BIODIVERSITY AND FUNCTIONAL METAGENOMIC PROFILING OF MICROBIAL COMMUNITIES IN TASIK KENYIR, TERENGGANU

MOHD EZHAR MOHD NOOR[1], SHARIFAH NOOR EMILIA SYED JAMIL FADAAK[1], MOHD NOOR MAT ISA[3], MOHD FAIZAL ABU BAKAR[3] and MUHD DANISH-DANIEL[1,2*]

*[1]School of Fisheries and Aquaculture Sciences, Universiti Malaysia Terengganu,
21030 Kuala Nerus, Terengganu, Malaysia
[2]Institute of Marine Biotechnology (IMB), Universiti Malaysia Terengganu,
21030 Kuala Nerus, Terengganu, Malaysia
[3]Malaysia Genome Institute, Jalan Bangi, 43000 Kajang, Selangor, Malaysia
E-mail: mdda@umt.edu.my

ABSTRACT

Tasik Kenyir located in the state of Terengganu is the largest artificial lake in Southeast Asia with mean-depth of 37 m. This lake plays an important role in maintaining the biodiversity in the surrounding environment. Microbial communities in the lake are important, as most of the nutrients are recycled through the "microbial loop". Thus, understanding the connection between the diversity composition and functional role of aquatic microbial community is crucial for proper lake management. This study aims to determine the diversity and functions of microbial assemblages in Tasik Kenyir by means of shotgun metagenomics analysis. Briefly, water samples were collected from pristine and disturbed areas. Metagenome DNA were then extracted directly and subjected to clone-independent sequencing. Data sequences of all samples were analyzed and functional annotated using bioinformatics software MEGAN6. Analysis showed up to 41 phyla that had been detected from the water samples with the presence of dominant bacterial populations of more than 90% in all samples. Proteobacteria was the most dominant phylum, representing more than 70% of the microbiome in all samples. Other taxa such as Bacteriodetes, Terrabacteria group, Verrucomicrobia, Planctomycetes and Chloroflexi were also found as part of the microbial communities. The first sample from the disturbed area, TKSA1 had 3% of total contigs read assigned to genera *Pseudomonas* while the other samples appeared to be more homogeneous. The lake also appeared to contain a mixture of autotrophs and heterotrophs capable of performing main biogeochemical cycles. Findings of the present study has provided us valuable information on the microbial diversity structures and their functions in the nutrient processing pathways that occurred in the Tasik Kenyir environment and thus sheds light on the importance of freshwater microbial communities for ecosystem and human health.

Key words: Metagenomic, biodiversity, functional profiling, microbial communities, tasik Kenyir

INTRODUCTION

Tasik Kenyir (TK) is the largest artificial lake in South East Asia and also is one of the most important freshwaters in Malaysia. It serves as a natural ecological habitat and provides rich environment resources to flora and fauna. Some information had been documented by previous studies, which have shown the importance of studying the lake ecosystem and some suggested that lakes can be considered as systems of biological activity (Battin *et al*., 2009), and are central to many biogeochemical processes (Tranvik *et al*., 2009), and all biogeochemical cycles on biosphere involved with the microbes and major redox reactions such as the carbon cycle is catalysed by a set of key microbial enzymes (Falkowski *et al*., 2008). Microbes can be found almost everywhere in the biosphere as they play many important roles in the ecosystem. This claim was supported by a previous study where microbes are widely known for their part in controlling the ecosystem function and biogeochemical cycling in the biosphere (Szabo-Taylor *et al*., 2010). Microbes are known to perform many critical biogeochemical processes and thus also have the potential to modify and control water

---

* To whom correspondence should be addressed.

quality in these ecosystems. Despite some previous studies on microbes in the freshwater lakes environment using metagenomic analysis like in Lake Gatun (Rusch *et al.*, 2007), Lac du Bourget (Debroas *et al.*, 2009) and Lake Lanier (Oh *et al.*, 2011) which have provided some information of the functional diversity of freshwater bacterioplankton in single lake ecosystems, we still have limited understanding of their functional potential, genetic variability and community interactions.

In an earlier microbial diversity study, Amann *et al.* (1995) had suggested that approximately less than 1% of earth's microorganisms are cultivable in the laboratory. Later on, SILVA database version 1.15 had the existence of at least 63 bacterial phyla and candidate groups as reported by Pruesse *et al.* (2007) and Quast *et al.* (2013). These reported studies have given a better insight into the undiscovered genetic potential concealed by microbes that evolved in billions of years. Along with world modernization, new technologies in culture independent approach have enabled researchers to analyze in detail about microbial communities and its functions in the natural environment. Next-generation sequencing (NGS) metagenomics has become a powerful tool to investigate the biodiversity of complex microbial communities and for studying its metabolic pathways. The new generation of this sequencing technology, with its ability to sequence possibly thousands of organisms in parallel, has proved to be uniquely suited to this application. Metagenomics can be used to study the whole microbial community without the need to adopt the culture techniques. According to Schloss and Handelsman, (2003), metagenomics was built on advances in microbial genomics, polymerase chain reaction (PCR) amplification and cloning of genes. The field of metagenomics has played a pivotal role for significant progress in microbial ecology, evolution, and diversity over the past five to ten years. This development has opened new ways of understanding microbial diversity and functions in the environment. This study is important because freshwater lakes are underrepresented among the growing numbers of environmental metagenomics datasets. Hence, the characterization of the microbial community of Tasik Kenyir has the potential to determine the microbial diversity structures and their metabolic functions in the Tasik Kenyir environment as shown in several lake studies by Rusch *et al.* (2007) and Oh *et al.* (2011).

MATERIALS AND METHODS

Sampling Sites and Sample Collection

The studied area was located in the state of Terengganu, which is on the east coast of Malaysia. Water samples were collected three times (January, June October) at specific site throughout the year of 2015 from Sungai Como area, (aquaculture activities area; N05°01.887 E102°50.600) and Sungai Lasir area, (pristine area; N04°59.997 E102°50.133). A brief water parameters like dissolved oxygen, pH and water temperature in the lake were measured on-site. A horizontal water sampler was used to collect approximately five liters of water samples at a depth of 5 meters from the surface, which represents a well-oxygenated, highly productive layer of the water column. All collected water samples were kept in sterile bottles and maintained at 4°C to keep the samples fresh before being transported back to the laboratory for further analysis. About 0.5 liters to 5 liters of water were filtered in the analysis laboratory through a vacuum filter cartridge Sterifil® Aseptic System and Holder (Merck, Germany), with a 0.22 µm pore size nitrocellulose microbiology grade. Membrane filters (Merck, Germany) were aseptically placed on the filter holder. The filtered membrane filters for all samples were stored at -80°C until nucleic acid extraction.

Metagenomic Analysis

Nucleic acid extractions were performed by using Metagenome DNA isolation kit (Epicentre, Wisconsin, USA) with minor modifications on the sample quantity used in the cell lysis procedure. Whole shotgun metagenome sequencing was carried out on Illumina HiSeq 2500 (San Diego, CA, USA). The generated multi-million paired-end sequencing reads were then filtered and trimmed with the built-in program in SolexaQA software, DynamicTrim based on base quality and LengthSort based on sequence length (Cox *et al.*, 2010). De novo assembly of good-quality reads into contiguous sequences (contigs) representing DNA fragments in the metagenome was performed using Metagenomic Assembler program, MetaVelvet (Namiki *et al.*, 2012). Open reading frames (ORFs) in the assembled contigs were predicted using Prodigal gene prediction software (Hyatt *et al.*, 2012). Functions of the predicted proteins similarity searches was performed against the NCBI GenBank non-redundant (nr) protein sequence database to

identify the best hits for each gene. The similarity search results were analyzed using MEGAN6 (Huson *et al*., 2016) by assigning BLAST results to NCBI taxonomies with the default Lowest Common Ancestor (LCA) algorithm parameters.

RESULTS

Whole Genome Shotgun Metagenome Analysis

The present study datasets has provided enormous data for taxonomy profiling compared to other techniques. More than 850,000 contigs read from all six samples were trimmed and assembled. MEGAN6 (Huson *et al*., 2016) software was primarily used to display and analyze the data. The detailed information regarding the metagenomic libraries is summarized in Table 1. Phylogenetic analysis for all samples revealed that the overwhelming majority of the phyla abundance recovered in the Tasik Kenyir data set were bacterial. Overall, three samples from the disturbed area, TKSA1, TKSB1 and TKSC1 combined had 80.1% of the predicted ORFs, which belonged to bacteria, and 0.06% and 0.69% were from Archaea and Eukaryota, respectively. On the other hand, three samples from the pristine area TKSX1, TKSY1 and TKSZ1 combined had 81.1% of the predicted ORFs, which belonged to bacteria, and 0.1% and 0.3% were from Archaea and Eukaryota, respectively. Another fraction of the total reads belonged to viruses and unclassied sequences. A total of 41 phyla (Figure 1) were identied from all samples, and the top 10 phyla included (in decreasing order of prevalence) Proteobacteria, Verrucomicrobia, Cyanobacteria, Bacteroidetes, Actinobacteria, Planctomycetes, Firmicutes, Armatimonadetes, Chloroflexi and Acidobacteria.

Microbial Functional Gene Diversity

A total of 1,346,841 full-length protein-coding genes identied within the shotgun metagenome dataset were analyzed. Among these, 1,093,267 ORFs were annotated based on the closest match in the GenBank NR protein database. Using the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al*., 2004) and SEED database (Overbeek *et al*., 2005) function of MEGAN6 (Figure 2 and Figure 3), the Open Reading Frames (ORFs) were classied according to their putative functions. Sequence affiliations were used to further understand the relationship between the geochemical parameters and the population diversity within Tasik Kenyir. Interestingly, about 30% of genes in all samples in the present study constituted of the new putative genes and hypothetical proteins with unknown functions. Data from this study also revealed that only less than 50% genes read showed the result when blasted against NCBI GenBank database. This is because all available genes in the database mainly revolved around the pathogens towards human.

DISCUSSION

Environmental samples are expected to have a diverse genetic diversity. But in Malaysia, many previous studies on ecology microbes had been based on culture-dependant approach. Only some studies on microbes and their function using metagenomic analysis like study on microbes (thermophiles) in freshwater by Chan *et al*. (2015) was reported. This situation has led to the lack of study on ecology of microbes especially in freshwater in Malaysia. Approximately 3% of the total ORFs were assigned to genera *Pseudomonas* specifically in the disturbed area sample TKSA1. This data indicated that the impact of aquaculture activities had changed the composition of microbial community at the disturbed area in the early sampling (January). Nevertheless, microbial communities in the disturbed area appeared to be increasing over the sampling timeline (January, June and October) meanwhile the microbial communities in the pristine area were more homogeneous throughout the sampling timeline. This may be due to aquaculture activities occurring at that area had

Table 1. Characteristics of shotgun metagenomic libraries and phylogenetic analysis reads for all samples

| Characteristic | Sample designation | | | | | |
| | Disturbed area (Aquaculture) | | | Pristine area | | |
Sample Name	TKSA1	TKSB1	TKSC1	TKSX1	TKSY1	TKSZ1
Contigs Reads	29,593	160,657	345,696	48,433	62,724	206,929
Gene Reads	66,362	276,147	448,979	106,097	148,084	301,172
Species Detected	4,153	41,691	55,421	32,421	43,753	51,507
Bacteria (Phyla)	4	17	24	19	22	23
Archaea (Phyla)	0	1	2	2	3	2
Eukarya (Phyla)	1	6	14	13	7	7
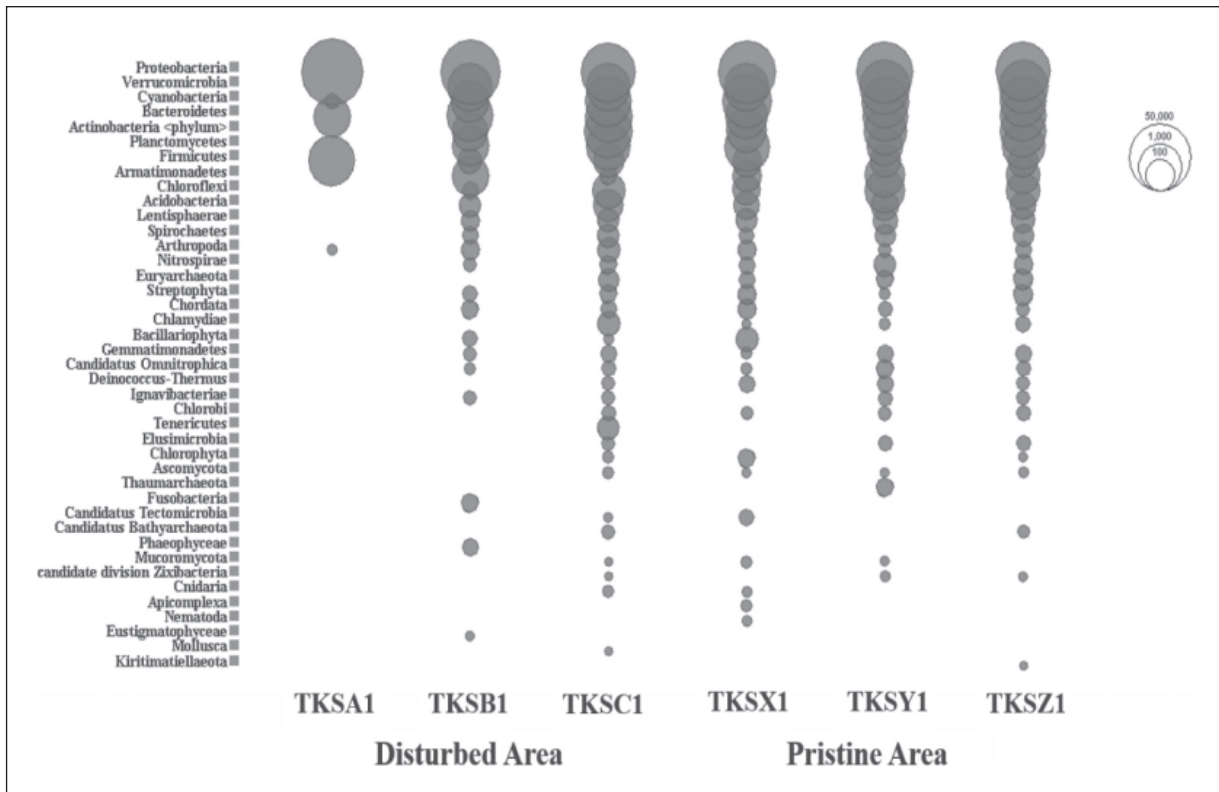
Fig. 1. Log scale value of phyla abundance of microbial community detected from the metagenome sequences of all samples.
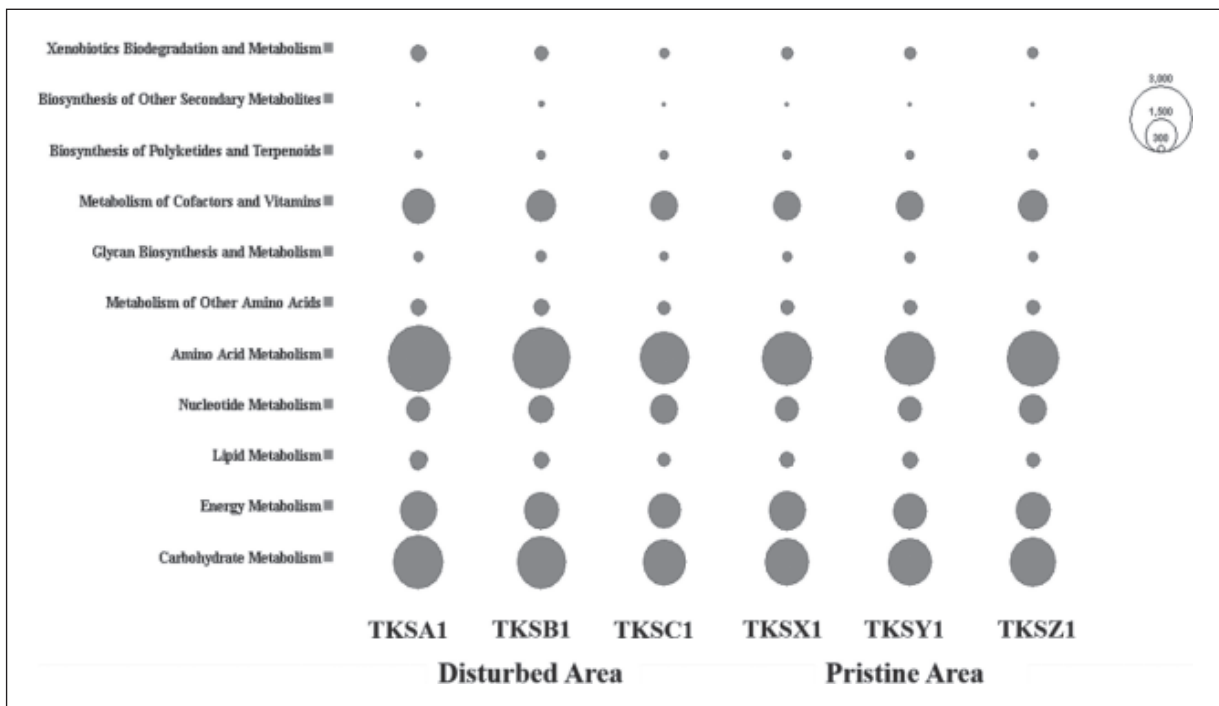

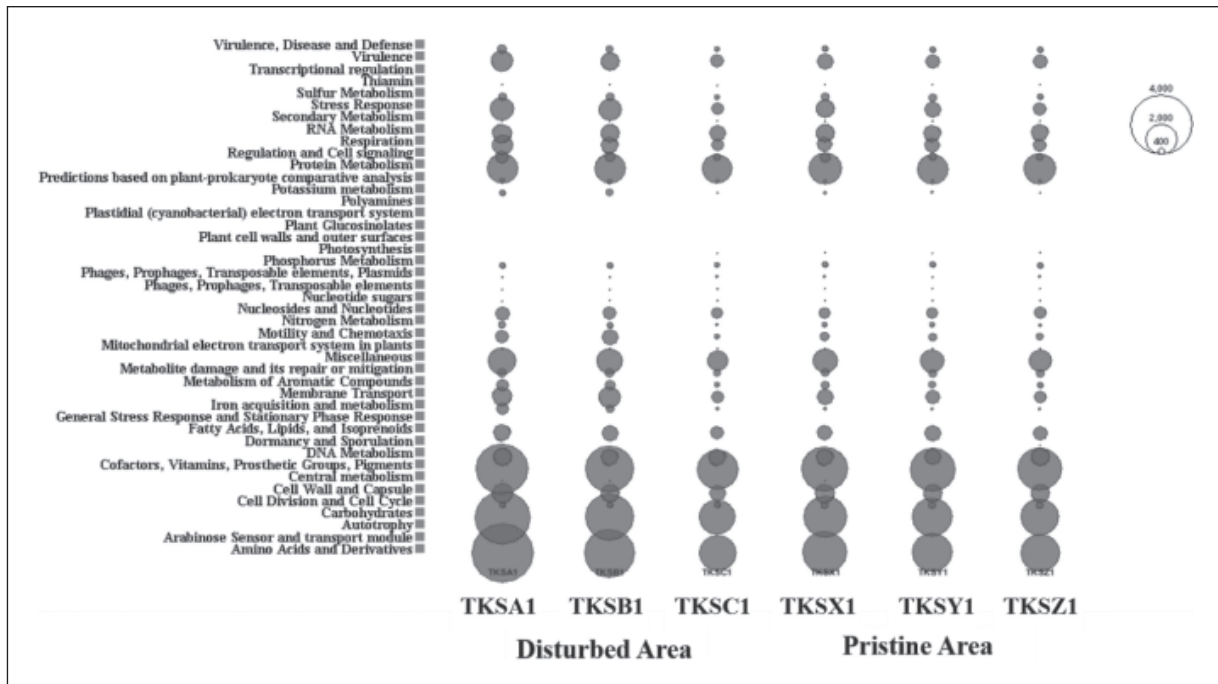**Fig. 2.** Number of predicted ORFs that matched metabolic categories, based on KEGG analysis.

Fig. 3. The number of predicted ORFs that matched metabolic categories, based on SEED subsystem.

affected the composition of the microbes. It also can be found out that although aquaculture had affected the microbes but the communities seemed to be generally recovered after a certain period of time. This was shown as the third sampling data from the disturbed area TKSC1 is comparable to the data from the pristine area. This may be due to the stream of the lake flow from the pristine area all the way down to disturbed area. Hence, introducing the original community of the microbes into the affected area. This diversity recovery is crucial because the in findings by Kennedy (1999), it was mentioned that the diversity of microbes is critical to the functioning of the ecosystem, because there is the need to maintain ecological processes like controlling pathogens within the ecosystems. Furthermore, study by Yamanaka *et al.* (2003) also stated that microbial diversity is linked to the ecosystem stability. All samples from both sites had been predominated by Proteobacteria with more than 70% of the total reads and samples from disturbed area specifically TKSA1 had been dominated by the known pathogens like *Pseudomonas* species and *Aeromonas* species. All major bacterial groups commonly found in freshwater ecosystems that has been reported by Humbert *et al.* (2009) were present in Tasik Kenyir. This study data also is consistent with findings by Kersters *et al.* (2006), which reported that Proteobacteria, the largest and most phenotypically diverse phylum, accounted for atleast 40% of all known genera. In addition, present study, phyla in common to all samples can be found

in a variety of aquatic environments similar to the finding by Dillon *et al.* (2009).

From the datasets, the reactions involved in carbon metabolism were carbon fixation and methane metabolism and the key enzymes also were shown to be involved in autotrophic and heterotrophic carbon fixation pathways. Carbon cycling is a key metabolic feature of members of the alpha and gamma Proteobacteria, including *Alteromonadaceae, Pseudomonadaceae, Sphingo-monadaceae*, and *Vibrionaceae* as well as *Flavo-bacteriaceae*, which is a member of *Cytophaga-Flavobacteria-Bacteroides* (CFB) clade. On the other hand, the nitrogen cycle is known to be a complex biological process and it requires interplay of many microorganisms. Shotgun metagenome datasets in this study revealed the presence of key enzymes such as ammonia monooxygenase and nitrate reductase involved with complex nitrogen cycle. This indicates that conversion of nitrogen related compounds such as nitrate into nitrite, occurred in the Tasik Kenyir communities. Similar to the earlier studies on the linkage between the bacteria and nitrogen cycling by Cole (1996), Chou *et al.* (2008) and Rajakumar *et al.* (2008) that had showed that the bacteria involved were from members of the *Comamonadaceae*, *Entero-bacteriaceae* and *Pseudomonadaceae*. It can be said that in this study the bacterial community is absolutely dominant over archaea and eukarya in the freshwater samples. Tasik Kenyir metagenomic data sets can provide some useful data analysis on how

natural microbial population changes when being impacted by human activities (aquaculture site). Hence, it enabled researchers to identify some of the crucial microbial players and its functions that could be the targets for future work. It also must be addressed that metagenomic data only discovers the functional potential of microbes in the environment and may need the combination of meta-transcriptomic and metaproteomic analysis to facilitate the study of environmental biology. Such studies will significantly advance the understanding of the importance of microbial communities for the ecosystem and human health especially in the freshwater environment.

ACKNOWLEDGEMENTS

REFERENCES

Amann, R.I., Ludwig, W. & Schleifer, K.H. 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiology and Molecular Biology Reviews*, **59(1)**: 143-169.

Battin, T.J., Luyssaert, S., Kaplan, L.A., Aufdenkampe, A.K., Richter, A. & Tranvik, L.J. 2009. The boundless carbon cycle. *Nature Geoscience*, **2(9)**: 598.

Chan, C.S., Chan, K.G., Tay, Y.L., Chua, Y.H. & Goh, K.M. 2015. Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing. *Frontiers Microbiology*, **6**: 177.

Chou, Y.J., Chou, J.H., Lin, K.Y., Lin, M.C., Wei, Y.H., Arun, A.B., Young, C.C. & Chen, W.M. 2008. *Rothia terrae* sp. nov. isolated from soil in Taiwan. *Internatinal Journal of Systematic and Evolutionary Microbiol*ogy, **58**: 84-88.

Cole, J. 1996. Nitrate reduction to ammonia by enteric bacteria: redundancy, or a strategy for survival during oxygen starvation? *FEMS Microbiology Letters,* **136**: 1-11.

Cox, M.P., Peterson, D.A. & Biggs, P.J. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, **11(1)**: 485.

Debroas, D., Humbert, J.F., Enault, F., Bronner, G., Faubladier, M. & Cornillot, E. 2009. Meta-genomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget - France). *Environmental Microbiology*, **11(9)**: 2412-2424.

Dillon, J.G., McMath, L.M. & Trout, A.L. 2009. Seasonal changes in bacterial diversity in the Salton Sea. *Hydrobiologia*, **632**: 49-56.

Falkowski, P.G., Fenchel, T. & Delong, E.F. 2008. The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, **320(5879)**: 1034-1039.

Humbert, J.F., Dorigo, U., Cecchi, P., Le Berre, B., Debroas, D. & Bouvy, M. 2009. Comparison of the structure and composition of bacterial communities from temperate and tropical freshwater ecosystems. *Environmental. Microbiol*ogy, **11**: 2339-2350.

Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S. & Tappu, R. 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology*, **12(6)**: 1-12.

Hyatt, D., Locascio, P.F., Hauser, L.J. & Uberbacher, E.C. 2012. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, **28(17)**: 2223-2230.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Research,* **32**: 277-280.

Kennedy, A.C. 1999. Bacterial diversity in agro-ecosystems. *Agriculture, Ecosystems & Environment*, **74(1)**: 65-76.

Kersters, K., De Vos, P., Gillis, M., Swings, J. Vandamme, P. & Stackebrandt, E. 2006. Introduction to the Proteobacteria. In: The Prokaryotes. (Eds.: Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K. H. and Stackebrandt, E.) Springer, New York, p. 3-37.

Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. 2012. MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, **40(20)**: e155.

Oh, S., Caro-Quintero, A., Tsementzi, D., DeLeon-Rodriguez, N., Luo, C., Poretsky, R. & Konstantinidis, K.T. 2011. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Applied and Environmental Microbiology*, **77(17)**: 6000-6011.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R. & Fonstein, M. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, **33**: 5691-5702.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B., Ludwig, W., Peplies, J. & Glöckner, F.O. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data. *Nucleic Acids Research,* **35(21)**: 7188-7196.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P. & Glöckner, F.O. 2012. The SILVA ribosomal RNA gene database project: imporved data processing and web-based tools. *Nucleic Acids Research*, **41(D1)**: D590-D596.

Rajakumar, S., Ayyasamy, P.M., Shanthi, K., Thavamani, P., Velmurugan, P., Song, Y.C. & Lakshmanaperumalsamy, P. 2008. Nitrate removal efficiency of bacterial consortium (*Pseudomonas* sp. KW1 and *Bacillus* sp. YW4) in synthetic nitrate-rich water. *Journal of Hazardous Materials*, **157**: 553-563.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S. & Venter, J.C. 2007. The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, **5(3)**: 0398-0431.

Schloss, P.D. & Handelsman, J. 2003. Biotechnological prospects from metagenomics. *Current Opinion in Biotechnology*, **14(3)**: 303-310.

Szabó-Taylor, K.É., Kiss, K.T., Logares, R., Eiler, A., Ács, É., Tóth, B. & Bertilsson, S. 2010. Composition and dynamics of microeukaryote communities in the River Danube. *Fottea*, **10(1)**: 99-113.

Tranvik, L.J., Downing, J.A., Cotner, J.B., Loiselle, S.A., Striegl, R.G., Ballatore, T.J., Dillon, P., Finlay, K., Fortino, K., Knoll, L.B. & Kortelainen, P.L. 2009. Lakes and reservoirs as regulators of carbon cycling and climate. *Limnology and Oceanography*, **54(6part2)**: 2298-2314.

Yamanaka, T., Helgeland, L., Farstad, I.N., Fukushima, H., Midtvedt, T. & Brandtzaeg, P. 2003. Microbial colonization drives lymphocyte accumulation and differentiation in the follicle-associated epithelium of Peyer's patches. *Journal of Immunology*, **170(2)**: 816-822.