Sci. Bull. (2015) 60(3):348–355
DOI 10.1007/s11434-014-0658-y

www.scibull.com
www.springer.com/scp

Article

Life & Medical Sciences

Functional genomic analysis of Hawaii marine metagenomes

Xiaoqi Wang · Qi Wang · Xiao Guo · Luying Liu · Jiangtao Guo · Jinxian Yao · Huaiqiu Zhu

Received: 17 July 2014 / Accepted: 13 August 2014 / Published online: 21 January 2015

Abstract Using high-throughput sequencing on metagenome to analyze marine microbial community, it is one of current main issues in the field of environmental microbe research. In this paper, we conducted the functional analysis on seven samples of metagenomic data from different depth seawater in Hawaii. The results of gene prediction and function annotation indicate that there are large amounts of potential novel genes of which functions remain unknown at present. Based on the gene annotation, codon usage bias is studied on ribosomal protein-related genes and shows an evident influence by the marine extreme environment. Furthermore, focusing on the marine environmental differences such as light intensity, dissolved oxygen, temperature and pressure among various depths, comparative analysis is carried out on related genes and metabolic pathways. Thus, the understanding as well as new insights into the correlation between marine environment and microbes are proposed at molecular level. Therefore, the studies herein afford a clue to reveal the special living strategies of microbial community from sea surface to deep sea.

Keywords Marine microorganism · Metagenome · Gene annotation · Codon usage bias · Metabolic pathway

1 Introduction

It is well known that marine microorganisms are in charge of ~98 % of marine main productivity and play an important role in global carbon and energy cycles [1]. Since they have to adapt to various environmental differences such as temperature, pressure and light intensity, the biodiversity makes most of the marine microorganisms not identified yet, including many typical extreme environmental microorganisms [2, 3]. The genomic research on marine microorganisms can not only help to understand the origin, evolution, diversity and adaptation of lives to complex and extreme environment, but also have important implications for biosynthesis and the application of synthetic biology to energy, pharmacy and environmental management. However, since more than 99 % of microorganisms are unculturable with current techniques, the microbial genomics research is highly limited [4]. With the expansion of metagenomic techniques that do not rely on culturing, the metagenomic research can be conducted using high-throughput sequencing, thus promoting the progress of marine microbial genomic research greatly.

In recent years, marine microbial metagenomic research has made notable progress, some metagenomic sequencing and analysis work on surface water, mesopelagic water, deep sea and subfloor sediments have been completed and published successively [5–8], among which the research on

Electronic supplementary material The online version of this article (doi:10.1007/s11434-014-0658-y) contains supplementary material, which is available to authorized users.

X. Wang · Q. Wang · L. Liu · J. Guo · H. Zhu (✉)
Department of Biomedical Engineering, College of Engineering, Peking University, Beijing 100871, China
e-mail: hqzhu@pku.edu.cn

X. Wang · Q. Wang · X. Guo · L. Liu · J. Guo · H. Zhu
Center for Quantitative Biology, Peking University, Beijing 100871, China

J. Yao
School of Life Sciences, Peking University, Beijing 100871, China

Hawaii marine microorganisms is a good example [8]. These researches have broken through the traditional analysis based on 16S rRNA and turned to structural and functional analysis of high-throughput whole genome based on next-generation sequencing (NGS). However, on account of both the challenges brought by metagenomes and high-throughput sequencing and the limitation of conventional genomics methods, the analyses of metagenomes are actually at an initial stage of exploration [9].

In this paper, metagenomes from Hawaii marine, which were sequenced by Delong et al. [8] and Hawaii Ocean Times-series station (HOT), are performed a comprehensive analysis of environmental metagenomes. According to biological, physical and chemical characteristics, they have sampled, sequenced, and assembled seven metagenomes from different depth seawater, and then carried out species binning and functional analysis [8]. However, with the limitation of analysis method and tools, only similarity search by BLASTX was applied in the original study, which brings two fatal problems to the marine microbial research. First, identifying genes by simply finding open reading frames (ORFs) could include fake genes into the analysis. Second, gene annotation by BLASTX limits the results only to the genes collected in the public database, the potential novel genes have been seriously ignored. With the recent development of bioinformatics on metagenomes, especially a series of metagenomic analysis tools designed by our lab [10–12], the latest methods can thus be applied to analyze the data from Hawaii sea microorganisms more accurately and deeply. Therefore, using series of metagenomic analysis methods and tools, we carried out gene prediction and function annotation on these seven datasets. The results indicate that there exist large amount of potential novel genes with unknown functions in marine microbiota. Based on gene annotation, we analyzed the codon usage bias using genes related to ribosomal proteins and found that marine extreme environments strengthen codon bias. Moreover, focusing on the marine environmental differences such as light intensity, dissolved oxygen, temperature and pressure among various depths, comparative analysis was performed on related genes and metabolic pathways. Our study revealed new insights into correlation and adaption between marine environment and microorganisms.

2 Materials and methods

2.1 Dataset

Hawaii marine metagenomes were collected from HOT station ALOHA, which is one of the most representative sampling sites in North Pacific Subtropical Gyre (NPSG). The seven samples are classified as follows: the upper euphotic zone (10 and 70 m), the base of chlorophyll maximum (130 m), below the base of the euphotic zone (200 m), well below the upper mesopelagic zone (500 m), the dissolved oxygen minimum layer (770 m) and the abyss 750 m above the seafloor (4,000 m). The various environmental factors of the seven samples are shown in Table 1. After sequencing, Delong et al. [13] published the data on integrated microbial genomes (IMG) database.

2.2 Gene and function annotation

Firstly, we used MetaGUN [10] to identify protein coding genes in the seven samples. After that, MetaTISA [11] was applied to improve the predictions of translation initiation site (TIS) of all genes. MetaGUN not only predict genes efficiently, but also discover more potential novel genes [10]. MetaTISA is a metagenomic TIS annotator for improving gene start prediction as a post-processor to a gene predicting program [11].

Functional information of the genes was obtained based on the amino acid sequence translated from the prediction result of MetaGUN and MetaTISA. All of the amino acid sequences were BLASTed against COG [14] and KO (KEGG Orthology) [15] database with e-value $\leq$0.01 and assigned to different orthology groups with the highest score.

2.3 Codon usage bias analysis

Relative synonymous codon usage (RSCU) is commonly used to evaluate the degree of codon bias in protein coding sequences [16–18]. On the assumption that each synonymous codon for an amino acid is used equally, RSCU is the observed frequency of codon divided by average frequency:

$$RSCU_{ij} = \frac{x_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}}, \quad (1)$$

where x_{ij} is the occurrences of codon j for the amino acid i, and n_i is the number (from 1 to 6) of codons for the ith amino acid. Obviously, RSCU value more than 1.0 indicates that the corresponding codon is used more frequently than expected, whereas <1.0 means the reverse [19].

2.4 Pathway comparative analysis

After function annotation by KO database, the occurrence of each protein in each sample was calculated, *i.e.,* gene copy number. In view of the big environmental difference among samples in light intensity, dissolved oxygen, temperature and pressure, different pathways were selected to perform comparative analysis, focusing on the difference