Journal of Microbiological Methods 107 (2014) 30–37

Contents lists available at ScienceDirect

Journal of Microbiological Methods

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

Data analysis for 16S microbial profiling from different benchtop sequencing platforms

Victor S. Pylro [a,c,*,1], Luiz Fernando W. Roesch [b,**,1], Daniel K. Morais [a,c], Ian M. Clark [c], Penny R. Hirsch [c], Marcos R. Tótola [a]

[a] Microbiology Department, Universidade Federal de Viçosa, Viçosa, MG 36570-900, Brazil
[b] Universidade Federal do Pampa, São Gabriel RS, 97300-000, Brazil
[c] AgroEcology Department, Rothamsted Research, Harpenden, Herts AL52JQ, United Kingdom

ARTICLE INFO

Article history:
Received 12 August 2014
Received in revised form 26 August 2014
Accepted 27 August 2014
Available online 3 September 2014

Keywords:
Next generation sequencing
Microbial community analysis
Amplicons
Alpha diversity
Beta diversity

ABSTRACT

Progress in microbial ecology is confounded by problems when evaluating results from different sequencing methodologies. Contrary to existing expectations, here we demonstrate that the same biological conclusion is reached using different NGS technologies when stringent sequence quality filtering and accurate clustering algorithms are applied.

1. Introduction

The study of microbial communities in relationship with their environment/host is essential for understanding ecosystem dynamics. Scientific and technological advances including metagenomics and metatranscriptomics have revolutionized the traditional approaches used to study biological resources over the past decade. Recent advances in nucleic acid extraction procedures and next generation sequencing (NGS), allow comparative analysis of whole microbial community diversity, abundance and functional genes at far greater sequencing depths than ever before.

Initially, high-throughput sequencing was used mainly for large-scale applications, since its development was focused on the race to the '$1000 human genome' (Hayden, 2014; Loman et al., 2012). However, currently at least two different benchtop high-throughput sequencing instruments compete for smaller scale applications such as sequencing bacterial genomes and amplicons. The Ion Torrent Personal Genome Machine (PGM) works in a similar way to the recently discontinued 454 GS platform (replaced by other NGS platforms). This technology exploits emulsion PCR and also employs a sequencing-by-synthesis approach, but uses a modified silicon chip to detect hydrogen ions released during base incorporation by DNA polymerase (Rothberg et al., 2011). The Illumina MiSeq technology performs solid-surface PCR amplification, resulting in clusters of identical DNA fragments. It is based on the reversible-terminator sequencing by synthesis technology used by all Illumina sequencing platforms, but reduces the run time by using a smaller flow cell, reduced imaging time and faster microfluidics, making it useful for medium and small-scale sequencing projects. The current cost per megabase sequenced has fallen to a level where these analyses are routinely performed in many research laboratories around the world.

Currently, microbial ecologists have many open-source software packages available to perform robust statistical methods to explore their datasets (Angiuoli et al., 2011; Caporaso et al., 2010; Edgar, 2013; Quast et al., 2013; Schloss et al., 2009). While combinations of data analysis methods provide new insights when assessing quite detailed information on the microbial communities, the different steps, parameters and algorithms adopted by each study make it hard to compare approaches. This critical issue highlights the need for different analyses to follow the same workflow. The Brazilian Microbiome Project (BMP - http://www.brmicrobiome.org/) (Pylro et al., 2014) has proposed the assembly of a Brazilian Metagenomic Consortium/Database. At present, many metagenomic projects underway in Brazil are widely known, and BMP's main goal is to co-ordinate and standardize approaches within these, together with future projects. One of the challenges is the development/dissemination of uniform standards for

* Correspondence to: V. Pylro, Department of Microbiology, Universidade Federal de Viçosa, Av. P.H. Rolfs, SN, Viçosa, MG, 36570-900, Brazil. Tel.: +55 31 38992903.
** Corresponding author. Tel.: +55 55 32326075.
E-mail addresses: victor.pylro@gmail.com (V.S. Pylro), luizroesch@unipampa.edu.br (L.F.W. Roesch).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.mimet.2014.08.018
0167-7012/

experimental design and data analysis that can be integrated with existing consortia and standards. The Earth Microbiome Project (EMP — http://www.earthmicrobiome.org/) is a global effort to characterize microbial communities in a systematic way. EMP applies, as a standard for 16S rRNA profiling, the protocol proposed by Caporaso et al. (2012) for paired-end 16S rRNA community sequencing on the Illumina HiSeq/MiSeq platform, using bacteria/archaeal primers 515F/806R, together with the open-reference OTU picking protocol on QIIME (Caporaso et al., 2010). However, the choice of one particular NGS platform is a limiting factor when attempting to employ this kind of analysis routinely, especially in developing countries such as Brazil. It is therefore also important to evaluate other methods using different NGS platforms, and to ensure that data already generated is compatible.

QIIME default parameters were initially established for 454 pyrosequencing raw data (http://qiime.org/tutorials/tutorial.html) and later to Illumina technology (Caporaso et al., 2012). Recently, Bokulich et al. (2013) presented new guidelines and defined improved default parameters for quality-filtering of Illumina reads, evaluated by testing their effects on taxonomic classification, alpha and beta diversity estimations, using the QIIME pipeline. However, until now, no standard pipeline to analyze PGM 16S reads was available.

Two recent reviews and comparative studies of NGS technologies/platforms are valuable (Jünemann et al., 2013; Loman et al., 2012), but compare only whole genome sequencing of *Escherichia coli* isolates. Prior reports evaluating 16S phylogenetic profiling outcomes do not encompass the new technologies now available, such as PGM. The rate at which sequencing technologies are evolving, with increased throughput, read length, and base quality, highlights the need for ongoing evaluation (Pallen, 2013).

Here, we evaluate the 16S rRNA phylogenetic profiling of two benchtop NGS platforms, Illumina MiSeq and PGM, and the associated data analysis, aiming to make them comparable. This validation is essential for countries like Brazil that must ensure existing investments in different NGS platforms by government scientific agencies are used effectively and collaboratively.

2. Material and methods

2.1. Sampling site and soil analysis

Soils were sampled at 3 sites on Trindade Island (coordinates: 20°29′–20°32′ S and 29°17′–29°21′ W) in an expedition supported by the Brazilian Navy and PROTRINDADE Research Program through March and April 2011 (Fig. 1). The National Council for Scientific and Technological Development (CNPq) provided all approvals and permits (project grant number 405544/2012-0 and authorization access to genetic resources process number 010645/2013-6) to conduct the study within this protected area. The field study did not involve endangered or protected species. Fifteen soil cores (1.5 cm in diameter and 10 cm in depth) were collected from each sampling point and then cores were bulked, sieved (<2 mm), resulting in one replicate. Samples were taken in duplicates, stored at 4 °C, and then transported to Scientific Station in Trindade Island (ECIT), and kept at −20 °C.

2.2. Molecular analyses

2.2.1. DNA extraction and quality check

Genomic DNA was extracted and purified from each soil sample (10 g), using the PowerMax® Soil DNA Isolation Kit (MoBio Laboratories, Carlsbad, CA), following the manufacturer's instructions. The purity of the extracted DNA was checked with the Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA) (260/280 nm ratio), and it was quantified by Qubit® 2.0 fluorometer using the dsDNA BR Assay kit (Invitrogen™). The integrity of the DNA was also confirmed by electrophoresis in a 0.8% agarose gel with 1× TAE buffer.

2.2.2. Illumina® and Ion Torrent® high-throughput sequencing of bacterial/archaeal 16S

Bacterial and archaeal 16S rRNA genes were amplified using primers 515F (5′-GTGCCAGCMGCCGCGGTAA-3′) and 806R (5′-GGACTACHVG GGTWTCTAAT-3′) for paired-end microbial community and sequenced on the Illumina MiSeq platform (Caporaso et al., 2012) at the High-throughput Genome Analysis Core (HGAC), Argonne National Laboratory (USA). Similarly, bacterial and archaeal 16S rRNA genes were also amplified using the same primer set, but for sequencing on the Ion Torrent® (PGM) platform at the Life Technologies Training Center (Brazil). Briefly, amplicons containing the adaptors and the Ion Xpress barcode (001–006) sequences were purified using the E-Gel® SizeSelect™ 2% Agarose, and concentrated with the AMPure Beads 1.2× (Beckman Coulter). Emulsion PCR was carried out using the Ion OneTouch 2™ with the Ion Template PGM™ OT2 400 Kit (Life Technologies) according to the manufacturer's instructions. Sequencing of the amplicon libraries was carried out on an Ion 318™ Chip Kit v2 using the Ion Torrent PGM system according to the supplier's instructions. After sequencing, all reads were filtered by the PGM software to remove low quality and polyclonal sequences.

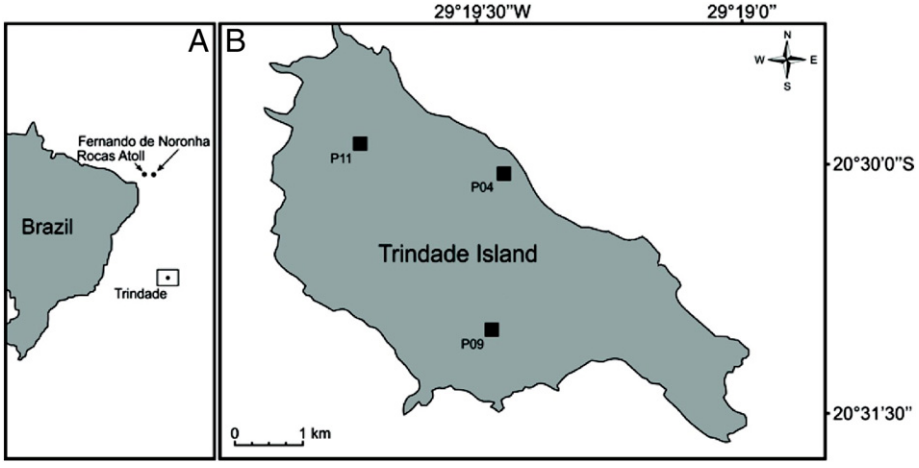

Fig. 1. Map of the sample sites. (A) The Brazilian coast and Atlantic Ocean with the Trindade Island in relief. (B) Trindade Island. Sample sites are indicated by black squares and respective identifications.

2.2.3. Data analysis and standardization

We applied four different bioinformatic strategies to empirically find the best pipeline to compare NGS data for 16S profile analysis irrespective of the sequencing technology, as follows:

a) *Default parameters on QIIME pipeline*: the 16S Illumina paired-end reads data of both, forward and reverse amplicons were assembled in contigs using "fastq-join" (Aronesty, 2013). The output file (*.fastq*) together with barcode file (*.fastq*), plus its respective mapping file (*.txt*) were then processed and sorted using the default parameters on QIIME version 1.8.0 (*split_libraries_fastq.py*) (Caporaso et al., 2012). Quality filtered reads were assigned to OTUs applying the open-reference OTU picking protocol using the QIIME toolkit (Caporaso et al., 2010). Briefly, demultiplexed sequences were clustered into OTUs at 97% similarity (*pick_open_reference_otus.py*). In this step, reads are clustered against a reference collection, and any reads, which do not match the reference data are subsequently clustered by using a *de novo* approach. Moreover, this script also applies a taxonomy assignment, sequence alignment, and tree-building steps. Further, diversity analyses were performed by running a workflow on QIIME, with the script *core_diversity_analyses.py*. Similarly, the 16S PGM reads data (*.fasta* and *.qual* files) were also processed using the default parameters, but applying a different script to demultiplexing and quality filtering steps on QIIME (*split_libraries.py*), appropriated to *.fasta* and *.qual* files (minimum sequence length = 200; minimum average quality score = 25). These parameters are currently proposed as a standard EMP Protocol for 16S taxonomic assignments.

b) *Modified parameters on QIIME pipeline*: this strategy comprises the same approach used before, but quality filtering parameters were changed, as follows: the minimum average quality score was kept equal or greater than Q30, and *usearch_qf* (usearch quality filter) pipeline script, built using USEARCH (Edgar, 2010), was used to perform filtering of noisy sequences, chimera checking, and OTU picking at 90% and 97% sequence similarity thresholds (separately). The output file was then used to pick a representative set of sequences using the *pick_rep_set.py* script, and further assigned to taxonomy using *uclust*, and Greengenes database (13_08) as a reference on QIIME (*assign_taxonomy.py*). Output files (seqs_otu.txt and tax_assignments.txt) were used to construct an OTU table (BIOM format) (*make_otu_table.py*). Further, diversity analyses were performed by running a workflow on QIIME, with the script *core_diversity_analyses.py*. We hypothesized that more stringent parameters to filter low-quality and noisy sequences could overcome the different errors/bias delivered by each sequencing technology.

c) *UPARSE + QIIME pipeline*: the recently published OTU clustering method, UPARSE (Edgar, 2013), together with final steps on QIIME, was also applied to the same datasets. The UPARSE standard pipeline was modified to work with both, Illumina Miseq and PGM data. This pipeline produced two output files, an OTU table in txt format (further converted into *.biom* format) and a set of representative sequences for each OTU in *fasta* format. The representative sequences were then assigned to taxonomy using *uclust*, and Greengenes database (13_08) as a reference on QIIME (*assign_taxonomy.py*) and the taxonomy was added to the OTU table by using the set of scripts described in http://biom-format.org/. Further, diversity analyses were performed by running a workflow on QIIME, with the script *core_diversity_analyses.py*. These complete pipelines are available on http://www.brmicrobiome.org/.

2.2.4. Comparing technologies and bioinformatic strategies

In order to compare the performance of each pipeline in clustering data from different sequencing platforms, we performed a Procrustes analysis using QIIME (http://qiime.org/tutorials/procrustes_analysis.html). Briefly, two coordinate matrices (one from each dataset: Illumina or PGM) generated by the script *beta_diversity_through_plots.py* (within the *core_diversity_analyses.py* workflow) were transformed by the script *transform_coordinate_matrices.py*. The results were visualized using QIIME by running *compare_3d_plots.py* script.

Moreover, OTU tables from different sequencing technologies, obtained by the same bioinformatic strategy, were then combined (*merge_otu_tables.py* script on QIIME), and further submitted to phylogenetic beta diversity analysis using UniFrac. The resulting distance matrix was used to cluster quality evaluation (*cluster_quality.py* script on QIIME).

We calculated the alpha diversity by using four different metrics as follow: Phylogenetic Diversity Whole Tree, Shannon, Simpson, and Equitability. Data were expressed as the average of two replicates, and submitted to analysis of variance. The results from each bioinformatic strategy were contrasted using Tukey's test ($p < 0.05$). Taxonomic summary files (resulting files from *core_diversity_analyses.py* workflow) were also compared by computing the correlation coefficient between pairs of samples from different sequencing platforms, in each taxonomic level, separately (*compare_taxa_summaries.py* on QIIME).

2.2.5. Supporting data

All sequence data and metadata were deposited in the Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/Traces/sra/) under BioProject PRJNA241041 ("Brazilian Microbiome Project: Standardizing 16S profiling data analysis from different benchtop sequencing platforms"; http://www.ncbi.nlm.nih.gov/bioproject/241041). Mapping files are provided as Supplementary information.

3. Results

Taxon-based and phylogeny-based approaches are commonly used in microbial ecology studies. We applied these analytical methods to verify the robustness of our bioinformatic strategies and determine the best approach for reconciling data from different benchtop sequencing platforms. These comparisons formed two groups: beta diversity (between-sample diversity comparison) and alpha diversity (within-sample diversity) approaches. As the sequencing depth can affect both, all diversity-related experiments were performed by using rarefaction (a random collection of sequences from a sample) ensuring that the same number of reads, equivalent to those in the smallest sample, were compared. Moreover, taxonomic classification was also evaluated.

3.1. Comparing technologies using beta diversity approaches

One of the first tests performed by microbial ecologists dealing with environmental samples is to determine the broad trends of similarities and differences among samples through cluster analysis. The Principal Coordinates Analysis (PCoA) is among the most used techniques (Bokulich et al., 2013). To test whether the beta diversity conclusions are consistent regardless of sequencing technology, the Procrustes analysis was applied to all of the bioinformatic strategies using unweighted and weighted UniFrac metrics (Figs. 2 and 3). The criterion to select the best fit adopted by Procrustes analysis is the minimization of a residual sum of squares after matching (M^2) which measures the remaining "lack of fit" of one configuration to the other (Krzanowski, 1990). Then the M^2 values might be used as a distance measure between any two PCoA comparisons where the smallest M^2 indicates a small difference between two plots.

The Procrustes analysis of unweighted (Fig. 2) and weighted (Fig. 3) UniFrac principal coordinate matrices revealed that samples sequenced on both Illumina and PGM were significantly correlated irrespective of the pipeline adopted. Nevertheless, the M^2 calculated for the data analyzed by using QIIME 1.8.0 with the default parameters presented small values ($M^2 = 0.019$, $p < 0.005$ — unweighted UniFrac; $M^2 = 0.024$, $p < 0.001$ — weighted UniFrac metric), followed by UPARSE ($M^2 = 0.030$, $p < 0.011$ — unweighted UniFrac; $M^2 =$

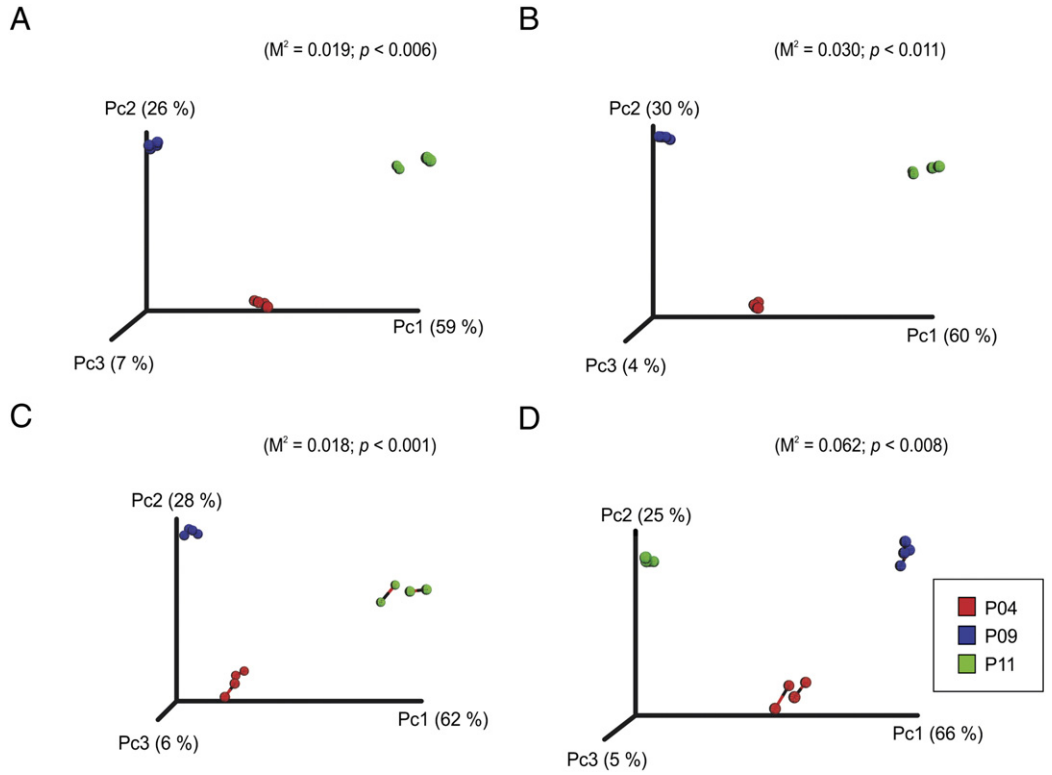

Fig. 2. Procrustes plot comparing principal coordinates of unweighted UniFrac distances, from paired sample sequences on PGM and Illumina, connected by the black line. Default parameters on QIIME pipeline (A); UPARSE + QIIME pipeline (B); and modified parameters on QIIME pipeline: OTU picking at 90% (C) and 97% (D) similarity thresholds. M^2 = minimization of residual sum of squares after matching. p = Monte Carlo p-value.

0.031, p < 0.002 — weighted UniFrac metric) indicating a very good fit, meaning that the sequencing technologies produced results that are in close agreement when analyzed by this pipeline. It is also important to highlight the effect of the relative abundance of sequences as contrasted by unweighted and weighted UniFrac. If 90% of sequence similarity threshold was applied, compared to the more stringent parameter

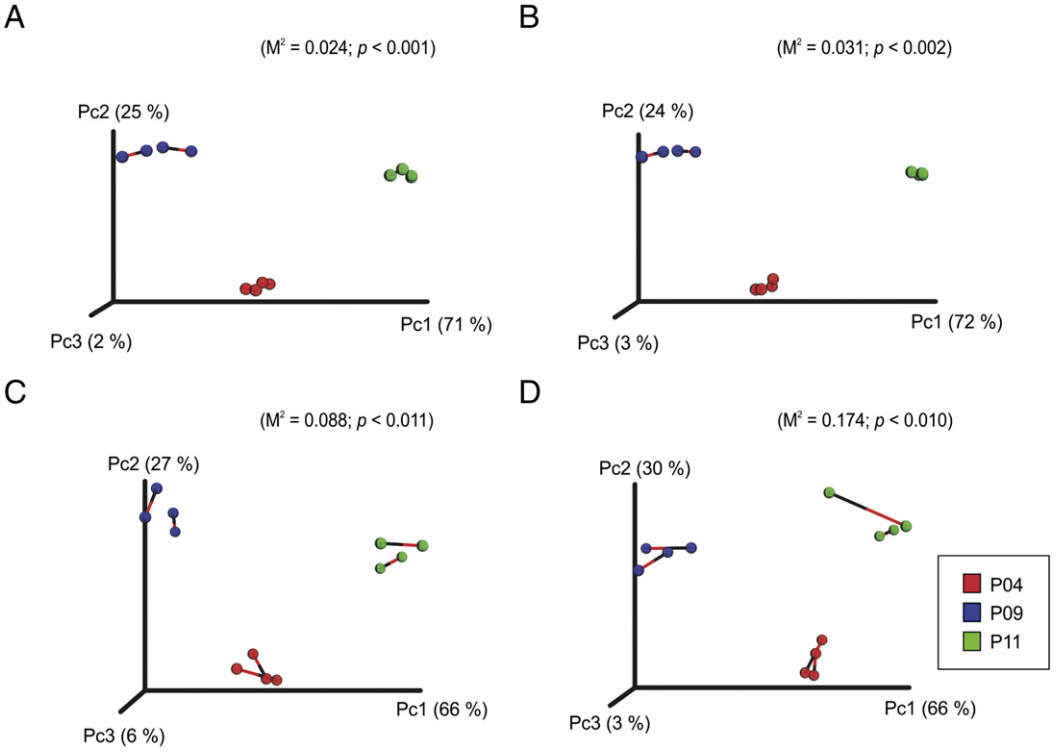

Fig. 3. Procrustes plot comparing principal coordinates of weighted UniFrac distances, from paired sample sequences on PGM and Illumina, connected by the black line. Default parameters on QIIME pipeline (A); UPARSE + QIIME pipeline (B); and modified parameters on QIIME pipeline: OTU picking at 90% (C) and 97% (D) similarity thresholds. M^2 = minimization of residual sum of squares after matching. p = Monte Carlo p-value.

of 97%, an even better fit was apparent. The M^2 was 0.018, $p < 0.001$ — unweighted UniFrac; 0.062, $p < 0.008$ — weighted UniFrac metric for the 90% pipeline (Figs. 1C and 2C), 0.018, $p < 0.001$ — unweighted UniFrac; and 0.174, $p < 0.010$ — weighted UniFrac metric for the 97% pipeline (Figs. 2D and 3D).

To confirm the results obtained by the Procrustes analysis, the dissimilarity within samples from each cluster in the PCoA was also computed (cluster quality evaluation). This analysis measures the dissimilarity ratio between and within clusters. Considering the natural biological variation found within our soil samples, the cluster quality was calculated for samples separated by each technology and by merging the data from different sequencing technologies within the same soil sample. To validate the hypothesis of no difference between sequence technologies, the dissimilarity within a cluster calculated based on data produced by each sequence technology must be lower or equal to the dissimilarity calculated when data from both technologies were combined.

Both Procrustes and cluster quality analyses converged. The environmental variation was greater than the variation between sequencing technologies, confirming our hypotheses that outcomes from each sequencing platform could provide compatible data (Table 1). The dissimilarity calculated for individual sequencing technologies was about two times higher than those from combined data, meaning more clustering in the latter.

3.2. Comparing technologies using alpha diversity approaches

Alpha diversity indices measure the taxon diversity within an individual sample. The metric Phylogenetic Diversity (PD) Whole Tree, is based on a phylogenetic tree and adds up all the branch lengths as a measure of diversity. If a new OTU is found, and is closely related to another OTU in the sample, it will generate a relatively small increase in diversity. However, new OTUs coming from a unique lineage in the sample will contribute significantly to increase the diversity. The PD Whole Tree index was higher in the default QIIME pipeline ($p < 0.05$), for all samples (Fig. 4A).

The Shannon index measures the average degree of uncertainty in predicting to what species an individual chosen at random from a collection of *S* species and *N* individuals will belong. The value increases as the number of species increases and as the distribution of individuals among the species becomes even. It will be zero if the sample in consideration has only one species, and would be maximal when all species in the sample have even abundances. Similar to PD Whole Tree, the Shannon index showed the same trend, with higher diversities being computed in the default QIIME pipeline ($p < 0.05$), for all samples (Fig. 4B).

The Simpson index indicates species dominance and reflects the probability of two individuals that belong to the same species being randomly chosen. The index increases as the diversity decreases (Simpson, 1949). It is represented as "1 — Dominance" (D) and varies from 0 to 1; where, zero represents no diversity and 1, the maximum diversity. According to the Simpson index results, a high diversity and low dominance were observed, irrespective of the samples or pipelines ($p < 0.05$) (Fig. 4C). Also, similar patterns in microbial evenness were observed for all samples and pipelines ($p < 0.05$), with the equitability ranging from 0.73 to 0.84 (Fig. 4D).

Table 1
Dissimilarity ratio of mean distances within clusters.

Pipelines	PGM	Illumina	PGM–Illumina combined
	Weighted UniFrac		
Default QIIME	4.03	4.86	2.19
Modified QIIME 90%	3.81	3.54	1.31
Modified QIIME 97%	3.38	3.11	1.28
UPARSE	4.20	4.99	1.44
	Unweighted UniFrac		
Default QIIME	2.02	4.86	1.12
Modified QIIME 90%	2.65	2.60	1.28
Modified QIIME 97%	2.94	2.68	1.30
UPARSE	2.56	2.80	1.27

3.3. Comparing technologies using taxonomic correlation

Taxonomic classification was also evaluated to compare sequencing technologies and data analysis pipelines. The script (*compare_taxa_summary.py*) sorted the taxa in the same order in each sample and compared the information provided by two taxonomic summary files (from phyla to genus levels) by computing the Pearson's correlation coefficient (Table 2). This step was useful to check if the output from both technologies and from different pipelines of analysis were correlated or not. The Pearson's correlation coefficient among pipelines varied from 0.9426 to 0.8790 for the phylum level and from 0.9189 to 0.7989 for the genus level showing that both technologies were highly correlated. The default QIIME followed by the UPARSE pipeline showed the best Pearson's correlations, indicating that these two strategies were more reliable to recover taxonomic information between sequencing technologies.

4. Discussion

The goal of this study was to find an effective bioinformatic pipeline to provide comparable outcomes using data from different sequencing platforms. To do this, we analyzed six natural microbial communities by using both PGM and Illumina MiSeq sequencing of the V4 hypervariable region of the 16S rRNA gene. Every sequencing technology presents positive and negative aspects as already observed in other studies. In 2012, Loman and co-workers compared the performance of the 454 GS Junior (Roche), MiSeq (Illumina) and Ion Torrent PGM (Life Technologies) by sequencing an isolate of *E. coli* O104:H4. They found that Illumina MiSeq presented the lowest error rates (0.1 substitutions and <0.001 indels per 100 bases) while both Ion Torrent PGM and 454 GSJ produced homopolymer-associated indel errors (1.5 and 0.38 errors per 100 bases, respectively). Further, Jünemann et al. (2013), reproduced the experiments of Loman et al. (2012), providing new perspectives, especially showing improvements regarding indel errors, using new available sequencing kits and chip, on Ion Torrent platform (0.3955 errors per 100 bases for 200 bp kit, and 0.6722 errors per 100 bases for 400 bp kit). Similar error rates were found in 454 GSJ using Titanium Sequencing kit (0.4011 errors per 100 bases). It is important to highlight that SNP variants are usually almost 10× more frequent in nature than indel variants, and this update showed that new sequencing kits from PGM perform better than Illumina MiSeq regarding substitutions, with the PGM delivering up to three times fewer SNP-associated errors. In summary, the Illumina MiSeq performed better regarding throughput per run and indel errors, but the PGM was well suited for sequencing amplicons, showing greater recent improvements to the technology (Jünemann et al., 2013).

Raw data from next-generation sequencing platforms, such as PGM and Illumina, uses quality scores (Q), commonly expressed as *Phred* scores, to predict the base calling error probability. A stringent quality filtering is not required for genome sequencing since multiple reads are assembled to create a consensus sequence. On the other hand, marker-gene-based studies are dependent on high quality reads because consensus methods cannot be applied for error correction or removal of chimeric sequences. In this regard, it is very challenging to discriminate true biological data and between-sample distinctions from sequencing or PCR artifacts (chimeras and errors during amplification). Generally, read quality filters, such as that utilized by QIIME, uses an average Q score to filter reads by quality. The importance of how to include the Q score parameter in the read filter quality has been discussed when evaluating data from the same sequencing technology

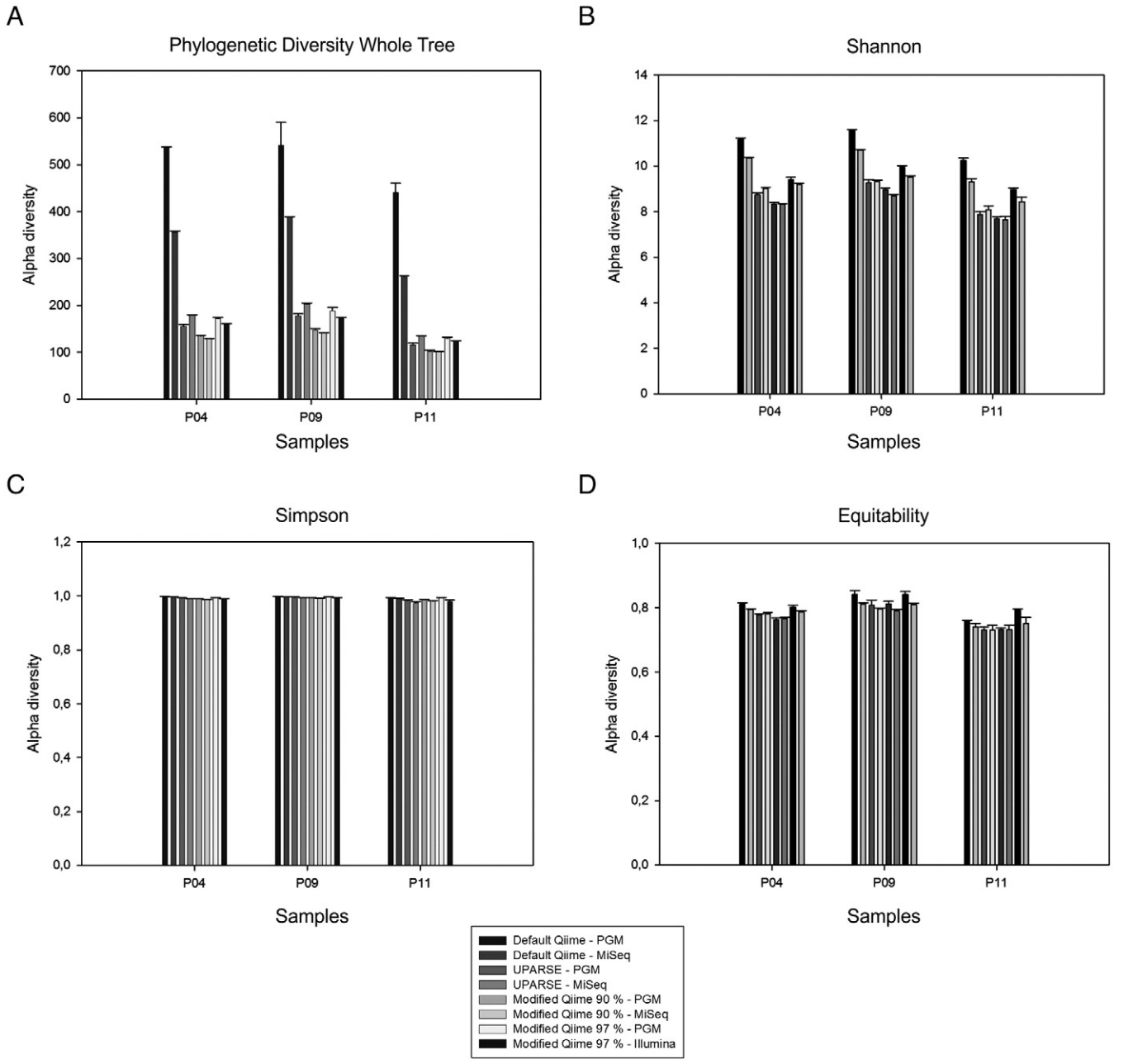

Fig. 4. Alpha diversity indices calculated for each sample in each bioinformatic strategy. Phylogenetic Diversity Whole Tree (A); Shannon (B); Simpson (C); and Equitability (D).

(Illumina HiSeq/MiSeq/GAIIx) (Bokulich et al., 2013). However, Q scores generated by each sequencing platform are not calculated in the same way, making further analysis using data from different technologies difficult. Those aspects associated with the errors involved in amplicon sequencing reinforce the need for stringent processing to reduce the percentage of misclassified reads and spurious sequences.

Table 2
Pearson's correlation between taxa found when comparing Illumina and PGM data.

	Default QIIME	Modified QIIME (90%)	Modified QIIME (97%)	UPARSE
	Pearson's correlation coefficient			
Phylum	0.9426	0.8833	0.8790	0.9265
Class	0.9168	0.8139	0.8115	0.8916
Order	0.9189	0.8008	0.7976	0.8910
Family	0.9222	0.8074	0.8075	0.8960
Genus	0.9189	0.7989	0.8050	0.8959

Stringent quality filtering parameters were shown to produce high-quality data (Bokulich et al., 2013; Jumpstart Consortium Human Microbiome Project Data Generation Working Group, 2012; Schloss, 2010; Schloss et al., 2011). However, bioinformaticians should develop platform-specific approaches for the PGM (Bragg et al., 2013), or universal standard data analysis that adequately accounts for the majority of errors introduced by different platforms.

The UPARSE clustering method applies a different strategy of quality filtering based not on average Q score, but on the maximum expected error, a better indicator of read accuracy. Moreover, it uses UPARSE-OTU clustering, a new algorithm able to perform chimera filtering and OTU clustering at the same time, and unlike QIIME it does not require any technology- or gene-specific parameters (Edgar, 2013). Given different types of reads (454 pyrosequencing and Illumina), UPARSE was reported to be more reliable for recovering biological sequences from mock communities than AmpliconNoise, mothur and QIIME, with improved accuracy of clustering OTUs.

Our analysis showed that the UniFrac distances between samples sequenced on both Illumina MiSeq and PGM were significantly correlated, as determined by Procrustes analysis of weighted and unweighted UniFrac principal coordinate matrices (Figs. 2 and 3). In general, our results display the same beta diversity trends when the default QIIME and UPARSE pipeline strategies were applied denoting a closer agreement between the data generated by PGM and Illumina MiSeq platforms. The UniFrac is a method for computing differences between microbial communities based on phylogenetic information (Lozupone and Knight, 2005). In contrast to OTU-based approaches, it does not need a rigid OTU definition based on a cutoff distance because it measures the phylogenetic distance between sets of taxa in a phylogenetic tree. Thus, the UniFrac is a robust method, able to detect the variation accounted by environment type, rather than by methodological artifacts (Liu et al., 2007). Caporaso et al. (2011) already confirmed the reliability of UniFrac distances to capture biological information irrespective of the sequencing platform. They found that Illumina and 454 pyrosequencing were significantly correlated, as determined by Procrustes analysis of unweighted UniFrac principal coordinate matrices, and Pearson correlation of UniFrac distances for pairs of samples.

The Shannon, Simpson and Equitability indices as well as the community comparisons based on taxonomic classifications are collectively called OTU-based approaches. For such approaches, the sequencing error rates are especially important because it increases the number of predicted OTUs and inflates richness estimates. The PD whole tree is not based on OTU counts, instead it quantifies the branch diversity of the phylogenetic tree (Chao et al., 2010), but it is also subject to bias caused by sequencing artifacts since base call errors can artificially inflate the number of branches in a phylogenetic tree. Despite QIIME default parameters and UPARSE provide similar outcomes for beta diversity, the first showed higher alpha diversity values for both indices, PD Whole Tree and Shannon. This might be explained by the elimination of singletons and chimera filtering adopted by the UPARSE algorithm. UPARSE discards singletons by default, which may eliminate few rare taxa (Edgar, 2013). According to Holmes and McMurdie (2012), filtering out species that are very rare is crucial because these species may appear with inflated influence under some re-weighting schemes and it can be beneficial to delineate true presence from simple noise effects. Using a synthetic 16S microbial community (mock community), Edgar (2013) reported that 25–67% of the QIIME OTUs were chimeric. Moreover, comparing trimmed OTUs with 150 nucleotides, he found a high identity between QIIME and the reference database OTUs at the beginning of the sequence, and lower towards the end, showing that the increased number of OTUs is likely to be a consequence of high error rates, predominantly at the end of the read where quality tends to drop.

We expected to improve the outcomes by using modified parameters on QIIME (90% and 97% sequence similarity thresholds) because both strategies involve the maintenance of only high quality reads (Q30), besides filtering out noisy sequences, and chimera checking (de novo approach followed by reference-based approach) before picking OTUs. Also, by applying less stringent cut-off criteria (90% similarity), we expected differences between datasets to be reduced allowing for an increased similarity between samples [although it is well known that any universal identity cut-offs fail to capture all putative ecotypes (Koeppel and Wu, 2013) mainly when less stringent cut-off criteria are applied]. Contrary to our expectations, those modified parameters did not perform well for the diversity calculations implemented in our study. Bokulich et al. (2013) presented similar results when applying an increased *Phred* quality (Q) score threshold and higher OTU filtration. This strategy probably resulted in an excessive reduction of the detectable diversity, leading to the observation of phylogenetic similarity where difference should actually occur. However, comparing only the modified parameters on QIIME pipelines, a more flexible sequence similarity threshold while picking OTUs, such as 90%, showed better results than 97%. It is well known that distinct sequence technologies introduce different biases and errors during short-amplicon sample analysis. For this reason, it is likely that less strict sequence similarity thresholds also tend to cause a clustering of slightly different OTUs, leading to the same effect observed when increased Q scores were chosen. In this case, this can be considered a positive phenomenon because biases and errors are then masked, and consequently not accounted during the OTU picking and clustering process.

The correlation between taxonomic classifications of data from different technologies was consistent with those results from beta and alpha analyses, supporting our hypothesis that differences between sequence technologies can be adjusted by adopting the correct pipeline of analysis.

QIIME is open source and USEARCH v7 (where the UPARSE algorithm is implemented) is freely available for academic purposes (32 bit version), and both are easy-to-install software packages. The time of analysis depends on the amount of user data and informatics infrastructure power, but generally, the analysis workflow on UPARSE is faster than on QIIME (data not shown). Intermediate bioinformatic skills are also required, because both packages work with a command line interface. Given the overall results, we suggest the use of UPARSE clustering method, in order to recover the most reliable 16S rRNA microbial community profiling results, when working with data from distinct sequencing technologies. Despite our findings that the default QIIME parameters show similar beta diversity to UPARSE, the lack of a chimera filtering step must be considered a significant problem, especially concerning alpha diversity approaches. We also agree that additional work is also required to address the effect of other parameters, such as the minimum number of consecutive high-quality base calls to retain a read, maximum number of consecutive low-quality base calls allowed before truncating a read, and the minimum number of representative sequences required to retain an OTU, on the analysis of data from different sequencing technologies.

This is the first effort to evaluate the 16S rRNA phylogenetic profiling of the two benchtop NGS platforms, Illumina MiSeq and PGM, in order to provide users with an effective pipeline to make these data comparable. All the command lines used to analyze our data are available in the BMP website (http://brmicrobiome.org) and we recommend their use for 16S rRNA data analysis.

Conflict of interest statement

The authors declare no conflict of interest.

Authors' contributions

V.S.P. and L.F.W.R. designed the experiment. VSP and D.K.M. collected the samples. VSP, D.K.M., I.M.C. and P.R.H. processed the samples and/or extracted DNA. V.S.P., D.K.M., I.M.C. and P.R.H. prepared *amplicon* libraries. V.S.P., L.F.W.R., I.M.C. and M.R.T. analyzed the results. V.S.P., L.F.W.R., I.M.C., P.R.H. and M.R.T. wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We would like to thank the Brazilian Navy and Captain Rodrigo Otoch Chaves for the logistic support while collecting samples. We are also grateful to Sarah Owens (Argonne National Laboratory/USA) for helping with Illumina 16S sequencing and Alysson Silvano, Alexia Leite, Raphael Fonseca and Beatriz Pinto from Life Technologies/Brazil, for providing all structures for 16S sequencing on PGM Ion torrent platform. CNPq grant 405544/2012-0, FAPEMIG (PACCSS program) and CAPES (PROEX program) funded this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.mimet.2014.08.018.

References

Angiuoli, S.V., Matalka, M., Gussman, G., Galens, K., Vangala, M., Riley, D.R., et al., 2011. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinforma. 12, 356. http://dx.doi.org/10.1186/1471-2105-12-356.

Aronesty, E., 2013. Comparison of sequencing utility programs. Open Bioinforma. J. 7, 1–8. http://dx.doi.org/10.2174/1875036201307010001.

Bokulich, N.A., Subramanian, S., Fait, J.J., Gevers, D., Gordon, J.I., Knight, R., Mills, D.A., Caporaso, J.G., 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat. Methods 10, 57–59.

Bragg, L.M., Stone, G., Butler, M.K., Hugenholtz, P., Tyson, G.W., 2013. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput. Biol. 9, e1003031. http://dx.doi.org/10.1371/journal.pcbi.1003031.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., et al., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7 (5), 335–336.

Caporaso, J.G., Lauber, C.L., Costello, E.K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., et al., 2011. Moving pictures of the human microbiome. Genome Biol. 12 (5), R50. http://dx.doi.org/10.1186/gb-2011-12-5-r50.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., et al., 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 6, 1621–1624.

Chao, A., Chiu, C., Jost, L., 2010. Phylogenetic diversity measures based on Hill numbers. Philos. Trans. R. Soc. B 365 (1558), 3599–3609.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26 (19), 2460–2461. http://dx.doi.org/10.1093/bioinformatics/btq461.

Edgar, R.C., 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Methods 10, 996–998.

Group Jumpstart Consortium Human Microbiome Project Data Generation Working Group, 2012. Evaluation of 16S rDNA-based community profiling for human microbiome research. PLoS ONE 7 (6), e39315. http://dx.doi.org/10.1371/journal.pone.0039315.

Hayden, E.C., 2014. Technology: the $1000 genome. Nature 507, 294–295. http://dx.doi.org/10.1038/507294a.

Holmes, S., McMurdie, P.J., 2012. Statistical data analysis challenges from the microbiome. In: Olsen, L., Choffnes, E.R., Mack, A. (Eds.), The Social Biology of Microbial Communities, pp. 275–303.

Jünemann, S., Sedlazeck, F.J., Prior, K., Albersmeier, A., John, U., Kalinowski, J., et al., 2013. Updating benchtop sequencing performance comparison. Nat. Biotechnol. 31, 294–296.

Koeppel, A.F., Wu, M., 2013. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial operational taxonomic units. Nucleic Acids Res. 41 (10), 5175–5188.

Krzanowski, W.J., 1990. Between-group analysis with heterogeneous covariance matrices: the common principal component model. J. Classif. 7, 81–98.

Liu, Z., Lozupone, C., Hamady, M., Bushman, F.D., Knight, R., 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. 35, e120.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J., et al., 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 30, 434–439.

Lozupone, C., Knight, R., 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. 71, 8228–8235.

Pallen, M.J., 2013. Updating benchtop sequencing performance comparison reply. Nat. Biotechnol. 31 (296-296).

Pylro, V.S., Roesch, L.F., Ortega, J.M., Amaral, A.M., Tótola, M.R., Hirsch, P.R., et al., 2014. Brazilian microbiome project: revealing the unexplored microbial diversity — challenges and prospects. Microb. Ecol. 67, 237–241.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41 (D1), D590–D596.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., et al., 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348–352.

Schloss, P.D., 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput. Biol. 6 (7), e1000844. http://dx.doi.org/10.1371/journal.pcbi.1000844.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., et al., 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541.

Schloss, P.D., Gevers, D., Westcott, S.L., 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE 6, e27310.

Simpson, E.H., 1949. Measurement of diversity. Nature 163, 688.