and pooled (MG-P) from whole guts of the APW larvae according to the instructions in the kit user guide briefly explained and summarised in the chart below (Figure 3.1).

The metagenomic DNA (MG-P) was diluted and quantified by the dsDNA (double strand DNA) assay method using a Qubit fluorimeter to a concentration of ~2.5ng/µl. Twenty-microlitre (20 µl) of this DNA (~50ng total) was tagmented (a process where

the DNA is fragmented and tagged with adapter sequences) with the help of an engineered transposase enzyme in a limited PCR cycle. A clean up step to separate the tagmented DNA from the Nextera transposome was carried out using a ZYMO purification kit: ZR-96 DNA Clean & Concentrator™-5 (ZYMO research, Cambridge Bioscience, UK) as specified in the Nextera library prep protocol. The product of this tagmentation reaction was confirmed using a tape station (Agilent technologies 2200) in which a broad distribution of DNA fragments with size ranging between 150bp-1kb is expected. Next, the purified tagmented DNA was amplified and a selected combination of index 1(i7) and index 2(i5) adapters were added in a 5 cycle PCR reaction. The reaction mixture and thermocycling conditions for the amplification are shown on tables 3.1 and 3.2 respectively.

Table 3.1 PCR reaction mixture for amplification of tagmented DNA for addition of index adapters

Component	Amount (µL)
Cleaned tagmented DNA	20
Nextera XT index 1 Primer	5
Nextera XT index 2 Primer	5
Nextera PCR Master Mix (NPM)	15
PCR Primer Cocktail (PPC)	5
Total reaction volume	50

Table 3.2 Thermocyling conditions for PCR amplification of tagmented DNA for addition of index adapters

Stage	Temperature (°C)	Time (seconds)	Cycles
Initial denaturation	72	180	1
	98	30	1
Denaturation	98	10	
Annealing	63	30	5
Elongation	72	180	
Hold	10		∞

Indexed libraries were further purified to remove contaminants, and size selected using AMPure XP beads according to the stipulated protocol in the Nextera DNA library prep

reference guide to achieve a size distribution of ~250bp to 1000bp which is the recommended range required to create optimum cluster densities and achieve highest quality data on the Illumina sequencing platform (Head *et al.,* 2014). The libraries were quantified using a Qubit fluorimeter and fragment size verified by tape station (Agilent technologies, UK).

In preparation for sequencing, the libraries were normalised to 2nM using Tris-Cl with 0.1%tween 20, denatured using 0.2M NaOH, and diluted to 8pM (Nextera DNA library prep reference guide). Sequencing control (PhiX) was also added to the diluted libraries and 600 µl of this mixture was loaded unto the sequencing cartridge and run on an Illumina Miseq using the V2 chemistry and 2x250 bp read length at the genomics facility of the University of Salford.

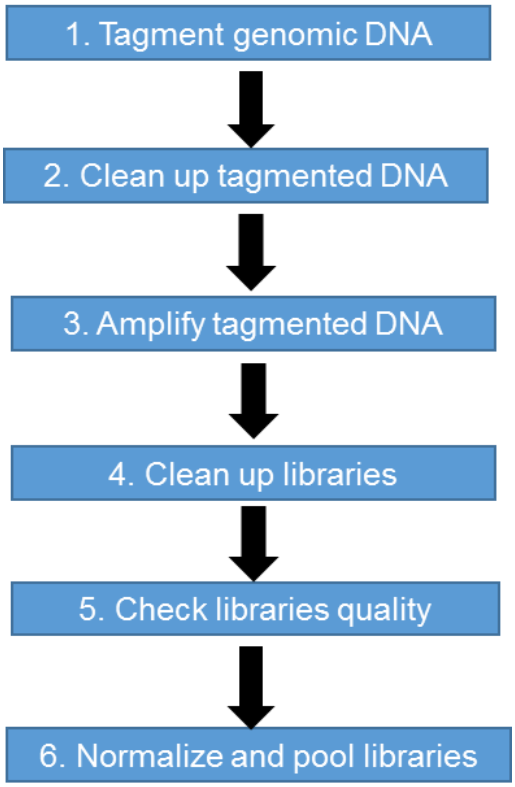


Figure 3.1 Metagenomic library preparation workflow
Illustration of the workflow for metagenomics library preparation using the Nextera DNA library prep kit from Illumina (Adapted from the Nextera DNA Library Prep Reference Guide, document # 15027987 v01, January 2016).

3.3.3 Quality control of sequence data

The raw data generated above was organised in two files; forward (R1) and reverse (R2) and will be referred to as R1 and R2 respectively going forward. Raw data quality was assessed visually by running a FastQC analysis (version 0.11.8) (Andrew 2010; FastQC manual). Illumina sequencing adapters, and low-quality bases were trimmed off using trimgalore version 0.6.3 with default settings (quality score parameter set to 20 and a stringency of 4 as per published guidance). Trimming occurs in the form of cutting off bits of low-quality sequences at the set threshold from the ends of reads and discarding sequences that align with Illumina adapters from the start of reads (Kreuger, 2015). The quality profiles for both R1 and R2 were re-inspected by running another FastQC analysis post trimming.

3.3.4 Assembly of shotgun metagenomic data using MEGAHIT

The quality-trimmed short sequence reads were assembled *de novo* into longer contiguous sequences called “contigs” by basically identifying and merging overlapping read pairs for maximum quality of ORF prediction (Kunath *et al., 2017*; Knight *et al.,* 2018, Thomas *et al.,* 2012). The MEGAHIT metagenome assembler version 1.2.9 set to default parameters with the following set of k-mers: 29, 49, 69, 89, 109, 129, 149, 169, 189 was used as MEGAHIT adopts a multiple k-mer strategy (Li *et al.,* 2015). The quality of the generated assembly was assessed using the Quality Assessment Tool for Genome Assemblies -QUAST version 5.0.2 (Gurevich *et al.,* 2013; Mikheenko *et al.,* 2016, Mikheenko *et al.,* 2018).

3.3.5 Taxonomic profiling of APW whole gut metagenomic data

The MEGAHIT assembled contigs were uploaded onto the Pathosystems Resources Integration Center (PATRIC) workstation (Wattam *et al.,* 2014), and the data were taxonomically classified using Kraken 2 on PATRIC version 3.6.8 with default settings searching against all genomes database, RDP (SSU rRNA) (Maidak *et al.,* 1997), SILVA (SSU rRNA) (Yilmaz *et al.,* 2014), and the output data was visualized by Krona software (Ondov *et al.,* 2011).

3.3.6 Open reading frame prediction, metagenome annotation and functional assignment

The assembled contigs were subjected to open reading frame prediction whereby sequences that contain coding regions (open reading frames) were identified. This was achieved using the prokaryotic genomes gene-prediction software 'Prodigal' v2.6.3 (PROkaryotic DYnamic programming Gene-finding ALgorithm) (Hyatt *et al.,* 2010). Metagenome annotation which entails identifying and labelling the relevant coding and non-coding features and assigning functions to the CDS in a set of sequences was performed using the command line software tool for rapid annotation of prokaryotic genomes "Prokka" v1.14.0. with default E-value threshold of 10^{-6} and series of databases as detailed in Seeman, 2014. FASTA output files of all genomic nucleotide features (.ffn) and translated coding genes (.faa) among others were generated from the ORF input sequences. A (.tsv) file of all features such as the gene ID (locus tag), feature type (ftype), length (bp), gene, EC_number, COG, and product information was also generated (Seeman, 2014). Having identified thousands of genes with predicted functions and EC_numbers, we manually used the EC_numbers of enzymes from the 5 different classes of the CAZy database (Lombard *et al.,* 2014, Levasseur *et al.,* 2013) to filter out members of these classes present in our (.tsv) annotation output file.

3.3.7 Selection of genes of interest

From the members of the lignin active 'Auxiliary Activities' class of the CAZy database found, we focused on selecting full-length genes predicted as potential lignin modifying enzymes (Polyphenol oxidases and peroxidases found in the AA1 and AA2 families of the CAZy database respectively) for PCR amplification, to validate our sequencing and gene prediction analyses. Sequence similarity and conserved domain searches were carried out on NCBI employing the BLASTp (standard protein BLAST) tool against the UniProtKB/Swiss-Prot(swissprot) manually curated database to compare and confirm the identity/ activity of the predicted gene sequences. The UniProtKB/Swiss-Prot(swissprot) database contains all known relevant information about proteins extracted from scientific literature and biocurator-evaluated computational analysis hence is more streamlined to partially characterised proteins (UniProt, 2021). The

results from the searches were analysed using statistical scores based on the E-value, the query coverage, the alignment quality and percent identity, and similarity in conserved domain architecture to existing proteins (Pearson, 2013). These guided our decision to select 3 genes in the first instance that showed significant homology and matches to suggested conserved domains indicating putative dye decolourising peroxidase (2 genes) and polyphenol oxidase (1 gene) functions for PCR validation and subsequent cloning and recombinant expression.

3.3.8 Whole genome amplification (WGA) of metagenomic DNA

Due to the quantity of the sequenced metagenomic DNA sample being small posing the risk of sample inadequacy for PCR amplifications for downstream experiments, the metagenomic DNA was subjected to whole genome amplification (WGA) reactions to increase the amount of DNA template and allow for unlimited PCR amplification attempts using the Repli-G mini kit (Qiagen, UK). The kit employs an isothermal genome amplification methodology, called Multiple Displacement Amplification (MDA), where hexamer oligonucleotides act as random primers for amplification of up to 70 kbp fragments by Phi29 polymerase, a proof-reading polymerase derived from *Bacillus subtilis* phage phi29 (Garmendia *et al.*, 1992; Lasken 2009) thus reducing bias. Approximately 10ng of metagenomic DNA was incubated in the buffers provided for 3 minutes at room temperature to denature the double stranded DNA according to the manufacturer's instructions. The amplification of metagenomic DNA employing the REPLI-g mini kit was carried out for 16 hours in a thermocycler at 30°C in the presence of Phi29 polymerase, at the end of which the polymerase was deactivated by heating at 65°C for 3 minutes. The purity and final concentration of the amplified metagenomic DNA was assessed using a nanodrop spectrophotometer (Nanodrop 2000, Thermo Fisher Scientific, UK). From a 2.5 µl starting material of extracted metagenomic DNA, a 50 µl final volume containing 482.7 ng/ µl of metagenomic DNA was achieved. A 1:20 dilution of the amplified metagenomic DNA was performed in TE buffer and was assessed to have a concentration of 23.6 ng/ µl and a purity value of 1.82 based on the ($A_{260}/_{280}$) ratio. This diluted DNA was used as template DNA for the routine PCR amplifications of selected genes.

3.3.9 Primer design and PCR amplification of selected genes from whole genome amplified DNA

Forward and reverse gene specific primers were designed to selectively amplify the selected genes of interest from the whole genome amplified metagenomic DNA template. Guidelines for designing primers for amplification of genes to be cloned into our vector of choice (Champion pET151/D-TOPO) were adhered to by adding the CACC overhang sequence to the 5' end of the forward primer and ensuring that the reverse primer had a stop codon and was not complementary to the GTCC (reverse complement of the added CACC) overhang sequence at the 5'-end in order to enable directional cloning with maximum efficiency (Champion pettopo user manual, MAN0000214). The primers were ordered from Eurofins genomics, Germany. One hundred (100) pmol/µl of stock primer solutions were reconstituted in TE buffer as instructed in the oligonucleotide synthesis report and a 1:10 diluted working solution (10 pmol/µl) was prepared. Annealing temperatures for each gene were estimated using the Tm Calculator offered via the NEB interactive tools for Q5 High-Fidelity 2X master mix polymerase. A summary of the selected gene IDs, designed primer sequences, putative functions, expected sizes of PCR products, and calculated annealing temperatures are presented in Table 3.3

Table 3.3 The primer sets designed for amplification of the 3 selected putative lignin degrading genes

Gene ID number/ Primer name	Primer sequence (5'- 3')	Putative function	Expected size of PCR product	Annealing temperature
A-30342/F	CACCATGAGTAAGCTAATTG	Polyphenol oxidase	721bp	59°C
A-30342/R	TTAACTTGCCATACGACC			
B-38773/F	CACCATGAACAGCAAGCAACAGGGA	Deferrochelatase/ peroxidase EfeB	1285bp	72°C
B-38773/R	TTACAATGCCCTGGCTGCGGAAATC			
C-08687/F	CACCATGTCTCAGGTTCAGAG	deferrochelatase/ peroxidase YfeX	895bp	66°C
C-08687/R	TTAGATACGCTCCAGCGACG			

In order to validate the bioinformatic annotations and certify that the selected gene sequences from the metagenomic annotation output file occur naturally within the gut metagenome of our study insect, the above designed primer sets were used to amplify the targeted genes using the whole genome amplified DNA as template in a PCR. The reaction was set up according to the NEB protocol for Q5® High Fidelity 2X Master Mix kit which uses a Q5 high fidelity polymerase with a 3' – 5' exonuclease activity and over 280 times higher fidelity than *Taq* hence achieving robust and error proof amplification (New England Biolabs, UK). Also, the Q5 polymerase was chosen because it generates blunt ended PCR products which are required for cloning into our vector of choice. Modifications were made to the protocol based on the size and annealing temperature of each gene sequence which impacts the annealing and extension conditions of the reaction. The reaction mixture and thermocycling conditions for PCR amplification of the selected genes are shown on Tables 3.4 and 3.5 respectively.

Table 3.4 The PCR reaction mixture for amplification of selected genes

Component	Amount (µL)
Q5 2X Master Mix	25
Forward Primer (10pM)	1.25
Reverse Primer (10pM)	1.25
Template DNA	~100ng
Nuclease-free water	Up to 50
Total volume	50

Table 3.4 Thermocycling conditions for amplification of selected genes

Stage	Temperature (°C)	Time (seconds)	Cycles
Initial denaturation	98	30	1
Denaturation	98	10	35
Annealing A B C	 59 72 66	 20 35 20	
Elongation (A and C) B	 72 72	 30 60	
Final extension	72	240	1
Hold	4		∞

To check the amplification success and integrity of the PCR products, 1% agarose gel supplemented with SYBR™ Safe DNA Gel Stain (Fisher Scientific, UK) was prepared, samples were mixed with appropriate amounts of 6X purple loading dye and loaded unto the gels alongside a 1Kb DNA ladder (Lee *et al.,* 2012). The gel was run at a constant volt of 100V in an electrophoretic tank for about an hour and images were captured using the G: Box (Syngene).

3.4 Results

3.4.1 Metagenomic DNA quality

Bacterial metagenomic DNA extracted from whole guts of APW larvae was run on a 1% agarose gel alongside an extraction control to assess its size and integrity (Figure 3.2).

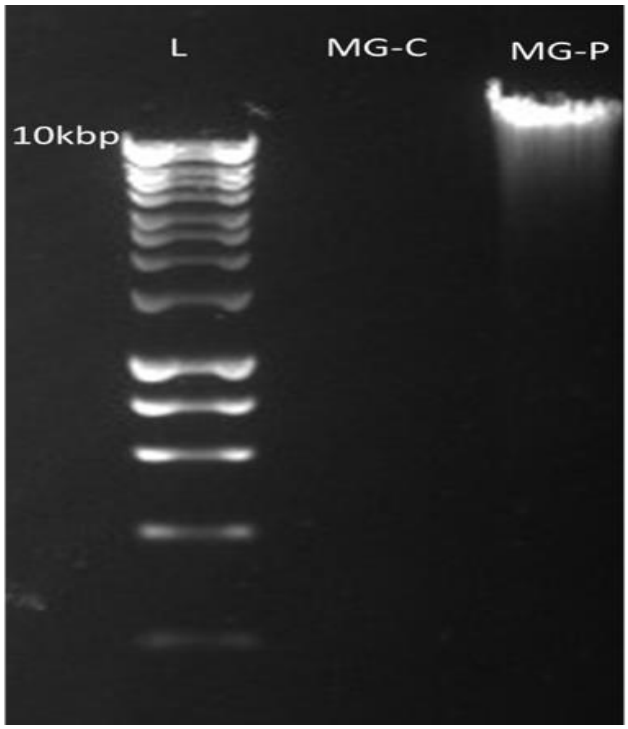


Figure 3.2 1% agarose gel image of bacterial metagenomic DNA samples from the APW gut tissue

In lane L, 2µl of 1kb DNA ladder was loaded as DNA marker. In lanes MG-C and MG-P, 6µl (5 µl sample +1 µl 6X purple loading dye) of the extraction control (MG-C) and metagenomic DNA pool (MG-P) samples were loaded.

From the gel, a thick band of size >10kbp with little smearing was observed in the lane containing the pooled metagenomic DNA sample (MG-P) indicating that the extracted DNA was of high molecular weight but only slightly degraded. There was no band appearance in the well containing extraction control sample (MG-C) as the DNA concentration was very low (0.004ng/µl on nanodrop and undetected on Qubit), indicating the absence of contamination during the extraction process. The determined concentration of the MG-P sample was 72 ng/µl and purity($A_{260/280}$) of 1.92 on nanodrop, and a concentration of 3.93 ng/µl on Qubit, suitable for shotgun library preparation.

3.4.2 Shotgun metagenomic sequencing data statistics

The bacterial metagenomic DNA sample shown in figure 3.2 was used to create a 2x250bp paired-end shotgun metagenomic DNA library that was sequenced using an Illumina® Miseq V2 kit. A summarised statistic of the raw sequencing data is presented in Table 3.6

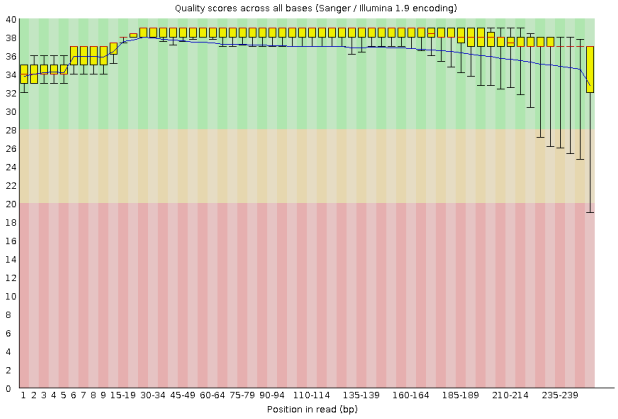
Table 3.5 Shotgun sequencing data statistics of the *R. phoenicis* gut metagenome

Sequencing statistics	Raw data
Raw sequence data (Gbp)	2.89
Total number of reads	11,725,946
Sequence length distribution(bp)	35-251 (Majority at 251)
Mean GC content (%)	47

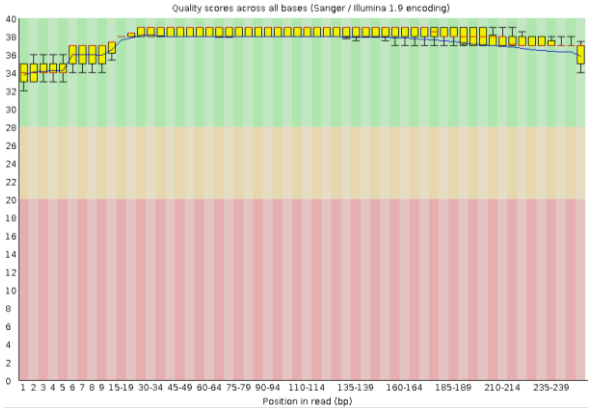
The information contained in the table above indicates that the sequencing run was successful. Over 11.7 million reads were generated from 2.89 Gbp of data. A GC content of 47% also implies that the data was not skewed in favour of genomes with extremely high or low GC contents (FastQC manual).

3.4.3 Quality assessment and quality control of sequence data

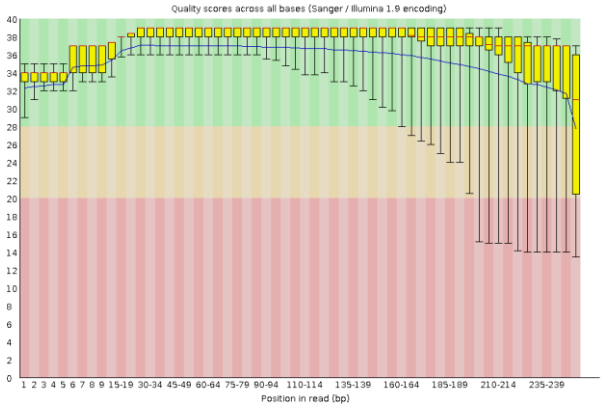
The quality of the data was visually assessed by running a FastQC analysis before and after quality trimming. Quality trimming of the data was performed using Trim galore as earlier described. The results showing the 'per base' sequence quality before and after trimming for both forward and reverse reads are presented below (Figure 3.3)



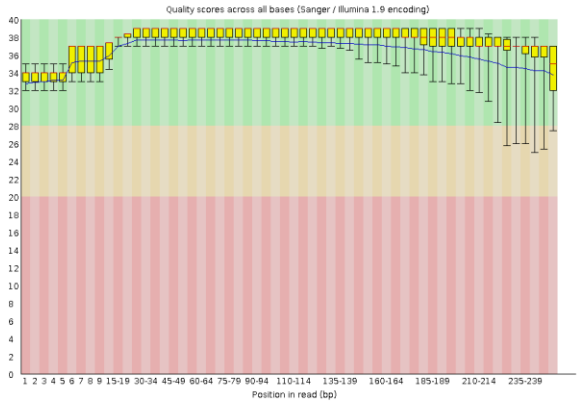
R1 reads before trimming



R1 reads after trimming



R2 reads before trimming



R2 reads after trimming

Figure 3.3 Quality profile of sequence read pairs before and after quality trimming.

Box plots showing the phred quality score (Q) distribution (on the Y-axis) at different base positions (X-axis) along the sequence reads in the forward (a) and reverse (b) reads before and after quality trimming. The yellow boxes, whiskers, blue and red lines represent the 25-75th percentile, 10-90th percentile, mean and median of the distribution respectively. The quality regions are partitioned and coloured, Q0-20: Red; Q21-30: Orange; Q31-Q40: Green representing 99%, 99.9% and 99.99% accuracy of base calling. Ideally, a mean phred quality score above Q30 is considered ideal (Andrew, 2010)

In the forward reads, the plot shows upper and lower values (denoted by whiskers) of Q39 and Q19 with the majority of base calls above Q20, and only a few towards the end (245-250) falling below Q20. Post trimming, the quality of reads improved. The least quality was seen now at around Q34 and all bases were distributed only in the green region (above Q30, corresponding to 99.9% accuracy) indicating very good

quality base calls, evidencing the effectiveness of the trimming exercise in cutting off poor quality reads. Mean sequence quality scores increased from Q37 to Q38 after trimming.

For the reverse reads, the quality before trimming was much worse than that of the forward reads as the plot shows upper and lower values of Q39 and Q14 having bases between positions 190-250 falling below Q20. This situation of having poorer base calls in the reverse reads and especially towards the end is typical with Illumina sequencing as continuous exposure to laser light at each cycle begins to damage the DNA strands hence increasing chances of incorrect base calling (McElhoe *et al.,* 2014). Trimming also improved the quality of the reverse reads. Quality scores went up now ranging between Q39-26, with majority of base calls (positions1-220) in the green region above Q30. Mean sequence quality scores for the reverse reads also increased from Q32 to Q36 after trimming. The results show that both forward and reverse reads had mean quality scores above Q30 which is the ideal quality score required for further analyses. In total, about 342,000 reads were lost to quality trimming leaving behind about 11.4 million high quality reads with size of 2.71Gbp. The reads were adjudged to have passed the quality control processing (Q score > 30) and were carried on to sequence assembly (Andrew, 2010; FastQC manual).

3.4.4 Data sequence assembly

The quality statistics of the metagenome assembly performed by MEGAHIT and assessed using QUAST is presented in Table (3.7).

Table 3.6 Summary of basic sequence assembly statistics

Parameter	Metric
Number of input reads (reads post QC)	11,383640
Total size of assembled metagenome (Kbp)	49,774
Total number of Contigs (≥ 0 bp)	79,690
Total number of Contigs (≥ 500 bp)	26,995
Total number of bases in assembly	28,319,722
Largest Contig length (bp)	244,560
N50	1,063
L50	6,057
N75	674
L75	14,647
GC (%)	57.05

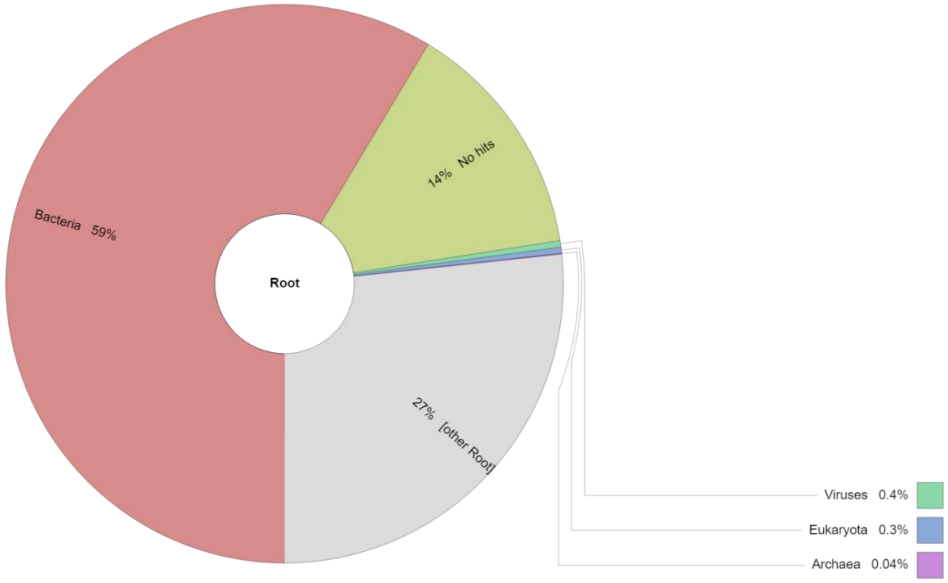
All statistics are based on contigs of size ≥ 500 bp, unless otherwise noted.
N50 is the length for which the collection of all contigs of that length or longer covers at least half (50%) the total base content of the assembly**.**
N75 is defined similarly to N50 but with 75% instead of 50%.
L50 (L75) is the number of contigs equal to or longer than N50 (N75).
GC (%) is the total number of G and C nucleotides in the assembly, divided by the total length of the assembly.

The total size of the assembled metagenome was 49,774 Kbp with the largest contig of length 244,560 bp. The generated statistics indicate the assembly was of good quality having produced 79,690 contigs out of which about one-third (26,995) are longer or equal to 500bp. Six thousand and fifty-seven (6,057bp) contigs of length 1,063bp or longer made up at least 50% of the assembled reads and a GC content of 57% represents a balanced GC content which is not skewed, eliminating the risk of generating misassembled reads (Lapidus and Korobeynikov, 2021; Chen *et al.,* 2013).

3.4.5 Taxonomic profiling of APW whole gut shotgun metagenomic data

The taxonomic profile of the APW gut metagenome as identified by analysis of shotgun metagenomic data of the whole gut using Kraken 2 on PATRIC can be seen in Figure 3.4.

A.



B.

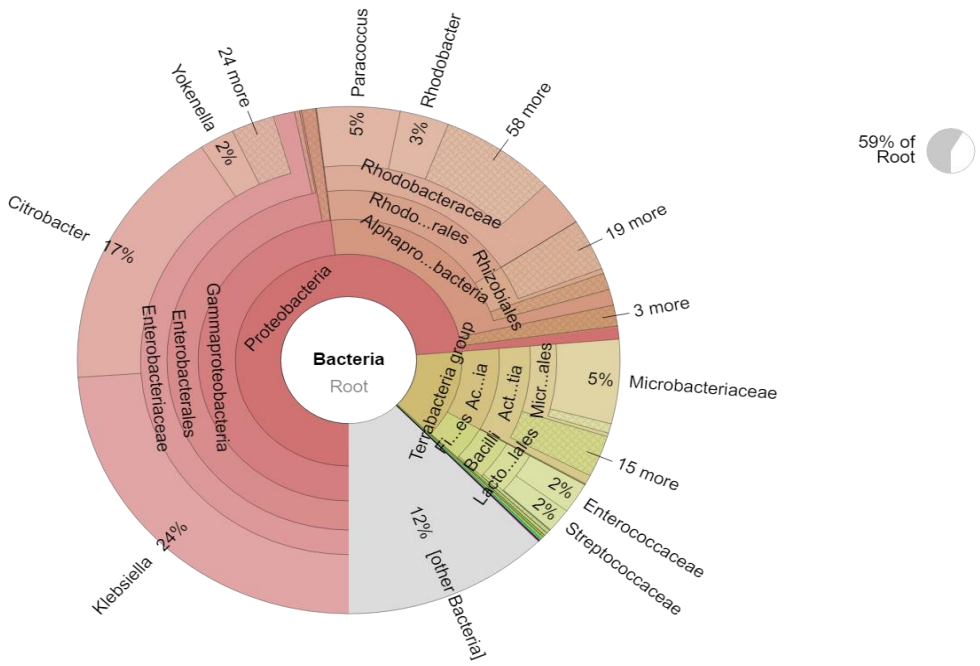


Figure 3.4 Taxonomic classification of the APW gut bacterial metagenome.
Krona plots showing (A.) Root level and (B.) Genus level taxonomic classification of MEGAHIT assembled contigs of APW gut bacterial metagenome data analyzed using kraken 2 on PATRIC genome analyzing platform.

The results of taxonomic classification of assembled metagenome revealed that 14% of the sequences were unclassified having no hits to any sequences in the reference database, 27% belonged to other roots which are not of cellular organisms while the remaining 59% of the assigned hits were rooted to cellular organism with bacteria comprising about 58%, viruses, eukaryota, and archaea constituting less than 1% (Figure 3.4A).

From Figure 3.4B, it is obvious that the APW gut metagenome harboured bacteria predominantly (about 58%), and the most dominant bacterial taxa were mainly from the phylum *Proteobacteria* that represented 74% of all identified bacteria and 43% of the entire root. Within the *Proteobacteria* phylum, *Klebsiella* (24%), *Citrobacter* (17%) and *Yokenella* (2%) were the predominant genera which belong in the *Gammaproteobacteria* class while *Paracoccus* and *Rhodobacter* were the most abundant genera in *Alphaproteobacteria* class. *Actinobacteria* (9% of bacteria and 6% root) and *Firmicutes* (4% bacteria and 2% root) were the next most abundant phyla belonging to the *Terrabacteria* group. Listed in order of prominence are the genera *Microbacteria, Streptomyces*, *Leucobacter* and *Nocardia* (belonging to *Actinobacteria*), and then *Enterococcus, Lactococcus*, *Bacillus*, *Erysipelothrix* and *Paenibacillus* (Belonging to *Firmicutes*) respectively. Other unclassified bacteria made up 12% of the hits. Zooming out on the krona plot and viewing at species level (species level classification not captured on krona plots), the dominant species within the APW gut metagenome across all phyla were *Klebsiella pneumoniae*, *Klebsiella quasipneomoniae*, *Klebsiella aerogenes*, *Klebsiella variicola*, *Citrobacter koseri*, *Citrobacter freundii*, *Yokenella regensburgei*.

Other bacterial genera of interest (lignin degradation associated bacteria as listed on table 1.4) which have not been captured in this Krona plots due to having very low abundances (<1%) were also identified. These include *Pseudomonas* (0.24%), *Escherichia* (0.21%), *Ochrobactrum* (0.16%), *Brucella* (0.10%), *Sphingomonas* (0.10%), *Corynebacterium* (0.08%), *Rhodococcus* (0.07%), *Clostridia* (0.06%), *Raoultella* (0.06%), *Variovorax* (0.06%), *Sphingobium* (0.04%), Novosphingobium (0.04%), Rhizobium (0.30%), Pantoea (0.02%), Serratia (0.02%), *Xanthomonas*

(0.02%), *Pandoraea* (0.02%), *Delftia* (0.01%), *Acinetobacter* (0.01%), *Aeromonas* (0.01%), *Shigella* (0.01%), *Amycolatopsis* (0.01%).

3.4.5 ORF prediction, annotation, and functional assignment

In total, 60,615 ORFs were predicted from the assembled contigs and annotated. A summary of the gene prediction statistics is presented in table 3.8.

Table 3.7 Summary of ORF prediction and functional annotation output

ORFs identified	Number
Total	60615
rRNAs	139
tmRNA	8
tRNAs	663
Hypothetical	43,913
Putative functions assigned	15,892
EC number assigned	8,275
COG assignment	12,020

The gut bacterial metagenome of the APW larvae contained 60,615 protein coding sequences (ORFs) in total, of which 15,892 had a functional prediction and 43,913 were labelled ‘hypothetical’ while RNA genes were 810. From the number of genes with functions predicted, 8,275 sequences had predicted enzyme classifications with assigned EC numbers and 12,020 had COG assignments. One thousand, one hundred and forty-one (1141) genes predicted to belong to enzymes in the CAZy database were filtered out by using the EC numbers of the enzymes as recorded in the CAZy database to compare against the genes with predicted EC numbers from the functional annotation output file. The number of genes found for each identified CAZy class, with details of each family for members of the AA class (lignin active class) are presented in table 3.9.

Table 3.8 CAZymes identified within the APW gut bacterial metagenome

CAZy class	Description of activity/ EC number	Total genes
GH	Hydrolysis and/or rearrangement of glycosidic bonds (3.2.1)	191
GT	Formation of glycosidic bonds (2.4.-)	276
PL	Non-hydrolytic cleavage of glycosidic bonds (4.2.2)	32
CE	Hydrolysis of carbohydrate esters (3.1.1; 2.3.1; 3.5.1)	423
AA	Redox enzymes that act in conjunction with CAZymes	219
AA Family identified		
AA1	Laccase / p-diphenol: oxygen oxidoreductase / ferroxidase (EC 1.10.3.2); ferroxidase (EC 1.10.3.-); Laccase-like multicopper oxidase (EC 1.10.3.-)	30
AA2	Manganese peroxidase (EC 1.11.1.13); versatile peroxidase (EC 1.11.1.16); lignin peroxidase (EC 1.11.1.14); peroxidase (EC 1.11.1.-), Dye decolourising peroxidase (EC 1.11.1.19)	61
AA3	cellobiose dehydrogenase (EC 1.1.99.18); glucose 1-oxidase (EC 1.1.3.4); aryl alcohol oxidase (EC 1.1.3.7); alcohol oxidase (EC 1.1.3.13); pyranose oxidase (EC 1.1.3.10)	55
AA5	Oxidase with oxygen as acceptor (EC 1.1.3.-); galactose oxidase (EC 1.1.3.9); glyoxal oxidase (EC 1.2.3.15); alcohol oxidase (EC 1.1.3.13)	
AA7	Glucooligosaccharide oxidase (EC 1.1.3.-); chitooligosaccharide oxidase (EC 1.1.3.-)	
AA9	copper-dependent lytic polysaccharide monooxygenases (LPMOs)	103
AA10	copper-dependent lytic polysaccharide monooxygenases (LPMOs)	

Our findings reveal an abundance of 1,390 carbohydrate active proteins comprising 1141 genes corresponding to enzymes directly involved with carbohydrate metabolism (GH, GT, CE, and PL), and 249 from groups associated with lignin breakdown (AA- auxiliary activities class members).

3.4.6 Result of BLASTp search of selected genes

Results of the BLASTp query and conserved domain search using amino acid sequences of the selected genes of interest against the UniProtKB/Swiss-Prot(swissprot) database are shown below (Figures 3.5, 3.6 and 3.7). Genes with ORF IDs 30342, 38773 and 08687 were selected based on sequence homology and conserved domain similarity to characterised polyphenol oxidases and DyP-type peroxidases.

A protein BLAST of gene A-30342 as query against the UniProt database yielded 33 sequences similar to purine nucleoside phosphorylases. We observed 100% sequence coverage, excellent alignment scores of >= 200, high percentage identity (90.12% and 89.71%) and low E-values (1e-167 and 2e-166) indicating biologically significant alignment which occurred in nature and not by chance as seen with the top two hits.

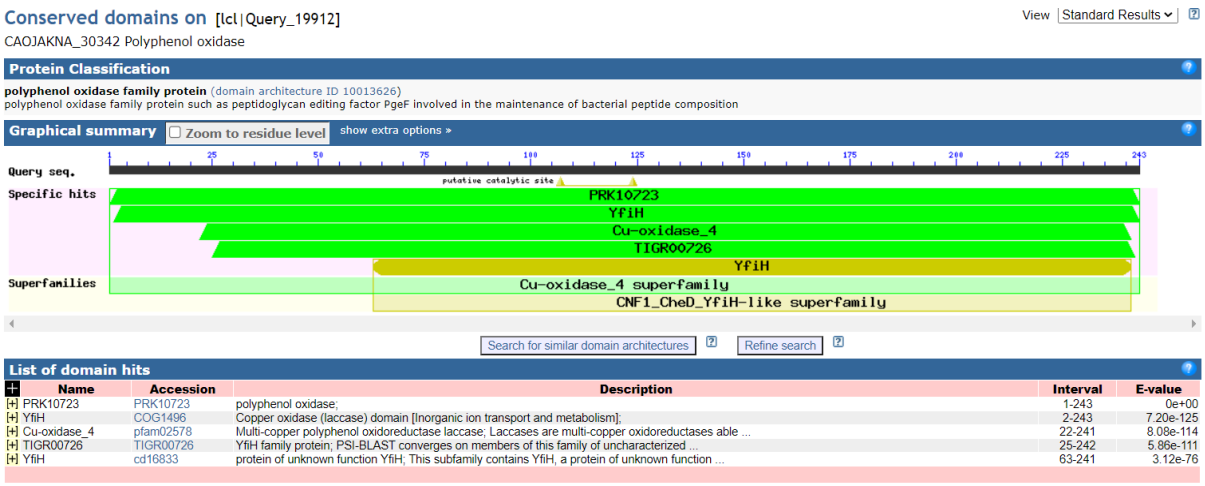


Figure 3.5 BLASTp search results for gene A-30342.
BLASTp search output for gene A-30342 amino acid sequences aligned against the UniProtKb/Swiss-Prot database showing the conserved domain architecture with specific hits to the PRK10723, YfiH, Cu-oxidase_4 and TIGR domains and Cu-oxidase_4 superfamily of proteins.

Conserved domain search revealed best hits to Polyphenol oxidase, multicopper/ copper oxidase (laccase) domains. A putative catalytic site was also observed involving residues at positions between 100-125 from the domain architecture. We therefore selected this gene to verify its potential polyphenol oxidase (laccase) activity as predicted and confirmed by sequence homology and conserved domain search.

A protein BLAST of gene B-38773 as query against the uniport database yielded 12 sequences with 11 similar to deferrochelatase/ peroxidases and one to dye decolourising peroxidase. We observed 96% sequence coverage, excellent alignment scores of >= 200, moderate percentage identity: between 33-55% for deferrochalatases and 31% for the dye decolourising peroxidase hits.

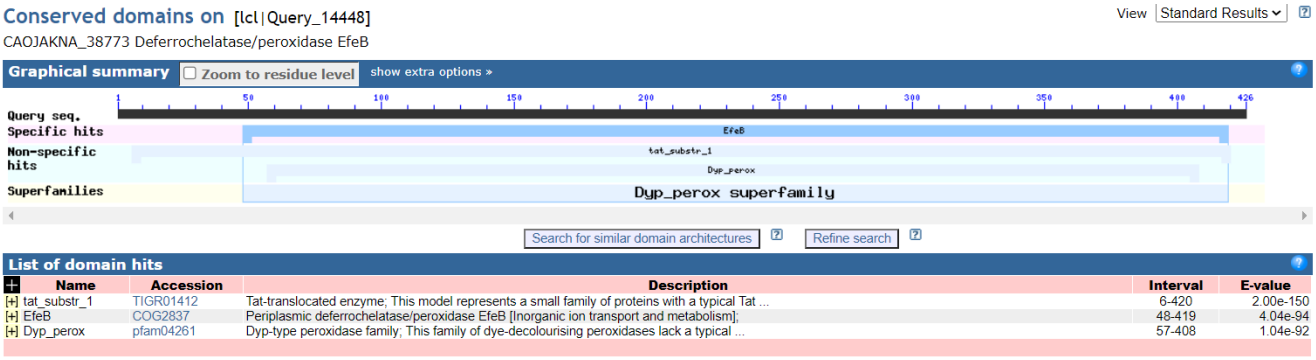


Figure 3.6 BLASTp search results for of gene B-38773.
BLASTp search output for gene B-38773 amino acid sequences aligned against the UniProtKb/Swiss-Prot database showing the conserved domain architecture with specific hits to EfeB, nonspecific hits to tat_subsr_1 domains and Dyp_peroxidase superfamily of proteins.

Conserved domain search revealed similarity to periplasmic deferrochelatase/ peroxidase and DyP-type (dye-decolourizing) peroxidase family. We selected this gene to verify its potential dye-decolourizing peroxidase activity as predicted and confirmed by sequence homology and conserved domain search.

A protein BLAST of gene C-08687 as query against the uniport database yielded 3 sequences with the top 2 hits similar to dye decolourising peroxidase and 1 to deferrochelatase/ peroxidases. We observed 100% sequence coverage, excellent alignment scores of >= 200 and 88% similarity with the dye decolourising peroxidases, and 52% coverage, poor alignment score <40, 25% similarity with the deferrochelatase/peroxidase hits.

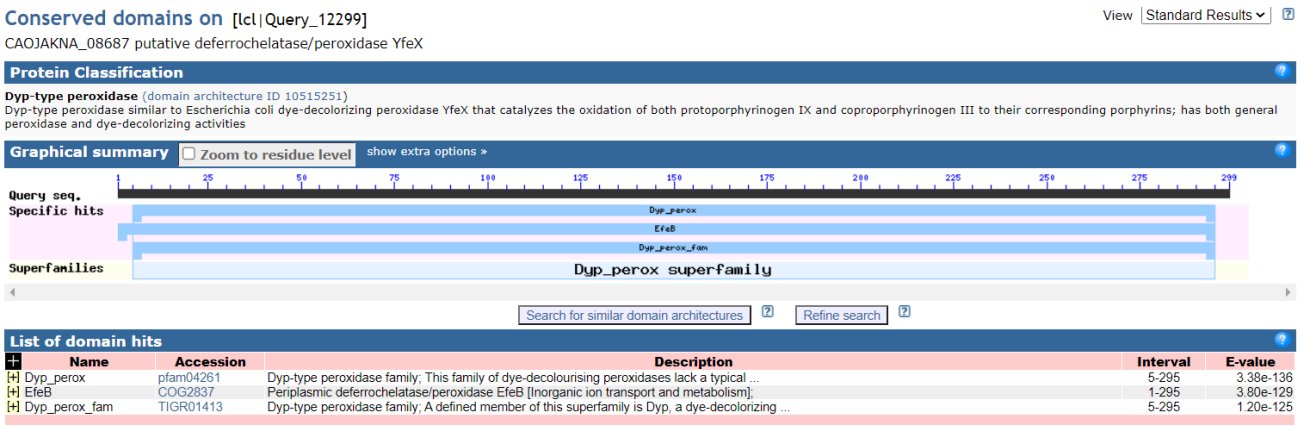


Figure 3.7 BlastP search results for gene C-08687.
BLASTp search output for gene C-08687 amino acid sequences aligned against the UniProtKb/Swiss-Prot database showing the conserved domain architecture with specific hits Dyp_perox and EfeB domains.

From figure 3.7, conserved domain search revealed similarity to periplasmic deferrochelatase/ peroxidase and Dyp-type (dye-decolourizing) peroxidase family. We selected this gene to verify its potential predicted Dyp-type (dye-decolourizing) peroxidase activity as predicted and confirmed by sequence homology and conserved domain search.

A summary of the selected genes and their characteristics as predicted by functional annotations are presented in table 3.10.

Table 3.10 List of selected genes with their predicted characteristics and putative functions

Gene ID	Nucleotide sequence Length (bp)	Amino acid sequence Length	CAZyme family	EC-number	Putative function
A-30342	732	243	AA1	1.10.3.-	Polyphenol oxidase
B-38773	1281	426	AA2	1.11.1.-	Deferrochelatase/ peroxidase EfeB
C-08687	900	299	AA2	1.11.1.-	Deferrochelatase/ peroxidase

3.4.7 Whole genome amplification (WGA) of metagenomic DNA

In order to obtain enough metagenomic DNA for a large number of PCR reactions, metagenomic DNA was subjected to whole genome amplification reactions. Figure 3.8. shows the agarose gel of the whole genome amplified DNA.

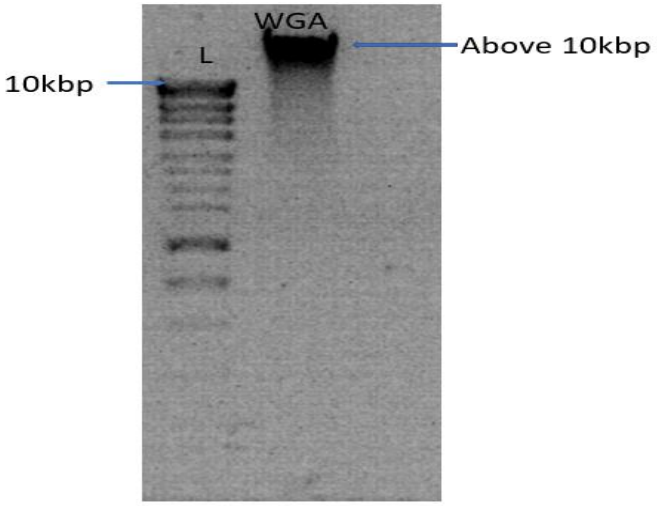


Figure 3.8 Agarose gel image of Whole genome amplified DNA from metagenomic DNA template

Figure 3.8 shows 1 µL of Whole genome amplified DNA (Lane WGA) created from approximately 10ng of the original metagenomic DNA as template, alongside 2.5 µL of 1kb DNA ladder (Bioline UK) loaded in lane L.

The gel shows successful amplification of DNA of length comparable to the original metagenomic DNA using the WGA method as seen by band showing a high molecular weight DNA of size above 10 kbp.

3.4.8 Results for PCR amplification of selected genes

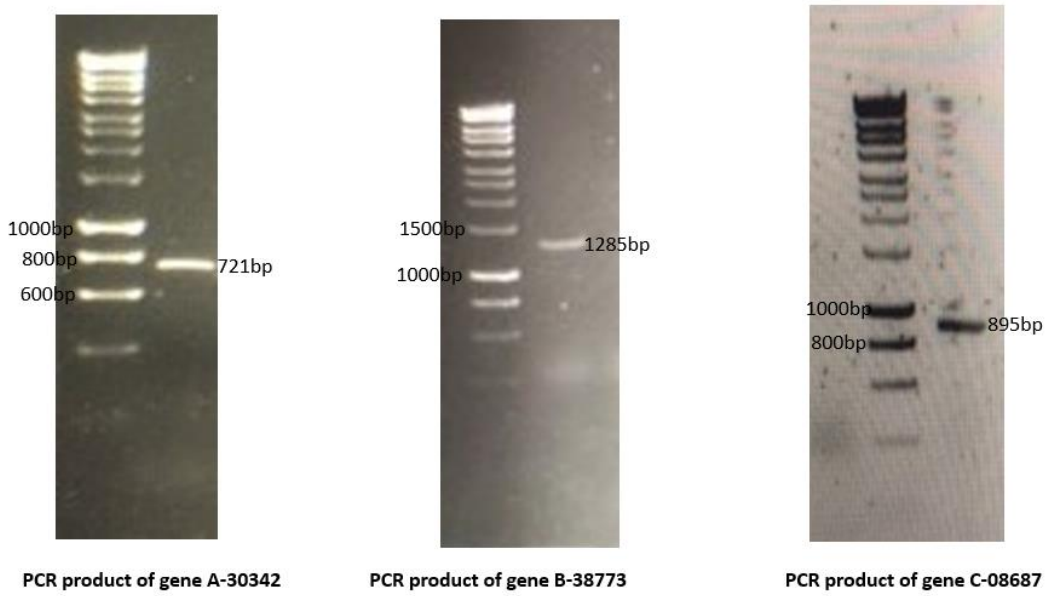


Figure 3.9 Agarose gels of PCR products of selected genes of interest.
DNA bands of PCR products of genes A-30342, B-38773, and C-08687 run on 1% agarose gels alongside 1kb DNA ladder.

All three (3) selected genes were successfully amplified from the WGA DNA template as predicted from the functional annotation of the APW gut metagenome. Bands corresponding to the expected sizes for each gene (A-30342; 721 bp, B-38773; 1285 bp, and C-08687; 895 bp) were seen validating the gene prediction metagenome analysis.

3.5 Discussion

Microorganisms are the oldest life forms and the richest source of genetic diversity on earth hence they have been exploited for identification and isolation of potent and novel biocatalysts of industrial relevance using metagenomics (Madhavan *et al.,* 2017; Brune *et al.,* 2014). A function or sequence based metagenomic approach can be employed in screening biomolecules from microbial environments (Ferrer *et al.,* 2009; Madhavan *et al.,* 2017; Knight *et al.,* 2018). Several studies have demonstrated the application of sequence-based metagenomic techniques in profiling the metabolic capacity and mining various genes encoding functions of interest from different environments e.g. glycoside hydrolases from slug gut metagenome (Joynson *et al.,* 2017), biomass degrading genes from cow rumen (Hess *et al.,* 2011), long horned beetle (Scully *et al.,* 2013), and termites (Do *et al.,* 2014), lipases from pond water (Ranjan *et al.,* 2005), chitinases from tomato moth (Fitches *et al.,* 2005), antibiotic resistance genes from arable field (Courtois *et al.,* 2003), dietary fibre catabolic genes from human gut microbiome (Tasse *et al.,* 2010).

Following the identification of abundant and diverse bacteria associated with lignin degradation in the gut of APW in the preceding chapter of this study, we adopted a sequence-based functional metagenomic approach, in conjunction with bioinformatics to identify and isolate ligninolytic genes of bacterial origin from the APW gut metagenome.

In this chapter, we detail the processes we employed and our findings; from DNA extraction and quality control, assembly of metagenome sequences, ORF prediction, functional annotation, and validation of bioinformatics by PCR amplification of selected genes. When extracting DNA for whole metagenomic studies, care must be taken to ensure the extracted DNA is representative of all the microorganisms within the community, is of high molecular weight and undegraded, and has high purity without contamination (Thomas *et al.,* 2012; Egert *et al.,* 2003) in order to achieve good quality libraries, adequate cluster density generation and the sequencing by synthesis process on the flow cell employed by Illumina platforms (Illumina website). Our DNA extraction strategy ensured that the DNA was representative enough of the microbial population by pooling DNA from a total of 15 guts to make a composite sample. DNA was of high

molecular weight and intact (above 10kb with only a little smear) and was of high purity without contamination (Figure 3.2). The extracted metagenomic DNA was therefore of good quality and quantity. The extraction control sample had almost no DNA indicating that DNA extractions were carefully carried out without contamination, therefore, we didn't consider it necessary to sequence the control sample. Unwanted host DNA which has the potential to overwhelm the metagenomic DNA and complicate bioinformatic analysis (Thomas *et al.,* 2012; Sharpton, 2014) was depleted prior to bacterial DNA extraction with the help of benzonase from the QIAamp microbiome kit as explained in section 2.3.1.1.

NGS technologies (particularly the 454/Roche and the Illumina/Solexa systems) are now the most extensively applied methods for metagenomic sequencing over the past 10 years, shifting away from sanger sequencing technology (Thomas *et al.,* 2012; Head *et al.,* 2014; Quinn *et al.,* 2018). We have determined the whole metagenomic sequences of the bacterial community in the gut of APW larvae to the extent that our sequencing depth covered using Illumina technology. The statistics of the raw data generated (Table 3.6) is an indication that the DNA template and libraries prepared from them were of good quality and sequencing run was a success. However, the total reads generated fell short of what similar sequencing efforts by other researchers have obtained e.g., 43.5 million reads from the silkworm gut metagenome (Chen *et al.,* 2018), 94.6 million reads from whole springtails guts metagenome (Le, 2021), 44.1 million and 58 million from cutworm and grasshopper metagenomes respectively (Shi *et al.,* 2013) using illumine Hiseq sequencers. This could be due to the enrichment strategy we adopted during DNA extraction which was designed to extract only bacterial DNA while eliminating other microbial and host DNA and also the use of the illumine Miseq as against the Hiseq used in the above cited projects. We believe this may impact on our results and thus we include a caveat to say the taxonomic profile as well as the functional annotation may not be the best representation of the APW bacterial gut communities. FastQC analysis of the data before and after trimming further confirms the quality of the sequencing process as shown in figure 3.3 for both the forward and reverse reads. Typically, NGS sequence data are subjected to quality control to yield high quality sequences for efficient downstream analyses (Bragg and

Tyson, 2014). Quality trimming with Trim galore removed about 342,000 sequences (adapter sequences and poor-quality reads) yielding 11.4 million reads of higher quality from an initial 11.7 million raw reads.

We subjected the quality trimmed reads to *de novo* assembly to obtain longer contiguous sequences (contigs) in order to increases the quality of annotation (Vestergaard *et al.,* 2017; Do *et al.,* 2014; Thomas *et al.,* 2012). The longer the contigs, the more information is available for annotation especially via homology search (Wommack *et al.,* 2008). Numerous approaches for computationally reconstructing microbial community composition from a pool of sequence reads have been published, choosing the 'best' is a daunting task and depends largely on the aims of the study (Quince *et al.,* 2017). We leveraged on reports from studies where different assemblers have been compared and we found more reports providing great evidence in support of MEGAHIT assembler's efficiency (Van der Walt *et al.,* 2017; Li *et al.,* 2016, Olson *et al.,* 2019; Lapidus and Korobeynikov, 2021) and so we assembled our data using the MEGAHIT assembler. Metrics such as number of contigs, average or maximum contig size can be misleading in assessing the contiguity of assembled reads, which is the attempt to regenerate one contig per chromosome, because a large number of small contigs are often generated due to sequencing errors or other artifacts. A more significant metric is the N50 which is the minimum contig from the entire set of contigs that make up 50% of the assembly considering contigs of size 500bp and above in most cases (Olson *et al.,* 2019). Although a total of 79,690 contigs were generated, only 26,995 (about one-third) were >= 500bp. Six thousand and fifty-seven (6,057) of these made up 50% of the assembly (L50) with the minimum of them having a length of 1,063bp (N50) as seen on the assembly parameters summary in Table 3.7.

We considered it prudent to validate the taxonomic classification of APW gut microbiome obtained using 16S rRNA gene amplicon sequencing reported in previous chapter by carrying out another taxonomic classification of the gut metagenome using contigs from assembled whole metagenome sequence data generated in this chapter. The taxonomic distribution based on the metagenome assembled contigs (Figure 3.4A and 3.4B) identified a high percentage of hits belonging to the bacterial domain (58%) and less than 1% belonging to Viruses, Eukaryote and Archaea, 14% unclassified

sequences and 27% belonging to other not describable root. However, the identification of less than 1% of total hits belonging to other cellular organisms aside bacteria in the WGS method compared to 100% of classified sequences being of bacterial origin in the 16S rRNA gene analysis method attests to the efficiency of the QIAamp microbiome kit at selectively enriching the extracted DNA for bacterial metagenomes prior to sequencing (section 2.3.1.1). Also, having 41% of the WGS data as unclassified and unidentified is not unexpected, though this appears to be comparably on the high side to findings from other studies (Le, 2021; Chen *et al.,* 2018; Jovel *et al.,* 2016; Mhuantong *et al.,* 2015). This observation could be explained to be the consequence of the difference in fundamental principle between the sequencing methods employed where the 16S rRNA gene method specifically targets the bacterial SSU-rRNA gene producing an amplicon via PCR amplification contrary to the WGS method where randomly-sheared fragments of DNA are sequenced and the resulting reads assembly poses a great challenge and could result in mis assemblies and artefacts of false prediction or noise sequences (Ranjan *et al.,* 2016; Jovel *et al.*, 2016; Rosnow *et al.,* 2017). Also, the differences in gut dissection method (gut segments used for 16S rRNA gene sequencing and whole guts used for WGS) and employed in extracting DNA for sequencing could be a contributing factor.

However, we observed similar bacterial taxa in the APW gut metagenome regardless of the difference in sequencing method employed. *Proteobacteria, Firmicutes,* and *Actinobacteria* were the predominant phyla with *Proteobacteria* (46%) being most abundant for WGS data and Firmicutes (63%) for 16S rRNA gene sequencing data. The most abundant genera that are considered to constitute the core bacterial community of the larval guts of the APW as identified from both sequencing methods include *Klebsiella*, *Enterobacter*, *Citrobacter*, *Enterococcus*, *Lactococcus* and *Yokenella* in varying abundances. The prevalence of these genera are consistent with those from the metagenomes of other herbivorous organisms that depend on their microbiomes to degrade the lignocellulose present in their feedstuff such as woodcutter ants (Auer *et al.,* 2017), slugs (Joynson *et al.,* 2017), snails (Cardoso *et al.,* 2012) and beetles (Franzini *et al.,* 2016; Mohammed *et al.,* 2018; Bozorov *et al.,* 2018). A large number of other bacterial genera were detected in much lower abundances by both

methods, with more diverse taxa and species level identification seen in the WGS based profile. This is because shotgun sequencing generated reads capture all available microbiota in the metagenome thereby providing a potentially more accurate and in-depth characterization and representation of uncultured species (Sharpton *et al.,* 2011, Shi *et al.,* 2013). Overall, the taxonomic profiles of bacteria obtained between 16S rRNA gene sequencing and shotgun metagenomic sequencing are comparably similar. Other factors that could be responsible for observed discrepancies in bacterial diversity and abundance include priming and PCR amplification bias associated with 16S rRNA gene amplicon sequencing method with differences in the amplification efficiency of DNA from different bacterial species, biases caused by chosen pipelines, software, reference databases etc employed for taxonomic assignment possibly leading to misidentifications (Ranjan *et al.,* 2016; Jovel *et al.,* 2016; Do *et al.,* 2018).

Gene prediction determines which metagenomic reads contain coding sequences (see Richardson and Watson, 2013, and Yandell *et al.,* 2012 for detailed reviews). Because of the considerable diversity of genomes in nature compared to those in sequence databases (Wu *et al.,* 2009), not all predicted genes will exhibit homology to known sequences. Some of these predictions may be spurious, while others will represent novel or highly diverged proteins. Thus, gene prediction is critical to the identification of novel genes (Sharpton, 2014). Once identified, coding sequences can be functionally annotated.

From the total 79,690 contigs assembled, 60,615 (76%) ORFs were predicted. Of these reads, only 15,892 (28%) had best alignment scores to annotated proteins, the remaining 43,913 (72%) had highest scoring BLAST alignments to hypothetical or uncharacterized proteins (Table 3.8).

Because annotation is not done *de novo* but via mapping to genes or protein libraries with existing knowledge (i.e., a non-redundant database), ORFs that have been predicted by software, but have no homolog of known function in the databases, and no known functional domain, are referred to as “hypothetical”. They may be real genes coding for proteins with unknown biochemical functions or could code for known proteins but might not have sequence homology with such known proteins in available databases at present, or they may just be artifacts of the gene prediction process

(Richardson and Watson, 2013). Targeting hypothetical proteins for biochemical characterisation studies, though a very difficult process, is largely responsible for the seemingly never-ending genetic novelty in microbial metagenomics driven by the quest to unravel their potential functions and contributions to their inhabitant communities (Ni and Tokuda, 2013; Thomas *et al.,* 2012; Bragg and Tyson, 2014). Beloqui *et al.,* 2006 have characterised two proteins initially designated to be hypothetical proteins from the metagenome library of bovine rumen and found them to be multicopper oxidases with polyphenol oxidase/ laccase activity (Beloqui *et al.,* 2006). The identification of large numbers of these hypothetical proteins is not surprising or abnormal as currently, it is estimated that only 20 to 50% of a metagenomic sequence can be annotated (Thomas *et al.,* 2012). Similar findings have been reported in functional metagenomic analysis where large numbers of reads have been aligned to hypothetical proteins e.g. 21% of total reads annotated from *A. glabripennis* metagenome (Scully *et al.,* 2013), 24% of candidate carbohydrate active genes from cow rumen metagenome (Hess *et al.,* 2011), about 64% of total annotated genes from the sargasso sea (Venter *et al.,* 2004), 1740 out of 5356 predicted proteins from genome of *Pandoraea* sp. ISTKB (Kumar *et al.,* 2018), 13 nickel resistant clones from the rhizosphere microbial community of an acid-mine drainage (AMD)-adapted plant, *Erica andevalensis*, (Gonzalez and Mirete, 2010). For metagenomic projects dedicated to the discovery of specific functions as is the case in our study, it is recommended that ORFs are further annotated using a specialized database for that function (Madhavan *et al.,* 2017; Escobar-Zepeda *et al.,* 2015) e.g CAZy and dbCAN databases for carbohydrate active genes (Cantrel *et al.,* 2009; Kunath *et al.,* 2017; Do *et al.,* 2014; Ransom-Jones *et al.,* 2017; Edwards *et al.,* 2010; Joynson *et al.,* 2017; Hess *et al.,* 2011; Busk *et al.,* 2017; Kanokratana *et al.,* 2013; Rosnow *et al.,* 2017; Cardoso *et al.,* 2012; Jia *et al.,* 2013; Ameri *et al.,* 2018), (CARD), antibiotics resistance genes database (ARDB- no longer maintained) and Resfams databases for Antibiotic resistance genes (Jia *et al.,* 2017; Quince *et al.,* 2017; Gibson, 2015; Liu and Pop, 2009), Fungal oxidative lignin enzymes (FOLy), Lignin degrading enzymes (LD^2L) and eLignin databases for lignin degrading genes (Levasseur *et al.,* 2008; Levasseur *et al.,* 2013; Kameshwar and Qin, 2017a; Arumugam *et al.,* 2014; Brink *et al.,* 2019), and MetaBioMe for commercially useful

enzymes (Sharma *et al.,* 2010), etc. Although we would have preferred to use databases specific for lignin degrading genes such as the LD^2L, FOLy, and eLignin which contain detailed information about lignin modifying and lignin degrading accessory enzymes and the microbes that produce them, these databases are either at very early stages of manual curation and hence not up to date (eLignin) or have been currently discontinued or publicly not accessible (LD^2L and FOLy). We therefore used the CAZy database which specializes in the display and analysis of genomic, structural, and biochemical information on carbohydrate-active enzymes, but also contains the auxiliary activities (AA) lignin-active class and has been employed in functional assignment by other researchers to identify lignin degrading genes (Joynson *et al.,* 2017; Scully *et al.,* 2013).

The identification of an abundance of potential polysaccharide and plant cell wall biomass-degrading enzymes (CAZYmes) indicates the ability of the APW to metabolise all parts of plant biomass. Among the ORFs identified to belong to the CAZy database, we found 1,141 Genes encoding glycoside hydrolases, glycosyl transferases, polysaccharide lyases, and carbohydrate esterases which are active against carbohydrates while 249 genes encoding enzymes that act on or consort with lignin (AA class members) were identified in the microbiomes affiliated with the APW gut. However, results of some metagenomic studies have reported the absence of lignin degrading genes in the midgut and hindguts of termites that are well known plant biomass degraders (Do *et al.,* 2014, He *et al.,* 2013; Warnecke *et al.,* 2007).This may be as a result of these studies targeting mostly the mid and hindgut segments of the termite gut where oxygen requirement for the oxidative reactions that breakdown lignin are limiting or it may suggest that the organisms produce endogenous enzymes themselves that facilitate lignin degradation or may adopt other methods. From the lignin modifying group of enzymes, genes for laccases, multicopper oxidases, polyphenol oxidases, and several peroxidases were present, but none were found for LiP, MnP, and VP, members of the class II superfamily of plant peroxidase (Welinder, 1992) which are well known for their role in lignin degradation. This finding was rather disappointing but not unexpected as other researchers have reported that homologs of genes encoding these enzymes have not been identified from biochemical studies of

bacterial ligninolytic enzymes from sequenced genomes or proteomes (Davis *et al.,* 2013; Brown *et al.,* 2012) and it may seem that these lignin-degrading enzymes are restricted to fungi as they may be difficult to express in bacterial systems due to their complex, heavily glycosylated, multiple disulphide bonds and the presence of several calcium ions and a heme cofactor (de Gonzalo *et al.,* 2016). A number of other extracellular peroxidases that are often highly expressed by lignin degrading microbes during periods of active lignin degradation were however detected. These include iron-dependent peroxidases, thiol peroxidases, catalase-peroxidases, thioredoxin/glutathione peroxidases, and cytochrome c peroxidases. The potential participation of these peroxidases in large-scale lignin degradation is also supported by the detection of a number of peroxide-generating enzymes including aryl alcohol oxidases, FAD oxidoreductases, glyoxal oxidases, and pyranose oxidases.

We screened several polyphenol oxidases (which could have laccase-like activities) and dye decolourising peroxidases to select the most suitable genes for further experiments in line with our study aim which is the identification of lignin degrading enzymes.

In this study, we demonstrate the use of basic BLAST bioinformatics tool, including blast algorithms as described by (Altschul *et al.*, 1997). We exploited the sensitivity and flexibility of BLASTp in the effective confirmation and selection of genes with biologically meaningful homology and likely conserved domains similarity by comparisons between query sequences (genes with predicted functions) and proteins of known functions in the UniProtKb/(swissprot) database (UniProt consortium, 2019) based on inference from alignment scores, query coverage and E-value among other statistical parameters.

The selected candidate genes (Laccase and Dye decolourising peroxidases) have been shown to be capable of disrupting β-aryl ether bonds which are the most dominant linkages in hardwood lignin either directly or in most cases with the help of natural redox mediators. The disruption of these β-aryl ether linkages represents a critical step in lignin degradation (Hatfield and Vermerri, 2001; Ahmad *et al.,* 2011).

Following putative gene identification and selection, the metagenomic DNA was multiplied by whole genome amplification (Figure 3.8) which navigates the problem of

small sample size often seen with environmental samples and overcomes the challenge of having to use expensive gene synthesis methods in ensuring an adequate supply of DNA for downstream experiments and analyses (Czyz *et al.,* 2015; Borgstrom *et al.,* 2017).

Here, the 3 selected gene sequences from the generated metagenomic annotation output files were amplified from the whole genome amplified metagenomic DNA and the integrity assessed by running samples on agarose gels as depicted in figure 3.9. The presence of bands of the predicted size on the agarose gels serves to validate the assembly suggesting that, the predicted sequences do in fact exist in nature.

3.6 Conclusion

Without a doubt, the exploration of metagenomes from natural and active biomass utilising systems such as the guts of wood feeding organisms has proved to be very useful in extending the scope of our understanding of lignin metabolism by host associated microbes, and for bioprospecting novel genes from these microbial habitats.

In this chapter, we used whole metagenome shotgun sequencing and analysis to describe the structure and functional capability of the APW gut with particular emphasis on its lignocellulose/ lignin degrading potential. Taxonomic profiling using WGS data revealed a similar community structure compared to the profile obtained by 16S rRNA sequencing in the previous chapter. A large number of CAZy genes (1,141) were identified out of which 249 belonged to the AA class of lignin degrading enzymes. The identification of these large and diverse sets of genes cutting across the different classes of lignocellulose/lignin degrading enzymes demonstrates that the APW gut is well equipped to breakdown plant matter and contribute to the digestion of woody tissue. In order to verify the quality of sequence assembly and subsequently discover novel lignin degrading enzymes, 3 predicted coding genes for a polyphenol oxidase and 2 dye decolorising peroxidases were successfully amplified from a template of whole genome amplified metagenomic DNA. In the next chapter of this study, we will describe the cloning, recombinant expression, and characterization of these gene products for functional validation of predicted activities.

Chapter 4: Recombinant protein expression, activity testing and characterisation of gene B-38773 construct

4.1 Abstract

The functional characterisation of the whole gut bacterial metagenome of the African palm weevil larvae has revealed a rich reservoir of diverse putative lignin degrading genes cutting across different classes of lignin degrading enzymes. However, these annotations are a result of computational analysis dependent on the information available in databases for identified, and in some cases, characterised proteins. Therefore, in this chapter, we sought to validate the bioinformatics analyses performed by cloning and heterologous expression of the successfully amplified products of the selected genes generated from the functional annotation analysis in the previous chapter. Only gene B-38773 which putatively encodes a dye decolourising peroxidase was successfully cloned as verified by presence of insert containing colonies, restriction enzyme digest and sanger sequencing, and hence was carried forward unto expression in BL21 *E. Coli* cells. The recombinant protein produced was purified by a two-step process employing affinity and ion exchange chromatographic techniques and identified to be approximately 46kDa in size (as predicted by prot-param tool on ExPasy server online) through SDS-PAGE and western blot analysis. Peroxidase, dye decolourising and lignin degrading potentials of the recombinant enzyme were tested by assaying against the peroxidase substrate ABTS, the anthraquinone dye RB19 and Alkali kraft lignin respectively. The enzyme was optimally active at a pH of 4 and temperature of 40°C, displayed Michaelis Menten kinetics with an estimated K_m value of 1.089mM, and V_{max} of 3.68 µMol/min when tested against ABTS. Other kinetic parameters K_{cat} and catalytic efficiency were calculated and values of $540.9S^{-1}$ and of 4.96×10^5 $M^{-1}S^{-1}$ respectively were obtained while specific activity of the B-38773 protein was $12.9Umg^{-1}$. On dye decolourising efficiency, B-38773 mildly decolourised RB19 dye (up to 15.9% in 10 minutes). However, no activity was observed with alkali kraft lignin in order to show direct evidence of lignin degradation potential. Therefore, in this chapter, we successfully validated the whole metagenome sequencing and functional annotation analyses by cloning and expressing gene B-38773 obtained from

the APW gut metagenome as a recombinant protein and confirmed that its function matched what was predicted (a dye decolourising peroxidase) as it efficiently oxidized the peroxidase substrate ABTS and decolourised RB19 dye.

4.2 Introduction

The increased interest in the microbial valorisation of lignin, being a renewable source of aromatic chemicals has led to the identification and heterologous expression of many genes encoding enzymes that are involved in lignin degradation resulting from metagenomic exploration of natural eco-systems where lignin degradation is known to occur (Sahinkaya *et al.,* 2019; Silva *et al.,* 2018; Chen and Wan, 2017; Chen *et al.,* 2015; Ameri *et al.,* 2018; Ali *et al.,* 2019; Bugg *et al.,* 2020; Munoz-Benavent, 2021; Robinson *et al.,* 2021; Arnau *et al.,* 2020).

Generally, multicopper dependent laccases and heme containing peroxidases (classified as lignin degrading/ modifying enzymes-LMEs) are the major groups of enzymes that have been identified as capable of oxidatively degrading the recalcitrant bonds holding the lignin molecule together with the help of a variety of accessory enzymes (classified as lignin degrading accessory enzymes-LDAs) such as glyoxal oxidase, aryl-alcohol oxidase, cellobiose dehydrogenase, quinone oxidoreductase, glutathione dependent etherases etc that play supporting roles necessary for the complete deconstruction of lignin (Rashid and Bugg, 2021; Bugg *et al.,* 2020; de Gonzalo *et al.,* 2016; Janusz *et al.,* 2017; Xie *et al.,* 2014; Datta *et al.,* 2017). Current knowledge on the model of microbial lignin degradation involves the oxidative combustion of lignin mediated by a range of small molecular weight compounds rather than a direct degradation by the enzymes themselves (Cagide and Castro-Sowinski 2020; Chan *et al.,* 2020; Asina *et al.,* 2017). These diffusible mediators make their way into the enzyme's active site where they are oxidized into more stable, high redox intermediates which can penetrate and react directly with lignin to generate radical sites within the substrate and trigger a cascade of bond breaking reactions that ultimately leads to lignin's decomposition into smaller aromatic compounds, CO_2, and water (Cagide and Castro-Sowinski 2020; de Gonzalo *et al.,* 2016; Brown and Chang, 2014; Weiss *et al.,* 2020). Although direct oxidation of substrates without the need for

mediators have also been recently reported to be possible (Choolaei *et al.,* 2020; Perna *et al.,* 2020; Vuong *et al.,* 2021).

The mechanism of lignin degradation by laccases is facilitated by the presence of four copper atoms located in three distinct binding sites within their active centre which are critical for the catalytic activity of laccases to oxidise a wide variety of phenolic and non-phenolic compounds to their corresponding radical species with the concomitant reduction of atmospheric oxygen which serves as an electron acceptor, to water (Janusz *et al.,* 2017, Choolaei *et al.,* 2021; Mayr *et al.,* 2021; Chauhan *et al.,* 2017; Li *et al.,* 2009; Fisher and Fong, 2014; Datta *et al.,* 2017; Neeraas, 2019). On the other hand, peroxidases, most of which contain heme molecules as the essential prosthetic group for activity catalyse the oxidation of a wide range of substrates using hydrogen peroxide (H_2O_2) as an electron acceptor with concomitant reduction of H_2O_2 to water (Xu *et al.,* 2021; Liu *et al.,* 2017; Qin *et al.,* 2018). The strategy adopted by ligninolytic peroxidases is based on non-specific, one electron oxidation of the benzenic rings in the different lignin substrates in synergy with oxidases that generate hydrogen peroxide (Ahmad, 2010). They can catalyse the cleavage of α, β and β-ether bonds (including β-O-4 linkages) leading to the efficient degradation of lignin into mono-aromatic structures, which has been demonstrated using lignin model compounds (Schoenherr *et al.,* 2018; Sahinkaya *et al.,* 2019).

The classical and efficient lignin degrading peroxidases (LiP, MnP and VP; members of the class II superfamily of plant peroxidases) are commonly found in fungi but lacking in most bacterial genomes and metagenomes (Welinder, 1992; Davis *et al.,* 2013; Brown *et al.,* 2012; Le, 2021; Adamo *et al.,* 2022). Research efforts geared towards the discovery of bacterial lignin degraders for industrial depolymerization of lignin has intensified in recent years due to the apparent advantages bacterial enzymes hold over their fungal counterparts (Chen and Li 2016; Chen *et al.,* 2015). This has resulted in the identification of an abundance of a new superfamily of heme peroxidases (the DyP type peroxidases) from several bacteria such as *Bacillus* (Min *et al.,* 2015; Adamo *et al.,* 2022; Zhu *et al.,* 2017; Raj *et al.,* 2007; Mayr *et al.,* 2021), *Rhodococcus* (Ahmad *et al.,* 2011; Sahinkaya *et al.,* 2019), *Pandoraea* (Chen *et al.,* 2012), *Klebsiella* (Angzass *et al.,* 2016), *Pseudomonas* (Rahmanpour and Bugg 2015; Loncar *et al.,*

2019; Yang *et al.,* 2018; Pour and Bugg, 2015), *Streptomyces* (Buraimoh *et al.,* 2017, Riyadi *et al.,* 2020) *etc* that have shown evidence of the ability to readily degrade single ring aromatic substrates and lignin like compounds (van Bloois *et al.,* 2010, Chen *et al.,* 2015, Liu *et al.,* 2017; Uchida *et al.,* 2015; Loncar *et al.,* 2019; Le NG, 2021; Yang *et al.,* 2018; Vuong et *al.,* 2021). These bacterial peroxidases do not share sequence and phylogenetic similarity to the better understood and well-studied fungal enzymes hence only scanty literature about them is available presenting a significant challenge in understanding bacterial lignin degradation and a set back to the use of bioinformatics in the identification of potential bacterial lignin degrading enzymes (Sahinkaya *et al.,* 2019; Loncar *et al.,* 2019; Li *et al.,* 2009).

Structurally, DyPs show a dimeric ferredoxin-like fold consisting of two domains made up of a four-stranded anti-parallel β-sheet surrounded by α-helices with a non-covalently bound heme b cofactor located at the cavity between the two domains, a highly conserved GXXDG-motif, and a conserved proximal histidine, which acts as the fifth ligand of the heme iron (Janusz *et al.,* 2017; de Gonzalo *et al.,* 2016; Chen *et al.,* 2015; Chauhan 2020). Yet, while DyPs are structurally unrelated to the common fungal peroxidases, some bacterial DyPs are secreted via the Tat secretion machinery just as with secreted fungal peroxidases (de Gonzalo *et al.,* 2016; Ahmad 2010) and the similarity of catalytic sites (heme pocket) which is considered as a type of convergent evolution explain them having similar catalytic properties (Rahmanpour and Bugg, 2015; Janusz *et al.,* 2017). Therefore, DyPs can be regarded as the bacterial equivalent of the fungal lignin degrading peroxidases, but they are much easier to manipulate as their functional expression does not involve post translational modification (Chen and Li, 2016).

The peroxibase database classifies DyP-type peroxidases into four phylogenetically distinct subfamilies; A, B, C, and D (Janusz *et al.,* 2017; Pour and Bugg 2015; Sahinkaya *et al.,* 2019) with those in classes A, B, and C predominantly from bacteria, while those in class D are largely from fungi (Pour and Bugg 2015; Chen and Li, 2016; Liu *et al.,* 2017). Although all of the four classes of DyPs belong to a common family and have similar tertiary structures, there isn't very high similarity in amino acid sequences between different classes and the different classes exhibit characteristic

features distinct from one another (Yoshida and Sugano, 2015; Xu *et al.,* 2021). Because of these discrepancies, DyP-type peroxidases were re-classified using structure-based sequence alignments. In the new classification, DyP-type peroxidases were subdivided into three classes: Class P (former class B which carry the most compact structures due to having shorter amino acid sequences and lower catalytic efficiency than other classes and are mostly encoded in bacteria and lower eukaryotes), Class V (former classes C and D that have the many extra sequences in same regions), and Class I (former class A that have the extra sequences fewer than class V) (Sahinkaya *et al.,* 2019; Xu *et al.,* 2021; Lauber *et al.,* 2017).

DyPs are multifunctional enzymes representing not only typical peroxidase activity but also dye-decolorizing activity. Hence, they have broad substrate specificity, catalysing the oxidation of a wide spectrum of industrially relevant peroxidase substrates, anthraquinone and azo dyes, and lignin derived chemicals (Chen and Li, 2016; de Gonzalo *et al.,* 2016; Chauhan, 2020). Due to their peroxidase and dye decolourising activities, they have been found to be a rich source of potent biocatalysts with potential biotechnological application in lignin degradation and bioremediation of dye-contaminated wastewater, enzymatic whitening of whey-containing foods and beverages and as antimicrobial (pro)drug targets (Loncar *et al.,* 2019; Chen and Li, 2016).

To ascertain the functionality of metagenomics predicted genes that are continuously being discovered and documented in pfam, peroxibase and other databases, and to study their properties relevant to biocatalysis, it is pertinent to clone and produce the proteins of these genes and to perform in vitro assays that can be employed to test and compare the predicted activities independent of the system where it was discovered from (Fisher and Fong, 2014; Robinson *et al.,* 2021). Recombinant protein expression in microbial hosts is being widely employed in the production of large amounts of enzymes replacing the use of huge amounts of animal and plant tissue extract that were hitherto required to produce small amounts of a protein of interest (Nevalainen *et al.,* 2005; Hempel *et al.*, 2011; Su *et al.*, 2012). Purification of recombinant proteins is one of the most critical steps because of how important it is to ensure all impurities and contaminants are eliminated while conserving the functional properties of the protein

of interest. Protein purification methods exploit the general properties of proteins such as size, solubility, charge and binding affinity, and several methods based on these properties have been designed and described (Berg *et al*., 2002; Janson, 2011). The specific binding abilities of biological molecules such as ligands, antibodies, inhibitors are exploited in the isolation of target proteins from crude samples. Immobilized metal affinity chromatography (IMAC) is the most widely used binding affinity technique in research for single step purification of recombinant proteins, directed towards protein side chains (usually polyhistidine-tag). In IMAC, proteins or peptides with polyhistidine tag for example are separated according to their affinity for metal ions that have been immobilized by chelation to an insoluble matrix (Porath, 1992; Saraswathy and Ramalingam, 2011). Using an organic compound such as imidazole of different concentrations, polyhistidine-tagged proteins can be eluted via competitive interaction between imidazole and the metal-charged resins in either stepwise manner or gradient concentration of imidazole (Schmitt *et al.,*1993). In some cases, the target protein may be eluted alongside other contaminant proteins after IMAC purification, therefore, additional purification steps using size exclusion or ion-exchange chromatography can be employed. Ion exchange chromatography (IEX) is another frequently used technique for purification of proteins, peptides, nucleic acids, and other charged biomolecules based on differences in their charge properties (Fekete *et al.,* 2015). The technique can separate molecular species that have only minor differences in their charge properties, for example two proteins differing by one charged amino acid. These features make IEX well suited for capture, intermediate purification, or polishing steps in a purification protocol.

Developing assays for quantifying ligninolytic activity is a challenge as lignin is heterogeneous and enzymatic degradation thereby occurs by multiple mechanisms (Fisher and Fong, 2014). A bottom-up approach consists of testing a variety of monomer compounds that represent a specific class, for instance, phenolic or nonphenolic compounds such as ABTS, guaiacol, 2.4-diclorophenol, syringaldazine, veratryl alcohol, DMP, catechol, 2,6- dimethoxyphenol, Congo Red, Reactive Blue, Reactive Black 5, hydroquinone, etc (Janusz *et al.,* 2017; Fisher and Fong, 2014; Reid, 2011; Catucci *et al.,* 2020). Although these substrates

may not be specific to lignin, they give good indication of general peroxidase activities and can be useful for studying the kinetics of purified enzymes (Ahmad, 2010). A second level of complication entails the testing of the activity of the enzyme against one or more dimeric lignin model compounds such as veratrylglycerol-β-guaiacol ether (VGE) and guaiacylglycerol-β-guaiacol ether (GGE). Further analysis requires the assessment of enzymatic activity against polymeric lignin preparations such as kraft lignin, organosolv lignin, nitrated lignin, or wheat straw lignocellulose (Chen *et al.,* 2015; Linde *et al.,* 2021; Brown *et al.,* 2012; Min *et al.,* 2015; Catucci *et al.,* 2020). Spectrophotometric, mass spectrometric or fluorescent assays are mostly employed.

4.3 Methods

In this section, we detail the methods and experiments we performed to clone, express, purify and characterize a putative dye-decolorizing peroxidase (B-38773), one among the selected genes of interest that were amplified in the previous chapter. Blunt-ended PCR products were directionally cloned into pET151/D-TOPO vector, and the gene construct used to transform *E. coli* TOP10 competent cells for maintenance and propagation. Bacterial plasmid DNA extracted from insert containing cells were cultured and used to transform *E. coli* BL21 Star™(DE3) cells for IPTG (isopropyl-1-thio-β-D-galactopyranoside) induced heterologous expression. A two-step protein purification employing affinity and ion exchange chromatography was used to isolate the pure recombinant protein. The purified protein was quantified and tested for the predicted activity as a dye-decolorizing peroxidase against typical peroxidase, anthraquinone dye, and lignin-like substrates (ABTS, RB19 and alkali kraft lignin respectively). Biochemical characterization (pH and temperature profiles and optima, and kinetic parameters; V_{max}, K_m, and K_{cat}) of the recombinant protein B-38773 were preliminarily investigated and determined using ABTS as substrate.

4.3.1 Cloning of PCR products into pET151/D-TOPO vector

Cloning and expression of our amplified genes of interest were performed using the TOPO cloning strategy of the Champion™ pET Directional TOPO$_R$ Expression Kit

(Invitrogen) into expression vector pET151/D-TOPO in a series of steps summarised in the flow chart (Figure 4.1) and described in greater details below.

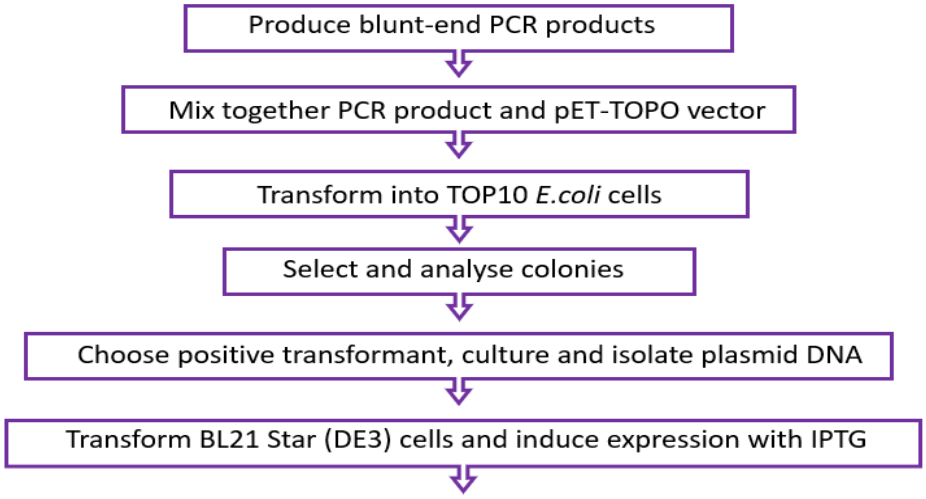


Figure 4.1 Cloning and expression steps
Flowchart of general steps required for cloning and expression of blunt-end PCR products with the Champion™ pET Directional TOPO$_R$ Expression Kit.

Having successfully produced blunt-end PCR products of our genes of interest by amplification with Q5 polymerase (Section 3.3.9), the PCR products were cloned into pET151/D-TOPO vector in a ligation-independent method that does not require the use of restriction endonucleases and DNA ligase. The following reaction in table 4.1 was set up for each blunt-end PCR product of our 3 genes of interest (A-30342, B-38773, and C-08687).

Table 4.1 TOPO cloning reaction mixture

Reagents	Volume (µl)
Fresh PCR product Salt solution Sterile water TOPO vector	2 1 2 1
Total Volume	6

The components were gently mixed, incubated for 5 minutes at room temperature and placed on ice afterwards. Three microlitres (3 µl) of the cloning reaction mixture for

each gene from above was added to one vial of one shot TOP10 chemically competent *E. coli* cells (kit supplied), mixed gently and incubated on ice for 30 minutes. These cells were used for propagation and maintenance of recombinant plasmids only. The cells were heat shocked at 42°C for 30 seconds without shaking, then immediately transferred to ice. Two-hundred and fifty microlitres (250 µl) of room temperature S.O.C medium was added to each vial and placed horizontally in a shaking incubator at 37°C and 200rpm (rotations per minute) for 1 hour. Twenty and eighty microlitres (20 and 80µl) of the transformation reaction mixture was each spread onto freshly prepared Luria–Bertani (LB) + Agar selective plates containing the antibiotic carbenicillin (50 µg/ml) in a sterile environment using a sterile L-shaped spreader. Plates were incubated at 37°C overnight (about 16 hours).

4.3.2 Screening colonies for positive clones

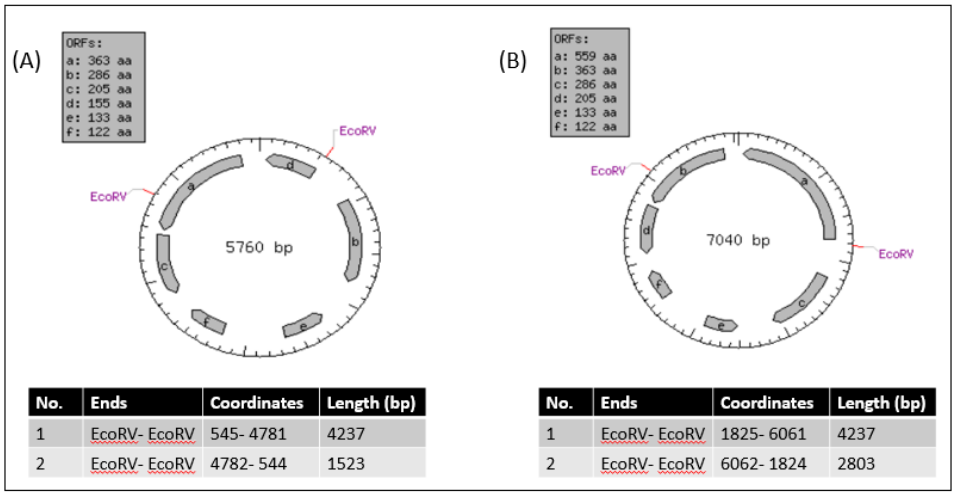
In molecular cloning, there is no one-shot method of identifying positive clones which contain the gene insert from the many clones that usually grow following an efficient cloning reaction and incubation. We used colony PCR to rapidly screen large numbers of colonies per time using the gene specific primers and *Taq* polymerase. Only gene B-38773 showed positive results with colony PCR screening and so it was carried on for further validation of insert presence by restriction enzyme digest and DNA (sanger) sequencing. Genes A-30342 and C-08687 were not further processed after this point.

4.3.3 Plasmid DNA extraction

One positive colony for gene B-38773 from the culture plate was picked with a sterile loop, inoculated into 10ml of LB containing the appropriate antibiotic (50 µg/ml carbenicillin) and cultured overnight at 37°C with shaking at 190rpm. Plasmid DNA was isolated from the overnight culture using the ISOLATE II plasmid mini kit (Bioline, UK), low copy plasmid protocol. Concentration and purity ($A_{260}/_{280}$) of extracted DNA was measured spectrophotometrically using the Nanodrop 2000 to be 270ng/ µl and 1.84 respectively. Plasmid DNA was stored at -20°C.

4.3.4 Verifying insert using restriction enzyme digest and Sanger sequencing

For further confirmation of the presence of inserts as suggested by colony PCR, the extracted plasmid DNA from the positive clone was subjected to a diagnostic restriction enzyme digest. This is frequently done to confirm presence of insert before going on to further verification by DNA sequencing. Here, we performed a custom digest on NEBcutter vs 2.0 (New England biolabs) and chose to use the restriction enzyme EcoRV to digest our plasmid construct. The restriction sites and expected fragment sizes for a successfully digested non-insert containing (empty), and an insert-containing plasmid using this enzyme are shown below.



No.	Ends	Coordinates	Length (bp)
1	EcoRV- EcoRV	545- 4781	4237
2	EcoRV- EcoRV	4782- 544	1523

No.	Ends	Coordinates	Length (bp)
1	EcoRV- EcoRV	1825- 6061	4237
2	EcoRV- EcoRV	6062- 1824	2803

Figure 4.2 Custom digest of empty and insert-containing plasmids with EcoRV
Visual representation of restriction sites, coordinates and expected sizes of fragments from a custom digest using EcoRV on the empty (A) and insert-containing (B) pET151/D-TOPO vector generated on NEBcutter Vs 2.0.

To determine appropriate restriction enzyme digest protocols for the chosen endonuclease, NEBtools™ (NEB, UK) via the NEBcloner® (v 1.3.13), was consulted and the protocol shown below was used (Table 4.2.)

Table 4.2 Protocol for restriction enzyme digest of gene B-38773 plasmid DNA construct.

Component	Amount (µl)
Plasmid DNA (350 - 500 ng)	2
NEB 3.1 buffer	2.5
EcoRV enzyme	0.5
Nuclease Free Water	20
Total	25

We incubated the reactions in a thermocycler for 1 hour at 37°C after which the reaction was terminated by the addition of 5 µl of 6X purple loading dye. Plasmid DNA of the empty vector was also digested to serve as control. For detection by agarose gel electrophoresis, 10 µl of this total reaction (for both B-38773 and control) was each loaded and run on a 1 % agarose gel alongside a 1Kb DNA ladder as marker.

Upon confirming the presence of insert by observing gel pattern of the expected fragment sizes, the extracted plasmid DNA sample was sent for sanger sequencing (Source Bioscience, UK). Sample was prepared (5.5ul of ~100ng/ µl DNA) according to the specified guidelines of source biosciences using the T7 forward (5´-TAATACGACTCACTATAGGG-3´) and reverse (5´-TAGTTATTGCTCAGCGGTGG-3´) primers supplied with the champion pET151/D-TOPO$_R$ expression kit. Sequence fidelity was analysed by aligning the data from Sanger sequencing against the WGS derived nucleotide sequence for gene B-38773 using the EMBOSS Needle multiple alignment tool offered via EMBL-EBI (https://www.ebi.ac.uk/Tools/psa/emboss_needle/).

4.3.5 Recombinant protein expression of gene B-38773 construct

BL21 Star™(DE3) One Shot *E. coli* cells supplied with the champion pET151/D-TOPO® expression kit was used as host cells for the expression of pure plasmid DNA of our pET-TOPO gene construct. pET151/D-TOPO® allows expression of recombinant protein with an N-terminal tag containing the V5 epitope and a 6xHis tag. Also supplied in the kit is the expression positive control vector;

pET151/D//*lacZ* (8832 bp) which contains a *lacZ* gene coding for a β-galactosidase that has been directionally TOPO Cloned into pET151/D-TOPO vector in frame with the N-terminal peptide containing the V5 epitope and the 6xHis tag. The size of the β-galactosidase fusion protein is approximately 120 kDa.

The recombinant plasmids harbouring gene B-38773 and positive control (pET151/D//*lacZ*) were transformed into BL21 Star™(DE3) cells by mixing ~10 ng of each plasmid DNA to one vial of BL21 cells, incubated on ice for 30 minutes, and heat shocked for 30 seconds at 42°C in a water bath. Room temperature S.O.C medium (250 µl) was added to the vials and placed horizontally in a shaking incubator at 37°C with 200rpm (rotations per minute) for 30 minutes. The entire transformation reactions (gene B-38773 plasmid construct and positive control) were each added to 10 ml of LB broth containing 50 µg/ml carbenicillin and incubated to saturation overnight at 37°C with shaking at 190rpm.

Pilot protein expression was performed in small volumes to determine the optimum expression conditions from which a scale up expression can be done to produce higher amounts of the recombinant protein for biochemical testing. The overnight cultures for gene B-38773 and *lacZ* positive control were diluted in LB containing 50 µg/ml carbenicillin to an OD_{600nm} of 0.1. The diluted cells were then grown until OD_{600nm} was ~0.53 (mid-log phase) for approximately 1.5 hours. Each 10ml culture was split into two 5ml volumes and isopropyl-1-thio-β-D-galactopyranoside (IPTG) to a final concentration of 0.7mM was added to one of each of the 5ml cultures to induce expression while the other set of 5ml cultures served as non-induced samples. 500 µl aliquots were removed from each 5ml cultures of induced and non-induced samples for the gene B-38773 and *lacZ* positive control, cells were pelleted at 11,000 x g for 30 seconds and stored frozen at -20°C as time point zero samples (i.e., $T0_{induced}$ and $T0_{non\text{-}induced}$). Samples were incubated at 37°C with shaking and after every hour, 500 µl aliquots were taken from each sample for 4 hours, cell pellets collected and denoted as time points T1- T4 with subscripts (i) or (ni) to represent induced and non-induced samples.

Cell pellets from each time point were lysed in 200 µl of lysis buffer (see recipe in appendix) to which protease inhibitor, lysozyme and benzonase were added to

enhance cell lysis. Three freeze-thaw cycles in dry-ice and water bath at 42°C were performed to achieve complete lysis and release of periplasmic proteins. The soluble protein (supernatant) was then harvested from the lysate following centrifugation at 4,000Xg in a temperature regulated sorvall legend centrifuge (Thermofischer scientific, UK) at 4°C, for 30 minutes.

4.3.6 Analysing samples from pilot expression by SDS-PAGE and Western blot

To determine expression success and integrity of the expressed proteins, the clarified lysates from the time point expression studies for gene B-38773 and the *lacZ* control (both induced and non-induced for each) were analysed using Sodium Dodecyl Sulphate – Poly Acrylamide Gel Electrophoresis (SDS-PAGE) with slight modifications to the Laemlli protocol (Laemlli, 1970), and then by western blot guided by the presence of bands of the expected size of the protein as predicted by the Prot-Param webtool on ExPASy server online (ExPASy; web.expasy.org).

Denaturing gels were prepared using the recipe detailed in Appendix 4, and the Mini-PROTEAN hand-caste system from Bio-Rad (Bio-Rad, UK). A reducing sample was prepared by mixing equal amounts of clarified lysate and 2X SDS-PAGE sample Buffer (with B-mercaptoethanol freshly added). The samples were incubated in a thermocycler at 95°C for 5 minutes to denature the proteins, and 15 µl each of the denatured proteins were loaded onto 2 SDS-PAGE gels producing mirror copies of each other alongside 5 µl of the pre-stained protein molecular weight marker (Bioline, UK). Gels were run at constant 100 V until samples started separating in the resolving gel and then increased to 120 V for ~1.5 hours or until molecular weight markers were well separated on the gel. At the end of electrophoresis, one of the gels was stained in approximately 10 ml of Coomassie InstantBlue™ Protein Stain (Expedeon) and allowed on a shaker until bands developed. The gel was de-stained by washing severally with water to make the bands more visible. Gel images were captured in the G-box image documentation system (Syngene, UK).

Western blot technique which involves the transfer of proteins from an SDS-PAGE gel onto a membrane and detection via chemiluminescence after reacting with antibodies

specific for that protein was used to confirm that the overexpressed bands seen on the stained SDS gel corresponds to our expressed His-tagged recombinant protein in this study (modified protocol from Yang and Mahmood, 2012). The duplicate gel from the SDS-PAGE run was transferred onto a PVDF membrane (0.45µm) in cold western transfer buffer at a constant current of 400 mA for 1 hour using an ice pack and magnetic stirrer to avoid heat accumulation in the buffer and ensure proper cooling. After protein transfer and immobilisation onto the membrane, the membrane was covered in a blocking solution made up of 5% dried skimmed milk in a mixture of 1x PBS supplemented with 0.1%Tween (PBST buffer) for 2 hours. The membrane was washed using 10ml PBST buffer on a shaker three times for 10 minutes per wash. After this, the membrane was incubated overnight with slow shaking at 4°C 1 hr in a solution of 2.5% milk in PBST buffer to which the primary antibody- Monoclonal Anti-HIS Tag mouse antibody (Sigma-Aldrich) was added. The primary antibody was collected, and the membrane washed 3 times with PBST at 10 minutes intervals per wash. Incubation in secondary antibody solution (Anti-mouse IgG produced in goat from Thermo fisher Scientific + 2.5% milk in buffer PBST) followed for another 1 hour. The tween was rinsed off the membrane by washing in 1xPBS only before visualisation by chemiluminescence. The membrane was incubated in the dark for 5 minutes after adding equal volumes of the Stable Peroxide Solution and the Luminol/Enhancer Solution of the West Dura chemiluminescence detection kit (Thermofischer scientific, UK). Images were captured using the G: Box (Syngene, UK). The results from the captured images here guided our decision on the time point at which expression was best under the specified conditions. These determined optimum conditions were applied in the scale up expression to produce large volumes of our protein of interest for downstream assays.

4.3.7 Scale up expression

Having determined the optimum conditions for successful expression of our gene of interest from the pilot study, we proceeded to scale up expression to a 1Litre bacterial culture. 1ml of the culture of the BL21 cells transformation reaction in section 4.2.5 was used to inoculate 50ml of LB containing the appropriate antibiotic and grown overnight

incubated at 37°C with shaking (190 rpm). The entire 50ml overnight starter culture was used to inoculate 1Litre LB medium (1:20 dilution). The 1L culture was grown to an OD_{600nm} of ~0.55 (midlog phase). Expression was induced with 0.7mM IPTG at 37°C with shaking at 225 rpm for 4 hours being the optimum conditions determined from pilot studies in section 4.2.5.

Cells were harvested by centrifugation (10,000×g for 40 min at 4°C), lysed, supernatant collected, and expression success validated by SDS-PAGE and western blot analyses as previously described in section 4.2.6.

4.3.8 Purification of the recombinant protein

The clarified lysate obtained following the optimised large-scale expression in section 4.2.7 which have a 6XHis-tag at the N-terminal was prepared for purification. Purification was carried out by employing immobilised metal affinity column chromatography (IMAC) using the ÄKTA start chromatography system (GE Healthcare, UK). We used the HisPur™ Cobalt Chromatography cartridges (1ml) specific for His-tagged proteins (Fischer scientific, UK) for the purification of the recombinant protein and elution was based on competitive dislodgement of His-tag bound proteins using increasing concentrations of imidazole (Shalini, Sharma and Kumar, 2010). The lysate was diluted with an equal volume of equilibration buffer and filtered through a 0.45 µM membrane filter before loading onto the column that has been pre-equilibrated with equilibration buffer (20 mM NaH_2PO_4, 5 mM imidazole, 500 mM NaCl, pH 7.4). Unbound proteins were washed off the column with 15 ml of wash buffer (20 mM NaH_2PO_4, 20 mM imidazole, 500 mM NaCl, pH 7.4) and the targeted protein was eluted with 20ml elution buffer (20 mM NaH_2PO_4, 500 mM imidazole, 500 mM NaCl, pH 7.4) applying gradient elution at a flow rate of 0.5 ml/min and collecting 1 ml fractions of eluates.

The purification run was monitored by observing peaks detected by the UV flow cell in the AKTA system at A_{280}. Aliquots of the flow through fraction (FT), washed off unbound proteins fraction (W), and the protein fractions eluted at A_{280} were collected and analysed by SDS-PAGE and western blot to monitor the success of the purification process before use in any downstream assays. SDS PAGE and WB images of the

purification run showed the presence of high amounts of an unexpected protein with a band size of about 41kDa alongside the expected 46 kDa band size predicted for our protein of interest thereby necessitating further purification.

The impure eluted fractions containing the His-tagged protein were pooled together and subjected to desalting and buffer exchange through a PD-10 pre-packed Sephadex G-25 column (GE Healthcare, UK) pre-equilibrated with 1X start buffer (20mM Tris, pH 8.0) to get rid of the imidazole and other salts from IMAC purification and to ensure the sample was adjusted to the chosen starting pH and ionic strength for another purification step by IEX chromatography. Ion-exchange chromatography (IEX) separates charged biomolecules such as proteins, peptides, and nucleic acids according to differences in their net surface charge. The predicted PI of our recombinant protein according to ExPASy was 5.85, so we chose a pH of 8.0 which is higher than the PI and thus conferring an overall negative charge on the proteins and we ran our partially purified protein sample through the anion exchanger HiTrap Q-Sepharose Fast Flow (QFF) 1ml column (Cytiva, UK) on the ÄKTA start chromatography system (GE healthcare, UK). The column was equilibrated with 10ml of the 1X start buffer, sample applied at 0.5ml/min and unbound substances were washed off using 15ml of 1X start buffer (same used for equilibration). Elution was performed using stepwise ionic salt gradients of the elution buffer (20mM Tris, 1M NaCl, pH 8.0) at 10% concentration intervals starting from 10-100% and collecting 1ml fractions for each percentage concentration of elution buffer. The flowthrough, wash and eluted fractions were run on SDS gel to monitor the purification and western blot was performed to reconfirm the elution of our protein of interest.

The purified protein was desalted and subjected to buffer exchange through a PD-10 gel filtration column into a storage suitable buffer (20mM tris, 10mM NaCl at pH 8.0) and concentrated by loading onto Amicon ultra-2 centrifugal filter columns (10K MWCO, Millipore) by centrifuging at 4000xg for 20 minutes using a swinging bucket rotor. 50% glycerol was added to concentrated proteins and stored in aliquots at -20°C until ready to use in enzymatic assays. The buffer exchanged and concentrated fractions of the protein were analyzed by SDS-PAGE and immunoblotting to visualise the final purity of the isolated recombinant protein.

4.3.9 Estimation of protein concentration by Bradford assay method

The concentration of the purified and concentrated recombinant protein was determined by the Bradford assay method (Bradford, 1976) before it was used in any enzymatic assays. We employed the standard microplate protocol described in the Coomasie (Bradford) protein assay kit instructional guide with slight modifications explained below. Refer to https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0011181_Coomassie_Bradford_Protein_Asy_UG.pdf for detailed instructions on protein quantification by the standard micro plate protocol.

Five microlitres of bovine serum albumin (BSA) protein standards of known concentrations (0.2, 0.4, 0.6, 0.8 and 1.0 µg/µl), recombinant protein of unknown concentration (B-38773) and water (for blanks) were pipetted into wells of a 96-well microplate in triplicates. Two hundred and fifty microlitres of diluted Coomasie dye reagent prepared by mixing 1 part of dye to 4 parts of water was added to each well and mixed properly. The plate containing the reaction components was incubated at room temperature for 10 minutes and then absorbance was measured at 595nm using the Varioskan LUX Multimode Microplate Reader (ThermoFisher Scientific, UK). The average 595nm measurement for the blank replicates were determined and subtracted from the average 595nm measurements of all other individual standard and unknown sample replicates and a standard curve was prepared by plotting the average blank-corrected 595nm measurement for each BSA standard vs. its concentration in µg/µl. Standard error of the mean ($n=3$) was calculated and has been indicated as error bars on the standard curve. The concentration of the unknown protein sample was determined by extrapolating from the equation of linear regression ($y = mx+c$) retrieved from the plotted standard curve. The undiluted protein was used for all enzyme assays.

4.3.10 Enzyme activity and characterisation assays

4.3.10.1 Determining reaction velocity of the enzyme catalysed reaction

Preliminary assays were performed using protocols and conditions adapted from published literature to determine if a linear relationship could be established when

protein B-38773 is reacted against the different substrates (ABTS, RB19, KL) we aim to test its activity on. Horseradish peroxidase (HRP) was used as positive standard in the determinations with each substrate. Enzymatic activity can be observed by the change in absorbance over time (reaction velocity) and estimations are usually done by choosing time points where a linear relationship is maintained (Neeraas, 2019).

A continuous assay method using the recombinant enzyme (B-38773) and horseradish peroxidase (HRP) as standard was performed for all 3 substrates. For ABTS, the reaction contained 1mM ABTS (30 µl), 50mM sodium acetate buffer pH 5 (120 µl), 1mM H_2O_2 (10 µl), and 5 µl of B-38773 active enzyme (1.125 µg protein) incubated at 30°C and absorbance due to the production of the ABTS cation radical was monitored at 420nm (Sahinkaya *et al.,* 2019; Pour and Bugg, 2015; Lauber *et al.,* 2017; Raj *et al.,* 2007; Yang *et al.,* 2018). The oxidative degradation and decolourisation of RB19 dye was assayed in a reaction containing 0.5mM RB19 (30 µl), 50mM sodium acetate buffer pH 4 (120 µl), 1mM H_2O_2 (10 µl), and 5 µl of B-38773 active enzyme (1.125 µg protein) incubated at 30°C and absorbance monitored at 595nm (Celebi *et al.,* 2013; Sahinkaya *et al.,* 2019). For kraft lignin, 10mg amount of KL was dissolved in 1 ml DMSO to make a 10mg/ml stock solution. Reaction mixture containing 10mg/ml KL stock (5 µl), DMSO (10 µl), 50mM sodium phosphate buffer pH 7 (160 µl), 2mM H_2O_2 (10 µl), and enzyme B-38773 (10 µl) incubated at 30°C and absorbance monitored at 465nm (Guo *et al.,* 2021). Other published protocols (Ahmad *et al.,* 2011; Raj *et al.,* 2007; Loncar et al., 2019, Rhamanpour and Bugg 2015) and variations of the above protocol for the kraft lignin assay using different buffers of different pHs, different concentrations of substrate and enzyme and different temperatures were also tried in a troubleshooting effort, but none worked except the one reported. Blank and standard reactions were also set up containing the same reaction components as prepared for each substrate except that the enzyme amount was replaced by the corresponding volumes of buffer (for blank), 0.01mg/ml HRP (for standard) for all three substrates. All the reactions were performed in triplicates and absorbances monitored continuously at 1-minute intervals at the specified wavelengths for 10 minutes using the Varioskan LUX Multimode Microplate Reader (ThermoFisher Scientific, UK). The plot of linear

relationship was constructed using Microsoft excel by plotting the absorbances obtained (a.u) versus time (minutes).

4.3.10.2 Enzyme characterization using ABTS as substrate

4.3.10.2.1 Determination of pH and temperature profiles and optima

The pH profile and optimum for substrate oxidation activity of the purified enzyme was determined using 1mM ABTS as substrate with the same reaction mix described in section 4.2.10.1 above using a set of buffers ranging in pH from 1- 9 and measuring absorbance at 420nm after 5 minutes incubation at 30°C in triplicates. The buffer systems used were 50mM KCl-HCl (pH 1.0 and 2.0), 50mM sodium citrate (pH 3), 50mM sodium acetate (pH 4.0 and 5.0), 50mM sodium phosphate (pH 6 and 7), and Tris-HCl (pH 8.0) (Afreen et al., 2017).

The profile and optimum temperature for activity of the purified enzyme B-38773 was tested at temperatures in the range of 10-70°C in increments of 10°C using the same reaction components described above and at the optimum pH determined. Reactions were incubated in a thermocycler for 5 minutes and stopped by placing on ice. The reaction mixtures were then transferred immediately to micro-well plates and absorbances measured with the varioskan plate reader at 420nm. Blank reactions lacking the active enzyme were also set up and absorbance values of blank reactions were subtracted from absorbances of enzyme containing reactions.

The rate of product formation at each pH and temperature in µMol/min was derived by calculating the concentration of product formed per minute using beer's law from the absorbance measurements and molar extinction coefficients of ABTS (E_{420}=36,000 L M^{-1} cm^{-1}).

A=ECl

*Therefore, C=A/E*l*

Where;

A= Absorbance

E= molar extinction coefficient (in μMol^{-1} cm^{-1})

L= pathlength of light (0.56cm for pathlength correction)

Therefore, we defined one unit of enzyme activity (1U) as the amount of enzyme that oxidises the substrate to produce 1 µMol of product per minute per ml of reaction at 40°C degrees and pH 4 in the presence of 1mM H_20_2.

The activity at the optimal pH and temperature were considered as 100% and relative activity was calculated and plotted against the corresponding pH and temperatures for comparison. Standard error of the mean (*n=3*) was indicated as error bars on each plot.

4.3.10.2.2 Determination of specific activity of B-38773 enzyme

Having determined the optimum conditions of activity for the enzyme with ABTS as substrate, a reaction containing the same components mentioned above at the optimum pH and temperature was set up and incubated for 5 minutes. The change in absorbance monitored at 420nm was determined. Enzyme activity was calculated as reported above and specific activity (Umg^{-1}) was determined.

4.3.10.2.3 Kinetic parameters (Steady state kinetics)

To determine steady-state kinetic parameters of B-38773, the reactions were performed in the same way as described above in section 4.2.10.1 at the determined optimum conditions of activity of the enzyme except that the concentration of substrate was varied (0.1- 5.0mM ABTS) while maintaining same conc. of enzyme (1.125 µg). The Michaelis- Menten constant (K_m which is a measure of an enzyme's affinity for its substrate), and the maximal reaction velocity at a given enzyme concentration (V_{max}) for B-38773 were determined by extrapolating in line with the equation of straight line ($y = mx + c$) derived from a Lineweaver Burk's plot (double reciprocal plot) compared to the Lineweaver Burk equation (Neeraas, 2019, Sahinkaya *et al.,* 2019, Lauber *et al.,* 2017).

4.3.10.3 Dye decolourising activity

Dye-decolorizing activity of the B-38773 enzyme was determined by spectrophotometrically measuring the rate of H_2O_2-mediated decomposition of reactive blue 19 (RB19). The reaction was set up in the same way as described for velocity determination assay of RB19 using a buffer of pH 4 and temperature of 30°C. The

reaction was performed in triplicates and absorbances monitored continuously at 1-minute intervals at wavelengths of 595nm for 10 minutes using the Varioskan LUX Multimode Microplate Reader (ThermoFisher Scientific, UK). The dye decolourising efficiency was determined as the percentage rate of RB19 decomposition using the equation below.

Percentage decolourisation (%) = $(A_{initial} - A_{final})/ A_{initial}*100$

where $A_{initial}$ is the initial absorbance at 595nm and A_{final} refers to the absorbance at 595nm at incubation time.

4.4 Results

4.4.1 Restriction endonuclease digest

Isolated plasmid DNA of the gene construct of B-38773 and the empty pET151/D-TOPO vector (Control) were digested using the restriction endonuclease enzyme EcoRV to confirm the presence of insert in the gene construct.

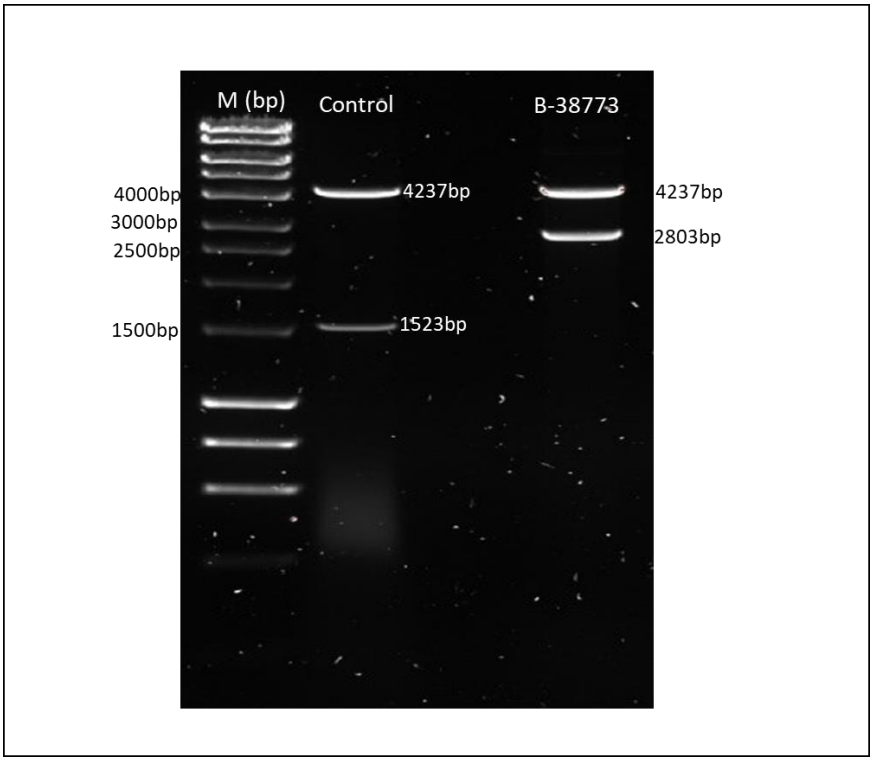


Figure 4.3 Agarose gel image of restriction digest of gene B-38773 construct
The 1% agarose gel was loaded with 1.5 µl of 1kb DNA ladder as DNA marker in lane M, and 10 µl of restriction digest product (25 µl of restriction digest product + 5 µl of purple loading dye) for gene B-38773 construct and empty vector (control) in lanes B-38773 and Control respectively using Eco-RV endonuclease.

From Figure 4.3 above, we confirmed that our gene B-38773 of interest was successfully cloned into the pET151/D-TOPO vector as we observed bands of the expected fragment sizes (4237bp and 2803bp) as predicted by a custom digest of the gene construct with EcoRV restriction endonuclease enzyme. Expected fragment sizes of 4237bp and 1523bp for the empty pET151/D-TOPO vector were also seen confirming the accuracy of the digestion experiment.

Further confirmation by sanger sequencing revealed a 90.4% identity and similarity when obtained sequences from sequencing run were aligned against the original WGS derived sequences for gene B-38773. The successful alignment also confirms that the insert was cloned in frame with the vector in the right orientation for expression to occur (See appendix 3).

4.4.2 Time points protein expression

Aliquots of induced and non-induced cultures of B-38773 and *lacZ* control transformed BL21(DE3) cells from a pilot time point expression were lysed, and the soluble fractions (supernatant) were analysed by SDS-PAGE and western blot to visualise the expression outcome at the chosen conditions.

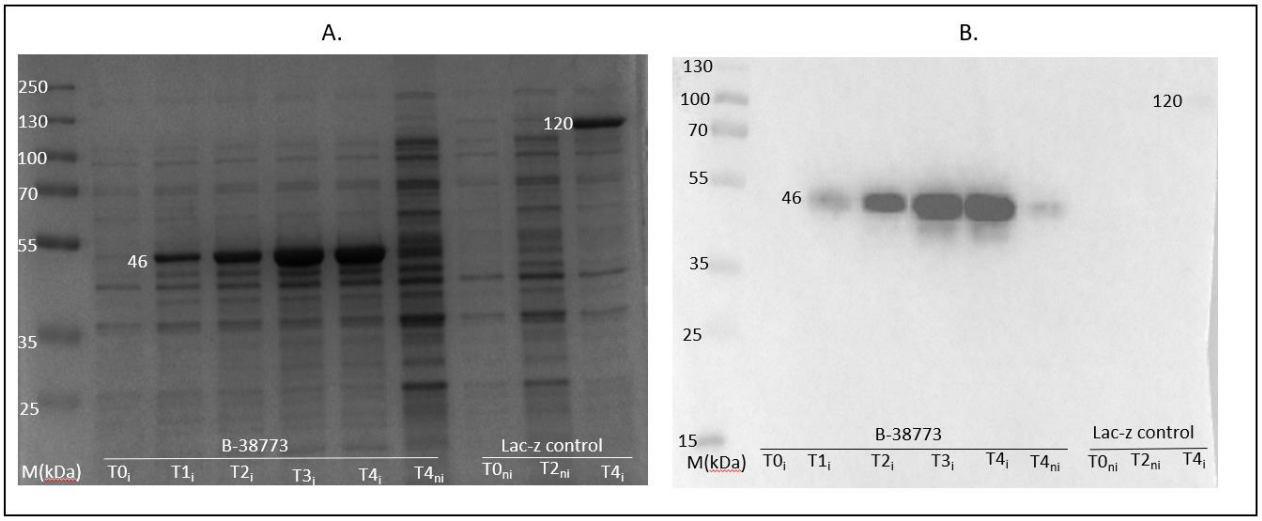


Figure 4.4 SDS-PAGE and Western blot images of expressed proteins of gene B-38773 and lacz control at different time points

(A)SDS PAGE gel loaded with 3µl Page Ruler Plus Pre-stained Protein ladder (lane M) for Protein size estimation. In subsequent lanes, 15µl of equally mixed amounts of cell lysate and 2XSDS loading buffer to which fresh BME was added and denatured by heating were loaded as follows: IPTG induced expression products of gene B-38773 at time 0-4 hrs (Lanes T0i-T4i) and non-induced gene B-38773 product after 4 hours (Lane T4ni). Products of lacZ gene at time 0hrs and 2hrs without IPTG induction and time 4hrs with IPTG induction were loaded in lanes T0ni, T4ni, and T4i respectively. After electrophoresis, the gel was stained with Coomassie instant blue stain for development of protein bands, de-stained by rinsing with water, and visualised in a G:box gel documentation system (B) Western blot image from an unstained duplicate mirror copy of the SDS gel in (A) from which proteins have been transferred unto a PVDF membrane and incubated with antibodies specific for the his-tagged proteins (B-38773 product and β-glucosidase product of lacZ gene) thereby confirming successful expression.

According to Prot-param tool on ExPasy server, protein B-38773 was predicted to have 426 amino acids, a molecular weight of 45850.97 (approximately 46kDa) and a theoretical pI of 5.85. Bands of the expected sizes for the expressed proteins (46kDa and 120kDa for B-38773 and lacz control respectively) were seen on the SDS gel alongside other *E. coli* proteins with increasing intensity from T0 - T4. Western blot images showed bands at same positions for the expected His-tagged proteins. The presence of the 120kDa size β-glucosidase product of the lacz control gene confirms

the efficiency of the BL21 expression system. A band corresponding to the size of our protein of interest was also observed in the lanes where non-induced sample of B-38773 after 4 hours expression was loaded on the SDS gel and western blot images indicating that basal expression occurs even in the absence of IPTG induction.
From the results, expression after 4 hours of IPTG induction (T4) yielded high amounts of undegraded soluble proteins. Thus, we adopted optimal expression conditions at 37°C with 0.7mM IPTG induction for 4 hours.

4.4.3 Protein purification

Clarified cell lysate obtained following scale-up expression of gene B-38773 were purified by a two-step purification process using immobilised metal affinity chromatography (IMAC), and then by ion exchange chromatography (IEX) on the AKTA as described in section 4.2.8.

4.4.3.1 Purification by IMAC

Lysates from BL21 *E. coli* cells induced to express gene B-38773 were loaded onto cobalt charged columns using the AKTA, and His-tagged proteins eluted with gradient concentrations of imidazole. Protein purification was monitored by observing peaks detected by the UV flow cell embedded in the AKTA system at A_{280nm} and collecting the fractions eluted at that absorbance. Aliquots of the lysate (L), flow-through (F), wash (W) and proteins eluted at A_{280nm} were run on SDS-PAGE and western blot to monitor the purification process and assess the purity of the eluted protein fractions. See SDS-PAGE and western blot images below

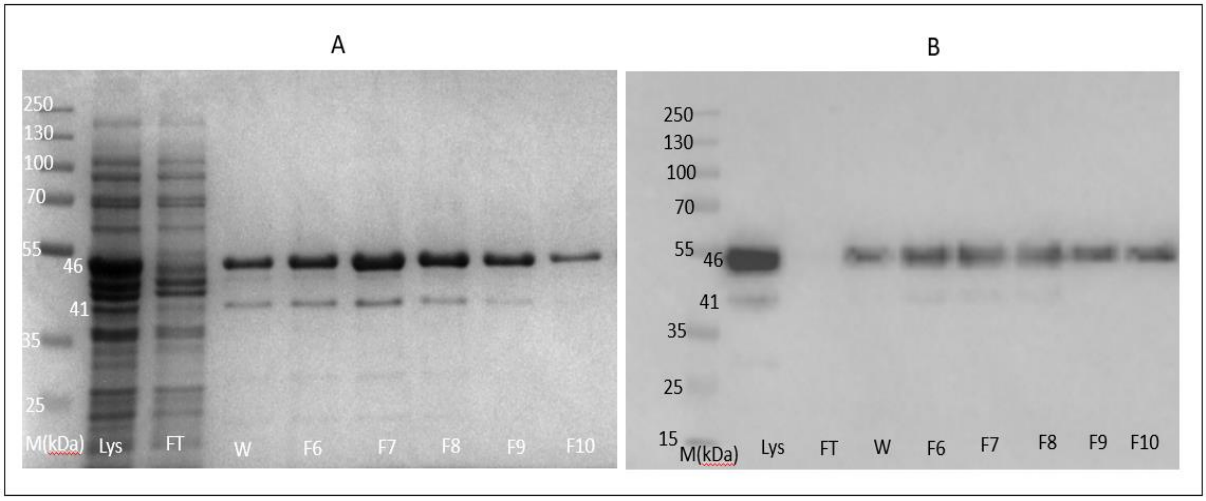


Figure 4.5 SDS-PAGE and western blot analyses of protein purification by IMAC

(A)SDS gel loaded with 3µl PageRuler Plus Prestained Protein ladder (lane M) for Protein size estimation. 15 µl of denatured samples (equally mixed amounts of samples and 2XSDS loading buffer to which fresh BME was added and denatured by heating) of the cell lysate, flowthrough, wash, and fractions eluted at A_{280nm} (F6-F10) were loaded in lanes, Lys, FT, W, F6, F7, F8, F9, F10 respectively. After electrophoresis, the gel was stained with Coomassie instant blue stain for development of protein bands, destained by rinsing with water, and visualised in a G:box gel documentation system (B) Western blot image from an unstained duplicate mirror copy of the SDS gel in (A) from which proteins have been transferred unto a PVDF membrane and incubated with antibodies specific for the His-tagged protein product of B-38773.

Figure 4.5. shows the SDS-PAGE and western blot images of purification steps of protein B-38773 by IMAC. The band for purified B-38773 was observed at approximately 46kDa matching the molecular weight predicted by ExPASy, but an additional band of approximately 41kDa was also observed co-purifying alongside our protein of interest in all fractions eluted at A_{280nm} (Fractions 6-10). We also observed the presence of our protein of interest being eluted at the washing off step suggesting the concentration of imidazole in the wash buffer was too high and was capable of dislodging some of the His-tagged proteins at that stage.

The result above showed that our eluted proteins were not very pure, so we employed ion exchange chromatography to further purify the wash and eluted fractions obtained from this purification step.

4.4.3.2 Purification by IEX

Following IEX purification run of the partially purified protein samples through a Sepharose QFF anion exchanger column, aliquots of the partially purified sample (PPS), flow-through (FT), wash (W) and fractions eluted at A_{280nm} were collected and run on SDS-PAGE and western blot to monitor the purification process and assess the purity of the eluted protein fractions. Only eluted fractions containing protein of interest are shown here.

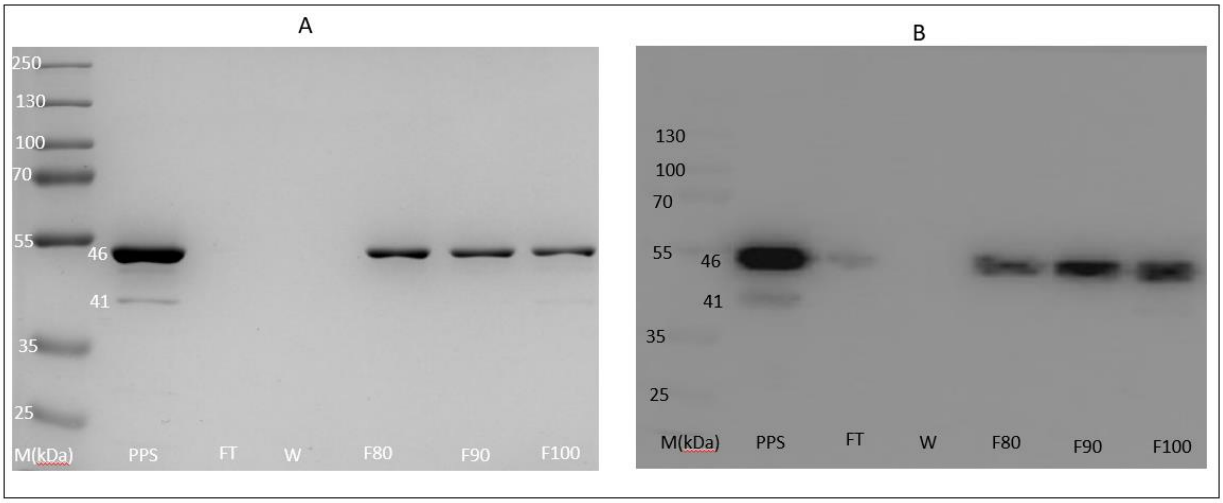


Figure 4.6 SDS-PAGE and western blot analyses of re-purification of partially purified protein B-38773 by IEX

(A)SDS gel loaded with 3µl Page Ruler Plus Pre-Stained Protein ladder (lane M) for Protein size estimation. 15 µl of denatured samples (equally mixed amounts of samples and 2XSDS loading buffer to which fresh BME was added and denatured by heating) of the partially purified protein sample, flowthrough, wash, and eluted fractions at 80%, 90% and 100% concentration of 1M NaCl ionic strength elution buffer (F80, F90, F100) were loaded in lanes PPS, FT, W, F80, F90 and F100. Gel was stained with Coomassie instant blue stain for development of protein bands, destained by rinsing with water and visualised in a G:box gel documentation system. (B) Western blot image from an unstained duplicate mirror copy of the SDS gel in (A) from which proteins have been transferred unto a PVDF membrane and incubated with antibodies specific for the his-tagged protein product of B-38773.

From figure 4.6, Highly pure fractions of protein B-38773 were observed as indicated by single bands of the expected protein size (approximately 46kDa) in the fractions eluted at 80%, 90% and 100% concentration of salt in the elution buffer (i.e., 0.8M, 0.9M and 1M concentrations of NaCl).

The F80-100 fractions were pooled together, desalted and buffer-exchanged into a suitable buffer for storage as one sample and then concentrated. The pooled, buffer-

exchanged, and concentrated proteins were again assessed by SDS-PAGE and western blot to ensure the final protein to be used for downstream assays is still pure and intact.

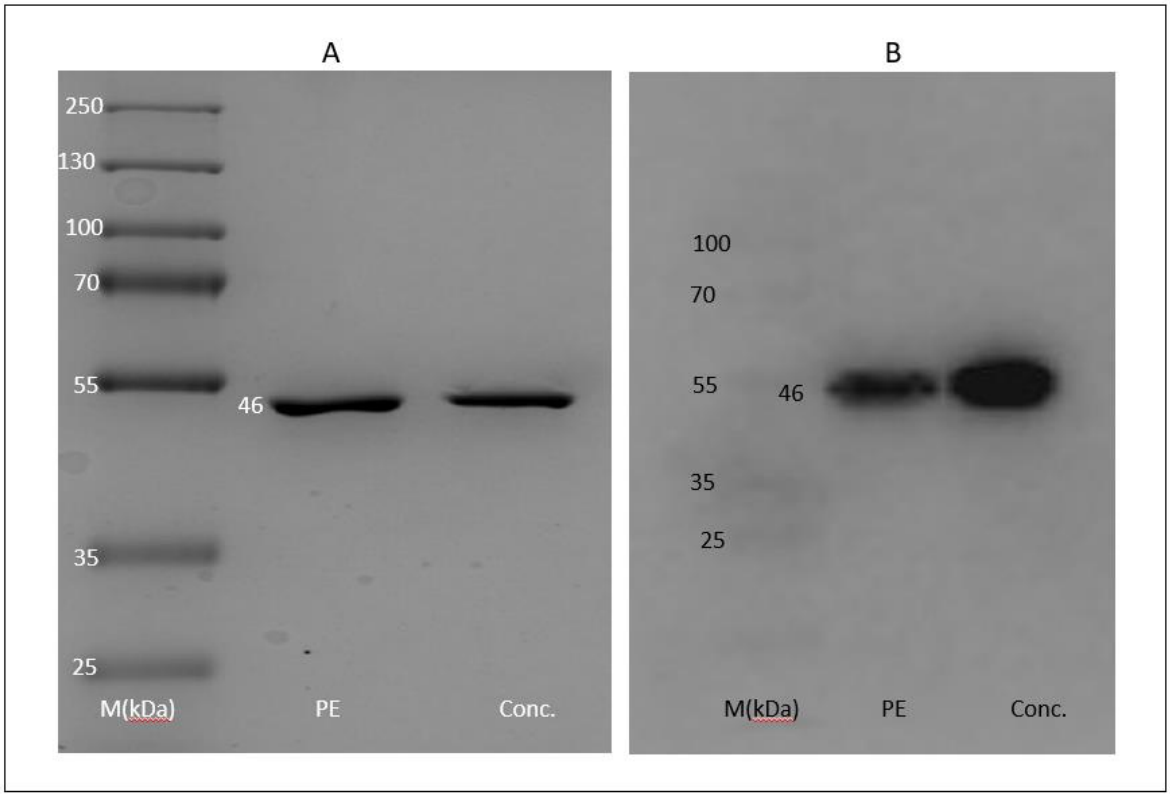


Figure 4.7 SDS-PAGE and Western blot images showing buffer-exchanged and concentrated fractions of re-purified B-38773 protein

(A)SDS gel loaded with 3µl Page Ruler Plus Pre-Stained Protein ladder (lane M) for Protein size estimation. 15 µl of denatured samples (equally mixed amounts of samples and 2XSDS loading buffer to which fresh BME was added and denatured by heating) of the purified and buffer-exchanged (PE) and concentrated (Conc.) protein samples were loaded in lanes PE and Conc. Gel was stained with Coomassie instant blue stain for development of protein bands, destained by rinsing with water and visualised in a G:box gel documentation system. (B) Western blot image from an unstained duplicate mirror copy of the SDS gel in (A) from which proteins have been transferred unto a PVDF membrane and incubated with antibodies specific for the his-tagged protein product of B-38773.

In Figure 4.7, we show SDS-PAGE and western blot images of protein B-38773 fractions that were desalted, and buffer-exchanged to remove the high salt concentrations introduced from the IEX elution buffer, and subsequently concentrated. Samples of buffer exchanged, and concentrated proteins were run alongside a protein

marker and single protein bands of the expected size (approximately 46kDa) were observed for both the buffer exchanged and concentrated protein fractions. From the SDS-PAGE (A) and western blot images (B) we successfully conclude the isolation of pure protein.

4.4.4 Protein estimation by Bradford assay method

Prior to enzyme activity testing, the concentration of the purified recombinant protein was determined by the Bradford assay method using bovine serum albumin (BSA) as a protein standard. A standard curve was generated to estimate the amount of protein within the linear working range of 0.2 - 1.0 µg/µl of BSA standards. The amount of protein following absorbance measurement at 595 nm was calculated using the linear regression equation deduced from the plot of absorbance against concentration.

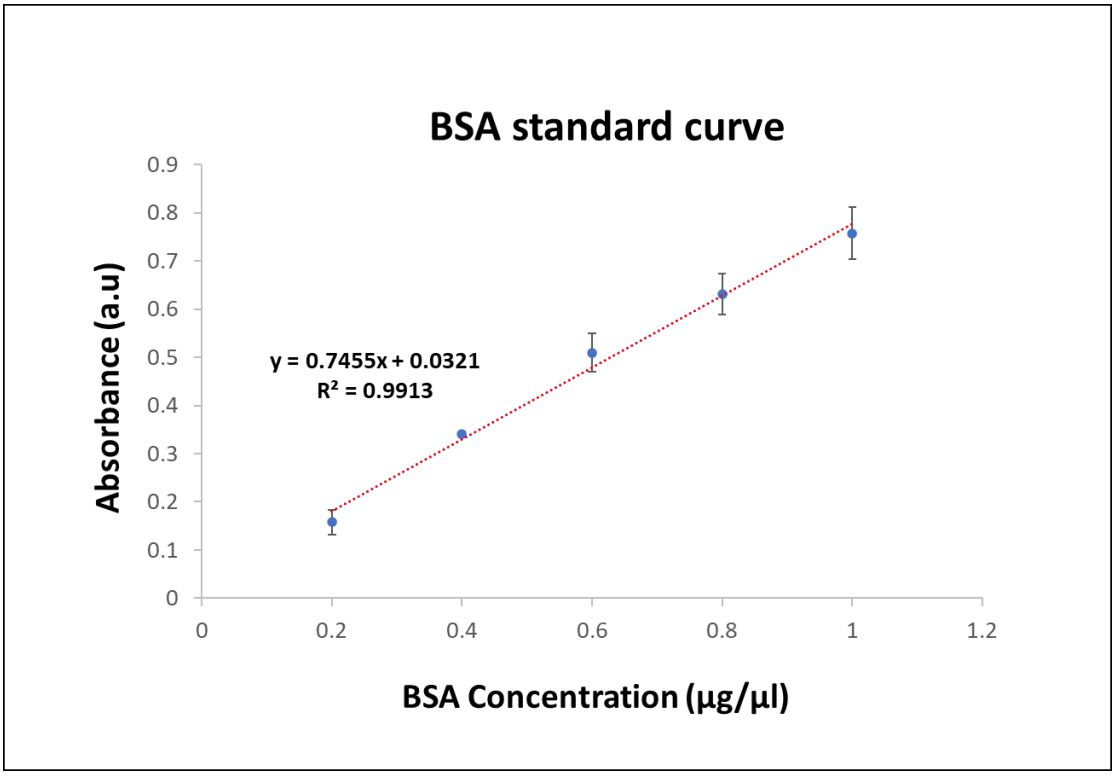


Figure 4.8 BSA standard curve for estimation of protein concentration

Standard curve for estimation of protein concentration where absorbance measurements have been plotted against their corresponding concentrations of BSA standards (0.2- 1.0 µg/µl) derived from a standard micro plate Bradford assay. The equation of linear regression was given as y = 0.7455x+0.0321 (R^2= 0.9913). Error bars represent the standard error mean from triplicates experiments.

Concentration of protein B-38773 was estimated to be 0.225 µg/µl as extrapolated from a blank corrected absorbance value of 0.192 using the equation of linear regression y = 0.7455x+0.0321

4.4.5 Enzyme activity and characterisation assays of enzyme B-38773

4.4.5.1 Reaction velocity of enzyme assay against the substrates ABTS, RB19 and KL.

Results of assays to establish linear reaction velocities for the protein B-38773 against the substrates ABTS, RB19, and KL as described in section 4.2.10.1. Horseradish peroxidase (HRP) was used as positive standard in the determination with each substrate to validate the protocols used.

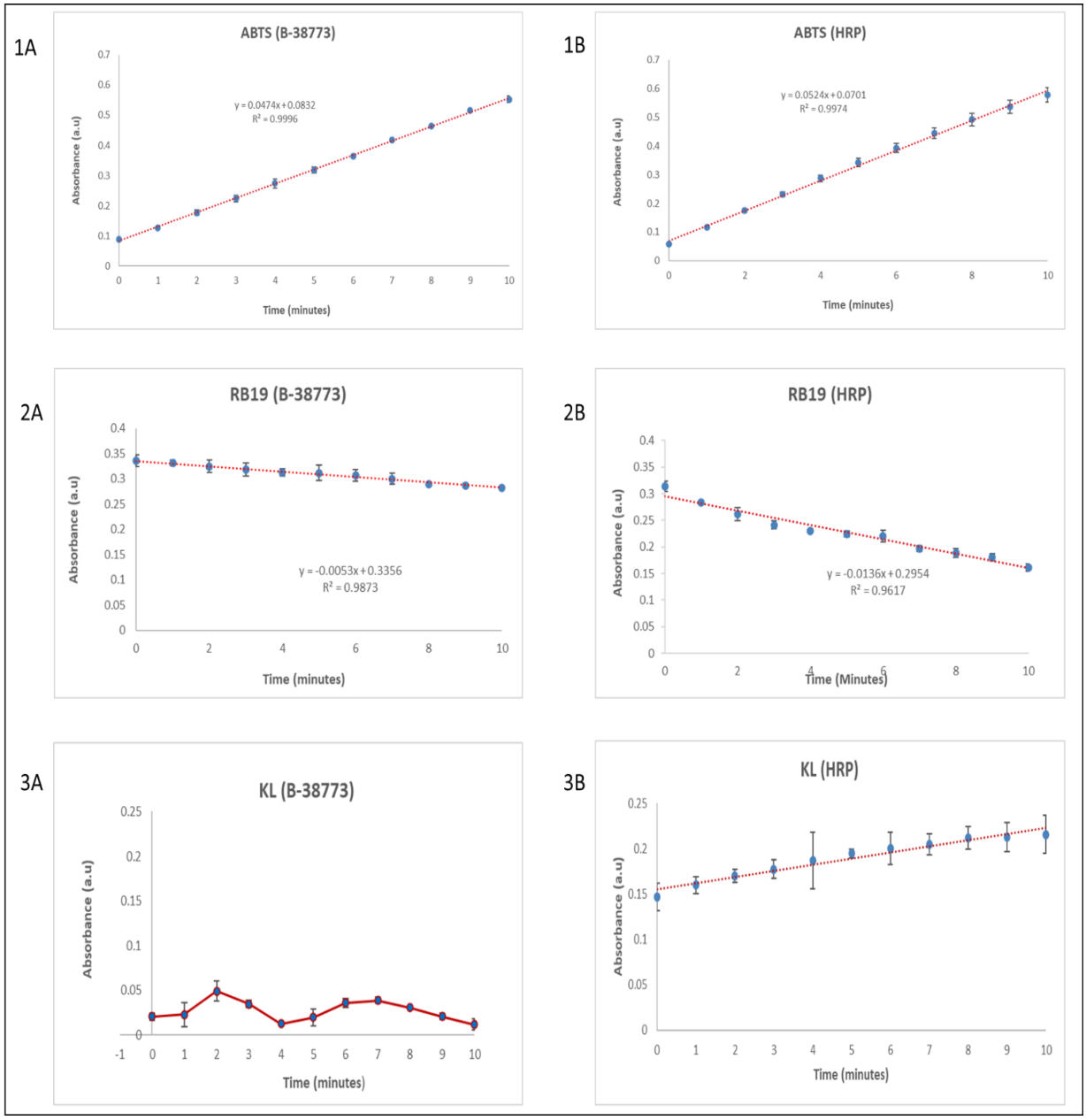


Figure 4.9 Reaction velocities of B-38773 enzyme assay against ABTS, RB19 and KL substrates

Linear plots of absorbances against time for reactions of enzyme B-38773 against the substrates ABTS(1A), RB19 (2A) and KL (3A) with corresponding reactions using HRP as positive control (1B, 2B and 3B) respectively plotted on Microsoft excel. Enzymatic assay on each substrate was carried out within a linear range of 10 minutes. Error bars represent the standard error mean from triplicates experiments.

In Figure 4.9, a linear and consistent increase and decrease in absorbance was observed for ABTS (1) and RB19 (2) respectively as expected for both our enzyme B-38773 and HRP over 10 minutes thereby allowing for the determination of reaction

velocities indicating enzyme activity. In the case of KL (3), an expected linear increase was observed with the positive standard (3B), but no linear relationship of enzyme activity was observed using 10 µl of enzyme B-38773 sample, hence, the velocity of the B-38773 reaction cannot be derived suggesting an absence of enzymatic activity

4.4.5.2 Enzyme characterisations using ABTS as substrate

Following enzyme activity linearity determination, we partially characterised B-38773 spectrophotometrically using ABTS as substrate. We determined the pH and temperature profiles and optima for activity of B-38773. We then determined the kinetic parameters (V_{max}, K_m and K_{cat}) and specific activity from experiments using the optimum conditions determined.

4.4.5.2.1 pH and temperature profiles and optima

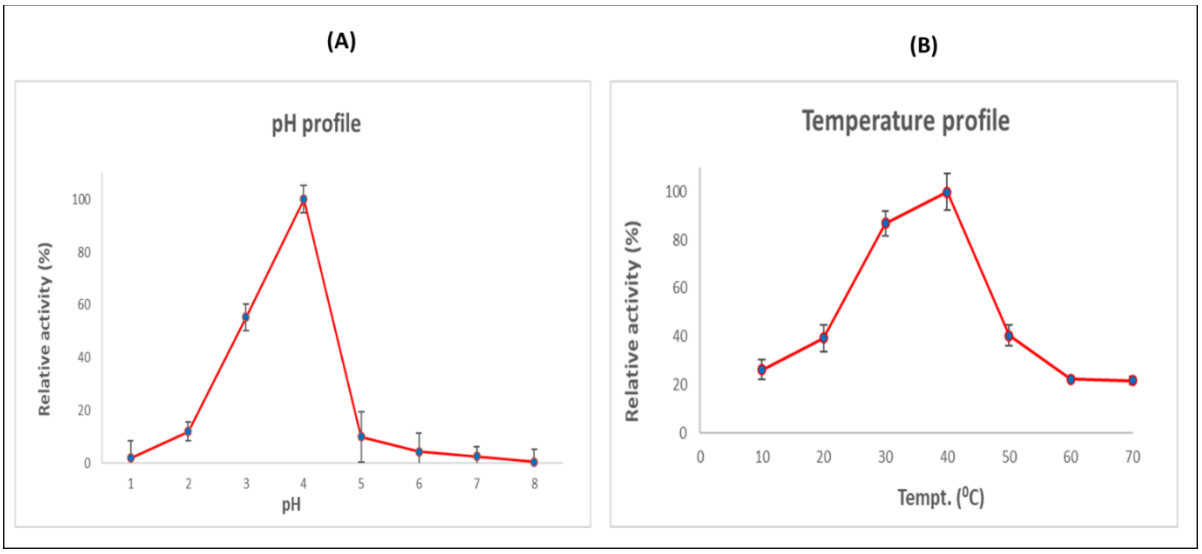


Figure 4.10 pH and temperature profiles of B-38773 using ABTS as substrate
Plots of relative enzyme activities of B-38773 on ABTS against a range of corresponding (A) pHs (1-8) and (B) temperatures (10-70°C) determined after 5 minutes reaction time. Standard error of the mean (*n=3*) has been indicated as positive and negative error bars. Error bars represent the standard error mean from triplicates experiments.

Figure 4.10. shows the pH and temperature profiles of B-38773 determined by enzyme activity assay against ABTS as described in the experimental section. The enzyme was active between pH 2-5 and temperatures of 20-50°C but the optimum pH and temperature for activity were observed at pH 4 and 40°C respectively.

4.4.5.2.2 Enzyme Kinetics parameters

We set up enzymatic experiments using a constant enzyme B-38773 concentration of 1.125ug against varying concentrations of ABTS (0.1-5.0 mM) at pH 4 and 40°C. We measured the change in absorbance after 5 minutes incubation and determined the reaction velocities. Taking reciprocals of the reaction velocities and corresponding substrate concentration, we made a double reciprocal plot (Lineweaver-Burk plot) from which we extrapolated the Km and Vmax.

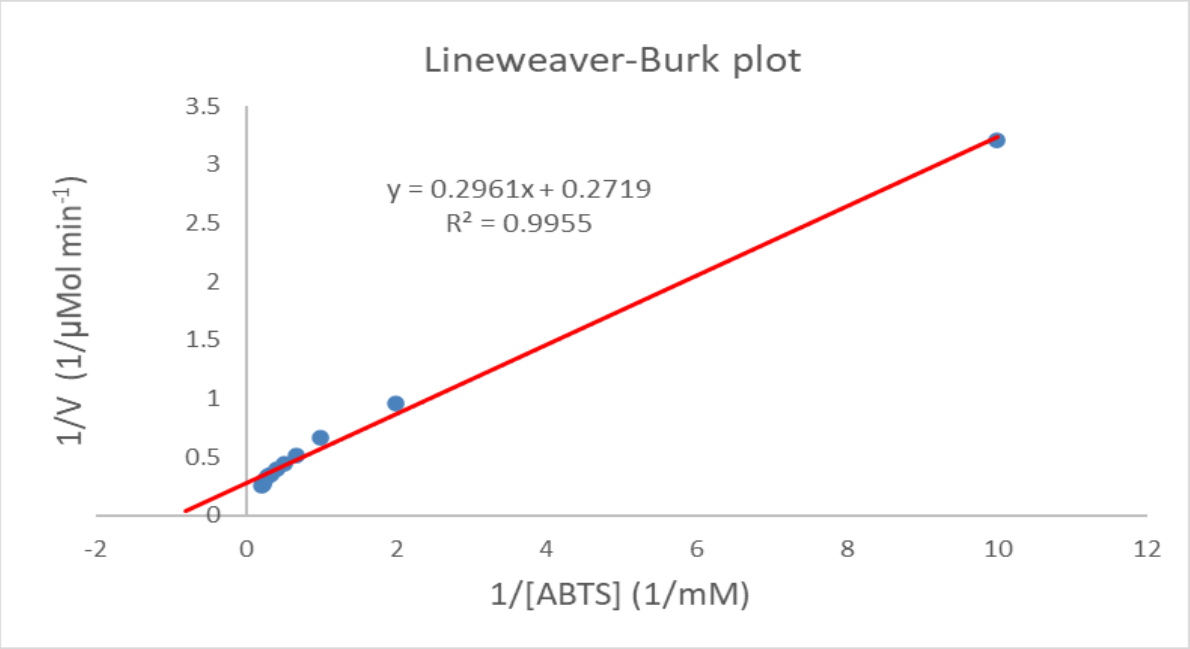


Figure 4.11 Lineweaver-Burk plot

A plot of the reciprocals of the reaction velocities (1/V) and corresponding substrate concentration (1/[ABTS]) using a constant concentration of enzyme B-38773. The linear equation deduced from the plot was y=0.2961x +0.2719 (R^2= 0.9955).

B-38773 displayed simple Michaelis–Menten kinetics. The kinetic parameters determined through the Lineweaver-Burk's plot were 1.089 mM and 3.678µmol min^{-1} for K_m and V_{max}, respectively. K_{cat} was calculated to be 540.9S^{-1} (3.3) and K_{cat}/K_m was 496.7 (3.0). Specific activity was obtained as 12.9U/mg protein.

4.4.5.3 Dye decolourising activity

The efficiency of enzyme B-38773 to decompose the anthraquinone dye RB19 was assayed. A decrease in absorbance for each substrate was monitored up to 10 min at 1 min interval, and the percentage decolourisation at each minute was calculated and presented below.

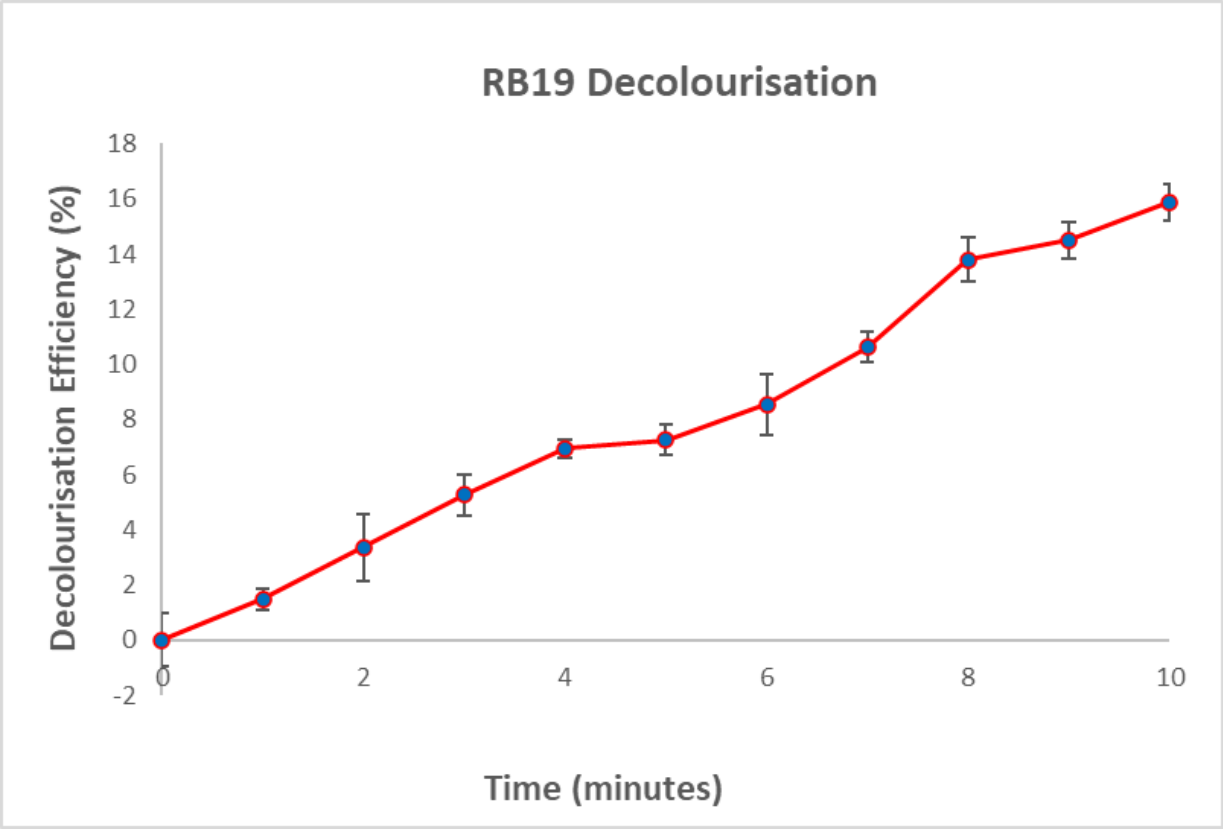


Figure 4.12 Decolourising efficiency of B-38773 on RB19 dye.
A plot of percentage decolourisation of RB19 per minute over ten minutes.
From Figure 4.12, we observe an increase in decolourisation efficiency over time. The longer the reaction was allowed to proceed, the more the percentage decolourisation activity of B-38773 on RB19. At the 10th minute, the percentage decolourisation was 15.9%.

4.5 Discussion

Presented in this chapter is the process employed in the recombinant production, activity testing and characterisation of a putative dye decolourising peroxidase which sequence was obtained from the gut metagenome of the African palm weevil (*Rhynchophorus phoenicis*). This has been made possible by the technique of molecular DNA cloning where any protein from a cell can be produced in nearly unlimited amounts for analysis of biochemical activities, sequence and structural studies which was previously impossible due to the extreme difficulty in obtaining more than a few micrograms of the pure proteins from the cell as they are only present in very small amounts (Alberts *et al.,* 2002). Initially, we selected 3 genes from classes of enzymes associated with lignin degradation (2 putative dye decolourising peroxidases and a laccase-like multicopper oxidase) with the aim to validate the metagenomic analysis and to subsequently produce and test the protein product of at least one of the three. DyP type peroxidases and laccases have been reported to be the two major classes of lignin modifying enzymes found in bacteria (de Gonzalo *et al.,* 2016). All three genes were successfully amplified as reported in the previous chapter but after cloning, only gene B-38773, a putative dye decolourising peroxidase, was carried forward as the other genes were not successfully cloned as observed from colony PCR (data not shown). Troubleshooting the cloning experiment of gene A-30342 and C-08687 will be considered in future work.

We performed further checks by restriction enzyme digest. According to the gel image in figure 4.3., the expected sizes of fragments from a digest of the recombinant plasmid with EcoRV enzyme following a successful ligation and transformation reaction was observed (4237bp and 2803bp) confirming the presence of cloned insert. The efficiency of the digest reaction was also validated by an observation of expected fragment sizes (4237bp and 1523bp) in the digest of the empty vector (control) with EcoRV restriction endonuclease. Orientation of ligation and sequence fidelity were assessed by analysis of sanger sequencing data compared against original WGS sequences, and a similarity of 90.4% was obtained. Although the percentage similarity is expected to be nearer 100%, we found gaps of 5.8% (75 nucleotides) from the pairwise alignment, and the sequence data from sanger sequences was about 90

nucleotides short. This could be due to sequencing error, or PCR bias as PCR product from whole genome amplified DNA was cloned rather than the original extracted metagenomic DNA. However, the observed difference may not account for a significant difference in overall protein function as all sequences in the active site were intact.

The production of recombinant proteins from sequence based expression clones generated from metagenomic DNA libraries of cow rumen (Hess *et al.,* 2011), goatrumen (Le, 2021; Do *et al.,* 2018), slugs (Joynson *et al.,* 2017), soils (Bergmann *et al.,* 2014) and microbial genomes (Rahmanpour and Bugg 2015; Neeraas, 2019; Ahmad *et al.,* 2011; Uchida *et al.,* 2015; Chen *et al.,* 2015; Qin *et al.,* 2018; Loncar *et al.,* 2019) have been demonstrated to be possible and allows scale-up for potential characterisation in several industrial applications (Schumann and Ferreira, 2004; Robinson *et al.,* 2021). These proteins have also been tested to validate their activities as predicted by bioinformatic analysis. In this study, we also employed these techniques to produce and isolate the protein product of gene B-38773, a putative dye decolourising peroxidase selected from the metagenomic library of the African palm weevil. We used The Champion™ pET expression kit for heterologous expression of the recombinant construct of gene B-38773. The vectors in this kit system takes advantage of the high activity and specificity of the bacteriophage T7 RNA polymerase to allow regulated expression of heterologous genes in *E. coli* from the T7 promoter originally developed by Studier and colleagues (Rosenberg *et al.*, 1987; Studier & Moffatt, 1986).

The expression of this approximately 46kDa protein was observed using polyacrylamide gels and confirmed using monoclonal His tag and secondary anti-his-tag antibodies in a western blot experiment (Figure 4.4) as predicted by ExPASy molecular weight calculator tool, an online program that estimates the molecular mass of the protein based on molecular weight of each amino acid residues included into the protein (Sahinkaya *et al.,* 2019). DyP type peroxidases typically have been reported to have molecular weights in the range of 40-67kDa (Chauhan *et al.,* 2020; Xu *et al.,* 2021; Lauber *et al.,* 2017), therefore B-38773 having a molecular weight of 46kDa falls within the range for an enzyme of the DyP-type peroxidase family.

The purification process required significant optimization as pure protein could not be obtained by IMAC despite varying imidazole concentration, pH of buffers, time, amount of sample loaded, etc. Contaminating protein bands corresponding to a protein of size approximately 41kDa was observed co-eluting alongside our protein of interest (Figure 4.5). This protein size coincides with the DnaJ protein of *E. coli* which has been identified as a common protein that co-purifies during IMAC purification (Graslund *et al.,* 2008). The DnaJ protein revealed has 10 histidine residues in the 376 amino acid residues and a predicted pI of 7.98 (Riley *et al.,* 2005; Hayashi *et al.,* 2006) which probably aided its ability to be bound to the purification column and be co-eluted in the presence of high imidazole concentration in the wash and elution buffers just as with our protein of interest. Subsequently, purification of B-38773 which has a predicted PI of 5.85 was achieved with ion exchange chromatography (Figure 4.6). The principle of buffer exchange is to replace one set of buffers in a solution of soluble protein with another (Phillips and Signs, 2005). This is easily accomplished by size exclusion chromatography (SEC) allowing small molecules, such as salts, to be efficiently separated from higher molecular weight substances of interest, such as proteins/enzymes or by dialysis (Neeraas, 2019). In this study, all buffer exchanges were performed by size exclusion chromatography using a PD-10 column. Buffer exchange, as well as concentration had no consequence on the purity and molecular weight of the protein (Figure 4.7) as single intact bands were seen following analysis by SDS-PAGE and western blot.

Expressed proteins were quantified using microliter amounts by the Bradford assay (Bradford, 1976) within the linear range of the assay depicted in (Figure 4.8) indicating sensitivity of Bradford as an assay for quantification of mg amounts of protein. The transformation of the dark brown dye reagent to a blue colour is indicative of a positive reaction and the intensity of the blue colour is said to be directly proportional to the concentration of protein within the sample (Bradford, 1976).

DyPs have been reported to represent both general peroxidase and dye-decolourised activity as evidenced in their ability to oxidise non-phenolic methoxylated aromatics, manganese, high redox synthetic azo and anthraquinone dyes (Datta *et al.,* 2017; Liu *et al.,* 2017), and there is increasing discovery of bacterial DyPs from *R. jostii RHA1,*

DyP2 from *Amycolatopsis sp*. *75iv2* and *AauDyPs* from *A. auricula-judae* that have shown noteworthy evidence for lignin degrading potential (Rahmanpour and Bugg, 2015; Ahmad *et al.,* 2011; Chen *et al.,* 2015, Datta *et al.,* 2017).

Here, we used the typical peroxidase substrate ABTS, the anthraquinone dye RB19 and alkaline kraft lignin to assay for the peroxidase, dye decolourising and lignin degrading potential of our recombinant putative DyP and limited all assays to spectrophotometric determinations considering that this activity study is only a preliminary investigation. From our results of observed reaction velocity determination presented in Figure 4.9, we could only observe linear relationships over time for ABTS and RB19 with our enzyme B-38773 but not for kraft lignin. Only the reaction with the standard enzyme HRP showed a linear trend with kraft lignin as with the other two substrates, indicating that our protocols were working fine and the lack of linearity in the kraft lignin reaction with our enzyme was due to lack of activity. Therefore, no further characterisation of our enzyme was attempted with kraft lignin as substrate.

ABTS is the most widely patronised substrate used in the assay for peroxidase activity of both crude and pure proteins in the presence of H_2O_2 due to its non-toxic nature (Rodríguez-López *et al.,* 2000; Groome N, 1980; Yang *et al.,* 2018; Xu et al., 2021; Childs and Bradsley, 1975). When ABTS is oxidized by hydrogen peroxide via the peroxidase-catalyzed reaction mechanism, it yields the corresponding cation radical ($ABTS^{\bullet+}$) which has a blue-green colour, the absorbance of which corresponds to the amount of the $ABTS^{\bullet+}$ released and can be monitored at 420nm (Duan *et al.,* 2018; Liu *et al.,* 2017; de Gonzalo *et al.,* 2016; Neeraas, 2019).

The pH and temperature profiles and optima, kinetic parameters, and specific activity of protein B-38773 were determined using ABTS as substrate. The optimum pH of B-38773 for the oxidation of ABTS was found to be 4.0. with over 10% activity seen at Ph 2, 3 and 5 (Figure 4.10A) which agrees with other studies that have reported that most DyPs are most active at acidic pH with isoelectric points in the range 3-5 (Liu *et al.,* 2017, Dhankhar *et al.,* 2020). For example, the optimal pH observed with BsDyP, PfDyP, and PpDyP based on ABTS oxidation was mainly distributed between pH levels 2.0–4.0, 4.0–5.0, and 3.0–5.5, while the optimum pH for each varied (Santos *et al.,* 2014; Xu *et al.,* 2021). Also, optimum pH values of 5.5 for DyP1B (Pour and Bugg,

2015), 3.5 for TfuDyP (Bloois *et al.,* 2010), 4 for DyPT1 (Sahinkaya *et al.,* 2019), 3.8 for rPsaDyP (Lauber *et al.,* 2017), 3.5 for *Il*-DyP4 have all been reported (Hofrichter *et al.,* 2010; Sahinkaya *et al.,* 2019; Lauber *et al.,* 2017; Janusz *et al.,* 2017; Colpa *et al.,* 2014; Xu *et al.,* 2021). The presence of aspartic acid and arginine in its heme pocket (the GXXDG motif) that function as acid-base catalyst and involved in the formation of compound I differentiates DyPs from other peroxidases from the plant superfamily and this difference has been reported to be responsible for the catalytic activity of DyPs in acidic pH (Chauhan 2020, Rahmanpour and Bugg 2015; Welinder, 1992; Sugano *et al.,* 2007; Santos *et al.,* 2014).

B-38773 showed activity with ABTS of over 20% at all temperatures tested but the optimum temperature was 40°C (figure 4.10B) which falls within the mostly 40-60°C range reported for most DyPs (Xu *et al.,* 2021; Colpa *et al.,* 2014) and is same as for DyP1B and DyPA (Santos *et al.,* 2014), although a significant 86% activity was also observed at 30°C with B-38773.

Literature reviews indicate highly variable reports of kinetic parameters for the DyPs from various sources. B-38773 exhibited Michealis-Menten like behaviour against ABTS showing that the enzyme can catalyse the oxidation of the substrate to its cation radical. We recorded V_{max} of 3.7µmol/min. K_m value was 1.089mM (Figure 4.11), same as for PflDyP1B from *Pseudomonas fluorescens* (Rashid and Bugg, 2021) and similar to those observed with DyP1B, DyP2B and DyPA from *Pseudomonas fluorescens* and also TfuDyP from *Thermobifida fusca* (Pour and Bugg, 2015) depicting high binding affinity of B-38773 towards ABTS. B-38773 had a turnover number (K_{cat}) of $540.9S^{-1}$ and showed catalytic efficiency of 4.96×10^5 $M^{-1}S^{-1}$ with ABTS, slightly higher than what was reported for some characterised bacterial Dyps (Qin *et al.,* 2018; Bloois *et al.,* 2010; Sugano, 2009; Liers *et al.,* 2010; Zubieta *et al.,* 2007). Although bacterial DyPs are generally known to possess lower oxidizing ability than fungal DyPs (de Gonzalo *et al.,* 2016), the catalytic efficiency of B-38773 was close to that of the fungal *Il*DyP from Irpex lacteus (Qin *et al.,* 2018). The specific activity of B-38773 which is a measure of the purity of the enzyme was determined as $12.9Umg^{-1}$.

Dye decolourising activity of B-38773 was assessed by determining the percentage decolourisation of the anthraquinone dye RB19 over time. DyPs have been shown to

have different decolourisation effects on different types of dyes and higher efficiencies on anthraquinone dyes which are known to be more difficult to degrade by other general peroxidases (Chauhan, 2020; Chen *et al.,* 2015; Duan *et al.,* 2018). It was observed that the enzyme decolourised the dye with increased efficiency over time at pH 4 as seen in figure 4.12. The decolourisation efficiency (15.9% after 10 minutes incubation) can best be described as “moderate” when compared to between 70-97% rates observed with other bacterial Dyps and HRP (Celebi *et al.,* 2013; Blooise *et al.,* 2010; Uchida *et al.,* 2015; Qin *et al.,* 2018; Sahinkaya *et al.,* 2019). However, we believe that the incubation time of 10 minutes as against the much longer times used in the other studies may be partly responsible for this, and better decolourisation efficiency could be recorded for B-38773 with RB19 if the reaction time is increased.

After studying typical peroxidation and dye decolourising activities, the capability of B-38773 to oxidize and degrade lignin was evaluated using Kraft lignin. However, Despite the reports and mounting experimental evidence of bacterial DyPs showing potential for KL degradation (Yang *et al.,* 2018; Riyadi *et al.,* 2020; Chen *et al.,* 2012; Loncar *et al.,* 2019; Catucci *et al.,* 2020; Zhu *et al.,* 2017; Ahmad *et al.,* 2011; Rahmanpour *et al.,* 2016), we found it disappointing that our enzyme B-38773 showed no activity towards kraft lignin despite several troubleshooting efforts where we varied the temperature, pH, amount of enzyme and substrate, solvent for dissolving KL, reaction time and volume etc, from different spectrophotometry based published protocols that have been reported to be effective with other DyPs while monitoring reactivity at 465nm (Ahmad *et al.,* 2011; Raj *et al.,* 2007; Loncar *et al.,* 2019; Rhamanpour and Bugg, 2015). With more optimisations, we eventually realised a protocol (Guo *et al.,* 2021) whereby we observed an increase in absorbance with KL and HRP monitoring at 465nm, but no activity was observed with our enzyme B-38773 as shown in figure 4.9 (3A and 3B). However, it is not totally abnormal that our DyP showed no activity against a polymeric lignin substrate such as kraft lignin as the general understanding of how DyPs attack polymeric lignin is still very incomplete (Bugg *et al.,* 2020) and they very often demonstrate activity towards phenolic compounds but not always on non-phenolic or more complicated lignin molecules (Catucci *et al.,* 2020). Of the three DyPs from *Pseudomonas fluorescens* Pf-5 overexpressed and characterised by Pour and

Bugg, all were active against different peroxidase substrates, but only DyP1B showed activity for the oxidation of kraft lignin and was suggested that the reactivity of DyP1B with kraft lignin might be the result of oxidation of hydroxyl groups present in lignin monomers that leads to formation of quinone products that absorbs in 400 nm region. No activity was observed for DyP2B and DyPA against kraft lignin (Pour and Bugg, 2015). However, if time had permitted, more experiments using other polymeric lignin and/or lignin dimer model compounds and employing other analytical methods could have been attempted.

4.6 Conclusion

From our findings in this chapter, we have demonstrated the successful application of functional metagenomics in discovery of novel enzymes by heterologous expression of a metagenome derived gene sequence producing its recombinant protein product. We have further validated the functional annotation analysis beyond just successful PCR amplification to confirming its predicted function via activity assays. Unfortunately, we could not exhaustively exploit all analysis to precisely conclude if our recombinant protein B-38773 has direct lignin degrading abilities especially because our main aim is the identification of lignin degrading enzymes from the gut metagenome of the African palm weevil. However, the evidence we have presented show that B-38773 is a typical dye decolourising peroxidase based on its ability to oxidise the classical peroxidase substrate ABTS, the decolourisation of the anthraquinone dye RB19 and the pH and temperature profiles and kinetic parameters that are consistent with other characterised bacterial DyPs. Therefore, B-38773 can find potential biotechnological applications in dealing with environmental problems such as bioremediation of waste, enzymatic whitening of whey-containing foods and beverages and as antimicrobial (pro)drug targets.

On a positive note, though, the data from the reservoir of metagenome derived putative lignin degrading genes is a valuable resource that can be explored in future studies from which more genes can be selected and expressed and tested for lignin degrading activity.

Chapter 5: Thesis summary, contribution to knowledge and future research

5.1 Thesis summary

The exploitation of lignocellulosic biomass as a potential fossil fuel replacement to curb global over reliance on fossil fuels for the production of biofuels and biobased materials which are more cost effective, environmentally friendly, and sustainable has been largely constrained by its recalcitrance to degradation (Chan *et al.,* 2020; Ekas *et al.,* 2019; Takkellapati *et al.,* 2018). To circumvent the currently employed pre-treatment methods that involve the use of harsh chemicals and require high energy input, biological methods that leverage on the natural ability of microorganisms to produce enzymes that facilitate the bioconversion of lignocellulose are preferred and sought after. Furthermore, most enzymes already identified are of fungal origin and have not been applied at industrial scale to achieve the efficient degradation of the aromatic lignin polymer which is the component primarily responsible for the complexity and resistance of lignocellulose to degradation. It is therefore believed that discovery of new, highly active, highly stable and end-product-inhibition resistant CAZymes of bacterial origin could revolutionise the production of biobased products from lignocellulosic biomass at lower cost and higher efficiency compared to what is obtained currently with fossil fuels. To that end many studies have been carried out in search of novel highly active bacterial lignin degrading enzymes present in nature, especially from the guts of herbivorous mammals and insects as reported in the following reviews (Chauhan, 2020; Chukwuma *et al.,* 2021; de Gonzalo *et al.,* 2016; Bugg *et al.,* 2020). Therefore, the aim of this research was to explore the gut metagenome of the African palm weevil to mine lignin degrading genes and enzymes in line with the need identified, and the research efforts reviewed above using a combination of microbial metagenomic, molecular biology and biochemical techniques. We choose the African palm weevil because it is a pest of palm trees that excavates and lives its whole life cycle (Thomas and Dimkpa, 2016, Montagna *et al.,* 2015) in the high lignin containing trunks of different palm species (See Table 1.7). We collected the larvae of APW from Ejekimomi forest in Amukpe village of Delta state, Nigeria. Prior to field collection, we performed an experiment to compare the efficacies of the

commonly used ethanol and NAP buffer to determine which of them we could use to best preserve the quality and quantity of insect gut DNA in storage. NAP buffer showed better parameters and coupled with it being safer and more convenient for transport on a commercial flight, we chose it as the preservative for field collection of insects. As a step towards achieving our aim, we performed a preliminary bacterial community profiling of the larval guts of the APW using the easier and less expensive 16S rRNA sequencing in order to ascertain if the bacteria which will produce our sought-after lignin degrading enzymes are present within the APW gut. We compartmentalised the gut into segments (foregut, midgut, and hindgut) to gain insight into which gut compartment of the larvae harboured more abundant and diverse bacteria associated with lignin degradation. Having identified a large number of lignin degrading bacteria in the gut, we proceeded to shotgun sequencing of the whole gut metagenome library prepared from bacteria enriched DNA in order to specifically target bacterial genes and to lessen the burden of bioinformatic analysis due to host DNA contamination. Reads were assembled and functionally annotated, CAZy genes were identified, and 3 genes (2 deferrochelatase/ peroxidases and a polyphenol oxidase) were selected from the AA family members following a BLASTp and conserved domain search and match to characterised dye decolourising peroxidase and laccase proteins in the UniProt/Swiss-Prot database. We also taxonomically classified assembled contigs from the shotgun metagenomic sequencing and compared to the profile obtained using 16S rRNA sequencing and analysis method. The 3 genes we selected (A-30342, B-38773, and C-08687) were all successfully amplified by PCR from whole genome amplified products of the original DNA that was sequenced. We attempted to clone all 3 genes into pET151 vector but only B-38773 was successfully cloned as confirmed by colony PCR and furthermore by restriction digest and sanger sequencing. The gene construct of B-38773 was heterologously expressed in BL21 star *E.coli* cells. The recombinant protein was fractionated and purified by a two-step chromatography process initially employing IMAC and subsequently IEX after which pure protein fraction of the predicted size of 46kDa was obtained (Figure 4.7). We tested the dye decolourising activity of the recombinant protein by assaying against the peroxidase substrate-ABTS, the anthraquinone dye- RB19 and the commercially prepared polymeric lignin-

Alkaline kraft lignin using horseradish peroxidase (HRP) as standard, and we determined the enzymatic characteristics (optimum pH, optimum temperature, specific activity, specific activity, V_{max}, K_m, K_{cat} and catalytic efficiency) of the protein using ABTS as substrate.

5.2 Main findings and contribution to knowledge

Preservation of larval samples for a later time extraction of gut DNA in NAP buffer yielded higher quantity and quality of DNA compared to samples preserved in ethanol (Table 2.6). This finding is very significant and contributes valuable information showing evidence in support of using NAP buffer as a cheap and safe buffer for storage of field collected samples. The effectiveness of this buffer has been described by other studies (Camacho-Sanchez *et al.,* 2013; Kilpatrick 2002) but none has directly compared it to ethanol which is the predominantly used solvent that is comparatively cheap.

The most dominant bacterial phyla identified in the APW gut by 16S rRNA amplicon sequencing are *Firmicutes* (63.7%), *Proteobacteria* (33.2%), *Bacteriodetes* (1.9%) and *Actinobacteria* (1.0%) (Table 2.10). The genera that were predominant across all gut segments are *Enterococcus, Lactococcus, Shimwellia, Lelliotia, Klebsiella* and *Enterobacter*. 16 genera implicated in lignin degradation were identified (Figure 2.10A) and the foregut had the most diverse and highest abundance of these lignin degrading phylotypes (Figure 2.13). This finding proves that the APW gut harbours a diverse community of bacteria with an abundance of ligninolytic genera that most likely play significant roles in producing enzymes that facilitate the weevil's ability to metabolise and obtain the nourishment it requires for its survival within the high lignin environment of the palm trunk where it lives and thrives. In this study we also re-confirmed the presence of similar bacterial phyla using shotgun sequencing method (*Proteobacteria*-74%, *Firmicutes*-4%, and *Actinobacteria*-9%) and genera (*Klebsiella*, *Citrobacter, Yokenella, Enterococcus, Enterobacter, Bacillus*)(Figures 3.4A and 3.4B) with those previously seen in chapter 2 where 16S sequencing method was used, giving confidence to identifications made and suggesting that some of these microbes present in the gut may be permanent members of the gut microbiome. These predominant

phyla have also been identified in the guts of most herbivores (Cardoso *et al.,* 2012; Franzini *et al.,* 2016; Joynson *et al.,* 2017) and within other lignocellulose degrading environments (Kanokratana *et al.,* 2015; Ransom-Jones *et al.,* 2017; Mhuantong *et al.,* 2015) suggesting a more general association of these phyla with facilitation of lignocellulose degradation. To the best of our knowledge, this research serves as the first attempt at describing the microbiome associated with the gut of *R. phoenicis*, and our findings agree with preliminary studies of the gut bacterial communities associated with a close relative of *R. phoenicis*; the red palm weevil (*Rhynchophorus ferrugineus*) where *Klebsiella, Lactococcus, Enterococcus,* and *Enterobacter* are the most recurring bacterial genera identified (Angzzass *et al.,* 2016; Montagna *et al.,* 2015, Jia *et al.,* 2013; Muhammad *et al.,* 2017, Tagliavia *et al.,* 2014, Valzano *et al.,* 2012). We also suggest from the data that the foregut could be the primary site of lignin catabolism as it harboured more diverse and abundant ligninolytic bacterial genera (Figure 2.13). The discovery of ligninolytic bacteria was a positive confirmation towards validating the hypothesis that this insect could be a potential source of lignin degrading enzymes and thus justifies the progression to carrying out a whole metagenomic sequencing in order to mine lignin degrading genes.

From the shotgun metagenomic sequencing data, we also carried out in depth analysis of the functional capabilities of the microbes present in the gut with particular interest in their lignin degrading ability. We identified 15,892 genes with putative functions assigned and 43,913 (72% of total ORFs identified) hypothetical genes (Table 3.8). A total of 1,141 genes from the function assigned genes had EC numbers matching proteins in different families of the CAZy database, out of which 249 auxiliary activity linked genes are thought to be involved in the degradation of lignin (Table 3.9). Although a large number of identified genes were hypothetical, there was an abundance of CAZy genes indicating that the APW gut harbours a consortium of genes with not only ligninolytic potential (as seen from numbers of AA family) but capable of breaking down the other polysaccharide components of lignocellulose. This indicates that our hypothesis made after identification of ligninolytic bacteria was correct, that the *APW* gut environment was harbouring bacteria that potentially contributed greatly to lignocellulose/ lignin degradation. We make bold to say that this is the first time the

APW gut microbiome has been subjected to such high-resolution analysis for both information of the composition of the gut bacterial consortium and its lignocellulose/ lignin degrading capability in particular.

We also validated the annotation predicted gene sequences through amplification of these from whole genome amplified metagenomic sample. Successful amplification of 3 glycoside hydrolase genes and observation that the predicted sequences truly exist in nature gives strength to the bioinformatics analyses and the expression of the B-38773 gene, shows that these homology-based annotation methods are capable of identifying novel functioning gene sequences and proves the possibility of bioprospecting novel enzymes from understudied environments using a metagenomics approach.

The recombinant protein B-38773 exhibited typical peroxidase and dye decolorising activity against the substrates ABTS and RB19 respectively but showed no ligninolytic activity against kraft lignin (figure 4.9). This finding may point to our enzyme acting more as a deferrochelatase/ peroxidase than a dye decolourising peroxidase as many enzymes could have multiple functions (Le, 2021). The enzyme showed optimum activity at pH 4 and 40°C temperature (Figure 4.10) and may be suited for some industrial applications. Specific activity was estimated as 12.9U/mg, Km of 1.089mM, Vmax of 3.678µmol min^{-1}, K_{cat} of 540.9S^{-1} and catalytic efficiency of 496.7 indicating a high affinity for its substrate and good catalytic activity.

5.3 Future Research

From the outcomes of this study, several areas of future work can be identified and have been summarised under the two themes below.

5.3.1 More in-depth description of structure and function of APW gut bacterial community.

Being the very first exploration of its kind into the APW gut, the sequence information obtained (both 16S and shotgun) can be further analysed for a more robust description of the bacterial community structure and their associated functions.

- The full functional profile of the identified genes can be analysed in order to determine what specific functions the bacterial community in the APW gut help their host to perform and how they influence other characteristics exhibited by their host. This analysis is underway and may be included in this thesis before the examination.
- Determination of host-microbiome relationship by identifying the bacterial sources of identified genes.

5.3.2 Exploitation of metagenome predicted genes

The metagenomic library generated from functional annotation of the APW gut metagenome is a great asset to the Natalie Ferry lab and can be further explored for the following

- As a matter of priority, further activity testing of the expressed B-38773 protein against other lignin substrates and employing other methods of analysis need to be carried out to tentatively determine if the protein has any activity towards lignin as would be expected of a typical dye decolourising peroxidase. This research has focused on selecting and expressing members of the lignin modifying group of enzymes (peroxidases and laccases) which are oxidative enzymes that are involved directly in breaking the C-O and C-C bonds of lignin to yield lower molecular weight aromatic compounds.
- Secondly, troubleshooting of the other two genes (A-30342 and C-08687) which were amplified in this research but not successfully cloned should be performed to produce their recombinant products and subsequently test their predicted laccase and dye decolourising activities respectively against suitable substrates and their activities compared against standard enzymes that are currently commercially available.
- With over 200 lignin associated genes identified (Genes predicted to belong to both lignin modifying enzymes and lignin degrading accessory enzymes groups) that are required for complete deconstruction of the lignin polymer, continuous selection, amplification and cloning of these genes presents a

great opportunity for identification of highly active cocktail of ligninolytic enzymes that could be applied to act in synergy towards the pre-treatment and bioconversion of lignocellulose.

- Further investigation should also be carried out into the high numbers of other CAZy class genes identified in this study such as the glycoside hydrolases, polysaccharide lyases and carbohydrate binding modules. The presence of a vast majority of the identified genes being assigned hypothetical functions presents an inexhaustible library where more genes can be selected and studied with great opportunity for the discovery of novel enzymes and functionalities.

Furthermore, the entire process of bioprospecting by target identification, metagenomic DNA extraction, bioinformatics analysis and eventual expression of enzymes of interest could be replicated for the total APW gut microbiota (not being restricted to bacteria as focused by this research), other eukaryote guts or other environments as validated methods that have yielded a great deal of information here. Unfortunately, we couldn’t advance to performing more assays to validate the lignin degrading capability of our expressed enzyme (B-38773) due to time constraint.

References

Abas, N., Kalair, A., & Khan, N. (2015). Review of fossil fuels and future energy technologies. *Futures*, *69*, 31- 49.

Abdullah, B., Muhammad, S. A. F. A. S., Shokravi, Z., Ismail, S., Kassim, K. A., Mahmood, A. N., & Aziz, M. M. A. (2019). Fourth generation biofuel: A review on risks and mitigation strategies. *Renewable and sustainable energy reviews*, *107*, 37-50.

Acharya, S., & Chaudhary, A. (2012). Bioprospecting thermophiles for cellulase production: a review. *Brazilian Journal of Microbiology*, *43*(3), 844-856.

Adamo, M., Comtet-Marre, S., Büttner, E., Kellner, H., Luis, P., Vallon, L., ... & Marmeisse, R. (2022). Fungal dye-decolorizing peroxidase diversity: roles in either intra-or extracellular processes. *Applied Microbiology and Biotechnology*, 1-15.

Afreen, S., Shamsi, T. N., Baig, M. A., Ahmad, N., Fatima, S., Qureshi, M. I., ... & Fatma, T. (2017). A novel multicopper oxidase (laccase) from cyanobacteria: purification, characterization with potential in the decolorization of anthraquinonic dye. *PloS one*, *12*(4), e0175144.

Ahmad, M (2010). Development of Novel Assays for Lignin Breakdown and Identification of a New Bacterial Lignin Degrading Enzyme (Doctoral thesis, University of Warwick). http://go.warwick.ac.uk/wrap/4477

Ahmad, M., Roberts, J. N., Hardiman, E. M., Singh, R., Eltis, L. D., & Bugg, T. D. (2011). Identification of DypB from Rhodococcus jostii RHA1 as a lignin peroxidase. *Biochemistry*, *50*(23), 5096-5107.

Ajai, O. (1997). Access to Genetic Resources and Biotechnology Regulation in Nigeria. *Rev. Eur. Comp. & Int'l Envtl. L.*, *6*, 42.

Ajanovic, A. (2011). Biofuels versus food production: Does biofuels production increase food prices?. *Energy*, *36*(4), 2070-2076.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). Garland Science. *Molecular Biology of the Cell*.

Alcon-Giner, C., Caim, S., Mitra, S., Ketskemety, J., Wegmann, U., Wain, J., ... J. (2017). Optimisation of 16S rRNA gut microbiota profiling of extremely low birth weight infants. *BMC genomics*, *18*(1), 841.

Alex Barret, 2018 https://bioplasticsnews.com/2018/06/01/uk-top-10-biobased-chemicals-green- chemicals/

Ali, S. S., Al-Tohamy, R., Sun, J., Wu, J., & Huizi, L. (2019). Screening and construction of a novel microbial consortium SSA-6 enriched from the gut symbionts of wood-feeding termite, Coptotermes formosanus and its biomass-based biorefineries. *Fuel*, *236*, 1128- 1145.

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389-3402.

Al-Zuhair, S., Abualreesh, M., Ahmed, K., & Abdul Razak, A. (2015). Enzymatic delignification of biomass for enhanced fermentable sugars production. *Energy Technology*, *3*(2), 121-127.

Ameh, A.O., Ojo, A.A. & Gaiya, J. (2016) Preliminary investigation into the synthesis of furfural from sugarcane bagasse. *Cellulose*, *33*, pp.34-06.

Ameri, R., Laville, E., Potocki-Véronèse, G., Trabelsi, S., Mezghani, M., Elgharbi, F., & Bejar, S. (2018). Two new gene clusters involved in the degradation of plant cell wall from the fecal microbiota of Tunisian dromedary. *Plos one*, *13*(3), e0194621.

Ammar, M., Khiari, R., Berrima, B., Belgacem, M.N. & Elaloui, E. (2014) Isolation and characterization of lignin from Stipa tenacissima L. and Phoenix dactylifera. *Cellul Chem Technol*, 48, pp.255-263.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Angzzas, S. M. K., Ashuvila, M. A., & Dayang, N. F. A. Z. (2016, March). Potential lignin degraders isolated from the gut of rhynchophorus ferrugineus. In *International*

Conference on Mechanics, Materials and Structural Engineering (ICMMSE 2016). Atlantis Press.

Ansari, K. B., Arora, J. S., Chew, J. W., Dauenhauer, P. J., & Mushrif, S. H. (2019). Fast pyrolysis of cellulose, hemicellulose, and lignin: effect of operating temperature on bio-oil yield and composition and insights into the intrinsic pyrolysis chemistry. *Industrial & Engineering Chemistry Research*, *58*(35), 15838-15852.

Antwis, R. E., Griffiths, S. M., Harrison, X. A., Aranega-Bou, P., Arce, A., Bettridge, A. S., ... & Sutherland, W. J. (2017). Fifty important research questions in microbial ecology. *FEMS Microbiology Ecology*, *93*(5), fix044.

Antwis, R. E., Harrison, X. A., & Cox, M. J. (Eds.). (2020). *Microbiomes of soils, plants and animals: an integrated approach*. Cambridge University Press.

Arnau, J., Yaver, D., & Hjort, C. M. (2020). Strategies and challenges for the development of industrial enzymes using fungal cell factories. In *Grand challenges in fungal biotechnology* (pp. 179-210). Springer, Cham.

Aro, E. M. (2016). From first generation biofuels to advanced solar biofuels. *Ambio*, *45*(1), 24-31.

Arumugam, N. N., & Mahalingam, P. U. (2015). Lignocellulose plant biomass; an emerging alternative fuel resource. *ResearchGate*, *49*, 291-295.

Arumugam, N., Kalavathi, P., & Mahalingam, P. U. (2014). Lignin database for diversity of lignin degrading microbial enzymes (LD2L). *Research in Biotechnology*, *5*(1), 13-18.

Auer, L., Lazuka, A., Sillam-Dussès, D., Miambi, E., O'Donohue, M., & Hernandez-Raquet, G. (2017). Uncovering the potential of termite gut microbiome for lignocellulose bioconversion in anaerobic batch bioreactors. *Frontiers in microbiology*, *8*, 2623.

Avanthi, A., & Banerjee, R. (2016). A strategic laccase mediated lignin degradation of lignocellulosic feedstocks for ethanol production. *Industrial Crops and Products*, *92*, 174-185.

Ayadi, M., Sarma, S.J., Pachapur, V.L., Brar, S.K. & Cheikh, R.B. (2016) History and global policy of biofuels. In *Green Fuels Technology* (pp. 1-14). Springer, Gewerbestrasse, Switzerland.

Bai, D. P., Lin, X. Y., Hu, Y. Q., Chen, Z. Z., Chen, L., Huang, Y. F., ... & Li, J. (2021). Metagenomics approach to identify lignocellulose-degrading enzymes in the gut microbiota of the Chinese bamboo rat cecum. *Electronic Journal of Biotechnology*, *50*, 29-36.

Baker, G. C., Smith, J. J., & Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *Journal of microbiological methods*, *55*(3), 541-555.

Baldrian, P., & López-Mondéjar, R. (2014). Microbial genomics, transcriptomics and proteomics: new discoveries in decomposition research using complementary methods. *Applied microbiology and biotechnology*, *98*(4), 1531-1537.

Bamidele, O., Ajele, J., Kolawole, A., & Oluwafemi, A. (2013). Changes in the tissue antioxidant enzyme activities of palm weevil (Rynchophorous phoenicis) larva by the action of 2, 2-dichlorovinyl dimethyl phosphate. *African Journal of Biochemistry Research*, *7*(7), 128-137.

Banjo, A. D., Lawal, O. A., & Songonuga, E. A. (2006). The nutritional value of fourteen species of edible insects in south-western Nigeria. *African Journal of Biotechnology*, *5*(3), 298- 301.

Barakat, A., De Vries, H., & Rouau, X. (2013). Dry fractionation process as an important step in current and future lignocellulose biorefineries: a review. *Bioresource Technology*, *134*, 362-373.

Bartels, J. R., Pate, M. B., & Olson, N. K. (2010). An economic survey of hydrogen production from conventional and alternative energy sources. *International journal of hydrogen energy*, *35*(16), 8371-8384.

Batista-García, R. A., del Rayo Sánchez-Carbente, M., Talia, P., Jackson, S. A., O'Leary, N. D., Dobson, A. D., & Folch-Mallol, J. L. (2016). From lignocellulosic metagenomes to lignocellulolytic genes: trends, challenges, and future prospects. *Biofuels, Bioproducts and Biorefining*, *10*(6), 864-882.

Behera, S., Singh, R., Arora, R., Sharma, N. K., Shukla, M., & Kumar, S. (2015). Scope of algae as third generation biofuels. *Frontiers in bioengineering and biotechnology*, *2*, 90.

Beloqui, A., Pita, M., Polaina, J., Martínez-Arias, A., Golyshina, O. V., Zumárraga, M., metagenome expression library of bovine rumen: biochemical properties, structural analysis, and phylogenetic relationships. *Journal of Biological Chemistry*, *281*(32), 22933-22942.

Ben Guerrero, E., Soria, M., Salvador, R., Ceja-Navarro, J. A., Campos, E., Brodie, E. L., & Talia, P. (2016). Effect of different lignocellulosic diets on bacterial microbiota and hydrolytic enzyme activities in the gut of the cotton boll weevil (Anthonomus grandis). *Frontiers in microbiology*, *7*, 2093.

Bendahou, A., Dufresne, A., Kaddami, H. & Habibi, Y. (2007) Isolation and structural characterization of hemicelluloses from palm of Phoenix dactylifera L. *Carbohydrate Polymers*, 68(3), pp.601- 608.

Bensah, E.C., Kádár, Z. & Mensah, M.Y. (2015) Ethanol production from hydrothermally- treated biomass from West Africa. *BioResources*, 10(4), pp.6522-6537.

Ben-Yosef, M., Pasternak, Z., Jurkevitch, E., & Yuval, B. (2014). Symbiotic bacteria enable olive flies (B actrocera oleae) to exploit intractable sources of nitrogen. *Journal of evolutionary biology*, *27*(12), 2695-2705.

Berg, A. H., Combs, T. P., & Scherer, P. E. (2002). ACRP30/adiponectin: an adipokine regulating glucose and lipid metabolism. *Trends in Endocrinology & Metabolism*, *13*(2), 84-89.

Bergmann, J. C., Costa, O. Y. A., Gladden, J. M., Singer, S., Heins, R., D'haeseleer, P., ... & Quirino, B. F. (2014). Discovery of two novel β-glucosidases from an Amazon soil metagenomic library. *FEMS Microbiology Letters*, *351*(2), 147-155.

Berlyn, M. K., Plunkett, G., Wishart, D. S., Nori, H., Riley, M., Rudd, K. E., ... & Chaudhuri, R. R. (2006). Escherichia coli K-12: A cooperatively developed annotation snapshot-2005.

Bhalla, A., Bischoff, K. M., Uppugundla, N., Balan, V., & Sani, R. K. (2014). Novel thermostable endo-xylanase cloned and expressed from bacterium Geobacillus sp. WSUCF1. *Bioresource technology*, *165*, 314-318.

Bodor, A., Bounedjoum, N., Vincze, G. E., Erdeiné Kis, Á., Laczi, K., Bende, G., ... & Rákhely, G. (2020). Challenges of unculturable bacteria: environmental perspectives. *Reviews in Environmental Science and Bio/Technology*, *19*(1), 1-22.

Borgström, E., Paterlini, M., Mold, J. E., Frisen, J., & Lundeberg, J. (2017). Comparison of whole genome amplification techniques for human single cell exome sequencing. *PloS one*, *12*(2), e0171566.

Bozorov, T. A., Rasulov, B. A., & Zhang, D. (2019). Characterization of the gut microbiota of invasive Agrilus mali Matsumara (Coleoptera: Buprestidae) using high-throughput sequencing: uncovering plant cell-wall degrading bacteria. *Scientific reports*, *9*(1), 1-12.

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical biochemistry*, *72*(1-2), 248-254.

Bragg, L., & Tyson, G. W. (2014). Metagenomics using next-generation sequencing. In *Environmental microbiology* (pp. 183-201). Humana Press, Totowa, NJ.

Brink, D. P., Ravi, K., Lidén, G., & Gorwa-Grauslund, M. F. (2019). Mapping the diversity of microbial lignin catabolism: experiences from the eLignin database. *Applied microbiology and biotechnology*, *103*(10), 3979-4002.

Brown, M. E., & Chang, M. C. (2014). Exploring bacterial lignin degradation. *Current opinion in chemical biology*, *19*, 1-7.

Brown, M. E., Barros, T., & Chang, M. C. (2012). Identification and characterization of a multifunctional dye peroxidase from a lignin-reactive bacterium. *ACS chemical biology*, *7*(12), 2074-2081.

Brune, A. (2014). Symbiotic digestion of lignocellulose in termite guts. *Nature Reviews Microbiology*, *12*(3), 168.

Bugg, T. D., & Rahmanpour, R. (2015). Enzymatic conversion of lignin into renewable chemicals. *Current opinion in chemical biology*, *29*, 10-17.

Bugg, T. D., Ahmad, M., Hardiman, E. M., & Rahmanpour, R. (2011)a. Pathways for degradation of lignin in bacteria and fungi. *Natural Product Reports*, *28*(12), 1883- 1896.

Bugg, T. D., Ahmad, M., Hardiman, E. M., & Singh, R. (2011)b. The emerging role for bacteria in lignin degradation and bio-product formation. *Current Opinion in Biotechnology*, *22*(3), 394- 400.

Bugg, T. D., Williamson, J. J., & Rashid, G. M. (2020). Bacterial enzymes for lignin depolymerisation: new biocatalysts for generation of renewable chemicals from biomass. *Current opinion in chemical biology*, *55*, 26-33.

Bugg, T. D., Williamson, J. J., & Rashid, G. M. (2020). Bacterial enzymes for lignin depolymerisation: new biocatalysts for generation of renewable chemicals from biomass. *Current opinion in chemical biology*, *55*, 26-33.

Bundhoo, Z. M. (2018). Microwave-assisted conversion of biomass and waste materials to biofuels. *Renewable and Sustainable Energy Reviews*, *82*, 1149-1177.

Bundhoo, Z. M., & Mohee, R. (2018). Ultrasound-assisted biological conversion of biomass and waste materials to biofuels: A review. *Ultrasonics sonochemistry*, *40*, 298-313.

Buraimoh, O. M., Ilori, M. O., Amund, O. O., Isanbor, C., & Michel Jr, F. C. (2017). The degradation of coniferyl alcohol and the complementary production of chlorogenic acids in the growth culture of Streptomyces albogriseolus KF977548 isolated from decaying wood residues. *Process Biochemistry*, *52*, 22-29.

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.

Cagide, C., & Castro-Sowinski, S. (2020). Technological and biochemical features of lignin-degrading enzymes: A brief review. *Environmental Sustainability*, *3*(4), 371-389.

Callahan, B.J., Sankaran, K., Fukuyama, J.A., McMurdie, P.J. & Holmes, S.P. (2016) Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. F1000Research, 5.

Camacho-Sanchez, M., Burraco, P., Gomez-Mestre, I., & Leonard, J. A. (2013). Preservation of RNA and DNA from mammal samples under field conditions. *Molecular Ecology Resources*, *13*(4), 663-673.

CamKowsari's site, 2012: (https://biofuel.webgarden.com/sections/blog/pictures-for-lignocellulose)

Campbell, C. L., Mummey, D. L., Schmidtmann, E. T., & Wilson, W. C. (2004). Culture-independent analysis of midgut microbiota in the arbovirus vector Culicoides sonorensis (Diptera: Ceratopogonidae). *Journal of Medical Entomology*, *41*(3), 340-348.

Cardoso, A. M., Cavalcante, J. J., Cantão, M. E., Thompson, C. E., Flatschart, R. B., Glogauer, A., Scapin, S. M. N., Sade, Y., Beltra˜o, P. J. M. S. I., Gerber, A. L., Martins, O. B., Garcia, E. S. de Souza, W., Vasconcelos, A. R. (2012). *PLoS One*, *7*(11), e48505.

Cassland, P., & Jönsson, L. J. (1999). Characterization of a gene encoding Trametes versicolor laccase A and improved heterologous expression in Saccharomyces cerevisiae by decreased cultivation temperature. *Applied Microbiology and Biotechnology*, *52*(3), 393-400.

Catucci, G., Valetti, F., Sadeghi, S. J., & Gilardi, G. (2020). Biochemical features of dye-decolorizing peroxidases: current impact on lignin degradation. *Biotechnology and Applied Biochemistry*, *67*(5), 751-759.

Ceballos, S. J., Yu, C., Claypool, J. T., Singer, S. W., Simmons, B. A., Thelen, M. P., ... & VanderGheynst, J. S. (2017). Development and characterization of a thermophilic, lignin degrading microbiota. *Process Biochemistry*, *63*, 193-203.

Celebi, M., Altikatoglu, M., Akdeste, Z. M., & Yıldırım, H. (2013). Determination of decolorization properties of Reactive Blue 19 dye using Horseradish Peroxidase enzyme. *Turkish Journal of Biochemistry/Turk Biyokimya Dergisi*, *38*(2).

Chan, J. C., Paice, M., & Zhang, X. (2020). Enzymatic oxidation of lignin: challenges and barriers toward practical applications. *ChemCatChem*, *12*(2), 401-425.

Chang, C. J., Wu, C. P., Lu, S. C., Chao, A. L., Ho, T. H. D., Yu, S. M., & Chao, Y. C. (2012). A novel exo- cellulase from white spotted longhorn beetle (Anoplophora malasiaca). *Insect biochemistry and molecular biology*, *42*(9), 629-636.

Chapman, R. F., Kerkut, G. A., & Gilbert, L. I. (2013). Structure of the digestive system. *Comprehensive insect physiology, biochemistry, and pharmacology*, 165-211.

Chauhan, P. S. (2020). Role of various bacterial enzymes in complete depolymerization of lignin: A review. *Biocatalysis and agricultural biotechnology*, *23*, 101498.

Chauhan, P. S., Goradia, B., & Saxena, A. (2017). Bacterial laccase: recent update on production, properties, and industrial applications. *3 Biotech*, *7*(5), 1-20.

Chen, B., Yu, T., Xie, S., Du, K., Liang, X., Lan, Y., ... & Shao, Y. (2018). Comparative shotgun metagenomic data of the silkworm Bombyx mori gut microbiome. *Scientific data*, *5*(1), 1-10.

Chen, C., & Li, T. (2016). Bacterial dye-decolorizing peroxidases: Biochemical properties and biotechnological opportunities. *Physical Sciences Reviews*, *1*(9).

Chen, C., Shrestha, R., Jia, K., Gao, P. F., Geisbrecht, B. V., Bossmann, S. H., ... & Li, P. (2015). Characterization of dye-decolorizing peroxidase (DyP) from Thermomonospora curvata reveals unique catalytic properties of A-type DyPs. *Journal of Biological Chemistry*, *290*(38), 23447-23463.

Chen, H. (2014). Chemical composition and structure of natural lignocellulose. In *Biotechnology of lignocellulose* (pp. 25-71). Springer, Dordrecht.

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., & Hwang, C. C. (2013). Effects of GC bias in next-generation- sequencing data on de novo genome assembly. *PloS one*, *8*(4), e62856.

Chen, Y. H., Chai, L. Y., Zhu, Y. H., Yang, Z. H., Zheng, Y., & Zhang, H. (2012). Biodegradation of kraft lignin by a bacterial strain Comamonas sp. B-9 isolated from eroded bamboo slips. *Journal of applied microbiology*, *112*(5), 900-906.

Chen, Z., & Wan, C. (2017). Biological valorization strategies for converting lignin into fuels and chemicals. *Renewable and Sustainable Energy Reviews*, *73*, 610-621.

Cheng, F., & Brewer, C. E. (2017). Producing jet fuel from biomass lignin: Potential pathways to alkyl- benzenes and cycloalkanes. *Renewable and Sustainable Energy Reviews*, *72*, 673-722.

Cherubini, F. (2010). The biorefinery concept: using biomass instead of oil for producing energy and chemicals. *Energy conversion and management*, *51*(7), 1412-1421.

Chew, Y. M., Lye, S., Salleh, M. M., & Yahya, A. (2018). 16S rRNA metagenomic analysis of the symbiotic community structures of bacteria in foregut, midgut, and hindgut of the wood- feeding termite Bulbitermes sp. *Symbiosis*, *76*(2), 187-197.

Chien, A., Edgar, D. B., & Trela, J. M. (1976). Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *Journal of bacteriology*, *127*(3), 1550-1557.

Childs, R. E., & Bardsley, W. G. (1975). The steady-state kinetics of peroxidase with 2, 2′-azino-di-(3-ethyl- benzthiazoline-6-sulphonic acid) as chromogen. *Biochemical journal*, *145*(1), 93-103.

Choi, J. M., Han, S. S., & Kim, H. S. (2015). Industrial applications of enzyme biocatalysis: current status and future aspects. *Biotechnology Advances*, *33*(7), 1443-1454.

Choinowski, T., Blodig, W., Winterhalter, K. H., & Piontek, K. (1999). The crystal structure of lignin peroxidase at 1.70 Å resolution reveals a hydroxy group on the Cβ of tryptophan 171: a novel radical site formed during the redox cycle. *Journal of molecular biology*, *286*(3), 809-827.

Choolaei, Z., Flick, R., Khusnutdinova, A. N., Edwards, E. A., & Yakunin, A. F. (2021). Lignin-oxidizing activity of bacterial laccases characterized using soluble substrates and polymeric lignin. *Journal of Biotechnology*, *325*, 128-137.

Chung, G. F. (2012). Effect of pests and diseases on oil palm yield. In *Palm Oil* (pp. 163-210). AOCS Press.

Coatsworth, H., Caicedo, P. A., Van Rossum, T., Ocampo, C. B., & Lowenberger, C. (2018). The composition of midgut bacteria in Aedes aegypti (Diptera: Culicidae) that are naturally susceptible or refractory to dengue viruses. *Journal of Insect Science*, *18*(6), 12.

Cole, J. R., Chai, B., Marsh, T. L., Farris, R. J., Wang, Q., Kulam, S. A., ... & Tiedje, J. M. (2003). The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic acids research*, *31*(1), 442-443.

Colombo, L. T., de Oliveira, M. N. V., Carneiro, D. G., de Souza, R. A., Alvim, M. C. T., Dos Santos, J. C., ... & Passos, F. M. L. (2016). Applying functional metagenomics to search for novel lignocellulosic enzymes in a microbial consortium derived from a thermophilic composting phase of sugarcane bagasse and cow manure. *Antonie Van Leeuwenhoek*, *109*(9), 1217-1233.

Colpa, D. I., Fraaije, M. W., & van Bloois, E. (2014). DyP-type peroxidases: a promising and versatile class of enzymes. *Journal of Industrial Microbiology and Biotechnology*, *41*(1), 1-7.

Courtois, S., Cappellano, C. M., Ball, M., Francou, F. X., Normand, P., Helynck, G., ... & Pernodet, J. L. (2003). Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Applied and environmental microbiology*, *69*(1), 49-55.

Cragg, S. M., Beckham, G. T., Bruce, N. C., Bugg, T. D., Distel, D. L., Dupree, P., ... & McQueen- Mason, S. J. (2015). Lignocellulose degradation mechanisms across the Tree of Life. *Current Opinion in Chemical Biology*, *29*, 108-119.

Czyz, Z. T., Kirsch, S., & Polzer, B. (2015). Principles of whole-genome amplification. *Whole Genome Amplification*, 1-14.

da Costa, R. R., Hu, H., Pilgaard, B., Vreeburg, S. M., Schückel, J., Pedersen, K. S., ... & Poulsen, M. (2018). Enzyme activities at different stages of plant biomass decomposition in three species of fungus-growing termites. *Applied and environmental microbiology*, *84*(5), e01815-17.

da Costa, R. R., Hu, H., Pilgaard, B., Vreeburg, S. M., Schückel, J., Pedersen, K. S., ... & Poulsen, M. (2018). Enzyme activities at different stages of plant biomass decomposition in three species of fungus-growing termites. *Applied and environmental microbiology*, *84*(5), e01815-17.

da Silva, C. G., Grelier, S., Pichavant, F., Frollini, E., & Castellan, A. (2013). Adding value to lignins isolated from sugarcane bagasse and Miscanthus. *Industrial Crops and Products*, *42*, 87-95.

Dada, N., Jupatanakul, N., Minard, G., Short, S. M., Akorli, J., & Villegas, L. M. (2021). Considerations for mosquito microbiome research from the Mosquito Microbiome Consortium. *Microbiome*, *9*(1), 1-16.

Datta, R., Kelkar, A., Baraniya, D., Molaei, A., Moulick, A., Meena, R. S., & Formanek, P. (2017). Enzymatic degradation of lignin in soil: a review. *Sustainability*, *9*(7), 1163.

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, *6*(1), 1-14.

de Gonzalo, G., Colpa, D. I., Habib, M. H., & Fraaije, M. W. (2016). Bacterial enzymes involved in lignin degradation. *Journal of Biotechnology*, *236*, 110-119.

DEFRA -Department for Environment, Food and Rural Affairs (2011). Biodiversity 2020: A strategy for England's wildlife and ecosystem services. 2-20

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*, *72*(7), 5069-5072.

Dhankhar, P., Dalal, V., Mahto, J. K., Gurjar, B. R., Tomar, S., Sharma, A. K., & Kumar, P. (2020). Characterization of dye-decolorizing peroxidase from Bacillus subtilis. *Archives of Biochemistry and Biophysics*, *693*, 108590.

Didane, D. H., Ab Wahab, A., Shamsudin, S., & Rosly, N. S. (2016). Wind as a sustainable alternative energy source in Malaysia-a review. *ARPN Journal of Engineering and Applied Sciences*, *11*(10), 6442-6449.

Dimkpa, S.O.N., Appiah, S.O., Afreh-Nuamah, K. and Yawson, G.K. (2010). The Susceptibility of Some Oil Palm *Elaeis guineensis* Jacq Progenies to *Coelaenomenodera lameensis* Berti and Mariau, (Coleoptera: Chrysomelidae). *Current Research Journal of Biological Sciences* 2(3): 168-172.

Do, T. H., Le, N. G., Dao, T. K., Nguyen, T. M. P., Le, T. L., Luu, H. L., ... & Truong, N. H. (2018). Metagenomic insights into lignocellulose-degrading genes through Illumina-based de novo sequencing of the microbiome in Vietnamese native goats' rumen. *The Journal of general and applied microbiology*, *64*(3), 108-116.

Do, T. H., Nguyen, T. T., Nguyen, T. N., Le, Q. G., Nguyen, C., Kimura, K., & Truong, N. H. (2014). Mining biomass-degrading genes through Illumina-based de novo sequencing and metagenomic analysis of free-living bacteria in the gut of the lower termite Coptotermes gestroi harvested in Vietnam. *Journal of Bioscience and Bioengineering*, *118*(6), 665-671.

Dresselhaus, M. S., & Thomas, I. L. (2001). Alternative energy technologies. *Nature*, *414*(6861), 332-337.

Duan, Z., Shen, R., Liu, B., Yao, M., & Jia, R. (2018). Comprehensive investigation of a dye-decolorizing peroxidase and a manganese peroxidase from Irpex lacteus F17, a lignin-degrading basidiomycete. Editor(s): Michael Stephenson, Energy and Climate Change,Elsevier,2018, Pages 1-26,

Edwards, J. L., Smith, D. L., Connolly, J., McDonald, J. E., Cox, M. J., Joint, I., ... & McCarthy, A. J. (2010). Identification of carbohydrate metabolism genes in

the metagenome of a marine biofilm community shown to be dominated by Gammaproteobacteria and Bacteroidetes. *Genes*, *1*(3), 371-384.

Egert, M., Wagner, B., Lemke, T., Brune, A., & Friedrich, M. W. (2003). Microbial community structure in midgut and hindgut of the humus-feeding larva of Pachnoda ephippiata (Coleoptera: Scarabaeidae). *Applied and Environmental Microbiology*, *69*(11), 6659- 6668.

Ekas, H., Deaner, M., & Alper, H. S. (2019). Recent advancements in fungal-derived fuel and chemical production and commercialization. *Current opinion in biotechnology*, *57*, 1-9.

Ekpo, K. E., & Onigbinde, A. O. (2005). Nutritional potentials of the larva of Rhynchophorus phoenicis (F). *Pakistan Journal of Nutrition*, *4*(5), 287-290.

Engel, P., & Moran, N. A. (2013). The gut microbiota of insects–diversity in structure and function. *FEMS microbiology reviews*, *37*(5), 699-735.

Escobar, J. C., Lora, E. S., Venturini, O. J., Yáñez, E. E., Castillo, E. F., & Almazan, O. (2009). Biofuels: environment, technology, and food security. *Renewable and sustainable energy reviews*, *13*(6-7), 1275-1287.

Escobar-Zepeda, A., Vera-Ponce de León, A., & Sanchez-Flores, A. (2015). The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in genetics*, *6*, 348.

Ewuim, S. C., Akunne, C. E., Anumba, A. I., & Etaga, H. O. (2016). Insects associated with wine from Raffia palm (Raphia Hookeri) in Alor, Nigeria. *Animal Research International*, *8*(1).

Ezeilo, U. R., Zakaria, I. I., Huyop, F., & Wahab, R. A. (2017). Enzymatic breakdown of lignocellulosic biomass: the role of glycosyl hydrolases and lytic polysaccharide monooxygenases. *Biotechnology & Biotechnological Equipment*, *31*(4), 647-662.

Fadele, O., Oguocha, I.N., Odeshi, A. & Soleimani, M. (2017) The effect of alkalization on properties of raffia palm fiber. Proceedings of the 26th CANCAM

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., ... & Heath, A. C. en Leibel, RL (2013)'The long-term stability of the human gut microbiota'. *Science*, *341*.

Faunce, T. A., Lubitz, W., Rutherford, A. B., MacFarlane, D., Moore, G. F., Yang, P., ... & Styring, S. (2013). Energy and environment policy case for a global project on artificial photosynthesis. *Energy & Environmental Science*, *6*(3), 695-698.

Feehery, G. R., Yigit, E., Oyola, S. O., Langhorst, B. W., Schmidt, V. T., Stewart, F. J., ... & Pradhan, S. (2013). A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PloS one*, *8*(10), e76096.

Fekete, S., Beck, A., Veuthey, J. L., & Guillarme, D. (2015). Ion-exchange chromatography for the characterization of biopharmaceuticals. *Journal of pharmaceutical and biomedical analysis*, *113*, 43-55.

Ferrer, M., Beloqui, A., Timmis, K. N., & Golyshin, P. N. (2009). Metagenomics for mining new genetic resources of microbial communities. *Journal of Molecular Microbiology and Biotechnology*, *16*(1-2), 109-123.

Ferrer, M., Golyshina, O. V., Chernikova, T. N., Khachane, A. N., Reyes-Duarte, D., Santos, V. A. M. D., ... & Golyshin, P. N. (2005). Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. *Environmental Microbiology*, *7*(12), 1996-2010.

Field, D., Amaral-Zettler, L., Cochrane, G., Cole, J. R., Dawyndt, P., Garrity, G. M., ... & Wooley, J. (2011). The genomic standards consortium. *PLoS biology*, *9*(6), e1001088.

Fisher, A. B., & Fong, S. S. (2014). Lignin biodegradation and industrial implications. AIMS Bioengineering. 1(2), 92-112.

Fitches, E., Wilkinson, H., Bell, H., Bown, D. P., Gatehouse, J. A., & Edwards, J. P. (2004). Cloning, expression and functional characterisation of chitinase from larvae of tomato moth (Lacanobia oleracea): a demonstration of the insecticidal activity of insect chitinase. *Insect Biochemistry and Molecular Biology*, *34*(10), 1037-1050.

Franzini, P. Z., Ramond, J. B., Scholtz, C. H., Sole, C. L., Ronca, S., & Cowan, D. A. (2016). The gut microbiomes of two Pachysoma MacLeay desert dung beetle species (Coleoptera: Scarabaeidae: Scarabaeinae) feeding on different diets. *PloS one*, *11*(8), e0161118.

Gallagher, S. (1998). Quantitation of nucleic acids with absorption spectroscopy. *Current protocols in protein science*, *13*(1), A-4K.

Galli, C., & Gentili, P. (2004). Chemical messengers: mediated oxidations with the enzyme laccase. *Journal of physical organic chemistry*, *17*(11), 973-977.

García-López, R., Cornejo-Granados, F., Lopez-Zavala, A. A., Sánchez-López, F., Cota-Huízar, A., Sotelo- Mundo, R. R., ... & Ochoa-Leyva, A. (2020). Doing more with less: A comparison of 16S hypervariable regions in search of defining the shrimp microbiota. *Microorganisms*, *8*(1), 134.

Garcia-Ruiz, E., Gonzalez-Perez, D., Ruiz-Dueñas, F. J., Martínez, A. T., & Alcalde, M. (2012). Directed evolution of a temperature-, peroxide-and alkaline pH-tolerant versatile peroxidase. *Biochemical Journal*, *441*(1), 487-498.

Garmendia, C., Bernad, A., Esteban, J. A., Blanco, L., & Salas, M. (1992). The bacteriophage phi 29 DNA polymerase, a proofreading enzyme. *Journal of Biological Chemistry*, *267*(4), 2594-2599.

Geib, S. M., Filley, T. R., Hatcher, P. G., Hoover, K., Carlson, J. E., del Mar Jimenez-Gasco, M., Nakagawa- Izumi, A., Sleighter, R.L., & Tien, M. (2008). Lignin degradation in wood- feeding insects. *Proceedings of the National Academy of Sciences*.

Geib, S. M., Tien, M., & Hoover, K. (2010). Identification of proteins involved in lignocellulose degradation using in gel zymogram analysis combined with mass spectroscopy- based peptide analysis of gut proteins from larval Asian longhorned beetles, Anoplophora glabripennis. *Insect Science*, *17*(3), 253- 264.

Ghosh, Purnendu, and Tarun K. Ghose. "Bioethanol in India: recent past and emerging future." *Biotechnology in India II* (2003): 1-27.

Gibson, M. K., Forsberg, K. J., & Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME journal*, *9*(1), 207-216.

Gilbert, H. J. (2010). The biochemistry and structural biology of plant cell wall deconstruction. *Plant Physiology*, Vol 153, 444-455.

Glasel, J. A. (1995). Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios. *Biotechniques*, *18*(1), 62-63.

Godbole, V., Pal, M. K., & Gautam, P. (2021). A critical perspective on the scope of interdisciplinary approaches used in fourth-generation biofuel production. *Algal Research*, *58*, 102436.

Gomes, S. I., Kielak, A. M., Hannula, S. E., Heinen, R., Jongen, R., Keesmaat, I., ... & Bezemer, T. M. (2020). Microbiomes of a specialist caterpillar are consistent across different habitats but also resemble the local soil microbial communities. *Animal microbiome*, *2*(1), 1-12.

Gong, G., Kim, S., Lee, S. M., Woo, H. M., Park, T. H., & Um, Y. (2017). Complete genome sequence of Bacillus sp. 275, producing extracellular cellulolytic, xylanolytic and ligninolytic enzymes. *Journal of Biotechnology*, *254*, 59-62.

Gong, Q., Cao, L. J., Sun, L. N., Chen, J. C., Gong, Y. J., Pu, D. Q., ... & Wei, S. J. (2020). Similar gut bacterial microbiota in two fruit-feeding moth pests collected from different host species and locations. *Insects*, *11*(12), 840.

Gräslund, S., Nordlund, P., Weigelt, J., Hallberg, B. M., Bray, J., Gileadi, O., ... & Gunsalus, K. C. (2008). Protein production and purification. *Nature methods*, *5*(2), 135-146.

Groome, N. P. (1980). Superiority of ABTS over Trinder reagent as chromogen in highly sensitive peroxidase assays for enzyme linked immunoadsorbent assay.

Guo, W. J., Xu, J. K., Liu, J. J., Lang, J. J., Gao, S. Q., Wen, G. B., & Lin, Y. W. (2021). Biotransformation of Lignin by an Artificial Heme Enzyme Designed in Myoglobin With a Covalently Linked Heme Group. *Frontiers in bioengineering and biotechnology*, *9*, 664388.

Gupta, A., & Verma, J. P. (2015). Sustainable bio-ethanol production from agro-residues: a review. *Renewable and Sustainable Energy Reviews*, *41*, 550-567

Gupta, J., Rathour, R., Kumar, M., & Thakur, I. S. (2017). Metagenomic analysis of microbial diversity in landfill lysimeter soil of Ghazipur landfill site, New Delhi, India. *Genome Announcements*, *5*(42), e01104-17.

Gupta, R., & Lee, Y. Y. (2010). Investigation of biomass degradation mechanism in pre- treatment of switchgrass by aqueous ammonia and sodium hydroxide. *Bioresource technology*, *101*(21), 8185-8191.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.

Gurung, N., Ray, S., Bose, S., & Rai, V. (2013). A broader view: microbial enzymes and their relevance in industries, medicine, and beyond. *BioMed research international*, *2013*.

Guzman, J., & Vilcinskas, A. (2020). Bacteria associated with cockroaches: health risk or biotechnological opportunity?. *Applied microbiology and biotechnology*, 1-19.

Hammer, T. J., Dickerson, J. C., & Fierer, N. (2015). Evidence-based recommendations on storing and handling specimens for analyses of insect microbiota. *PeerJ*, *3*, e1190.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured Microorganisms. *Microbiology and molecular biology reviews*, *68*(4), 669-685.

Hansen, A. K., & Moran, N. A. (2014). The impact of microbial symbionts on host plant utilization by herbivorous insects. *Molecular ecology*, *23*(6), 1473-1496.

Harnpicharnchai, P., Thongaram, T., Sriprang, R., Champreda, V., Tanapongpipat, S., & Eurwilaichitr, L. (2007). An efficient purification and fractionation of genomic DNA from soil by modified troughing method. *Letters in applied microbiology*, *45*(4), 387-391.

Harris, M. N., Norzainih, J. J., & Wahida, O. N. (2015). Morphology and Histology of the Digestive System of the Red Palm Weevil Larva, Rhynchophorus ferrugineus, Olivier (Coleoptera: Dryophthoridae). *Interaction*, *22*, 24.

Harrison, X. A., McDevitt, A. D., Dunn, J. C., Griffiths, S. M., Benvenuto, C., Birtles, R., ... & Antwis, R. E. (2021). Fungal microbiomes are determined by host phylogeny and exhibit widespread associations with the bacterial microbiome. *Proceedings of the Royal Society B*, *288*(1957), 20210552.

Hartl, D. L., & Jones, E. W. (2009). *Genetics: analysis of genes and genomes*. Jones & Bartlett Learning.

Hatakka, A., & Hammel, K. E. (2011). Fungal biodegradation of lignocelluloses. In *Industrial applications* (pp. 319-340). Springer, Berlin, Heidelberg.

Havlík, P., Schneider, U. A., Schmid, E., Böttcher, H., Fritz, S., Skalský, R., ... & Obersteiner, M. (2011). Global land-use implications of first and second generation biofuel targets. *Energy policy*, *39*(10), 5690-5702.

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., ... & Horiuchi, T. (2006). Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Molecular systems biology*, *2*(1), 2006-0007.

Hays, S. G., & Ducat, D. C. (2015). Engineering cyanobacteria as photosynthetic feedstock factories. *Photosynthesis research*, *123*(3), 285-295.

He, S., Ivanova, N., Kirton, E., Allgaier, M., Bergin, C., Scheffrahn, R. H., ... & Hugenholtz, P. (2013). Comparative metagenomic and metatranscriptomic analysis of hindgut paunch microbiota in wood- and dung-feeding higher termites. *PloS one*, *8*(4), e61126.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, *56*(2), 61-77.

Hempel, F., Lau, J., Klingl, A., & Maier, U. G. (2011). Algae as protein factories: expression of a human antibody and the respective antigen in the diatom Phaeodactylum tricornutum. *PloS one*, *6*(12), e28424.

Hess, M., Sczyrba, A., Egan, R., Kim, T. W., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S., Chen, F., Zhang, T., Mackie, R. I., Pennacchio, L. A., Tringe, S. G., Visel, A., Woyke, T., Wang, Z., Rubin, E. M. (2011). Metagenomic discovery of biomass- degrading genes and genomes from cow rumen. *Science*, *331*(6016), 463-467.

Hirani, A. H., Javed, N., Asif, M., Basu, S. K., & Kumar, A. (2018). A review on first-and second-generation biofuel productions. *Biofuels: greenhouse gas mitigation and global warming*, 141-154.

Hofrichter, M., Ullrich, R., Pecyna, M. J., Liers, C., & Lundell, T. (2010). New and classic families of secreted fungal heme peroxidases. *Applied microbiology and biotechnology*, *87*(3), 871-897.

Howe, P.D. & Leiserowitz, A. (2013) Who remembers a hot summer or a cold winter? The asymmetric effect of beliefs about global warming on perceptions of local climate conditions in the US. *Global environmental change*, *23*(6), pp.1488-1500.

https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

https://www.bbnet-nibb.co.uk/wp-content/uploads/2019/08/LBNet-UKBioChem10_UK-Top- Bio-based- Chemicals-Opportunities_Dec2017.pdf)

https://www.theguardian.com/environment/2008/jul/03/biofuels.renewableenergy

Huang, S., Sheng, P., & Zhang, H. (2012). Isolation and identification of cellulolytic bacteria from the gut of Holotrichia parallela larvae (Coleoptera: Scarabaeidae). *International Journal of Molecular Sciences*, *13*(3), 2563- 2577.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, *11*(1), 1-11.

IATA (2019). Dangerous goods regulation 60th edition 2019 manual. https://www.iata.org/publications/Pages/standards-manuals.aspx. The International Air Transport Association).

Inganäs, O., & Sundström, V. (2016). Solar energy for electricity and fuels. *Ambio*, *45*(1), 15-23.

International Energy Agency- IEA bioenergy- task42. Bio-based chemicals: Value added products from biorefineries ISBN 9780128120217, https://doi.org/10.1016/B978-0-12-812021-7.00001-4.

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

Israel, A.U., Obot, I.B., Umoren, S.A., Mkpenie, V. & Asuquo, J.E. (2008) Production of cellulosic polymers from agricultural wastes. *Journal of Chemistry*, *5*(1), pp.81-85.

Jamal, M. A. H. M., Sharma, S. P., Chung, H. J., Kim, H. J., Hong, S. T., & Lee, S. (2017). Ultra-high efficient colony PCR for high throughput screening of bacterial genes. *Indian journal of microbiology*, *57*(3), 365- 369.

Janson, J. C. (Ed.). (2012). Protein purification: principles, high resolution methods, and applications. John Wiley & Sons.

Janusz, G., Pawlik, A., Sulej, J., Świderska-Burek, U., Jarosz-Wilkołazka, A., & Paszczyński, A. (2017). Lignin degradation: microorganisms, enzymes involved, genomes analysis and evolution. *FEMS microbiology reviews*, *41*(6), 941-962.

Jia, S., Zhang, X., Zhang, G., Yin, A., Zhang, S., Li, F., ... & Sun, G. (2013). Seasonally variable intestinal metagenomes of the red palm weevil (Rhynchophorus ferrugineus). *Environmental microbiology*, *15*(11), 3020-3029.

Jia, T., Dai, Y., & Wang, R. (2018). Refining energy sources in winemaking industry by using solar energy as alternatives for fossil fuels: A review and perspective. *Renewable and Sustainable Energy Reviews*, *88*, 278-296.

Jimenez, D. J., Andreote, F. D., Chaves, D., Montaña, J. S., Osorio-Forero, C., Junca, H., ... & Baena, S. (2012). Structural and functional insights from the metagenome of an acidic hot spring microbial planktonic community in the Colombian Andes. *PloS one*, *7*(12), e52069.

Jones, R. T., Sanchez, L. G., & Fierer, N. (2013). A cross-taxon analysis of insect-associated bacterial diversity. *PLoS one*, *8*(4), e61218.

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., ... & Wong, G. K. S. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in microbiology*, *7*, 459.

Joynson, R. E. (2015). Analysis of the gut microbiome of the common black slug Arion ater: In search of novel lignocellulose degrading enzymes (Doctoral dissertation, University of Salford).

Joynson, R., Pritchard, L., Osemwekha, E., & Ferry, N. (2017). Metagenomic analysis of the gut microbiome of the common black slug Arion ater in search of novel lignocellulose degrading enzymes. *Frontiers in Microbiology*, *8*, 2181.

Joynson, R., Swamy, A., Bou, P. A., Chapuis, A., & Ferry, N. (2014). Characterization of cellulolytic activity in the gut of the terrestrial land slug Arion ater: biochemical identification of targets for intensive study. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *177*, 29-35.

Kameshwar, A. K. S., & Qin, W. (2017a). Metadata Analysis of Phanerochaete chrysosporium gene expression data identified common CAZymes encoding gene expression profiles involved in cellulose and hemicellulose degradation. *International journal of biological sciences*, *13*(1), 85.

Kameshwar, A. K. S., & Qin, W. (2017b). Qualitative and quantitative methods for isolation and characterization of lignin-modifying enzymes secreted by microorganisms. *BioEnergy Research*, *10*(1), 248-266.

Kane, S. D., & French, C. E. (2018). Characterisation of novel biomass degradation enzymes from the genome of Cellulomonas fimi. *Enzyme and microbial technology*, *113*, 9-17.

Kanokratana, P., Eurwilaichitr, L., Pootanakit, K., & Champreda, V. (2015). Identification of glycosyl hydrolases from a metagenomic library of microflora in sugarcane bagasse collection site and their cooperative action on cellulose degradation. *Journal of bioscience and bioengineering*, *119*(4), 384-391.

Kanokratana, P., Mhuantong, W., Laothanachareon, T., Tangphatsornruang, S., Eurwilaichitr, L., Pootanakit, K., & Champreda, V. (2013). Phylogenetic analysis and metabolic potential of microbial communities in an industrial bagasse collection site. *Microbial ecology*, *66*(2), 322-334.

Kassim, A. S. M., Ishak, N., Aripin, A. M., & Zaidel, D. N. F. A. (2016). Potential Lignin Degraders Isolated from the Gut of Rhynchophorus Ferrugineus.

Kersten, P., & Cullen, D. (2014). Copper radical oxidases and related extracellular oxidoreductases of wood- decay Agaricomycetes. *Fungal Genetics and Biology*, *72*, 124-130.

Khalil, H.S.A., Alwani, M.S. & Omar, A.K.M. (2007) Chemical composition, anatomy, lignin distribution, and cell wall structure of Malaysian plant waste fibers. *BioResources*, *1*(2), pp.220-232

Kilpatrick, C. W. (2002). Noncryogenic preservation of mammalian tissues for DNA extraction: an assessment of storage methods. *Biochemical genetics*, *40*(1), 53-62.

Kim, O. S., Cho, Y. J., Lee, K., Yoon, S. H., Kim, M., Na, H., ... & Won, S. (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *International journal of systematic and evolutionary microbiology*, *62*(3), 716-721.

Kim, S. J., & Shoda, M. (1999). Purification and characterization of a novel peroxidase from Geotrichum candidum Dec 1 involved in decolorization of dyes. *Applied and Environmental Microbiology*, *65*(3), 1029- 1035.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., & Glöckner, F. O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing- based diversity studies. *Nucleic acids research*, *41*(1), e1-e1.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J. A., ... & Bailey, M. J. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nature biotechnology*, *30*(6), 513.

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., ... & Melnik, A. V. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, *16*(7), 410-422.

Koffi, D. M., Cisse, M., Koua, G. A., & Niamke, S. L. (2017). Nutritional and functional properties of flour from the palm (Elaeis guineensis) weevil Rhynchophorus phoenicis larvae consumed as protein source in south Côte

d'Ivoire. *Annals of the University Dunarea de Jos of Galati Fascicle VI--Food Technology*, *41*(1).

Kougias, P. G., Campanaro, S., Treu, L., Tsapekos, P., Armani, A., & Angelidaki, I. (2018). Spatial distribution and diverse metabolic functions of lignocellulose-degrading uncultured bacteria as revealed by genome-centric metagenomics. *Applied and environmental microbiology*, *84*(18), e01244-18.

Krueger, F. (2015). Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, 516(517).

Kumar, A., & Chandra, R. (2020). Ligninolytic enzymes and its mechanisms for degradation of lignocellulosic waste in environment. *Heliyon*, *6*(2), e03170.

Kumar, D., Singh, B., & Korstad, J. (2017). Utilization of lignocellulosic biomass by oleaginous yeast and bacteria for production of biodiesel and renewable diesel. *Renewable and Sustainable Energy Reviews*, *73*, 654-671.

Kumari, D., & Singh, R. (2018). Pretreatment of lignocellulosic wastes for biofuel production: A critical review. *Renewable and Sustainable Energy Reviews*, *90*, 877-891.

Kunath, B. J., Bremges, A., Weimann, A., McHardy, A. C., & Pope, P. B. (2017). Metagenomics and CAZyme discovery. In *Protein-Carbohydrate Interactions* (pp. 255-277). Humana Press, New York, NY.

Kurian, J. K., Nair, G. R., Hussain, A., & Raghavan, G. V. (2013). Feedstocks, logistics and pre- treatment processes for sustainable lignocellulosic biorefineries: a comprehensive review. *Renewable and Sustainable Energy Reviews*, *25*, 205- 219.

Lai, C. M. T., Chua, H. B., Danquah, M. K., & Saptoro, A. (2017, June). Isolation of thermophilic lignin degrading bacteria from oil-palm empty fruit bunch (EFB) compost. In *IOP conference series: materials science and engineering* (Vol. 206, No. 1, p. 012016). IOP Publishing.

Lambertz, C., Ece, S., Fischer, R., & Commandeur, U. (2016). Progress and obstacles in the production and application of recombinant lignin-degrading peroxidases. *Bioengineered*, *7*(3), 145-154.

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., ... & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature biotechnology*, *31*(9), 814-821.

Lapidus, A. L., & Korobeynikov, A. I. (2021). Metagenomic data assembly–the way of decoding unknown microorganisms. *Frontiers in Microbiology*, *12*, 653.

Lasken, R. S. (2009). Genomic DNA amplification by the multiple displacement amplification (MDA) method. *Biochemical Society Transactions*, *37*(2), 450-453.

Lauber, C., Schwarz, T., Nguyen, Q. K., Lorenz, P., Lochnit, G., & Zorn, H. (2017). Identification, heterologous expression, and characterization of a dye-decolorizing peroxidase of Pleurotus sapidus. *AMB Express*, *7*(1), 1-15.

Lazarevic, V., Gaïa, N., Girard, M., & Schrenzel, J. (2016). Decontamination of 16S rRNA gene amplicon sequence datasets based on bacterial load assessment by qPCR. *BMC microbiology*, *16*(1), 1-8.

Lazaridis, P. A., Fotopoulos, A. P., Karakoulia, S. A., & Triantafyllidis, K. S. (2018). Catalytic fast pyrolysis of kraft lignin with conventional, mesoporous and nanosized ZSM-5 zeolite for the production of alkyl- phenols and aromatics. *Frontiers in chemistry*, *6*, 295.

LBNet and BBSRC- NIBB (2019). The ten green chemicals which can create growth, jobs and trade for the UK.

Le, N. G. (2021). Isolation and characterization of novel enzymatic activities from gut metagenomes to support lignocellulose breakdown.

Lee, H. V., Hamid, S. B. A., & Zain, S. K. (2014). Conversion of lignocellulosic biomass to nanocellulose: structure and chemical process. *The Scientific World Journal*, *2014*.

Lee, P. Y., Costumbrado, J., Hsu, C. Y., & Kim, Y. H. (2012). Agarose gel electrophoresis for the separation of DNA fragments. *JoVE (Journal of Visualized Experiments)*, (62), e3923.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M., & Henrissat, B. (2013). Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for biofuels*, *6*(1), 1-14.

Levasseur, A., Piumi, F., Coutinho, P. M., Rancurel, C., Asther, M., Delattre, M., ... & Record, E. (2008). FOLy: an integrated database for the classification and functional annotation of fungal oxidoreductases potentially involved in the degradation of lignin and related aromatic compounds. *Fungal genetics and biology*, *45*(5), 638-645.

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674-1676.

Li, J., Yuan, H., & Yang, J. (2009). Bacteria and lignin degradation. *Frontiers of Biology in China*, *4*(1), 29-38.

Li, Y., Hu, X., Yang, S., Zhou, J., Qi, L., Sun, X., ... & Lin, S. (2018). Comparison between the fecal bacterial microbiota of healthy and diarrheic captive musk deer. *Frontiers in microbiology*, *9*, 300.

Liers, C., Aranda, E., Strittmatter, E., Piontek, K., Plattner, D. A., Zorn, H., ... & Hofrichter, M. (2014). Phenol oxidation by DyP-type peroxidases in comparison to fungal and plant peroxidases. *Journal of Molecular Catalysis B: Enzymatic*, *103*, 41-46.

Liers, C., Bobeth, C., Pecyna, M., Ullrich, R., & Hofrichter, M. (2010). DyP-like peroxidases of the jelly fungus Auricularia auricula-judae oxidize nonphenolic lignin model compounds and high-redox potential dyes. *Applied microbiology and biotechnology*, *85*(6), 1869-1879.

Lin, D., Zhang, L., Shao, W., Li, X., Liu, X., Wu, H., & Rao, Q. (2019). Phylogenetic analyses and characteristics of the microbiomes from five mealybugs (Hemiptera: Pseudococcidae). *Ecology and evolution*, *9*(4), 1972-1984.

Linard, B., Arribas, P., Andújar, C., Crampton-Platt, A., & Vogler, A. P. (2016). Lessons from genome skimming of arthropod-preserving ethanol. *Molecular ecology resources*, *16*(6), 1365-1377.

Linde, D., Ayuso-Fernández, I., Laloux, M., Aguiar-Cervera, J. E., de Lacey, A. L., Ruiz-Dueñas, F. J., & Martínez, A. T. (2021). Comparing ligninolytic capabilities of bacterial and fungal dye-decolorizing peroxidases and class-ii peroxidase-catalases. *International journal of molecular sciences*, *22*(5), 2629.

Liu, B., & Pop, M. (2009). ARDB—antibiotic resistance genes database. *Nucleic acids research*, *37*(suppl_1), D443-D447.

Liu, P. F., Avramova, L. V., & Park, C. (2009). Revisiting absorbance at 230 nm as a protein unfolding probe. *Analytical biochemistry*, *389*(2), 165-170.

Liu, X., Yuan, Z., Wang, J., Cui, Y., Liu, S., Ma, Y., ... & Xu, S. (2017). Crystal structure and biochemical features of dye-decolorizing peroxidase YfeX from Escherichia coli O157 Asp143 and Arg232 play divergent roles toward different substrates. *Biochemical and biophysical research communications*, *484*(1), 40-44.

Lluch, J., Servant, F., Païssé, S., Valle, C., Valière, S., Kuchly, C., ... & Amar, J. (2015). The characterization of novel tissue microbiota using an optimized 16S metagenomic sequencing pipeline. *PloS one*, *10*(11), e0142334.

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*(D1), D490-D495.

Lončar, N., Drašković, N., Božić, N., Romero, E., Simić, S., Opsenica, I., ... & Fraaije, M. W. (2019). Expression and Characterization of a Dye-decolorizing Peroxidase from Pseudomonas fluorescens Pf0-1. *Catalysts*, *9*(5), 463.

Lu, F., Karlen, S. D., Regner, M., Kim, H., Ralph, S. A., Sun, R. C., ... & Ralph, J. (2015). Naturally p-hydroxybenzoylated lignins in palms. *BioEnergy Research*, *8*(3), 934-952.

Lucena-Aguilar, G., Sánchez-López, A. M., Barberán-Aceituno, C., Carrillo-Avila, J. A., López-Guerrero, J. A., & Aguilar-Quesada, R. (2016). DNA source selection for downstream applications based on DNA quality indicators analysis. *Biopreservation and Biobanking*, *14*(4), 264-270.

Luo, C., Li, Y., Chen, Y., Fu, C., Nong, X., & Yang, Y. (2019). Degradation of bamboo lignocellulose by bamboo snout beetle Cyrtotrachelus buqueti in vivo and vitro: efficiency and mechanism. *Biotechnology for biofuels*, *12*(1), 1-14.

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.

Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., & Woese, C. R. (1997). The RDP (ribosomal database project). *Nucleic acids research*, *25*(1), 109-110.

Majidian, P., Tabatabaei, M., Zeinolabedini, M., Naghshbandi, M. P., & Chisti, Y. (2018). Metabolic engineering of microorganisms for biofuel production. *Renewable and Sustainable Energy Reviews*, *82*, 3863-3885.

Martin-Dominguez, V., Estevez, J., Ojembarrena, F. D. B., Santos, V. E., & Ladero, M. (2018). Fumaric acid production: a biorefinery perspective. *Fermentation*, *4*(2), 33.

Martins, L. F., Antunes, L. P., Pascon, R. C., de Oliveira, J. C. F., Digiampietri, L. A., Barbosa, D., ... & Setubal, J. C. (2013). Metagenomic analysis of a tropical composting operation at the São Paulo Zoo Park reveals diversity of biomass degradation functions and organisms. *PloS one*, *8*(4), e61928.

Mateo, Nicolás, Werner Nader, and Giselle Tamayo. "Bioprospecting." *Encyclopedia of Biodiversity* 1 (2001): 471-487.

Mathews, S. L., Grunden, A. M., & Pawlak, J. (2016). Degradation of lignocellulose and lignin by Paenibacillus glucanolyticus. *International Biodeterioration & Biodegradation*, *110*, 79-86.

Mayr, S. A., Subagia, R., Weiss, R., Schwaiger, N., Weber, H. K., Leitner, J., ... & Guebitz, G. M. (2021). Oxidation of various kraft lignins with a bacterial laccase enzyme. *International Journal of Molecular Sciences*, *22*(23), 13161.

Mba, A. R. F., Kansci, G., Viau, M., Hafnaoui, N., Meynier, A., Demmano, G., & Genot, C. (2017). Lipid and amino acid profiles support the potential of Rhynchophorus phoenicis larvae for human nutrition. *Journal of Food Composition and Analysis*, *60*, 64-73.

Mba, A. R. F., Kansci, G., Viau, M., Hafnaoui, N., Meynier, A., Demmano, G., & Genot, C. (2017). Lipid and amino acid profiles support the potential of Rhynchophorus phoenicis larvae for human nutrition. *Journal of Food Composition and Analysis*, *60*, 64-73.

McClenaghan, B., Gibson, J. F., Shokralla, S., & Hajibabaei, M. (2015). Discrimination of grasshopper (Orthoptera: A crididae) diet and niche overlap using next-generation sequencing of gut contents. *Ecology and evolution*, *5*(15), 3046-3055.

McDonald, J. E., Allison, H. E., & McCarthy, A. J. (2010). Composition of the landfill microbial community as determined by application of domain-and group-specific 16S and 18S rRNA-targeted oligonucleotide probes. *Applied and environmental microbiology*, *76*(4), 1301-1306.

McElhoe, J. A., Holland, M. M., Makova, K. D., Su, M. S. W., Paul, I. M., Baker, C. H., ... & Young, B. (2014). Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. *Forensic Science International: Genetics*, *13*, 20-29.

Mei, H. Z., Xia, D. G., Zhao, Q. L., Zhang, G. Z., Qiu, Z. Y., Qian, P., & Lu, C. (2016). Molecular cloning, expression, purification and characterization of a novel cellulase gene (Bh-EGaseI) in the beetle Batocera horsfieldi. *Gene*, *576*(1), 45-51.

Menon, V., & Rao, M. (2012). Trends in bioconversion of lignocellulose: biofuels, platform chemicals & biorefinery concept. *Progress in energy and combustion science*, *38*(4), 522-550.

Mhuantong, W., Charoensawan, V., Kanokratana, P., Tangphatsornruang, S., & Champreda, V. (2015). Comparative analysis of sugarcane bagasse metagenome reveals unique and conserved biomass- degrading enzymes

among lignocellulolytic microbial communities. *Biotechnology for biofuels*, *8*(1), 1- 17.

Michaelides, E. E. S. (2012). *Alternative energy sources*. Springer Science & Business Media.

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, *34*(13), i142-i150.

Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, *32*(7), 1088-1090.

Min, K., Gong, G., Woo, H. M., Kim, Y., & Um, Y. (2015). A dye-decolorizing peroxidase from Bacillus subtilis exhibiting substrate-dependent optimum temperature for dyes and β-ether lignin dimer. *Scientific Reports*, *5*(1), 1-8.

Mofijur, M., Rasul, M. G., Hyde, J., & Bhuyia, M. M. K. (2015). Role of biofuels on IC engines emission reduction. *Energy Procedia*, *75*, 886-892.

Mohammed, W. S., Ziganshina, E. E., Shagimardanova, E. I., Gogoleva, N. E., & Ziganshin, A. M. (2018). Comparison of intestinal bacterial and fungal communities across various xylophagous beetle larvae (Coleoptera: Cerambycidae). *Scientific reports*, *8*(1), 1-12.

Mohr, A., & Raman, S. (2013). Lessons from first generation biofuels and implications for the sustainability appraisal of second generation biofuels. *Energy Policy*, *63*, 114- 122.

Montagna, M., Chouaia, B., Mazza, G., Prosdocimi, E. M., Crotti, E., Mereghetti, V., ... & Cervo, R. (2015). Effects of the diet on the microbiota of the red palm weevil (Coleoptera: Dryophthoridae). *PLoS One*, *10*(1), e0117439.

Moreau, C. S., Wray, B. D., Czekanski-Moir, J. E., & Rubin, B. E. (2013). DNA preservation: a test of commonly used preservatives for insects. *Invertebrate systematics*, *27*(1), 81-86.

Muafor, F. J., Gnetegha, A. A., Le Gall, P., & Levang, P. (2015). *Exploitation, trade and farming of palm weevil grubs in Cameroon* (Vol. 178). CIFOR.

Muhammad, A., Fang, Y., Hou, Y., & Shi, Z. (2017). The gut entomotype of red palm weevil Rhynchophorus ferrugineus Olivier (Coleoptera: Dryophthoridae) and their effect on host nutrition metabolism. *Frontiers in microbiology*, *8*, 2291.

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Katta, H. Y., Mojica, A., ... & Reddy, T. B. K. (2019). Genomes OnLine database (GOLD) v. 7: updates and new features. *Nucleic acids research*, *47*(D1), D649-D659.

Munoz-Benavent, M., Pérez-Cobas, A. E., García-Ferris, C., Moya, A., & Latorre, A. (2021). Insects' potential: understanding the functional role of their gut microbiome. *Journal of Pharmaceutical and Biomedical Analysis*, *194*, 113787.

Mussatto, S. I., & Dragone, G. M. (2016). Biomass pretreatment, biorefineries, and potential products for a bioeconomy development. In *Biomass fractionation technologies for a lignocellulosic feedstock based biorefinery* (pp. 1-22). Elsevier.

Mustafa, G. R., Li, C., Zhao, S., Jin, L., He, X., Shabbir, M. Z., ... & Zou, L. (2021). Metagenomic analysis revealed a wide distribution of antibiotic resistance genes and biosynthesis of antibiotics in the gut of giant pandas. *BMC microbiology*, *21*(1), 1-18.

Mwangi, J.K., Lee, W.J., Chang, Y.C., Chen, C.Y. & Wang, L.C., (2015) An overview: Energy saving and pollution reduction by using green fuel blends in diesel engines. *Applied energy*, ***159***, pp.214-236.

Naik, S. N., Goud, V. V., Rout, P. K., & Dalai, A. K. (2010). Production of first and second generation biofuels: a comprehensive review. *Renewable and Sustainable Energy Reviews*, *14*(2), 578-597.

Narra, M., James, J. P., & Balasubramanian, V. (2015). Simultaneous saccharification and fermentation of delignified lignocellulosic biomass at high solid loadings by a newly isolated thermotolerant Kluyveromyces sp. for ethanol production. *Bioresource Technology*, *179*, 331-338.

Nasser, R. A., Salem, M. Z., Hiziroglu, S., Al-Mefarrej, H. A., Mohareb, A. S., Alam, M., & Aref, I. M. (2016). Chemical analysis of different parts of date palm (Phoenix dactylifera L.) using ultimate, proximate and thermo-gravimetric techniques for energy production. *Energies*, *9*(5), 374.

Neeraas, P. C. (2019). Discovery, expression, and characterization of novel laccases derived from marine bacteria (Master's thesis, Norwegian University of Life Sciences, Ås).

Nevalainen, K. H., Te'o, V. S., & Bergquist, P. L. (2005). Heterologous protein expression in filamentous fungi. *Trends in biotechnology*, *23*(9), 468-474.10.1016/j.tibtech.2005.06.002.Review. *Biocatalysis and agricultural biotechnology*, *23*, 101498.

Ni, J., & Tokuda, G. (2013). Lignocellulose-degrading enzymes from termites and their symbiotic microbiota. *Biotechnology advances*, *31*(6), 838-850.

NIFOR. (2015). A Manual on Oil Palm Production. 9th Edn., Extension Division, Nigerian Institute For Oil Palm Research. P57.

Nimchua, T., Thongaram, T., Uengwetwanit, T., Pongpattanakitshote, S., & Eurwilaichitr, L. (2012). Metagenomic analysis of novel lignocellulose-degrading enzymes from higher termite guts inhabiting microbes. *Journal of microbiology and biotechnology*, *22*(4), 462-469.

Obahiagbon, F. I. (2012). A review: aspects of the African oil palm (Elaeis guineesis jacq.) and the implications of its bioactives in human health. *Am J Biochem Mol Biol*, *2*(3), 106-119.

Ogola, H. J. O., Kamiike, T., Hashimoto, N., Ashida, H., Ishikawa, T., Shibata, H., & Sawa, Y. (2009). Molecular characterization of a novel peroxidase from the cyanobacterium Anabaena sp. strain PCC 7120. *Applied and environmental microbiology*, *75*(23), 7509-7518.

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

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B., ... & Wagner, H. (2013). Community ecology package. *R package version*, *2*(0).

Olguín, E. J. (2012). Dual purpose microalgae–bacteria-based systems that treat wastewater and produce biodiesel and chemical products within a Biorefinery. *Biotechnology advances*, *30*(5), 1031- 1046.

Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., & Pop, M. (2019). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in bioinformatics*, *20*(4), 1140-1150.

Olsson, N. (2016). Lignin degradation and oxygen dependence. SLU, Swedish University of Agricultural Sciences Faculty of Landscape Architecture, Horticulture and Crop Production Science Department of Plant Breeding

Omotoso, O. T., & Adedire, C. O. (2007). Nutrient composition, mineral content and the solubility of the proteins of palm weevil, Rhynchophorus phoenicis f. (Coleoptera: Curculionidae). *Journal of Zhejiang University Science B*, *8*(5), 318-322.

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, *12*(1), 385. https://doi.org/10.1186/1471-2105-12-385

Overholt, W. A., Diaz, R., Rosskopf, E., Green, S. J., & Overholt, W. A. (2015). Deep characterization of the microbiomes of Calophya spp.(Hemiptera: Calophyidae) gall-inducing psyllids reveals the absence of plant pathogenic bacteria and three dominant endosymbionts. *PLoS One*, *10*(7), e0132248.

Pagani, I., Liolios, K., Jansson, J., Chen, I. M. A., Smirnova, T., Nosrat, B., Markowitz, V. M., & Kyrpides, N. C. (2011). The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, *40*(D1), D571-D579.

Pal, S., & Karmakar, P. (2018). Symbionts associated with insect digestive system and their role in insect nutrition. *Journal of Entomology and Zoology Studies*, *6*(5), 421-425.

Paniagua Voirol, L. R., Frago, E., Kaltenpoth, M., Hilker, M., & Fatouros, N. E. (2018). Bacterial symbionts in Lepidoptera: their diversity, transmission, and impact on the host. *Frontiers in microbiology*, *9*, 556.

Park, R., Dzialo, M. C., Spaepen, S., Nsabimana, D., Gielens, K., Devriese, H., ... & Verstrepen, K. J. (2019). Microbial communities of the house fly Musca domestica vary with geographical location and habitat. *Microbiome*, *7*(1), 1-12.

Patil, I. (2021). Visualizations with statistical details: The'ggstatsplot'approach. *Journal of Open Source Software*, *6*(61), 3167.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, *42*(1), 3-1.

Perez-Boada, M., Ruiz-Duenas, F. J., Pogni, R., Basosi, R., Choinowski, T., Martínez, M. J., ... & Martínez, A. T. (2005). Versatile peroxidase oxidation of high redox potential aromatic compounds: site-directed mutagenesis, spectroscopic and crystallographic investigation of three long-range electron transfer pathways. *Journal of molecular biology*, *354*(2), 385-402.

Perna, V., Meyer, A. S., Holck, J., Eltis, L. D., Eijsink, V. G., & Wittrup Agger, J. (2019). Laccase-catalyzed oxidation of lignin induces production of H2O2. *ACS Sustainable Chemistry & Engineering*, *8*(2), 831-841.

Phillips, A. T., & Signs, M. W. (2004). Desalting, concentration, and buffer exchange by dialysis and ultrafiltration. *Current Protocols in Protein Science*, *38*(1), 4-4.

Poelchau, M. F., Coates, B. S., Childers, C. P., de Leon, A. A. P., Evans, J. D., Hackett, K., & Shoemaker, D. (2016). Agricultural applications of insect ecological genomics. *Current opinion in insect science*, *13*, 61-69.

Pollegioni, L., Tonin, F., & Rosini, E. (2015). Lignin-degrading enzymes. *The FEBS journal*, *282*(7), 1190-1213.

Pope, P. B., Denman, S. E., Jones, M., Tringe, S. G., Barry, K., Malfatti, S. A., Mchardy, A. C., Cheng, J. F., Hugenholtz, P., McSweeney, C. S., & Morrison, M. (2010). Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other

herbivores. *Proceedings of the National Academy of Sciences*, USA. 201005297.

Popp, J., Lakner, Z., Harangi-Rakos, M., & Fari, M. (2014). The effect of bioenergy expansion: Food, energy, and environment. *Renewable and sustainable energy reviews*, *32*, 559-578.

Porath, J. (1992). Immobilized metal ion affinity chromatography. *Protein expression and purification*, *3*(4), 263- 281.

Pour, R. R., & Bugg, T. D. (2015). Enzymology and Structural Enzymology of Dye-decolorizing Peroxidases and a Primary Study of Encapsulin (Doctoral dissertation, University of Warwick).

Prasad, R. K., Chatterjee, S., Sharma, S., Mazumder, P. B., Vairale, M. G., & Raju, P. S. (2018). Insect Gut bacteria and their potential application in degradation of Lignocellulosic biomass: a review. *Bioremediation: applications for environmental protection and management*, 277-299.

Prawitwong, P., Kosugi, A., Arai, T., *et al.* (2012) Efficient ethanol production from separated parenchyma and vascular bundle of oil palm trunk. *Bioresource technology*, *125*, pp.37-42.

Pu, Y., Jiang, N., & Ragauskas, A. J. (2007). Ionic liquid as a green solvent for lignin. *Journal of Wood Chemistry and Technology*, *27*(1), 23-33.

Qin, X., Luo, H., Zhang, X., Yao, B., Ma, F., & Su, X. (2018). Dye-decolorizing peroxidases in Irpex lacteus combining the catalytic properties of heme peroxidases and laccase play important roles in ligninolytic system. *Biotechnology for biofuels*, *11*(1), 1-11.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), 833-844.

Quinn, O. (2017). Microbiota of an Invasive Wasp Vespula vulgaris and Hymenopteran relatives: Interpreting the microbiome.

Quinn, T. P., Erb, I., Richardson, M. F., & Crowley, T. M. (2018). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, *34*(16), 2870-2878.

R Core Team. 2013. R: a language and environment for statistical computing. R Foundation

Rabelo-Fernandez, R. J., Santiago-Morales, K., Morales-Vale, L., & Rios-Velazquez, C. (2018). The metagenome of Caracolus marginella gut microbiome using culture independent approaches and shotgun sequencing. *Data in brief*, *16*, 501-505.

Radakovits, R., Jinkerson, R. E., Darzins, A., & Posewitz, M. C. (2010). Genetic engineering of algae for enhanced biofuel production. *Eukaryotic cell*, *9*(4), 486-501.

Ragauskas, A. J., Beckham, G. T., Biddy, M. J., Chandra, R., Chen, F., Davis, M. F., Davison, B.H., Dixon, A. R., Gilna, P., Keller, M., Langan, P., Naskar, A. K., Saddler, J. N., Tschaplinski, T. J., Tuskan., G A., Wyman, C. E.,(2014). Lignin valorization: improving lignin processing in the biorefinery. *Science*, *344*(6185), 12468431-10.

Rahmanpour, R., & Bugg, T. D. (2015). Characterisation of Dyp-type peroxidases from Pseudomonas fluorescens Pf-5: oxidation of Mn (II) and polymeric lignin by Dyp1B. *Archives of biochemistry and biophysics*, *574*, 93-98.

Rahmanpour, R., Rea, D., Jamshidi, S., Fülöp, V., & Bugg, T. D. (2016). Structure of Thermobifida fusca DyP- type peroxidase and activity towards Kraft lignin and lignin model compounds. *Archives of biochemistry and biophysics*, *594*, 54-60

Raj, A., Reddy, M. K., & Chandra, R. (2007). Identification of low molecular weight aromatic compounds by gas chromatography–mass spectrometry (GC–MS) from kraft lignin degradation by three Bacillus sp. *International biodeterioration & biodegradation*, *59*(4), 292-296.

Rajagopal, R. (2009). Beneficial interactions between insects and gut bacteria. *Indian Journal of Microbiology*, *49*(2), 114-119.

Ramalho, M. O., Bueno, O. C., & Moreau, C. S. (2017). Microbial composition of spiny ants (Hymenoptera: Formicidae: Polyrhachis) across their geographic range. *BMC Evolutionary Biology*, *17*(1), 1-15.

Rana, V. & Rana, D., (2017). Role of Microorganisms in Lignocellulosic Biodegradation. *Renewable Biofuels: Bioconversion of Lignocellulosic Biomass by Microbial Community*, pp 19-68. Springer, Gewerbestrasse, Switzerland.

Ranjan, R., Grover, A., Kapardar, R. K., & Sharma, R. (2005). Isolation of novel lipolytic genes from uncultured bacteria of pond water. *Biochemical and Biophysical Research Communications*, *335*(1), 57-65.

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and biophysical research communications*, *469*(4), 967-977.

Ransom-Jones, E., McCarthy, A. J., Haldenby, S., Doonan, J., & McDonald, J. E. (2017). Lignocellulose- degrading microbial communities in landfill sites represent a repository of unexplored biomass- degrading diversity. *Msphere*, *2*(4), e00300-17.

Rashid, G. M., & Bugg, T. D. (2021). Enhanced biocatalytic degradation of lignin using combinations of lignin- degrading enzymes and accessory enzymes. *Catalysis Science & Technology*, *11*(10), 3568-3577.

Rasiravuthanahalli, K. G., Revathi, S., Rameshkumar, N., Krishnan, M., & Kayalvizhi, N. (2017). Digestion of Tannin by bacteria Enterobacter cloacae from the gut of Indian mole cricket (Gryllotalpa krishnani). *J Bioprocess Biotech*, *7*(302), 2.

Reddy, T. B., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., ... & Kyrpides, N. C. (2015). The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta) genome project classification. *Nucleic acids research*, *43*(D1), D1099-D1106.

Reich, I., Ijaz, U. Z., Gormally, M., & Smith, C. J. (2018). 16S rRNA sequencing reveals likely beneficial core microbes within faecal samples of the EU protected slug Geomalacus maculosus. *Scientific reports*, *8*(1), 1-9.

Reid, I. D. (1995). Biodegradation of lignin. *Canadian Journal of Botany*, *73*(S1), 1011-1018.

Richardson, E. J., & Watson, M. (2013). The automatic annotation of bacterial genomes. *Briefings in bioinformatics*, *14*(1), 1-12.

Richardson, E. J., Escalettes, F., Fotheringham, I., Wallace, R. J., & Watson, M. (2013). Meta4: a web application for sharing and annotating metagenomic gene predictions using web services. *Frontiers in genetics*, *4*, 168.

Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T., Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H., Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L. Escherichia coli K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Res.* 34 (1), 1-9 (2006) 16397293

Riyadi, F. A., Tahir, A. A., Yusof, N., Sabri, N. S. A., Noor, M. J. M. M., Akhir, F. N., ... & Hara, H. (2020). Enzymatic and genetic characterization of lignin depolymerization by Streptomyces sp. S6 isolated from a tropical environment. *Scientific reports*, *10*(1), 1-9.

Robinson, S. L., Piel, J., & Sunagawa, S. (2021). A roadmap for metagenomic enzyme discovery. *Natural Product Reports*, *38*(11), 1994-2023.

Rodríguez-López, J. N., Gilabert, M. A., Tudela, J., Thorneley, R. N., & García-Cánovas, F. (2000). Reactivity of horseradish peroxidase compound II toward substrates: kinetic evidence for a two-step mechanism. *Biochemistry*, *39*(43), 13201-13209.

Romero, S., Nastasa, A., Chapman, A., Kwong, W. K., & Foster, L. J. (2019). The honey bee gut microbiota: strategies for study and characterization. *Insect molecular biology*, *28*(4), 455-472.

Rosano, G. L. and Ceccarelli, E. A. (2014) 'Recombinant protein expression in Escherichia coli: advances and challenges.', *Frontiers in microbiology*. Frontiers Media SA, 5(172), pp. 1–17. doi: 10.3389/fmicb.2014.00172.

Rosenberg, A. H., Lade, B. N., Dao-shan, C., Lin, S. W., Dunn, J. J., & Studier, F. W. (1987). Vectors for selective expression of cloned DNAs by T7 RNA polymerase. *Gene*, *56*(1), 125-135.

Rosnow, J. J., Anderson, L. N., Nair, R. N., Baker, E. S., & Wright, A. T. (2017). Profiling microbial lignocellulose degradation and utilization by emergent omics technologies. *Critical reviews in biotechnology*, *37*(5), 626-640.

Ross, K., & Mazza, G. (2011). Comparative analysis of pyrolysis products from a variety of herbaceous Canadian crop residues. *World Journal of Agricultural Sciences*, *7*(6), 763-776.

Ruiz-Dueñas, F. J., & Martínez, Á. T. (2009). Microbial degradation of lignin: how a bulky recalcitrant polymer is efficiently recycled in nature and how we can take advantage of this. *Microbial biotechnology*, *2*(2), 164- 177.

Saha, B. C., Iten, L. B., Cotta, M. A., & Wu, Y. V. (2005). Dilute acid pretreatment, enzymatic saccharification, and fermentation of rice hulls to ethanol. *Biotechnology Progress*, *21*(3), 816- 822.

Sahadevan, L. D. M., Misra, C. S., & Thankamani, V. (2016). Characterization of lignin-degrading enzymes (LDEs) from a dimorphic novel fungus and identification of products of enzymatic breakdown of lignin. *Biotech*, *6*(1), 56.

Sahinkaya, M., Colak, D. N., Ozer, A., Canakci, S., Deniz, I., & Belduz, A. O. (2019). Cloning, characterization and paper pulp applications of a newly isolated DyP type peroxidase from Rhodococcus sp. T1. *Molecular biology reports*, *46*(1), 569-580.

Saini, S., Meghendra, S., & Kumar, A. (2018). Global warming and climate change: next generation biofuels and role of biotechnology. *Intl J Life Sci Pharma Res*, *8*(2), 52-57.

Saka, S., Munusamy, M. V., Shibata, M., Tono, Y., & Miyafuji, H. (2008). Chemical constituents of the different anatomical parts of the oil palm (Elaeis guineensis) for their sustainable utilization.

Sanders, J. G., Powell, S., Kronauer, D. J., Vasconcelos, H. L., Frederickson, M. E., & Pierce, N. E. (2014). Stability and phylogenetic correlation in gut microbiota: lessons from ants and apes. *Molecular Ecology*, *23*(6), 1268-1283.

Santo Domingo, J. W. (1998). Use of 16S rDNA community fingerprints to study cricket hindgut microbial communities. *The Journal of General and Applied Microbiology*, *44*(2), 119- 127.

Santos, A., Mendes, S., Brissos, V., & Martins, L. O. (2014). New dye-decolorizing peroxidases from Bacillus subtilis and Pseudomonas putida MET94: towards biotechnological applications. *Applied microbiology and biotechnology*, *98*(5), 2053-2065.

Santos, R. B., Hart, P., Jameel, H., & Chang, H. M. (2013). Wood based lignin reactions important to the biorefinery and pulp and paper industries. *BioResources*, *8*(1), 1456-1477.

Saraswathy, N., & Ramalingam, P. (2011). *Concepts and techniques in genomics and proteomics*. Elsevier.

Sarma, S. J., Brar, S. K., Le Bihan, Y., Buelna, G., & Soccol, C. R. (2014). Mitigation of the inhibitory effect of soap by magnesium salt treatment of crude glycerol–A novel approach for enhanced biohydrogen production from the biodiesel industry waste. *Bioresource technology*, *151*, 49- 53.

Scaife, M. A., Nguyen, G. T., Rico, J., Lambert, D., Helliwell, K. E., & Smith, A. G. (2015). Establishing Chlamydomonas reinhardtii as an industrial biotechnology host. *The Plant Journal*, *82*(3), 532- 546.

Scheller, H. V., & Ulvskov, P. (2010). Hemicelluloses. *Annual Review of Plant Biology*, *61*, 263- 289.

Schloss, P. D., Delalibera Jr, I., Handelsman, J. O., & Raffa, K. F. (2006). Bacteria associated with the guts of two wood-boring beetles: Anoplophora glabripennis and Saperda estita (Cerambycidae). *Environmental Entomology*, *35*(3), 625-629.

Schmeisser, C., Steele, H., & Streit, W. R. (2007). Metagenomics, biotechnology with non- culturable microbes. *Applied microbiology and biotechnology*, *75*(5), 955-962.

Schmitt, J., Hess, H., & Stunnenberg, H. G. (1993). Affinity purification of histidine-tagged proteins. *Molecular biology reports*, *18*(3), 223-230.

Schoenherr, S., Ebrahimi, M., & Czermak, P. (2018). Lignin degradation processes and the purification of valuable products. *Lignin-Trends and Applications*.

Schumann, W., & Ferreira, L. C. S. (2004). Production of recombinant proteins in Escherichia coli. *Genetics and Molecular Biology*, *27*(3), 442-453.

Scully, E. D., Geib, S. M., Hoover, K., Tien, M., Tringe, S. G., Barry, K. W., Glavina del Rio, T., Chovatia, M., Herr, J. R., & Carlson, J. E. (2013). Metagenomic profiling reveals lignocellulose degrading system in a microbial community associated with a wood-feeding beetle. *PLoS One*, *8*(9), e73827.

Seeman, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068-2069.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, *9*(8), 811.

Shan, H., Wu, W., Sun, Z., Chen, J., & Li, H. (2021). The Gut Microbiota of the Insect Infraorder Pentatomomorpha (Hemiptera: Heteroptera) for the Light of Ecology and Evolution. *Microorganisms*, *9*(2), 464.

Sharma, N., Kumar, J., Abedin, M., Sahoo, D., Pandey, A., Rai, A. K., & Singh, S. P. (2020). Metagenomics revealing molecular profiling of community structure and metabolic pathways in natural hot springs of the Sikkim Himalaya. *BMC microbiology*, *20*(1), 1-17.

Sharma, V. K., Kumar, N., Prakash, T., & Taylor, T. D. (2010). MetaBioME: a database to explore commercially useful enzymes in metagenomic datasets. *Nucleic acids research*, *38*(suppl_1), D468-D472.

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in plant science*, *5*, 209.

Sharpton, T. J., Riesenfeld, S. J., Kembel, S. W., Ladau, J., O'Dwyer, J. P., Green, J. L., ... & Pollard, K. S. (2011). PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS computational biology*, *7*(1), e1001061.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, *26*(10), 1135-1145.

Shewa, W. A., Lalman, J. A., Chaganti, S. R., & Heath, D. D. (2016). Electricity production from lignin photocatalytic degradation byproducts. *Energy*, *111*, 774-784.

Shi, W., Xie, S., Chen, X., Sun, S., Zhou, X., Liu, L., ... & Yuan, J. S. (2013). Comparative genomic analysis of the endosymbionts of herbivorous insects reveals eco-environmental adaptations: biotechnology applications. *PLoS genetics*, *9*(1), e1003131.

Shi, Y., Chai, L., Tang, C., Yang, Z., Zheng, Y., Chen, Y., & Jing, Q. (2013). Biochemical investigation of kraft lignin degradation by Pandoraea sp. B-6 isolated from bamboo slips. *Bioprocess and biosystems engineering*, *36*(12), 1957-1965.

Sigoillot, J. C., Petit-Conil, M., Ruel, K., Moukha, S., Comtat, J., Laugero, C., ... & Asther, M. (1997). Enzymatic treatment with manganese peroxidase from Phanerochaete chrysosporium for enhancing wheat straw pulp characteristics.

Silva, C. O., Vaz, R. P., & Filho, E. X. (2018). Bringing plant cell wall-degrading enzymes into the lignocellulosic biorefinery concept. *Biofuels, Bioproducts and Biorefining*, *12*(2), 277-289.

Simon, C., & Daniel, R. (2009). Achievements and new knowledge unraveled by metagenomic approaches. *Applied microbiology and biotechnology*, *85*(2), 265-276.

Simon, C., & Daniel, R. (2011). Metagenomic analyses: past and future trends. *Appl. Environ. Microbiol.*, *77*(4), 1153-1161.

Sindhu, R., Binod, P., & Pandey, A. (2016). Biological pretreatment of lignocellulosic biomass– An overview. *Bioresource Technology*, *199*, 76-82.

Singh, K. M., Reddy, B., Patel, D., Patel, A. K., Parmar, N., Patel, A., ... & Joshi, C. G. (2014). High potential source for biomass degradation enzyme discovery and environmental aspects revealed through metagenomics of Indian buffalo rumen. *BioMed research international*, *2014*.

Sleator, R. D., Shortall, C., & Hill, C. (2008). Metagenomics. *Letters in Applied Microbiology*, *47*(5), 361- 366.

Socha, A. M., Parthasarathi, R., Shi, J., Pattathil, S., Whyte, D., Bergeron, M., ... & Hahn, M. G. (2014). Efficient biomass pretreatment using ionic liquids derived from lignin and hemicellulose. *Proceedings of the National Academy of Sciences*, *111*(35), E3587- E3595.

Sontowski, R., & van Dam, N. M. (2020). Functional variation in Dipteran gut bacterial communities in relation to their diet, life cycle stage and habitat. *Insects*, *11*(8), 543.

Steele, H. L., Jaeger, K. E., Daniel, R., & Streit, W. R. (2009). Advances in recovery of novel biocatalysts from metagenomes. *Journal of molecular microbiology and biotechnology*, *16*(1- 2), 25-37.

Stefan Tangermann. What's causing global food price inflation? Vox,http://www.voxeu.org/index.php?q¼node/1437; 22 July 2008

Stephenson, M. (2018). The Carbon Cycle, Fossil Fuels and Climate Change. *Energy Clim. Chang*, 1-26.

Stewart, E. J. (2012). Growing unculturable bacteria. *Journal of bacteriology*, *194*(16), 4151-4160.

Stewart, J. J., Akiyama, T., Chapple, C., Ralph, J., & Mansfield, S. D. (2009). The effects on lignin structure of overexpression of ferulate 5-hydroxylase in hybrid poplar1. *Plant Physiology*, *150*(2), 621-635.

Streit R. W. and Daniel R. (2010). Metagenomics; Methods and protocols. Springer protocols. Humana press.

Studier, F. W., & Moffatt, B. A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of molecular biology*, *189*(1), 113-130.

Su, X., Schmitz, G., Zhang, M., Mackie, R. I., & Cann, I. K. (2012). Heterologous gene expression in filamentous fungi. *Advances in applied microbiology*, *81*, 1-61.

Sugano, Y., Muramatsu, R., Ichiyanagi, A., Sato, T., & Shoda, M. (2007). DyP, a unique dye-decolorizing peroxidase, represents a novel heme peroxidase family:

ASP171 replaces the distal histidine of classical peroxidases. *Journal of Biological Chemistry*, *282*(50), 36652-36658.

Sugio, A., Dubreuil, G., Giron, D., & Simon, J. C. (2015). Plant–insect interactions under bacterial influence: ecological implications and underlying mechanisms. *Journal of experimental botany*, *66*(2), 467-478.

Sun, J., & Zhou, X. J. (2011). Utilization of lignocellulose-feeding insects for viable biofuels: an emerging and promising area of entomological science. In *Recent advances in entomological research* (pp. 434-500). Springer, Berlin, Heidelberg.

Suranovic, S., (2013) Fossil fuel addiction and the implications for climate change policy. *Global Environmental Change*, *23*(3), pp.598-608.

Sutherland, G. R., Zapanta, L. S., Tien, M., & Aust, S. D. (1997). Role of calcium in maintaining the heme environment of manganese peroxidase. *Biochemistry*, *36*(12), 3654-3662.

Tagliavia, M., Messina, E., Manachini, B., Cappello, S., & Quatrini, P. (2014). The gut microbiota of larvae of Rhynchophorus ferrugineus Oliver (Coleoptera: Curculionidae). *BMC microbiology*, *14*(1), 136.

Takkellapati, S., Li, T., & Gonzalez, M. A. (2018). An overview of biorefinery-derived platform chemicals from a cellulose and hemicellulose biorefinery. *Clean technologies and environmental policy*, *20*(7), 1615-1630.

Tanveer, A., Yadav, S., & Yadav, D. (2016). Comparative assessment of methods for metagenomic DNA isolation from soils of different crop growing fields. *3 Biotech*, *6*(2), 220.

Tasse, L., Bercovici, J., Pizzut-Serin, S., Robe, P., Tap, J., Klopp, C., ... & Potocki-Veronese, G. (2010). Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome research*, *20*(11), 1605-1612.

Temitope, O. O. (2013). Morphology and histology of the alimentary tract of adult palm weevil, Rhynchophorus phoenicis Fabricius (Coleoptera:

curculionidae). *Journal of Developmental Biology and Tissue Engineering*, *5*(2), 13-17.

Tenenbaum, D. J. (2008). Food vs. fuel: diversion of crops could cause more hunger. Environ Health Perspect. 2008 Jun; 116(6): A254–A257. doi: 10.1289/ehp.116-a254

Thomas, C. N., & Dimkpa, S. O. N. (2016). Nutrients associated with breeding of African Palm weevil (Rhynchophorus phoenicis) in oil palm (Elaeis guineensisjacq.). *Acta Agronomica Nigeriana*, *16*(1/2), 71-80.

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics-a guide from sampling to data analysis. *Microbial informatics and experimentation*, *2*(1), 3.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., ... & Knight, R. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, *551*(7681), 457-463.

Tien, M., & Kirk, T. K. (1983). Lignin-degrading enzyme from the hymenomycete Phanerochaete chrysosporium Burds. *Science*, *221*(4611), 661-663.

Tinker, K. A., & Ottesen, E. A. (2018). The hindgut microbiota of praying mantids is highly variable and includes both prey-associated and host-specific microbes. *PloS one*, *13*(12), e0208917.

Tong, W. (2010). Wind power generation and wind turbine design. WIT press.

Trache, D., Hussin, M. H., Haafiz, M. M., & Thakur, V. K. (2017). Recent progress in cellulose nanocrystals: sources and production. *Nanoscale*, *9*(5), 1763-1786.

Tsegaye, B., Balomajumder, C., & Roy, P. (2019). Microbial delignification and hydrolysis of lignocellulosic biomass to enhance biofuel production: an overview and future prospect. *Bulletin of the National Research Centre*, *43*(1), 1-16.

Uchida, T., Sasaki, M., Tanaka, Y., & Ishimori, K. (2015). A dye-decolorizing peroxidase from Vibrio cholerae. *Biochemistry*, *54*(43), 6610-6621.

UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, *47*(D1), D506- D515.

UniProt: the universal protein knowledgebase in 2021." *Nucleic acids research* 49, no. D1 (2021): D480-D489.

USDOE- United States Department of Energy- Biomass Multi-Year Program plan, 2008.

Usino, D. O., Ylitervo, P., Dou, J., Sipponen, M. H., & Richards, T. (2020). Identifying the primary reactions and products of fast pyrolysis of alkali lignin. *Journal of Analytical and Applied Pyrolysis*, *151*, 104917.

Valzano, M., Achille, G., Burzacca, F., Damiani, C., Scuppa, P., Ricci, I., & Favia, G. (2012). Deciphering microbiota associated to Rhynchophorus ferrugineus in Italian samples: a preliminary study. *Journal of Entomological and Acarological Research*, *44*(3), e16-e16.

van Bloois, E., Torres Pazmiño, D. E., Winter, R. T., & Fraaije, M. W. (2010). A robust and extracellular heme- containing peroxidase from Thermobifida fusca as prototype of a bacterial peroxidase superfamily. *Applied microbiology and biotechnology*, *86*(5), 1419-1430.

Van der Lelie, D., Taghavi, S., McCorkle, S. M., Li, L. L., Malfatti, S. A., Monteleone, D., ... & Tringe, S. G. (2012). The metagenome of an anaerobic microbial community decomposing poplar wood chips. *PloS one*, *7*(5), e36740.

Van der Walt, A. J., Van Goethem, M. W., Ramond, J. B., Makhalanyane, T. P., Reva, O., & Cowan, D. A. (2017). Assembling metagenomes, one community at a time. *BMC genomics*, *18*(1), 1-13.

Vanholme, R., Demedts, B., Morreel, K., Ralph, J., & Boerjan, W. (2010). Update on Lignin Biosynthesis and Structure Lignin Biosynthesis and Structure 1.

Vares, T., Kalsi, M., & Hatakka, A. (1995). Lignin peroxidases, manganese peroxidases, and other ligninolytic enzymes produced by Phlebia radiata during solid-state fermentation of wheat straw. *Applied and Environmental Microbiology*, *61*(10), 3515-3520.

Vasco-Correa, J., Ge, X., & Li, Y. (2016). Biological Pretreatment of Lignocellulosic Biomass. In *Biomass Fractionation Technologies for a Lignocellulosic Feedstock Based Biorefinery* (pp. 561-585).

Vatanparast, M., Hosseininaveh, V., Ghadamyari, M., & Sajjadian, S. M. (2014). Plant cell wall degrading enzymes, pectinase and cellulase, in the digestive system of the red palm weevil, *Rhynchophorus ferrugineus* (Coleoptera: Curculionidae). *Plant Protection Science*, *50*(4), 190-198.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., ... & Fouts, D. E. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *science*, *304*(5667), 66-74.

Vestergaard, G., Schulz, S., Schöler, A., & Schloter, M. (2017). Making big data smart—how to use metagenomics to understand soil quality. 479- 484.

Vuong, T. V., Singh, R., Eltis, L. D., & Master, E. R. (2021). The comparative abilities of a small laccase and a dye-decoloring peroxidase from the same bacterium to transform natural and technical lignins. *Frontiers in microbiology*, 3101.

Walker, A. W., Martin, J. C., Scott, P., Parkhill, J., Flint, H. J., & Scott, K. P. (2015). 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome*, *3*(1), 1-11.

Walters, W., Hyde, E. R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., ... & Knight, R. (2016). Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. *Msystems*, *1*(1), e00009-15.

Wang, X. V., Blades, N., Ding, J., Sultana, R., & Parmigiani, G. (2012). Estimation of sequencing error rates in short reads. *BMC bioinformatics*, *13*(1), 1-12.

Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., ... & Leadbetter, J. R. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, *450*(7169), 560-565.

Watanabe, H., & Tokuda, G. (2010). Cellulolytic systems in insects. *Annual review of entomology*, *55*, 609-632.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., ... & Kenyon, R. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. 42 (D1), D581–D591.

Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of bacteriology*, *173*(2), 697-703.

Weiss, R., Guebitz, G. M., Pellis, A., & Nyanhongo, G. S. (2020). Harnessing the power of enzymes for tailoring and valorizing lignin. *Trends in Biotechnology*, *38*(11), 1215-1231.

Welinder, K. G. (1992). Superfamily of plant, fungal and bacterial peroxidases. *Current Opinion in Structural Biology*, *2*(3), 388-393.

West, T. P. (2017). Microbial production of malic acid from biofuel-related coproducts and biomass. *Fermentation*, *3*(2), 14.

Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016 2016.

Willis, J. D. (2009). Identification and characterization of novel cellulases from Dissosteira carolina (Orthoptera: Acrididae) and molecular cloning and expression of an endo- beta- 1, 4- glucanase from *Tribolium castaneum* (Coleoptera: Tenebrionidae).

Willis, J. D., Klingeman, W. E., Oppert, C., Oppert, B., & Jurat-Fuentes, J. L. (2010). Characterization of cellulolytic activity from digestive fluids of Dissosteira carolina (Orthoptera: Acrididae). *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *157*(3), 267-272.

Woese, C. R. (1987). Bacterial evolution. *Microbiological reviews*, *51*(2), 221-271.

Womeni, H. M., Tiencheu, B., Linder, M., Nabayo, E. M. C., Tenyang, N., Mbiapo, F. T., Villeneuve, P., Fanni, J., & Parmentier, M. (2012). Nutritional value and effect of cooking, drying and storage process on some functional properties of Rhynchophorus phoenicis. *International Journal of Life Science and Pharma Research*, *2*(3), 203-219.

Wommack, K. E., Bhavsar, J., & Ravel, J. (2008). Metagenomics: read length matters. *Applied and environmental microbiology*, *74*(5), 1453-1463.

Wong, D. W. (2009). Structure and action mechanism of ligninolytic enzymes. *Applied biochemistry and biotechnology*, *157*(2), 174-209.

World Bank. Rising food prices: policy options and world bank response,http://siteresources.worldbank.org/NEWS/Resources/risingfoodprices_backgroundnote _apr08.pdf; 2008.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., ... & Eisen, J. A. (2009). A phylogeny-

Xia, Y., Ju, F., Fang, H. H., & Zhang, T. (2013). Mining of novel thermo-stable cellulolytic genes from a thermophilic cellulose-degrading consortium by metagenomics. *PloS one*, *8*(1), e53779.

Xie, S., Syrenne, R., Sun, S., & Yuan, J. S. (2014). Exploration of natural biomass utilization systems (NBUS) for advanced biofuel—from systems biology to synthetic design. *Current opinion in biotechnology*, *27*, 195-203.

Xu, L., Sun, J., Qaria, M. A., Gao, L., & Zhu, D. (2021). Dye decoloring peroxidase structure, catalytic properties and applications: Current advancement and futurity. *Catalysts*, *11*(8), 955.

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329-342.

Yang, B., Wang, Y., & Qian, P. Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC bioinformatics*, *17*(1), 135.

Yang, C., Yue, F., Cui, Y., Xu, Y., Shan, Y., Liu, B., ... & Lü, X. (2018). Biodegradation of lignin by Pseudomonas sp. Q18 and the characterization of a novel bacterial DyP-type peroxidase. *Journal of Industrial Microbiology and Biotechnology*, *45*(10), 913-927.

Yang, Y. S., Zhou, J. T., Lu, H., Yuan, Y. L., & Zhao, L. H. (2011). Isolation and characterization of a fungus Aspergillus sp. strain F-3 capable of degrading alkali lignin. *Biodegradation*, *22*(5), 1017-1027.

Yao, S., Yang, Y.Y., Song, H., Wang, Y. & Wan, H.Q., (2015) Quantitative industrial analysis of lignocellulosic composition in typical agro-residues and extraction of inner hemicelluloses with ionic liquid.

Yasuda, S., Suenaga, T., Orschler, L., Agrawal, S., Lackner, S., & Terada, A. (2021). Metagenomic Insights into Functional and Taxonomic Compositions of an Activated Sludge Microbial Community Treating Leachate of a Completed Landfill: A Pathway-Based Analysis. *Frontiers in microbiology*, *12*, 640848.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., ... & Glöckner, F. O. (2014). The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. *Nucleic acids research*, *42*(D1), D643- D648.

Yoshida, T., & Sugano, Y. (2015). A structural and functional perspective of DyP-type peroxidase family. *Archives of biochemistry and biophysics*, *574*, 49-55.

Yu, Z., Li, S., Li, Y., Jiang, Z., Zhou, J., & An, Q. (2018). Complete genome sequence of N2- fixing model strain Klebsiella sp. nov. M5al, which produces plant cell wall- degrading enzymes and siderophores. *Biotechnology Reports*, *17*, 6-9.

Yun, J. H., Roh, S. W., Whon, T. W., Jung, M. J., Kim, M. S., Park, D. S., ... & Kim, J. Y. (2014). Insect gut bacterial diversity determined by environmental habitat, diet, developmental stage, and phylogeny of host. *Appl. Environ. Microbiol.*, *80*(17), 5254-5264.

Zheng, X., Zhu, Q., Zhou, Z., Wu, F., Chen, L., Cao, Q., & Shi, F. (2021). Gut bacterial communities across 12 Ensifera (Orthoptera) at different feeding habits and its prediction for the insect with contrasting feeding habits. *Plos one*, *16*(4), e0250675.

Zhu, D., Zhang, P., Xie, C., Zhang, W., Sun, J., Qian, W. J., & Yang, B. (2017). Biodegradation of alkaline lignin by Bacillus ligniniphilus L1. *Biotechnology for biofuels*, *10*(1), 1-14.

Zhu, L., Wu, Q., Dai, J., Zhang, S., & Wei, F. (2011). Evidence of cellulose metabolism by the giant panda gut microbiome. *Proceedings of the National Academy of Sciences*, USA. *108*(43), 17714-17719.

Zubieta, C., Krishna, S. S., Kapoor, M., Kozbial, P., McMullan, D., Axelrod, H. L., ... & Wilson, I. A. (2007). Crystal structures of two novel dye-decolorizing peroxidases reveal a β-barrel fold with a conserved heme- binding motif. *Proteins: Structure, Function, and Bioinformatics*, *69*(2), 223-233.

Appendices

Appendix 1: ASV table generated from analysis of 16S amplicon sequencing data

Column1	Kingdom	Phylum	Class	Order	Family	Genus	Total	TotalPercentage	Foregut	ForegutPercentage	Midgut	MidgutPercentage	Hindgut	HindgutPercentage	Column2
SV 2	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Family_En	126461	25.99	48423	23.57	41271	24.85	36767	31.97	
SV 5	Bacteria	Firmicutes	Bacilli	Lactobacil	Lactobacil	Lactobacil	77524	15.93	49750	24.22	26942	16.22	832	0.72	
SV 1	Bacteria	Firmicutes	Bacilli	Lactobacil	Enterococ	Enterococ	56119	11.54	23618	11.5	13771	8.29	18730	16.29	
SV 3	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Salmonell	38431	7.9	15988	7.78	18363	11.06	4080	3.55	
SV 11	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Klebsiella	36984	7.6	15471	7.53	12563	7.56	8950	7.78	
SV 4	Bacteria	Firmicutes	Negativici	Selenomc	Veillonell	Megaspha	26107	5.37	13174	6.41	12754	7.68	179	0.16	
SV 7	Bacteria	Firmicutes	Bacilli	Lactobacil	Streptoco	Lactococcu	20674	4.25	4774	2.32	8072	4.86	7828	6.81	
SV 10	Bacteria	Firmicutes	Negativici	Selenomc	Veillonell	Pectinatus	18405	3.78	10872	5.29	7395	4.45	138	0.12	
SV 9	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Lelliottia	18207	3.74	8399	4.09	6039	3.64	3769	3.28	
SV 20	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Enterobac	9979	2.05	3704	1.8	3850	2.32	2425	2.11	
SV 26	Bacteria	Tenericute	Mollicute	Entomopl	Entomopl	Entomopl	7183	1.48	1203	0.59	5217	3.14	763	0.66	
SV 32	Bacteria	Proteobact	Alphaprot	Rhodobac	Rhodobac	Family_Rh	7069	1.45	461	0.22	468	0.28	6140	5.34	
SV 39	Bacteria	Bacteroide	Bacteroidi	Bacteroidi	Dysgonon	Dysgonon	6380	1.31	28	0.01	441	0.27	5911	5.14	
SV 29	Bacteria	Firmicutes	Erysipelot	Erysipelot	Erysipelot	Erysipelot	5314	1.09	0	0	49	0.03	5265	4.58	
SV 31	Bacteria	Firmicutes	Negativici	Selenomc	Veillonell	Family_Ve	4084	0.84	1895	0.92	2172	1.31	17	0.01	
SV 53	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Citrobacte	3723	0.77	948	0.46	1308	0.79	1467	1.28	
SV 41	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Cronobact	3650	0.75	1529	0.74	1569	0.94	552	0.48	
SV 58	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Morganell	2094	0.43	94	0.05	198	0.12	1802	1.57	
SV 47	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Serratia	1992	0.41	26	0.01	987	0.59	979	0.85	
SV 45	Bacteria	Actinobact	Actinobac	Coryneba	Coryneba	Coryneba	1711	0.35	1350	0.66	15	0.01	346	0.3	
SV 55	Bacteria	Actinobact	Actinobac	Micrococc	Promicror	Family_Pr	1704	0.35	0	0	19	0.01	1685	1.47	
SV 107	Bacteria	Bacteroide	Bacteroidi	Bacteroidi	Bacteroidi	Bacteroidi	1543	0.32	179	0.09	385	0.23	979	0.85	
SV 56	Bacteria	Actinobact	Actinobac	Bifidobact	Bifidobact	Gardnerel	992	0.2	423	0.21	427	0.26	142	0.12	
SV 60	Bacteria	Proteobact	Alphaprot	Rhizobiale	Rhizobiac	Ochrobact	895	0.18	99	0.05	113	0.07	683	0.59	
SV 59	Bacteria	Firmicutes	Bacilli	Lactobacil	Leuconost	Leuconost	851	0.17	353	0.17	427	0.26	71	0.06	
SV 84	Bacteria	Bacteroide	Bacteroidi	Flavobacti	Weeksell	Family_W	791	0.16	642	0.31	118	0.07	31	0.03	
SV 79	Bacteria	Proteobact	Alphaprot	Rhodobac	Rhodobac	Gemmoba	619	0.13	0	0	77	0.05	542	0.47	
SV 140	Bacteria	Proteobact	Gammapr	Pseudom	Moraxella	Acinetoba	476	0.1	341	0.17	82	0.05	53	0.05	
SV 81	Bacteria	Proteobact	Alphaprot	Rhodobac	Rhodobac	Defluviim	441	0.09	0	0	23	0.01	418	0.36	
SV 121	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Yokenella	418	0.09	33	0.02	79	0.05	306	0.27	
SV 111	Bacteria	Proteobact	Alphaprot	Rhizobiale	Rhizobiac	Allorhizot	415	0.09	30	0.01	32	0.02	353	0.31	
SV 85	Bacteria	Bacteroide	Bacteroidi	Sphingobi	Sphingobi	Sphingoba	396	0.08	173	0.08	102	0.06	121	0.11	
SV 90	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Providenc	357	0.07	0	0	0	0	357	0.31	
SV 274	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Mangrovil	321	0.07	70	0.03	14	0.01	237	0.21	
SV 97	Bacteria	Bacteroide	Bacteroidi	Flavobacti	Weeksell	Moheibac	297	0.06	177	0.09	51	0.03	69	0.06	
SV 98	Bacteria	Proteobact	Gammapr	Betaprote	Burkholde	Family_Bu	290	0.06	69	0.03	25	0.02	196	0.17	
SV 130	Bacteria	Proteobact	Alphaprot	Rhodobac	Rhodobac	Paracoccu	252	0.05	72	0.04	0	0	180	0.16	
SV 272	Bacteria	Bacteroide	Bacteroidi	Bacteroidi	Prevotella	Prevotella	237	0.05	222	0.11	15	0.01	0	0	
SV 147	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Pseudocit	233	0.05	26	0.01	22	0.01	185	0.16	
SV 117	Bacteria	Firmicutes	Bacilli	Bacillales	Staphyloc	Staphyloc	202	0.04	202	0.1	0	0	0	0	
SV 124	Bacteria	Firmicutes	Negativici	Selenomo	Acidamin	Family_Ac	189	0.04	0	0	0	0	189	0.16	
SV 199	Bacteria	Actinobact	Actinobac	Micrococc	Order_Mi	Order_Mi	164	0.03	10	0	0	0	154	0.13	
SV 175	Bacteria	Actinobact	Actinobac	Coryneba	Nocardiac	Gordonia	163	0.03	88	0.04	20	0.01	55	0.05	
SV 134	Bacteria	Proteobact	Alphaprot	Acetobact	Acetobact	Roseomor	159	0.03	17	0.01	22	0.01	120	0.1	
SV 225	Bacteria	Firmicutes	Clostridia	Clostridial	Ruminoco	Faecalibac	149	0.03	41	0.02	34	0.02	74	0.06	
SV 170	Bacteria	Firmicutes	Clostridia	Clostridial	Lachnospi	Blautia	131	0.03	20	0.01	111	0.07	0	0	
SV 229	Bacteria	Actinobact	Actinobac	Micrococc	Microbact	Microbact	119	0.02	31	0.02	0	0	88	0.08	
SV 115	Bacteria	Proteobact	Alphaprot	Sphingom	Sphingom	Sphingobi	119	0.02	31	0.02	16	0.01	72	0.06	
SV 228	Bacteria	Proteobact	Alphaprot	Acetobact	Acetobact	Acetobact	89	0.02	0	0	89	0.05	0	0	
SV 168	Bacteria	Proteobact	Alphaprot	Rhodobac	Rhodobac	Haematot	84	0.02	0	0	0	0	84	0.07	
SV 226	Bacteria	Firmicutes	Clostridia	Clostridial	Lachnospi	Family_La	82	0.02	21	0.01	54	0.03	7	0.01	
SV 173	Bacteria	Proteobact	Alphaprot	Rhodobac	Rhodobac	Thioclava	77	0.02	0	0	0	0	77	0.07	
SV 185	Bacteria	Bacteroide	Bacteroidi	Chitinoph	Order_Chi	Order_Chi	74	0.02	59	0.03	11	0.01	4	0	
SV 179	Bacteria	Actinobact	Actinobac	Micrococc	Rarobacte	Rarobacte	70	0.01	0	0	0	0	70	0.06	
SV 180	Bacteria	Actinobact	Actinobac	Micrococc	Demequir	Demequir	69	0.01	0	0	0	0	69	0.06	
SV 213	Bacteria	Proteobact	Gammapr	Betaprote	Neisseria	Snodgrass	43	0.01	32	0.02	11	0.01	0	0	
SV 216	Bacteria	Proteobact	Alphaprot	Sphingom	Sphingom	Novosphi	42	0.01	0	0	19	0.01	23	0.02	
SV 219	Bacteria	Firmicutes	Bacilli	Lactobacil	Streptoco	Streptoco	39	0.01	0	0	39	0.02	0	0	
SV 285	Bacteria	Firmicutes	Bacilli	Lactobacil	Streptoco	Anthococ	37	0.01	12	0.01	6	0	19	0.02	
SV 220	Bacteria	Proteobact	Gammapr	Betaprote	Burkholde	Corticibac	37	0.01	0	0	0	0	37	0.03	
SV 237	Bacteria	Actinobact	Actinobac	Propionib	Nocardioi	Nocardioi	37	0.01	0	0	0	0	37	0.03	
SV 231	Bacteria	Proteobact	Deltaprot	Desulfovil	Desulfovil	Desulfovil	31	0.01	10	0	21	0.01	0	0	
SV 214	Bacteria	Bacteroide	Bacteroidi	Bacteroidi	Tannerell	Parabacte	31	0.01	9	0	0	0	22	0.02	
SV 252	Bacteria	Actinobact	Actinobac	Micrococc	Microbact	Pseudocla	31	0.01	31	0.02	0	0	0	0	
SV 232	Bacteria	Firmicutes	Clostridia	Clostridial	Peptostre	Rombouts	31	0.01	19	0.01	12	0.01	0	0	
SV 235	Bacteria	Firmicutes	Clostridia	Clostridial	Lachnospi	Fusicateni	29	0.01	17	0.01	0	0	12	0.01	
SV 277	Bacteria	Firmicutes	Bacilli	Lactobacil	Enterococ	Melissoco	26	0.01	6	0	7	0	13	0.01	
SV 346	Bacteria	Firmicutes	Clostridia	Clostridial	Lachnospi	Lachnospi	25	0.01	5	0	0	0	20	0.02	
SV 267	Bacteria	Firmicutes	Bacilli	Lactobacil	Order_Lac	Order_Lac	25	0.01	15	0.01	4	0	6	0.01	
SV 247	Bacteria	Epsilonbac	Campylob	Campylob	Arcobacte	Arcobacte	23	0	3	0	20	0.01	0	0	
SV 265	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Kosakonia	23	0	6	0	17	0.01	0	0	
SV 249	Bacteria	Bacteroide	Bacteroidi	Cytophag	Order_Cyt	Order_Cyt	22	0	22	0.01	0	0	0	0	
SV 250	Bacteria	Actinobact	Actinobac	Micrococc	Dermabac	Brachybac	21	0	21	0.01	0	0	0	0	
SV 255	Bacteria	Firmicutes	Clostridia	Clostridial	Lachnospi	Lachnospi	20	0	0	0	0	0	20	0.02	
SV 373	Bacteria	Bacteroide	Bacteroidi	Bacteroidi	Order_Ba	Order_Ba	19	0	4	0	2	0	13	0.01	
SV 260	Bacteria	Firmicutes	Clostridia	Clostridial	Lachnospi	Anaerosti	18	0	0	0	0	0	18	0.02	
SV 259	Bacteria	Proteobact	Alphaprot	Rhizobiale	Rhizobiac	Family_Rh	18	0	0	0	0	0	18	0.02	
SV 261	Bacteria	Firmicutes	Clostridia	Clostridial	Lachnospi	Lachnoclo	18	0	0	0	0	0	18	0.02	
SV 263	Bacteria	Proteobact	Alphaprot	Rhodobac	Rhodobac	Ketogulon	17	0	17	0.01	0	0	0	0	
SV 331	Bacteria	Bacteroide	Bacteroidi	Bacteroidi	Prevotella	Prevotella	17	0	10	0	7	0	0	0	
SV 296	Bacteria	Cyanobact	Oxyphoto	Chloropla	Order_Chl	Order_Chl	16	0	5	0	6	0	5	0	
SV 269	Bacteria	Proteobact	Alphaprot	Caulobact	Caulobact	Brevundir	15	0	0	0	6	0	9	0.01	
SV 360	Bacteria	Proteobact	Gammapr	Betaprote	Burkholde	Verticia	15	0	3	0	0	0	12	0.01	
SV 273	Bacteria	Firmicutes	Clostridia	Clostridial	Lachnospi	Lachnospi	14	0	0	0	0	0	14	0.01	
SV 283	Bacteria	Firmicutes	Clostridia	Clostridial	Clostridia	Clostridiu	13	0	0	0	13	0.01	0	0	
SV 288	Bacteria	Actinobact	Actinobac	Propionib	Nocardioi	Aeromicro	12	0	0	0	0	0	12	0.01	
SV 289	Bacteria	Firmicutes	Clostridia	Clostridial	Lachnospi	Agathoba	12	0	0	0	12	0.01	0	0	
SV 287	Bacteria	Patescibac	Saccharim	Saccharim	Order_Sac	Order_Sac	12	0	0	0	0	0	12	0.01	
SV 298	Bacteria	Proteobact	Alphaprot	Acetobact	Acetobact	Acidispha	11	0	0	0	11	0.01	0	0	
SV 297	Bacteria	Firmicutes	Clostridia	Clostridial	Lachnospi	Lachnospi	11	0	0	0	0	0	11	0.01	
SV 309	Bacteria	Bacteroide	Bacteroidi	Bacteroidi	Tannerell	Family_Ta	10	0	0	0	10	0.01	0	0	
SV 305	Bacteria	Bacteroide	Bacteroidi	Sphingobi	Sphingobi	Olivibacte	10	0	7	0	3	0	0	0	
SV 239	Bacteria	Firmicutes	Clostridia	Clostridial	Ruminoco	Ruminoco	10	0	0	0	0	0	10	0.01	
SV 312	Bacteria	Proteobact	Gammapr	Pseudom	Pseudom	Azorhizop	9	0	0	0	0	0	9	0.01	
SV 314	Bacteria	Actinobact	Coriobacti	Coriobacti	Order_Co	Order_Co	9	0	0	0	9	0.01	0	0	
SV 311	Bacteria	Proteobact	Alphaprot	Rhizobiale	Rhizobiac	Parameso	9	0	9	0	0	0	0	0	
SV 315	Bacteria	Firmicutes	Clostridia	Clostridial	Ruminoco	Ruminoco	9	0	0	0	9	0.01	0	0	
SV 332	Bacteria	Actinobact	Coriobacti	Coriobacti	Atopobiac	Atopobiu	7	0	0	0	7	0	0	0	
SV 334	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Atlantibac	6	0	6	0	0	0	0	0	
SV 335	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	endosymt	6	0	6	0	0	0	0	0	
SV 336	Bacteria	Proteobact	Gammapr	Xanthomo	Xanthomc	SN8	6	0	6	0	0	0	0	0	
SV 355	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Trabulsiel	6	0	0	0	6	0	0	0	
SV 339	Bacteria	Proteobact	Alphaprot	Caulobact	Hyphomo	UKL13-1	6	0	0	0	0	0	6	0.01	
SV 350	Bacteria	Kingdom_E	Kingdom_	Kingdom_	Kingdom_	Kingdom_	5	0	0	0	5	0	0	0	
SV 343	Bacteria	Firmicutes	Clostridia	Clostridial	Ruminoco	Ruminoco	5	0	5	0	0	0	0	0	
SV 352	Bacteria	Firmicutes	Clostridia	Clostridial	Clostridia	Family_Cl	4	0	4	0	0	0	0	0	
SV 376	Bacteria	Proteobact	Gammapr	Enterobac	Enterobac	Franconib	4	0	0	0	2	0	2	0	
SV 361	Bacteria	Firmicutes	Clostridia	Clostridial	Order_Clo	Order_Clo	3	0	3	0	0	0	0	0	
SV 364	Bacteria	Proteobact	Gammapr	Betaprote	Burkholde	Parasutter	3	0	0	0	0	0	3	0	
SV 365	Bacteria	Proteobact	Gammapr	Xanthomc	Xanthomc	Stenotrop	3	0	0	0	3	0	0	0	
SV 375	Bacteria	Bacteroide	Bacteroidi	Bacteroidi	Rikenellac	Alistipes	2	0	0	0	2	0	0	0	
SV 377	Bacteria	Proteobact	Gammapr	Betaprote	Burkholde	Ottowia	2	0	0	0	2	0	0	0	
SV 368	Bacteria	Proteobact	Gammapr	Diplorickе	Diplorickе	Rickettsie	2	0	0	0	0	0	2	0	
SV 366	Bacteria	Proteobact	Gammapr	Xanthomc	Xanthomc	Thermom	2	0	2	0	0	0	0	0	
SV 382	Bacteria	Firmicutes	Bacilli	Lactobacil	Carnobact	Family_Ca	1	0	0	0	0	0	1	0	
SV 381	Bacteria	Firmicutes	Phylum_F	Phylum_F	Phylum_F	Phylum_F	1	0	0	0	0	0	1	0	

Appendix 2: Nucleotide sequences of selected genes amplified in this study

Below are the nucleotide sequences of genes selected from the functional annotation output file of APW gut bacterial metagenome for PCR amplification.

Appendix 2.1: Gene ID_30342 Polyphenol oxidase (732bp)

ATGAGTAAGCTAATTGTTCCGCAGTGGCCAATCCCGGAGGGGGTTGCAGCCTGTAGTTCTG
TACGTACAGGCGGTGTCAGCTTACCGCCTTACGATTCTCTGAACCTGGGCGCCCATTGCGG
CGATAACCCGGAACATGTGGAAGAGAATCGCAAGAGACTCTTTGCTGCGGGCAATTTGCCG
TCAAAACCCGTCTGGCTTGAACAGGTGCATGGCAAGGACGTGCTGAAACTCACCGGCGAG
CCTTATGCATCCAAACGTGCGGATGCGTCGTACAGCAATACCCCCGGCACCGTTTGCGCGG
TGATGACGGCGGACTGTTTGCCCGTGTTGTTTTGTAATCGCGCGGGAACGGAAGTGGCGG
CGGCTCATGCGGGCTGGCGCGGATTATGTGAAGGTGTGCTGGAAGAGACGGTCGCCTGCT
TTGCCGATAAACCTGAAAACATTATCGCCTGGCTGGGTCCGGCTATTGGCCCTGCTGCTTTT
GAAGTGGGGGCTGAAGTGCGTGACGCGTTTATCGCAAAAGATGCGAAGGCGAGCAGCGCA
TTTCAGGCGCGCGGTGAAAAATATCTGGCGGATATTTATCAGCTTGCGCGTCAACGTCTGGC
GAATGTGGGAGTTGAAAGCGTCTATGGCGGCGATCGCTGTACCTACAGCGAAAGTGAGACT
TTCTTCTCTTATCGTCGCGACAAGATCACAGGTCGTATGGCAAGTTTTATTTGGCTGATATAA

Appendix 2.2: Gene ID_38773 Deferrochelatase/peroxidase EfeB (1281bp)

ATGAACAGCAAGCAACAGGGACCAAGCCGGCGCCACGTCCTGATGGGGCTGGGCGCGGGGGCGGT
GGGTGCGATGGCGCCGCTGGCCGCCGCGGCGCAACAGGTGAATGACGCGCCCCATGGCGACGCCG
GCACCGCGGCGCAGCGGGTGGCGTTCTTCGGCCGCCATCAGGCGGGGGTCACCACGCCGCGCCCG
GCTTCGGGCATCGTTGCGGCCTTCGATTTGGCGGGCACCAGCCCGGACGGTTTCGCGCGGGTCATGC
GCGCGCTGACCGCACCCGCGCTGTTCCTGACGCAGGGGGGGGCGGTGCCCGGGCGCGACCCGAAG
TGGCCGCCCGCCGATTCCGGGCTGCTGGGGCCGGTCGTCGCGCCGGACAACCAGACGATCACCGTC
AGCCTCGGCAACGGGCTGTTCGAGCGCTTCGACTGGCTGCGGCCGCTGAAGCCCGTGCGCCTGCAG
CAGATGGTGCAGTTCCCGAATGACGCGCTGGTCGCGGATCTTTGCCACGGCGACATGACGATCCAGTT
CTGCGCGAACCTGCAGGACACGAATATCCACGCCCTGCGCGACCTGATGAAGAACCTGTCGGAATTCC
TCGTGATCCGCTGGATGCAGGAGGGCGACGTGCCCCCGGTGCCGCCCGCGCCGGATGGCTCGACGC
CTTCGGCACGGAATTTCCTGGGCTTCCGCGACGGGTCTGCGAACCCGGATTCGAACGATGCCGCGCT
GATGGAAAAGGTGGTCTGGGTCGGCGCCGGGGATGGCGAGCCGGTCTGGGCCGAGGGCGGCAGCT
ATCAGGTGGTGCGGCTGATCCGGAACATGGTCGAACGCTGGGACCGCACGCCGCTGCAGGAGCAGG
AGCGCGATTTCGGCCGGCGCAAGATGTCGGGCGCGCCGCTGGACGGCGGCGCGGATGCGACCGAG
CGCGACGTTCCCGATTATGCGCGGGACCCGGAGGGCAAGGCCACGCATCTGCTGTCGCACATCCGGC
TGGCCAATCCGCGCACGGCCGAGACCCAGAAGAACCTGATCCTGCGCCGCGCCTTCAACTACACCAA
CGGCGTGATGAAGAACGGGCAGCTGGATCAGGGCCTGCTGTTCATCTGCTATCAGGCCGATCTGGAG
GCCGGGTTCATCACCGTGCAGAACCGGCTGAACGGCGAGTTGATGGAAGAATATTTCAAACCGATCGG
CGGGGGGTATTTCTTCACCCTGCCGGGGGTGACGGGGCCGGGCGATTTCCTGGGCTCCGGCCTGATT
TCCGCAGCCAGGGCATTGTGA

Appendix 2.3: Gene ID_08687 putative deferrochelatase/peroxidase YfeX (900bp)

ATGTCTCAGGTTCAGAGCGGCATTTTGCCGGAACATTGCCGCGCGGCGATTTGGATTGAAGCCAATGTCA
AAGGGGACGTTAACGCCCTGCGCGAAGCGAGCAAAATTTTTGTCGATAAAGTGGCCACCTTTCAGGCTAA
ATTCCCTGACGCCAAACTCGGTGCCGTGGTGGCGTTCGGCAATAACGTCTGGCGTCAGCTGAGCGGCG
GCGAAGGGGCGGAAGAGTTAAAAGATTTTCCGGTCTACGGCAAAGGGCTGGCGCCGTCCACCCAGTAT
GACCTGCTGATTCATATTTTATCCGCCCGCCATGAAGTGAACTTCTCGGTAGCGCAGGCCGCGCTGGCCG
CCTTTGGCGACGCTATCCAGGTGAAAGAAGAGATCCACGGCTTCCGCTGGGTGGAAGAGCGTGACCTCA
GCGGCTTCGTCGACGGCACCGAAAACCCGGCGGGGGAAGAAACCCGGCGTGAAGTGGCGGTGATCAA
AGACGGCGTGGACGCGGGCGGCAGCTACGTGTTCGTCCAGCGCTGGGAGCATAATCTCAAACAGCTGA
ACCGCATGAGCGTGCCGGATCAGGAGATGATGATCGGCCGTACTAAAGACGCTAACGAAGAGATCGATG
GCGATGAACGTCCGGTCACGTCGCCCCTGAGCCGCGTGGACTTAAAAGAAGAGGGAAAAGGGCTGAAA
ATCGTCCGTCAGAGCCTGCCGTACGGCACCGCCAGCAGCACCCATGGCCTCTATTTCTGCGCCTACTGC
GCGCGCCTGTATAACATCGAGCAGCAGCTGCTGAGCATGTTTGGCGATACCGACGGTAAACGCGACGCG
ATGCTGCGCTTCACCAAACCGGTGACCGGCGGGTATTACTTCGCGCCGTCGCTGGAGCGTATCCCGGCG
CTGTAA

Appendix 3: Sequence Alignment

Pairwise alignment of sanger sequences on EMBOSS NEEDLE of gene B-38773 construct (B_T7F_F10) against original gene B-38773 sequences to confirm accuracy of cloning experiment carried out.

```
# Length: 1288
# Identity:    1164/1288 (90.4%)
# Similarity:  1164/1288 (90.4%)
# Gaps:          75/1288 ( 5.8%)
# Score: 7441.5
#
#
#=======================================

B_28_T7F_F10       1 CACCATGAACAGCAAGCAACAGGGACCAAGCCGGCGCCACGTCCTGATGG     50
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
B-38773            1 CACCATGAACAGCAAGCAACAGGGACCAAGCCGGCGCCACGTCCTGATGG     50

B_28_T7F_F10      51 GGCTGGGCGCGGGGGCGGTGGGCGCCATGGCGCCGCTGGCCGCCGCGGCG    100
                     ||||||||||||||||||||||.||.||||||||||||||||||||||||||
B-38773           51 GGCTGGGCGCGGGGGCGGTGGGTGCGATGGCGCCGCTGGCCGCCGCGGCG    100

B_28_T7F_F10     101 CAACAGGTGAATGACGCGCCCCACGGCGACGCCGGCACCGCGGCGCAGCG    150
                     ||||||||||||||||||||||.||||||||||||||||||||||||||||
B-38773          101 CAACAGGTGAATGACGCGCCCCATGGCGACGCCGGCACCGCGGCGCAGCG    150

B_28_T7F_F10     151 GGTGGCGTTCTTCGGCCGTCATCAGGCGGGGGTCACCACGCCGCGCCCGG    200
                     ||||||||||||||||||.||||||||||||||||||||||||||||||||
B-38773          151 GGTGGCGTTCTTCGGCCGCCATCAGGCGGGGGTCACCACGCCGCGCCCGG    200

B_28_T7F_F10     201 CTTCGGGCATCGTTGCGGCCTTCGATCTGGCGATCACCAGCTTGGACGAT    250
                     ||||||||||||||||||||||||||.||||||..||||||..||||||.|
B-38773          201 CTTCGGGCATCGTTGCGGCCTTCGATTTGGCGGGCACCAGCCCGGACGGT    250

B_28_T7F_F10     251 TTCGAGCGGATGATGCGCGCGCTGACCGAACGCGCGCTGTTCCTGACGCA    300
                     ||||.||||.|.|||||||||||||||.||.||||||||||||||||||||
B-38773          251 TTCGCGCGGGTCATGCGCGCGCTGACCGCACCCGCGCTGTTCCTGACGCA    300
```

```
B_28_T7F_F10     301 GGGCGGGGCCGTGCCCGAGCGCGACCCGAAGTTGCCGCCGGCGGATTCCG    350
                     |||.|||||.|||||||.|||||||||||||||.|||||||.||.|||||||
B-38773          301 GGGGGGGGCGGTGCCCGGGCGCGACCCGAAGTGGCCGCCCGCCGATTCCG    350

B_28_T7F_F10     351 GGCTGCTGGGGCCGGTCGTCGCGCCGGACAATCAGACGATCACCGTCAGC    400
                     |||||||||||||||||||||||||||||||.||||||||||||||||||
B-38773          351 GGCTGCTGGGGCCGGTCGTCGCGCCGGACAACCAGACGATCACCGTCAGC    400

B_28_T7F_F10     401 CTCGGCAACGGGCTGTTCGAGCGCTTCGACTGGCTGCGGCCGCTGAAGCC    450
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
B-38773          401 CTCGGCAACGGGCTGTTCGAGCGCTTCGACTGGCTGCGGCCGCTGAAGCC    450

B_28_T7F_F10     451 CGTGCGTCTGCAGCAGATGGTGCAGTTCCCGAATGACGCGCTGGTCGCGG    500
                     |||||||.|||||||||||||||||||||||||||||||||||||||||||
B-38773          451 CGTGCGCCTGCAGCAGATGGTGCAGTTCCCGAATGACGCGCTGGTCGCGG    500

B_28_T7F_F10     501 ATCTTTGCCACGGCGACATGACGATCCAGTTCTGCGCGAACCTGCAGGAC    550
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
B-38773          501 ATCTTTGCCACGGCGACATGACGATCCAGTTCTGCGCGAACCTGCAGGAC    550

B_28_T7F_F10     551 ACCAATATCCATGCCCTGCGCGACCTGATGAAGAACCTGTCGGAATTCCT    600
                     ||.||||||||.|||||||||||||||||||||||||||||||||||||||
B-38773          551 ACGAATATCCACGCCCTGCGCGACCTGATGAAGAACCTGTCGGAATTCCT    600

B_28_T7F_F10     601 CGTTATCCGCTGGATGCAGGAGGGCGACGTGCCCCCGGTGCCGCCCGCGC    650
                     |||.|||||||||||||||||||||||||||||||||||||||||||||||||
B-38773          601 CGTGATCCGCTGGATGCAGGAGGGCGACGTGCCCCCGGTGCCGCCCGCGC    650

B_28_T7F_F10     651 CGGATGGATCGACGCCTTCGGCACGGAATTTCCTGGGCTTCCGCGACGGG    700
                     |||||||.|||||||||||||||||||||||||||||||||||||||||||||
B-38773          651 CGGATGGCTCGACGCCTTCGGCACGGAATTTCCTGGGCTTCCGCGACGGG    700

B_28_T7F_F10     701 TCTGCGAACCCGGATTCGAACGATGCCGCGCTGATGGAAAAGGTGGTCTG    750
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
B-38773          701 TCTGCGAACCCGGATTCGAACGATGCCGCGCTGATGGAAAAGGTGGTCTG    750

B_28_T7F_F10     751 GGTCGGCGCCGGGGATGGCGAGCCGGCCTGGGCCGAGGGCGGCAGCTATC    800
                     ||||||||||||||||||||||||||.||||||||||||||||||||||||||
B-38773          751 GGTCGGCGCCGGGGATGGCGAGCCGGTCTGGGCCGAGGGCGGCAGCTATC    800

B_28_T7F_F10     801 AGGTGGTGCGGCTGATCCGGAACATGGTCGAACGCTGGGACCGCACGCCG    850
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
B-38773          801 AGGTGGTGCGGCTGATCCGGAACATGGTCGAACGCTGGGACCGCACGCCG    850

B_28_T7F_F10     851 CTGCAAGAGCAGGAGCGCGATTTCGGCCGGCGCAAGATGTCGGGCGCGCC    900
                     ||||||.|||||||||||||||||||||||||||||||||||||||||||||
B-38773          851 CTGCAGGAGCAGGAGCGCGATTTCGGCCGGCGCAAGATGTCGGGCGCGCC    900

B_28_T7F_F10     901 GATGGATGGCGGCCCCGGGGCGACCGAGCGCGACGTTCCCGATTATGCGC    950
                     |.||||.||||||.|.|..||||||||||||||||||||||||||||||||||
B-38773          901 GCTGGACGGCGGCGCGGATGCGACCGAGCGCGACGTTCCCGATTATGCGC    950

B_28_T7F_F10     951 GGGATCCGGAGGGCAAGGTCACGCATCTGCTGTCGCACATCCGGCTGGCC   1000
                     ||||.|||||||||||||.|||||||||||||||||||||||||||||||||
B-38773          951 GGGACCCGGAGGGCAAGGCCACGCATCTGCTGTCGCACATCCGGCTGGCC   1000

B_28_T7F_F10    1001 AATCCGCGCACGGCCGANACCCAGAANAACCTGATCCTGCGCCGCGCCTT   1050
                     ||||||||||||||||.||||||||.||||||||||||||||||||||||||
B-38773         1001 AATCCGCGCACGGCCGAGACCCAGAAGAACCTGATCCTGCGCCGCGCCTT   1050

B_28_T7F_F10    1051 CAACTACACCAACGGCGTGATGAANAACGGGCAGCTTGATCAGGGCCTGC   1100
                     |||||||||||||||||||||||||.||||||||||.||||||||||||||
```

```
B-38773         1051 CAACTACACCAACGGCGTGATGAAGAACGGGCAGCTGGATCAGGGCCTGC   1100

B_28_T7F_F10    1101 TGTTCATCTGCTATCAGGCCGATCTGGAGGCCGGGTTCATCACCGTGCAG   1150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
B-38773         1101 TGTTCATCTGCTATCAGGCCGATCTGGAGGCCGGGTTCATCACCGTGCAG   1150

B_28_T7F_F10    1151 AACCGGCTGAACGGNNNGTTGATGGAAAGANNATTTCAAACCGATCGGGG   1200
                     ||||||||||||||...|||||||| ||||..||||||||||||||||.|
B-38773         1151 AACCGGCTGAACGGCGAGTTGATGG-AAGAATATTTCAAACCGATCGGCG   1199

B_28_T7F_F10    1201 GGGGGAATTTNNNTTC---------------------------------   1216
                     |||||.||||  .|||
B-38773         1200 GGGGGTATTT--CTTCACCCTGCCGGGGGTGACGGGGCCGGGCGATTTCC   1247

B_28_T7F_F10    1217 ---------------------------------------   1216

B-38773         1248 TGGGCTCCGGCCTGATTTCCGCAGCCAGGGCATTGTAA   1285


#---------------------------------------
#---------------------------------------
```

Appendix 4: Denaturing SDS-PAGE gels (Resolving and Stacking).

Volumes of reagent components required for casting 4 denaturing SDS gels (1.5mm) of the indicated percentages.

Components	Volumes for Resolving gel (12%)	Volumes for Stacking gel (4%)
30% acrylamide	8 ml	1.32 ml
0.5M Tris-Hcl, pH 6.8	----	1.26 ml
1.5M Tris-Hcl, pH 8.8	5 ml	----
10% SDS	100 µl	50 µl
Distilled water	7ml	3 ml
TEMED	10 µl	5 µl
10% Ammonium per sulphate (APS)	100 µl	25 µl
Total volume	20 ml	5 ml

Appendix 5: Recipes for buffer preparations

Appendix 5.1: NAP buffer (1 Litre)

Weigh and dissolve the following reagents in 800ml distilled water in a volumetric flask.

Reagent	Amount required
EDTA	7.44g
Sodium citrate trisodium salt dihydrate	7.35g
Ammonium sulfate	700g

Use H_2SO_4 to adjust the pH of the solution to 5.2 and make up to 1L with distilled water.

Appendix 5.2: Cell lysis buffer (100ml)

Add the following reagents in a volumetric flask and add 80ml distilled water

Reagent	Amount needed
1M KH_2PO_4	0.3ml
1M K_2HPO_4	4.7ml
NaCl	2.3g
KCl	0.7g
Glycerol	10ml
Triton X-100	0.5ml
Imidazole	68mg

Adjust pH to 7.8 with HCl and make up to 100ml volume with distilled water.

Appendix 5.3: SDS-PAGE gel running buffer (10X) (1L)

Dissolve the following reagents in 1 litre of distilled water

Reagent	Amount
Tris base	30.3g
Glycine	144.1g
SDS	10g

Store at +4°C and dilute to (1X) before use.

Appendix 5.4: SDS-PAGE Sample Loading Buffer (2X) (30ml)

Prepare the following

Reagent	Amount needed (ml)
0.5M Tris-HCl (pH 6.8)	3.75
Glycerol (50%)	15.0
SDS (10%)	6.0
Bromophenol blue (1%)	0.3
Distilled water	0.25

Store at -20°C and add β-mercaptoethanol (1part to 19 parts of buffer) fresh before use.

Appedix 5.5: Western blot transfer buffer (500ml)

To prepare 500ml of a 10X western transfer buffer, add the following in a bottle.

Reagent	Amount needed
Tris-HCl	15.2g (25mM)
Glycine	72.1g (190mM)
Distilled water	To 500 ml

To dilute to a working concentration of 1X solution (500ml) from the 10X stock solution, the following were added.

Reagent	Amount needed
Ccccc/10X Western transfer buffer	50ml
Methanol (20%v/v)	100ml
Distilled water	350 ml