

Article

TaxCollector: Modifying Current 16S rRNA Databases for the Rapid Classification at Six Taxonomic Levels

Adriana Giongo, Austin G. Davis-Richardson, David B. Crabb and Eric W. Triplett *

Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences, University of Florida, 1052 Museum Road, Gainesville, FL 32611-0700, USA; E-Mails: giongo@ufl.edu (A.G.); adavisr@ufl.edu (A.G.D.-R.); dcrabb@ufl.edu (D.B.C.)

* Author to whom correspondence should be addressed; E-Mail: ewt@ufl.edu;
Tel.: +1-352-392-5430; Fax: +1-352-846-0950.

Received: 16 June 2010; in revised form: 14 July 2010 / Accepted: 19 July 2010 /

Published: 21 July 2010

Abstract: The high level of conservation of 16S ribosomal RNA gene (16S rRNA) in all Prokaryotes makes this gene an ideal tool for the rapid identification and classification of these microorganisms. Databases such as the Ribosomal Database Project II (RDP-II) and the Greengenes Project offer access to sets of ribosomal RNA sequence databases useful in identification of microbes in a culture-independent analysis of microbial communities. However, these databases do not contain all of the taxonomic levels attached to the published names of the bacterial and archaeal sequences. TaxCollector is a set of scripts developed in Python language that attaches taxonomic information to all 16S rRNA sequences in the RDP-II and Greengenes databases. These modified databases are referred to as TaxCollector databases, which when used in conjunction with BLAST allow for rapid classification of sequences from any environmental or clinical source at six different taxonomic levels, from domain to species. The TaxCollector database prepared from the RDP-II database is an important component of a new 16S rRNA pipeline called PANGEA. The usefulness of TaxCollector databases is demonstrated with two very different datasets obtained using samples from a clinical setting and an agricultural soil. The six TaxCollector scripts are freely available on <http://taxcollector.sourceforge.net> and on <http://www.microgator.org>.

Keywords: 16S rRNA gene; microbial diversity; taxonomy

1. Introduction

The sequencing and PCR amplification of 16S ribosomal RNA gene (16S rRNA) sequences have become the basis for the rapid identification and classification of Prokaryotes. Culture-independent methods take advantage of the robustness of the highly conserved 16S rRNA gene for phylogenetic assignments within microbial communities from any environment. The presence of this highly conserved gene in all Prokaryotes enables it to be used as a molecular chronometer in evolutionary studies and its length is ideal for all sequencing platforms [1]. Next generation sequencing produces many thousands of small 16S rRNA reads per sample [2–5] that require a further classification step using specific 16S rRNA databases. Projects such as the Ribosomal Database Project II (RDP-II) at Michigan State University [6,7] and the Greengenes Project, maintained by the Lawrence Berkeley National Laboratory [8] offer access to sets of rRNA sequence databases useful in microbial population studies. Those databases can be downloaded directly from the database project websites and can be tailored to the interests of the user such as the selection of the sequences in the database from cultured and/or uncultured organisms. However, these databases do not include all of the taxonomic information that is available for these sequences in their headers.

To address this problem we created TaxCollector (Taxonomy Collector), a set of Python scripts that attach taxonomic information from domain to species levels acquired from NCBI to RDP-II and Greengenes 16S rRNA databases. Thus, TaxCollector creates modified RDP or Greengenes databases with all of the taxonomic information to allow the user to rapidly identify taxonomic differences between communities at all levels. In addition to providing the code, up-to-date TaxCollector-modified 16S rRNA databases are available for download at <http://www.microgator.org/>. The TaxCollector code can also be quickly adapted to any other gene with an available database.

2. Experimental Section

2.1. Sampling, DNA Extraction and Sequencing

Two independent pyrosequencing-generated, 16S rRNA fragment libraries were used to demonstrate TaxCollector. The first set of sequences contained barcoded sequences amplified from DNA isolated from a sugar cane field in the Everglades Agricultural Area in Florida, first published by Roesch *et al.* [2]. Details on the sampling, DNA extraction, PCR amplification and sequencing for this library were described previously [2] with the exceptions that only one round of PCR was done with primers containing the pyrosequencing 454 A and B adaptors and 454 FLX sequencing was done which resulted in an average read length of 223 bases.

The second set of sequences was obtained from the 16S rRNA amplification products of DNA isolated from fresh stool samples obtained from seven children at the Shands Hospital, University of Florida, Gainesville, FL. Samples were collected in sterile flask containers and kept at $-20\text{ }^{\circ}\text{C}$ for further DNA extraction. DNA isolation and purification, PCR amplification and sequencing were performed as described by Roesch *et al.* [9]. After sequencing, pre-processing of the 454-sequences datasets was performed using to remove short sequences and trim those sequences that contain bases with low quality scores using PANGEA [10].

2.2. Files to Attach Taxonomic Information to 16S rRNA Sequences and Algorithm Description

The incorporation of the taxonomic names in known 16S rRNA databases requires downloading of the database of interest from public sources such as RDP-II or Greengenes plus additional files containing the taxon information from National Center for Biotechnology Information (NCBI) [11]. The 16S rRNA database containing the sequences and the published names of Prokaryotes was downloaded from Ribosomal Database Project II [<http://rdp.cme.msu.edu>] release 10.14 from August 2009 and Greengenes [<http://greengenes.lbl.gov>] from January 2009.

Files from NCBI containing the taxonomy information, called names.dmp and nodes.dmp files, were obtained at <ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz>. Those files are used to create taxonomic descriptions that are associated with the information available in the RDP II and Greengenes FASTA file databases. Each node is a "child" node and has a "parent" node. The script traverses the nodes list until it finds the highest parent node. It then uses the names found in the database to create the taxonomic description. However, the script is restricted to seven taxonomic levels (kingdom, phylum, class, order, family, genus and species). In the case of finding some gap in the taxonomic description (*i.e.* NCBI's database lack the Class level), the script re-uses the previous name and surrounds it by quotes. TaxCollector is written to fill a single gap only. In cases where there is a larger gap, the script simply moves on to the next available name. This creates homogenous taxonomic descriptions, which can easily be imported into spreadsheet software.

2.3. Attaching Taxonomic Information to 16S rRNA Databases—the TaxCollector Scripts

A combination of the database containing the sequences (obtained from RDP-II or Greengenes) with the taxon information found in NCBI was obtained using a Python script called *TaxCollectorRDP.py*.

Besides the RDP-II database, two taxonomic information files obtained from NCBI, called names.dmp and nodes.dmp are required (Table 1).

Table 1. Scripts and command lines used to modify databases using TaxCollector.

	Script	Function	Example of command line
RDP	<i>TaxCollectorRDP</i>	Attaches self-explanatory taxonomic information headers	python <i>TaxCollectorRDP.py</i> names.dmp nodes.dmp RDPoriginal_database.fas RDP_TaxCollector.fas
Greengenes	<i>acctotax</i>	Creates intermediate file containing the ProkMSA IDs	python <i>acctotax.py</i> greengenes_export_0.txt
	<i>TaxCollectorGreengenes</i>	Attaches self-explanatory taxonomic information headers	python <i>TaxCollectorGreengenes.py</i> names.dmp nodes.dmp greengenes_export_0.txtout greengenes.fas greengenesTC.fas
	<i>filter</i>	Filters undesired records from the database based on a single word in the headers	python <i>filter.py</i> greengenesTC.fas greengenesTC_isolatesonly.fas Eukarya
	<i>remdup</i>	Removes duplicated sequences between two databases	python <i>remdup.py</i> greengenesTC.fas rdpTC.fas rdp+greengenesTC.fas

The original OTU names in the Greengenes database do not contain the identification necessary to match the taxon information released from NCBI files. A script called *acctotax.py* was used to generate an intermediate file to be used in *TaxCollectorGreengenes.py*. That file contains a list of ProkMSA IDs from Greengenes database and their corresponding NCBI Tax IDs. To generate this file, the Greengenes format file called *greengenes_export_0.txt* was downloaded from http://greengenes.lbl.gov/Download/Sequence_Data/Greengenes_format/.

In addition to the names and nodes obtained files from NCBI and the full taxonomic information obtained from the Greengenes database in a fasta format, available at http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/, the file generated by *acctotax.py* (called *greengenes_export_0.txtout*) was used by the *TaxCollectorGreengenes.py* to generate *greengenesTC.fas* database.

Numbers were inserted between brackets representing seven levels of taxonomy: [0] for Domain, [1] for Phylum, [2] for Class, [3] for Order, [4] for Family, [5] for Genus (or correspondent OTU) and [6] for Species (or corresponding OTU).

The 16S rRNA sequences from uncultured strains and isolates can be downloaded separately from the RDP-II website. However, this cannot be done at the Greengenes website. Thus, the Greengenes dataset requires a further step to separate the uncultured 16S rRNA sequences from the isolates. This is done using a Python script called *filter.py*. A fifth script was written to filter undesired records from the database based on a single word. This command is case insensitive and only filters based on what is in the header of the sequence. For example, sequences belonging to the Eukaryota domain were deleted from the Greengenes database modified by TaxCollector and saved in a new FASTA file. The RDP database does not require this step since only bacterial and archaeal sequences are available for download. This script allows the user to easily eliminate any sequences from unwanted organisms present in the database.

2.4. Concatenation of the RDP and Greengenes Databases

After processed using TaxCollector, RDP and Greengenes databases were concatenated using a Python script called *remdup.py*. Upon encountering records with duplicate (2 or more) headers, the script keeps the one that contains the largest sequence (Table 1).

2.5. Utility of the TaxCollector modified databases using two datasets

The sequences were taxonomically classified using standalone BLAST search using Megablast. Megablast is a program inserted into the package called BLAST [11] available in the NCBI website [<http://www.ncbi.nlm.nih.gov/blast/download.shtml>]. Local blast analysis was performed against the five databases generated by TaxCollector (Table 2). The closest bacterial relatives were assigned according to their best matches to sequences in the database.

Table 2. Number of 16S rRNA sequences in six databases obtained before and after TaxCollector.

<i>Database</i>	<i>Number of sequences</i>		
	<i>Downloaded</i>	<i>After TaxCollector</i>	<i>% of recovery sequences</i>
RDP (isolates only)	167,313	164,476	98.30
RDP	924,043	919,524	99.51
Greengenes (isolates only)	302,066	83,263*	27.56
Greengenes	302,066	302,066	100.00
RDP + Greengenes (isolates only)	247,739	165,107	66.65
RDP (Bacteria + Archaea, isolates only)	169,386	166,543	98.32

* after TaxCollector + unclassifiedRemover

To get an overall comparison and to determine differences in the number of classifiable sequences from each database, queries were grouped into Operational Taxonomic Units (OTU) based on the relatedness of sequences (Table 3). In this study, queries/subjects were grouped into OTU exhibiting similarity values depending on the desired taxonomic level, *i.e.*, 80% at Domain/Phylum, 90% to Class/Order/Family, 95% to Genus (or corresponding OTU) and 99% of similarity to Species (or corresponding OTU) levels [12]. The output file generated by Megablast was processed using PANGEA [10].

Table 3. Number of OTU identified in each taxonomic level, using six different databases modified by TaxCollector compared to the results obtained from the online RDP Classifier.

<i>Database</i>	<i>Number OTU identified in each taxonomic level</i>					
	<i>Domain</i>	<i>Phylum</i>	<i>Class</i>	<i>Order</i>	<i>Family</i>	<i>Genus</i>
	<i>(80%)</i>	<i>(80%)</i>	<i>(90%)</i>	<i>(90%)</i>	<i>(90%)</i>	<i>(95%)</i>
RDP						
human gut	1	4	5	11	17	25
Florida soil	1	18	20	33	50	60
RDP (isolates only)						
human gut	1	4	5	12	21	31
Florida soil	1	21	33	64	155	180
Greengenes						
human gut	1	4	6	9	14	26
Florida soil	1	19	21	36	61	69
Greengenes (isolates only)						
human gut	1	4	5	11	20	30
Florida soil	1	22	36	66	157	191

Table 3. Cont.

Database	<i>Number OTU identified in each taxonomic level</i>					
	<i>Domain</i> (80%)	<i>Phylum</i> (80%)	<i>Class</i> (90%)	<i>Order</i> (90%)	<i>Family</i> (90%)	<i>Genus</i> (95%)
RDP + Greengenes						
(isolates only)						
human gut	1	4	5	11	20	34
Florida soil	1	22	37	66	147	180
RDP Bac/Arch						
(isolates only)						
Florida soil	2	22	37	70	158	184
RDP Classifier						
human gut	1	4	5	11	20	19
Florida soil	2	13	15	15	16	17

2.6. RDP Classifier

The datasets were submitted to the RDP Classifier program of the Ribosomal Database Project (RDP-II) release 10 [<http://rdp.cme.msu.edu>] to obtain the closest matches to known organisms using 16S rRNA gene fragments. Taxonomic hierarchy was performed on the Phylum, Class, Order, Family and Genus levels. Sequences were clustered at 80% of similarity for Phylum, 90% of similarity for Class/Order/Family and 95% of similarity for Genus [11]. Tables were generated containing the percentage of each taxonomic level in each sample using text editors.

2.7. Megan

Reads were assigned using Megan, a metagenome package, which classifies DNA fragments based on a lowest common ancestor algorithm [13]. Prior to the Megan analysis, the sequences were classified using the standalone BLAST search [11]. Local blast analysis was performed against the RDP database containing sequences belonging to the isolates only. The BLAST output was processed by Megan to assess the taxonomy for each sequence. Results were exported and then placed into a text editor.

3. Results

The databases modified by TaxCollector were tested using two 454 pyrosequencing libraries derived from two very different sources: human stool and soil sample. The identification of these sequences from the TaxCollector-modified databases was compared with the classification provided by RDP Classifier and Megan. A total of 8,006 16S rRNA sequences were analyzed for each sample.

The databases used for classification consisted of variants of the RDP-II and Greengenes databases modified to contain the taxonomic information in the header of each sequence's published name (Figure 1). In addition, the total number of sequences available from RDP-II and Greengenes for this work and the number of sequences obtained after TaxCollector processing are listed (Table 2). After

processing using TaxCollector, a small proportion of sequences was not classified and thus is not inserted in the new database.

Figure 1. Example of RDP-II database modified by TaxCollector. Top: a single 16S rRNA sequence downloaded directly from RDP-II. Bottom: the same sequence modified using TaxCollector.

```
>S000608701 Bacillus subtilis; B-1082; AM110930
ggccagcgcgctatctgcagtcgagcggacagatgggagccttgctccctgatgtagcggcggacgggtgagtaacacgt
gggtaacctgcctgtaagactgggataactccgggaacccgggctaataccggatgggtggttgaaccgcatggttcaa
acataaaagggtggcttcggctaccacttacagatggaccgcggcgccattagctagttggtgaggtaacggctcaccaag
gcaacgatgcgtagccgacctgagaggggtgatcggccacactgggactgagacacggccagactcctacgggaggcagc

>[0]Bacteria;[1]Firmicutes;[2]Bacilli;[3]Bacillales;[4]Bacillaceae;[5]Bacillus;[6]Bacillus_subtilis
ggccagcgcgctatctgcagtcgagcggacagatgggagccttgctccctgatgtagcggcggacgggtgagtaacacgt
gggtaacctgcctgtaagactgggataactccgggaacccgggctaataccggatgggtggttgaaccgcatggttcaa
acataaaagggtggcttcggctaccacttacagatggaccgcggcgccattagctagttggtgaggtaacggctcaccaag
gcaacgatgcgtagccgacctgagaggggtgatcggccacactgggactgagacacggccagactcctacgggaggcagc
```

The sequences from the stool sample contain only bacterial sequences since the primers used for the amplification of 16S rRNA are bacterial-specific. In contrast, the soil sample sequences were generated using a primer set that amplifies 16S rRNA genes from both archaeal and bacterial groups.

As the Greengenes and RDP-II databases are both widely used, TaxCollector was used to make a set of modified databases that are specific to either Greengenes or RDP-II as well as additional databases that contain sequences from both sources. RDP-II allows the user to download sequences derived from only cultured isolates or from uncultured strains as well. The Greengenes databases include all Archaea and Bacteria while sequences from both domains are downloaded separately from the RDP-II databases.

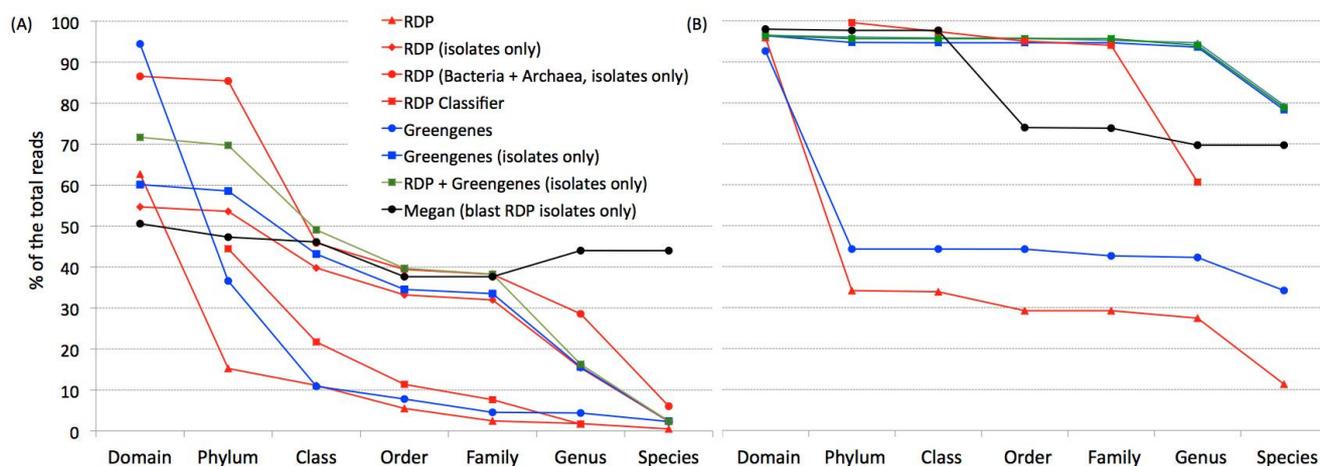
Automated monthly updates of ten TaxCollector modified databases are available for download at <http://www.microgator.org/>. These databases include all of the combinations described above.

When classifying the sequences from the two datasets used here, a far greater proportion of sequences are classified when the database includes only isolates (Figure 2) since a large proportion of sequences are most closely related to a sequence from an uncultured organism. However, these uncultured sequences mask the identity of closely related organisms at traditional taxonomic levels.

These data also show that there is no value added to the taxonomic identification of the sequences when using databases from Greengenes compared to RDP-II. This is expected as the RDP-II database has more sequences than the Greengenes database. Tremendous added value was obtained for the soil sample when the archaeal sequences were included with the bacterial sequences in the RDP-II isolates database.

A far greater proportion of the human samples were classified at all taxonomic levels compared to the soil sample. As the RDP Classifier [14] also uses sequences from uncultured organisms, it was unable to classify nearly as many sequences as the RDP databases designed to include sequences only from cultured organisms. Hence, RDP Classifier identified far fewer sequences at all taxonomic levels with the soil sample and at the genus level for the human sample. The same was observed in the classification using Megan. Both RDP Classifier and Megan classified more sequences from clinical samples than from soil sample.

Figure 2. Microbial classifications obtained from modified RDP and Greengenes databases. Percentage of classifiable sequences in soil (A) and human (B) samples using six databases-header modified by TaxCollector, RDP Classifier and Megan.



4. Discussion

The 16S rRNA gene continues to play a very valuable role in the identification of Prokaryotes even with the increasing use of metagenomic sequencing for environmental and clinical samples [15]. The advent of culture-independent techniques allows obtaining a large number of DNA sequences directly from environment samples. Tools such as BLAST, a similarity-based binning method, allow the classification of thousands of 16S rRNA reads using databases such as RDP-II and Greengenes databases [11]. Depending on the database chosen by the user, the closest match for each sequence may be a cultured or uncultured organism but often these matches do not include useful taxonomic information. TaxCollector was designed to attach self-explanatory taxonomic information to RDP-II and Greengenes databases thus providing the ability to classify reads from domain to species level, thus providing self-explanatory taxonomic headers in the 16S rRNA gene databases. This is a significant advantage over the currently available databases where taxonomic information above the genus level is lacking.

Current tools for the classification of 16S rRNA do not provide all phylogenetic levels from domain to species directly in the database sequences' headers. The tools provided within TaxCollector allows the creation of an up-to-date 16S rRNA database that ties NCBI taxonomic information with rRNA sequences from the Ribosomal Database Project (<http://rdp.cme.msu.edu>) and/or Greengenes (<http://greengenes.lbl.gov>). The code provided with TaxCollector is freely available and can be used for the creation of databases for other genes or sets of genes. The code can also be used to construct user-defined 16S rRNA databases.

The taxonomic information for the construction of TaxCollector databases is derived from NCBI and includes files called names and nodes. The taxonomic information in these files is highly curated by the NCBI Taxonomy team and is based on information from The Prokaryotes ([http://www.springerlink.com/reference-works/?sortorder=asc&mode=boolean&k=ti:\(prokaryotes\)](http://www.springerlink.com/reference-works/?sortorder=asc&mode=boolean&k=ti:(prokaryotes))), Bergey's manual (<http://www.cme.msu.edu/bergeys/>), the up-to-date DSMZ Bacterial Nomenclature list (<http://www.dsmz.de/bactnom/bactname.htm>), the Greengenes 16S rRNA database and workbench

(<http://greengenes.lbl.gov>), the International journal of Systematic and Evolutionary Microbiology (<http://ijs.sgmjournals.org/>), the Official List of Bacterial Names with Standing in Nomenclature (<http://www.bacterio.cict.fr/>), the Ribosomal Database Project (<http://rdp.cme.msu.edu>), and the Taxonomic Outline of Bacteria and Archaea (<http://www.taxonomicoutline.org/>). In addition, the NCBI taxonomic files are commonly used as the backbone of bacterial and archaeal taxonomic information for several international resources such as Megan [13], the PhyloGenie [16], and the Biopathway Workbench [17].

Here TaxCollector was used to classify 16S rRNA pyrosequences from two very different environments: human stool and soil samples. As the soil dataset has many more OTU than the stool dataset, a much higher percentage of the sequences were identified to genus and species in the stool samples (Figure 2). The data presented here illustrate that soil bacteria are far more diverse than gut bacteria (Table 2). Five times more phyla are observed in the soil environment than in the human gut. Hence, the soil sample is phylum rich and species rich while the human sample is phylum poor and species rich.

Other tools can provide taxonomic information for 16S rRNA sequences but they are specifically designed for metagenome studies such as Megan [13] and Sort-ITEMS [18]. Using the TaxCollector modified database in PANGEA [10], the classification of each sequence occurs during the MEGABLAST analysis and prior to clustering. In Megan, the BLAST analysis and classification are separate events requiring more processing time and more computer power.

5. Conclusions

The tools designed within TaxCollector allow for the creation of an up-to-date 16S rRNA database that ties NCBI taxonomic information with rRNA sequences from 16S rRNA databases. The six TaxCollector scripts are freely available at <http://www.microgator.org>. This web-tool can be used to create the headers in Greengenes or RDP databases depending on the user needs. These scripts can also be used for the creation of databases for other genes or sets of genes.

Acknowledgements

This work was supported by the National Science Foundation (grant number MCB-0454030); and the United States Department of Agriculture [grant numbers 2005-35319-16300, 00067345).

References and Notes

1. Wu, D.; Hartman, A.; Ward, N.; Eisen, J.A. An automated phylogenetic tree-based Small Subunit rRNA Taxonomy and Alignment Pipeline (STAP). *PLoS ONE* **2008**, *3*, e2566.
2. Roesch, L.F.W.; Fulthorpe, R.R.; Riva, A.; Casella, G.; Hadwin, A.K.M.; Kent, A.D.; Daroub, S.H.; Camargo, F.A.O.; Farmerie, W.G.; Triplett, E.W. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.* **2007**, *1*, 283–290.

3. Roesch, L.F.W.; Lorca, G.L.; Casella, G.; Giongo, A.; Naranjo, A.; Pionzio, A.M.; Li, N.; Mai, V.; Wasserfall, C.H.; Schatz, D.; Atkinson, M.A.; Neu, J.; Triplett, E.W. Culture-independent identification of gut bacteria correlated with the onset of diabetes in a rat model. *ISME J.* **2009**, *3*, 536–548.
4. Hamady, M.; Walker, J.; Harris, J.K.; Gold, N.J.; Knight, R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* **2008**, *5*, 235–237.
5. Liu, Z.; DeSantis, T.Z.; Andersen, G.L.; Knight, R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucl. Acids Res.* **2008**, *36*, e120.
6. Cole, J.R.; Chai, B.; Farris, R.J.; Wang, Q.; Kulam, S.A.; McGarrell, D.M.; Garrity, G.M.; Tiedje, J.M. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucl. Acid Res.* **2005**, *33*, D294–D296.
7. Cole, J.R.; Chai, B.; Farris, R.J.; Wang, Q.; Kulam, S.A.; McGarrell, D.M.; Bandela, A.M.; Cardenas, E.; Garrity, G.M.; Tiedje, J.M. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucl. Acid Res.* **2007**, *35*, D169–D172.
8. DeSantis, T.Z.; Hugenholtz, P.; Larsen, N.; Rojas, M.; Brodie, E.L.; Keller, K.; Huber, T.; Dalevi, D.; Hu, P.; Andersen, G.L. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **2006**, *72*, 5069–5072.
9. Roesch, L.F.; Casella, G.; Simell, O.; Krischer, J.; Wasserfall, C.H.; Schatz, D.; Atkinson, M.A.; Neu, J.; Triplett, E.W. Influence of fecal sample storage on bacterial community diversity. *Open Microbiol. J.* **2009**, *3*, 40–46.
10. Giongo, A.; Crabb, D.B.; Davis-Richardson, A.G.; Chauliac, D.; Mobberley, J.M.; Gano, K.A.; Mukherjee, N.; Casella, G.; Roesch, L.F.W.; Walts, B.; Riva, A.; King, G.; Triplett, E.W. PANGEA: Pipeline for analysis of next generation amplicons. *ISME J.* **2010**, *4*, 852–861.
11. Zhang, Z.; Schwartz, S.; Wagner, L.; Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **2000**, *7*, 203–214.
12. Hong, S.H.; Bunge, J.; Jeon, S.O.; Epstein, S.S. Predicting microbial species richness. *P. Natl. Acad. Sci. USA* **2006**, *103*, 117–122.
13. Huson, D.H.; Auch, A.F.; Qi, J.; Schuster, S.C. MEGAN analysis of metagenomic data. *Genome Res.* **2007**, *17*, 377–386.
14. Wang, Q.; Garrity, G.M.; Tiedje, J.M.; Cole, J.R. Naïve bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **2007**, *73*, 5261–5267.
15. Tringe, S.G.; Hugenholtz, P. A renaissance for the pioneering 16S rRNA gene. *Curr. Opin. Microbiol.* **2008**, *11*, 442–446.
16. Frickey, T.; Lupas, A.N. PhyloGenie: automated phylome generation and analysis. *Nucl. Acid Res.* **2004**, *32*, 5231–5238.
17. Byrnes, R.W.; Cotter, D.; Maer, A.; Li, J.; Nadeau, D.; Subramaniam, S. An editor for pathway drawing and data visualization in the Biopathway Workbench. *BMC Syst. Biol.* **2009**, *3*, 99.

18. Haque, M.M.; Ghosh, T.S.; Komanduri, D.; Mande, S.S. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* **2009**, *25*, 1722–1730.

© 2010 by the authors; licensee MDPI, Basel, Switzerland. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).