Malaysian Journal of Microbiology, Vol 13(4) December 2017, pp. 350- 362

Malaysian Journal of Microbiology
Published by Malaysian Society for Microbiology
(In SCOPUS since 2011)

Shotgun metagenomic analysis of microbial communities in the surface waters of the Eastern South China Sea

Jessica Song[1], Aazani Mujahid[2], Po-Teen Lim[3], Azizan Abu Samah[4], Birgit Quack[5], Klaus Pfeilsticker[6], Sen-Lin Tang[7], Elena Ivanova[8], and Moritz Müller[1*]

[1]Faculty of Engineering, Computing, and Science, Swinburne University of Technology, Sarawak Campus, 93350 Kuching, Sarawak, Malaysia.
[2]Department of Aquatic Science, Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, 93400 Kota Samarahan, Sarawak, Malaysia.
[3]Bachok Marine Research Station, Institute of Ocean and Earth Sciences, University of Malaya, 16310 Bachok, Kelantan, Malaysia.
[4]National Antarctic Research Center, Institute of Postgraduate Studies Building, University of Malaya,50603, Kuala Lumpur, Malaysia.
[5]Marine Biogeochemistry, GEOMAR, Helmholtz Centre for Ocean Research, Kiel, Germany.
[6]Institute of Environmental Physics, University of Heidelberg, Heidelberg, Germany.
[7]Biodiversity Research Center, Academia Sinica, Taipei 115, Taiwan.
[8]Faculty of Science, Engineering and Technology, Swinburne University of Technology, Hawthorn, Victoria, Australia.
Email: mmueller@swinburne.edu.my

Received 6 September 2016; Received in revised form 9 May 2017; Accepted 15 May 2017

ABSTRACT

Aims: The South China Sea (SCS) harbours a rich biodiversity. However, few studies have been published on its diverse communities, particularly its microbial counterparts. As key players behind many of the vital processes carried out in the ocean, microbes are the focus of this study, placing particular emphasis on community composition, structure, and function.
Methodology and results: By employing next generation shotgun sequencing technologies (Illumina HiSeq2000), we assessed the taxonomic structure and functional diversity of the prokaryotic communities in surface waters collected from 3 representative sites in the Eastern SCS: Sarawak (Kuching), Sabah (Kota Kinabalu), and Philippines (Manila). Comparisons were undertaken to similar studies from coastal and open ocean environments. All 3 locations were dominated by members of the *Proteobacteria* (*Alpha-* and *Gamma-*) and *Cyanobacteria* (*Synechococcus* sp. and *Prochlorococcus* sp.). The highest proportion of *Gammaproteobacteria* was found in Sarawak, representing an approximate 20% of total sequences. Archaeal assemblages were made up largely of *Euryarchaeota* and unclassified sequences, while *Crenarchaeota* and *Thaumarchaeota* were present in much smaller proportions, except in the Philippines where *Thaumarchaeota* made up almost 40% of the entire taxa.
Conclusion, significance and impact of study: The majority of the microbial communities adhered to a core set of functional genes across the different locations. However, differences existed particularly in Sarawak waters which are hypothesized to be due to local environmental parameters such as riverine influence. The results obtained from this study provide the first comparison of prokaryotic communities in the surface waters of the eastern SCS and will serve as a good platform for prospective studies in the field of environmental science.

Keywords: Metagenomics, microbial communities, South China Sea

INTRODUCTION

Microbes make up a vast proportion of the marine community, far exceeding their multi-cellular counterparts in terms of abundance, biomass and activity (Pomeroy and Darwin, 2007). As key players in the marine ecosystem, microbes mediate a large percentage of the vital biogeochemical processes carried out in the ocean — all of which bear great impact on the marine community as a whole (Das *et al.*, 2006; Dang *et al.*, 2008). Processes such as nutrient cycling, toxin neutralization and degradation, and other biogeochemical cycles carried out by these microbes mediate the flow of energy and matter within the different trophic levels that exist, which on a

***Corresponding author**

larger scale contribute to maintaining the oxidative state of our planet (DeLong and Karl, 2005; Gianoulis *et al.*, 2009).

In an attempt to shed some light on the underexplored microbial communities in the ocean, scientists worldwide have worked towards assessing the functional capabilities of these microbial populations and the factors involved in shaping their structure. One such example would be the Global Ocean Sampling (GOS) expedition conducted in 2003 by the Craig Venter Institute (Venter *et al.*, 2004), where next generation sequencing technologies were employed to sample the microbes that inhabit the Earth's oceans. Marine microbiota which were sampled from surface waters across both the Atlantic and Pacific Ocean were studied for both their structural diversity as well as their functional adaptations across the different environments (Rusch *et al.*, 2007; Yooseph *et al.*, 2010). Based on the large collection of marine genomic sequences, communities from both nutrient-rich and nutrient-deficit environments were compared, revealing differences in genome size, genetic composition, and metabolic potential (Yooseph *et al.*, 2010).

The South China Sea (SCS) is said to harbour a wide diversity of marine ecosystems, providing rich biological resources to many countries across the globe (Ng and Tan, 2000; Liao *et al.*, 2009). Its promising diversity, however, remains largely unexplored. Due to a lack of information in this area, many initiatives have since been organized to compile findings and studies carried out on the SCS by scientists from all over the world in an effort to better understand its biodiversity (Ng and Tan, 2000). While several studies have been conducted on plant and animal communities, very few surveys have been carried out on its microbial diversity, the few existing having focused mainly on sediments, estuaries, and open ocean environments (Jiang *et al.*, 2007; Liao *et al.*, 2009; Zhang *et al.*, 2011; Zhang *et al.*, 2014). In this study, however, we focus mainly on the diversity and function of surface coastal water prokaryotic communities present along the equatorial regions of the SCS, fringing off the coast of Sarawak and extending towards the Philippine islands.

This research aims to conduct a metagenomic analysis of the prokaryotic communities that inhabit the surface coastal waters of the eastern region of the SCS through the employment of next generation sequencing techniques. We analyzed the composition of marine microbial communities in an effort to pinpoint the key players as well as their roles in their respective environments. The datasets containing genomic and physicochemical information that were collected along the different sampling sites (Sarawak, Sabah and the Philippines) are also compared to similar studies conducted in other open ocean and coastal waters in order to study the possible effects of spatial variation on community structure and function.

MATERIALS AND METHODS

Sampling sites and collection

Samples were collected onboard the Sonne Cruise No.SO218 research vessel during the SHIVA-Malaysia Campaign (Stratospheric Ozone: Halogens Impacts in a Varying Atmosphere) conducted in November 2011. The transect began from 1°15'36.0"N 103°49'12.0"E (Singapore) to 14°35'24.0"N 120°58'12.0"E (Manila), where surface water samples were collected from a depth of 5 m. Water samples of approximately 150 L were collected from the moon pool onboard the ship, through a silicone tube connected to a peristaltic pump, at 3 to 4 h intervals. The collected seawater was then filtered through a 10 µm plankton net, followed by a second filtration step using 0.2 µm nylon membrane filters (GE Healthcare Bio-Sciences, Pittsburgh, USA). Filters were then stored at room temperature in 50 mL screw cap bottles and fixed in 15 mL of saline ethanol.

CTD stations were set up at each sampling location to measure the physicochemical parameters along the cruise. The 3 representative locations chosen for this study are situated along the coast of Borneo, Malaysia, and the Philippines. Samples S2005, S2405, and S2705 were collected 20 nautical miles off the coast of Sarawak (Kuching), Sabah (Kota Kinabalu) and Philippines (Manila), respectively (Table 1).

Nitrate, phosphate, nitrite, and silicate concentrations (Table 1) were quantified using a QuAAtro auto-analyzer (SEAL Analytical, UK) following protocols provided in the SEAL analytical operation manual and Grashoff *et al.* (1999).

Table 1: Sampling site locations, in situ parameters, and sea water composition in Sarawak, Sabah, and the Philippines.

Sample	Location	Temp (°C)	Depth (m)	Distance from coast (km)	Salinity	Nitrate (µmol/L)	Phosphate (µmol/L)	Nitrite (µmol/L)	Silicate (µmol/L)
Sarawak S2005	3º27'28.2"N 111º49'52.8"N	29.2	5	65.12	32.0	0.041	0.031	0	3.426
Sabah S2405	7°53'56.4"N 118°03'13.8"E	28.2	5	95.03	32.5	0.056	0.007	0	1.855
Philippines S2705	9°23'09.6"N 120°17'41.4"E	29.1	5	144.03	33.1	0.136	0.017	0.015	2.174

NA, Not available

DNA extraction and amplification

DNA was extracted using the PowerWater® DNA Isolation kit (MoBio, Carlsbad, CA, USA), according to the instructions manual provided. Multiple displacement amplification (MDA) was carried out on the extracted genomic DNA using the phi29 DNA Polymerase kit (New England Biolabs, Inc.) following the protocol described by Wang *et al.* (2004), with slight modifications.

First, the denaturation step was carried out by incubating 4 µL of the genomic DNA with 0.5 µL of random hexamers (400 µg/mL) and 9 mL of sample buffer (50mM Tris HCl pH 8.2, 0.5 mM EDTA) at 95 °C for 3 min. A master mix consisting of 0.3 µL of phi29 DNA polymerase (100 units/mL), 2 µL of 10× phi29 DNA polymerase buffer, 0.2 µL of 100× BSA and 3.2 µL of dNTPs (2.5 mM) was then added to the denatured template and made up to a total volume of 20 µL using filtered deionized water.

The mixture was put through an amplification step, where it was incubated at 30 °C for 16 h, followed by an enzyme inactivation step at 65 °C for 10 min. The whole procedure was then repeated on the final amplified product but with 5× the original volume of reagents and starting template to improve yield. Final MDA products were then run on 1% agarose gel and visualized using the Gel Doc™ XR+ System (Bio-Rad Laboratories, Inc.). Multiple displacement amplification (MDA) has been shown to sometimes result in a potential bias in the results that exhibit characteristics that show a lesser correlation to its local environment (De Bourcy *et al.*, 2014).

Duplicates were prepared for each sample and pooled to obtain a final concentration of more than 2 µg of genomic DNA per sample. Shotgun metagenomic libraries were constructed and sequenced in the NGS Lab of Beijing Genomics Institute (BGI), Beijing using high-throughput Illumina HiSeq2000 2×100 bp paired-end sequencing technology.

Data and statistical analyses

Approximately 1 GB of sequencing data was generated for each of the 4 sites. Duplicate and adapter sequences were removed from the raw data, as well as reads with a phred score of ≤20, through an analysis pipeline carried out in the NGS Lab (BGI, Beijing).

The unassembled sequencing reads were uploaded directly to MG-RAST (Meyer *et al.*, 2008) where a normalization step was carried out, assigning each metagenome with a unique internal ID. Following this, a round of QC was performed, bypassing both the demultiplexing and screening steps, where the reads were filtered to remove any sequencing artifacts and ambiguous basepairs that exceed 5 bp in length.

Sequences were screened for potential coding elements using the BLASTX search tool (Altschul *et al.*, 1997), referenced against a comprehensive nonredundant (nr) SEED database, with 10^{-5} expect value (E) cut-off, 80% minimum identity cutoff, and a minimum alignment length cutoff of 50 (modified from *Mason et al.*, 2014). Sequences were referenced against the BlastX database alongside other accessory databases such as GREENGENES (DeSantis *et al.*, 2006), RDP-II (Cole *et al.*, 2007), and the European 16S RNA database (Wuyts *et al.*, 2002) to identify candidate RNA genes, while functional classifications were executed using external databases such as eggNOG (Powell *et al.*, 2013) and KEGG Orthology (Kanehisa and Goto, 2000), and mapped against SEED Subsystems to suggest the possible metabolic pathways and enzymes encoded within the genome.

Six additional shotgun metagenomes representing Ocean and Coastal Environments were collected (see Table 2) and compared using principal component analysis based on the relative abundance of SEED functional categories (normalized against the total gene count of each metagenome).

The overall species richness in each sample was estimated using rarefaction curves calculated based on the annotated species abundance counts (data not shown), while its α-diversity was measured to obtain the mean number of species in each site (Table 2).

Table 2: Metagenome IDs and alpha diversity scores calculated based on normalized sequence abundance counts depicting overall species richness of all sample and reference metagenomes. Data obtained from http://metagenomics.anl.gov/

Sampling Environment	Sample (Metagenome ID MG-RAST)	Alpha diversity (α)
Open Ocean	Philippines (4579203.3)	535.69
	Sarawak (4557808.3)	440.32
	Sabah (4579202.3)	220.93
	Indian Ocean 1 (4441607.3)	502.71
	Indian Ocean 2 (4441609.3)	502.75
	North Atlantic Ocean – Sargasso Sea 1 (4441573.3)	633.46
	North Atlantic Ocean – Sargasso Sea 2 (4441574.3)	925.92
	Caribbean Sea – Atlantic Ocean (4441589.3)	464.39
Coastal	Pacific Ocean – Gulf of Panama (4441591.3)	764.05
	Pacific Ocean – Galapagos Island (4441596.3)	675.96
	Pacific Ocean - Monterey Bay (4443713.3)	567.94

RESULTS

Sampling and metagenome summary

Surface water temperature and salinity across all 3 sites appeared to be constant except in Sabah where there was a slight drop in temperature below 29 °C, and in the Philippines where there was an increase in salinity (Table 1). Spectrophotometric measurements revealed that the distribution of nitrates and nitrites were relatively uniform throughout, with the highest concentration detected in Philippine waters, south of the Mindoro Islands; which was double that of the other regions (Table 1). Phosphate and silicate concentrations followed a similar pattern except in the coastal waters of Sarawak where maximum concentrations were detected.

Metagenomes from each site were sequenced through the use of a whole genome shotgun approach. High-throughput Illumina sequencing technology was employed for this study, generating an average of 5,000,000 raw reads per sampling site; each sample set approximately 1 GB in size.

Raw sequences were uploaded to MG-RAST and subject to the QC and processing pipeline designed by the analysis platform. An estimated 70-90% of total sequences were retrieved for further processing and annotation.

Species diversity

Rarefaction curves generated all arrived at curvilinear or plateau phase, indicating that the microbial communities in all 3 metagenomes were well-represented. The alpha-diversity of the microbial metagenomes in each site was calculated to assess both species richness and evenness. Philippine waters demonstrated the greatest diversity of all the sites (α = 535), followed by Sarawak (α = 440), making Sabah (α = 221) the least diverse waters of the 3 samples (Table 2).

Community composition

Prokaryotic communities were analyzed based on relative sequence abundance, of which bacterial assemblages represented an average of 50-60% of total sequences while *Archaea* made up 1-2%. Unassigned sequences occupied approximately 15% of the overall metagenome, while another 1% comprised of unclassified reads. Sequences belonging to *Eukaryota* made up an average of 10-20% of the total sequences, except in Sabah where it only constituted 3%. However, the sampling strategy did not select for the eukaryotic community (a substantial percentage of *Eukaryotes* were removed during the pre-filtration step) and thus will only be discussed briefly in this study.

Bacterial communities in all 3 metagenomes were predominantly represented by 5 major classes comprising of *Alphaproteobacteria*, *Gammaproteobacteria*, *Cyanobacteria*, *Flavobacteriia* and *Actinobacteria*. The *Alphaproteobacteria* class (20%), made up mostly of *Rhodobacterales* and *Rhizobiales*, along with a high ratio of the SAR11 clade, was found to be present in highest proportion in all samples. *Cyanobacteria* made up an average of 10-15% of the overall community, with *Synechococcus* sp. and *Prochlorococcus* sp. having the highest sequence counts in all 3 sampling sites. *Flavobacteriia* (4%) was present in similar proportions across all sites. Other classes, such as *Beta-* and *Deltaproteobacteria*, *Actinobacteria*, and *Planctomycetia*, were also present in all 3 sites (Table 3).

Archaeal assemblages consisted largely of *Euryarchaeota* and unclassified sequences, while *Crenarchaeota* and *Thaumarchaeota* were present in much smaller proportions, except in Philippines where *Thaumarchaeota* made up almost 40% of the entire taxa (Table 3).

Table 3: Proportion of top Prokaryotic sequences found in all 3 metagenomes based on relative abundance counts.

Prokaryotes	Proportion of Total Sequences(%)		
	Sarawak S2005	Sabah S2405	Philippines S2705
Proteobacteria	44.87	39.08	30.83
Alphaproteobacteria	21.84	26.22	16.90
Gammaproteobacteria	19.76	9.03	10.07
Cyanobacteria	9.19	16.66	8.05
unclassified (derived from Cyanobacteria)	9.16	16.63	8.01
Bacteroidetes	5.33	5.93	4.52
Flavobacteriia	3.92	4.13	3.07
Actinobacteria	1.53	1.51	2.57
Euryarchaeota	0.46	0.65	1.00
Halobacteria	0.09	0.09	0.18
Methanomicrobia	0.10	0.15	0.23
unclassified (derived from Eurarychaeota)	0.10	0.14	0.23
unclassified (derived from Archaea)	0.14	0.18	0.25
Thaumarchaeota	0.02	0.07	0.81
Crenarchaeota	0.05	0.06	0.13

Both heatmap and principal component analysis (PCoA) of the 3 metagenomes plotted alongside the 6 additional reference metagenomes showed that the bacterial sequences within the sample and reference datasets differed substantially (Figures 1 and 2).

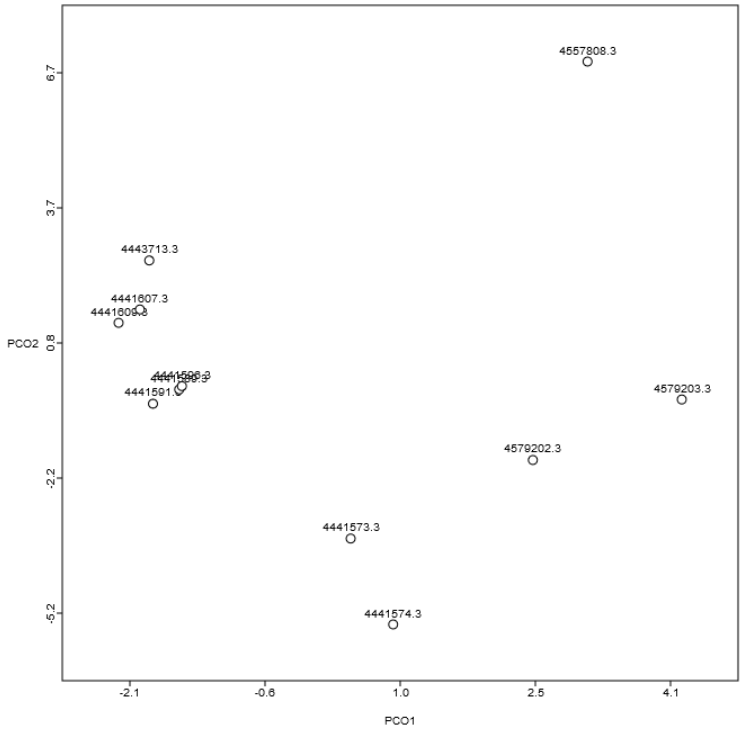

Figure 1: Principal component analysis (PCoA) scatterplot showing the percent of variation of the taxonomic groups among the sample and reference metagenomes.

Top functional genes and biogeochemically-relevant genes

The total processed reads for each sample set consisted of a calculated average of 80% ORFs, producing 2 to 3 million predicted protein coding features. Approximately 15 to 30% of these features were annotated using at least one of the protein databases in MG-RAST's collective M5NR database. 70% of all annotated features were then assigned to functional categories, as shown in Figure 3.

All 3 metagenomes were dominated by carbohydrate metabolism, protein metabolism, and amino acids and derivative reads. Carbohydrate and protein metabolism were equally dominant across the sampled metagenomes, both making up around 9% of total reads respectively (Figure 3).

Fatty acids, lipids, and isoprenoid reads found in all 3 sites encoded mostly for fatty acid and isoprenoid biosynthesis as well as phospholipid metabolism. Reads from the cell wall and capsule subsystem were composed of similar reads among all metagenomes and highest in the Sabah (S2405) sample. The sequences found across all 3 samples were mostly related to peptidoglycan biosynthesis, sialic acid metabolism, and mycolic acid synthesis. Reads encoding for rhamnose containing glycans, however, were most dominant in the Philippines (S2705) sample.

Several genes encoding for some of the more important biogeochemical processes were also present in slightly lower frequencies. The frequency of phosphate metabolism genes was more or less consistent across all sites, however slightly higher in the Sabah (S2405) sample, while nitrogen cycle genes were slightly less abundant in these waters compared to the other locations (data not shown). Reads belonging to the phosphorus cycle were mainly involved in the oxidative phosphorylation pathway, and nicotinate and nicotinamide metabolism, and were present in all 3 sample locations. The sample from Sabah (S2405) appeared to have the highest frequency of phage proteins of all the locations, consisting of phosphate ABC transporter proteins and phosphate starvation-inducible protein, *PhoH*. The highest counts of phosphate metabolism genes were also found in Sabah.

Nitrogen metabolism genes present were identical throughout all 3 metagenomes from Sarawak (S2005), Sabah (S2405), and the Philippines (S2705), and were involved predominantly in amino acid biosynthesis, glyoxylate and dicarboxylate metabolism, and ammonium transport. Sulfur and iron cycle genes were also present throughout all 3 sampled locations. Iron acquisition and metabolism reads were predominantly made up of ABC transporters, electron transfer flavoprotein subunits, and light-harvesting complex proteins in samples from Sarawak (S2005), Sabah (S2405), and the Philippines (S2705).

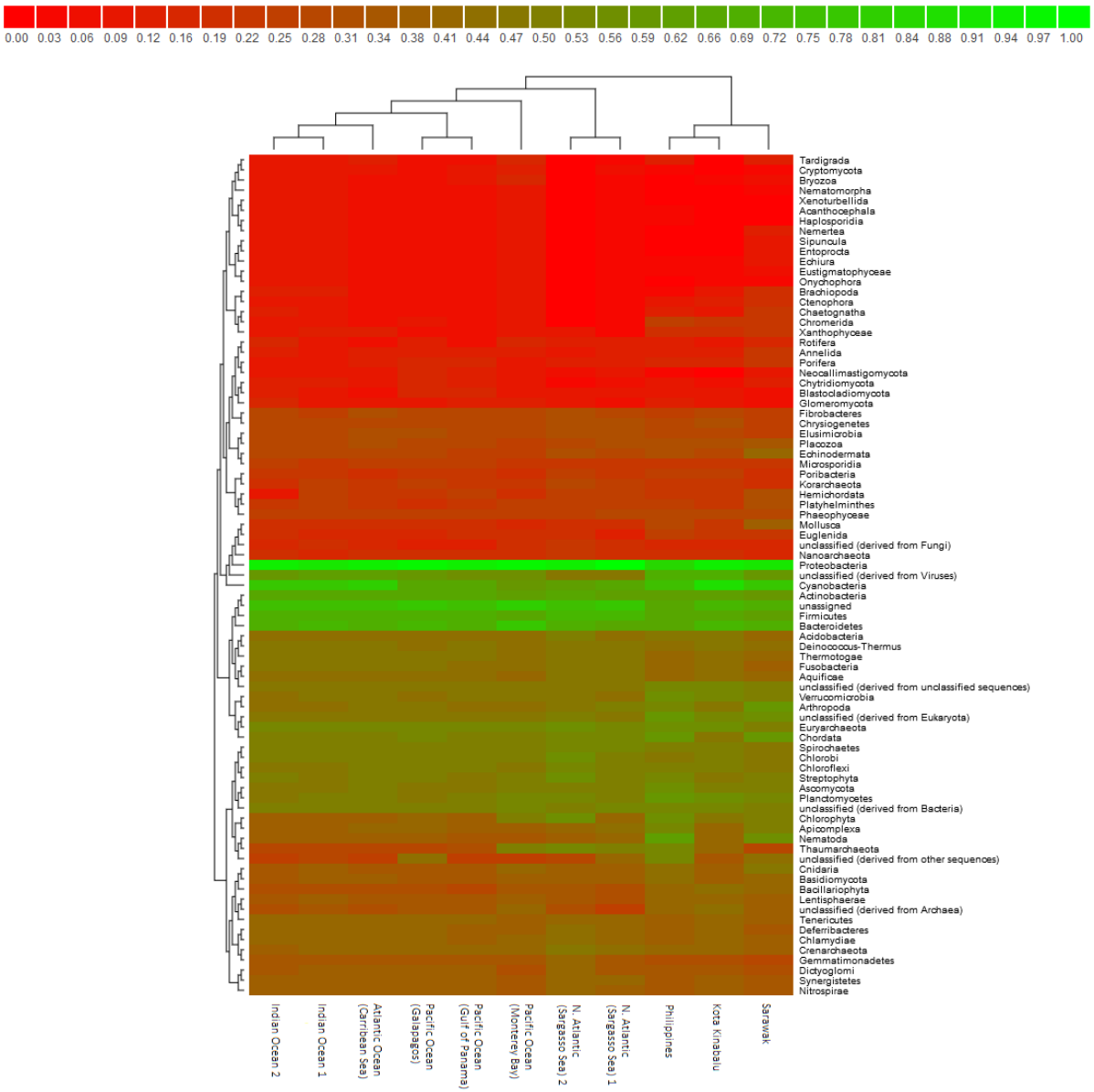

Figure 2: Heatmap comparing the different taxonomic groups (based on abundance counts) of the study metagenomes with reference metagenomes. The y-axis dendrogram shows the similarity/dissimilarity between the different taxonomic groups, whereas the x-axis dendrogram indicates the similarity/dissimilarity between the different metagenomes (Source: MG-RAST server, ver. 3.5).

Sulfur metabolism reads encoded mostly for oxidoreductases involved in energy and nucleotide metabolism, and dimethylsulfoniopropionate demethylase across all sites. A comparison between the 3 sample metagenomes and the additional 6 reference metagenomes through both heatmap and principal component analysis demonstrated a significant difference in the functional genes found in both sample and reference datasets (Figures 4 and 5).

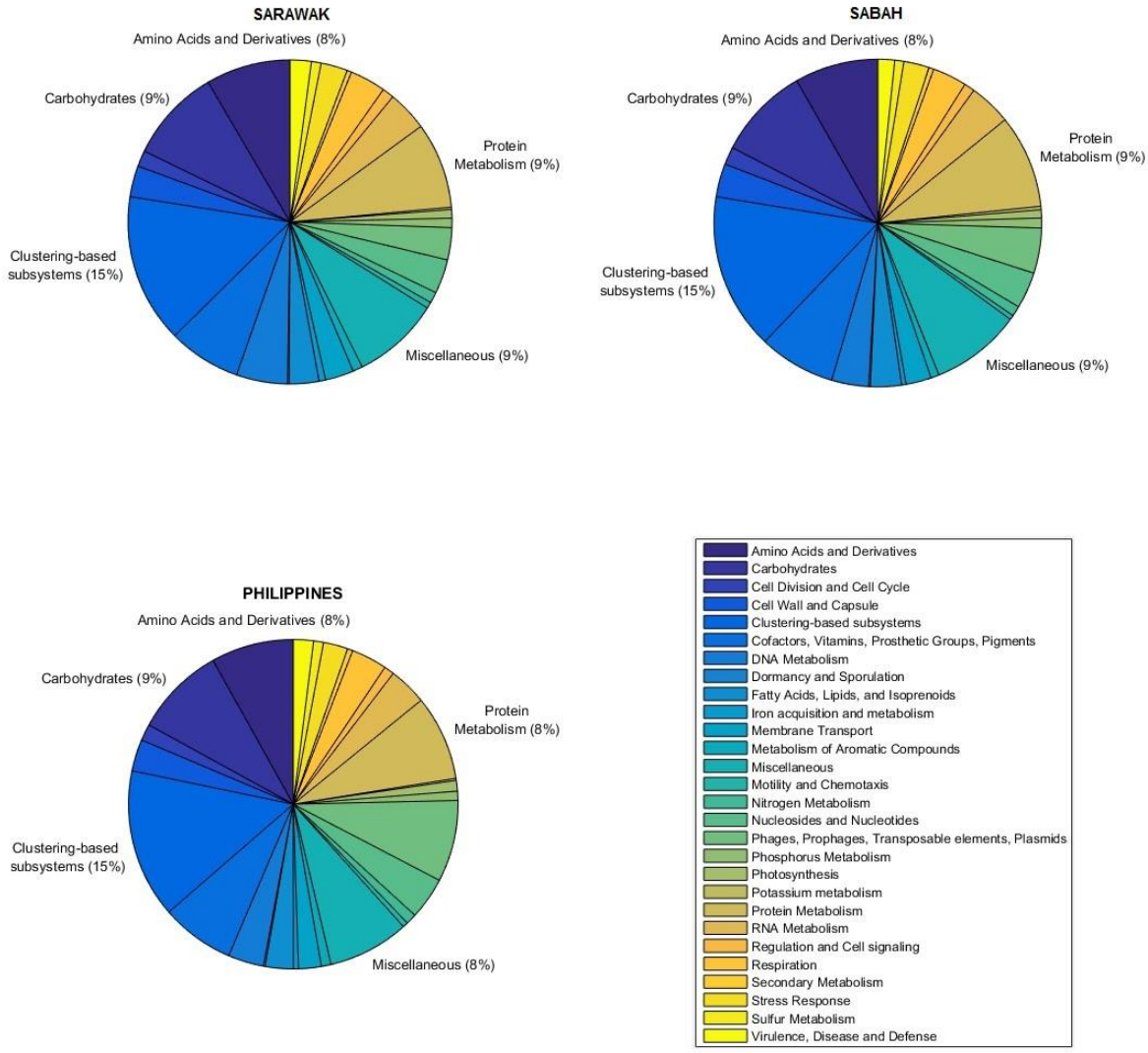

Figure 3: Functional category hits distribution annotated using MG-RAST Subsystems classification.

DISCUSSION

Through the employment of next generation sequencing technologies, this study aims to provide an assessment of the phylogenetic and functional diversity of the microbial communities that exist within the eastern SCS surface waters, and the possible impacts of spatial variation on their community patterns.

Sarawak waters (S2005) exhibited slightly higher concentrations of phosphates (compared to the other sampling sites), which can be indicators of anthropogenic activity (Liu *et al.*, 2012), and were possibly introduced through the Rajang river. The Rajang river is the longest river in Malaysia and discharges into the SCS nearby the sampling site. This correlates positively with the comparatively elevated silicate concentrations which is characteristic of riverine influence (Moore *et al.*, 1986). Other studies conducted along the same transect also suggest an anthropogenic input in the coastal waters of Sarawak, recording the highest concentration of total Chlorophyll a (TChla) of all the sites (unpublished date). Maximum nitrate concentrations were recorded in the southern regions of the Philippines (S2705), which demonstrated the highest alpha diversity of all 3 metagenomes studied. This may be indicative of nutrient enrichment in the waters surrounding the Philippine islands as a result of human-derived impacts such as urban development or aquaculture —a long-standing, steady industry in the country (Irz *et al.*, 2007; Nogales *et al.*, 2011).

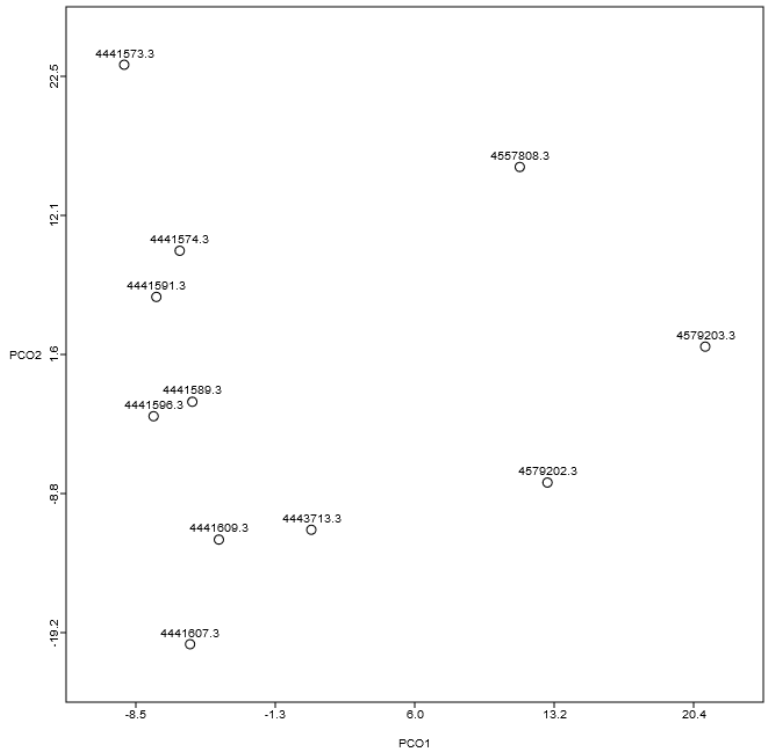

Figure 4: Principal component analysis (PCoA) scatterplot showing the percent of variation of the functional genes found among the sample and reference metagenomes.

Principal Component Analysis (PCoA) revealed that the microbial diversity observed in the 3 samples from the eastern SCS is significantly different from other similar shotgun sequencing studies from the Indian Ocean, Northern Atlantic, and the Pacific (Figure 1). Among the 3 study sites, the microbial communities found in Sabah and Philippines were more closely related to each other than to Sarawak. The 3 samples were most closely related to samples collected in the Sargasso Sea, whereas metagenomes collected from the Indian Ocean were grouped with coastal water samples from the Galapagos, Caribbean Sea, and Monterey Bay (Figure 1). In addition to the variation in its prokaryotic community, Sarawak waters also appeared to harbour eukaryotes distinctive of the other sites, particularly those belonging to the phyla *Chromerida*, *Xantophyceae*, and *Porifera*. The observed pattern was further supported by a heatmap (Figure 2), which similarly grouped the 2 sample metagenomes away from the reference genomes and Sarawak on its own.

Despite them being grouped on their own, the dominance of *Alphaproteobacteria* (comprising an average 20 to 30% of the overall microbial communities in Sarawak, Sabah and the Philippines) is consistent with other studies such as the Sorcerer II Global Ocean Sampling (GOS) expedition. A large percentage of the sequences belonged to *Candidatus Pelagibacter* (average 30% of *Alphaproteobacteria*), a prominent member of the SAR11 clade, commonly known to be one of the most abundant groups of Bacteria found in marine surface waters (Dang *et al.*, 2008; Brown *et al.*, 2012). This was again consistent with findings reported from the reference metagenomes, where *Pelagibacter* sequences were found in all surface water samples from across different marine habitats in the Atlantic and Pacific oceans, namely coastal, estuary and open ocean environments (Rusch *et al.*, 2007; Brown *et al.*, 2009). *Rhodobacterales*, which was the second largest group of *Alphaproteobacteria*, comprised of a diverse range of species with a relatively even distribution in terms of abundance. *Gammaproteobacteria* consisted of *Alteromonodales*, *Pseudomonadales*, *Oceanospirillales* and *Enterobacteriales* across all sites, which is also similar to previous studies (Zhang *et al.*, 2007).

The apparent core community composition patterns that extend across samples from Sarawak, Sabah, and the Philippines, were shared in other recent studies conducted in the SCS as well. In studies by both Zhang *et al.* (2014) and Tseng *et al.* (2015), the prokaryotic communities in surface waters were dominated mainly by *Proteobacteria* and *Cyanobacteria*. *Alphaproteobacteria*, making up the largest class, was composed mainly of members of the SAR11 clade and *Rhodobacteraceae*, followed by high counts of *Prochlorococcus*, *Flavobacteria* and *Actinobacteria*. *Gammaproteobacteria* were also represented by a similar make-up to those in our study, consisting largely of *Alteromonadales*, *Oceanospirillales*, *Pseudomonadales*, and *Vibrionales* (Zhang *et al.*, 2014; Tseng *et al.*, 2015).

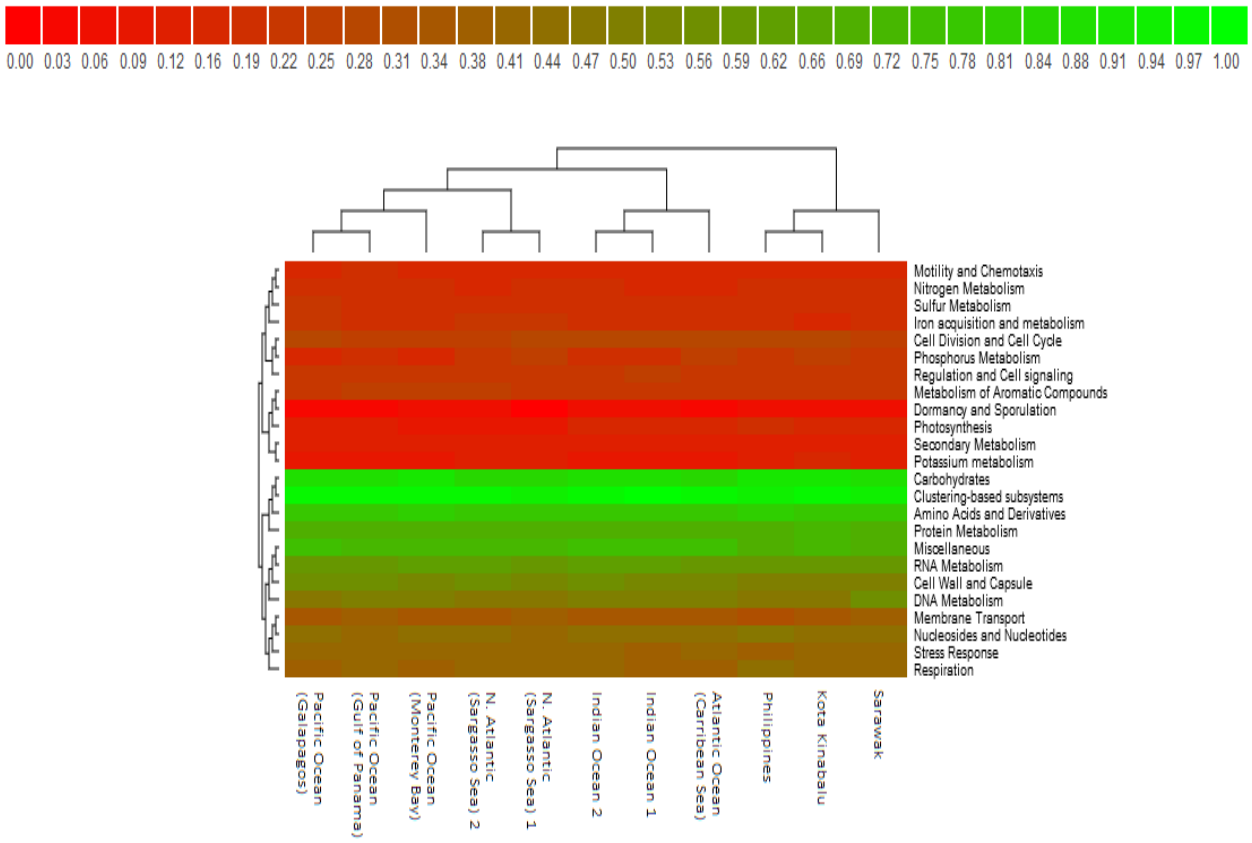

Figure 5: Heatmap comparing the different metabolic processes encoded for within the different metagenomes based on abundance counts. The y-axis dendrogram shows the similarity/dissimilarity between the different functional categories, whereas the x-axis dendrogram indicates the similarity/dissimilarity between the different metagenomes (Source: MG-RAST server, ver. 3.5).

Despite a general similarity with other samples from the SCS, the Sarawak metagenome did show several features that distinguished it from all other samples. The highest proportion of *Gammaproteobacteria*, for example, was found in Sarawak, representing an approximate 20% of total sequences (Table 3). This is not typical of *Gammaproteobacteria* as they are mostly found in higher numbers in lower pelagic and benthic environments (Zinger *et al.*, 2011). However, other studies on culturable bacteria from coastal waters off Sarawak coast also revealed high occurrences of *Vibrio* (Kuek *et al.*, 2015; Kuek *et al.*, 2016), demonstrating unusually high counts in these waters. A possible explanation could be that the coastal waters are anthropogenically polluted and that enhanced nutrient concentrations lead to high *Vibrio* proliferation (Eilers *et al.*, 2000; Pinhassi and Berman, 2003), however, this seems unlikely as other bacterial groups did exhibit similar patterns suggestive of a positive correlation. *Synechococcus* for example are known to thrive more in nutrient rich environments as compared to *Prochlorococcus* which, in contrast, appear to be more abundantly found in oligotrophic environments off-coast (Partensky *et al.*, 1999; Zhang *et al.*, 2009). However, in Sarawak (and also the other two SCS metagenomes), *Prochlorococcus* accounted for one of the highest proportion of sequences, indicating oligotrophic conditions. Sarawak's coastline is dominated by large rivers (including the abovementioned Rajang, the largest river in Malaysia) which input significant amounts of sediment and it seems likely that the affiliation of *Vibrio* with sediment particles (Zinger *et al.*, 2011) leads to the high counts in the area.

Archaeal assemblages represented a much smaller proportion of the metagenomes, making up an average of 1 to 2% of total sequences. These results are in agreement with previous findings reported of *Archaea* communities in marine near surface environments where these microbes are typically present in smaller proportions (Rusch *et al.*, 2007; Biers *et al.*, 2009). Instead, *Archaea* are known to exhibit a preference for deeper waters where they are more commonly found in higher numbers (Luria *et al.*, 2014; Signori *et al.*, 2014). The dominant clade, *Euryarchaeota*, which are more typically known to inhabit surface waters, made up an average of 40 to 70% of total archaeal sequences in all 3 sites, while the common deep-sea inhabitants, *Crenarchaeota*, only comprised 5%

(Brown *et al.*, 2009). This corresponded with results obtained by Chan *et al.* (2013) where *Euryarchaeota* made up 65.2% of Archaea communities found in the surface seawater surrounding the coast of West Malaysia. The community in Philippine (S2705) waters, which generally demonstrated higher counts of *Archaea* overall, experienced a *Thaumarchaeota* bloom which made up 37% of total archaeal sequences in contrast to the average of 5% found in the other 2 sites. Usually present more dominantly in both meso- and bathypelagic regions, *Thaumarchaeota* populations have been observed to thrive in the presence of high concentrations of chlorophyll-a, which may suggest a positive relationship between *Thaumarchaeal* abundance and the eukaryotic phytoplankton communities that sustain them (Robidart *et al.*, 2012). The surface waters along all 3 sampling sites contained high concentrations of TChla, which coincides with high diatom concentrations found in these waters (Cheah *et al.*, forthcoming 2015). This was confirmed by high silicate concentrations in regions surrounding the Mindoro Islands, suggesting a positive correlation with the elevated diatom concentrations that may be sustaining the *Thaumarchaeota* bloom.

Principal component analysis (PCoA) and heatmaps generated allow for a direct comparison between the different metagenomes. The 3 SCS samples were grouped more closely together compared to the other reference metagenomes (Figure 4). Among the 3 sites, the metabolic properties of the microbial communities found in Sabah (S2405) and Philippine (S2705) waters were most closely related to each other (Figures 4 and 5).

Through the large collection of sequences obtained, we were able to observe a similar pattern in these communities in terms of their functional characteristics across the different sites. The top functional genes appeared to be similar in all 3 locations, mainly consisting of genes involved in carbohydrate, protein, and amino acid metabolism. These findings were consistent with those of the reference metagenomes, where the surface prokaryotic communities found in the pelagic region of the Indian, Pacific, and Atlantic Ocean (Venter *et al.*, 2004; Rusch *et al.*, 2007; Hewson *et al.*, 2009) were also observed to share a similar distribution of functional genes. This may suggest that the prokaryotes that dominate surface water environments possibly share a core set of genes that are essential to the adaptation and survival of these communities (Hewson *et al.*, 2009).

A more significant variation was observed between sites for the less abundant genes encoding for biogeochemically-relevant processes. Iron metabolism genes were found in highest abundance in Sarawak (S2005) waters, as opposed to the other 2 sites which shared a similar composition but in slightly lesser counts. These genes were made up predominantly of siderophores, in the form of ABC transporter proteins involved in binding Fe^{3+}, and light-harvesting complex proteins found mainly in the SAR11 clade and *Prochlorococcus marinus*, respectively. Prevalence in siderophore uptake genes is indicative of a low iron environment as is commonly the case with open ocean environments where Fe^{3+} uptake is typically higher as compared to coastal environments where Fe^{2+} uptake is more common (Toulza *et al.*, 2012). Phosphate metabolism genes found in all 3 locations belonged to similar pathways which consist mainly of the oxidative phosphorylation pathway, and nicotinate and nicotinamide metabolism. The source of these genes was predominantly found to be *Candidatus pelagibacter*, which correlates to previous studies whereby *C. pelagibacter* was found to utilize these pathways as central metabolism for energy production in aerobic conditions (Tripp, 2007). Nitrogen metabolism genes, conversely, did not appear to vary substantially in terms of proportion throughout all 3 locations. The majority of the functional genes found in Sarawak (S2005), Sabah (S2405), and the Philippines (S2705) appeared to come largely from the SAR11 clade which, incidentally, is present in the highest numbers across all 3 sites. Similarly, the reference metagenomes also displayed an identical metabolic make-up with very little variation across the sites. This is hypothesized to be due to the similarities in the dominant taxons found within the microbial communities in these waters.

CONCLUSION

Slight variations in surface microbial community patterns were observed across the different sampling locations. This correlates to findings by Zhang *et al.* (2014), where the total bacterial communities sampled along the SCS were found to be strongly affected by environmental factors in terms of their diversity and biogeographic patterns. Sarawak (S2005) waters exhibited a more distinctive community composition and metabolism compared to the other samples as these waters were hypothesized to be subject to higher anthropogenic input in the form of the Rajang river. Further and more extensive studies must thus be carried out to obtain more definitive results in an effort to better understand prokaryotic community patterns and its relationship with spatial variation.

ACKNOWLEDGEMENT

We thank graduate student, Ching-Hung Tseng, for his help with the bioinformatics analysis carried out in this study, and Samson Lee, Juliana Ho, and Nastassia Denis for their assistance in the lab work conducted, as well as to those involved on field for the collection of samples. We are also grateful to the Sarawak Biodiversity Centre (SBC-RA-0091-MM) for their kind permission to conduct research on the collected water samples. The research leading to these results has received funding from the European Union's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 226224 - SHIVA.

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997).

Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* **25(17), 3389-3402.**

Biers, E. J., Sun, S. and Howard, E. C. (2009). Prokaryotic genomes and diversity in surface ocean waters: Interrogating the global ocean sampling metagenome. *Applied Environmental Microbiology* **75(7), 2221-9.**

Brown, D. W., Butchko, R. A., Busman, M. and Proctor, R. H. (2012). Identification of gene clusters associated with fusaric acid, fusarin, and perithecial pigment production *in Fusarium verticillioides*. *Fungal Genetic Biology* **49(7), 521-532.**

Brown, M. V., Philip, G. K., Bunge, J. A., Smith, M. C., Bissett, A., Lauro, F. M., Fuhrman, J. A. and Donachie, S. P. (2009). Microbial community structure in the North Pacific ocean. *The ISME Journal* **3(12), 1374-86.**

Chan, X. Y., Arumugam, R., Choo, S. W., Yin, W. F. and Chan, K. G. (2013) Metagenomic sequencing of prokaryotic microbiota from tropical surface seawater. *Genome Announcements* **1(4), e00540-13.**

Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam-Syed-Mohideen, A., McGarrell, D. M., Bandela, A., Cardenas, E., Garrity, G. M. and Tiedje, J. M. (2007). The ribosomal database project (RDP-II): Introducing myRDP space and quality controlled public data. *Nucleic Acids Research* **35(suppl 1), D169-D172.**

Dang, H., Li, T., Chen, M. and Huang, G. (2008). Cross-ocean distribution of Rhodobacterales bacteria as primary surface colonizers in temperate coastal marine waters. *Applied Environmental Microbiology* **74(1), 52-60.**

Das, S., Lyla, P. S. and Ajmal Khan, S. (2006). Marine microbial diversity and ecology: Importance and future perspectives. *Current Science* **90(10), 1325-1335.**

De Bourcy, C. F., De Vlaminck, I., Kanbar, J. N., Wang, J., Gawad, C. and Quake, S. R. (2014). A quantitative comparison of single-cell whole genome amplification methods. *PLoS One* **9(8), e105585.**

DeLong, E. F. and Karl, D. M. (2005). Genomic perspectives in microbial oceanography. *Nature* **437, 336-342.**

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. and Andersen, G. L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied Environmental Microbiology* **72(7), 5069-5072.**

Eilers, H., Pernthaler, J. and Amann, R. (2000). Succession of pelagic marine bacteria during enrichment: A close look at cultivation-induced shifts. *Applied and Environmental Microbiology* **66(11), 4634-4640.**

Gianoulis, T. A., Raes, J., Patel, P. V., Bjornson, R., Korbel, J. O., Letunic, I., Yamada, T., Paccanaro, A., Jensen, L. J., Snyder, M., Bork, P. and Gerstein, M. B. (2009). Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proceedings of the National Academy of Sciences* **106(5), 1374-1379.** doi: 10.1073/pnas.0808022106.

Grashoff, K., Kremling, K. and Ehrhardt, M. (1999). Methods of seawater analysis. Wiley-VCH. Weinheim, Germany. **pp 160.** doi: 10.1002/9783527613984.

Hewson, I., Paerl, R. W., Tripp, H. J., Zehr, J. P. and Karl, D. M. (2009). Metagenomic potential of microbial assemblages in the surface waters of the central Pacific Ocean tracks variability in oceanic habitat. *Limnology and Oceanography* **54(6), 1981-1994.** doi: 10.4319/lo.2009.54.6.1981.

Irz, X., Stevenson, J. R., Tanoy, A., Villarante, P. and Morissens, P. (2007). The equity and poverty impacts of aquaculture: insights from the Philippines. *Development Policy Review* **25(4), 495-516.**

Jiang, H., Dong, H., Ji, S., Ye, Y. and Wu, N. (2007). Microbial diversity in the deep marine sediments from the Qiongdongnan Basin in South China Sea. *Geomicrobiology Journal* **24(6), 505-517.**

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28(1), 27-30.**

Kuek, F. W. I., Ngu, L. H., Lim, L. F., Mujahid, A., Lim, P. T., Leaw, C. P. and Müller, M. (2016). Diversity and DMS(P)-related genes in culturable bacterial communities in Malaysian coastal waters. *Sains Malaysiana* **45(6), 915-931.**

Kuek, F. W. I., Ngu, L. H., Lim, L. F., Mujahid, A., Lim, P. T., Leaw, C. P. and Müller, M. (2015). The potential roles of bacterial communities in coral defence: A case study at Talang- talang reef. *Ocean Science Journal* **50(2), 269-282.**

Liao, L., Xu, X. W., Wang, C. S., Zhang, D. S. and Wu, M. (2009). Bacterial and archaeal communities in the surface sediment from the northern slope of the South China Sea. *Journal of Zhejiang University Science B* **10(12), 890-901.**

Liu, C., Kroeze, C., Hoekstra, A. Y. and Gerbens-Leenes, W. (2012). Past and future trends in grey water footprints of anthropogenic nitrogen and phosphorus inputs to major world rivers. *Ecological Indicators* **18, 42-49.**

Luria, C. M., Ducklow, H. W. and Amaral-Zettler, L. A. (2014). Marine bacterial, archaeal and eukaryotic diversity and community structure on the continental shelf of the western Antarctic Peninsula. *Aquatic Microbial Ecology* **73, 107-121.**

Mason, O. U., Scott, N. M., Gonzalez, A., Robbins-Pianka, A., Bælum, J., Kimbre, J., Bouskill, N. J., Prestat, E., Borglin, S., Joyner, D. C., Fortney, J. L., Jurelevicius, D., Stringfellow, W. T., Alvarez-Cohen, L., Hazen, T. C., Knight, R., Gilbert, J. A. and Jansson, J. K. (2014). Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *The ISME Journal* **8, 1464-1475.**

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A.,

Stevens, R., Wilke, A., Wilkening, J. and Edwards, R. A. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9, 386**. doi: 10.1186/1471-2105-9-386.

Moore, W. S., Sarmiento, J. L. and Key, R. (1986). Tracing the Amazon component of surface Atlantic water using 228Ra, salinity and silica. *Journal of Geophysical Research: Oceans (1978–2012)* **91(C2), 2574-2580.**

Ng, P. K. and Tan, K. S. (2000). The state of marine biodiversity in the South China Sea. *The Raffles Bulletin of Zoology* **8, 3-7.**

Nogales, B., Lanfranconi, M. P., Pina-Villalonga, J. M. and Bosch, R. (2011). Anthropogenic perturbations in marine microbial communities. *FEMS Microbiology Review* **35(2), 275-98.**

Partensky, F., Blanchot, J. and Vaulot, D. (1999). Differential distribution and ecology of *Prochlorococcus* and **Synechococcus** in oceanic waters: A review. *Bulletin-Institut Oceanographique Monaco-Numero Special-* **457-476**.

Pinhassi, J. and Berman, T. (2003). Differential growth response of colony-forming α-and γ-proteobacteria in dilution culture and nutrient addition experiments from Lake Kinneret (Israel), the Eastern Mediterranean Sea, and the Gulf of Eilat. *Applied and Environmental Microbiology* **69(1), 199-211.**

Pomeroy, L. R. and Darwin, C. (2007). The microbial loop. *Oceanography* **20(2), 28.**

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldón, T., Rattei, T., Creevey, C. and Kuhn, M. (2013). eggNOG v4. 0: Nested orthology inference across 3686 organisms. *Nucleic Acids Research* **42(D1), D231-D239.**

Robidart, J. C., Preston, C. M., Paerl, R. W., Turk, K. A., Mosier, A. C., Francis, C. A., Scholin, C. A. and Zehr, J. P. (2012). Seasonal Synechococcus and Thaumarchaeal population dynamics examined with high resolution with remote in situ instrumentation. *The ISME Journal* **6(3), 513-523.**

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J. E., Li, K., Kravitz, S., Heidelberg, J. F., Utterback, T., Rogers, Y. H., Falcon, L. I., Souza, V., Bonilla-Rosso, G., Eguiarte, L. E., Karl, D. M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M. R., Strausberg, R. L., Nealson, K., Friedman, R., Frazier, M. and Venter, J. C. (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biology* **5(3), e77.**

Signori, C. N., Thomas, F., Prast, A. E., Pollery, R. C. G. and Sievert, S. M. (2014). Microbial diversity and community structure across environmental gradients in Bransfield Strait, Western Antartic Peninsula. *Frontiers in Microbiology* **5, 647.**

Toulza, E., Tagliabue, A., Blain, S. and Piganeau, G. (2012). Analysis of the global ocean sampling (GOS) project for trends in iron uptake by surface ocean microbes. *PLoS ONE* **7(2), e30931.**

Tripp, H. J. (2007). Genomic-assisted determination of the natural nutrient requirements of the cosmopolitan marine bacterium '*Candidatus* Pelagibacter ubique'. Ph.D. Thesis. Oregon State University, Corvallis, OR.

Tseng, C. H., Chiang, P. W., Lai, H. C., Shiah, F. K., Hsu, T. C., Chen, Y. L., Wen, L. S., Tseng, C. M., Shieh, W. Y., Saeed, I., Halgamuge, S. and Tang, S. L. (2015). Prokaryotic assemblages and metagenomes in pelagic zones of the South China Sea. *BMC Genomics* **16, 219.**

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H. and Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304(5667), 66-74.**

Wang, G., Maher, E., Brennan, C., Chin, L., Leo, C., Kaur, M., Zhu, P., Rook, M., Wolfe, J. L. and Makrigiorgos, G. M. (2004). DNA amplification method tolerant to sample degradation. *Genome Research* **14(11), 2357-2366.**

Wuyts, J., Van de Peer, Y., Winkelmans, T. and De Wachter, R. (2002). The European database on small subunit ribosomal RNA. *Nucleic Acids Research* **30(1), 183-185.**

Yooseph, S., Nealson, K. H., Rusch, D. B., McCrow, J. P., Dupont, C. L., Kim, M., Johnson, J., Montgomery, R., Ferriera, S., Beeson, K., Williamson, S. J., Tovchigrechko, A., Allen, A. E., Zeigler, L. A., Sutton, G., Eisenstadt, E., Rogers, Y. H., Friedman, R., Frazier, M. and Venter, J. C. (2010). Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468(7320), 60-66.**

Zhang, R., Lau, S. C., Ki, J. S., Thiyagarajan, V., Qian, P. Y. (2009). Response of bacterioplankton community structures to hydrological conditions and anthropogenic pollution in contrasting subtropical environments. *FEMS Microbiology Ecology* **69(3), 449-460.**

Zhang, R., Liu, B., Lau, S. C., Ki, J. S. and Qian, P. Y. (2007). Particle-attached and free-living bacterial communities in a contrasting marine environment: Victoria Harbor, Hong Kong. *FEMS Microbiology Ecology* **61(3), 496-508.**

Zhang, Y., Zhao, Z., Dai, M., Jiao, N. and Herndl, G. J. (2014). Drivers shaping the diversity and biogeography of total and active bacterial communities in the South China Sea. *Molecular Ecology* **23, 2260-2274.**

Zhang, Y., Jiao, N., Sun, Z., Hu, A. and Zheng, Q. (2011). Phylogenetic diversity of bacterial communities in South China Sea mesoscale cyclonic eddy pertubations. *Research in Microbiology* **162(2011), 320-329.**

Zinger, L., Amaral-Zettler, L. A., Fuhrman, J. A., Horner-Devine, M. C., Huse, S. M., Welch, D. B., Martiny, J. B., Sogin, M., Boetius, A. and Ramette, A. (2011). Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS ONE* **6(9), e24570.**