Optimization of DNA Extraction and Library Preparation for Metagenomic Analysis of Giant Panda in Malaysia

by

SURAJ ARVIND A/L V RAJASEGAR

A report submitted in fulfillment of the requirements for the degree of Bachelor of Applied Science (Natural Resources Science) with Honours

FACULTY OF EARTH SCIENCE

UNIVERSITI MALAYSIA KELANTAN

2017

DECLARATION

I declare that this thesis entitled "Optimization of DNA Extraction and Library Preparation for Metagenomic Analysis of Giant Panda in Malaysia" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : _____

Name : Suraj Arvind A/L V Rajasegar

Date : _____

ACKNOWLEDGEMENT

Firstly, I would like to take this opportunity to be heartily thankful to my supervisor, Dr. Jayaraj Vijaya Kumaran for his constant encouragement and guidance from the beginning until the final steps in perfecting and completing my final year project. I would like to also show my appreciation to my co-supervisor, Dr. Suganthi Appalasamy in guiding me in the field of bioinformatics throughout this research project. I would also like to express my gratitude to Associate Professor Dr. Aweng A/L Eh Rak, Dean of the Faculty of Earth Sciences, University Malaysia Kelantan (UMK); and Encik Zulhazman in Hamzah, Head of Department of Natural Resources Sciences, and UMK for their support throughout my 4 –year study period in UMK by ensuring a proper study environment.

The writing of this thesis would not have been possible without the assistance of my research mate, Yamunaa Sangaya to support and exchange information throughout the project. I would also like to extend my gratitude to the laboratory staffs in UMK for assisting me in acquiring the necessary chemicals, equipments, and apparatus on time. Last but not least, I offer my regards and blessings to all those who have supported me in any respect during the completion of my final year research, especially my family members.

Optimization of DNA Extraction and Library Preparation for Metagenomic Analysis of Giant Panda in Malaysia

ABSTRACT

The objective of this study is to determine the expected minimum number of sequence reads read needed to achieve full coverage of the gut microbial species of the giant panda metagenome, firstly by analyzing 5 different mammalian metagenomes, namely that of the horse, coyote, whitetail deer, humpback whale and the bottlenose dolphin as reference; based on an approximate of 1 000 000 sequence read estimation. After rarefaction analysis using MG-RAST Version 3.0 analysis pipeline, an average value of 775 075 reads, which is close to expected 1 000 000 reads, is deduced. Next, the fecal matter of the giant panda is sampled at Zoo Negara, and metagenomic DNA extraction is performed to recover four final samples of raw metagenomic DNA, namely Panda A, Panda B, Panda C and Panda D. These four samples are then subjected to qualification and quantification analysis; where the results show that the samples are still viable and sufficient in yield to be used for library preparation, despite slight degradation and contamination during the extraction process. However, only Panda A and Panda B are selected for library preparation through six phase namely; DNA fragmentation, end repair and size selection, adenylate 3' ends, adapter ligation, library validation, as well as library normalization and pooling. Panda C and Panda D are kept as reserves. After library preparation, the pooled samples are sequenced using the Illumina™ MiSeq® next generation sequencer. The results of this research serve as a foundation for further studies of the giant panda metagenome.

Optimizasi Pengesktrakan DNA dan Penyediaaan Perpustakaan Untuk Analisis Metagenomik Beruang Panda Gergasi di Malaysia

ABSTRAK

Matlamat penyelidikan ini adalah untuk mengenalpasti bilangan bacaan penjujukan minimum yang diperlukan untuk mencapai liputan species mikrob sistem penghadaman yang penuh dalam metagenom beruang panda gergasi, pertama sekali dengan membuat analisis metagenom daripada lima species mamalia yang berbeza, iaitu kuda, anjing hutan, rusa "whitetail", ikan paus kelasa, dengan ikan lumba lumba "bottlenose" sebagai rujukan; berdasarkan anggaran dalam lebih kurang 1000 000 bacaan penjujukan. Selepas menjalankan analisis perjernihan menggunakan saluran analisis atas talian MG-RAST Versi 3.0, 775 075 bacaan penujukan telah dikenalpasti sebagai nilai purata, di mana nilai tersebut adalah hampir dengan 1 000 000 bacaan penjujukan yang diramalkan sebelum itu. Selepas itu, najis beruang panda gergasi telah disampel di Zoo Negara, dan pengekstrakan metagenomic telah dijalankan, di mana empat sampel DNA metagenomik mentah telah dihimpunkan; berlabel Panda A, Panda B, Panda C, dan Panda D. Analisis kualifikasi dan kuantifikasi telah dijalankan atas keempat-empat sampel tersebut, di mana keputusan menunjukkan bahawa sampel-sampel tersebut adalah sah dan mencukupi, walaupun terdapat degradasi DNA yang sedikit dan pencemaran daripada proses pengekstrakan. Bagaimanapun, hanya Panda A dan Panda B sahaja dipilih untuk menjalani proses penyediaan perpustakaan melalui enam fasa; iaitu pembaikan penghujung dan pemilihan saiz, penghujung "adenylate 3'", ligasi penyesuai, pengesahan perpustakaan, dan juga normalisasi perpustakaan dengan pengumpulan. Panda C dan Panda D disimpan sebagai simpanan. Selepas penyediaan perpustakaan, sampel yang terkumpul dijujuk menggunakan penjujuk "next generation", iaitu Illumina™ MiSeq®. Hasil penyelidikan ini berfungsi sebagai dasar untuk mendalami lagi penyelidikan mengenai metagenom beruang panda gergasi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

16S	16 Svenberg unit(s)
AGE	Agarose gel electrophoresis
ALP	Adapter Ligation Plate
ATL	A-Tailing Mix
BAC	Bacterial artificial chromosome
bp	Base pair
CAMERA	Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis
CAP	Clean-up ALP Plate
CEP	Clean-up End Repair Plate
CFP	Covaris Fragmentation Plate
CH_3COOH	Glacial acetic acid
CSP	Clean-up Sheared DNA Plate
DAP	DNA Adapter Plate
DCT	Diluted Cluster Template
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dsDNA	double-stranded DNA
EDTA	Ethylenediaminetetraacetic acid
et al.	Et alii

EtBr	Ethidium bromide
EtOH	Ethanol
g	Gram(s)
Gb	Gigabase(s)
g/cm3	Gram(s) per cubic centimeter
g/mol	Gram(s) per mole
HCl	Hydrochloric acid
*Hind*III	Hin D Three
IMG/M	Integrated Microbial Genomes and Metagenomes
IMP	Insert Modification Plate
IUCN	International Union for Conservation of Nature
KCl	Potassium chloride
kg	Kilogram(s)
kg/day	Kilogram(s) per day
KH_2PO_4	Potassium dihydrogen phosphate
M	Mega, molar
m	Meter(s)
MGI	Malaysia Genome Institute
MG-RAST	Metagenomic Rapid Annotations using Subsystems Technology
ml	Milliliter(s)
mM	Millimolar
mm	Milimeters(s)

M.W.	Molar weight
MYA	Million years ago
µg	Microgram(s)
µg/ml	Microgram(s) per milliliter
µl	Microliter(s)
MGI	Malaysia Genome Institute
MYA	Million years ago
NaCl	Sodium chloride
NaOH	Sodium hydroxide
n.d.	No date
ng/µl	Nanogram(s) per microliter
NGS	Next generation sequencing
nM	Nanomolar
nm	Nanometer(s)
ORF	Open reading frame
OTU	Operational taxonomic unit(s)
PBS	Phosphate buffered saline
PDP	Pooled DCT Plate
pH	*Potentia hydrogenii*
QC	Quality control
qPCR	Quantity polymerase chain reaction
rDNA	Ribosomal deoxyribonucleic acid

rpm	Rotation(s) per minute
rRNA	Ribosomal ribonucleic acid
SDS	Sodium dodecyl sulfate
SPB	Sample Purification Beads
spp.	Species
STE	Sodium chloride-Tris-EDTA
TAE	Tris-acetate-EDTA
TE	Tris-EDTA
Tris	tris(hydroxymethyl)aminomethane
TSPI	Target Sample Plate
UK	United Kingdom
USA	United States of America
UV	Ultraviolet
V	Volt(s)
x	Time(s)

LIST OF SYMBOLS

β	Beta
°C	Degree(s) Celsius
%	Percent
®	Registered
™	Trademark

CHAPTER 1

INTRODUCTION

1.1 Background of Study

A flagship species in wildlife conservation and an adored icon worldwide, *Ailuropoda melanoleuca*, otherwise known as the giant panda, is one of the most endangered animals on the planet, where there are only 2500 to 3000 left in the wilderness (Wei *et al*., 2014; Zhan *et al*., 2006).

The giant panda population, however is suffering grave threats of both anthropological and disease factors, making it an endangered species according to the World Conservation Union's (IUCN's) Red List of Threatened Species (Smithsonian National Zoological Park, n.d.). Examples of the former include habitat loss, fragmentation, and bamboo harvesting, poaching, and mass tourism whereas the latter include communicable disease, parasitic disease, obstetrical disease, and internal disease, particularly gastrointestinal ones like intussusceptions, volvulus, and ileus (Chengdu Research Base of Giant Panda Breeding, 2007).

Metagenomics is a scientific term coined in 1998 (Handelsman *et al*., 1998), which means "genome study that transcends". This is because metagenomics focuses on analyzing deoxyribononucleic acid (DNA) extracted directly from mixed microbial communities in a particular environment instead of culturing, due to the fact that not all microbes are able to culture on its own in laboratory conditions (Handelsman, 2004). This is because many microbial species live in symbiosis with

other microbial species and need optimal natural conditions in order to thrive. Metagenomics serves to determine environmental microbial community diversity and activity, biosynthetic pathways, and to detect novel and functional individual genes; where all these are possible by analyzing the 16S ribosomal DNA (rDNA) genes or sequences present in the metagenomic sample (Woese, 1987). By understanding metagenomics, the phylogenetic diversity and relative abundance of microbial species in the gut of both the male and female giant panda can be studied, indirectly analyzing and monitoring their activity on cellulose, hemicellulose, and lignin, ultimately providing vital information for the diet management and conservation of the endangered giant panda.

1.2 Problem Statement

Recent research has found out that despite the giant panda's specialized bamboo diet, its gut microbiota is found to be closer to that of the gastrointestinal system of carnivores. However, other recent researches has also found out that there is a small percentage of microbes in giant panda gut that are capable of digesting cellulose and hemicelluloses, which comprises the majority percentage of bamboo dry weight. To study the relative abundance and phylogenetic diversity of these gut microbes, metagenomic analysis of the feces of the male giant panda only, and then the relative abundance and phylogenetic diversity comparison of both male and female giant panda will be done. This approach is important because previously, no research on the comparison of male and female giant panda gut microbiota structure is conducted. Thus, data analysis from this research is capable in captive and overall health management of the endangered species.

1.3 Objectives

The purpose of this research is to estimate the minimum number of sequence reads necessary to achieve bacterial species richness coverage of the giant panda metagenome, based on reference metagenomic analysis of coyote, horse, whitetail deer, humpback whale, and bottlenose dolphin; in order to optimize phenol chloroform DNA extraction and library preparation of giant panda metagenome.

CHAPTER 2

LITERATURE REVIEW

2.1 Characteristics of Giant Panda

The giant panda belongs to the family Ursidae, in the order Carnivora, where it has evolved from omnivorous bears (Xue *et al*., 2015). According to Pastor *et al.* (2008), the giant panda has a carnivorous ancestor due to the structure, distribution, and morphological lingual papillae characteristics resembling more to that of a carnivore. The earliest known ancestor of the giant panda is *Ailurarctos lufengensis*, where it had appeared during the late Miocene period, about 8 million years ago (MYA) (Qiu and Qi, 1989). Later during the early Pleistocene, *Ailuropoda microta* appeared, having the smallest body size for any known giant panda. In the middle of the Pleistocene period however, it has evolved into the largest known giant panda, named *Ailuropoda baconi* (Pei, 1974; Tsiang-ke, 1974; Jin *et al*., 2007), before eventually evolving into the smaller *Ailuropoda melanoleuca*.

Morphologically and proportionally, the giant panda is similar to other bears, but its distinctive feature is its thick, woolly coat, where it is black around the eyes, ears, limbs, shoulders, and tail on white (Davis, 1964 and Hendey, 1972). It has a head-body length of 1.2 to 1.5m, tail length of approximately 130mm, and weighs around 75 to 160kg (Chorn and Hoffman, 1978).

The giant panda in China mainly live in the old-growth, subalpine coniferous forests in Gansu, Sichuan, and Shaanxi provinces. According to Brambel (1976), the giant panda range in these forests is from 2600m to 3500m, where their main food *Sinarundinaria* bamboo can be found growing abundantly and otherwise known as China cane. Climate-wise, the forests are damp, where it is cool in summer and cold in winter, together with rain, hail, and snow (Wang Sung and Lu Chang-kun, 1973; Morris and Morris, 1966). When above freezing point in winter, temperature and humidity range from approximately -10°C and 90% relative humidity , to 25°C and 60% in summer (Brambell, 1976).

FIGURE 2.1 The *Sinarundinaria nitida* bamboo, otherwise known as China cane. (Source: Yang Shijian, 2012)

About 99% of the herbivorous giant panda food source is bamboo, with the remaining percentage of 1% being eggs, fish, fruits, honey, shrub leaves and yams, in

contrast to all other ursids (Schaller *et al*., 1985; Schaller *et al*., 1989). It was estimated that the omnivorous ancestors of the giant panda started eating bamboo about 7 MYA and completed its dietary switch about 2 to 2.4 MYA (Schaller *et al*., 1985). Several mutations in the genome of the giant panda has a probable contribution to the phenomenon, including the pseudogenization of the umami taste receptor gene *TAS1R1* since about 4.2 MYA and dopamine metabolism defects in the appetite-reward system (Jin *et al*., 2011). The panda may have also switched to bamboo due to a possibility in meat scarcity.

However due to the fact that the bear is actually a carnivore by phylogeny, it retains a carnivorous digestive system; a simple stomach, degenerate cecum, and a short, straight colon with rapid transit interval, unlike other herbivores (Beijing Zoo, 1986). In addition, the giant panda genome also encodes the digestive enzymes necessary for a carnivorous diet only (Li *et al*., 2010), which makes it genetically deficient in digesting cellulose, lignocelluloses, and hemicellulose, having no alternative but to rely on gut symbionts to adapt to its highly fibrous diet. Metagenomic research of 112 fecal samples by Xue *et al.* (2015) proved the giant panda however has extremely low diversity of gut microbiota and an overall structure that diverged from the gut microbial communities of non-panda herbivores, resembling more of carnivorous and omnivorous bears. Nevertheless, Zhu *et al*. (2011) and Fang *et al*. (2012) have discovered that the giant panda has also a small percentage of cellulose and lignin degrading bacteria in its digestive tract. These putative microbes together with special morphological modifications, including pseudothumbs; well developed skull and chewing muscles; high amount of food ingestion (12.5kg/day); rapid transit time; and thick mucus lining the digestive tract,

likely have emerged as a result of adapting to a highly fibrous bamboo diet within the constraints imposed by the panda's carnivorous digestive system (Zhu *et al*., 2011).

2.2 Overall Gut Microbiota Structure Of Giant Panda

Diet and phylogeny are the two known, main factors that influence gut microbiota structure at various taxonomic scales (Delsuc *et al*., 2014). In this case, the phylogeny of the giant panda overpowers the influence of diet. Despite the specialized bamboo diet, the giant panda still retains a carnivorous digestive system; a simple stomach, degenerate cecum and a short, straight colon. According to a recent study by Xue *et al*. (2015), the bamboo-eating giant panda harbors low-diversity, carnivorous gut microbiota with excessive variation according to seasons, deviating from the general postulate that "microbiota adaptation to diet is consistent across different mammalian lineages". This harboring of simple gut microbiota indicates that the microbial community has not fully evolved in adapting to the newly adopted diet, which may affect the co-evolutionary fitness of the ursid (Xue *et al*., 2015). Dominated mainly by phyla Firmicutes and Proteobacteria (Zhu *et al*., 2011), the gut microbiota has an overall structure that diverged from that of non-panda herbivores, but converged with that of carnivorous and omnivorous bears. Instead of harboring Ruminococcaceae and Bacteroidales that are abundant in the digestive system of other herbivores, the giant panda gut has a high relative abundance of *Clostridium* and *Streptococcus*, Gram-positive bacteria of the phylum Firmicutes; as well as *Escherichia/Shigella* and Enterobacteriaceae, Gram-negative bacteria of the phylum Proteobacteria (Susan, 2005) – almost none of which aids in polysaccharide digestion. Nevertheless, putative cellulose and hemicellulose degrading Clostridia

bacteria, including 13 “operational taxonomic units” (OTUs) have recently been found. In fact, 7 of these 13 OTUs were unique to pandas and not other mammals (Zhu *et al*., 2011). Two phylotypes were also found to be affiliated with the known lignin digesting bacterium *Pseudomonas putida* and the mangrove forest bacteria (Fang *et al*., 2012). In short, these discoveries reflect the low digestibility of the cellulose and hemicellulose of the giant pandas bamboo diet, despite the fact that its carnivorous gut harbors carnivore-like microorganisms.

2.3 Cellulose Metabolism in the Gut of Giant Panda

The stringy polysaccharide cellulose is the main component of the giant pandas bamboo diet. Though the giant panda can absorb non-cellulosic material from bamboo using enzymes coded from its own genomes, digestion of cellulose and hemicelluloses is impossible with the absence of genes in the giant panda genome that codes cellulose-digesting enzymes. Yet, the bear can only digest 17% of the dry mater it consumes daily, where it has low digestion coefficients for bamboo hemicelluloses (27%) and cellulose which is 8% (Dierenfeld *et al*., 1982). This reaches a postulate that there are cellulose-digesting microbes in the digestive system. Zhu *et al*. (2011) have collected fecal samples of captive and wild giant pandas and performed analysis of 16S ribosomal RNA (rRNA) gene sequences using *de novo* sequencing to determine the functional attributes of the gut symbionts. Based on the analysis, the majority of the microbes were of phyla Firmicutes and Proteobacteria, with the remaining percentage belonging to phyla Acidobacteria, Actinobacteria, Cyanobacteria, and Bacteroidetes. Nevertheless, the major focus of the research was the detection of 13 Clostridian OTUs of phyla Firmicutes, which are

closely related to *Clostridium* group I and group XIVa, microbes that are known to have cellulose-digesting properties (Collins *et al*., 1994). Moreover, out of these 13 OTUs, 7 were endemic only to the giant panda compared to other mammals. These findings are result of the detection of genes coding for hemicelluloses and cellulose-degrading enzymes within the genus *Clostridium*. These putative genes code for one hemicelluloses-digesting enzyme and two cellulose-digesting enzymes, namely xylan-1,4-β-xyloxidase; cellulase and β-glucosidase; all within *Clostridium* group I(Zhu *et al*., 2011). The researches then compared the glycoside hydrolase profiles of the giant panda with that of other herbivores and omnivores, including humans; and discovered that the ursid has a moderate abundance of oligosaccharide-digesting enzymes (36%), close to that of humans which is 37%. However, as of endocellulases and cellulases, the giant panda scored the lowest abundance among the herbivores. Nevertheless, this fact also implies that giant panda is capable of digesting cellulose and hemicellulose at a low degree, despite its genome which cannot code polysaccharide-digesting enzymes. Therefore, the cellulose metabolism by the gut microbiota of the giant panda is still primitive and inefficient.

2.4 Lignin Oxidation in the Gut of Giant Panda

Previously in a study to prove there is evidence for lignin oxidation by the giant panda fecal microbiome in 2011, Zhu *et al*. has revealed that there is small abundance of cellulose and hemicellulose-digesting gut microbiota closely-related to the *Clostridium* group I and XIVa, found in the digestive tract of the giant panda, capable of cellulose and hemicellulose digestion at a small degree. However, Fang *et al*. (2012) argued that cellulose and hemicellulose are embedded in a matrix of

lignin, where the latter is 25% of the dry weight of bamboo (Bugg *et al*., 2011). These three biopolymers form a complex structure together, called lignocelluloses – the "key" component of the bamboo diet. Lignin is a primary barrier in the digestion of cellulose-based biopolymers. Nevertheless, since the digestion of cellulose and hemicellulose is possible, it is speculated that the digestion of lignin is also possible. Possible presence of lignin-degrading bacteria and lignolysis-enzyme-coding gene is screened out using the 16S rRNA gene library constructed from the pooled microbial genome of the giant panda dung. Evaluation on the bacterial laccase activity against phenolic lignin-related moieties and its effect on lignin with or without mediators is also carried out. Based on the clones obtained, 53% belonged to the phyla Proteobacteria, where a particular phylotype, B_35 is closely linked to uncultured to mangrove forest bacteria and *Pseudomonas spp.*, particularly *P. putida*, both capable of degrading lignin (Ahmad *et al*., 2010). Laccase-coding genes were also screened and two multicopper oxidoreductase genes were detected, one of them being lac51, clustering with multicopper oxidases from *Pseudomonas*. At 38°C and pH of 7.5 which simulates the conditions in the gastrointestinal tract of the giant panda, lac51 showed about 60% of its maximum activity, metabolizing aldehydes, aromatic acids, and alcohols such as 2,6-dimethoxyphenol, guaiacol, veratryl alcohol, sinapinic acid, and ferulic acid – all of which are lignin-derived moieties. In short, Lac51 has been shown to metabolize lignin.

2.5 Morphological and Genetic Adaptations of the Giant Panda to Bamboo Diet

Despite the poor efficiency of the gut microbial community of the giant panda to breakdown the low-nutrition and low-calorific bamboo matter, it has developed morphological and genetic adaptations in order to compensate the digestive inefficiency of its carnivorous digestive system on cellulose, hemicellulose, and lignin. One example is the evolution of the pseudothumb; an enlarged, modified radial sesamoid bone, used for grasping and manipulating bamboo while feeding (Endo, Hayashi, *et al*., 1999). Giant pandas also possess skulls that are dense in bone mass; featuring expanded zygomatic arches and well-developed mandible structure (Zhang *et al*., 2007), linked with zygomatic-mandibular muscle attachment. They also possess large, flat teeth which have elaborate crown patterns. All these features are prominent in the mastication of tough, fibrous bamboo during feeding, resulting in efficient physical breakdown of bamboo matter. Adaptation has also occurred at the molecular level where the umami taste receptor *TAS1R1*, has undergone pseudogenization due to a 2 base pair (bp) insertion in exon 3 and a 6-bp deletion in exon 6 (Li *et al*., 2010). The said receptor is responsible for sensing components of meat and other protein-rich foods. Therefore, the loss of function of the *TAS1R1* gene may have contributed to the panda's dietary switch, which happened about 4.2 MYA (Zhao *et al*., 2010).

2.6 Principles of Metagenomics

In 1998, Handelsman coined "metagenomics", a scientific term used to describe the direct genetic analysis of mixed genomes sampled from the

environment. This field of bioinformatics is developed as means to study the phylogenetic and functional diversity, of the uncultured microbial world. As of today, 52 phyla of microorganisms have been delineated, where the majority is dominated by uncultured microbiota (Rappe and Giovannoni, 2003). This world of unseen microbes may be resistant to culturing due to lack of necessary surfaces, nutrients, or symbionts, incorrect combinations of atmospheric gas composition, pressure, or temperature, and accumulation of toxic by-products from their own metabolism (Simu and Hagstrom, 2004). The advent in the field started off with the cloning of the environmental DNA into a bacterial artificial chromosome (BAC), fosmid, or cosmid vectors producing large insert libraries, representing the complementary genetic whole of a single environment. This was followed by the screening of functional gene manifestation (Handelsman *et al*., 1998), and was finally supplemented by the metagenomic DNA direct random shotgun sequencing (Tyson *et al*., 2004; Venter *et al*., 2004). These milestones not only fortified the approach of metagenomics, but also proved that the microbial world is rich in functional gene diversity (Simon and Daniel, 2011); providing a clearer insight rather than evolution and phylogenetic analysis alone, which is usually rooted on the idea of singular-gene diversity, like the 16S rRNA gene.

Research on the functional gene diversity provides vital information on evolutionary profiles of environmental community structure and collective function, correlations between phylogeny and function of uncultured microbes, as well as genetic information of novel biocatalysts; hence serving as a vital instrument in formulating general hypotheses on microbial function (Beja *et al*., 2000; Nicol and Schleper, 2006). This is made possible by using the 16S rDNA as a phylogenetic marker when identifying clones belonging to a specific organism, and then the gene

which is flanked by the said DNA is sequenced to gain information about the microbial physiology. Through this technique, scientists are able to understand the ability of prokaryotes to thrive, adapt, and populate any natural environment through their physiological and metabolic diversity. Thus, this diversity can be exploited and applied in many lucrative fields; medicine and healthcare particularly, which is of high demand today to facilitate novel drug discovery. This is supported by the fact that symbiotic bacteria that are resistant to pure culture techniques are one of the major sources of therapeutic products, which leads to the point that an increased understanding of microbial antibiotic resistance, enabling the development of antibiotic resistance-inhibiting compounds (Riesenfeld *et al*., 2004). Metagenomic analysis includes environmental sample DNA isolation, splicing the DNA into an appropriate vector, insertion into host bacterium for cloning, and transformant screening. By understanding metagenomics, the phylogenetic diversity and relative abundance of microbial species in the gut of both the male and female giant panda can be studied, indirectly analyzing and monitoring their activity on cellulose, hemicellulose, and lignin, ultimately providing vital information for the diet management and conservation of the endangered giant panda.

The basic protocol for a metagenomic project starts with environmental community sampling and DNA extraction, DNA sequencing, annotation, and subsequently by statistical analysis and data sharing. In the sampling process, DNA extracted shall be of high quality and quantity, representing all types of cells present in sample for subsequent library preparation and sequencing. If a host-associated target community is involved, in this case where the gut microbiota of the giant panda is to be studied, fractionation is crucial to ensure maximal acquirement of target community DNA and minimum contamination from host DNA (Burke *et al*.,

2009; Thomas *et al*., 2010) and enzymatic inhibitors; all possible through physical, chemical separation and cell isolation. The next step is DNA sequencing, where libraries are constructed to amplify the previously extracted starting material in order to facilitate detection by the applied sequencing technologies. Next, the precise order of the nucleotides within the DNA fragments is determined using two-dimensional chromatography and fluorescence, particularly in the 454/Roche and Illumina/Solexa technology.

Annotation then comes in where detected genes are assigned with functions, which takes place in two steps; step 1 – feature prediction, where genes of interest are identified and step 2 – functional annotation, where taxonomic neighbors and gene are designated. All the detected genes are annotated by mapping against existing annotated data of gene or protein libraries. However, sequences which cannot be mapped to any known sequence space are assigned as "open reading frames"(ORF). ORFs basically are genes that code for unknown biochemicals, having no sequence similarity with known genes, but possessing structural similarity with known proteins in terms of folds or families. Currently, metadata bases like Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA), Integrated Microbial Genomes and Metagenomes (IMG/M), and Metagenomic Rapid Annotations using Subsystems Technology (MG-RAST) are available for processing and storing metagenome information (Glass *et al*., 2010; Markowitz *et al*., 2008; Sun *et al*., 2011), where MG-RAST will be the chosen program to analyze metagenomic data of the giant panda gut microbiota. MG-RAST basically stands for Metagenomic Rapid Annotations using Subsystems Technology; a comparative genomic space, data storehouse, and an automated analysis pipeline which carries out quality control, prediction of

features and annotation of gene functions, where the results are usually conveyed in the form of abundance profiles for a specific functional annotation or taxa. It is by far one of the most convenient metagenomic data server as products generated can be downloaded, shared and published within the same place; allowing integration of metadata into statistics by statistical technique comparisons, all in the effort to standardize annotations and centralize resources for future metagenomic discoveries.

2.7 Phenol-Chloroform DNA Extraction

The phenol-chloroform DNA extraction is a technique to purify DNA samples by removing proteins from an aqueous solution of homogenized tissue or lysed cells by adding equal volumes of phenol and chloroform: isoamyl alcohol mixture, taking advantage of the fact that protein removal is more efficient in the presence of two different organic solvents compared to one type alone (Kirby, 1956). Chloroform is added to ensure phase separation of phenol and water because the denser chloroform (1.47g/cm3) is miscible with phenol; forcing a sharper separation of the organic and aqueous phases, hence the easier removal of the aqueous phase and prevention of organic phase cross contamination. Since the said organic solvents and water are immiscible, two phases form, the heavier organic phase which sits at the bottom and the lighter aqueous phase which floats at the top. The phases are then mixed, forcing the organic phase to disperse into the aqueous phase to form a droplet emulsion. The proteins in the mixture is denatured and separated into the organic phase or phase interphase whereas the DNA remains in the aqueous phase after centrifugation. This is because the proteins, which are made of long chains of amino acids, have folded structures with side chains which are either polar or non-polar.

These structures are folded so that the hydrophobic, non-polar side chains are folded inside, away from the external water environment of cells and the hydrophobic parts are exposed, in opposite. However, when proteins are introduced to the organic solvents, the hydrophobic side chains start interacting with the non-polar organic solvent molecules outside whereas the hydrophilic side chains start to shield away into the inside of the globular structure, causing them to unfold and denature, as a result. DNA on the other hand, is a polar, hydrophilic molecule, due to the negative charges of its phosphate backbone, hence its dissolution in water. The nucleic acid sample is then precipitated using ethanol and salt.

CHAPTER 3

MATERIALS AND METHOD

3.1 Materials

3.1.1 Equipment

The equipments that will be used in this research project together with the manufacturers are as follows: autoclave machine (Hirayama, Japan), pH meter (Dakota Instruments, USA), electronic balance (Sun Lin Jewelry Instruments, China), hot stir plate (Thermo Fisher, USA) for preparations of buffer and reagents; -20°C freezer (Sanyo, Japan), -80°C freezer (Thermo Fisher, USA), vortex (USA Scientific, USA), mini centrifuge (Recenttec, Taiwan), microcentrifuge (Beckman Coulter, USA) and chilling/heating plate (Torrey Pines Scientific, USA) for metagenomic DNA extraction; water distilling system (Merck, Germany), water purification system (Merck, Germany), spectrophotometer (Thermo Fisher, USA), UV transilluminator (Aplegen, USA), microwave oven (Sharp, Japan), power supply (Cleaver Scientific, UK) gel electrophoresis tank (Cleaver Scientific, UK), hybrid graphic printer (Sony, Japan), for extracted DNA quality and quantity analysis; focused-ultrasonicator (Covaris, USA), bioanalyzer (Agilent Technologies, USA), vortex (IKA-Works, Germany), concentrator (Eppendorf, Germany), fluorometer (Thermo Fisher, USA), thermo mixer (Eppendorf, Germany), thermal cycler (Bio-Rad Laboratories, USA), and master cycler (Eppendorf, Germany) for metagenomic library preparation; and the Illumina™ MiSeq® sequencer for sequencing.

3.1.2 Apparatus

The apparatus that will be used in this research are 100ml, 200ml, 250ml and 1000ml beakers; 50ml and 200ml conical flasks; forceps; 10ml, 50ml, 100ml, and 1000ml measuring cylinders; 10µl, 100µl, and 1000µl micropipettes; 10µl, 100µl, and 1000µl micropipette tips, dropper, tile, spatula, 2ml microcentrifuge tubes, 50ml conical centrifuge tubes, conical centrifuge rack, 100ml and 1000ml reagent bottles, 1000ml graduated laboratory bottle, microcentrifuge tube rack, magnetic stand, thermo conductive tube rack, 0.3ml PCR plate, DNA chips, glass rod, 0.2 microns micro filter syringe, zipper bags, surgical masks, gloves, grafting tape, Microseal 'B' adhesive seal, aluminium foil, laboratory wipes, delicate task wipers, gel cast, comb, ice box, and weighing boat.

3.1.3 Reagents and Chemicals

In order to prepare buffers and reagents for metagenomic DNA extraction; tris(hydroxymethyl)aminomethane (Tris) base, ethylenediaminetetraacetic acid (EDTA), chloroform, isoamyl alcohol, double-distilled water, sodium chloride (NaCl), sodium hydroxide (NaOH), sodium dodecyl sulfate (SDS) powder, hydrochloric acid (HCl), and glacial acetic acid (CH_3COOH) , lysozyme, Proteinase K, absolute ethanol, 70% ethanol (EtOH), 80% EtOH, phenol, and sodium acetate is used. For AGE, SYBR® Green I nucleic acid stain 10 000x concentrate in dimethyl sulfoxide (DMSO), agarose gel powder, 1000bp DNA ladder, and loading dye are used. The Illumina™ TruSeq® polymerase chain reaction (PCR)-Free LT Kit, PCR-grade water, nuclease free water, KAPA™ Library Quantification Kit for Illumina™

platforms, and Qubit® double-stranded DNA (dsDNA) BR Assay Kit is used to prepare the metagenomic library.

3.2 Preparations of Buffer Solutions and Reagents

3.2.1 Preparation of 1M Tris-HCL pH 8.0 (100ml Stock Solution)

1M of Tris-HCL of pH 8.0 is prepared by weighing out 12.11g of Tris base (M.W. = 121.14g/mol) and adding to it an appropriate amount of HCl to adjust the pH to the desired level. Next, the volume is adjusted to 100ml with double distilled water, stored in a 100ml reagent bottle, autoclaved and kept at room temperature.

3.2.2 Preparation of 0.5M EDTA pH 8.0 (100ml Stock Solution)

0.5M of EDTA of pH 8.0 is prepared by weighing out 18.61g of disodium ethylenediamine tetraacetate.$2H_20$ (M.W= 372.2g/mol) and adding to it the appropriate amount of NaOH pellet to adjust the pH to the desired level. Next, the volume is brought up to 100ml with doubled distilled water, stored in a 100ml reagent bottle, autoclaved and kept at room temperature.

3.2.3 Preparation of Chloroform-Isoamyl Alcohol (CIA, 24:1, 50ml Stock Solution)

Chloroform-isoamyl alcohol mixture is prepared by adding 2ml of isoamyl alcohol to 48ml of chloroform in a 100ml reagent bottle. Next, the bottle is wrapped

in aluminium foil to prevent degradation of the light sensitive mixture, shaken well, and stored at room temperature.

3.2.4 Preparation of 10% SDS Solution (100ml Stock Solution)

10% of SDS solution is prepared by weighing out 10g of SDS powder into a conical flask and dissolved in 80ml of double distilled water at 68°C by heating and stirring. Then, the volume is adjusted to 100ml, filtered into a 100ml reagent bottle using 0.2 microns micro filter syringe and stored at room temperature. No autoclaving is needed.

3.2.5 Preparation of STE Buffer (100ml Stock Solution)

STE buffer is prepared by mixing well 2ml of 1M Tris-HCl, 1ml 5M NaCl, 0.2ml 0.5M EDTA, and 96.8ml double distilled and autoclaved before storing at room temperature in a 100ml reagent bottle.

3.2.6 Preparation 50x TAE Buffer (100ml Stock Solution)

50x TAE buffer is prepared by mixing together 24.2g Tris base and 5.71ml glacial acetic acid. Next, 10ml of 0.5M EDTA solution pH 8.0 is added and the volume is brought up to 100ml with double distilled water. Then, the mixture is stored in a 100ml reagent bottle and kept at room temperature.

3.2.7 Preparation of 1x TAE Buffer (1000ml Stock Solution)

1x TAE buffer is prepared by adding 20ml of 50x TAE buffer solutions into 980ml of double distilled water and mixed well. The mixture is then stored in 1000ml reagent bottle and kept at room temperature.

3.2.8 Preparation of $T_{10}E_1$ Buffer Solution pH 8.0 (100ml Stock Solution)

Preparation of $T_{10}E_1$ buffer solution pH 8.0 is done by adding 1ml of Tris-HCl pH 8.0 into 0.2ml of 0.5M EDTA solution pH 8.0. The volume is then adjusted to 100ml with double distilled water, autoclaved and stored in a 100ml reagent bottle at room temperature.

3.2.9 Preparation of 70% Ethanol (1000ml Stock Solution)

1000ml of 70% ethanol is prepared by adding 700ml absolute ethanol to 300ml of double distilled water and mixed well. Next, the solution is stored in 1000ml graduated laboratory bottle and stored at -20°C.

3.2.10 Preparation of 1x TAE Buffer with 5% SYBR® Green I (500ml Stock Solution)

1x TAE buffer with 5% SYBR® Green I is prepared by adding 5µl of SYBR Safe to 500ml of 1x TAE buffer, and stored in a 1000ml graduated laboratory bottle at room temperature. The bottle is wrapped with aluminium foil as the mixture is

light sensitive. This mixture is used as an ingredient for casting agarose gel for agarose gel electrophoresis.

3.2.11 Preparation of 3M Sodium Acetate (100ml Stock Solution)

To make 3M sodium acetate, 24.6g of anhydrous sodium acetate is added to 80ml of double distilled water. Once the salts have dissolved completely, the pH is adjusted to 5.2 with glacial acetic acid. Next, the volume is brought up to 100ml in a volumetric flask and is decanted into a 100ml reagent bottled, autoclaved, and stored at room temperature.

3.2.12 Preparation of Electrophoresis Agarose Gel (60ml Working Solution)

To prepare 60ml of electrophoresis agarose gel, 0.6g of agarose powder is weighed out and added to 60ml of 1x TAE buffer with 5% SYBR® Green I and micro waved for 2 minutes until completely dissolution. The mixture is left to cool down to a warm temperature for 5 minutes before being poured into the gel cast together with comb; to be set for another 5 minutes. Finally, the comb is removed, leaving behind wells.

3.3 Method

This experiment to estimate the minimum number of sequence reads that should be obtained to achieve bacterial species richness coverage of the giant panda metagenome includes several processes like the reference sequence read-species richness analysis, sampling of the male stool, phenol-chloroform DNA extraction,

quality control checking, metagenome library preparation, next generation sequencing (NGS), and online metagenome data analysis.

3.3.1 Reference Sequence Read - Species Richness Analysis

A metagenome analysis was done on five metagenomes of five different mammals, namely the coyote, horse, whitetail deer, humpback whale, and the bottlenose dolphin to set as a reference for the estimation of the minimal sequence reads needed to attain full gut microbial species richness coverage of the giant panda. The five datasets are analyzed in MG-RAST Version 3.0, using the best hit classification of the organism abundance mode; with maximum e-value cutoff set at 1e-5, minimum percentage identity cutoff at 60%, and minimum alignment length cutoff at 15 units in aa for protein and bp for RNA databases, with Kyoto Encyclopedia of Genes and Genomes (KEGG) as the annotation sources. The five metagenomes have the respective ID of 4526727.3, 4526729.3, 4526730.3, 4526723.3 and 4526724.3. Then, data visualization in the form of rarefaction plot, organism table, and organism tree is used to visualize and compare the five metagenomic data, which as illustrated in Section 4.1. Based on these reference metagenomes, an estimated of 1 000 000 sequence reads is needed to attain whole microbial species coverage in the metagenome of giant panda. Only after this analysis and estimation is where the sampling of the giant panda fecal matter follows.

3.3.2 Sampling

Fresh fecal sample of the adult male giant panda is collected by the zoo keeper at 08:00 hours within the giant panda enclosure in Zoo Negara and is immediately sealed in 2 sterile zip-lock bags kept together with ice packs and sealed in a styrofoam box to retain freshness. During sampling, it is ensured that the fresh feces did not come in contact with foreign environment as much as possible and is handled with disposable gloves to avoid contamination. One bag contains fecal sample composed primarily of pulverized bamboo stems whereas the other on primarily composed of pulverized bamboo leaves. After sampling, the samples are brought to Malaysia Genome Institute (MGI) to be frozen in the -20°C freezer. The sample of the female giant panda has been collected, subjected to phenol-chloroform DNA extraction, and sequenced at MGI in an earlier postgraduate study.

3.3.3 Sample Handling

Using a sterile spatula, about one third of the 50ml conical centrifuge tube is filled with the fecal sample; primarily composed of pulverized bamboo stems, which has been thawed earlier for 10 minutes. Next, the balance two thirds of the conical centrifuge tube is filled with double distilled water, capped and shaken well to form a fecal suspension. Next, 1000µl is drawn from the liquid part of the suspension and decanted into 2ml microcentrifuge tube and is centrifuged at 14 000 rpm for 10 minutes to form a fecal pellet, with the supernatant removed. This step is repeated for another 12 times using the same microcentrifuge tube to obtain a large cumulative pellet. After the final supernatant removal, 800µl of STE buffer is added and the tube is vortexed until the pellet is completely homogenized in the buffer.

3.3.4. Cell Lysis

5µl of lysozyme is added to the homogenized mixture and incubated at 45°C for 40 minutes. 5µl of 20µg/ml of Proteinase K is then added and is left to incubate at 55°C for another 40 minutes. Next, 50µl of 10% SLS buffer will be added and is incubated at 55°C for another 40 minutes.

3.3.5. Phenol-Chloroform DNA Extraction

After cell lysis, half of the lysate volume, which is about 500µl, is decanted into another new microcentrifuge tube. 425µl of phenol is then added, followed by 425µl of chloroform-isoamyl alcohol of ratio of 24:1; both into each of the two tubes. The tubes are then inverted 20 times to thoroughly mix the liquids and form emulsions. The resulting emulsions are then centrifuged at 14 000 rpm for 10 minutes. Next, carefully without touching the interphase and the organic layer the upper aqueous layer of the suspensions are pipetted into 2 new microcentrifuge tubes using a 1000µl micropipette tip, followed by a 10µl micropipette tip to recover the remaining aqueous layer; each containing 400µl of supernatant. The organic phase and the interphase are discarded.

3.3.6. DNA Precipitation

1200µl of cold, absolute ethanol will be added to each tube containing the supernatant, followed by 160µl of 3M sodium acetate to each tube. The tubes are then inverted gently for 20 times to mix well and stored in the -80°C freezer for 60 minutes. Next, the mixtures will be centrifuged at 14 000 rpm for 10 minutes. The

ethanol-sodium acetate mixture is then decanted, leaving behind a pellet in each tube. The pellets will be washed with 500µl of 70% ethanol and left to stand 30 minutes at room temperature. Centrifugation is then done at 14 000 rpm for 10 minutes. The ethanol will then be decanted. This step is repeated for another two times. All remaining ethanol will be pipetted out and air dried for 30 minutes. Finally, the remaining pellets is then rehydrated in 150µl TE buffer each and kept at -20°C overnight.

3.3.7 DNA Quantification

DNA can be quantified by analyzing its purity and concentration; all by taking absorbance readings at 260nm, 280nm, and 230nm. DNA is said to free of phenol and protein contamination if the reading of the 260/280 ratio falls within 1.8 to 2.0; whereas DNA is said to be free of salt contamination if the reading of the 260/230 ratio falls within 2.0 to 2.4 (UCSF Helen Diller Family Comprehensive Cancer Centre, 2015). If the DNA sample is pure and free from phenol, salt, or protein contamination, then the concentration can be determined by calculating the amount of UV radiation absorbed by the bases present in the DNA at 260nm. By performing DNA quantification, interference during metagenomic library preparation and sequencing can be avoided.

In order to determine DNA concentration and purity, the NanoDrop® ND-1000 spectrophotometer was used to measure the absorbance at 260nm, 280nm, and 230nm. First, the program is initialized on the computer; setting it to the nucleic acid mode. To initialize and clean the instrument, the lower pedestal is first loaded with nuclease free water twice, using 2µl each time. Then, the spectrophotometer is zeroed using TE buffer as the blank solution. Next, the samples are loaded onto the

very centre of the lower pedestal and the readings of the concentration, 260/280 and 260/230 ratios are displayed, recorded and showed in Table 4.1. After all samples are loaded, the lower pedestal is cleaned again with nuclease free water before shutting down the program. Delicate task wipes are used whenever to clean; and/or before loading samples onto the lower pedestal to avoid damage and inaccurate readings.

3.3.8. Agarose Gel Electrophoresis

60ml of 1% electrophoresis agarose gel is prepared by weighing out 0.6g of agarose powder and adding to 60ml of 1x TAE buffer with 5% SYBR® Green I and micro waved for 2 minutes until completely dissolution. The mixture is left to cool down to lukewarm temperature for 5 minutes before being poured into the gel cast together with comb, gently to avoid formation of air bubbles; to be set for another 5 minutes at room temperature. A pipette tip is used to remove any bubbles formed. Finally, the comb is removed, leaving behind wells in the solidified gel. The solidified gel is then placed into an electrophoresis tank containing 1x TAE running buffer. After that, DNA samples were loaded into the well by resuspending 1µl of 6x loading dye together with 5µl of each sample to form 6µl separate mixtures on a grafting tape. These sample mixtures are then loaded into the wells using 10µl micropipette. Then, 5µl of 1000bp DNA ladder is resuspended together with 1µl of 6x loading dye and pipetted into another well to serve as size reference. The agarose gel electrophoresis is then run at 90V for 90 minutes. Finally, the gel was removed and placed in a UV transilluminator, with the photograph of the gel taken and documented, as in **Figure 4.3.**

3.3.9 Metagenomic Library Construction

Two samples labeled Panda A and Panda B, with 143µl starting volume each is used to prepare the metagenomic library, whereas samples labeled Panda C and Panda D is kept as reserves. Panda A and Panda B has respective starting DNA yield of 31.06µg and 36.09µg. The kit used to prepare the metagenomic library is the KAPA Library Quantification Kit for Illumina™ platforms and the Illumina™ TruSeq® PCR-Free LT Kit. The process is divided into 6 phases; namely DNA fragmentation, end repair and size selection, adenylate 3' ends, adapter ligation, library validation, as well as library normalization and pooling.

(a) DNA Fragmentation

DNA fragmentation is where the metagenomic DNA is optimally fragmented using the Covaris shearing method to generate dsDNA with 3' or 5' overhangs; ultimately to obtain final libraries of the average insert sizes of 550bp. 2µg from each sample is withdrawn and is normalized to final volume of 55µl at 20ng/µl using "Resuspension Buffer" into each well of the new 0.3ml PCR plate labeled with DNA barcode. 52.5µl of sample from each well is then pipetted into new Covaris tubes placed in Covaris Fragmentation Plate (CFP) and centrifuged at 600x g for 5 seconds. Next, the DNA is sonicated using the focused ultrasonicator at temperature of 20°C for 45 seconds. The CFP is centrifuged again at 600 xg for 5 seconds and 50µl is pipetted to the corresponding wells of new 0.3ml PCR plate labeled with Clean-up Sheared DNA Plate (CSP) using single channel pipette. The balance 2.5µl of each sample is "bioanalyzed" using the bioanalyzer to measure fragment size.

Once the desired fragment size of average 550bp is achieved, this is immediately followed up by the clean up steps, where 80µl of well-mixed Sample Purification Beads (SPB) is added to each well containing the 50µl of fragmented metagenomic DNA and incubated at room temperature for 5 minutes. The CSP is then placed on a magnetic stand for 8 minutes. After that, 125µl of the supernatant is removed from each well, followed by the addition of 200µl of 80% EtOH to each well without disturbing the SPB pellets and incubation at 30 seconds at room temperature. The wash step is repeated another 2 times. The samples are then air-dried at room temperature for 5 minutes. Any remaining EtOH is removed using 10µl pipette. Next, resuspension buffer is added and the CSP is removed from the magnetic stand. The SPB pellet in each well is resuspended by repeatedly dispensing the resuspension buffer over the bead pellets until it is immersed in the solution; then the entire volumes are gently pipetted up and down 10 times to mix thoroughly, and incubated at room temperature for 2 minutes. It is the placed on the magnetic stand for 5 minutes. After that, 50µl of the clear supernatant in each well of the CSP is pipetted into the new corresponding wells of 0.3ml plate labeled with Insert Modification Plate (IMP) barcode, without disturbing the SPB pellets. The remaining 2.5µl is used for bioanalysis and Qubit analysis to analyze fragment size and DNA quantity respectively. If desired DNA quantity and fragment size is achieved, it is immediately followed by end repairing and size selection phase.

(b) End Repair and Size Selection

Performing end repair and size selection is essential to convert the overhangs resulting from the previous DNA fragmentation into blunt ends using and End Repair

Mix 2, where the 3' to 5' exonuclease activity removes the 3' overhangs and the 5' to 3' polymerase activity fills in the 5' overhangs. After that, the appropriate library sizes are sorted out using different ratios of SPB. First, 10µl of resuspension buffer is added each well of the IMP containing the 50µl of fragmented DNA, followed by 40µl of thawed End Repair Mix 2 and the whole volumes are gently pipetted up and down 10 times to mix thoroughly. The IMP is the sealed with Microseal 'B' adhesive seal and placed in a pre-programmed thermal cycler set at the ERP program. The thermal cycler is pre-programmed by first choosing the preheat lid option and set to 100°C,30°C for 1 hour, then is held at 4°C. Once the program reached 4°C, the IMP is removed from the thermal cycler.

The first step in removal of large DNA fragments is the preparation of diluted SPB mixture adding 184µl of SPB to 184µl of PCR grade water and vortexed for 5 seconds to disperse the beads thoroughly. 160µl of the diluted SPB mixture is the added to the wells of the IMP containing 100µl of the end repaired samples each. The entire volumes are gently pipetted up and down 10 times to mix thoroughly, and the IMP is incubated for 5 minutes at room temperature. The plate is then placed on a magnetic stand for 5 minutes. Next, using a 200µl single channel micropipette or multichannel set to 125µl; 125µl of the supernatant, containing the DNA of interest is transferred twice from each well of the IMP to the corresponding wells of the new 0.3ml PCR plate labeled with the Clean-up End Repair Plate (CEP) barcode, without disturbing the beads. The IMP plate containing the beads and any remaining diluted bead mixture is discarded.

To remove the small DNA fragments, 30µl of undiluted SPB is added to each well containing the 250µl of the supernatant with DNA of interest; gently the entire volume is pipetted up and down for 10 times to mix thoroughly, and the CEP is

incubated for 5 minutes. Next, the CEP is placed on a magnetic stand for 5 minutes. Then, using a 200µl single channel or multichannel pipette set to 138µl, 138µl of the supernatant is removed and discarded from each of CEP wells, without disturbing the beads. The previous step is repeated to remove and discard the balance 276µl of supernatant from each of the wells. After that, 200µl of 80% EtOH is added to each well of the CEP without disturbing the beads. The CEP is then left to incubate for 30 seconds, then all the supernatants in each well is removed without disturbing the beads. This wash step is repeated for another two times. The samples are then left to air dry for 5 minutes and any remaining EtOH is removed using a 10µl pipette. 17.5µl of resuspension buffer is then added to each well of the CEP. The plate is then removed from the magnetic stand. The bead pellets in each well is then resuspended by repeatedly dispensing the resuspension buffer over them until complete immersion in the solution, then the entire volume is gently pipetted up and down for 10 times to mix thoroughly. The CEP is then incubated for 2 minutes at 25°C. Then, the plate is placed back on the magnetic stand at room temperature for 5 minutes. Finally, 15µl of the clear supernatant from each well is transferred to the corresponding well of the new 0.3ml PCR plate labeled with Adapter Ligation Plate (ALP) barcode.

(c) Addition of Adenylate 3' Ends

The addition of adenylate 3' ends is where an ‘A’ single nucleotide is added to the 3' ends of the blunt fragments to prevent them from ligating to each other during adapter ligation reaction. A corresponding single ‘T’ nucleotide on the 3' end

of the adapter provides a complementary overhang for ligating the adapter to the fragment. Overall, this strategy is to ensure low chimera formation rate.

First, 2.5µl of resuspension buffer is added to each well, followed by 12.5µl of A-Tailing Mix (ATL). A 200µl pipette is then set to 20µl, and the entire volume is gently pipetted up and down 10 times to thoroughly mix, and the ALP is sealed with a Microseal 'B' adhesive seal. The ALP is then placed in a thermal cycler pre-programmed according to the following settings; pre-heat lid option set to 100°C, 37°C for 60 minutes, 70°C for 5 minutes, 4°C for 5 minutes, held at 4°C, and saved as ATAIL70. The lid is closed and he ATAIL70 program is selected and run. The ALP is then removed from the thermal cycler when it reaches 4°C for 5 minutes and is immediately followed by ligation of adapters.

(d) Ligation of Adapters

Ligation of adapters is the process where multiple indexing adapters are ligated to the end of DNA fragments, preparing them for hybridization on the flow cell during sequencing. The first step is where 2.5µl thawed DNA Adapter Index is added to each well of the DNA Adapter Plate (DAP). A 200µl pipette is then set to 35µl, and the entire volume is gently pipetted up and down 10 times to mix thoroughly. Using a 8-tip multichannel pipette, 2.5µl of thawed DNA Adapter from the DAP is transferred to each well of the ALP. A 200µl is set to 35µl, and the entire volume is then gently pipetted up and down 10 times to mix thoroughly, sealed with Microseal 'B' adhesive seal, and the ALP is centrifuged at 280x g for 1 minute. Next the ALP is placed in pre-programmed thermal cycler by setting it 100°C using the

pre-heat lid option, 30°C for 20 minutes, and hold at 4°C. ALP is the removed from the thermal cycler once the temperature reaches 4°C.

42.5µl of SPB is then added to each well. A 200µl pipette is set to 60µl, and the entire volume is then gently pipetted up and down 10 times to mix thoroughly. The plate is then incubated for 5 minutes at room temperature, and let on the magnetic stand or the subsequent 5 minutes. Then, 80µl of the supernatant from each well is then removed and discarded. With the plate on the magnetic stand, 200µl of freshly prepared 80% EtOH is then added to each well, incubated for 30 seconds, then all the supernatant from each well is removed and discarded. The wash step is repeated for another two times. Any remaining EtOH is then removed using a 10µl pipette, and the samples are air-dried for 5 minutes at room temperature. 52.5µl of resuspension buffer is then added to each well and the plate is removed from the magnetic stand. The beads in each well are resuspended by repeatedly dispensing the resuspension buffer over the bead pellet until it is immersed in the solution. The entire volume is gently pipetted up and down 10 times to mix thoroughly. The plate is then incubated at room temperature for 2 minutes, and subsequently placed on the magnetic stand for 5 minutes. Then, 50µl of the clear supernatant from each well of the ALP is then transferred to a new 0.3ml PCR plate labeled with the Clean-up ALP Plate (CAP) barcode. The remaining 2.5µl is used for bioanalysis to analyze if desired DNA and fragment size is achieved. If desired fragment size is achieved, it is moved on to the wash step.

50µl of mixed SPB is first added to each well for a second cleanup. A 200µl pipette is set to 75µl, and the entire volume is gently pipetted up and down 10 times to mix thoroughly. The plate is then incubated at room temperature for 5 minutes and placed on the magnetic stand for another 5 minutes. Then, 95µl of the supernatant is

the removed and discarded from each well. 200µl of freshly prepared 80% EtOH is then added to each well. After incubating the plate for 30 seconds, all the supernatant from the wells is removed and discarded. The wash step is repeated for another 2 times. Any remaining EtOH is then removed using a 10µl pipette. The samples are then left to air dry at room temperature for 5 minutes. 22.5µl of resuspension buffer is then added to each well and the plate is removed from the magnetic stand. The beads in each well are resuspended by repeatedly by dispensing the resuspension buffer over the bead pellet until fully immersed in the solution. The entire volume is gently pipetted up and down 10 times to mix thoroughly. The plate is then incubated for 2 minutes at room temperature, and placed on the magnetic stand for another 5 minutes. Finally, 20µl of the clear supernatant from each well of the CAP is transferred the corresponding well of the new 0.3ml PCR plate labeled Target Sample Plate (TSPI) barcode. Qubit analysis is then performed to keep track of DNA quantity.

(e) Validation

Validation is the process where the metagenomic DNA libraries prepared is quantified using "quantity polymerase chain reaction" (qPCR) technology. The metagenomic DNA libraries prepared in this research were quantified using the KAPA Library Quantification Kit for Illumina® platforms.

(f) Normalization and Pooling of Libraries

Normalizing and pooling of libraries is the final phase of the metagenomic library preparation, where DNA templates will be prepared to be applied in cluster

generation during sequencing. DNA libraries which are indexed are normalized to 2 nM in the "Diluted Cluster Template" (DCT) plate and then they will be pooled in equal volumes in the "Pooled DCT Plate" (PDP).

To make the DCT, firstly 5µl of sample library from each well of TSPI plate is transferred to the corresponding well of the new 0.3ml PCR plate labeled with the DCT barcode. The concentration of the sample library in each well are then normalized to 2nM using Tris-Cl 10mM, pH 8.5 with 0.1% Tween 20; then the entire normalized library sample volume is gently pipetted up and down 10 times to mix thoroughly.

Since the libraries are indexed, 5µl each of the normalized samples to be pooled from the DCT plate is transferred to one of the new well of the new PCR plate labeled with the PDP barcode. The total volume in each PDP plate well should be 5 times the number of the combined sample libraries and will be 10-120µl (2-24 libraries). Then, the entire volume is gently pipetted up and down 10 times to mix thoroughly. It is then proceeded to cluster generation, which is a phase of the sequencing process.

3.3.10 Sequencing

The pooled metagenomic DNA library are then sequenced using the Illumina™ MiSeq® system, where brand new reagents enable up to 15Gb of output with 2x300bp read lengths and 25M sequencing reads.

3.3.11 Data Analysis

The data will then be retrieved, uploaded and analyzed using the Metagenomic Rapid Annotations using Subsystems Technology (MG-RAST) online metagenomic analysis server. Before processing the raw data, quality control (QC) will be done by using the FASTQ sequence format to filter, analyze artifacts, and provide quality information by carrying out tasks such as demultiplexing, quality filtering, length filtering, dereplication, and model organism sequence removal. The sequences will then be analyzed based on organism classification by generating rarefaction curves, phylogenetic trees and bacteria classification tables.

CHAPTER 4

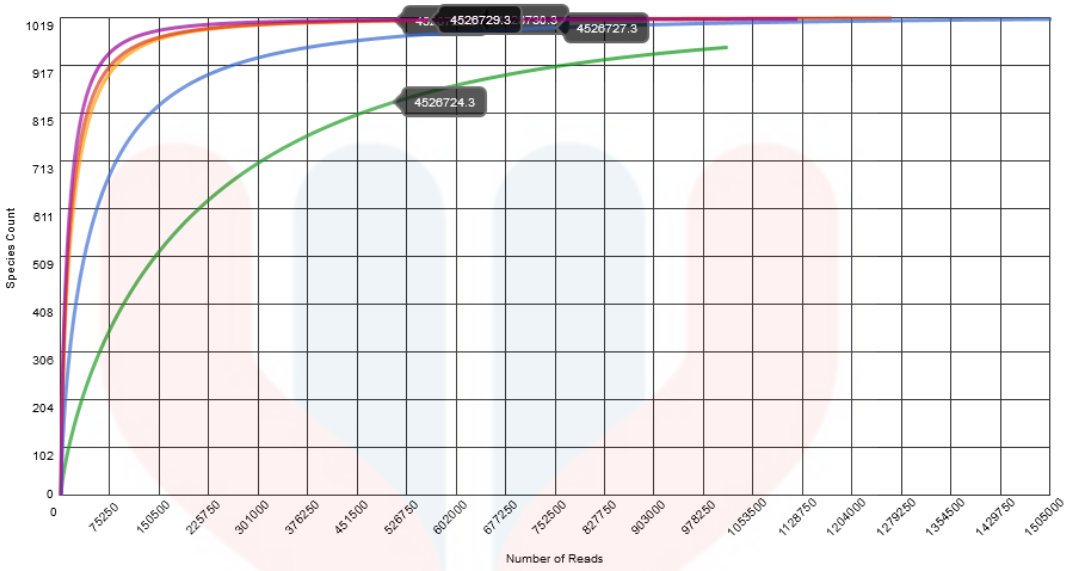
RESULTS AND DISCUSSION

4.1 Reference Metagenome Data Visualization and Analysis

As mentioned in Section 3.3.1 above, the five reference metagenomic datasets are visualized in three forms; namely the rarefaction plot, organism table, and organism tree.

4.1.1. Rarefaction Plot

A rarefaction plot of species count against number of reads is generated as shown in Figure 4.1 below, to analyze the efficiency of the sampling attempts of the five mammalian metagenomes in attaining whole microbial species coverage.



rarefaction curve	metagenome	alpha diversity
	4526727.3	81.04
	4526730.3	147.85
	4526723.3	140.64
	4526724.3	71.38
	4526729.3	305.15

FIGURE 4.1 Comparison of species count against number of reads between metagenomes of horse (4526729.3), whitetail deer (4526730.3), humpback whale (4526723.3), coyote (4526727.3), and the bottlenose dolphin (4526724.3) and their respective alpha diversity. (Source: MG-RAST Version 3.0, 2016)

Based on the figure above, the rarefaction curves of horse, whitetail deer, and humpback whale metagenome shows a similar trend, where they have a spike in species count of roughly 920 species of in the first 75250 reads, followed by a gradual rise in species count of lower rate, which finally levels off at 1019 species for 526750 reads. Among the 5 mammals, the horse scored the highest alpha diversity of 301.15. Thus, an inference can be made where sufficient sampling has been done for the coverage of the majority of the microbial species of the horse, whitetail deer, and humpback whale metagenome; which means that a few improvements need to be done in the sampling methods and attempts in order to

attain additional coverage of the much rare species present in the metagenome of said mammals.

As for the metagenome of the coyote however, the rarefaction curve shows a lower increment rate compared to that of the previous three organisms, scoring the second lowest alpha diversity of 81.04; nevertheless also leveling off at 1019, but for a higher sequence reads of 1 279 250 reads. Therefore, it can be inferred that there is much microbial species coverage to be done by increasing the efficiency of the metagenome sampling technique of the coyote. The metagenome of the bottlenose dolphin scores the lowest alpha diversity of 71.30, with a rarefaction curve of low but relatively steady increment rate, scoring a species count of 968 species at maximum sequence reads of 1 015 875 reads, although there is no leveling of the curve yet. This is probably because the metagenome has a relatively high microbial species, but insufficient sampling with low sequence reads has been done; resulting in relatively low coverage of the wide species diversity. Therefore, the sampling protocol of the bottlenose dolphin has also much to improve in efficiency.

Therefore, the average estimated minimum number of sequence reads needed to attain full microbial species coverage of the giant panda metagenome is calculated by adding each sequence read value of the 5 metagenomes and dividing the total by five, which will have a final value of 775 075 reads, which is close to expected 1 000 000 reads.

DNA sequence information size is able to be determined by representing in digital form in a computer. In general, DNA base pairs have for different possible combinations; namely “A-T”, “T-A”, “C-G”, “G-C”. These 4 different combinations are to be represented in a binary format, which includes the number “1” and “0”. The

two “bits” will form 4 corresponding different combinations; namely “00”, “01”, “10”, “11”.It takes 8 bits, or 4 DNA base pairs to form a single byte of digital information (Grigoryev, 2012). Therefore, the following formula is used to convert DNA sequences into digital storage size:

$$\text{DNA digital storage size} = (\text{Number of reads} \times \text{Read length}) \div 4 \quad (4.1)$$

Therefore, the digital storage size of the previous final average read value of 775 075 reads is calculated by multiplying the figure with the read length (in this case, 600bp is assigned due to the application of the Illumina™ MiSeq® system), and then dividing the product by 4, resulting in a value of 116Mb of data, which is close to the expected size of 150Mb (1 000 000 reads).

4.1.2. Organism Tree

An organism tree is generated with class being the maximum level of specification, and each class of similar colour is grouped together according to their common phyla. The leaf weights of each class are then displayed as stacked bars.

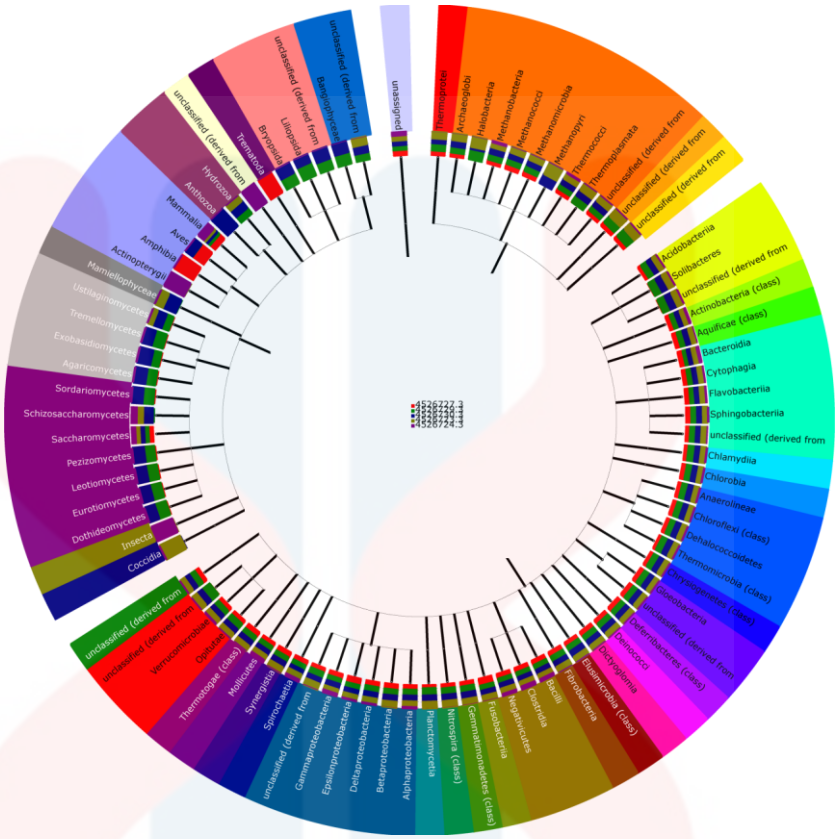


FIGURE 4.2 Organism tree of the microbial classes present and their respective relative abundance in the metagenomes of horse (4526729.3), whitetail deer (4526730.3), humpback whale (4526723.3), coyote (4526727.3), and the bottlenose dolphin (4526724.3). (Source: MG-RAST Version 3.0, 2016)

Based on the organism tree above, a total of 41 phyla are displayed, where each different colour of a leaf weight represents different phyla. Out of the 41 phyla, only 12 phyla are detectable in all the five different mammalian metagenomes, including the prominent phyla Firmicutes and Proteobacteria. However, phylum Euryarchaeota scores the highest number of different classes among all the phyla, which are 9 different classes of microbes.

4.1.3. Organism Table

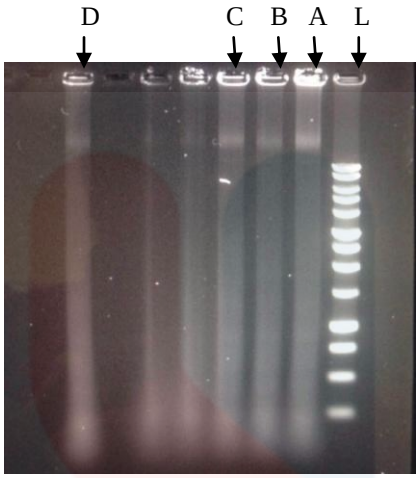
An organism table in Appendix B is also generated to study the abundance of microbes according to classification of phyla in the metagenomes of the five

mammalian species. As mentioned in Section 4.1.2, only 12 out of the 41 phyla detected are found in the metagenome of the entire five mammals; which includes Proteobacteria, Firmicutes, and Euryarchaeota, where phyla Firmicutes scores the highest abundance of 363 845 sequences feature with a hit.

4.2 Agarose Gel Electrophoresis for DNA Extraction Quality Analysis

Quality, quantity, and purity analysis via AGE is an essential procedure, where extraction quality and intensity of the metagenomic DNA can be estimated through gel electrophoresis by comparing the DNA bands with that of the 1000bp DNA ladder as a reference. Thus, the results obtained from AGE can be used to compare to that of the following spectrophotometry analysis by serving as a reference to determine if the absorbance results is under-estimated or over-estimated by the spectrophotometer; which may happen due to phenol, protein, or salt contamination in sample, slight difference between blank solution pH and sample pH or a dirty loading pedestal. Besides that, AGE also enables to detect DNA degradation or contamination by the presence of smearing, which validates the following spectrophotometry results.

In this procedure, all 4 samples of metagenomic DNA is loaded to test for extraction quality, as shown in Figure 4.1 below.



Legend

L: 1000bp DNA ladder lane
A: Panda A metagenomic DNA sample lane
B: Panda B metagenomic DNA sample lane
C: Panda C metagenomic DNA sample lane
D: Panda D metagenomic DNA sample lane

FIGURE 4.3 Agarose gel electrophoresis of 4 metagenomic DNA samples.

A total of 5µl of metagenomic DNA sample with 1µl of 6x loading dye were added carefully into 4 wells and electrophoresed on a 1% agarose gel at 90 volts for 90 minutes. From far right, lane L contains 1000bp DNA ladder as reference. Lane A: Panda A; Lane B: Panda B; Lane C: Panda C; Lane D: Panda D.

Based on the figure above, it can be observed that each lane loaded with samples have similar band patterns; where a thick but faint band is visible on top of each lane, all above the highest band of the DNA ladder. These bands are where the metagenomic DNA lie. After these bands, the DNA start to form smears at the bottom of the gel. This is probably cause by the presence of degraded DNA and carrier RNA due to nuclease contamination; as well as salt contamination (Fermentas Life Sciences, n.d.).

4.3 DNA Purity and Quantification by Spectrophotometry

As mentioned in section 3.3.6, DNA quality and quantity determination is a process whereby DNA samples is ensured to be free of contaminants such as salt, phenol, protein to avoid interference during downstream process, such as that of the metagenomic library preparation, cluster generation and ultimately, ensuring a smooth sequencing process. In addition, DNA yield of every sample is also plays a vital part in this study; as high, relatively equal quantities provides sufficient starting DNA to work on and enables less complicated protocol execution during the normalization and pooling steps of the metagenomic library preparation. In short, all the values of these parameters are determined by performing the NanoDrop® ND-1000 spectrophotometry analysis. All metagenomic DNA sample results are as shown in Table 4.1 below.

TABLE 4.1 NanoDrop® ND-1000 spectrophotometer results of four DNA sample purity, concentration, and yield at 230nm, 260nm, and 280nm

Sample	DNA purity (260nm/280nm)	DNA purity (260nm/230nm)	Concentration of DNA (ng/µl)	Volume (µl)	DNA yield (µg)
Panda A	2.06	3.08	217.2	143	31.06
Panda B	2.09	3.19	252.4	143	36.09
Panda C	2.09	3.14	207.4	143	29.66
Panda D	2.06	3.13	249.9	133	33.24

The following formula was used to determine the DNA yield of each sample.

DNA yield in µg = DNA concentration in ng/µl x volume in µl x 10^{-9} (4.2)

For example, for sample Panda D, the DNA yield is obtained by multiplying the value of DNA concentration, 249.9ng/µl; with that of volume, which is 133µl, and with 10^{-9} to obtain the final value in micrograms (µg).

As mentioned in Section 3.3.6, a good DNA extraction will result in pure metagenomic DNA with 260/280nm and 260/230nm absorbance ratio readings in the range of 1.8 to 2.0. However, based on 260/280nm readings in Table 4.1, all of the readings of the four samples; Panda A, Panda B, Panda C, and Panda D are close to 2.1, which is slightly over the recommended range of 1.8 to 2.0. This might be due to slight addition of carrier RNA to the purification procedure, which increases the nucleic acid yield (Qiagen, 2013); and all types of nucleic acids have peak absorbance at 260nm (UCSF Helen Diller Family Comprehensive Cancer Centre, 2015), hence the stated readings.

As for the 260/230nm readings, all the samples gave the average reading of about 3.14, which is about 0.7 units higher than the recommended range of 2.0 to 2.4. This is also caused by the presence of RNA contamination and probably because of the presence of relatively low quantity salt, phenol, and carbohydrate from the phenol-chloroform DNA extraction process. These substances have peak absorbance at 230nm (UCSF Helen Diller Family Comprehensive Cancer Centre, 2015), hence the tabulated readings.

In short, the MG-RAST analysis done on the metagenomes of the coyote, horse, whitetail deer, humpback whale, and bottlenose dolphin indicates that the gut

microbiota diversity of the 5 mammals is similar to that of the giant panda, the detection of bacteria from phyla Firmicutes and Proteobacteria in particular; which is probably due to their mammalian lineage. Based on the rarefaction analysis of the five model metagenomes of the five mammals, a mean of 775 075 sequence reads, or approximately 1 000 000 sequence reads is needed to achieve maximum bacterial species coverage in any metagenome of the 5 model, reference mammals. Thus, it can be deduced that 1 000 000 sequence reads, or approximately 150 Mb of metagenomic DNA data size also has to be retrieved in order to achieve maximum bacterial species coverage in the giant panda metagenome analysis. Thus; sampling, DNA extraction, and library preparation of giant panda metagenome have to be optimized to the point where at least 150Mb of metagenomic DNA data can be retrieved.

CHAPTER 5

CONCLUSION AND RECCOMENDATION

In conclusion, based on the analysis of the metagenomes of horse, coyote, whitetail deer, bottlenose dolphin, and humpback whale as benchmark, it can be hypothesized that approximately 1 000 000 sequence reads, or 150Mb of metagenomic DNA data shall be acquired from the sampling of giant panda fecal matter in order to achieve full microbial species coverage, which is where the giant panda metagenomic DNA extraction methods shall also improve significantly. Based on Section 4.2, all for DNA samples loaded for AGE displays bands of metagenomic DNA with smears or degraded DNA below, signifying major degradation of double stranded DNA during the extraction process; whereas in Section 4.3, all the four samples have a 260/280 readings of approximately 2.1, 260/230 readings of approximately 3.14, and roughly 32.5µg of DNA yield per sample.

In order to maximize acquisition of sequence reads with wide species coverage, relatively large-scale sampling shall be done, and microbial pellet sedimentation process should be done as many times as possible to recover rare microbial species. Samples shall also be handled at low temperatures at all times and shall not be exposed to frequent temperature change stress, especially during any centrifugal steps. This can be prevented by using a temperature-manipulative centrifuge. Clean-up steps shall also be repeated as many times as necessary to ensure excellent quality and purity of extracted DNA, as well as to prevent DNA degradation.

REFERENCES

Ahmad, M., Taylor, C. R., Pink, D., Burton, K., Eastwood, D., Bending, G. D., & Bugg, T. D. (2010). Development of novel assays for lignin degradation: comparative analysis of bacterial and fungal lignin degraders. *Molecular Biosystems*, *6*(5), 815-821.

Béja, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P., ... & Spudich, E. N. (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, *289*(5486), 1902-1906.

Bitesize Bio, (2012). How Much Information is Stored in the Human Genome. Available from <http://bitesizebio.com/8378/how-much-information-is-stored-in-the-human-genome/> [Accessed on 23 December 2016].

Brambell, M. R. (1976). The giant panda (*Ailuropoda melanoleuca*). *The Transactions of the Zoological Society of London*, *33*(2), 85-92.

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

Burke, C., Kjelleberg, S., & Thomas, T. (2009). Selective extraction of bacterial DNA from the surfaces of macroalgae. *Applied and Environmental Microbiology*, *75*(1), 252-256.

Chengdu Research Base of Giant Panda Breeding, (2007). Prevention and Treatment of Panda Disease. Available from <http://www.panda.org.cn/english/research/disease/2013-09-12/2449.html> [Accessed on 15 June 2016].

Chorn, J., & Hoffmann, R. S. (1978). *Ailuropoda melanoleuca*. Soc..

Collins, M. D., Lawson, P. A., Willems, A., Cordoba, J. J., Fernandez-Garayzabal, J., Garcia, P., ... & Farrow, J. A. E. (1994). The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *International Journal of Systematic and Evolutionary Microbiology*, *44*(4), 812-826.

Davis, D. D. (1964). *The giant panda: a morphological study of evolutionary mechanisms* (Vol. 3). Chicago Natural History Museum.

Delsuc, F., Metcalf, J. L., Wegener Parfrey, L., Song, S. J., González, A., & Knight, R. (2014). Convergence of gut microbiomes in myrmecophagous mammals. *Molecular Ecology*, *23*(6), 1301-1317.

Dierenfeld, E.S., Hintz, H.F., Robertson, J.B., Van Soest, P.J., Oftedal, O.T., (1982) Utilization of bamboo by the giant panda. *The Journal of Nutrition*, *112*:636–641.

Endo, H., Hayashi, Y., Yamagiwa, D., Kurohmaru, M., Koie, H., Yamaya, Y., & Kimura, J. (1999). CT examination of the manipulation system in the giant panda (*Ailuropoda melanoleuca*). *Journal of Anatomy*, *195*(2), 295-300.

Fang, W., Fang, Z., Zhou, P., Chang, F., Hong, Y., Zhang, X., ... & Xiao, Y. (2012). Evidence for lignin oxidation by the giant panda fecal microbiome. *PloS One*, *7*(11), e50312.

Fang, S. G., Wan, Q. H., & Fujihara, T. (2003). Loss of genetic variation in giant panda due to limited population and habitat fragmentation. *Journal of Applied Animal Research*, *24*(2), 137-144.

Fermentas Life Sciences, n.d. Troubleshooting Guide for DNA Electrophoresis. Available from < file:///C:/Users/User/Downloads/DNA_Troubleshooting.pdf> [Accessed on 16 November 2016].

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., & Meyer, F. (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols*, *2010*(1), pdb-prot5368.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669-685.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, *5*(10), 245-249.

Hendey, Q. B. (1972). *A Pliocene ursid from South Africa*. South African Museum.

Jin, C., Ciochon, R. L., Dong, W., Hunt, R. M., Liu, J., Jaeger, M., & Zhu, Q. (2007). The first skull of the earliest giant panda. *Proceedings of the National Academy of Sciences*, *104*(26), 10932-10937.

Jin, K., Xue, C., Wu, X., Qian, J., Zhu, Y., Yang, Z., ... & Zhong, Y. (2011). Why does the giant panda eat bamboo? A comparative analysis of appetite-reward-related genes among mammals. *PLoS One*, *6*(7), e22602.

Kirby, K. S. (1956). A new method for the isolation of ribonucleic acids from mammalian tissues. *Biochemical Journal*, *64*(3), 405.

Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., ... & Lykidis, A. (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Research*, *36* (suppl 1), D534-D538.

Morris, R., & D. Morris. (1966). Men and pandas. Hutchinson, London, 223 pp.

Nicol, G. W., & Schleper, C. (2006). Ammonia-oxidising Crenarchaeota: important players in the nitrogen cycle?. *Trends in Microbiology*, *14*(5), 207-212.

Pastor, J. F., Barbosa, M., & De Paz, F. J. (2008). Morphological study of the lingual papillae of the giant panda (*Ailuropoda melanoleuca*) by scanning electron microscopy. *Journal of Anatomy*, *212*(2), 99-105.

Pei, W. C. (1974). A brief evolutionary history of the giant panda. *Acta Zool. Sinica*, *20*, 188-190.

Qiagen, (2013). FAQs. Available from <https://www.qiagen.com/cn/resources/faq?id=8382dddc-2441-4687-9c61-b4e44a136ef4&lang=en> [Accessed on 7 November 2016].

Qiu, Z., & Qi, G. (1989). Ailuropod found from the Late Miocene deposits in Lufeng, Yunnan. *Vertebrata PalAsiatica*, *27*, 153-169.

Rappe, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Reviews in Microbiology*, *57*(1), 369-394.

Riesenfeld, C. S., Goodman, R. M., & Handelsman, J. (2004). Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental Microbiology*, *6*(9), 981-989.

Schaller, G. B. (1985). *Giant pandas of Wolong*. University of Chicago press.

Schaller, G. B., Qitao, T., Johnson, K. G., Xiaoming, W., Heming, S., & Jinchu, H. (1989). The feeding ecology of giant pandas and Asiatic black bears in the Tangjiahe Reserve, China. In *Carnivore Behavior, Ecology, and Evolution* (pp. 212-241). Springer US.

Simon, C., & Daniel, R. (2011). Metagenomic analyses: past and future trends. *Applied and Environmental Microbiology*, *77*(4), 1153-1161.

Simu, K., & Hagström, Å. (2004). Oligotrophic bacterioplankton with a novel single-cell life strategy. *Applied and Environmental Microbiology*, *70*(4), 2445-2451.

Smithsonian National Zoological Park, n.d. Giant Panda. Available from <https://nationalzoo.si.edu/animals/giantpandas/pandafacts/> [Accessed on 15 June 2016].

Sun, S., Chen, J., Li, W., Altinatas, I., Lin, A., Peltier, S., ... & Wooley, J. (2010). Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Research*, gkq1102.

Susan, L., (2005). Degradation of polymers, p 101–131. In Dürre P (ed), Handbook on clostridia, 1st ed. CRC Press, Boca Raton, FL.

Thomas, T., Rusch, D., DeMaere, M. Z., Yung, P. Y., Lewis, M., Halpern, A., ... & Kjelleberg, S. (2010). Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *The ISME Journal*, *4*(12), 1557-1567.

Tsiang-ke, W., (1974). On the taxonomic status of species, geological distribution and evolutionary history of ailuropoda. *Acta Zoologica Sinica*, *2*, 10.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., ... & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37-43.

UCSF Helen Diller Family Comprehensive Cancer Centre, (2015). Nucleic Acid Quality/ Concentration. Available from < http://cancer.ucsf.edu/research/cores/genome/services/genome-analysis-service-analyze> [Accessed on 7 November 2016].

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.

Wang Sung & Lu Chang-kun., (1973). Giant pandas in the wild . . . . Nat. Hist. 82:70-71.

Wei, F., Hu, Y., Yan, L., Nie, Y., Wu, Q., & Zhang, Z. (2014). Giant pandas are not an evolutionary cul-de-sac: evidence from multidisciplinary research. *Molecular Biology and Evolution*, msu278.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews*, *51*(2), 221.

Xu, W., Ouyang, Z., Viña, A., Zheng, H., Liu, J., & Xiao, Y. (2006). Designing a conservation plan for protecting the habitat for giant pandas in the Qionglai mountain range, China. *Diversity and Distributions*, *12*(5), 610-619.

Xue, Z., Zhang, W., Wang, L., Hou, R., Zhang, M., Fei, L., ... & Jiang, C. (2015). The bamboo-eating giant panda harbors a carnivore-like gut microbiota, with excessive seasonal variations. *MBio*, *6*(3), e00022-15.

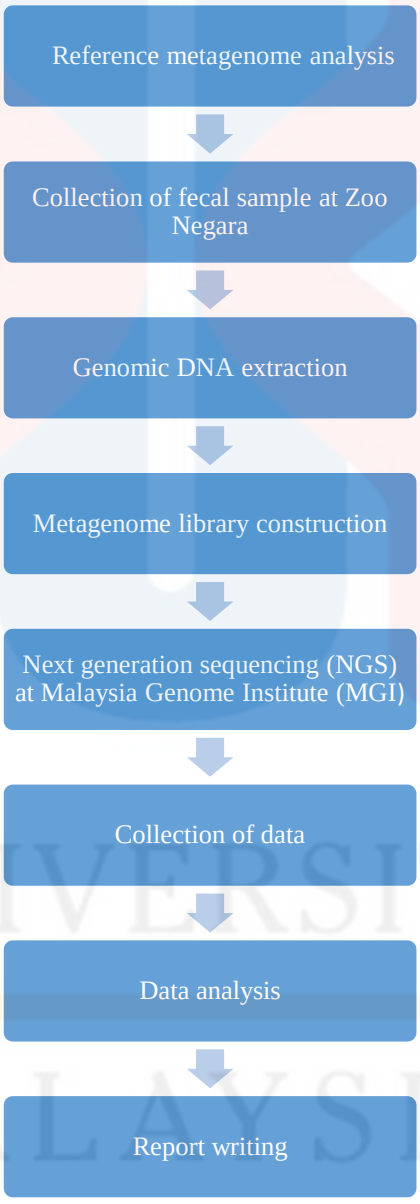
Zhan, X., Li, M., Zhang, Z., Goossens, B., Chen, Y., Wang, H., ... & Wei, F. (2006). Molecular censusing doubles giant panda population estimate in a key nature reserve. *Current Biology*, *16*(12), 451-452.

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*, *108*(43), 17714-17719.

Zoo, B. Beijing University, Beijing Agricultural University, Beijing Second Medical College, Beijing Natural History Museum, Shaanxi Zoology Institute (1986). Systematic anatomy and organ-histology. *Morphology of the Giant Panda. Science Press, Beijing*, 148-152.

APPENDIX A

Research Flow Chart



APPENDIX B

Organism Table

Metagenome	4526729	4526723	4526723	4526730	4526729	4526727
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Archaea	Archaea	Archaea	Archaea	Archaea	Archaea
Phylum	Euryarchaeota	Euryarchaeota	Crenarchaeota	Crenarchaeota	Crenarchaeota	Crenarchaeota
Abundance	4649	2221	52	21	91	17
Average e-Value	-9.12	-8.82	-7.14	-7.19	-5.67	-7.43
Average Percentage of Identity	70.83	69.56	67.66	67.97	68.9	70.34
Average Alignment Length	42.66	43.1	41.33	40.13	35.73	39.79
Number of Hits	3077	1515	52	21	91	17

Metagenome	4526723	4526727	4526729	4526723	4526727	4526730	4526727	4526724
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Archaea	Archaea	Archaea	Archaea	Archaea	Archaea	Archaea	Archaea
Phylum	Thaumarchaeota	Thaumarchaeota	Korarchaeota	Korarchaeota	Korarchaeota	Euryarchaeota	Euryarchaeota	Euryarchaeota
Abundance	12	4	10	19	4	5010	442	4
Average e-Value	-8.12	-10.5	-7.2	-6.05	-7	-8.84	-7.86	-6.25
Average Percentage of Identity	68.68	79.25	68.03	68.02	79.34	70.35	71.45	71.47
Average Alignment Length	41.5	43.25	39.4	38	36.75	42.54	39.71	36
Number of Hits	12	4	9	19	4	2970	410	4

Metagenome	4526727	4526723	4526730	4526729	4526730	4526723	4526727	4526729
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Archaea
Phylum	Actinobacteria	Actinobacteria	Actinobacteria	Acidobacteria	Acidobacteria	Acidobacteria	Acidobacteria	Thaumarchaeota
Abundance	1628	6950	9466	523	352	312	16	22
Average e-Value	-9.37	-8.68	-9.26	-7.79	-7.57	-8.67	-7.44	-6.54
Average Percentage of Identity	73.83	70.42	72.58	69.41	70.55	67.72	69.04	68.99
Average Alignment Length	41.38	42.35	41.92	40.39	39.59	43.38	39.81	38.31
Number of Hits	1278	3804	6774	444	352	276	16	22

Metagenome	4526727	4526729	4526724	4526727	4526730	4526723	4526724	4526729
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Bacteroidetes	Aquificae	Aquificae	Aquificae	Aquificae	Aquificae	Actinobacteria	Actinobacteria
Abundance	57322	423	2	126	497	303	20	10184
Average e-Value	-12.4	-8.09	-9.5	-8.32	-8.67	-9.04	-7.67	-8.64
Average Percentage of Identity	73.93	68.72	76.62	70.27	69.26	69.82	75.88	73.1
Average Alignment Length	46.86	41.73	43	42.23	43.15	43.79	35.12	40.64
Number of Hits	11441	423	2	126	432	303	20	6659

Metagenome	4526730	4526729	4526723	4526727	4526729	4526723	4526724	4526730
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Chlamydiae	Chlamydiae	Chlamydiae	Chlamydiae	Bacteroidetes	Bacteroidetes	Bacteroidetes	Bacteroidetes
Abundance	76	181	85	27	12170	89185	239	37129
Average e-Value	-7.08	-10.19	-7.45	-8.1	-10.98	-13.38	-9.46	-12.28
Average Percentage of Identity	70.05	70.93	67.76	79.29	71.2	71.34	79.06	72.06
Average Alignment Length	38.86	44.75	41.82	40.4	45.63	50.76	38.33	47.82
Number of Hits	76	138	83	3	7791	17416	239	13861

Metagenome	4526730	4526729	4526723	4526727	4526729	4526727	4526730	4526723
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Chloroflexi	Chloroflexi	Chloroflexi	Chloroflexi	Chlorobi	Chlorobi	Chlorobi	Chlorobi
Abundance	994	1193	773	159	816	125	624	641
Average e-Value	-8.92	-9.05	-8.95	-7.31	-9.16	-6.93	-8.46	-9.35
Average Percentage of Identity	70.48	69.47	69.45	72.57	69.81	72.08	70.65	69.56
Average Alignment Length	42.77	43.37	43.26	37.89	43.39	37.13	41.46	44.21
Number of Hits	880	1032	717	159	646	113	563	574

Metagenome	4526729	4526727	4526730	4526723	4526729	4526730	4526723	4526727
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Cyanobacteria	Cyanobacteria	Cyanobacteria	Cyanobacteria	Chrysiogenetes	Chrysiogenetes	Chrysiogenetes	Chrysiogenetes
Abundance	1199	169	1036	868	86	73	49	12
Average e-Value	-7.63	-7.98	-7.57	-8.19	-10	-9.74	-8.41	-7.79
Average Percentage of Identity	70.19	73	70.26	69.66	67.32	68.56	67.94	66.13
Average Alignment Length	40.2	39.29	40.15	41.71	46.74	45.87	42.5	42.86
Number of Hits	1199	169	1036	868	69	73	49	12

Metagenome	4526723	4526730	4526727	4526729	4526729	4526730	4526727	4526723
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Deinococcus-Thermus	Deinococcus-Thermus	Deinococcus-Thermus	Deinococcus-Thermus	Deferribacteres	Deferribacteres	Deferribacteres	Deferribacteres
Abundance	386	445	84	523	369	276	109	323
Average e-Value	-7.36	-7.34	-6.09	-6.81	-8.62	-9.18	-9.25	-11.42
Average Percentage of Identity	69.34	69.49	71.4	69.92	69.66	69.51	70.98	69.54
Average Alignment Length	40.63	39.97	37.27	38.5	42.21	43.55	43.62	47.88
Number of Hits	386	445	84	523	319	276	109	252

Metagenome	4526729	4526730	4526723	4526727	4526729	4526723	4526730	4526727
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Elusimicrobia	Elusimicrobia	Elusimicrobia	Elusimicrobia	Dictyoglomi	Dictyoglomi	Dictyoglomi	Dictyoglomi
Abundance	118	319	130	32	213	179	213	72
Average e-Value	-8.18	-14.35	-12.83	-13	-9.27	-10.47	-9.99	-6.82
Average Percentage of Identity	68.88	74.2	71.12	72.62	67.51	67.8	67.94	69.37
Average Alignment Length	41.03	50.52	49.5	49.12	45.13	47.5	46.31	39.31
Number of Hits	75	138	75	24	213	140	213	72

Metagenome	4526729	4526723	4526730	4526727	4526729	4526730	4526723	4526727
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Firmicutes	Firmicutes	Firmicutes	Firmicutes	Fibrobacteres	Fibrobacteres	Fibrobacteres	Fibrobacteres
Abundance	115508	91538	108860	47633	1423	813	1093	218
Average e-Value	-12.92	-11.92	-12.08	-11.71	-14.13	-14.69	-15.76	-14.53
Average Percentage of Identity	71.71	70.08	71.49	72.48	75.12	75.12	74.58	74.02
Average Alignment Length	49.89	48.94	48.27	46.78	49.43	50.72	53.03	51.3
Number of Hits	52425	35876	46364	24623	533	297	304	101

Metagenome	4526730	4526723	4526727	4526729	4526730	4526727	4526723	4526724
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Gemmatimonadetes	Gemmatimonadetes	Gemmatimonadetes	Fusobacteria	Fusobacteria	Fusobacteria	Fusobacteria	Firmicutes
Abundance	55	43	19	1696	1930	21224	1656	306
Average e-Value	-5.91	-6.93	-5.86	-11.74	-11.59	-16.62	-12.56	-9.07
Average Percentage of Identity	68.79	68.22	73.27	70.33	70.61	73.59	69.89	76.46
Average Alignment Length	37.44	40	36.05	48.33	47.7	55.91	50.34	39.48
Number of Hits	55	43	19	1052	1100	2851	994	306

Metagenome	4526729	4526723	4526727	4526723	4526729	4526730	4526727	4526729
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Planctomycetes	Planctomycetes	Planctomycetes	Nitrospirae	Nitrospirae	Nitrospirae	Nitrospirae	Gemmatimonadetes
Abundance	648	289	14	109	125	151	15	76
Average e-Value	-7.3	-8.96	-8.2	-7.65	-6.88	-8.44	-6.74	-6.07
Average Percentage of Identity	71.09	69.6	76.04	69.6	70.6	69.02	68.64	69.43
Average Alignment Length	38.14	42.79	39.33	40.8	38.45	43.25	38.9	37.64
Number of Hits	443	249	14	109	125	151	15	76

Metagenome	4526723	4526727	4526730	4526723	4526724	4526729	4526727	4526730
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Spirochaetes	Spirochaetes	Proteobacteria	Proteobacteria	Proteobacteria	Proteobacteria	Proteobacteria	Planctomycetes
Abundance	11118	1783	106503	13278	40	33276	10395	338
Average e-Value	-11.09	-11.8	-9.79	-8.99	-10.42	-8.77	-11.41	-9.54
Average Percentage of Identity	69.7	75.81	73.2	71	79.13	72.11	76.87	71.65
Average Alignment Length	47.5	45.02	42.81	42.4	39.96	41.43	43.86	43.28
Number of Hits	2741	998	33810	13278	40	23612	9099	338

Metagenome	4526723	4526730	4526723	4526729	4526727	4526729	4526724	4526730
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Tenericutes	Synergistetes	Synergistetes	Synergistetes	Synergistetes	Spirochaetes	Spirochaetes	Spirochaetes
Abundance	1264	208	353	247	185	4155	17	1823
Average e-Value	-10.57	-9.61	-13.22	-11.18	-12.03	-9.84	-9.47	-9.05
Average Percentage of Identity	70.41	68.85	70.64	70.39	70.7	70.74	73.98	70.7
Average Alignment Length	46.06	44.85	51.18	47.43	48.42	44.33	42.91	43.28
Number of Hits	846	192	228	191	87	2144	17	1088

Metagenome	4526730	4526727	4526729	4526723	4526730	4526724	4526729	4526727
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	Thermotogae	Thermotogae	Thermotogae	Thermotogae	Tenericutes	Tenericutes	Tenericutes	Tenericutes
Abundance	721	285	921	697	1106	19	1481	179
Average e-Value	-8.66	-8.63	-9.04	-9.16	-10.01	-10.23	-10.6	-9.77
Average Percentage of Identity	68.64	70.28	68.43	68.74	70.71	77.5	70.36	74.94
Average Alignment Length	42.93	42.18	43.77	44.16	44.44	41.84	46.31	41.94
Number of Hits	707	285	792	658	772	19	997	132

Metagenome	4526723	4526727	4526730	4526724	4526727	4526729	4526723	4526724
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
Phylum	unclassified (derived from Bacteria)	unclassified (derived from Bacteria)	Verrucomicrobia	Verrucomicrobia	Verrucomicrobia	Verrucomicrobia	Verrucomicrobia	Thermotogae
Abundance	78	26	663	3	126	1673	626	1
Average e-Value	-8.29	-7.07	-10.95	-10	-8.9	-10.89	-11.19	-10
Average Percentage of Identity	67	70.01	71.61	72.03	75.02	70.84	68.49	63.04
Average Alignment Length	43.37	39.25	45.06	40.67	39.65	45.81	47.76	46
Number of Hits	75	26	441	3	101	823	359	1

Metagenome	4526729	4526727	4526724	4526730	4526724	4526723	4526729	4526730
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Bacteria	Bacteria
Phylum	Ascomycota	Ascomycota	Ascomycota	Ascomycota	Arthropoda	Apicomplexa	unclassified (derived from Bacteria)	unclassified (derived from Bacteria)
Abundance	4771	1958	1400	10645	31	13	100	85
Average e-Value	-12.79	-11.21	-11.54	-11.6	-5.24	-10.2	-8.6	-8.88
Average Percentage of Identity	78.93	88.44	91.37	78.36	72.91	76.89	68.32	69.96
Average Alignment Length	44.98	37.38	36.81	43.02	34.34	41.6	43.05	42.47
Number of Hits	4292	1120	803	7796	27	13	100	85

Metagenome	4526723	4526727	4526730	4526723	4526723	4526729	4526730	4526723
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Eukaryota
Phylum	Chordata	Chordata	Chlorophyta	Chlorophyta	Basidiomycota	Basidiomycota	Basidiomycota	Ascomycota
Abundance	37	19873	62	17	19	247	668	1466
Average e-Value	-6.78	-8.69	-5.75	-6.78	-6.83	-8.73	-8.66	-8.91
Average Percentage of Identity	77.47	89.72	73.28	73.48	74.69	76.82	76.49	84.47
Average Alignment Length	34.2	32.65	35.56	37.28	37.61	38.33	38.53	34.97
Number of Hits	37	6948	62	17	15	247	668	868

Metagenome	4526727	4526724	4526729	4526723	4526730	4526724	4526729	4526730
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Eukaryota
Phylum	Platyhelminthes	Microsporidia	Cnidaria	Cnidaria	Cnidaria	Chordata	Chordata	Chordata
Abundance	69	3	44	33	216	10154	53	109
Average e-Value	-5.17	-5	-5.29	-7.27	-6.85	-8.49	-6.96	-6.29
Average Percentage of Identity	80.81	93.16	84.93	77.91	81.48	87.71	77.94	79.51
Average Alignment Length	30.09	25.33	27.52	36.5	32.27	32.74	35.56	32.47
Number of Hits	69	3	44	33	157	6203	53	109

Metagenome	4526724	4526729	4526723	4526730	4526723	4526729	4526730	4526729
Source	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG	KEGG
Domain	unassigned	unassigned	unassigned	Eukaryota	Eukaryota	Eukaryota	Eukaryota	Eukaryota
Phylum	unassigned	unassigned	unassigned	unclassified (derived from Eukaryota)	unclassified (derived from Eukaryota)	unclassified (derived from Eukaryota)	Streptophyta	Streptophyta
Abundance	45	4532	6178	429	281	295	549	628
Average e-Value	-10.96	-9.98	-11.75	-6.8	-7.12	-5.77	-7.56	-6.67
Average Percentage of Identity	84.4	71.52	70.72	72.67	71.69	73.85	78.53	76.51
Average Alignment Length	40.08	44.01	48.45	37.27	39.21	34.78	35.78	34.93
Number of Hits	45	2371	2509	429	281	295	549	628

Metagenome	4526727	4526730
Source	KEGG	KEGG
Domain	unassigned	unassigned
Phylum	unassigned	unassigned
Abundance	7066	6823
Average e-Value	-11.75	-10.42
Average Percentage of Identity	76.04	72.87
Average Alignment Length	44.96	44.14
Number of Hits	2566	3062