The ISME Journal (2010) 4, 852–861

www.nature.com/ismej

ORIGINAL ARTICLE

PANGEA: pipeline for analysis of next generation amplicons

Adriana Giongo[1], David B Crabb[1], Austin G Davis-Richardson[1], Diane Chauliac[1], Jennifer M Mobberley[1], Kelsey A Gano[1], Nabanita Mukherjee[2], George Casella[2,3], Luiz FW Roesch[4], Brandon Walts[3,5], Alberto Riva[3,5], Gary King[6] and Eric W Triplett[1,3]

[1]*Department of Microbiology and Cell Science, University of Florida, Gainesville, FL, USA;* [2]*Department of Statistics, University of Florida, Gainesville, FL, USA;* [3]*Genetics Institute, University of Florida, Gainesville, FL, USA;* [4]*Centro de Ciencias Agricolas, Universidade Federal do Pampa, São Gabriel, Rio Grande do Sul, Brazil;* [5]*Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL, USA and* [6]*Department of Biology, Louisiana State University, Baton Rouge, LA, USA*

High-throughput DNA sequencing can identify organisms and describe population structures in many environmental and clinical samples. Current technologies generate millions of reads in a single run, requiring extensive computational strategies to organize, analyze and interpret those sequences. A series of bioinformatics tools for high-throughput sequencing analysis, including pre-processing, clustering, database matching and classification, have been compiled into a pipeline called PANGEA. The PANGEA pipeline was written in Perl and can be run on Mac OSX, Windows or Linux. With PANGEA, sequences obtained directly from the sequencer can be processed quickly to provide the files needed for sequence identification by BLAST and for comparison of microbial communities. Two different sets of bacterial 16S rRNA sequences were used to show the efficiency of this workflow. The first set of 16S rRNA sequences is derived from various soils from Hawaii Volcanoes National Park. The second set is derived from stool samples collected from diabetes-resistant and diabetes-prone rats. The workflow described here allows the investigator to quickly assess libraries of sequences on personal computers with customized databases. PANGEA is provided for users as individual scripts for each step in the process or as a single script where all processes, except the χ^2 step, are joined into one program called the 'backbone'.
The ISME Journal (2010) **4,** 852–861; doi:10.1038/ismej.2010.16; published online 25 February 2010
Subject Category: microbial population and community ecology
Keywords: 16S rRNA; high throughput sequencing; microbial ecology; bioinformatics

Introduction

The analysis of amplified and sequenced 16S rRNA genes has become the most important single approach for the rapid identification and classification of prokaryotes. Amplicons from high-throughput sequencing by 454/Roche can generate many thousands of 16S rRNA sequences per sample, and unlike Sanger sequencing it does not require time-consuming clone library construction (Roesch *et al.*, 2007, 2009a; Hamady *et al.*, 2008; Liu *et al.*, 2008). These techniques provide short sequences (average of 100 to 400 bases) that have been applied to the study of microbial communities in aquatic and soil environments (Edwards *et al.*, 2006; Sogin *et al.*, 2006; Huber *et al.*, 2007; Roesch *et al.*, 2007; Brown *et al.*, 2009; Jones *et al.*, 2009; Miller *et al.*, 2009), analysis of rat, human and macaque gut microbiota (Liu *et al.*, 2007; Andersson *et al.*, 2008; Dethlefsen *et al.*, 2008; McKenna *et al.*, 2008; Roesch *et al.*, 2009a, b) and other human microniches (Dowd *et al.*, 2008; Luna *et al.*, 2007; Armougom and Raoult, 2008; Fierer *et al.*, 2008; Keijser *et al.*, 2008; Price *et al.*, 2009). As the use of high-throughput sequencing is increasing, user-friendly computing tools are needed to quickly and easily manipulate and analyze the results.

To reduce the cost of the new generation sequencing, a barcoding procedure was developed that incorporates an identifying nucleotide sequence at the 5′-end of every 454 read (Thomas *et al.*, 2006). The barcoding approach allows the construction of a single 454 library before pyrosequencing that contains sequences from many different samples. Over time, the barcoding method has been optimized especially for culture-independent analyses of microbial community composition by adding the barcode sequence to one of the primers used to amplify 16S rRNA (Parameswaran *et al.*, 2007; Hamady *et al.*, 2008).

Correspondence: EW Triplett, Department of Microbiology and Cell Science, University of Florida, 1052 Museum Road, Gainesville, FL 32611-0700, USA.
E-mail: ewt@ufl.edu
Received 26 November 2009; revised 22 January 2010; accepted 25 January 2010; published online 25 February 2010

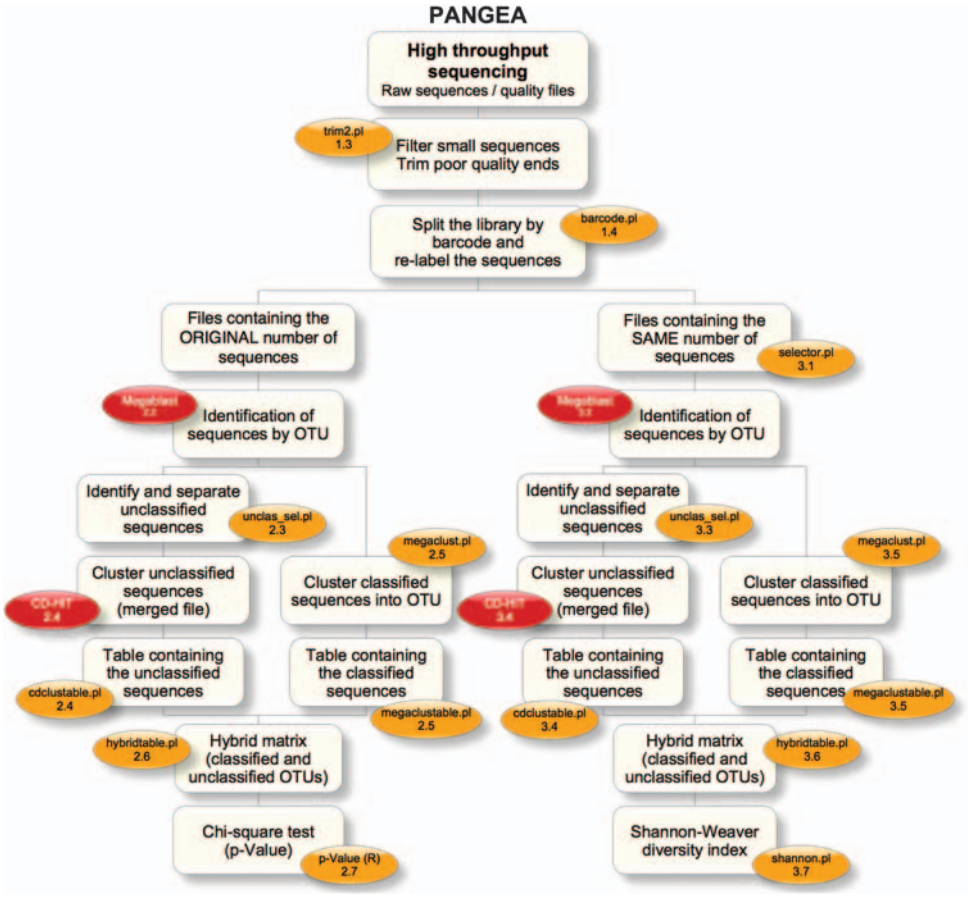

Figure 1 PANGEA workflow. Overview of the pipeline analyses that flow from the raw datasets to the χ^2 and Shannon diversity index results.

PANGEA (Pipeline for Analysis of Next Generation Amplicons) is a workflow designed to manipulate, analyze and identify high-throughput reads (Figure 1). PANGEA analyzes barcoded sequences and performs all necessary steps, from trimming the raw sequence data to the identification of each read in a sample. These tools also include statistical analysis to determine whether samples vary in the abundance of specific taxa. Although PANGEA is described in the context of 16S rRNA sequencing, the tools it provides can be used for the analysis of any barcoded, amplicon sequencing project and can be tailored to any database of interest to the user.

Materials and methods

PANGEA code was written in Perl, R, and Python because interpreters for these languages are available on all three major platforms: Mac OSX, Linux and Windows. The source codes are freely available in http://pangea-16s.sourceforge.net and in the website http://www.microgator.org/. Two data sets are used as examples to illustrate the usefulness of PANGEA. The number of reads available from each data set after trimming and barcode separation and removal is shown in Table 1. The numbering of the subheadings below refers to specific scripts used in PANGEA (Figure 1, Table 2). Example command lines and parameter definitions for Mac OSX are listed for each step below (Table 2).

1 Pre processing of the 16S rRNA barcoded nucleotide sequences

1.1 Data sets and barcoding previous sequencing. Two independent pyrosequencing-generated 16S rRNA fragment libraries are used to show PANGEA (Table 1). The first set of sequences contains barcoded sequences amplified from DNA isolated from 20 fecal samples collected from Bio-Breeding Diabetes-Resistant (BB-DR) and Bio-Breeding Diabetes-Prone (BB-DP) rats described previously (Roesch *et al.*, 2009a). Sampling, DNA extraction, PCR amplification and sequencing for this library were described previously (Roesch *et al.*, 2009a).

The second set of sequences was obtained from the 16S rRNA amplification products of DNA isolated from seven surface soil samples collected at Hawaii Volcanoes National Park in May 2008, from diverse altitudes, different ages since the last

Table 1 Number of pyrosequencing reads obtained from each sample after trimming and barcoding split and the percentage of sequences classified at 95% level of similarity

Rats	*Description*	*Number of sequences*	*Percentage of sequences classified at 95%*	*Shannon*	
				Original	*Normalized (1250 reads)*
BB-DP-01	Biobreeding diabetes prone, 70 days	6126	40.12	4.54	4.15
BB-DP-02	Biobreeding diabetes prone, 70 days	7073	49.02	4.32	4.17
BB-DP-03	Biobreeding diabetes prone, 70 days	6301	54.11	3.92	3.84
BB-DP-04	Biobreeding diabetes prone, 70 days	4554	49.65	4.40	4.13
BB-DP-05	Biobreeding diabetes prone, 70 days	3943	47.63	4.44	4.12
BB-DP-06	Biobreeding diabetes prone, 70 days	3926	36.82	4.67	4.30
BB-DP-07	Biobreeding diabetes prone, 70 days	2784	44.01	4.53	4.30
BB-DP-08	Biobreeding diabetes prone, 70 days	1250	45.48	4.50	4.39
BB-DP-09	Biobreeding diabetes prone, 70 days	1253	50.83	3.90	3.80
BB-DP-10	Biobreeding diabetes prone, 70 days	7965	48.02	4.43	4.34
BB-DR-01	Biobreeding diabetes-resistant, 70 days	2738	48.97	4.18	3.96
BB-DR-02	Biobreeding diabetes-resistant, 70 days	7679	46.26	4.67	4.42
BB-DR-03	Biobreeding diabetes-resistant, 70 days	3549	40.80	4.66	4.32
BB-DR-04	Biobreeding diabetes-resistant, 70 days	3741	45.85	4.42	4.19
BB-DR-05	Biobreeding diabetes-resistant, 70 days	2689	47.22	4.55	4.32
BB-DR-06	Biobreeding diabetes-resistant, 70 days	3379	60.76	3.90	3.77
BB-DR-07	Biobreeding diabetes-resistant, 70 days	2010	56.53	3.87	3.79
BB-DR-08	Biobreeding diabetes-resistant, 70 days	3152	53.65	4.00	3.84
BB-DR-09	Biobreeding diabetes-resistant, 70 days	1955	52.71	4.15	4.02
BB-DR-10	Biobreeding diabetes-resistant, 70 days	2740	51.33	3.83	3.66
Total of sequences		*78807*			

Hawaiian soils	*Altitude (m)*	*Age*	*Vegetation cover*	*Number of sequences*	*Percentage of sequences classified at 95%*	*Shannon*	
						Original	*Normalized (908 reads)*
Mauna Ulu summit	1300	1974	Unvegetated	20864	5.49	5.25	4.34
Mauna Ulu base	30	1974	Unvegetated	22076	8.29	5.22	3.85
Caldera Rim	1300	1790	Sparsely vegetated	2069	32.96	3.07	2.71
Mauna Ulu mid-alt CO	700		Unvegetated	28793	11.41	5.19	3.71
Pu'u Puai canopy	1300	1959	Vegetated tree 'island'	1611	12.66	3.20	2.96
CC Road Forest	1200	ca. 1700	Closed canopy forest	908	5.95	4.14	4.04
Pu'u Puai bare	1300	1959	Unvegetated cinders	53141	18.76	5.43	3.78
Total of sequences				*132231*			

Shannon indices are also provided with and without normalizing the pyrosequencing data sets.

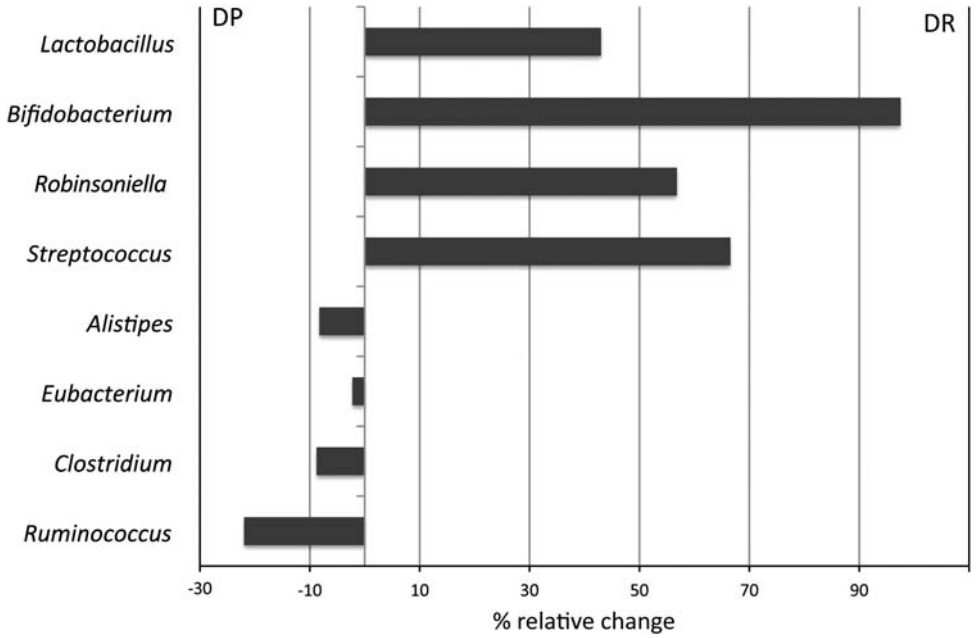

Figure 2 Percent relative change of eight bacterial genera that differ statistically in the χ^2-test (P-value $\leqslant 0.01$) in the BB-DP and BB-DR rat stool samples.

volcanic eruption and varying vegetation cover. Triplicate samples from each site were frozen (−20 °C) before transportation and then stored at −80°C until total DNA had been extracted using MoBio Power Soil extraction kit (MoBio, Carlsbad, CA, USA), amplified using primers described by Roesch *et al*. (2007), with the self-correcting barcode set from Hamady *et al*. (2008).

1.2 Sequencing and quality adjustment of sequences: trim2.pl. Sequences were generated using a GS FLX 454 DNA pyrosequencer (454 Life Sciences, Branford, CT, USA). A total of 89 847 and 275 529 reads were obtained for the rat and soil samples, respectively. Pre-processing of the 454-sequences data set was performed to remove short sequences and trim those sequences that contain bases with low quality scores. To perform this task, Xiaoqui Huang of Iowa State University provided a script called Trim2 (Huang *et al*., 2003), which was translated from C into Perl for use in PANGEA. This revised script also contains the function, 'chomp', which deletes information such as sequence length and ranking as well as data about the pyrosequencing plate.

1.3 Barcoding split and removal. A perl script named *barcode.pl* was written to search for the barcodes in the trimmed file and split them into different files for each barcode. This script splits the input.fas file according to the identified barcode sequence, removes the actual barcode sequence from each read and inserts the barcode number at the beginning of each sequence ID.

If the barcode is missing or present in the sequence at a location other than the beginning of the read, the sequence is removed and placed in a file of sequences lacking a barcode or in a file containing faulty sequences. The user may inspect these files to trouble-shoot errors in sequencing.

2 Identifying organisms responsible for variation

2.1 Working with the original number of sequences. For the identification of organisms present in a population, the files containing the sequences obtained after barcode split were used to run Megablast and the scripts that follow.

2.2 Standalone BLAST search - Megablast. Megablast is part of the package called BLAST (Zhang *et al*., 2000) and is available from NCBI. The sequences were phylogenetically classified using a standalone BLAST against a modified bacterial RDP-II database prepared using TaxCollector (http://www.microgator.org), which attaches complete taxonomic information from domain to species to each sequence in the database and can be obtained from http://www.microgator.org/. The closest bacterial relative was assigned to each sequence corresponding to the best match in the database.

The file generated by Megablast contains the Query ID (sequence name), Subject ID (name of the most closest related bacteria), percentage of identity between query and subject, alignment length, number of mismatches, gap openings, query start, query end, subject start, subject end, e-Value and bit score.

2.3 Unclassified sequences selection. The sequences not classified by Megablast (item 2.2) were captured using a perl script called unclassified

Table 2 Example of command lines and parameter definitions for Mac OSX used by PANGEA

Step in PANGEA/process	*Example command line*	*Parameter definitions*[a]	
1 Preprocessing			
1.2 Quality adjustment	Perl trim2.pl -is input.fas –iq input.qual –o input_trimmed.fas –l 100 –q 20	-is	Input sequence in fasta format
		-iq	Input quality file
		-l	Minimum length of the sequences
		-q	Minimum Phred quality score
1.3 Barcoding split and removal	Perl barcode.pl –s input.fas –b barcode_list.txt	-s	Input sequence file
		-b	Barcode list
2 Identifying organisms			
2.2 Standalone Megablast search	Megablast –d database.fas –i bar01_removed.fas –o bar01_megablast.txt –v 1 –b 1 –m 8	-d	Database
		-i	Input sequences in fasta format
		-v	Maximum number of database sequences used to report alignment
		-b	Maximum number or report alignments for a given database
		-m	Align view (tabular result)
2.3 Unclassified sequences selection	Perl unclassified_selector.pl –m bar01megablast_results.txt –s bar01.fas –t 95 –o bar01unclassified.fas –e -20 –b 200	-m	Megablast output file
		-s	Sequences in fasta file
		-t	Similarity level to be analyzed
		-e	e-value upper threshold (default e-20)
		-b	Bitscore lower threshold (default 200)
2.4 Clustering the unclassified sequences	cd-hit-est -i bar01.fas -o bar01.fas.clstr -c 0.95 -n 8 -g 1	-i	Input file
		-c	Sequence identity threshold
		-n	Word length
		-g	Clustering method
Table with cluster names	Perl clustertable.pl –c bar01.fas.clstr.clstr bar02.fas.clstr.clstr –n 2 –o clustertable_output.txt	-c	CD-HIT output file (.fas.clstr.clstr)
		-n	Number of columns in the table
2.5 Clustering the classified sequences	Perl megaclust.pl –i megablast.txt –o output.txt –s 95 –e -20 –b 200	-i	Megablast output file
		-s	Similarity lower threshold (percent)
		-e	e-value upper threshold (default e-20)
		-b	Bitscore lower threshold (default 200)
Table with sequences names	Perl megaclustable.pl -m bar01megaclust.txt bar02megaclust.txt bar03megaclust.txt –t 0 –o megaclustable_output.txt	-m	Output files from megaclust.pl
		-t	Taxonomic level to be achieved
2.6 Hybrid table with classified and unclassified OTUs	Perl hybridtable.pl –m megaclustable_output.txt –c cdclustable_output.txt –o hybrid_output.txt	-m	Megaclustable output file
		-c	Cdclustable output file
2.7 Differences between samples	Perl chi_square4mac.pl -l 1_4 2_3 5_6	-l	Pairs to be compared

Table 2 (Continued)

Step in PANGEA/process	*Example command line*	*Parameter definitions*[a]	
3 Community analysis			
3.1 Normalizing the number of sequences	Perl selector.pl –i bar01.fas –o bar01_1000.fas –s 1000	-i	Input sequence file
		-s	Number of sequences to be selected
3.7 Shannon diversity index	Perl Shannon.pl –t hybrid_output.txt –n 20 –o Shannon_output.txt	-t	Hybrid table output file
		-n	Number of samples
PANGEA backbone	Perl pangea4mac.pl –s input.fas –q input.fas.qual –d database.fas –b barcode.txt	-s	Input sequences file in fasta format
		-q	Input quality file
		-d	Database to be used by Megablast
		-b	Barcode list

[a]Wherever used, the -o definition denotes the output file.

selector (*unclas_sel.pl*), which recognizes the unclassified sequences directly from the Megablast output file at any given similarity level and generates a new file containing those sequences.

2.4 Clustering the unclassified sequences – CD-HIT and cdclustable.pl. Sequences obtained by the *unclas_sel.pl* were merged using a Unix *cat* command in Mac and Linux. In windows, a custom script performs this function. The sequences are then submitted to CD-HIT to be clustered into Operational Taxonomic Units (OTU) based on the relatedness of the sequences. CD-HIT (*Cluster Database at High Identity with Tolerance*; Li and Godzik, 2006) is a fast and flexible tool that uses a short word filter instead of many pairwise sequence alignments such as the BLAST algorithm (Li and Godzik, 2006). CD-HIT-EST is one of the scripts inserted into the CD-HIT and it is appropriate for non-intron containing sequences, such as prokaryote genomes. Two specific parameters were used in this step, the sequence identity threshold (-c), which was defined here as 0.80 similarity to Domain/Phylum, 0.90 to Class/Order/Family, 0.95 to Genus and 0.99 to Species level (CD-HIT can use several cutoffs ranging from 40 to 99% similarity) and word length (-n) defined here as 8 (8 for thresholds 0.9 to 1.0; word length decreases with the similarity).

2.5 Clustering the classified sequences – megaclust.pl *and* megaclustable.pl. Sequences classified by Megablast were grouped into OTUs based on the relatedness of classification. In this study, queries/subjects were grouped into OTUs at the above similarity levels. A perl script called *megaclust.pl* was written, which uses two different input files: the Megablast output file containing the best matches and a FASTA file containing the sequences from each sample. The output file generated is a tabular file containing the OTU name, number of sequences present in each OTU cluster based on a specific threshold, followed by the sequence ID of the longest query sequence that was retained as the representative sequence of the cluster.

2.6 Obtaining a hybrid table with classified and unclassified OTUs – hybridtable.pl. To access the microbial diversity among the samples, a hybrid table was prepared by combining the unclassified clusters obtained by CD-HIT (item 2.3) and the classified OTUs obtained by Megablast (item 2.4). For this purpose, tables generated by *cdclustable.pl* (item 2.3) and *megaclustable.pl* (item 2.4) were merged using a script called *hybridtable.pl*.

2.7 Taxa differences between samples, the χ^2-test with a P-*value*. A table containing the *hybridtable.pl* results was generated showing the number of OTUs present in the library as well as the number of sequences in each environment. To determine whether specific clusters of bacteria differ between

environments, an exact χ^2-test (based on 50 000 Monte Carlo iterations) was performed to get a *P*-value for the null hypothesis that there was no difference between all possible pairwise combinations of time points. The exact test, based on permutations, is not sensitive to zero counts in the bacterial relatives. The *P*-values were ordered and processed to obtain a false discovery rate of less than 1%.

The Chi-Square tool is used after the hybrid table prepared in the step above and automatically runs a script in R. The user defines the pairs to be compared.

3 Statistical analyses of microbial communities
3.1 Normalizing the number of sequences – selector.pl. To access microbial diversity, the number of reads analyzed was normalized to the same number of reads in each sample. This was done by identifying the sample with the smallest number of reads and selecting the number of sequences from all samples by randomly selecting sequences from the fasta file using a perl script called *selector.pl*.

3.2 to 3.7 Shannon diversity index. To assess the microbial diversity among the samples, a diversity index was calculated based on a hybrid table comprised by the classifiable sequences clusters obtained by Megablast and unclassified clusters obtained by CD-HIT. For this purpose, files containing the same number of sequences (generated by the script *selector.pl*) were submitted to the same methodology described in steps 2.2 to 2.6.

A hybrid table containing the number of sequences from all classified and unclassified clusters was built for each sample. The Shannon diversity index was determined for each sample using the script called *shannon.pl* and the hybrid table output file.

For comparison purposes, the Shannon index was also calculated using the results obtained in the hybrid table generated with the original number of sequences (see item 2.6).

4 PANGEA Backbone
The backbone was designed to integrate PANGEA scripts into a single system for rapid analysis of the sequences. The PANGEA backbone can be downloaded as a zip folder and requires the original fasta file containing the sequences (input.fas), the quality scores file (input.qual.fas), the database to be used by Megablast (database.fas) and a text file containing the barcodes number and their respective sequences (barcode.txt).

Results

Classification of the sequences and the χ^2-test
The most dominant Phylum detected in BB-DP and BB-DR rats were Firmicutes with 76% and 73.15% of the total sequences, and Bacteroidetes with 9.42% and 11.78% of the total sequences, respectively. A χ^2-test was used to determine which OTUs were different between BB-DP and BB-DR samples in all the taxonomic levels. Based on that, eight genera were found to be statistically more abundant in the BB-DP at 95% of similarity (Figure 2 and Supplementary Table S1).

The percent of sequences classified at genus level (95%) for BB-DP and BB-DR was 47.6% and 50.6%, respectively. A total of 164 bacterial OTUs were identified as inhabitants of the 20 rats stool samples. The most abundant genera found in rat samples were *Clostridium*, *Ruminococcus*, *Lactobacillus*, *Eubacterium* and *Bacteroides*. Of these, *Ruminococcus* and *Eubacterium* were more abundant in BB-DP and *Clostridium*, *Lactobacillus* and *Bacteroides* were more abundant in BB-DR samples.

For the Hawaiian soil samples, enormous differences among the seven sites were observed. Actinobacteria and Proteobacteria were found in all seven environments. The number of sequences classified at genus level, as defined by clustering at the 95% similarity level, in seven Hawaiian sites ranged from 5.49% in Mauna Ulu summit to 32.96% in Caldera Rim. A total of 98 bacterial genera were identified in Hawaiian samples.

Most of the families represent less than 1% of the total population in any of the soils. Bacteria belonging to the Family Acidobacteriaceae were present at all sites except by Mauna Ulu mid-altitude (CO site). Bacteria belonging to the Family Sphingomonadaceae were found only in the Caldera Rim site where they represented 2.6% of all reads.

Shannon diversity index
The Shannon diversity index was used to assess the diversity between BB-DP and BB-DR rats and between the seven Hawaiian soil microbial communities. To compare the diversity of these communities, Shannon diversity index was measured in two hybrid tables, one containing the original number of sequences and in the other where the sequences were normalized. The average of Shannon diversity index for BB-DP and BB-DR communities was $H' = 4.15$ and $H' = 4.03$, respectively, when the results were normalized to the same number of reads in each sample. When the files containing the original number of sequences were analyzed, the Shannon index was $H' = 4.36$ and $H' = 4.23$, for BB-DP and BB-DR, respectively. There were no significant differences among the BB-DR and BB-DP communities. However, as expected, the Hawaiian soil communities differed greatly from each other with the Shannon diversity indices ranging from $H' = 2.71$ in the Caldera Rim community to $H' = 4.34$ at the Mauna Ulu summit site, the results of which were obtained from a normalized data set (Table 1).

Discussion

Two 16S rRNA gene high-throughput sequencing datasets were analyzed using a workflow called PANGEA. The objective was to provide a specific set of new tools that take advantage of previously published tools to allow rapid characterization of microbial communities and identification of their members. The rat data set was analyzed in 24 h using a Mac Book Pro with Mac OSX version 10.6.2, 2.4 GHz Intel Core 2 Duo and 4 GB 667 MHz DDR2 SDRAM. This analysis included the initial steps for processing the sequences and the taxonomic classification using Megablast. Megablast is the most time consuming step and is used twice during the process. It is used to perform an accurate classification of the organisms and to calculate the Shannon diversity index of each community. In this work, a unique 16S rRNA database called RDP-TaxCollector was created using a set of scripts called TaxCollector (http://www.microgator.org). Each 16S rRNA sequence in the TaxCollector database is derived from classified isolates and has full taxonomic assignments, from Phylum to Species, for the majority of these sequences.

Classification of the 16S rRNA gene fragments from high-throughput sequencing requires the manipulation of over 300 000 000 sequences in a single file. Tools such as CD-HIT, the RDP Pipeline, and DOTUR (Schloss and Handelsman, 2005) cluster the sequences before reducing the number of sequences to be analyzed. For instance, CD-HIT organizes sequences into clusters and identifies the longest sequence as the representative sequence of the cluster.

However, the longest sequence might not be the closest best nucleotide match for taxonomic classification. PANGEA aligns and identifies the sequences within each library before the clustering step. This ensures that every sequence is classified at the genus and species levels before clustering. It also provides a more accurate identification of each sequence than that provided when clustering is done before classification. The use of the TaxCollector database inside PANGEA allows the user to classify the sequences at seven taxonomic levels, Domain to Species. These classifications can be used to enumerate the number of sequences found at each taxonomic level and determine any differences between treatments using the χ^2-test. Using the TaxCollector database, significant differences were found within each taxonomic level (Figure 2 and Supplementary Table S1) with the data set obtained from BB-DR and BB-DP rat stool samples. In addition, an R script to compute *P*-values allows the user to identify those taxa that differ significantly at a given *P*-value.

The consequences of clustering after classification as is done here, as opposed to clustering before classification as in Roesch *et al*. (2009a, b), can be significant. Clustering before classification using CD-HIT can artificially create clusters because of its dependence on sequence length. The sequences represented by a cluster sequence can then fall into a cluster that does not match its best match in the database. As a result, fewer, and in some cases different genera are observed differing significantly between DR and DP in this work than were observed in our previous work (Roesch *et al.*, 2009a, b). The main themes remain the same in both analyses. That is, probiotic genera such as *Lactobacillus* and *Bifidobacterium* are significantly higher in DR than DP, but subtle differences do occur between the two studies (Figure 2 and Supplementary Table S1). In addition, the RDP database includes new genera since the publication of Roesch *et al*. (2009a, b). In some cases, these new genera were once a subset of a genus described in the original paper.

Some barcoded sequence data sets have a high disparity between the number of fragments in each sample sequenced. This might be due to an error in quantification while creating a master DNA pool by combining the purified products in equimolar ratios before pyrosequencing library construction, resulting in different number of sequences for each sample. Another possibility in pyrosequencing data sets is that a significant proportion of sequences simply lacks the barcode and is discarded into separate files by *barcodes.pl*. To minimize the effect of this disparity in the number of sequences in each barcode file, the PANGEA workflow normalizes the data before community analysis. That is, the number of sequences in each sample within a barcoded set is identical and based on the number available in the least represented sample. The RDP Pipeline (Cole *et al.*, 2009) and mothur (Schloss *et al.*, 2009) manipulate files containing the original number of sequences without taking the disparity between the number of sequences between samples into consideration. This is particularly important when dealing with diversity indices. Diversity index values increase with sample size making normalization of the number of sequences in all samples crucial (Patil and Taillie, 1982).

The Shannon index was calculated for each data set in two situations to show the importance of normalization: first, using the hybrid table originated from the files containing the original number of sequences; second, using the hybrid table obtained from the files containing the same number of sequences in all the files (Table 1). To cluster the Megablast non-classified sequences, CD-HIT uses a merged file containing the sequences identified by name for each sample. As the clustering step by CD-HIT leads to a different number of clusters depending on the number of sequences in the data set to be analyzed, the non-normalized data set contains more sequences and consequently more clusters. This explains the increase in the Shannon indices when calculated from the original number of sequences (Table 1).

PANGEA is designed to manipulate high throughput sequences and can be used in combination with

other tools available to analyze and characterize microbial population, such as the RDP Pyrosequencing Pipeline (Cole *et al.*, 2009) and mothur (Schloss *et al.*, 2009). With PANGEA, each tool is available to the user separately so that the analysis can be adjusted to specific needs. The steps described in PANGEA should be followed in the prescribed order, but they can be tailored to fit the needs of any assortment of analyses. Although the two examples here are from 454-pyrosequencing libraries, these tools can be used to analyze different libraries independent of the high-throughput technology used. In addition, they can be applied to barcoded, amplified samples for any gene. The user need only define the database used to identify the organisms or genes present in the samples.

PANGEA offers the following advantages over all other tools currently available for 16S rRNA analyses. First, PANGEA normalizes the data sets to give equal sample sizes between treatments. This is essential to obtain valid diversity indices whether done by the Shannon method or other means. Second, PANGEA classifies each read in an automated fashion as opposed to clustering the reads first. Third, PANGEA clusters the unclassified sequences to get a complete sense of the diversity of a sample. Fourth, PANGEA has tremendous flexibility. It can be used as a complete pipeline taking raw reads through to the production of tables used in statistical analysis and the calculation of a diversity index. The programs can also be used individually to solve specific tasks. For example, if a user only needs to classify a set of reads, the Megablast tool can be used with our new TaxCollector databases. If a user simply wants to split the sequence set by the barcodes and remove the barcodes, the *barcode.pl* program is able to do that.

There are also several specific advantages of PANGEA over the RDP Pipeline. First, PANGEA is not web-based. Although at first this may appear to be a difficult hurdle, the format chosen for PANGEA leads to advantages, shared with other stand alone tools. The user does not need to upload data to a remote site, which is becoming increasingly difficult, as data sets get larger, and may not be desirable for privacy and confidentiality reasons. The user can devote as much or as little computer power to the job as needed. The user does not have to wait in a queue for his/her work to be finished. The user can modify the tools or use any subset of them for a particular task.

Second, the RDP Pipeline is not automated in the sense that a final output is provided to the user after a number of steps. Instead, an output is provided to the user at each step. That output must be resubmitted to the RDP Pipeline for the next step. In contrast, PANGEA can be used as an automated pipeline with a single input act from the user resulting in final files after a variety of analyses. Third, the source code for the RDP Pipeline software is not available and hence, cannot be modified by the user. Fourth, in the RDP Pipeline, the analysis must begin with raw sequence data that includes the quality scores. This prevents the user from analyzing data sets that have been submitted to GenBank that have been trimmed, lack barcodes and quality scores. Thus, the RDP Pipeline is not useful for the rapid re-analysis of data. In PANGEA, a fasta file obtained from GenBank can be entered into the process. Fifth, the RDP Pipeline requires that the primer sequences be known. This is not required in PANGEA. And finally, with the RDP Pipeline, the user is completely dependent on the databases provided by RDP. In PANGEA, the user defines the database.

Acknowledgements

This research was supported by grants from the National Science Foundation (MCB-0454030), United States Department of Agriculture (2005-35319-16300 and 00067345), and the Florida Agricultural Experiment Station.

References

Andersson AF, Lindberg M, Jakobsson H, Backhed F, Nyren P, Engstrand L. (2008). Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE* **3**: e2836.

Armougom F, Raoult D. (2008). Use of pyrosequencing and DNA barcodes to monitor variations in Firmicutes and Bacteroidetes communities in the gut microbiota of obese humans. *BMC Genomics* **9**: 576.

Brown MV, Philip GK, Bunge JA, Smith MC, Bissett A, Lauro FM *et al.* (2009). Microbial community structure in the North Pacific Ocean. *ISME J* **3**: 1374–1386.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**: d141–d145.

Dethlefsen L, Huse S, Sogin ML, Relman DA. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* **6**: e280.

Dowd SE, Sun Y, Secor PR, Rhoads DD, Wolcott BM, James GA *et al.* (2008). Survey of bacterial diversity in chronic wounds using Pyrosequencing, DGGE, and full ribosome shotgun sequencing. *BMC Microbiol* **8**: e43.

Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM *et al.* (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**: 57.

Fierer N, Hamady M, Lauber CL, Knight R. (2008). The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci USA* **105**: 17994–17999.

Hamady M, Walker J, Harris JK, Gold NJ, Knight R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* **5**: 235–237.

Huang X, Wang J, Aluru S, Yang SP, Hillier L. (2003). PCAP: a whole-genome assembly program. *Genome Res* **13**: 2164–2170.

Huber JA, Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA *et al.* (2007). Microbial population structures in the deep marine biosphere. *Science* **318**: 97–100.

Jones RT, Robeson MS, Lauber CL, Hamady M, Knight R, Fierer N. (2009). A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. *ISME J* **3**: 442–453.

Keijser BJF, Zaura E, Huse SM, vanderVossen JMBM, Schuren FHJ, Montijn RC *et al.* (2008). Pyrosequencing analysis of the oral microflora of healthy adults. *J Dent Res* **87**: 1016–1020.

Li W, Godzik A. (2006). CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* **35**: 1–10.

Liu Z, DeSantis TZ, Andersen GL, Knight R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* **36**: e120.

Luna RA, Fasciano LR, Jones SC, Boyanton Jr BL, Ton TT, Versalovic J. (2007). DNA pyrosequenging-based bacterial pathogen identification in a pediatric hospital setting. *J Clin Microbiol* **45**: 2985–2992.

McKenna P, Hoffmann C, Minkah N, Aye PP, Lackner A, Liu Z *et al.* (2008). The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathog* **4**: e20.

Miller SR, Strong AL, Jones CL, Ungerer MC. (2009). Bar-coded pyrosequencing reveals shared bacterial community properties along two alkaline hot spring temperature gradients in Yellowstone National Park. *Appl Environ Microbiol* **75**: 4565–4572.

Parameswaran P, Jalili R, Tao R, Shokralla S, Gharizadeh B, Ronaghi M *et al.* (2007). A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* **35**: e130.

Patil GP, Taillie C. (1982). Diversity as a concept and its measurement. *J Am Stat Assoc* **77**: 548–561.

Price LB, Liu CM, Melendez JH, Frankel YM, Engelthaler D, Aziz M *et al.* (2009). Community analysis of chronic wound bacteria using 16S rRNA gene-based pyrosequencing: impact of diabetes and antibiotics on chronic wound microbiota. *PLoS ONE* **4**: e6462.

Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD *et al.* (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283–290.

Roesch LFW, Lorca GL, Casella G, Giongo A, Naranjo A, Pionzio AM *et al.* (2009a). Culture-independent identification of gut bacteria correlated with the onset of diabetes in a rat model. *ISME J* **3**: 536–548.

Roesch LFW, Casella G, Simell O, Krischer J, Wasserfall CH, Schatz D *et al.* (2009b). Influence of fecal sample storage on bacterial community diversity. *Open Microbiol J* **3**: 40–46.

Schloss PD, Handelsman J. (2005). Introducing DOTUR, a computer program for dening operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71**: 1501–1506.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hillister EB *et al.* (2009). Introducing mothur: open source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored ''rare biosphere''. *Proc Natl Acad Sci USA* **103**: 12115–12120.

Thomas RK, Nickerson E, Simons JF, Janne PA, Tengs T, Yuza Y *et al.* (2006). Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat Med* **12**: 852–855.

Zhang Z, Schwartz S, Wagner L, Miller W.. (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**: 203–214.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)