Sequence composition diversity in Alaskan glacier and other metagenomes

Sulbha Choudhari[1], Roman J. Dial[2], Dibyendu Kumar[3], Daniel H. Shain[1], Andrey Grigoriev[1]

[1]*Department of Biology, Center for Computational and Integrative Biology, Rutgers University, Camden, New Jersey, USA*

[2]*Department of Environmental Science, Alaska Pacific University, Anchorage, Alaska, USA*

[3]*Waksman Genomics Core Facility, Rutgers University Busch Campus, Piscataway, NJ, USA*

Correspondence should be addressed to Andrey Grigoriev; andrey.grigoriev@rutgers.edu

Abstract

Metagenomics by next generation sequencing has become an important tool for interrogating complex microbial communities. In this study we analyzed several pairs of metagenomic samples obtained by different methods and observed biases, resulting in different nucleotide composition of the sequenced reads. The pairwise sample comparison was based on the principal component analysis of dinucleotide word frequencies in sequences obtained from different platforms. We found bias in the sequences obtained from the different platforms for the amplified hypervariable regions in 16S rRNA but not in shotgun metagenome reads aligned to such hypervariable regions. The differences and consistency of the distributions of the nucleotides suggest that the biases are likely due to a combination of biases introduced by PCR and different sequencing protocols, and they are related to the GC content of the reads produced. For this reason, caution should be exercised when interpreting the results of comparative metagenomics studies, as they may vary depending on the sequencing technology.

Introduction

In recent years, metagenomics has emerged as a powerful tool involving the study of genome of microbial communities by sequencing microbial DNA extracted directly from environmental samples. Metagenomics unravels the enormous uncultured microbial diversity present in the environment, irrespective of the ability of member organisms to grow in the laboratory (Wooley, Godzik & Friedberg, 2010). Since less than 1% of the microbes present in nature can be cultured, metagenomics allows us to acquire microbial genomes by circumventing the cultivation step (Amann, Ludwig & Schleifer, 1995). The advent of next generation sequencing represents great

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

advancements made in sequencing technologies over the last 30 years. SOLiD/Ion Torrent PGM from Life Sciences, Genome Analyzer/HiSeq 2000/MiSeq from Illumina, and GS FLX Titanium/GS Junior from Roche typically represent these systems (Liu et al., 2012). These platforms have been applied to metagenomics studies for the characterization of microbial communities in diverse environments, and have become the most widely used quantitative approaches for studying the uncultured microbes (Choudhari, Lohia & Grigoriev, 2014).

There are many challenges in analyzing metagenomic data generated by these platforms, such as the assessment of microbial abundance in environmental samples, which is based on the frequency of occurrence of an organism's DNA observed in sequencing reads (Morgan, Darling & Eisen, 2010). It has been shown that the relative frequencies of organisms depend significantly on the DNA extraction and sequencing protocol used. In a recent study, the data generated by Illumina MiSeq, and Ion Torrent PGM platforms were compared for bacterial community profiling (Salipante et al., 2014). Earlier studies have shown that the sequencing technologies have different biases, depending on the approach adopted to obtain sequence data. A few studies have investigated the bias imposed by various DNA extraction protocols using environmental samples (Morgan, Darling & Eisen, 2010; Abusleme et al., 2014); moreover, comparison of two next generation technologies based on phylogenetic profiling of reads derived from sequencing has been reported (Salipante et al., 2014; Claesson et al., 2010). The differences in DNA sequencing protocol may introduce biases in the resulting sequences, as highlighted by a recent comparison of Illumina and Roche 454 platforms in an analysis of a freshwater planktonic community (Luo et al., 2012). The study revealed bias towards A's and T's over C's and G's in homopolymers sequence generated by Roche 454. With Illumina, these patterns were less common, and the errors were more randomly distributed. The evaluation of base-call error, frameshift frequency, and contig length were also compared. The sequences produced by different platforms introduce systematic biases and unique patterns of biased sequence coverage (Harismendy et al., 2009). A recent study revealed that bacterial species representation alters when different DNA extraction protocols were used with the same sequencing platform (Abusleme et al., 2014). Furthermore, the approaches used for the taxonomic classification of metagenomic reads also introduced their own limitations (Mavromatis et al., 2007). However, a sequence composition-based comparison between different sequencing platforms has not been

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

done yet.

Recently, we have performed a sequence-based comparison of two geographically distinct glacier metagenomes (Choudhari, Lohia & Grigoriev, 2014), that of an Alaskan glacier (Choudhari et al., 2013) and an Alpine glacier (Simon et al., 2009), and observed a striking difference between them. Not only have we seen a significant difference in the numbers of operating taxonomic units between the two samples, but a sequence composition of their reads was very dissimilar (Choudhari, Lohia & Grigoriev, 2014). Although the same type of starting material (ice/snow) was used for the two samples, they were sequenced with different platforms. This motivated us to consider a broader picture of how metagenome sample composition may relate to the corresponding sequencing platform. The current study emphasizes the comparative analysis of metagenomic sequencing data from different platforms in order to understand the variance in the data generated by different sequencers. We compared sequence data generated by two different platforms from the same biological sample, as well as sequences of different samples generated by same platform. Additionally, here we examined the sequencing data of an environmental sample (glacier) generated via two different platforms, while keeping all other pre-sequencing steps similar. The general question is whether it is appropriate to use the information obtained via the two technologies for comparative purposes. The present study outlines how various sequencing tools generate differences in the distribution of nucleotides in a sequence. We compared the sequences generated by these platforms, and also classified the data by binning into distinct taxonomic groups. The approach used for the analysis of sequence data involved the generation of nucleotide word frequencies that represents the distribution of nucleotides in the reads produced by different technologies and was displayed using principal component analysis (PCA). We observed striking differences in the nucleotide composition of the reads generated by different sequencing platforms and related them to the PCA load factors and GC composition of the hypervariable regions of 16S rRNA.

Material and Methods

Datasets

The glacier samples were collected from Harding Icefield in Alaska in August 2013. Three liters of snow were scooped from accumulation zone of the glacier when the temperature was around

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

4ºC, with wind 16 m/s, and rain at 2.3 mm/hr. After the samples melted at room temperature, the pre-filtering of samples with coffee filters was done to remove unwanted materials. Further filtering was done using a hand pump with 1.6 µm filter, in order to filter out eukaryotes. An electric suction pump was used with three durapore membrane filter of 0.22 µm in parallel for final separation of prokaryotes from the filtered (at 1.6 µm) water. DNA extraction was carried out using bead beating method plus column filtration according to the manufacturer's instructions (MO BIO Laboratories, Inc). After isolation, DNA was fragmented using Covaris S2 and divided into two tubes for Illumina MiSeq and Ion Proton library preparation. Final library was enriched by running 15 PCR cycles prior to sequencing. We are currently analyzing the data from both platforms (beyond the compositional bias described here) for functional content and, upon completion of that analysis, will make it available at SRA.

The data used in the analysis of different metagenome and same sequencer was taken from two very different metagenomes, leech gut (Maltz et al., 2014) and deep sea (Sogin et al., 2006), sequenced by the 454 to examine if there was difference in the different metagenomes sequenced by the same platform. The content of the intestinum and intraluminal fluid of the crop from leech fed one meal of heparinized sheep blood was sequenced via 454 (SRR1157610-11). The sea DNA samples for the analysis were collected from Atlantic Deep Water at a depth of 4 m; the already trimmed data was taken as described in the paper (Sogin et al., 2006). Moreover, the same human urine sample, which was sequenced, using two different platforms (MiSeq and Ion Torrent), was also studied (Salipante et al., 2014). The accession numbers of the data for the human urine samples were SRR1204944 (MiSeq) and SRR1205111 (Ion Torrent).

Data Processing

Illumina MiSeq-

The data for glacier used in the study was shotgun metagenomic data, and not the 16S rRNA sequences. To access the taxonomic classification, reads from the data were compared against database comprised of V9 hypervariable sequences (refv9). Only 3,810,507 reads that passed the quality filtering and greater than 50 bp in length were taken for the comparison. The reads from data served as query to identify its closest match in self made reference database containing only V9 regions of 16S rRNA. 745 reads were found to match the V9 hypervariable

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

region. The reference sequences were downloaded from The Visualization and Analysis of Microbial Population Structures (VAMPS) (Huse et al., 2014).

The data for the human urine metagenome was generated via Illumina MiSeq platform containing the V1-V2 hypervariable regions of the SSU rRNA gene. After quality filtering, only 14,912 V1-V2 hypervariable regions of 16S were extracted from 887,900 reads for human urine samples. The sequences that passed the Phred scores greater than 20 and did not contain any 'N's were retained.

Ion Torrent PGM/Ion Proton-

The glacier metagenomic data comparison was done in similar manner as the MiSeq data. Out of 12,052,104 quality filtered reads greater than 50 bp in length, only 2,430 V9 hypervariable regions remained.

643,464 sequences were downloaded from SRA generated through Ion Torrent platform of human urine samples. The 518 unique V1-V2 regions of 16S rRNA amplicons remained after quality filtering.

454/Roche GS-FLX-

The medicinal leech gut microbiota included content of intestinum. After quality filtering, 2,195 V6 hypervariable region of the SSU rRNA gene was extracted from 30,938 reads. Out of 9,282 trimmed reads from the lower deep water, only 5,279 unique V6 hypervariable regions were selected for the analysis.

Principal component analysis (PCA)

The key task when analyzing any metagenomic dataset is the assignment of anonymous metagenomic sequences to a diverse microbial population. In the current study we grouped the reads (representing hypervariable regions and sequenced by different sequencing platforms) based on dinucleotide patterns, also referred as dinucleotide word frequency. This study focused on how the nucleotide composition of sequences varies when different sequencing platforms are used for metagenome analysis. A K-dimensional feature vector (K=16) represented each DNA fragment, where each element in a vector encodes the dimer occurring in the fragment. To reduce the high dimensionality of feature space, the PCA was utilized which is an orthogonal linear transformation to highlight the differences and similarities among the data. With the PCA, the original data was transformed to a new set of variables, also known as principal components,

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

ordered according to their corresponding variances, and we retained the three principal components that contribute most to the variance. Projections on these three principal components were plotted highlighting clustering of the sequences. We also analyzed the load factors of the first, second, and third principal components in order to see which frequencies were mainly contributing to the variance. Tables 1 and 2 display 16 dinucleotide frequencies and their corresponding load factors for the two components.

MOTHUR (Schloss et al., 2009) was utilized to describe the taxonomy of extracted hypervariable regions of 16S rRNA from different platforms. The taxonomy outline from SILVA (Pruesse et al., 2007) database was used to classify sequences into specific reference taxonomies. The k-nearest neighbor consensus and BLAST (Altschul et al., 1990) approach was used for taxonomic classification of the reads.

Results and Discussion

Deep sea and leech gut metagenome via 454

To examine the sequence data of different samples generated by single sequencing platform, we used data produced by 454 from deep sea metagenome and leech gut. Only the variable V6 region of the 16S rRNA was extracted from the reads and used for phylogenetic identification of species, and we analyzed nucleotide composition patterns using PCA. The sample sequences were also classified taxonomically into different bacterial groups. For this analysis, we selected the most dominant bacterial groups, including Bacteroidetes and Firmicutes (that constitute the vast majority of the dominant human gut microbiota (Arumugam et al., 2011)), as well as Actinobacteria and Proteobacteria (which dominate mostly all environmental metagenome (Choudhari, Lohia & Grigoriev, 2014; Lee et al., 2005). We found a good overlap of clusters between the two samples. The sequences cluster in separate taxonomic groups in the reduced sequence composition space represented by PCA, and the same bacterial groups from different samples formed overlapping clusters (Fig. 1). This was in agreement with our expectation that the sequences of same bacterial groups would be similar in composition, even for such very diverse sources as leech gut and deep sea metagenomes. This suggested that the origin of sample was not creating any kind of bias in sequences generated by the same sequencing platform.

One noticeable feature in the grouping of Firmicutes by both the metagenomes was

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

observed: It formed two separate clusters in the 3d space of PCA plot, with both clusters containing sequences from both samples. This behavior was likely caused by a bimodal GC content distribution for the reads from this group in each sample (Fig. 2). When we considered the PCA load factors, we saw that the main contributions in PC1 and PC3 were provided by dinucleotides that were either extremely GC-rich or extremely GC-poor (Table 1). By comparing the result of PCA plot, load factors, and GC content, it can be concluded that although the similar grouping effect for both the platforms was obtained, the separation of different bacterial groups was achieved for each platform, and the separation was due to difference in the GC-rich dinucleotides in different groups.

Human Urine metagenome via MiSeq and Ion Torrent

Further, we considered sequence data from a single source, but generated by two different platforms in order to examine if platforms introduced some kind of bias in the sequences. We analyzed human urine sample sequenced by MiSeq and Ion Torrent and found differences in nucleotide composition of the same bacterial groups for the two platforms. The V1-V2 hypervariable region of 16S rRNA (common for these samples) was used for the calculation of dinucleotide frequencies, and phylogenetic classification was performed as described earlier. In a PCA plot (Fig. 3) the two different technologies formed two distinct clusters. The same bacterial groups were considered as described earlier and all the groups clustered closer to their platform counterparts. Although the separation between the two platforms was small, there was no overlap between the two platforms, in contrast to the previous case. The first three principal components provided a good resolution, accounting for around 80 % of the variance. The top load factors again showed preference for word frequencies of extreme GC-rich or GC-poor dinucleotides (Table 2). The positive scatter region on PCA plot was occupied by Ion Torrent, indicating that this platform produced more GC-rich sequences compared to MiSeq. This was confirmed on the GC content plot (Fig. 4), where each individual bacterial group showed higher GC content when sequenced by Ion Torrent.

In general, these observations suggest that different sequencing technologies generate compositional bias in the sequences they produce. It is important to note that, in this section, significant bias in sequence composition of same sample was observed. This is in stark contrast to no bias observed when entirely different samples were sequenced using the same platform.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

However, GC content appeared to be the main contributor to the cluster separation in both cases (same platform or same sample).

Glacier Metagenome

As mentioned above, a motivation for this analysis was our earlier observation (Choudhari, Lohia & Grigoriev, 2014) of drastically different composition of hypervariable regions sequenced for two glaciers (in Alaska and Alps). We continued it in this study and compared the same snow samples from an Alaskan glacier using two sequencing platforms (Ion Proton and MiSeq). Although different sequencers were utilized, every other step prior to sequencing (i.e. the method of DNA isolation and preparation of libraries) was kept the same. The difference with the previous examples was that the sequences taken into account were from shotgun metagenomics, as opposed to specific PCR-amplified hypervariable regions of 16S rRNA in previous examples. From this shotgun dataset we selected V9 regions based on best match against refv9 database but such regions were of different length in the case of shotgun, possibly affecting the dinucleotide frequencies and taxonomic assignments as compared to the previous two cases, where hypervariable 16S rRNA regions were specifically sequenced.

The same bacterial groups were studied as described earlier. In a PCA plot (Fig. 5) two clusters are observed, and we used a 2d picture as it better shows a mirror symmetry along the PC1 axis. This symmetry is due to the fact that both plus and minus strand BLAST hits were considered together and it clearly illustrates that the PCA is very sensitive to detecting difference between reverse complimentary sequences. However, no platform-specific separation is seen. Instead, one can observe closeness of clusters corresponding to the same bacterial groups, although the clusters are not as clearly separated as in the case of the 454 sequencing.

Conclusion

In this paper we utilized PCA to show how the sequence composition changes with different sequencing technologies. We kept the same hypervariable regions of 16S rRNA while comparing two datasets in order to minimize any kind of regional bias. We compared V6 regions of 16S of two very different metagenomes: leech gut (Maltz et al., 2014) and deep-sea (Sogin et al., 2006), generated by 454 technology, and found the clustering of the data by phylogenetic group. We observed a good overlap of two metagenomes, and the same bacterial groups from both samples clustered together. Thus one should expect that if single sequencing technologies is used, no bias

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

is observed in the data no matter what is the origin of sample. Notably, we observed different clusters when same type of human urine sample (Salipante et al., 2014) was sequenced via MiSeq and Ion Torrent. The V1-V2 variable regions of 16S were utilized for the analysis. Here, the bacterial groups clustered together with their platform counterparts, rather than grouping with same bacterial groups. We found that the sequence composition changes when different sequencing platforms are used, so we need to be cautious when interpreting results in comparative studies. The variability in the data appears to be due to GC-content of the reads and it is also manifested in load factors for extreme GC-rich and GC-poor dinucleotides. Furthermore, with the exception of different sequencing platforms, we kept all aspects such as the sample, DNA extraction, and library preparation protocols identical when comparing glacier metagenomes and did not observe platform-specific separation. It appears that the platform bias is visible in the reads produced by amplifying hypervariable regions in 16S rRNA but not in shotgun metagenome reads aligned to such hypervariable regions.

One important conclusion from this study is that sequence composition of metagenomic data varied depending on the sequencing platforms used. Thus one needs to be sensitive to such differences while comparing and combining metagenomic data produced by different technologies.

References

Abusleme L, Hong BY, Dupuy AK, Strausbaugh LD, Diaz PI. 2014. Influence of DNA extraction on oral microbial profiles obtained via 16S rRNA gene sequencing. *Journal of Oral Microbiology* **6**:10.3402/jom.v6.23990. ECollection 2014.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215(3)**:403-10.

Amann RI, Ludwig W, Schleifer KH. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews* **59(1)**:143-69.

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, et al. 2011. Enterotypes of the human gut microbiome. Nature **473(7346)**:174-80.

Choudhari S, Smith S, Owens S, Gilbert JA, Shain DH, Dial RJ, Grigoriev A. 2013. Metagenome sequencing of prokaryotic microbiota collected from byron glacier, alaska. *Genome Announcements* **1(2)**:e0009913-13.

Choudhari S, Lohia R, Grigoriev A. 2014. Comparative metagenome analysis of an alaskan glacier. *Journal of Bioinformatics and Computational Biology* **12(2)**:1441003.

Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'Toole PW. 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research* **38(22)**:e200.

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, et al. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10(3)**:R32,2009-10-3-r32. Epub 2009 Mar 27.

Huse SM, Mark Welch DB, Voorhis A, Shipunova A, Morrison HG, Eren AM, Sogin ML. 2014. VAMPS: A website for visualization and analysis of microbial population structures. *BMC Bioinformatics* **15:**41,2105-15-41.

Lee KB, Liu CT, Anzai Y, Kim H, Aono T, Oyaizu H. 2005. The hierarchical system of the 'alphaproteobacteria': Description of hyphomonadaceae fam. nov., xanthobacteraceae fam. nov. and erythrobacteraceae fam. nov. *International Journal of Systematic and Evolutionary Microbiology* **55(Pt 5)**:1907-19.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. 2012. Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology* **2012**:251364.

Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. 2012. Direct comparisons of illumina vs. roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* **7(2)**:e30087.

Maltz MA, Bomar L, Lapierre P, Morrison HG, McClure EA, Sogin ML, Graf J. 2014. Metagenomic analysis of the medicinal leech gut microbiota. *Frontiers in Microbiology* **5**:151.

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, et al. 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods* **4(6)**:495-500.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

Morgan JL, Darling AE, Eisen JA. 2010. Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One* **5(4)**:e10209.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO. 2007. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35(21)**:7188-96.

Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, Harkins TT, Cookson BT, Hoffman NG. 2014. Performance comparison of illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Applied and Environmental Microbiology* **80(24)**:7583-91.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, et al. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75(23)**:7537-41.

Simon C, Wiezer A, Strittmatter AW, Daniel R. 2009. Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Applied and Environmental Microbiology* **75(23)**:7519-26.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* **103(32)**:12115-20.

Wooley JC, Godzik A, Friedberg I. 2010. A primer on metagenomics. *PLoS Computational Biology* **6(2)**:e1000667.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

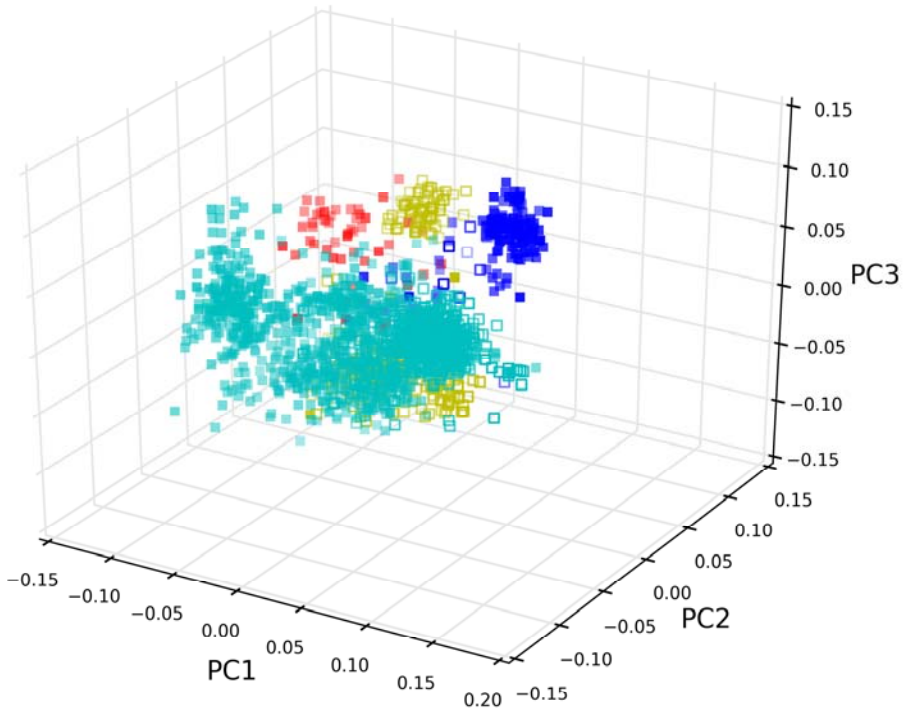

Figure 1: Deep sea metagenome (solid boxes) and Leech gut (empty boxes) data generated through 454: Nucleotide word frequency principal component analysis (PCA) of V6 hypervariable regions of 16S rRNA sequences. PCA orientation with phylogenetic classification at the phylum-level (Actinobacteria = blue, Bacteroidetes = red, Firmicutes = yellow, Proteobacteria = turquoise).

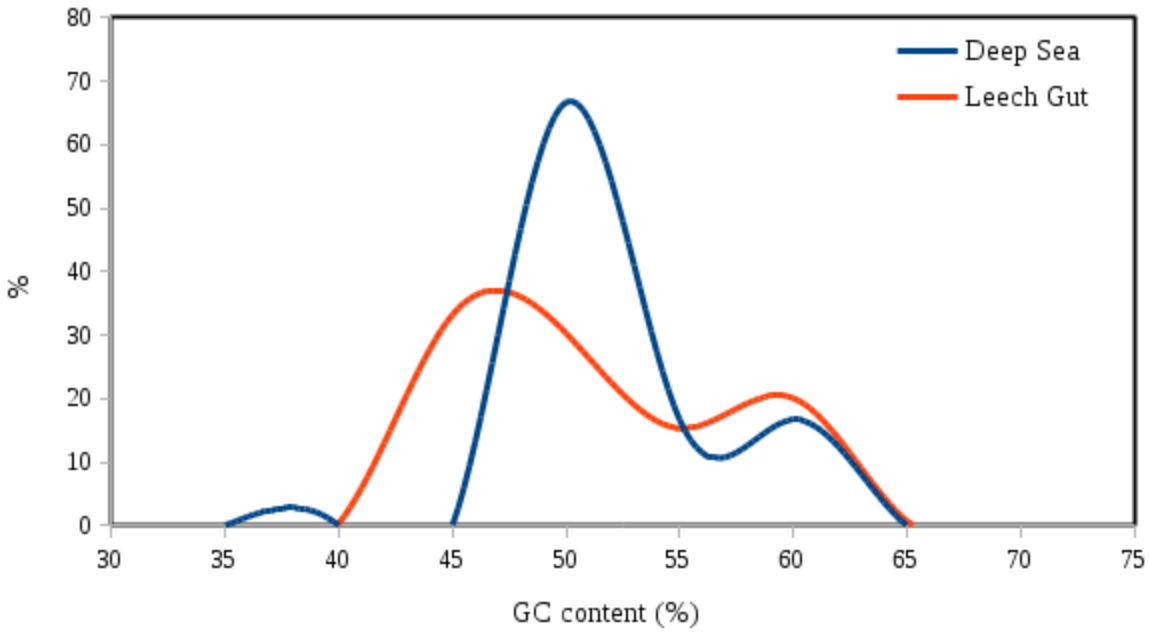

Figure 2: GC-curve of diverse metagenomic datasets: The GC% profile of bacterial group, Firmicutes of deep sea metagenome (blue) and leech gut metagenome (red).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

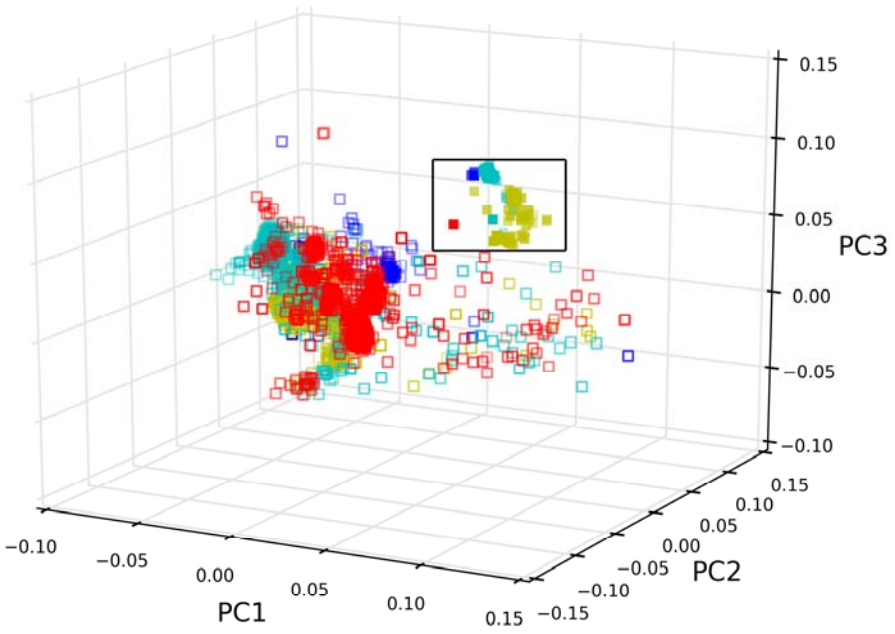

Figure 3: Human urine metagenome generated via Illumina MiSeq (empty boxes) and Ion Torrent (solid boxes): Dinucleotide word frequency principal component analysis (PCA) of V1-V2 hypervariable regions of 16S rRNA sequences. PCA orientation with phylogenetic classification at the phylum-level (Actinobacteria = blue, Bacteroidetes = red, Firmicutes = yellow, Proteobacteria = turquoise). The selected area is drawn around tightly clustered Ion Torrent sequences to highlight their separation from the MiSeq.

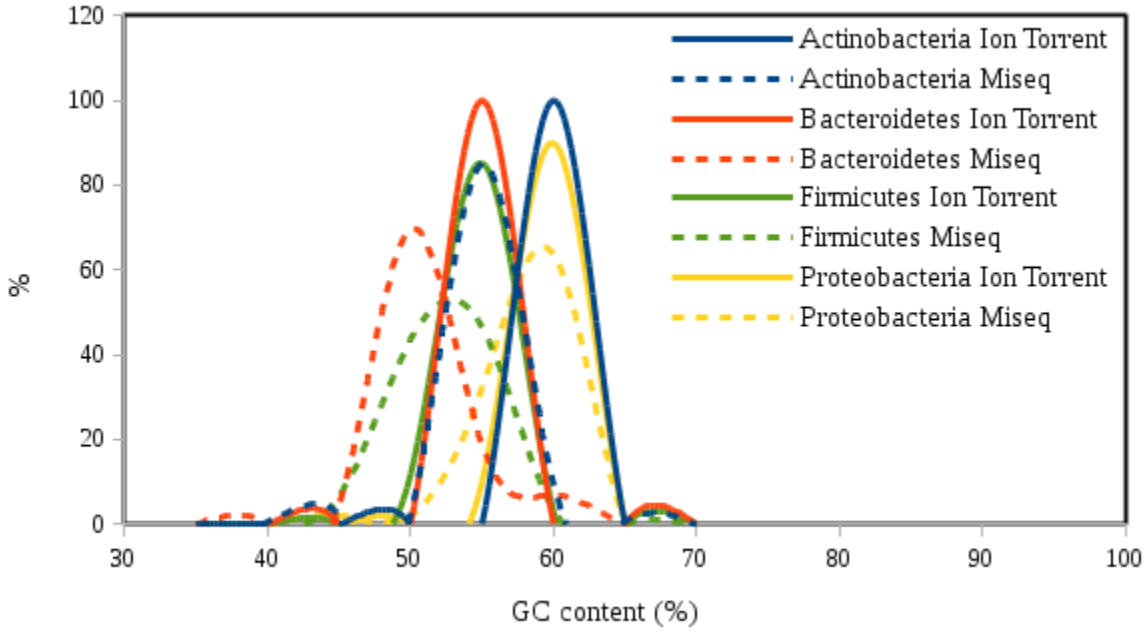

Figure 4: GC-curve of different bacterial groups: The GC% profile of bacterial group, Actinobacteria (blue), Bacteroidetes (red), Firmicutes (green) and Proteobacteria (yellow) sequenced via Ion Torrent (continuous) and Illumina MiSeq (dotted).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

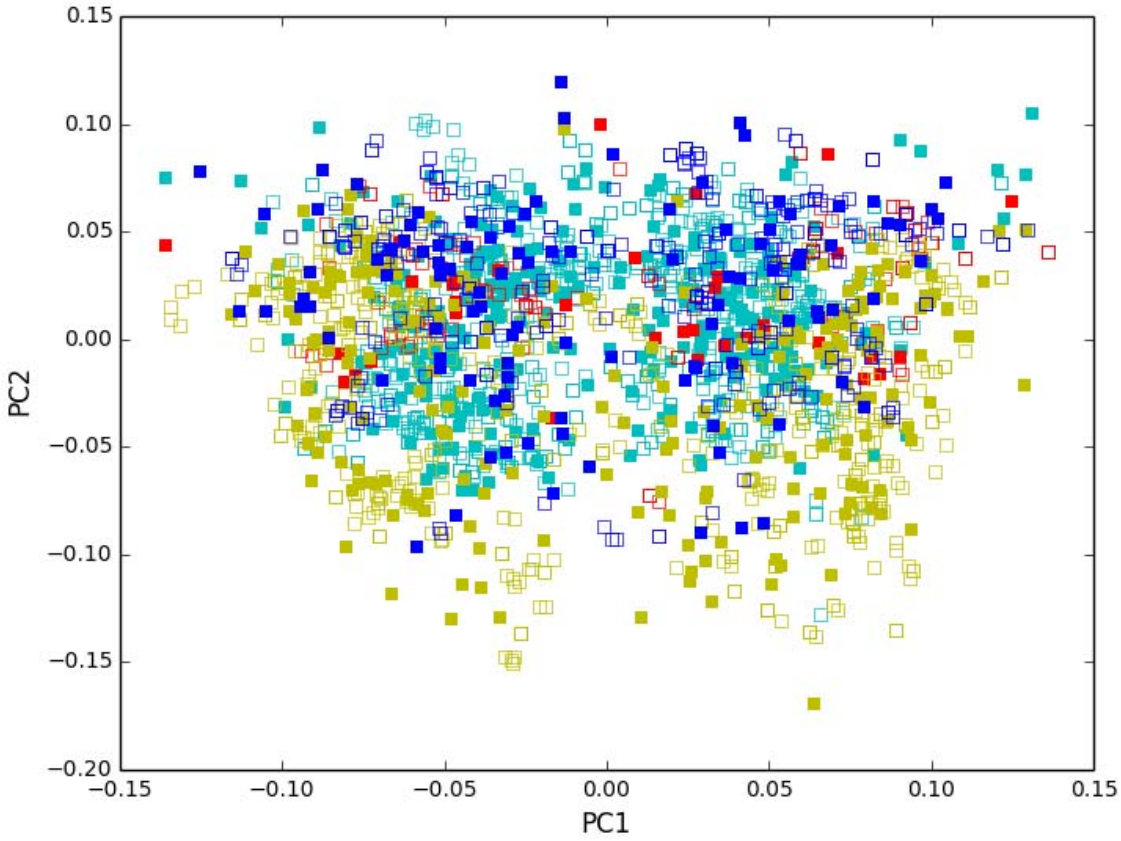

Figure 5: Glacier metagenome sequenced via Ion Proton (solid boxes) and MiSeq (empty boxes)- Dinucleotide word frequency principal component analysis (PCA) of V9 hyeprvariable region of 16S rRNA sequences. PCA orientation with phylogenetic classification at the phylum-level (Actinobacteria = blue, Bacteroidetes = red, Firmicutes = yellow, Proteobacteria = turquoise).

Table 1: First, second and third principal components of dinucleotide word frequencies and their corresponding load factors

	PC1 Component	PC2 Component	PC3 Component
	Deep Sea - Leech Gut[454]		
AA	-0.07	**0.28**	0.00
AC	-0.09	-0.27	-0.17
AG	0.07	-0.28	-0.03
AT	-0.20	0.10	**0.32**
CA	-0.01	0.05	-0.03
CC	**0.40**	-0.24	-0.11
CG	-0.21	-0.54	0.11
CT	0.02	0.12	-0.49
GA	-0.14	-0.20	-0.32
GC	0.14	-0.21	-0.16
GG	**0.65**	-0.20	**0.37**
GT	0.02	0.02	0.32
TA	-0.05	0.28	0.21
TC	-0.21	-0.31	-0.11
TG	0.16	**0.29**	0.03
TT	-0.45	0.12	0.30

The grays with boldface colored boxes correspond to frequencies having top two highest coefficients of variables (load factors) and the light gray colored represent bottom two lowest coefficients of variables.

454 - 454 Sequencing

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

Table 2: First, second and third principal components of dinucleotide word frequencies and their corresponding load factors

	PC1 Component	PC2 Component	PC3 Component
	Human urine [M, IT]		
AA	-0.16	-0.55	-0.18
AC	-0.15	-0.08	0.09
AG	-0.17	-0.28	-0.07
AT	0.14	-0.14	-0.21
CA	0.11	-0.37	**0.41**
CC	0.14	0.05	**0.44**
CG	-0.06	0.09	0.20
CT	0.08	0.28	-0.15
GA	-0.23	-0.08	-0.33
GC	-0.10	0.10	0.10
GG	-0.57	**0.40**	0.23
GT	0.08	0.14	0.05
TA	-0.05	-0.08	-0.32
TC	**0.36**	-0.01	0.28
TG	-0.01	**0.36**	-0.28

TT	**0.58**	0.18	-0.24

The grays with boldface colored boxes correspond to frequencies having top two highest coefficients of variables (load factors) and the light grays colored represent bottom two lowest coefficients of variables.

M - Illumina Miseq

IT - Ion Torrent

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.734v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014