Available online at http://www.ijims.com
ISSN - (Print): 2519 – 7908 ; ISSN - (Electronic): 2348 – 0343
IF:4.335; Index Copernicus (IC) Value: 60.59; UGC Recognized -UGC Journal No.: 47192. 1st July

Metagenomic Analysis: Microbial Diversity of Hot Spring Water Near Mumbai Region (India)

Deepak Mishra* and Neeta Khanolkar

Department of Microbiology, South Indian Welfare Society's Commerce and Science College, Major R. ParmeshwaramMarg, Mumbai,India.

Corresponding author: *Deepak Mishra

Abstract

A large diversity of microorganisms are being harbored by natural environments and most of them have not yet been identified /or characterized. Extremophilic microorganisms are known to survive in diverse extreme conditions, such as high or low temperatures, high salinity etc.Thermophiles and hyper-thermophiles have the ability to survive in environments with very high temperature such as in hot springs. Sample was collected from hot spring near Nimbavali village, Ganeshpuri, Mumbai region, Maharashtra, India. The bacterial diversity as determined by sequencing the hypervariable region V3 of the 16S rRNA belonging to the microbial community of the Ganeshpuri (Nimbavali) hot spring is presented here. In the taxonomy classification of OTUs, only top ten OTUs were selected and rests were shown as others. Top ten OTUs at phylum level wereProteobacteria, Firmicutes,Bacteriodetes;class- Gammaprotobactera, Betaprotobacteria, Clostridia; Order- Clostridiales, Bacteriodales, Rhodothermales; family- Halobacteriaece,Comamondaceae; genus- Protovella, Oscillospira;and species- *Bifidobacteriumadolescentis*. In all of the above significant percentage of OTUs were found to be unknown or not yet classified. These unknown microbes could serve as the source of novel products. Our data provide new information on hot spring bacteria and shed light on their abundance, diversity, distribution and coexisting organisms.

Keywords: Extremophiles, thermophiles, metagenomic analysis, hot springs, sequencing.

Introduction

A large diversity of microorganisms are being harbored by natural environments and most of them have not yet been identified /or characterized. Although numerous studies of bacterial communities in different natural habitats were performed, it is an acceptable norm that exploitable microbial diversity is in-exhaustive and microorganisms represent the largest reservoir of undescribed biodiversity.[1]Extremophilic microorganisms are known to survive in diverse extreme conditions, such as high or low temperatures, high salinity, acidic and alkaline pH-values, and high radiation. Among these extremophilic microbes, thermophiles, and hyper-thermophiles have the ability to survive in environments with very high temperature such as in hot springs, with the help of enzymes that remain catalytically active under such conditions.[2]

Extreme ecosystems such as hot springs are of great interest as a source of novel extremophilic species, enzymes, metabolic functions for survival and biotechnological products. India harbors hundreds of hot springs, the majority of which are not yet explored and require comprehensive studies to unravel their unknown and untapped phylogenetic and functional diversity.[2]Some studies focusing on the metagenomics analysis of hot springs located in India have reported *Bacillus licheniformis*, *Opitutusterrae*, *Rhodococcuserythropolis*, and *Cellovibriomixtus* as major bacterial genera.[2]

A culture independent approach "Metagenomic studies" to investigate the composition of mixed microbial communities from environmental niches, allows for the identification of not only culturable but also nonculturableundescribed bacteria. A better understanding of evolution, lifestyle and diversity of the microorganisms and direct access to the genomic DNA of co-existing microbial

species can give and uncover more of the hidden world of microbes. The use of metagenomics with post-Sanger sequencing offers the ability to quickly identify, characterize and understand microbial communities in detail.The GS-FLX-titanium sequencer based on pyrosequencing allows for rapid and inexpensive analysis of microbial diversity in multiple samples in a single run and without the need for cloning. The bacterial diversity as determined by sequencing the hypervariable region V3 of the 16S rRNA belonging to the microbial community of the Ganeshpuri(Nimbavali) hot spring is presented here. The seasonal microbial diversity of these hot water spring were not yet been studied and this report presents the first study to describe the bacterial diversity present at Ganeshpuri hot spring.[1]The remarkable genomic versatility and complexity of these largely unculturable extremophilic communities can be accessed using metagenomics and next-generation sequencing technologies.[2]In the last decade, metagenomics has emerged as one of the most incredible events in the study of microbial ecology which has made it possible to access, in-principle, almost 100% of the genetic material present in unculturable microbes. More than 98% of the bacteria which cannot be cultured using traditional methodologies can be directly sequenced from their natural environments using the metagenomic approaches.[6]

Materials and Methods

Site Description and Sample Collection

Ganeshpuri is located near the Mumbai region, under the district Thane, state Maharashtra, India.Sample was collected from hot spring near Nimbavali village (19°30’31.8”N73°00’54.6”E) which people were not using for bathe near Ganeshpuri (Fig. 01). Sample was named as C-Sample (the water of this hot-spring was clean and transparent).Sample was collected in a sterile plastic carboy (2L volume) which was rinsed with 0.05% bleach solution for disinfection. A total of 5L water sample was collected and brought in ice boxes to the laboratory within 12–18 h of collection.The sample was stored at −4◦C and processed for the extraction of metagenomic DNA within a week.[2,16]

The hot springs in this study have a temperature range from 43.5 to 56°C and neutral to slightly alkaline pH- values(7.0–9.4). The reservoir temperatures could be around 110 ± 20ºC.[2] Although, hot springs with similar temperature and pH-values have been studied globally for their phylogenetic and functional characteristics, the geographic allocation of the currently studied hot springs makesthemuniqueintheirgeochemicalsetup.16SrRNAV3hypervariableregionampliconsequencingandshotgunmetagenomesequencingofallthesampleswascarriedoutusingIlluminaNextSeq500fortheexplorationofmicrobialcommunitiesinthesampleandgainingnewinsightsintogenes, enzymes,andmetabolicpathways contributing to their survival in the thermophilic environment.[2]The Ganeshpuri hot-water springs were classified on the basis of their geo-tectonic setup and grouped ingeothermal provinces of west coast regions and Son-Narmada lineament.[6]

Samples summary

This report is based on the analysis done on sample C, refer Table 01.

Sequence quality checking

Raw read summary

We showed the summary of raw fastq files obtained from sequencer Table 02.

Fastq quality check

This step was performed to check quality parameters for the sequences obtained from sequencer. Following quality checks were performed for each sample.

- Base quality score distributions

- Average base content per read
- GC distribution in the reads

Base quality score distribution

Base quality of each cycle for all samples is shown in Figure 02-03. The x-axis represents sequencing cycle and y-axis represents percentage of total reads. The quality of left and right end of the paired-end read sequences of the sample is shown in these figures. It can be seen that more than 90% of the total reads have Phred score greater than 30 (>Q30; error-probability >= 0.001). The Phred score distribution of the sample is provided in Table 3.

Base composition distribution

The composition of nucleotides in the sequence read for each sample is shown in Figure 4-5. The x-axis represents read positions and y-axis represents nucleotide percentage. The base composition of left and right end of the paired-end read sequences are calculated. Since the target sequence is that of V3 region sequence composition bias is observed in the sample. Overall base compositions of these samples are provided in Table 4.

GC distribution

The average GC content distribution of the sequenced read of the samples is shown in Figure 05. The x-axis represents average GC content in the sequence and y-axis represents percentage of sequences. We observed that the reads have GC content in the range 50-60%.

Identification of V3 region from paired-end reads

A propriety wet-lab approach is followed to sequence 16S rRNA V3 region of bacteria. Following steps were performed to extract V3 region from Illumina paired-end sequences.

a) Trimming of spacer and conserved region

b) Building consensus V3 region from trimmed paired-end reads

c) Filters to identify high quality V3 region sequences

- Conserved region filter
- Spacer sequence filter
- Read quality filter
- Mismatch filter

Usually a paired-end sequence from V3 Metagenomics contains some portion of conserved region, spacer and V3 region. As a first step we removed the spacer and conserved region from paired-end reads. After trimming the unwanted sequences from original paired-end data a consensus V3 region sequence is constructed using ClustalO program. We applied multiple filters such as, conserved region filter, spacer filter and mismatch filter to take further only the high quality V3 region sequences for various downstream analyses(Table:5). While making consensus V3 sequence, All consensus reads were formed with 0 mismatches with an average contig length of ~145 to >175 shown in Figure:07.

Pre-processing of reads: Chimera filter

We have performed the following pre-processing steps before we start the analysis. Chimeras were also removed using the de-novo chimera removal method UCHIME implemented in the tool USEARCH.

OTUs and Taxonomy classification and relative abundance

This analysis was performed using the pre-processed consensus V3 sequences. Pre-processed reads from all samples were pooled and clustered into Operational Taxonomic Units (OTUs) based on their sequence similarity using Uclust program (similarity cutoff = 0.97). A total of 3,570 OTUs were identified from 248,453 reads (Figure. 08). OTUs with only one read in it are identified as Singletons OTUs and From 3,570 total OTUs, 923 singletons were removed and 2,647 OTUs were selected for further analysis.

QIIME program was used for the entire downstream analysis.[18] Representative sequence was identified for each OTU and aligned against Greengenes core set of sequences using PyNAST program.[19-20] Further we aligned this representative sequences against reference chimeric data sets. Then, taxonomy classification was performed using RDP classifier and SILVA OTUs database. The top 10 phylum, class, order, family, genus and species distribution for each sample based on OTU and reads are shown in Figure 9-14. It should be noted that the taxa other than top 10 are categorized as “Others”.

The sequences do not have any alignment against taxonomic database are categorized as “Unknown”. Further, a heat map was generated using QIIME program A detailed sample-wise OTUs, V3 sequences and taxonomical annotations based on SILVA database. [21,22]

Alpha diversity with samples and rarefaction curves

In this section we analyzed the microbial diversity within the sample by calculating Shannon, Chao1 and observed species metrics. The chao1 metric estimates the species richness while Shannon metric is the measure to estimate observed OTU abundances, and accounts for both richness and evenness. The observed species metric is the count of unique OTUs identified in the sample. The rarefaction curve for each of the metric is provided in Figure. 15-17. The metric calculation was performed using QIIME software.

Beta diversity between samples

In order to run beta diversity more than three or more samples were required, as it is the analysis between the samples, therefore not performed for this sample.

Discussion

The use of culture-independent molecular methodology has become an essential and reliable tool for surveying the diversity of microbial life in any given habitat[16].The word metagenomics was coined to capture the notion of analysis of a collection of similar but not identical items, as in a meta-analysis, which is an analysis of analyses. (Community genomics, environmental genomics, and population genomics are synonyms for the same approach.) [3.]

Microbes that are metabolically active in extremely thermoacidic environments have attracted for their unique ecology and physiology as well as sources of heat and acid stable biocatalysts. Solfataric fields are the most important biotopes thermoacidophilic microbes. Most of thermoacidophilic microbes that usually lived in solfataric fields belong to the archaea including the genera *Acidianus, Desulfurolobus, Metallosphaera, Stygiolobus, Sulfolobus, Sulfurisphaera, Sulfurococcus, Thermoplasma*, and *Picrophilus*.[4]

Maharashtra in India has following hot springs a) Unkeshwar (30°C – 40°C) lies in Nanded and Yeotmal District, b) Akoli (near Thane District), c) Vajreshwari Hot-water Spring (34 km. from Thane), d) Ganeshpuri i, e) Satvali, f) Sahada and Chopda (lies near Nandurbar District), g) Kundwa (44°C), h) Unabdeo (60°C), i) Ramtalab (40°C), j) Indave (41°C), k) Khadgaon (38°C).[6]

It is yet to be find that "what these thermophiles are doing in those extreme conditions". Entire genome analysis of all diverse microorganisms may lead to a set of novel genes which may be encoding for different functions such as enzymes (thermostable enzymes), peptides, different products and other types of fine chemicals which could take human life and health at a different height.

The results obtained were surprising, as the read GC percentage was between to <50 and read percentage was between 70-90% (fig.06). In the taxonomy classification of OTUs, only top ten OTUs were selected and rests were shown as others. Top ten OTUsat phylum level were Proteobacteria, Firmicutes, Bacteriodetes; class- Gammaprotobactera, Betaprotobacteria, Clostridia; Order- Clostridiales, Bacteriodales, Rhodothermales; family- Halobacteriaece,Comamondaceae; genus- Protovella, Oscillospira; and species- Bifidobacteriumadolescentis(Fig. 09-14). In all of the above several bacterial and archaeal sequences remained taxonomically unresolved, indicating potentially novel microorganisms in this geothermal ecosystem. Additional metagenomics of this habitat will facilitate identification of microorganisms possessing industrially relevant traits, such as enzymes and other compounds[5].

Although, the dominant groups are more emphasized, it is worth noting that the microorganisms found in low abundance in rare or extreme environment are important individuals of the community that could constitute an unexplored reservoir of genomic variation and novelty. A significant number of the sequences could not be assigned to any phyla and were grouped as unclassified bacteria. This group might represent bacteria that have not yet been classified or detected before. Contamination of the hot water spring by normal surface water, soil, and spores cannot be excluded, but it can, however, be concluded that the bacterial phylotypes detected can possibly all proliferate in this thermophilic environment.[1]Proteobacteria has also been reported from many studies based on the 16S rRNA analysis of hotsprings with moderately high and very high temperatures (44–110°C) at various geographic allocations, including India.[2]

Conclusion

In conclusion this study showed that a considerable diversity of microbial communities can be revealed by metagenomics using Illumina Next Seq500 of V3 region, and gives insight into microbial genetic diversity, community composition, distribution and abundance.

References

1. Tekere M, Lötter A, Olivier J, et al. 2011. Metagenomic analysis of bacterial diversity ofSiloam hot water spring, Limpopo, South Africa, African Journal of Biotechnology Vol. 10(78), pp. 18005-18012, DOI: 10.5897/AJB11.899.

2. Saxena R, Dhakan D, Mittal P, et al. 2017. Metagenomic Analysis of Hot Springs in Central India Reveals Hydrocarbon Degrading Thermophiles and Pathways Essential for Survival in Extreme Environments, Front. Microbiol. 7:2123. doi: 10.3389/fmicb.2016.02123.

3. HandelsmanJ, 2004. Metagenomics: Application of Genomics to Uncultured Microorganisms, Microbiology and Molecular Biology Reviews, Dec. 2004, p. 669–685 Vol. 68, No. 4, 1092-2172/04/$08.0, DOI: 10.1128/MBR.68.4.669–685.2004.

4. Bisht S, Das N, Tripathy N, 2011. Indian Hot- Water Springs: A Bird's Eye View, Journal of Energy, Environment & Carbon Credits Volume 1, Issue 1, Sep, 2011, Pages 1-15.

5. Bhatia S, Batra N, Pathak A, et al. 2015. Metagenomic evaluation of bacterial and archaeal diversity in the geothermal hot springs of Manikaran, India. Genome Announc 3(1):e01544-14. doi:10.1128/genomeA.01544-14.

6. Chaudhary N, Sharma AK, Agarwal P, et al. 16S Classifier: A tool for fast and accurate taxonomic classification of 16s rRNAhypervariable regions in metagenomic datasets. PLoS ONE 10(1). doi:10.1371/journal.pone.0116106.

07. Mehetre G, Paranjpe A, Dastager S, et al. 2016. Complete metagenome sequencing based bacterial diversity and functional insights from basaltic hot spring of Unkeshwar, Maharashtra, India. Genomics Data 7, 140–143, doi.org/10.1016/j.gdata.2015.12.031.

08. Xie W, Wang F, Guo L, et al. 2011. Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. The ISME Journal 5, 414–426, doi:10.1038/ismej.2010.144.

09. Saxena R, Chaudhary N, Dhakan D, et al. 2015. Draft genome sequence of Gulbenkianiamobilis strain MB1, a sulfur-metabolizing thermophile isolated from a hot spring in central India. Genome Announc 3(6):e01295-15. doi:10.1128/genomeA.01295-15.

10. Stetter K. 1999. Extremophiles and their adaptation to hot environments.FEBS Letters. 452 (1-2) 22-25, doi.org/10.1016/S0014-5793(99)00663-8.

11. Stetter KO. 1996. Hyperthermophilicprocaryotes. FEMS Microbiology Reviews. 18 (2-3), doi: 10.1111/j.1574-6976.1996.tb00233.x.

12. Solden L, Lloyd K, Wrighton K. 2016. The bright side of microbial dark matter: lessons learned from the uncultivated majority. Current Opinion in Microbiology 31, 217–226, http://dx.doi.org/10.1016/j.mib.2016.04.020.

13. Boutaiba S, Hacene H, Bidle K, et al. 2011. Microbial diversity of the hypersalineSidiAmeur and Himalatt salt lakes of the Algerian Sahara. J Arid Environ, 75(10), 909–916. doi:10.1016/j.jaridenv.2011.04.010.

14. DeCastro M, Rodríguez-Belmonte E, González-Siso, et al. 2016. Metagenomics of thermophiles with a focus on discovery of novel thermozymes. Front. Microbiol. 7:1521, doi: 10.3389/fmicb.2016.01521.

15. Zhou J, He Z, Yang Y, et al. 2015. High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. mBio 6(1):e02288-14. doi:10.1128/mBio.02288-14.

16. LozuponeC, StombaughJ, Gonzalez A, et al. 2013. Meta-analyses of studies of the human microbiota. Genome Res. Oct;23(10):1704-14.

17. D'Argenio V, Casaburi G, Precone V, et al. 2014. Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. Biomed Res Int.; 2014:325340.

18. Caporaso JG, Kuczynski J, Stombaugh J, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods.;7(5):335-6.

19. DeSantis T, Hugenholtz P, Larsen N, et al. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB.Appl Environ Microbiol 72:5069-72.

20. DeSantis T, Hugenholtz P, Keller K, et al. 2006. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res 34:W394-9.

21. Galand P, Casamayor E, Kirchman D, et. al. 2009.Ecology of the rare microbial biosphere of the Arctic Ocean.ProcNatlAcadSci U S A. 106(52):22427-32.

22. Chairmandurai A., Dharmaprakash V., and Shunmugiah K. 2013. Ultradeep 16S rRNA Sequencing Analysis of Geographically Similar but Diverse Unexplored Marine Samples Reveal Varied Bacterial Community Composition. PLoS ONE 8(10): e76724.

23. Ghelani A, Patel R, Mangrola A, Dudhagara P. 2015. Cultivation-independent comprehensive survey of bacterial diversity in TulsiShyam Hot Springs, India. Genomics Data 4:54–56. doi:10.1016/j.gdata.2015.03.003

Table 1: Sample Details

Sample names	Sample-C
Sequencing Platform	IlluminaHiSeq
Library type	Paired End (150bp x 2)
Project Type	Metagenomics Analysis (V3region)

Table 2: Raw read summary

ID	Total Reads (Paired-End)	Sequence Length (bp)	Total Data (Mb)	%GC	Average base quality (Phred score)
Sample-C	786,179	150	235.85	54.03	37.34

Table 3: Phred score distribution of the paired-end reads for the samples

Sample Name	Read Phred quality score distribution (%)			
	Q0-Q10	Q10-Q20	Q20-Q30	>= Q30
Sample-C	0.07	5.62	3.07	91.24

Table 4: Base composition distribution of the samples

Sample Name	Base Composition (%)			
	A	C	G	T
Sample-C	24.50	25.67	28.36	21.39

Table 5:Read Summary Table

Sample Name	Total Reads	Passed Conserved Region Filter	Passed Mismatch Filter
Sample-C	786,179 (100%)	296,028 (37.66%)	250,884 (31.92%)

Table 6: Summary of Singleton OTUs

Total Reads	248,453
Total OTUs Picked	3,570
Total Singleton OTUs	923
Total OTUs after Singleton removal	2,647

Figure 01: Sample collection site, hot-spring at Nimbavali village, Ganeshpuri, Maharashtra, India.

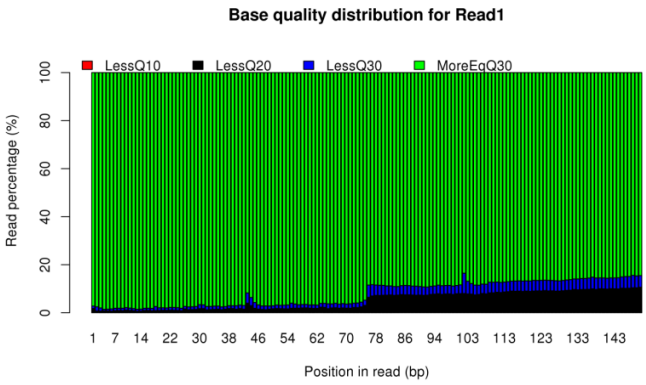

Figure 02. Base quality distribution of sample Sample-C(R1)

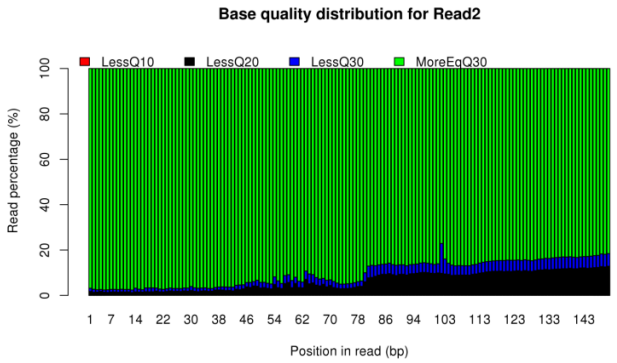

Figure 03. Base quality distribution of sample Sample-C (R2)

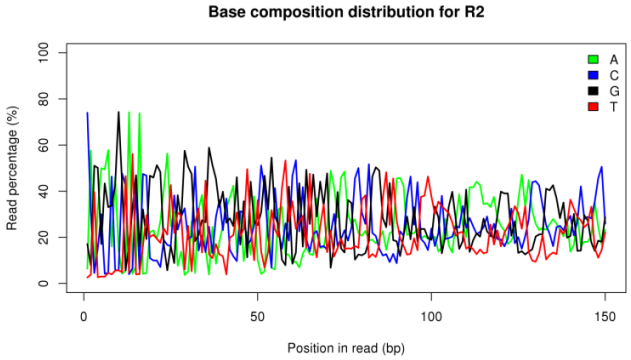

Figure 04.Base composition distribution of sample Sample-C(R1)

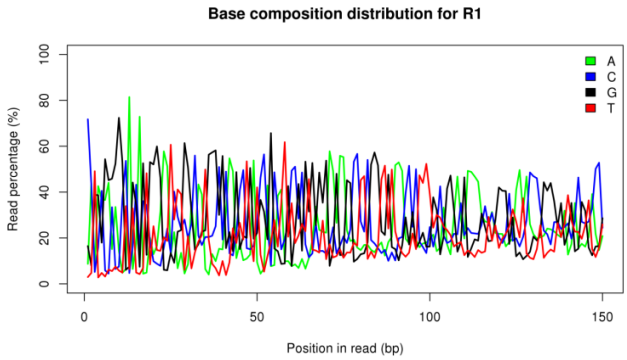

Figure 05. Base composition distribution of sample Sample-C(R2)

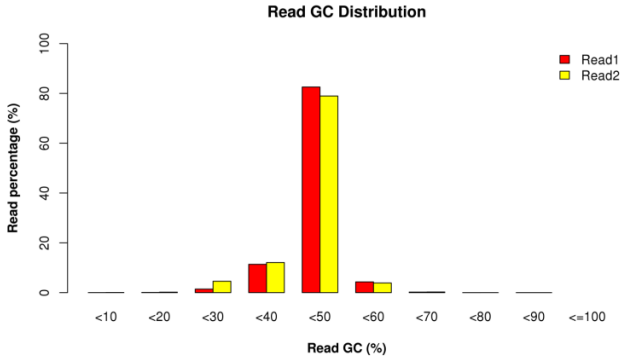

Figure 06.GC distribution of sample Sample-C

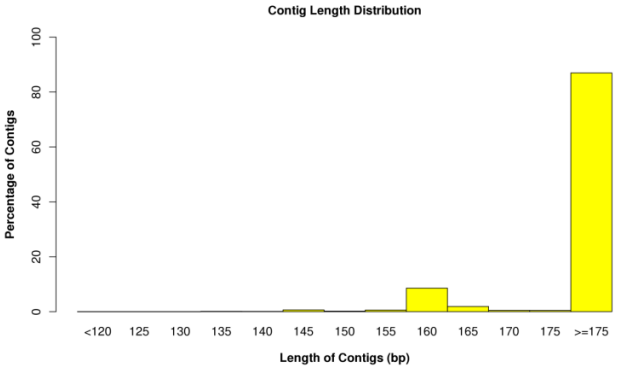

Figure07.Contig Length distribution of sample Sample-C

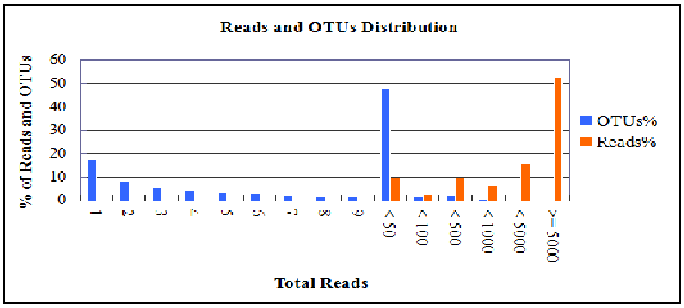

Figure 08.Shows a graphical representation of reads and OTU proportion.The red bar represents percentage of total OTUs in the read-count groups. The blue bar represents percentage of total read contributed by the OTUs in the read-count group.

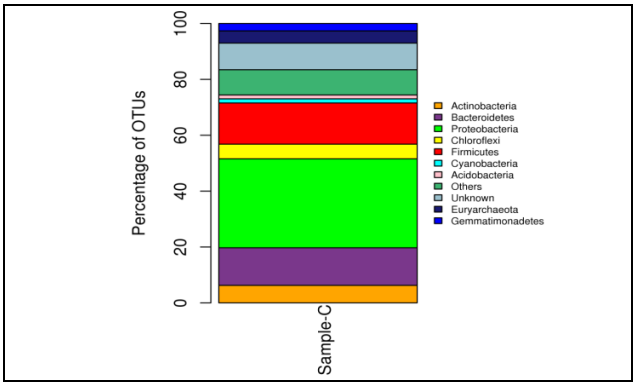

Figure 09.Taxonomy classification of OTUs at phylum level for the sample. Only top 10 enriched class categories are shown in the Figure.

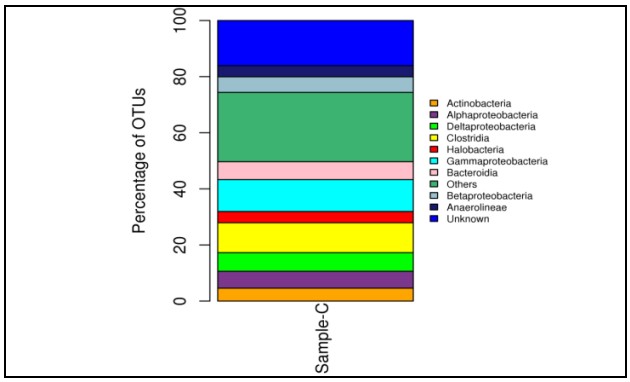

Figure 10. Taxonomy classification OTUs at class level for the sample. Only top 10 enriched class categories are shown in the Figure.

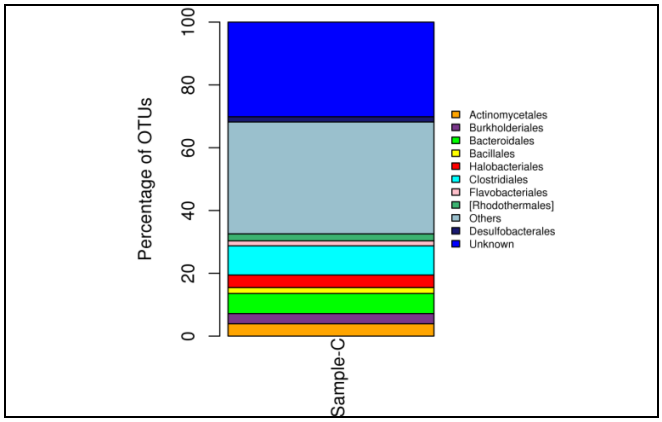

Figure 11.Taxonomy classification of OTUs at order level for the sample. Only top 10 enriched class categories are shown in the Figure.

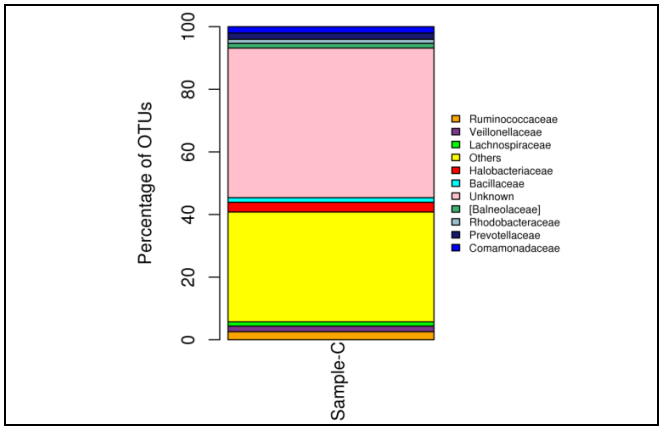

Figure 12.Taxonomy classification of OTUs at family level for the sample. Only top 10 enriched class categories are shown in the Figure.

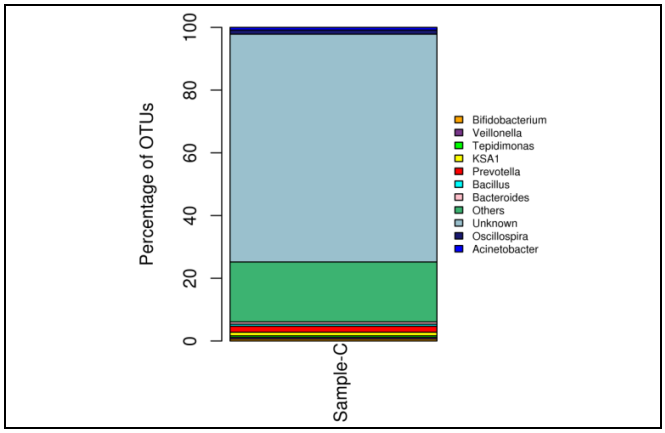

Figure 13.Taxonomy classification of OTUs at genus level for the sample. Only top 10 enriched class categories are shown in the Figure.

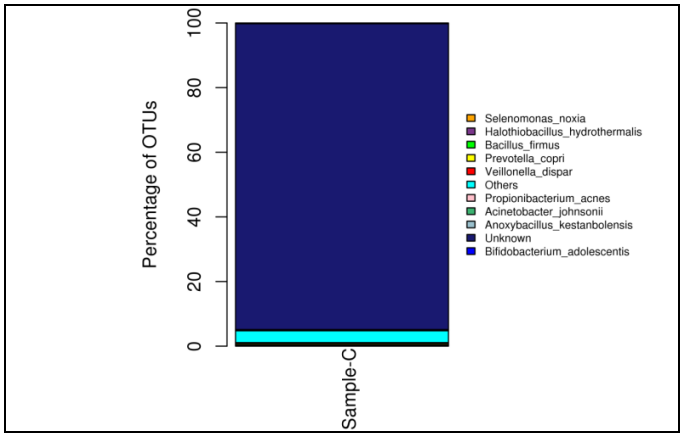

Figure 14.Taxonomy classification of OTUs at species level for sample. Only top 10 enriched class categories are shown in the Figure.

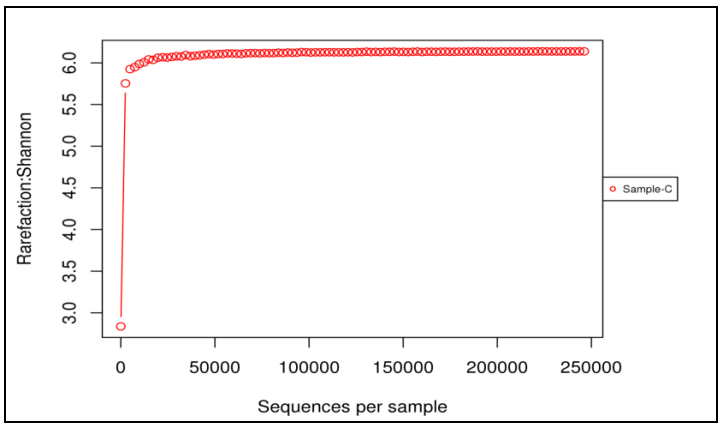

Figure 15.Shannon curve obtained for the samples. Alpha diversity was computed using Shannon metrics with rarefied OTU table size of 100.

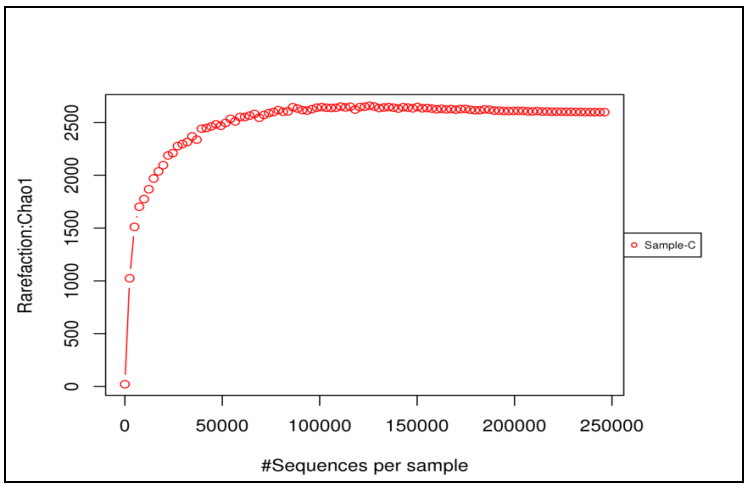

Figure 16.Chao1 curve obtained for the samples. Alpha diversity was computed using Chao1 metrics with rarefied OTU table size of 100.

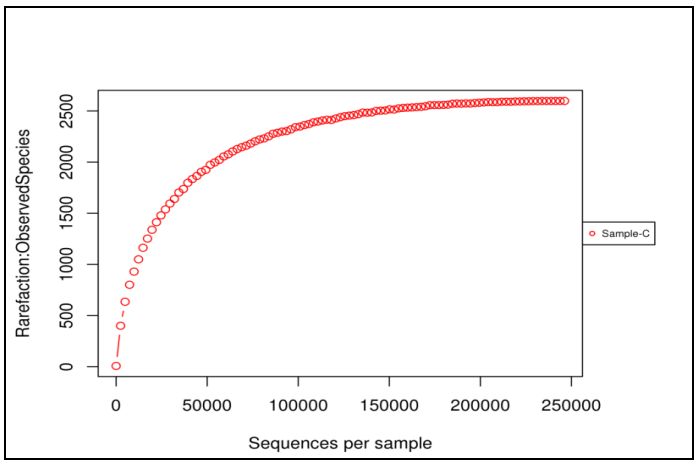

Figure 17.Observed species curve obtained for the samples. Alpha diversity was computed using observed species metrics with rarefied OTU table size of 100.