Genomics Data 12 (2017) 89–96

Contents lists available at ScienceDirect

Genomics Data

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

A metagenomic approach to decipher the indigenous microbial communities of arsenic contaminated groundwater of Assam

Saurav Das [a,b], Sudipta Sankar Bora [a], R.N.S. Yadav [b], Madhumita Barooah [a,*]

[a] Department of Agricultural Biotechnology, Assam Agricultural University, Jorhat, Assam, India
[b] Centre for Studies in Biotechnology, Dibrugarh University, Dibrugarh, Assam, India

ARTICLE INFO

Article history:
Received 19 October 2016
Received in revised form 22 March 2017
Accepted 26 March 2017
Available online 30 March 2017

Keywords:
Metagenomic
Arsenic
Assam
Groundwater
Bacteria
Proteobacteria
Siderophore

ABSTRACT

Metagenomic approach was used to understand the structural and functional diversity present in arsenic contaminated groundwater of the Ganges Brahmaputra Delta aquifer system. A metagene dataset (coded as TTGW1) of 89,171 sequences (totaling 125,449,864 base pairs) with an average length of 1406 bps was annotated. About 74,478 sequences containing 101,948 predicted protein coding regions passed the quality control. Taxonomical classification revealed abundance of bacteria that accounted for 98.3% of the microbial population of the metagenome. Eukaryota had an abundance of 1.1% followed by archea that showed 0.4% abundance. In phylum based classification, Proteobacteria was dominant (62.6%) followed by Bacteroidetes (11.7%), Planctomycetes (7.7%), Verrucomicrobia (5.6%), Actinobacteria (3.7%) and Firmicutes (1.9%). The Clusters of Orthologous Groups (COGs) analysis indicated that the protein regulating the metabolic functions constituted a high percentage (18,199 reads; 39.3%) of the whole metagenome followed by the proteins regulating the cellular processes (22.3%). About 0.07% sequences of the whole metagenome were related to genes coding for arsenic resistant mechanisms. Nearly 50% sequences of these coded for the arsenate reductase enzyme (EC. 1.20.4.1), the dominant enzyme of *ars* operon. Proteins associated with iron acquisition and metabolism were coded by 2% of the metagenome as revealed through SEED analysis. Our study reveals the microbial diversity and provides an insight into the functional aspect of the genes that might play crucial role in arsenic geocycle in contaminated ground water of Assam.

1. Introduction

Arsenic toxicity in drinking water is a serious human health concern affecting millions of people around the globe. The problem is especially acute in the Ganges-Brahmaputra Delta (GBD) region of India where geogenic groundwater arsenic concentration has been reported to be more than 50 fold higher than standard WHO limit [1]. There are strong evidences to suggest that microorganisms play crucial role in mobilizing arsenic in the groundwater through cascade of reduction and oxidation reactions. The secretion of siderophore by some bacteria affects this process by releasing the primary iron bound arsenic from the sediments. A detail insight of the microbial communities controlling the bio-geochemical cycle of arsenic in the natural system is challenging due to their extreme diversity and uncultivated status (Fig. 1). Metagenomic analysis has offered an unprecedented opportunity to examine the response and adaptation strategies of the microbial communities to the environmental toxicity [2,3,4]. Studies on microbial communities from several environments viz., acid-mine drainage [5], marine water and sediments [6,7] including arsenic contaminated soils [8,9] have provided novel insights on the microbial community structure their function along with evolution pattern and have led to the discovery of novel gene.

Arsenic contamination in groundwater of Assam, a north-eastern state of India was first reported in 2004 [10]. Since then studies indicating alarming increase in the arsenic content in the ground water of several districts in the state has been reported. Several sites (Titabor, Dhakgorah, Seleng-hat and Moriani) in the district of Jorhat of Assam have presence of very high arsenic content (194–657 g/μl) in the groundwater [1,11]. The level of arsenic in these localities is far above the WHO and BIS approved guidelines of 10 μg/l and 50 μg/μl respectively [12,13]. Such highly contaminated sites offer unique opportunity to investigate the role of microorganisms in arsenic geogenic cycle and its mobilization.

In this paper we report the microbial community structure and their function in a highly arsenic contaminated groundwater as revealed

* Corresponding author.
E-mail address: m.barooah@aau.ac.in (M. Barooah).

http://dx.doi.org/10.1016/j.gdata.2017.03.013

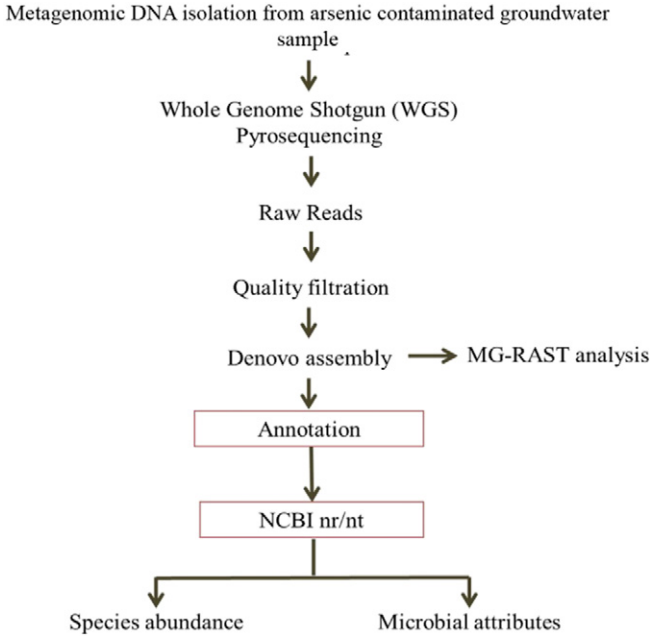

Fig. 1. Analysis strategy performed to analyze microbial diversity prevalent in the arsenic contaminated groundwater sample. DNA from composite groundwater sample was used for Whole Genome Shotgun (WGS) sequencing.

through shotgun sequencing method. The metagenomic library generated from our study also predicts of the roles of these microbes in arsenic geocycle.

2. Materials and methods

2.1. Ethics statement

No specific permits were required for the described field studies.

2.2. Sampling

Groundwater samples were collected from 5 different sites of Tanti Gaon, Titabor subdivision, Jorhat district (27°57′N, 94°16′E). All the samples were collected in sterile acid-washed Nalgene water bottles. Before collecting the water samples, hand-held tube-wells were pumped for 20 min to remove any unwanted residues present in the tube. Sampling was performed during November 2014. On field chemical parameters (pH and arsenic concentration) of the collected were recorded using portable pH meter (Spectronic Camspec Ltd., UK) and Arsenic Testing Kit (Merck, Germany) respectively. Samples were carried to the laboratory on ice packs and stored for further analyses using standard procedures. Concentration of arsenic was determined by atomic absorption spectrophotometer using protocol as described by Behari and Prakash [14]. Physicochemical parameters of the samples are presented in Table 1.

2.3. DNA extraction from contaminated water sample

Aliquots of 10 ml of water samples collected from 5 locations were thoroughly mixed to generate a 50 ml volume and considered as a composite sample for further analysis. The DNA was extracted from the filtrate using PowerWater® DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA, USA) in accordance with manufacturer's instructions. Extracted DNA was quantified by DNA (dsDNA)-binding dye assay on the Qubit Fluorometer which has a detection limit of as low as dsDNA at 10–100 pg/μl [15].

2.4. Preparation of 2 × 300 MiSeq libraries

A total of 3.0 μg of environmental DNA was extracted from the sample from which, 1.0 μg was subjected to restriction digestion and library construction The paired-end sequencing library was prepared using Illumina TruSeq DNA Library Preparation Kit, initiated with the fragmentation of 1.0 μg gDNA followed by paired-end adapter ligation. The ligated product was purified using 1× Ampure beads and elution of ~500–800 bp to further PCR amplify as described in the kit protocol. The amplified library was analyzed in Bioanalyzer 2100 (Agilent Technologies) using High Sensitivity (HS) DNA chip as per the manufacturer's instructions.

2.5. Cluster generation and sequencing

Based on the data obtained from the Qubit concentration for the library and the mean peak size (708 bp) from Bioanalyzer profile, 10 pM of the library was loaded onto Illumina MiSeq for cluster generation and sequencing. Paired-end sequencing allows the template fragments to be sequenced in both the forward and reverse directions on MiSeq. High-quality metagenome reads were assembled using CLC workbench (CLC bio, Denmark) with default parameter (minimum contig length: 200) for trimming and de novo assembly [16].

2.6. MG-RAST analysis

The MG-RAST portal offers automated quality control, annotation, comparative analysis and archiving services. The uploaded data is usually preprocessed through SolexaQA [17], to trim low-quality regions from FASTQ data. More than two standard deviations away from the mean read length are discarded [18]. A simple k-mer approach is used to rapidly identify all 20 character prefix identical sequences. This step is required in order to remove Artificial Duplicate Reads (ADRs) [19]. The set of ADRs is kept aside to be analyzed by DRISEE (Duplicate Read Inferred Sequencing Error Estimation) [20], in order to determine the degree of variation among prefix-identical sequences derived from the same template. The MG-RAST pipeline also provides the option of removing reads that are near-exact matches to the genomes of a handful of model organisms, including fly, mouse, cow, and human. The screening stage uses Bowtie [21] (a fast, memory-efficient, short read aligner), and only reads that do not match the model organisms pass into the next stage of the annotation pipeline.

Table 1
Physicochemical parameter of the contaminated groundwater sample collected for metagenomics analysis.

Sl. no.	Parameter	Ground-water sample 1	Ground-water sample 2	Ground-water sample 3	Ground-water sample 4	Ground-water sample 5
1.	pH	6.4	6.2	7.1	5.9	6.8
2.	Electrolytic conductivity (μS/m)	1783	1532	1572	1770	1814
3.	Temperature (°C)	22.0	24.0	22.0	21.6	22.0
4.	Dissolved oxygen (mg/l)	8.4	7.8	7.6	8.2	8.7
5.	Redox (mv)	187	172	167	183	181
6.	Arsenic concentration (μg/l)	217	50	20	156	112