Oxidation Communications 39, No 2, 1385–1391 (2016)

Biological and biochemical oxidation processes

APPLICATION OF SECOND-GENERATION HIGH-THROUGHPUT SEQUENCING BASED ON MiSeq SEQUENCER TO THE STUDY OF DIATOM SPECIES DIVERSITY OF WATER SAMPLES

JIAN-QIANG DENG[a*#], BAO-QIN LIU[b#], YING WANG[c], WEI LIU[d], JI-FENG CAI[e], REN LONG[a], ANWEN LIANG[a], JIBIN YIN[a]

[a]*Department of Forensic Medicine, Hainan Medical University, Haikou, China*
E-mail: *dengjianqiang2090@163.com*
[b]*Centre of Reproductive Medicine, Affiliated Hospital of Hainan Medical University, Haikou, China*
[c]*Department of Bio-chemistry and Molecular Biology, Hainan Medical University, Haikou, China*
[d]*Department of Medical Information Technology, Hainan Medical University, Haikou, China*
[e]*Department of Forensic Medicine, Xiangya Medical College of Central-South University, Changsha, China*

ABSTRACT

The aim of the paper was to investigate the feasibility of using high-throughput sequencing technique based on MiSeq sequencer in order to study the diatom species diversity of lake water samples. Lake water samples in 5 different positions of Haikou City of Hainan Province, China were randomly collected, and diatom DNA was extracted. The MiSeq second-generation sequencer was used to characterise the region V4 of the 18s rDNA with 512F/978R primer. The original sequence data were analysed by using QIIME Software, and the sequences were aligned against Silva database as the reference database. Species analysis was done by the criterion of sequence similarity ≥ 99%. The indicators detected included species classification, diversity and abundance. A total of 607 947 reads were obtained from 5 water samples with a coverage reaching over 0.999. Through alignment against Silva database, 538 species were identified and the sequences of many new species were obtained. Meanwhile, the data of chao1, ACE, Shannon index, Simpson index and PD_whole_tree were also acquired. MiSeq has the advantages of high throughput, fastness and convenience.

Keywords: diatom, species diversity, high throughput sequencing, 18s rDNA, lake water.

* For correspondence (# means co-first authors).

AIMS AND BACKGROUND

Algae are located at the beginning of food chain in a river ecosystem. Algal diversity is one major ecological feature of the algal population that reflects the response of an algal population to the changes of habitats. As primer producers, algae have a short life cycle and high sensitivity to pollutants, and algal composition varies from one water body to another. The properties and size of the population also change with the chemical composition of water. Therefore, algal features can be used for water quality monitoring and evaluation. Diatoms are autotrophic eukaryotic algae that can survive in most water bodies with great biodiversity. They are very sensitive to the changes of water temperature, pH value, electrical conductivity and concentration of nutrient salts. At present, diatoms are widely applied to the monitoring of water quality based on trophic status, acidification and pollutants of the water body[1].

Diatoms are commonly studied by morphological classification using indicators that are susceptible to environmental variations. However, species that are phylogenetically close vary little in terms of morphology or only differ in certain stages of their life history. Thus morphological classification alone can hardly differentiate between species or genera very accurately. The emerging molecular identification and classification techniques targeting at the differences in gene sequences provide new solutions. The sequencing of 18s rDNA, cytochrome c oxidase (COI) and internal transcribed spacer (ITS) genes has already been applied to diatom diversity researches globally with fastness and accuracy. Many attempts are made to establish the methods for rapid molecular identification of diatoms based on genetic features[2].

DNA sequencing is the molecular basis for species identification and considered the 'golden standard'. However, conventional sequencing techniques are unable to obtain all diatom sequences of the water samples in a specific location. This is because the water body usually contains many unknown diatom species and it is difficult to analyse the abundance of all diatom species[3]. Second-generation high-throughput sequencing has greatly reduced the cost, and Illumina MiSeq is the representative small-scale desktop sequencer advanced in February 2011 (Ref. 4). Integrating cluster generation and data generation, MiSeq is fast and accurate, which overcomes the defects of low throughput and complicated operations. Compared with Roche 454 sequencing based on pyrophosphoric acid principle, MiSeq can achieve synthesis and sequencing simultaneously, and the sequencing results are reliable. Miseq platform combines the advantages of Roche 454 and Illumina HiSeq 2500 and can sequence several variable regions in one run at faster speed and higher throughput[5–7]. Using the Miseq platform, the researchers can obtain satisfactory sequence data within 8 hours from the starting samples. Amplification, sequencing and data analysis are performed on a single Miseq platform, and the maximum amount of data acquired is 8Gb. Miseq is suitable for small-scale genome sequencing[8], and the microbial strains successfully sequenced include *Lactobacillus casei*[9–11], influenza A and influenza B (Refs 12–15). Although MiSeq has been recognised for its potential in microbial

community researches, MiSeq is rarely used in the research of algal diversity of local water bodies[16–18]. We performed sequencing of the region V4 of the 16s rDNA gene in diatoms using MiSeq and discussed the feasibility of MiSeq in this respect.

EXPERIMENTAL

Samples. Water samples were randomly collected from lakes at 5 different positions of Haikou City and detected at the laboratory.

DNA extraction. MoBio PowerSoil DNA Isolation Kit (ANBIOSCI TECH LTD.) was used for DNA extraction according to the manufacturer manual.

Sequencing. PCR amplifications were conducted in with the 512F/978R primer set that amplified the V4 region of the 18S rDNA gene. Sequencing was conducted on an Illumina MiSeq platform.

DNA detection: DNA concentration and purity were monitored on 1.5% agarose gels.

Amplicon generation: Primer 18S V3–V4: 512F-978R.16SrRNA genes were amplified using the specific primer with the barcode. All PCR reactions were carried out in 50 μl reactions with 38.8 μl of ddH2O; 5μl of 10×Buffer A; 1 μl 10 mM dNTP; 2 μl reverse primers; 0.2 μl of KAPA Taq and 1 μl template DNA. Thermal cycling consisted of initial denaturation at 95°C for 3 min, followed by 30 cycles of denaturation at 95°C for 30 s, annealing at 50°C for 30 s, elongation at 72°C for 60 s, and finally at 72°C for 7 min.

PCR products. Mix the same volume of 1X loading buffer (contained synaptobrevin green) with PCR products and perform electrophoresis on 1.5% agarose gel. Samples with bright main strip between 400–450 bp were chosen for further experiments. Polymerase chain reaction (PCR) products were mixed in equidensity ratios. Then, the mixture PCR products were purified with GeneJET Gel Extraction Kit (QIAGEN).

Library preparation and sequencing. Sequencing libraries were generated using NEXTflex™ Rapid DNA-seq kit for illumian (BIOO SCIENTIFIC, USA) following the manufacturer recommendations and index codes were added. The library quality was assessed on the Qubit® 2.0 Fluorometer (Thermo Scientific) and Agilent Bioanalyser 2100 system. At last, the library was sequenced on an Illumina MiSeq platform and 250 bp paired-end reads were generated.

DATA ANALYSIS

Paired-end reads from the original DNA fragments were merged by using FLASH-a very fast and accurate analysis tool which was designed to merge paired-end reads when there were overlaps between reads1 and reads2. Paired-end reads were assigned to each sample according to the unique barcodes. Sequences were analysed using QIIME software package (Quantitative Insights Into Microbial Ecology), and

in-house Perl scripts were used to analyse alpha- (within samples) and beta- (among samples) diversity. First, reads were filtered by QIIME quality filters. Then we used pick_de_novo_otus.py to pick operational taxonomic units (OTUs) by making operational taxonomic unit (OUT) table. Sequences with ≥ 99% similarity were assigned to the same OTUs. We picked a representative sequences for each OTU and used the ribosomal database project (RDP) classifier to annotate taxonomic information for each representative sequence. In order to compute Alpha Diversity, we rarified the OTU table and calculated three metrics: Chao1 estimated the species abundance; Observed Species estimated the amount of unique OTUs found in each sample, and Shannon index. Rarefaction curves were generated based on these three metrics.

RESULTS AND DISCUSSION

As shown by the rarefaction curves of all samples, the curves tended to parallel to the *X* axis, indicating the ample size of the samples.

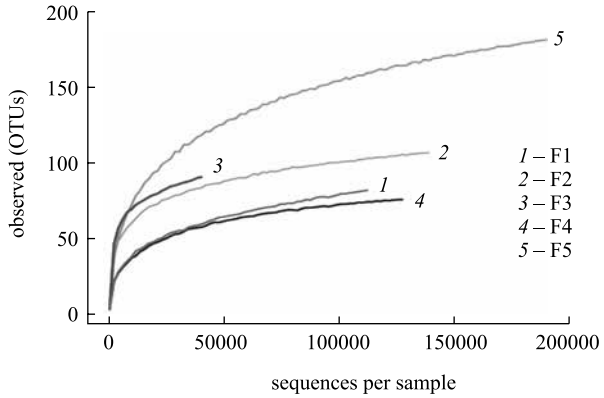

Fig. 1. Rarefaction curves of all samples

A total of 909 017 reads were obtained from 5 samples. They were aligned against Silva database. Through clustering by the criterion of sequence similarity above 99%, 607 947 sequences were selected for subsequent classification (Table 1).

Table 1. Reads obtained by high-throughput sequencing using MiSeq

Samples	Sequence numbers (original)	Sequence numbers (OTUs)	Average length
F1	203166	112068	436.45
F2	188970	138563	434.54
F3	150651	40403	438.72
F4	150594	128000	436.42
F5	215636	188913	435.94
Total	909017	607947	–

Species classification was done according to Silva reference taxonomy across 6 levels, which were kingdom, phylum, class, order, family, genus and species successively. For each sample and each level, the abundance of the sequences was calculated so as to construct the abundance matrix. As a result, 538 known algal species were identified on the species level (Table 2), and all remaining species were unknown algae.

Table 2. Alpha diversity of each sample

Samples	Goods_coverage	Chao1	ACE	Shannon index	Simpson index	PD_whole_tree	Observed_species
F1	0.999863	124.1	124.4227	2.691069	0.693318	0.40733	107
F2	0.999891	83.58333	86.26518	1.630827	0.529765	0.28130	76
F3	0.999767	241.125	226.5104	1.050462	0.231298	0.32307	182
F4	0.999579	104.6	117.2848	2.023669	0.577273	0.36853	91
F5	0.999777	112	119.4002	2.045990	0.673800	0.16444	82

Diversity on the species level was compared between the 5 samples. It was found that the species composition varied greatly in different water bodies and each water body had a unique algal composition.

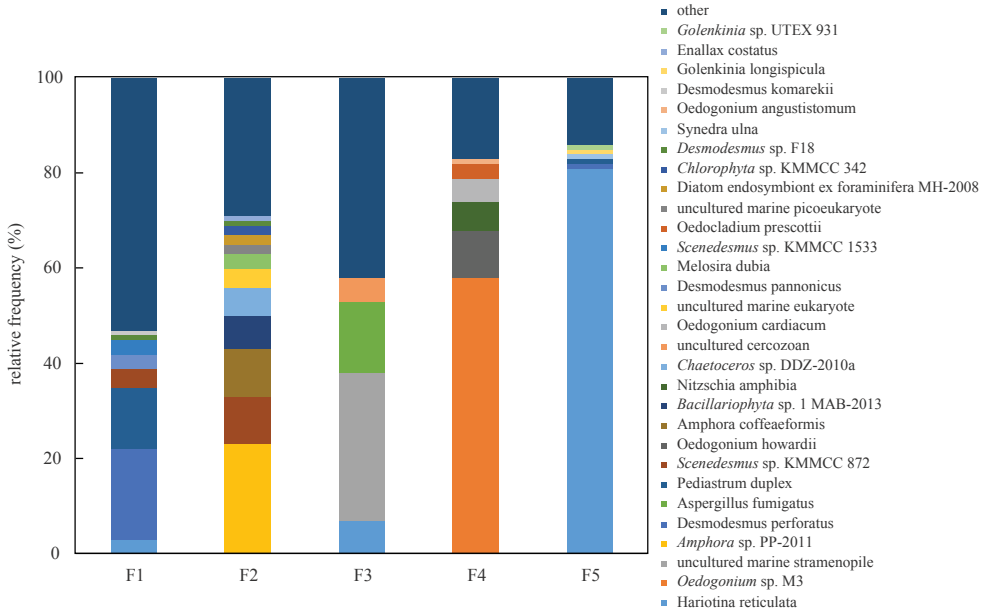

Fig. 2. Species diversity of each sample

Diatoms occupy a significant position in river health evaluation, and species classification and identification are the first step. As many diatoms species have disappeared or newly emerged throughout the world, we need to keep close track of the latest trends of the growth and proliferation of the diatoms and constantly update the

diatom species library. The conventional morphological classification can no longer satisfy the demands, and first-generation sequencing technique combined with DNA fragment cloning is unable to detect all diatom species in lakes or sea based on numerous amplifications. Thus, the second-generation high-throughput sequencing is invented, which overcomes these limitations.

We applied MiSeq to the study of the diatom species diversity of the lakes. After DNA extraction from water samples in 5 different positions, 607 947 reads were obtained with coverage exceeding 0.999. By alignment against Silva database[1], 538 species were identified. This is otherwise unconceivable using conventional morphological classification. Besides the features of fastness and high efficiency, MiSeq is a small-sized portal device. The output of MiSeq not only includes fragment coverage and species classification, but also the indicators of population analysis such as chao1, ACE, Shannon index, Simpson index and PD_whole_tree. All these indicators are desired in ecological survey. The detection results varied from one lake to another, and a large number of sequences could not be aligned to any known sequences in the reference databases, indicating that many species were newly discovered. The above results fully demonstrated the fastness and convenience of MiSeq in diatom survey and its ability to discover new species.

CONCLUSIONS

MiSeq is a new method for analysing diatoms in nature with the advantages of high throughput, fastness and convenience. However, due to the lack of experimental studies using MiSeq so far, it is necessary to carry out special design in light of the specific purpose before the diatom survey. This is important to give a full play to the strong points of MiSeq.

ACKNOWLEDGEMENTS

This paper is supported by the Natural Science Foundation of China (No 81260465, 81560304) and the Scientific Research Cultivation Fund of Hainan Medical University (#HY2012-012).

REFERENCES

1. E. SZABÓ, G. L. ZÜGNER, M. FARKAS, I. SZILÁGYI, S. DÓBÉ: Direct Kinetic Study of the OH-radical Initiated Oxidation of Pivalaldehyde, $(CH_3)_3CC(O)H$, in the Gas Phase. Oxid Commun, **35** (3), 538 (2012).
2. Y. LIANG, P. H. CHU, X. K. WANG: Health-related Quality of Life of Chinese Earthquake Survivors: A Case Study of Five Hard-hit Disaster Counties in Sichuan. Soc Indic Res, **119** (2), 943 (2014).
3. J. ZIMMERMANN, G. GLÖCKNER, R. JAHN, N. ENKE, B. GEMEINHOLZER: Metabarcoding vs. Morphological Identification to Assess Diatom Diversity in Environmental Studies. Mol Ecol Resour, **15** (3), 526 (2015).

4. Y. LIANG, W. WU: Exploratory Analysis of Health-related Quality of Life among the Empty-nest Elderly in Rural China: An Empirical Study in Three Economically Developed Cities in Eastern China. Health Qual Life Out, **12** (59), 1 (2014).
5. L. GUO, Z. SUI, S. ZHANG, Y. REN, Y. LIU: Comparison of Potential Diatom 'Barcode' Genes (the 18S rRNA Gene and ITS, COI, rbcL) and Their Effectiveness in Discriminating and Determining Species Taxonomy in the Bacillariophyta. Int J Syst Evol Microbiol, **65** (4), 1369 (2015).
6. Y. LIANG, W. Y. LU, W. WU: Are Social Security Policies for Chinese Landless Farmers Really Effective on Health in the Process of Chinese Rapid Urbanization? A Study on the Effect of Social Security Policies for Chinese Landless Farmers on Their Health-related Quality of Life. Int J Equity Health, **13** (5), 1 (2014).
7. S. T. WILLIAMS, P. G. FOSTER, D. T. LITTLEWOOD: The Complete Mitochondrial Genome of a Turbinid Vetigastropod from MiSeq Illumina Sequencing of Genomic DNA and Steps towards a Resolved Gastropod Phylogeny. Gene, **533** (1), 38 (2014).
8. Y. LIANG, R. X. CAO: Employment Assistance Policies of Chinese Government Play Positive Roles! The Impact of Post-earthquake Employment Assistance Policies on the Health-related Quality of Life of Chinese Earthquake Populations. Soc Indic Res, **120** (3), 835 (2015).
9. N. J. LOMAN, R. V. MISRA, T. J. DALLMAN, C. CONSTANTINIDOU, S. E. GHARBIA, J. WAIN, M. J. PALLEN: Performance Comparison of Benchtop High-throughput Sequencing Platforms. Nat Biotechnol, **30** (5), 434 (2012).
10. Y. LIANG, M. L. GUO: Utilization of Health Services and Health-related Quality of Life Research of Rural-to-Urban Migrants in China: A Cross-sectional Analysis. Soc Indic Res, **120** (1), 277 (2015).
11. W. RUTVISUTTINUNT, P. CHINNAWIROTPISAN: Simultaneous and Complete Genome Sequencing of Influenza A and B with High Coverage by Illumina MiSeq Platform. J Virol Methods, **193** (2), 394 (2013).
12. L. SU, H. W. WANG, J. W. MIAO, Y. LIANG: Clinicopathological Significance and Potential Drug Target of CDKN2A/p16 in Endometrial Carcinoma. Sci Rep, **5**, 13238 (2015).
13. C. QUAST, E. PRUESSE, P. YILMAZ, J. GERKEN: The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-based Tools. Nucleic Acids Res, **41**, 590 (2013).
14. Y. LIANG, X. K. WANG: Developing a New Perspective to Study the Health of Survivors of Sichuan Earthquakes in China: A Study on the Effect of Post-earthquake Rescue Policies on Survivors' Health-related Quality of Life. Health Res Policy Syst, **11** (41), 1 (2013).
15. M. J. FOLLOWS, S. DUTKIEWICZ: Modeling Diverse Communities of Marine Microbes. Ann Rev Mar Sci, **3**, 427 (2011).
16. Y. LIANG, D. M. ZHU: Subjective Well-being of Chinese Landless Peasants in Relatively Developed Regions: Measurement Using PANAS and SWLS. Soc Indic Res, **123** (3), 817 (2015).
17. J. ZIMMERMANN, R. JAHN, B. GEMEINHOLZER: Barcoding Diatoms: Evaluation of the V4 Subregion on the 18S rRNA Gene, Including New Primers and Protocols. Org Divers Evol, **11**, 173 ((2011).
18. Y. LIANG, S. Q. LI: Landless Female Peasants Living in Resettlement Residential Areas in China Have Poorer Quality of Life than Males: Results from a Household Study in the Yangtze River Delta Region. Health Qual Life Out, **12** (71), 1 (2014).

Received 4 August 2015
Revised 19 September 2015