

**METAGENOME OF MATANG MANGROVE  
FOREST REVEALS TREE HARVESTING ALTER  
SOIL MICROBIOME**

by

**PRIYA A/P GUNASELVAM**

**Thesis submitted in fulfillment of the requirements  
for the degree of  
Master of Science**

**May 2017**

## **ACKNOWLEDGEMENT**

Firstly, I would like to express my gratitude to two postdocs in CCB who have been my mentors during the project, Dr. Lau Nyok Sean and Dr. Go Furusawa. I would also like to thank my main supervisor, Prof. Alexander Chong Shu-Chien who has been the backbone of my research here in CCB. I am indeed very grateful for their vision and foresight in which inspired and motivated me throughout this study.

I would like to take this opportunity to record my sincere thanks to my CCB lab mates, Dinesh Balachandra, Valerie Chiam Nyet Cheng, Muhd. Khairul Luqman, Sam Ka Kei and Jess Loh Swee Cheng for their valuable help and support. Besides that, I would like to thank the past and present administration staffs who have facilitated my work here in CCB. I also thank my family for their unceasing support and encouragement. Last but not least, I would like to thank the Malaysian Government for funding my study by providing the MyMaster (MyBrain15) scholarship.

## TABLE OF CONTENTS

<b>ACKNOWLEDGEMENT</b> .....	ii
<b>TABLE OF CONTENTS</b> .....	iii
<b>LIST OF TABLES</b> .....	vii
<b>LIST OF FIGURES</b> .....	viii
<b>LIST OF SYMBOLS AND ABBREVIATIONS</b> .....	xii
<b>ABSTRAK</b> .....	xiv
<b>ABSTRACT</b> .....	xvi
<b>CHAPTER 1 INTRODUCTION</b> .....	1
1.1 Background of Research .....	1
1.2 Research Objectives .....	4
<b>CHAPTER 2 LITERATURE REVIEW</b> .....	5
2.1 Matang Mangrove Forest Reserve .....	5
2.2 Microbial diversity .....	7
2.2.1 Microbial communities in the soil environment .....	8
2.3 The limitations of culture-dependent analysis .....	8
2.4 Metagenomics .....	9
2.4.1 Sequence-based metagenomics .....	10
2.4.1(a) Amplicon sequencing .....	11

2.4.1(b) Whole genome shotgun (WGS) sequencing.....	12
2.4.2 Function-based metagenomics.....	13
2.5 Sampling and processing .....	13
2.6 Sequencing technology .....	15
2.7 Softwares for analyzing metagenomic sequences.....	19
2.7.1 MG-RAST.....	19
2.7.2 MEGAN.....	20
2.7.3 IMG/M.....	21
<b>CHAPTER 3 MATERIALS AND METHODOLOGY.....</b>	<b>23</b>
3.1 Materials.....	23
3.2 Methodology .....	24
3.2.1 Sample collection.....	24
3.2.2 DNA extraction .....	25
3.2.3 Illumina library preparation, high-throughput sequencing and quality assessment .....	27
3.2.4 Reads processing.....	28
3.2.5 Assembly.....	29
3.2.6 Taxonomic and functional analyses .....	31
3.2.7 Statistical analysis .....	32
3.2.8 Phylogenetic tree.....	32

3.2.9 Carbohydrate degradation enzymes .....	33
3.2.10 Principal Component Analysis (PCA) .....	33
<b>CHAPTER 4 RESULTS</b> .....	<b>35</b>
4.1 Physicochemical analysis of soil samples .....	35
4.2 DNA extraction .....	36
4.3 Metagenomics sequencing and analysis .....	37
4.3.1 Analysis statistics of PZ and VJF metagenomes .....	37
4.3.2 Source hits distribution of PZ and VJF metagenomic reads .....	39
4.3.3 Rarefaction curves of PZ and VJF metagenomes. ....	41
4.3.4 Length distributions of metagenomic reads from PZ and VJF. ....	43
4.3.5 GC distribution of metagenomic reads from PZ and VJF. ....	44
4.4 Taxonomic composition of mangrove metagenomes .....	45
4.4.1 Taxonomic profiles of PZ and VJF metagenomes at domain level .....	45
4.4.2 Taxonomic profiles of PZ and VJF metagenomes at phylum level .....	47
4.4.3 Taxonomic profiles of PZ and VJF metagenomes at class level .....	51
4.4.4 Taxonomic profiles of PZ and VJF metagenomes at order level .....	58
4.4.5 Taxonomic profiles of PZ and VJF metagenomes at family level .....	61
4.4.6 Taxonomic profiles of PZ and VJF metagenomes at genus level .....	63
4.4.7 Principal component analysis .....	65
4.5 Functional analyses of the metagenomes .....	68

4.5.1 Functional annotations using COG, NOG, KO and SEED Subsystems databases .....	68
4.5.2 Comparison of PZ and VJF metabolic profiles at hierarchical level .....	72
4.6 Mining for genes involved in carbohydrate degradation .....	77
4.7 Carbohydrate degradation genes analyses .....	79
4.8 Biogeochemical cycle analyses .....	89
<b>CHAPTER 5 DISCUSSION</b> .....	91
5.1 Taxonomic affiliations of the PZ and VJF metagenomic reads .....	91
5.2 Functional capabilities of the PZ and VJF metagenomic reads .....	96
<b>CHAPTER 6 CONCLUSION</b> .....	100
<b>REFERENCES</b> .....	101

## LIST OF TABLES

	<b>Page</b>
Table 4.1 Physicochemical characteristics of the PZ and VJF soil samples, revealing significant difference in terms of total organic carbon content.	35
Table 4.2 Metagenomes used for comparison and their characteristics (taxonomy assignment was performed using MG-RAST).	67

## LIST OF FIGURES

		<b>Page</b>
Figure 2.1	Map of peninsular Malaysia focusing on Matang	5
Figure 3.1	(A) Geographic map of Kuala Sepetang region, (B) photograph of PZ and (C) photograph of VJF	24
Figure 3.2	Workflow of genomic DNA extraction using the PowerSoil® DNA Isolation Kit	25
Figure 4.1	M, 1kb DNA ladder; Lane 1, extracted soil DNA of PZ; Lane 2, extracted soil DNA of VJF	36
Figure 4.2	Analysis statistics of (A) PZ and (B) VJF metagenomes, indicating the number of metagenomic reads that passed QC, predicted proteins, annotated sequences and number of reads assigned to functional categories	37
Figure 4.3	Graph displaying number of features in the (A) PZ and (B) VJF metagenomes annotated by several databases. The bars signifying annotated reads are coloured by <i>e</i> -value range	39
Figure 4.4	Rarefaction curves showing species richness identified in the (A) PZ and (B) VJF metagenomes. The number of reads sampled is indicated on the x-axis	42
Figure 4.5	Length distributions of reads post QC for (A) PZ and (B) VJF metagenomes	43
Figure 4.6	GC distribution of reads post QC for (A) PZ and (B) VJF	44



	metagenomes	
Figure 4.7	Taxonomic assignment of reads at domain level for (A) PZ and (B) VJF metagenomes, showing majority of sequences belonging to the domain Bacteria	46
Figure 4.8	Taxonomic assignment of reads at phylum level for (A) PZ and (B) VJF metagenomes	48
Figure 4.9	Comparative taxonomic profile of PZ (blue) and VJF (orange) metagenomes analysed using STAMP at the phylum level ( $p>0.05$ )	50
Figure 4.10	Taxonomic assignment of reads at class level for (A) PZ and (B) VJF metagenomes	52
Figure 4.11	Differential proportion of sequences assigned within the phylum: (A) and (B) <i>Proteobacteria</i> ; (C) and (D) <i>Firmicutes</i> ; (E) and (F) <i>Bacteroidetes</i>	55
Figure 4.12	Comparative taxonomic profile of PZ (blue) and VJF (orange) metagenomes using STAMP at the class level ( $p>0.05$ )	57
Figure 4.13	Comparative taxonomic profile of PZ (blue) and VJF (orange) metagenomes using STAMP at the order level ( $p>0.05$ )	59
Figure 4.14	Relative abundance of orders within the most dominant phylum, <i>Proteobacteria</i> , represented as percentages on the y-axis	60
Figure 4.15	Comparative taxonomic profile of PZ (blue) and VJF (orange) metagenomes using STAMP at the family level ( $p>0.05$ )	62

Figure 4.16	Comparative taxonomic profile of PZ (blue) and VJF (orange) metagenomes using STAMP at the genus level ( $p>0.05$ )	64
Figure 4.17	Principal component analysis of metagenomes based on taxonomic affiliation of reads determined using MG-RAST	66
Figure 4.18	Functional category hits distribution according to (A) COG, (B) NOG, (C) KO and (D) Subsystems of PZ and VJF metagenomes	69
Figure 4.19	Comparative functional profile of the PZ (blue) and VJF (orange) samples analysed using STAMP for SEED subsystem at level 1. (*) $p$ -value $\leq 0.05$	73
Figure 4.20	Comparative functional profile of the PZ (blue) and VJF (orange) samples analysed using STAMP for SEED subsystem at level 2 (carbohydrate). (*) $p$ -value $\leq 0.05$	74
Figure 4.21	Comparative functional profile of the PZ (blue) and VJF (orange) samples analysed using STAMP for SEED subsystem at level 3 (monosaccharides). (*) $p$ -value $\leq 0.05$	76
Figure 4.22	Barchart of carbohydrate degradation enzymes based on unique hits from SEED database	77
Figure 4.23	Phylogenetic analysis of bacterial sequences assigned as endo-1,4- $\beta$ -xylanase from PZ (red) and VJF (green) metagenomes	80
Figure 4.24	Phylogenetic analysis of bacterial sequences assigned as xylan 1,4- $\beta$ -xylosidase from PZ (red) and VJF (green) metagenomes	82
Figure 4.25	Phylogenetic analysis of bacterial sequences assigned as $\alpha$ -L-	84

	arabinofuranosidase from PZ (red) and VJF (green) metagenomes	
Figure 4.26	Phylogenetic analysis of bacterial sequences assigned as $\beta$ - glucosidase from PZ (red) and VJF (green) metagenomes.	86
Figure 4.27	Phylogenetic analysis of bacterial sequences assigned as pectate lyase from PZ (red) and VJF (green) metagenomes	88
Figure 4.28	KEGG pathway for (A) starch and sucrose metabolism and (B) sulphur metabolism. The colour scale represents the number of genes (in logarithmic scale) found for each KEGG entry; PZ (blue) and VJF (red)	90

## LIST OF SYMBOLS AND ABBREVIATIONS

%	Percentage
μL	Microlitre
ADR	Artificial duplicate reads
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like alignment tool
bp	Base pair
CAT	CAZymes Analysis Toolkit
CAZy	Carbohydrate-active enzymes
cm	Centimetre
COG	Clusters of orthologous groups
DNA	Deoxyribonucleic acid
g	Gram
Gb	Gigabyte
GH	Glycoside hydrolase
IMG/M	The integrated microbial genomes and metagenomes
km	Kilometre
KO	KEGG orthology
M5nr	M5 non-redundant protein database
MEGAN	Metagenome Analyzer
MG-RAST	Metagenomic rapid annotation using subsystem technology

mL	Millilitre
MMFR	Matang Mangrove Forest Reserve
ng	Nanogram
NGS	Next generation sequencing
NOG	Non-supervised orthologous groups
°C	Degree celcius
ORF	Open reading frame
PCA	Principal component analysis
pcr	Polymerase chain reaction
PZ	Productive Zone
QIIME	Quantitative Insights Into Microbial Ecology
rpm	Revolutions per minute
SDS	Sodium dodecyl sulfate
spf	STAMP profile file
STAMP	Statistical Analysis of Metagenomic Profiles
TAE	Tris-acetate-EDTA
Tb	Terabyte
tsv	Tab-separated value
qc	Quality control
VJF	Virgin Jungle Forest

# METAGENOM HUTAN PAYA BAKAU MATANG MENUNJUKKAN AKTIVITI PENEANGAN POKOK MENGUBAH MIKROB TANAH

## ABSTRAK

Hutan Simpan Paya Bakau Matang di Malaysia telah diiktiraf sebagai hutan bakau yang diurus terbaik di seluruh dunia. Pengetahuan sangat terhad mengenai kesan penggunaan tanah pada komuniti mikrobial tanah dan keupayaan fungsinya. Kajian ini telah menganalisis pemprosesan tinggi dataset metagenomik dari dua tapak persampelan di Hutan Simpan Paya Bakau Matang yang berbeza: Zon Produktif dan Hutan Dara. Zon Produktif ditetapkan untuk peneangan pokok manakala Hutan Dara adalah kawasan yang masih belum diterokai. Daun sampah daripada sisa selepas tuai boleh menyumbang untuk membina bahan organik di dalam tanah dan ia menjelaskan jumlah kandungan karbon 19 kali lebih tinggi di Zon Produktif berbanding dengan Hutan Dara. Penjajaran menggunakan Illumina HiSeq 2500 platform dan 150 bp kimia berpasangan akhir menghasilkan 30.8 dan 30.1 Gb untuk Zon Produktif dan Hutan Dara, masing-masing. Jujukan metagenomik dianalisis menggunakan MG-RAST dan digambarkan secara statistik menggunakan perisian STAMP. Klasifikasi taxonomi metagenomik menggunakan MG- RAST menunjukkan perbezaan dalam filum dominan bakteria yang terdapat dalam kedua-dua sampel. Filum penting yang dipamerkan oleh metagenome Zon Produktif ialah *Firmicutes*, *Bacteroidetes* dan *Chloroflexi*. Sebaliknya, *Proteobacteria*, *Actinobacteria* dan *Planctomycetes* adalah filum yang lazim dalam Hutan Dara. Kelas *Dehalococcoidetes*, *Clostridia*, *Flavobacteriia*, *Bacteroidia* dan

*Deltaproteobacteria* adalah dominan dalam sampel Zon Produktif manakala *Alphaproteobacteria*, *Gammaproteobacteria*, *Planctomycetia*, *Actinobacteria* dan *Betaproteobacteria* dominan dalam sampel Hutan Dara. Jujukan yang berkaitan dengan metabolisme karbohidrat, terutamanya enzim untuk degradasi dan penggunaan polisakarida pada dinding sel tumbuhan, adalah utama di Zon Produktif dan pemerhatian ini mungkin berkait rapat dengan kandungan karbon yang tinggi dalam tanah. Analisis fungsian memberi tumpuan pada enzim karbohidrat mendedahkan pelbagai enzim yang terlibat dalam hemiselulosa, selulosa dan pektin. Keputusan yang dibentangkan dalam kajian ini memberi gambaran tentang kepelbagaian mikrob dan potensi metabolik tanah bakau Malaysia dan mempunyai informasi penting dalam memahami kemungkinan kesan penuaian pokok pada masyarakat mikrob tanah.

# METAGENOME OF MATANG MANGROVE FOREST REVEALS TREE HARVESTING ALTER SOIL MICROBIOME

## ABSTRACT

Matang Mangrove Forest Reserve (MMFR) in Malaysia has been recognized as the best-managed mangrove forest in the world. There is limited knowledge about the effects of land use changes on soil microbial diversity and its functional capability. This study analysed high-throughput metagenomic datasets from two sampling sites at MMFR with distinct features: the Productive Zone and Virgin Jungle Forest. The former zone is designated for tree harvesting while the latter is a pristine area. Leaf litter from post-harvest residue could contribute to build up of organic matter in soil and it explained the 19 times higher total carbon content in the Productive Zone compared to the Virgin Jungle Forest. Sequencing using Illumina HiSeq 2500 platform and 150 bp paired-end chemistry resulted in 30.8 and 30.1 Gb bases for the Productive Zone and Virgin Jungle Forest samples, respectively. The metagenomic sequences were analyzed using MG-RAST and visualized statistically using STAMP software. Taxonomic classification of metagenomic reads using MG-RAST revealed differences in the dominant phylum of bacteria present in both samples. The salient phyla exhibited by the Productive Zone metagenome are *Firmicutes*, *Bacteroidetes* and *Chloroflexi*. In contrast, *Proteobacteria*, *Actinobacteria* and *Planctomycetes* were the prevalent phylum in the Virgin Jungle Forest. The classes *Dehalococcoidetes*, *Clostridia*, *Flavobacteriia*, *Bacteroidia* and *Deltaproteobacteria* were dominant in the Productive Zone sample



whereas the Virgin Jungle Forest sample had significantly more reads in *Alphaproteobacteria*, *Gammaproteobacteria*, *Planctomycetia*, *Actinobacteria* and *Betaproteobacteria*. Sequences related to carbohydrate metabolism, especially enzymes for degradation and utilization of polysaccharides from plant cell wall, were predominant in the Productive Zone and this observation possibly correlated with the high carbon content of the soil. Functional analysis focusing on carbohydrate degrading enzymes disclosed an array of enzymes involved in hemicellulose, cellulose and pectin utilization enzymes. The results presented in this study render insights into the microbial diversity and metabolic potential of Malaysian mangrove soil and have important entailments in understanding the possible effects of tree harvesting on soil microbial communities.

## CHAPTER 1 INTRODUCTION

### 1.1 Background of Research

Mangroves are the one and only ligneous salt-tolerant plants present at the coastal littoral zone (Alongi 2002). Almost 60-75% of the Earth's coastline in the tropical and subtropical regions are covered by tidal forests (Holguin, Vazquez, and Bashan 2001). The extensive biomass of mangrove forests located around the equator rivals that of the tropical forests (Alongi 2002). Mangrove forests are among the most productive ecosystem that significantly contribute to the carbon cycle, maintenance of coastlines, preserving water quality and supporting the fisheries industry (Kathiresan and Bingham 2001). These ecosystems consist of estuarine environment with brackish water that is capable of supporting niche populations of organisms (Thompson et al. 2013). Mangrove ecosystems revolve around factors that are constantly fluctuating such as salinity, water levels, temperature and nutrients (Gonzalez-Acosta et al. 2006, Gomes et al. 2008). Salinity and intertidal variation especially, creates redox potential allowing microbes to be ubiquitous in mangrove environments (Clark et al. 1998, Holguin, Vazquez, and Bashan 2001, Ferreira et al. 2010).

Mangrove sediments represent one of the most dynamic and diverse microbial habitat in which one gram of soil may harbour thousands of different bacterial species (Raynaud and Nunan 2014). Bacteria and fungi constitute most of the total biomass in tropical mangrove regions whereas algae and protozoa only represent a minor fraction (Alongi 1988, 1994, Bano et al. 1997). Soil microorganisms are the important

determinants of the functioning of terrestrial ecosystems, particularly in nutrient cycling (Zeller et al. 2008), decomposition of organic matter and polluting compounds (Herman et al. 2012), as well as soil development (Cotrufo et al. 2013). The highly productive nature of mangroves are mainly due to an efficient nutrient recycling systems capable of regenerating new nutrients decomposing mangrove leaves (Alongi, Christoffersen, and Tirendi 1993). Microbial communities in mangrove soil rely strongly on the biogeographical, ecological and anthropogenic properties presented by the ecosystem. Land use changes including selective logging and forest conversion for agriculture drive changes to the soil chemistry via alterations in pH, carbon content, carbon-to-nitrogen ratio, and the availability of phosphorus and calcium (McGrath et al. 2001, Murty et al. 2002). These physicochemical changes to the soil, in turn, influence the composition and diversity of soil microbiome. Nevertheless, the effects of land use changes on soil bacteria in terms of phylogeny and functionality are poorly understood.

Metagenome is the sequence-based study of genetic material recovered directly from environmental samples. The analysis of these sequence data provides a broad view on the diversity and functionality of the whole microbiota present in that environment. Metagenome of soil microorganisms from Brazilian mangroves has been studied in detail using culture-independent approaches (Andreote et al. 2012, Dias et al. 2010, Dias et al. 2011, Thompson et al. 2013). Despite the fact that these studies connote an important contribution to our knowledge of microbial life in mangrove ecosystems, further studies are needed for tapping into the microbial diversity from different zones in mangroves along with those subjected to distinct anthropogenic impacts (Strangmann, Bashan, and Giani 2008).

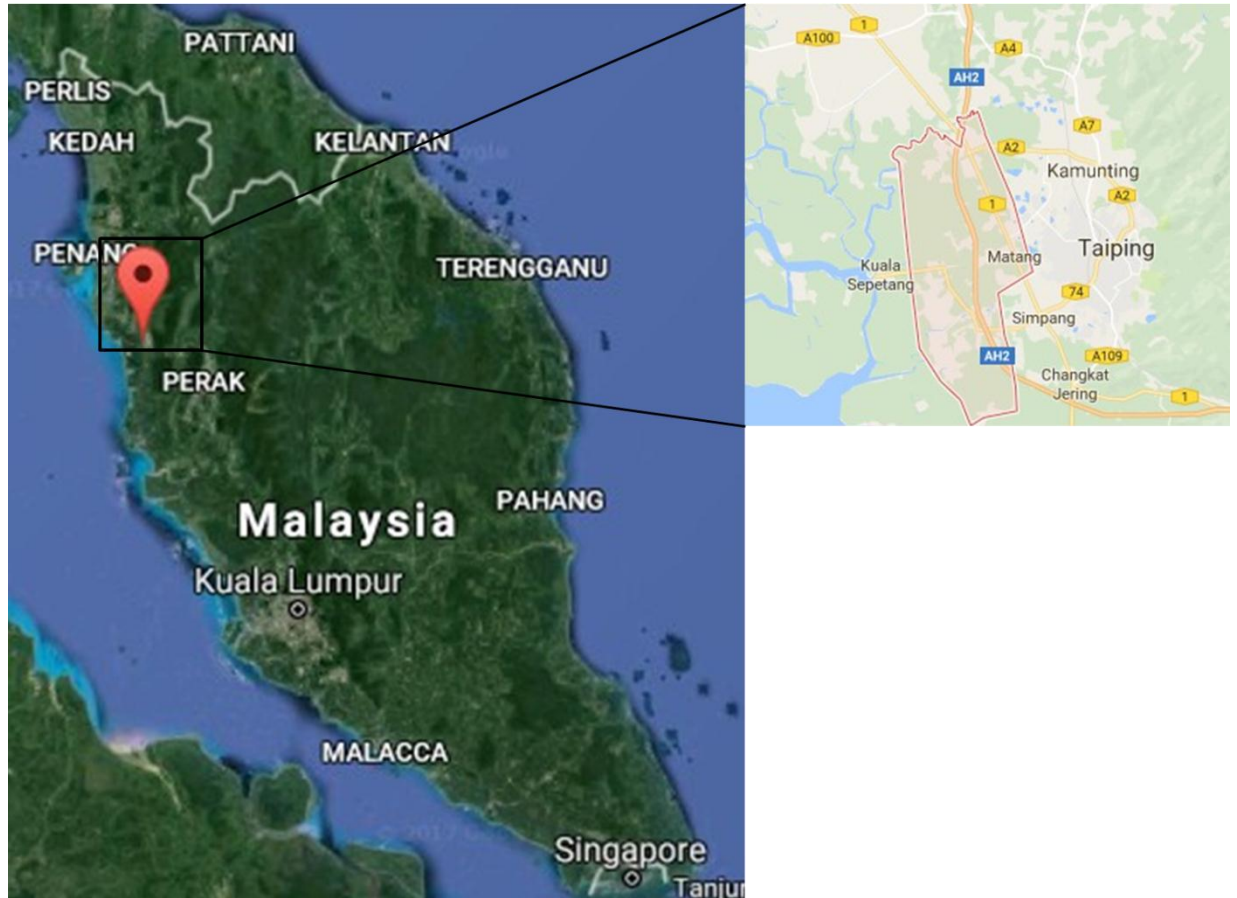
The restoration of mangroves has received a lot of attention worldwide and mangrove management has long been practised in South East Asia. Among these efforts, Matang mangrove forest (Malaysia) is considered the best managed mangrove in the world (Walters et al. 2008). Here, we performed an in-depth analysis of the microbial community found in two different sampling sites with distinct features in Matang mangrove forest. The first site being Productive Zone which is exposed to tree harvesting and human activities whereas the second site is the Virgin Jungle Forest which has been conserved and protected through systematic management for over a century. Tree harvesting at the Productive Zone involves opening of the canopy, removing certain tree species and resulting in post-harvest residues that affect the physicochemical properties of the soil. In this study, we present a robust description of microbes found in two sampling sites with distinct soil properties and the functional responses of microbiome to changes in soil chemistry caused by different land use. Moreover, the abundance of mangrove leaf litter in which the average total annual leaf litterfall ranged between 130 to 1870 g m<sup>-2</sup> (Kathiresan and Bingham 2001) has been shown to provide a favourable environment for the enrichment of microorganisms with polysaccharide-degrading capabilities (Rosado and Govind 2003). This encouraged the functional gene assignment analysis in terms of carbohydrate degradation enzymes of Matang mangrove forest sediments.

## **1.2 Research Objectives**

- I. To assess the taxonomical distribution of the microbial community in the Productive Zone and Virgin Jungle Forest.
- II. To explore the functional capability in terms of carbohydrate degradation enzymes, of the Productive Zone and Virgin Jungle Forest metagenomes.

## CHAPTER 2 LITERATURE REVIEW

### 2.1 Matang Mangrove Forest Reserve



**Figure 2.1:** Map of peninsular Malaysia focusing on Matang.

The Matang Mangrove Forest Reserve (MMFR), situated on the northwest coast of Peninsular Malaysia, is hailed as the best-managed mangrove forest in the world since early 20<sup>th</sup> century (Walters et al. 2008). The Mangrove forests are a stretch from Kuala Gula in the north to Panchor in the south, of 51.5 km in distance and 13 km in width. Matang mangroves span 40,288 hectares. The main genera of mangroves in MMFR revolve around the genus *Rhizophora*, *Bruguiera*, *Sonneratia*, *Avicennia* and *Ceriops*

(Goessens et al. 2014). The significance of mangrove forests is affirmed more than their status as home to many endangered flora and fauna species (Jusoff 2009). Being a precious natural resource with unique diversity, mangrove forests are blessed with high intrinsic natural productivity and unique habitat value (Ramu and Takeda 2003). Mangrove forests offer invaluable goods and services both economically and environmentally (Jusoff 2009). Other than the production of poles, charcoal and fuel wood, the mangrove ecosystem supports many functions such as the assimilation of waste, source of food, shelter and sanctuary for fauna, to name but a few (Jusoff 2013). The MMFR management system, regarded as the best in the world, utilize a highly regulated 30-year old rotation cycle in specific compartments (Goessens et al. 2014, Walters et al. 2008). Briefly, cleared areas are left for natural recruitments at the beginning of the cycle (Muda and Mustafa 2003). Assessments of survival and growth of seedlings are done annually, with planting of *Rhizophora apiculata* seedlings carried out, where needed. Two artificial thinning activities are done every 15 and 20 years to harvest poles for construction and also promote better growth for remaining trees. At completion of the cycle, mature trees are harvested for charcoal production, followed by replanting (Goessens et al. 2014). In MMFR, around 73.6 % of the total forest has been classified as productive forest whereas the remaining 26.4 % are non-productive or protected forest (Ibharim et al. 2015). The former is assigned as they are conducive for timber production and the non-productive forests for bio-diversity conservation, erosion mitigation, research and education, recreation, local community's needs and settlement (Ibharim et al. 2015).

## 2.2 Microbial diversity

The number of species present on Earth is astronomical and is composed of three domains of life namely *Bacteria*, *Archaea* and *Eukarya* (Woese, Kandler, and Wheelis 1990). Microorganisms occupy nearly all the habitats, ranging from deep ocean sediments (Li, Kato, and Horikoshi 1999), glaciers (Zhang et al. 2002), volcanic vents (Huber, Butterfield, and Baross 2003) to human gut (Suau et al. 1999). They are found in abundance essentially everywhere. The number of *Bacteria* and *Archaea* accounts for approximately  $4-6 \times 10^{30}$  cells and this figure is at least two to three times more than all of the eukaryotic cells on the biosphere (Whitman, Coleman, and Wiebe 1998). Microbes are the key players in biogeochemical cycles, converting carbon, oxygen, nitrogen and sulfur to forms accessible to all other living beings (Handelsman et al. 2007).

Not only microbes play a vital role in our ecological system, they are also used in various ways to enhance the condition of humankind. These tiny living beings are being used in the pharmaceutical industry for the production of drugs and antibiotics (Penesyan et al. 2009, Sykes et al. 1981). They are also involved in the biofuel production (Biello 2010), soil and water bioremediation (Christofi and Ivshina 2002, Groudeva, Groudev, and Doycheva 2001), fermentation of human foods (Fernandes, Shahani, and Amer 1987) as well as improving crop productivity (Kloepper, Lifshitz, and Zablotowicz 1989).



### **2.2.1 Microbial communities in the soil environment**

Soil is the Earth's most essential asset in managing all life in the physical environment (Violante, Huang, and Gadd 2007). There are approximately 600 million bacteria in one gram of high quality soil and the amount of distinct species ranges from 15,000 to 20,000 (de Souza Silva and Fay 2012). Among these, only a fraction has been known to be successfully grown under lab condition (Kakirde, Parsley, and Liles 2010). Biotic factors, and abiotic factors such as water, sunlight, oxygen, soil and temperature are known to have an effect on the bacterial communities' structure and diversity (Buckley and Schmidt 2002). Choosing a sampling site is very important in a metagenomic study as geographic location plays an important part in determining the phylogenetic makeup and growth of microorganisms (Kakirde, Parsley, and Liles 2010).

### **2.3 The limitations of culture-dependent analysis**

The most essential part in microbiology over the past 100 years has been to attain pure cultures for laboratory research. Normally, growth media like nutrient agar, Tryptic Soy agar and Luria-Bertani medium are used to isolate pure colonies of microorganisms and identify them (Kirk et al. 2004). Having said that, this technique is inadequate while dealing with the ecological community as most of the microorganism present cannot be grown in the laboratory. Only 1% or less of the microbes in the biosphere is known to be identified using culturing methods. This number was predicted through the “great plate-count anomaly” whereby the plate counts of microorganism were lesser than those directly observed with a microscope (Staley and Konopka 1985). To get the most out of the bacterial communities that can be cultivated, a few enhanced cultivation

methodology and growth media have been developed that imitates the environment's nutrients, oxygen gradient and pH (Rastogi and Sani 2011). Then again, complex communities of microbes or those that are strictly anaerobic may prove to be extremely difficult to be cultivated (Müller et al. 2012). The identification of new species using culturing methods is inevitable, but it will be intensely one-sided, delivering only a limited snapshot of a handful of amenable constituents and it cannot help to improve the understanding of the entire community or obtaining any sort of complete survey (Houghton 2013).

## **2.4 Metagenomics**

Metagenomics is defined as the direct genetic analysis of genomes contained within environmental samples, bypassing the need to isolate and culture individual microbial species (Thomas, Gilbert, and Meyer 2012). In its methodology and strategy, metagenomics rise above individual genes and genomes, empowering researchers to concentrate the majority of the genomes in a group all in all (Handelsman et al. 2007). The term metagenomics was first used in 1998 by Handelsmann et al. (Handelsman et al. 1998) with regards to soil as a home for microbes and was termed as “the collective genomes of soil microflora”. This method has also been coined as: environmental DNA cloning (Stein et al. 1996), multigenomic cloning (Cowan 2000), environmental genomics (Béja et al. 2000), eDNA cloning (Brady and Clardy 2000), recombinant environmental cloning (Courtois et al. 2003) and community genome analyses (Tyson et al. 2004). Metagenomics proves to be a good methodology for a quick analysis of the whole bacterial diversity in complex microbial communities, allowing for the

simultaneous assessment of community composition and structure, as well as phylotype richness to take place. Metagenomics is characterized into two different approaches, namely the sequence-based metagenomics and the function-based metagenomics (Handelsman et al. 2007).

#### **2.4.1 Sequence-based metagenomics**

In sequence-based metagenomics, DNA from environmental samples are extracted, sequenced randomly and analysed (Thomas, Gilbert, and Meyer 2012). Generally, the separation of bacteria, eukaryotic cells, viruses and free DNA by size are performed using centrifugation or filtration (Handelsman et al. 2007), and the extraction of total DNA from the appropriate fraction is done. The sequenced DNA sample is thought to be an arbitrary portion of the entire population (Fierer, Barberán, and Laughlin 2014). Then, the metagenomic sequences are compared to known sequences that are collected in international archives (Schmieder and Edwards 2012).

To assemble genomes, identify genes (Culligan et al. 2014) and find complete metabolic pathways, the sequence-based metagenomics studies may prove to be timely (Handelsman et al. 2007), and it can be used too, to analyze the genome of the community as a whole, which can give us some ideas about both the population ecology and evolution (Thomas, Gilbert, and Meyer 2012). The analysis of microbial diversity is not very expensive and less computer intensive than assembling genomes, and valuable information about the ecology of microbes in a sample can be provided (Sharpton 2014). There is a great deal of computer power required in genome assembly but it can facilitate the understanding of how certain genes contribute towards helping organisms

to survive in a certain environment. To this end, more than one thousand different metagenomes have been sequenced from such a large array of environments (Schmieder and Edwards 2012). Furthermore, extinct species such as the woolly mammoth (Poinar et al. 2006) and the Neanderthals (Noonan et al. 2006) have been examined using the sequence-based metagenomic approaches.

#### **2.4.1(a) Amplicon sequencing**

In this approach, a sample of the community as well as the DNA are extracted from all the sample cells. A taxonomically informative genomic marker that is no stranger to all organisms of interest is then targeted and amplified by PCR. The resultant amplicons are sequenced and bioinformatically characterized in order to see which microbes exist in the sample and at what relative abundance is their existence (Sharpton 2014). The amplicon sequencing of the 16S locus exposed a tremendous amount of microbial diversity on Earth (Pace 1997, Rappe and Giovannoni 2003, Lozupone and Knight 2007) and it serves to characterize the microbial biodiversity from various environments including the human gut (Yatsunenko et al. 2012), *Arabidopsis thaliana* roots (Lundberg et al. 2012), ocean thermal vents (McCliment et al. 2006), hot springs (De León et al. 2013), and Antarctic volcano mineral soils (Soo et al. 2009).

Despite it being powerful, the amplicon sequencing has its own shortcomings. Other than the fact that it may fail to resolve a substantial fraction of the diversity in a community given various biases that have to do with the PCR (Hong et al. 2009, Sharpton et al. 2011, Logares et al. 2014), the amplicon sequencing can generate widely varying diversity estimates (Jumpstart Consortium Human Microbiome Project Data

Generation Working 2012). Also, another flaw is that the amplicon sequencing only typically provides insight into the taxonomic aspect of the microbial community (Langille et al. 2013). Finally, amplicon sequencing is restricted to the taxa analysis for where the taxonomically informative genetic markers are known and amplifiable (Sharpton 2014).

#### **2.4.1(b) Whole genome shotgun (WGS) sequencing**

Shotgun metagenomic sequencing is another approach to the study of uncultured microbiota that is able to steer clear from these limitations. Here, again, DNA is extracted from all cells in a community. However, rather than targeting a specific genomic locus for amplification, all DNA is subsequently fragmented, where they are independently sequenced. This results in DNA sequences that align to multifarious genomic locations for the myriad genomes that can also be found in the sample. Some of these reads will be sampled from taxonomically informative genomic loci, and others will be sampled from coding sequences which offer an insight into the biological functions that have been encoded in the genome. Consequently, metagenomic data opens up the opportunity to explore two aspects of a microbial community simultaneously: the questions of “who is there?” and “what are they capable of doing?” are raised (Sharpton 2014).

### **2.4.2 Function-based metagenomics**

Function-based metagenomics allows scientists to straightforwardly extract and identify novel proteins and metabolites from a microbial community (Craig et al. 2010). In function-based metagenomics, researchers go through the metagenomic libraries for a number of functions, such as biocatalysts (Lorenz et al. 2002), industrial enzymes (Suenaga, Ohnuki, and Miyazaki 2007) and antibiotic production (D’Costa, Griffiths, and Wright 2007). Functions alien to the controlled number of microbes able to be grown in a laboratory can be identified by scientists using this approach (Jurkowski and Reid 2007). Millions of random DNA fragments in a library are translated into proteins by bacteria that proliferate in the laboratory. Clones capable of producing “foreign” proteins are then screened to see their many abilities. Researchers will be able to access the tremendous genetic diversity in a microbial community without much, or any, knowledge about the fundamental gene sequence, the structure of the desired protein, or the microbe of origin (Handelsman et al. 2007).

### **2.5 Sampling and processing**

Sample processing stands out to be the most prominent and functional step in any metagenomics project. The DNA extracted should depict all cells in the sample and sufficient amounts of high-quality nucleic acids must be obtained for the library production and sequencing that follow (Thomas, Gilbert, and Meyer 2012). Processing needs particular protocols for each sample type, and various robust methods for DNA extraction are ready to be adopted (Venter et al. 2004, Burke, Kjelleberg, and Thomas 2009, Delmont et al. 2011). There are also initiatives made to penetrate into the aspect of

the microbial biodiversity from tens of thousands of ecosystems using a single DNA extraction technology to make sure that they are comparable (Knight et al. 2012).

If the target community is linked with a host such as a plant or invertebrate, then either fractionation or selective lysis might be regarded as appropriate to ensure that minimal host DNA can be obtained (Burke, Kjelleberg, and Thomas 2009, Thomas et al. 2010). Several selective filtration or centrifugation steps, or even flow cytometry, can serve to enrich the target fraction (Venter et al. 2004, Palenik et al. 2009, Angly et al. 2006). Physical separation and isolation of cells from the samples might also be vital as they can maximize the DNA yield or avoid co-extraction of enzymatic inhibitors that might hinder the next process, which is the processing (Thomas, Gilbert, and Meyer 2012). In terms of microbial diversity and the DNA yield, the direct lysis of cells in the soil matrix versus indirect has a quantifiable bias, and resulting sequence fragment length (Delmont et al. 2011).

Certain types of samples such as biopsies often produce only very small amounts of DNA (Abbai et al. 2012). Library production for most sequencing technologies necessitates nanograms or micrograms amounts of DNA, and hence there may be the requirement for the amplification of the starting material. Similar to any amplification method, there are potential problems that have to do with reagent contaminations, chimera formation and sequence bias in the amplification, and their impact will rely on the starting material's amount and type and the required number of amplification rounds to yield enough nucleic acids (Thomas, Gilbert, and Meyer 2012). The impact of these issues on the following metagenomic community analysis (Abbai et al. 2012) can be

highly significant, therefore it is crucial to deliberate upon whether amplification would be allowed (Thomas, Gilbert, and Meyer 2012).

## **2.6 Sequencing technology**

The emergence of the next-generation sequencing (NGS) or high-throughput sequencing has paved the way for the field of microbial ecology and contributed to the establishment of the field metagenomics (Oulas et al. 2015). The first NGS technology, which could be materialized due to incredible amalgam of nanotechnology, organic chemistry, optical engineering, enzyme engineering, and robotics, became a viable commercial offering in 2005 (Kumar et al. 2015). The NGS platforms are functional when it comes to standard sequencing applications and also for novel applications previously untouched by Sanger sequencing (Morozova and Marra 2008). Prior to the arrival of NGS platforms, Venter et al. (Venter et al. 2004) in 2004 produced high magnitude metagenomics sequence data to the tune of 1.66 million reads, with 1.045 billion base pairs with an average read length of 818 bp from metagenomic samples gathered from Sargasso Sea. As a follow-up, during Sorcerer II Global Ocean Sampling expedition, Rusch et al. (Rusch et al. 2007) generated 7.7 billion sequencing reads, carrying 6.3 billion base pairs using Sanger sequencing. This large amount of sequence data using Sanger sequencing is not insignificant, but the magnitude of data which are produced in a single run of NGS machine marks to be several fold higher (Kumar et al. 2015). The large scale sequencing projects and consortia have already produced NGS derived huge sequence data sets, namely, The ENCODE project (Rosenbloom et al. 2011), 1000 Genomes (Lappalainen et al. 2013), Human Microbiome Project



(Turnbaugh et al. 2007), and Earth Microbiome Project (Gilbert et al. 2010). The NGS platforms have paved the way to sequence the metagenomic DNA directly, altogether circumventing the need for the difficult steps of cloning and library-preparing, and also allow massive parallel sequencing where hundreds of thousands to hundreds of millions of sequencing reactions are performed and detected at the same time, further causing very high throughput. As it is very crucial to make the decision about the suitability of a particular type of NGS platform for a metagenomic project, the selection of a particular NGS platform has to be made with consideration given to the varying features of NGS platforms like read length, degree of automation, and a few others (Kumar et al. 2015).

Going back to the year 2005, Roche introduced the 454 pyrosequencer, which could easily produce more data than 50 capillary sequencers at about one sixth of the cost. This was followed by the release of the Solexa Genome Analyzer by Illumina in 2006, which used sequencing by synthesis to produce tens of millions of 32 bp reads, and of the SOLiD platform by Applied Biosystems in 2007 (Rodríguez-Ezpeleta, Hackenberg, and Aransay 2011).

The longer read length resulting from the Roche chemistry allows genome assembly to be more feasible, giving Roche 454 platform an upper edge over other competitors (Kumar et al. 2015). The Illumina's offerings, HiSeq 1500/2500, HiSeq 2000/1000, and Genome Analyzer IIX are the prevalent NGS platforms useful for metagenomic research (Chan et al. 2015, Hasan et al. 2014). One of the most recent additions of Illumina, that is, HiSeq 1500/2500, offers two run modes (rapid run and high output run mode) (Valencia et al. 2013). The high output run mode necessitates only 1 ng of community DNA to get complete metagenomic sequence data using

reversible terminator chemistry of Illumina for their HiSeq 2500 able to generate 270–300 Gb of sequence data with read length of up to 200 bp and very high coverage of not more than 5 days, and hence is seen to adapt well to metagenomic investigations (Kumar et al. 2015). Illumina recently released the HiSeq X Ten, a set of ten HiSeq X sequencing machines, with the staggering capacity to generate up to 1.8 Tb of sequence per run (van Dijk et al. 2014).

After Roche 454 and Illumina’s NGS platforms, the polony sequencing based ABI (now Life Technologies) SOLiD platforms are also applied in metagenomic literature (Oulas et al. 2015). These NGS platforms can cater for deep sequencing which makes it possible to locate very low abundant members of complex populations in metagenomic samples. The actual read length and depth required will be based on the required sensitivity and complexity of the population. NGS technologies have paved the way for shotgun metagenomics to reconstruct the entire bacterial and archaeal genomes without a reference genome (or their genome sequence) being present, by making use of robust assembly algorithms that join short overlapping DNA fragments produced by the NGS sequencers (Kumar et al. 2015). Luo et al. (Luo et al. 2012) drawn a direct comparison of the two most popular NGS platforms, that is, Roche 454 FLX Titanium and Illumina Genome Analyzer (GA) II, on the same DNA samples obtained from Lake Lanier, Atlanta. They inferred that there is ~90% assembly overlap of total sequences and high correlation ( $R^2 > 0.9$ ) for the in situ abundance of genes and genotypes between two platforms and sequence assemblies produced by Illumina that have the same quality to Roche 454 as assessed based on the base call error, frame shift frequency, and contig length. Ion Torrent (and more recently Ion Proton), Pacific

Biosciences (PacBio) SMRT sequencing, and Complete Genomics offering DNA nanoball sequencing are several other emerging sequencing technologies, but none of these emerging sequencing technologies have been completely applied and tested with metagenomic samples. NGS platforms are open to the process of multiplexing where hundreds to thousands of samples can be sequenced in parallel by adding 9–12 bp DNA tag to each DNA fragment before sequencing (Kumar et al. 2015). This tag is then used to make identification of the origin of the fragment from pooled samples and this allows for the synchronized exploration of various of bacterial communities in a highly cost-saving manner (Caporaso et al. 2012). The sequence reads produced in NGS based sequencing are typically shorter (except for Pacific Biosciences) than traditional Sanger sequencing reads (Liu et al. 2012) and they have origin from genome of varying organisms, which makes more more daunting the assembly and analysis of metagenomic NGS sequence data. In a typical sequencing based metagenomic project, post-sequencing steps are the most crucial steps that determine the investigation outcome. Most of the current assembly programs are devised to gather the sequences derived from single genome and therefore, not effective for a common metagenomic sequence data set that have various sequences with various origins. Without any reference genome for the assembly of genome sequences from unculturable representatives of metagenomic sequence pool, the task is made to be more complicated (Kumar et al. 2015).

## 2.7 Softwares for analyzing metagenomic sequences

Although bioinformatic tools are available at the moment, efforts are continuously exerted at several places to improve the accuracy of the alignment of the next generation sequence data (Smith, Xuan, and Zhang 2008). The development of sequence assemblers like MetaVelvet (Namiki et al. 2012) and Meta-IDBA (Peng et al. 2011) known to be specifically designed for *de novo* assembly of metagenomic sequence reads and metagenomics analysis pipelines such as MG-RAST (Meyer et al. 2008), MetAMOS (Treangen et al. 2013), MEGAN (Huson et al. 2007) and IMG/M (Markowitz et al. 2008) has given way to the researchers with limited expertise in bioinformatics to take on elaborative projects in metagenomics (Kumar et al. 2015).

### 2.7.1 MG-RAST

MG-RAST refers to a web-based analytical system that provides fully automated pipeline for purposes of quality control, feature prediction, functional annotation, and genomic comparisons. It only requires a minimum read length of 75 bp for gene prediction, similarity analysis, taxonomic binning, and functional classification (Thomas, Gilbert, and Meyer 2012). FragGeneScan is used in MG-RAST as a gene prediction method for the protein-coding region prediction in short reads using sequencing error models and codon usages in a hidden Markov model for the betterment of the prediction (Rho, Tang, and Ye 2010). Results are expressed in abundance profiles for specific taxa or functional annotations (Thomas, Gilbert, and Meyer 2012). MG-RAST boasts off a large-scale database for keeping intact the statistical results and metagenomic datasets (Dudhagara et al. 2015). To this extent, MG-RAST has more than

12,000 users, and more than 249,615 metagenomes that have been uploaded and analyzed, and from the total number of metagenomes, about 35,295 are publicly accessible and 108.06 Terabases analyzed as of May 2016.

### **2.7.2 MEGAN**

MEGAN is an independent computer software tool that can be used to study metagenomic data (Huson et al. 2007) and to study the annotation results derived from BLAST searches in a functional or taxonomic dendrogram (Thomas, Gilbert, and Meyer 2012). One major plus point for MEGAN is that the use of dendrograms to exhibit the metagenomic data can enable user to collapse the network of interpretation at a desired level and it can also make the analysis and interpretation of particular functional or taxonomic groups fast and hassle-free (Huson et al. 2007). To conduct a functional analysis, MEGAN assigns each read to the functional role of the highest scoring gene in BLAST output against a protein database (e.g., NCBI-NR), and then different functional roles are sorted out according to several SEED subsystems (Mitra et al. 2011). However, several limitations in using MEGAN for metagenomic analysis on functional identification cannot simply be dismissed. Firstly, the best score assignment could overlook the many putative functions. Due to the existence of sequencing error (Hoff et al. 2009), for the same sequencing read, it could have a function with identical matches of 32 out of 33 codons and also have a function with match score of 31 out of 33 codons. The MEGAN method will miss the second or even third best scoring functions that the read potentially has. Even more so, MEGAN just assigns one of the best functions (that have the same largest match values) to the short read. However, a gene could tend to

multiple functions in one go. To put in other words, MEGAN underestimates of the various functional roles that exist (An et al. 2014).

### **2.7.3 IMG/M**

The software or in its full name, the Integrated Microbial Genomes and Metagenomes (IMG/M) is a data storage, management, and analysis system for metagenomes hosted by the Joint Genome Institute (JGI) of U.S. Department of Energy (DOE) (Markowitz et al. 2008). IMG/M brings together the metagenome datasets with isolated microbial genomes from the IMG system (Markowitz et al. 2009). Being a domain-specific tool of IMG that suits the sequencing data from microbial communities, the IMG/M allows for combined analyses with all available draft and complete genomes, plasmids, and viruses in the public domain (Markowitz et al. 2012, Markowitz et al. 2014). It also offers free support for genomic and metagenomic data annotation, integration, and comparative analyses of integrated genomic and metagenomics data. The data content and analytical tools are continually updated. Pre-processing, quality control, and annotation of input data are performed by JGI's metagenome annotation system (Dudhagara et al. 2015). The stored IMG/M data are annotated via the various reference datasets to conduct three-tier analyses: (i) phylogenetic composition, (ii) functional or metabolic potential within individual microbiomes, and (iii) comparisons across microbiomes. IMG/M gives support for such analyses by combining the metagenome datasets with isolated microbial genomes from the IMG system (Markowitz et al. 2012). IMG/M gives the output data in multiple-cluster forms, further spurring its use as an online tool for environmental and organismal metagenomics