Journal of Microbiological Methods 120 (2016) 23–28

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Distribution of cold adaptation proteins in microbial mats in Lake Joyce, Antarctica: Analysis of metagenomic data by using two bioinformatics tools

Hyunmin Koo [a], Joseph A. Hakim [a], Phillip R.E. Fisher [a], Alexander Grueneberg [b,1], Dale T. Andersen [c,*,2], Asim K. Bej [a,**,2]

[a] Department of Biology, University of Alabama at Birmingham, Birmingham, AL 35294, USA
[b] Department of Pathology, University of Alabama at Birmingham, Birmingham, AL 35294, USA
[c] Carl Sagan Center for the Study of Life in the Universe, Mountain View, CA 94043, USA

ARTICLE INFO

Article history:
Received 30 October 2015
Received in revised form 9 November 2015
Accepted 9 November 2015
Available online 11 November 2015

Keywords:
MG-RAST
R code
McMurdo Dry Valleys
Perennially ice-covered lakes
CspA
Ice nucleation protein
Trehalose synthase
Antifreeze protein

ABSTRACT

In this study, we report the distribution and abundance of cold-adaptation proteins in microbial mat communities in the perennially ice-covered Lake Joyce, located in the McMurdo Dry Valleys, Antarctica. We have used MG-RAST and R code bioinformatics tools on Illumina HiSeq2000 shotgun metagenomic data and compared the filtering efficacy of these two methods on cold-adaptation proteins. Overall, the abundance of cold-shock DEAD-box protein A (CSDA), antifreeze proteins (AFPs), fatty acid desaturase (FAD), trehalose synthase (TS), and cold-shock family of proteins (CSPs) were present in all mat samples at high, moderate, or low levels, whereas the ice nucleation protein (INP) was present only in the ice and bulbous mat samples at insignificant levels. Considering the near homogeneous temperature profile of Lake Joyce (0.08–0.29 °C), the distribution and abundance of these proteins across various mat samples predictively correlated with known functional attributes necessary for microbial communities to thrive in this ecosystem. The comparison of the MG-RAST and the R code methods showed dissimilar occurrences of the cold-adaptation protein sequences, though with insignificant ANOSIM ($R = 0.357$; p-value $= 0.012$), ADONIS ($R^2 = 0.274$; p-value $= 0.03$) and STAMP (p-values $= 0.521$–0.984) statistical analyses. Furthermore, filtering targeted sequences using the R code accounted for taxonomic groups by avoiding sequence redundancies, whereas the MG-RAST provided total counts resulting in a higher sequence output. The results from this study revealed for the first time the distribution of cold-adaptation proteins in six different types of microbial mats in Lake Joyce, while suggesting a simpler and more manageable user-defined method of R code, as compared to a web-based MG-RAST pipeline.

1. Introduction

The McMurdo Dry Valleys (MDV) of southern Victoria Land constitute the largest (~4800 km^2) ice-free expanse in Antarctica and include an unusual group of closed-basin, meromictic perennially ice-covered lakes (Doran et al., 1994; Wharton et al., 1993). The MDV lakes harbor microbially-dominated ecosystems, with the largest biomass consisting of benthic photosynthetic microbial mats (Sutherland and Hawes, 2009; Wharton et al., 1983; Wharton et al., 1993; Zhang et al., 2015). The physicochemical environment in these lakes often manifest strong microscale chemical gradients in water chemistry, which influences the inhabiting microbial metabolisms as well as the lake nutrient cycle (Hawes et al., 2011; Laybourn-Parry and Wadham, 2014). Due to the extreme conditions, the microorganisms inhabiting these Antarctic lakes encounter various physicochemical constraints such as compromised light, lack of wind-induced waves and currents, limited nutrient availability, seasonal light–dark cycles, and most importantly, a year-round near zero temperature regimen (Lauro et al., 2011; Laybourn-Parry, 2002; Laybourn-Parry and Pearce, 2007). This translates to a large abundance of the extremophile bacteria that are accustomed to optimal growth in Antarctic and other cold environments (D'Amico et al., 2006). Consequently, adaptation to perpetual cold temperatures is perhaps the most common and essential criterion necessary for their sustenance and propagation of life (Laybourn-Parry, 2002; Rodrigues and Tiedje, 2008; Shivaji and Prakash, 2010). Since the discovery of the major cold shock gene (*cspA*) in *Escherichia coli* (Goldstein et al., 1990), the physiological and cellular

* Correspondence to: D.T. Andersen, Carl Sagan Center, SETI Institute, 189 Bernardo Ave., Suite 100, Mountain View, CA 94043, USA.
** Correspondence to: A.K. Bej, 1300 University Blvd., CH464, University of Alabama at Birmingham, Biology, Birmingham, AL, 35294-1170, USA.
E-mail addresses: dandersen@seti.org (D.T. Andersen), abej@uab.edu (A.K. Bej).
URL's: http://www.seti.org/users/dale-andersen (D.T. Andersen), http://www.uab.edu/cas/biology/bej (A.K. Bej).
[1] Current address: Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA.
[2] Joint corresponding authors.

http://dx.doi.org/10.1016/j.mimet.2015.11.008
0167-7012/

adaptations with underlying genetic mechanisms to cold and subzero temperatures have been a popular topic of investigation in microorganisms inhabiting the cold ecosystems (Bárria et al., 2013; Singh et al., 2015). Recently, the application of the NextGen sequencing (NGS) technology on microbial metacommunity DNA has greatly advanced our ability to identify and map metabolic genes, including an insight into the genetic strategies in cold adaptive bacteria (De Maayer et al., 2014; Varin et al., 2012). It is to be noted that the analysis of the NGS data requires efficient and reliable bioinformatics tools to filter, organize, and validate the targeted functional genes and their proteins of interest (Magi et al., 2010). In this study, we have used Illumina HiSeq2000 shotgun metagenomic data to explore the distribution of bacterial cold-adaptation proteins in six different types of microbial mats (herein, six mats) at various depths in Lake Joyce. For filtering of the cold-adaptation proteins from the metagenomic data and subsequent downstream bioinformatics analysis, we have compared the efficacy of the popularly used Metagenomic Analysis Server (MG-RAST) (Meyer et al., 2008) with the R code (Team, 2014), which was written for this study.

2. Materials and methods

2.1. Sample collection, DNA extraction, and sequencing

In November–December 2010, samples of benthic microbial mats were obtained from Lake Joyce (77° 43.2′S, 161° 37.7′E), a proglacial, permanently ice-covered, meromictic lake located in Pearse Valley, north-northwest of the Taylor Glacier (Andersen et al., 1998; Hawes et al., 2011). To investigate and sample the benthic environment, a hole was melted through the ice cover along the southwestern shore (77° 43.355′S, 161° 36.878′E) using methods described by Andersen (2007). Tethered divers using scuba then collected mat samples from just beneath the ice-cover at 7 m along a straight line downslope to a depth of 30 m. Six mat samples were collected, and the water temperatures and depths at which they were collected are described in Table 1. After collection, the mat samples were stored at −20 °C in the Albert P. Crary Science and Engineering Center at McMurdo Station, and then transported (frozen at −20 °C) to the University of Alabama at Birmingham (UAB). Temperature profiles of the water column were measured using a YSI 6600 Sonde multiparameter probe (Yellow Springs Instruments Ltd., Yellow Springs, OH, USA).

To increase the sampling accuracy, three samples were taken from different parts of each mat sample and used for community DNA extraction by using MoBiO PowerSoil® DNA Isolation Kit (MoBio Laboratories Inc., CA; www.mobio.com; cat # 12888-100). The purity, concentration and the quality (high molecular weight with least fragmentation) of the purified DNA was determined first by using a Lambda 2 spectrophotometer (Perkin Elmer, Norwalk, Conn.), and then by agarose gel electrophoresis (Ausubel et al., 1987). An equal amount of high-quality DNA (~3 μg) from each of the triplicate mat samples was pooled into a single sample (White et al., 2015), which resulted in a total of six DNA samples to represent the six mats. These DNA samples were then subjected to NGS on an Illumina HiSeq2000 (paired-end, 2 by 50 bp) sequencing platform at the UAB Heflin Center for Genomic Science (http://www.uab.edu/hcgs/). The raw sequence reads were quality-checked and quality-filtered (>25 quality score and 80% coverage) by using FastQC (Babraham Bioinformtics, Cambridgeshire CB22 3AT, U.K.) (Andrews and FastQC, 2010), and FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). The filtered sequences were assembled using Velvet/Metavelvet (Namiki et al., 2012; Zerbino and Birney, 2008), and then contigs were generated. The contigs were then converted to putative protein sequences using FragGeneScan (Rho et al., 2010). For protein annotation and analysis, the assembled sequences were uploaded on the Metagenomics Analysis Server (MG-RAST) (Meyer et al., 2008), and subjected to similarity search using the SEED database, keeping 10^{-5}as the maximum E-value (Overbeek et al., 2005).

2.2. Filtering cold-adaptation proteins using MG-RAST and R code

A total of six metagenomic datasets from six mats were used to filter antifreeze proteins (AFPs), cold-shock DEAD-box protein A (CSDA), cold-shock family of proteins (CSPs; including CspA, CspB, CspC, CspD, CspE, and CspG), fatty acid desaturase (FAD), ice nucleation protein (INP), and trehalose synthase (TS) by using MG-RAST (Meyer et al., 2008) and R code (Team, 2014). First, we used the SEED (Overbeek et al., 2005) database to annotate the metagenomics sequences, then the MG-RAST bioinformatics pipeline to search for the cold-adaptation proteins within the six metagenomic datasets using the following parameters: >10 alignment lengths, >65% sequence identity, and E-value of $\leq 10^{-5}$. We then used a bioinformatics script written for this study using the R code (Team, 2014) to search for and filter the cold-adaptation proteins from the metagenomic data. The bioinformatics workflow and R code have been elaborated in Supplementary data 1 (S1.1, and S1.2), and Supplementary data 2. After filtering through the MG-RAST and the R code, the abundance of the cold-adaptation protein sequences in each type of mat were recorded.

2.3. Stacked column bar graph and statistical analyses using bioinformatics tools

The distribution and abundance of the cold-adaptation proteins in the Lake Joyce mat samples filtered through MG-RAST and R code were analyzed by stacked column bar graphs using Microsoft Excel software (Microsoft, Seattle, WA). In order to compare and obtain statistical significance of the data generated by MG-RAST and R code, we performed non-parametric statistical analyses, including Analysis of Similarities (ANOSIM; vegan library) (Oksanen et al., 2007) and Permutational Multivariate Analyses of Variance using Distance Matrices (ADONIS; vegan library) (Oksanen et al., 2007) through the 'vegan' package (Oksanen et al., 2007) of R software (Team, 2014). The distribution and relative occurrence of the metagenomic protein sequences derived from the two different methods were further compared using Statistical Analyses of Metagenomic Profiles (STAMP) (Parks et al., 2014). The extended error bar plot, which follows two-sided Welch's t-test (Bluman, 2007), was constructed in STAMP, and the significance was determined based on 95% confidence interval.

3. Results

3.1. Distribution and abundances of the cold-adaptation proteins analyzed by MG-RAST and R code

Overall, analysis through MG-RAST revealed a total of 1473 matches against the SEED metabolic profile subsystems database

Table 1
Depth and temperature profiles for six mats in Lake Joyce, Antarctica used in this study.

Mats	Description	Depth (m)	Temperature (°C)
Ice Mat (IM)	Pigmented mats growing on a shelf of submerged ice	7	0.08
Lift-Off Mat (LO)	Floated pigmented mats lifted from the bottom	7.3	0.08
Tent-Mix Mat (TM)	Calcified mats with peak-ridge and pigmentation	10.15	0.16
Bulbous Mat (BM)	Calcified, pigmented mats with bulbous-columnar shapes	18	0.29
Yellow Mat (YM)	Yellow pigmented mats growing in hypoxic zone	26	0.13
Black Hole Mat (BH)	Black mat distributed in the bottom hypoxic zone of the lake	26	0.13