Ashdin Publishing
Metagenomics
Vol. 1 (2012), Article ID 235551, 5 pages
doi:10.4303/mg/235551

Research Article

Evaluation of the RDP Classifier Accuracy Using 16S rRNA Gene Variable Regions

Claudia Vilo[1] **and Qunfeng Dong**[1,2]

[1] *Department of Biological Sciences, University of North Texas, 1155 Union Circle #305220, Denton, Texas 76203-5017, USA*
[2] *Department of Computer Science and Engineering, University of North Texas, 1155 Union Circle #305220, Denton, Texas 76203-5017, USA*
Address correspondence to Qunfeng Dong, qunfeng.dong@unt.edu

Received 20 February 2012; Accepted 12 March 2012

Abstract The RDP Classifier is a widely used bioinformatic program that performs taxonomic classification of 16S rRNA gene sequences. However, the accuracy of the program is not clear when it is applied to common PCR products of the 16S rRNA variable regions, which are heavily used in microbiome projects. In this study, full-length 16S rRNA gene alignments from the SILVA database were used to simulate the PCR products of the combined variable regions (i.e., V1–V3, V3–V5, and V6–V9). The classification accuracies obtained from RDP Classifier were evaluated for each of the simulated 16S rRNA regions, at different confidence score thresholds. Although minor bias was observed, the RDP Classifier achieved overall similar and accurate classification results when using the combined variable regions of the 16S rRNA gene, i.e., V1–V3, V3–V5, and V6–V9. Additional analysis showed that V2 and V4 were the most accurate among individual regions (i.e., V1 to V9).

Keywords microbiome; taxonomic classification

1 Introduction

Culture-independent 16S rRNA gene sequencing has been widely applied to examine microbial diversity [21]. The full-length 16S rRNA genes (about 1500 bp) can be used for accurate taxonomic identification based on the underlying sequence diversity among different bacterial species [2,3,17]. However, the current high-throughput DNA sequencing technologies can only produce 16S rRNA gene fragments, instead of full-length genes. For example, the Roche 454 [12] and Illumina technologies [7] typically produce sequence reads of 100–400 and 75–100 bp, respectively. Therefore, only fragments of the 16S rRNA gene can be obtained by using degenerate PCR primers designed to amplify selected variable regions. The 16S rRNA gene fragments containing the V1–V3, V3–V5, and V6–V9 regions have been extensively used in various human microbiome projects (e.g., [4,5,9,15,24,25]). The V1–V3, V3–V5, and V6–V9 regions correspond to the 16S rRNA gene fragment ranging from the V1 through V3 regions, V3 through V5 regions, and V6 through V9 regions, respectively. The gene fragments containing individual variable regions (e.g., V2, V3, or V4) have also been used [6,11,22] in other metagenomic studies. One critical concern is whether such partial 16S rRNA gene fragments can give an accurate microbial classification. Although such concern can be debated as a theoretical question (e.g., whether the phylogenetic resolution of each variable region is different); in practice, biologists are mostly interested in the performance of existing classification programs for the different 16S rRNA gene regions. One of the most extensively used bioinformatics programs for 16S rRNA classification is RDP Classifier, which is a naïve Bayesian classifier that provides taxonomic classification from domain to genus, as well as confidence estimates for each classification prediction [23]. The RDP Classifier has been widely used for rapid and accurate processing of high-throughput 16S rRNA datasets. Moreover, by the time this manuscript was prepared, the RDP Classifier program had been cited in more than 400 articles since its publication in 2007 [20]. Despite its tremendous popularity, there exist few published reports that have evaluated its classification accuracy using 16S rRNA gene fragments. For example, the study done by Wang et al. [23] reported the average classification accuracy of simulated 16S rRNA sequence fragments, whereas Liu et al. [10] reported that partial 16S rRNA gene sequences could achieve similar classification accuracy as the full-length gene. However, those studies did not specifically examine the classification accuracy of the V1–V3, V3–V5, and V6–V9 regions that are commonly used in various microbiome projects. In this study, we have evaluated the performance of RDP Classifier for those specific 16S rRNA gene fragments in the following aspects: (i) whether their taxonomic classification accuracies are similar, and (ii) whether they exhibit any classification bias towards certain taxonomic groups.

2 Materials and methods

The RDP Classifier program (version 2.2) and the SILVA rRNA database (SSURef version 102) [16] were used for this study. The SILVA rRNA database was chosen because of the high quality of its sequence alignments [18, 19]. The downloaded SILVA database consists of a multiple sequence alignment of 391,167 full-length, or near full-length, 16S rRNA gene sequences. For each sequence in the SILVA database, we obtained its taxonomic classification, from the genus to the phylum level, by using the RDP Classifier program. For subsequent analysis, we selected 274,196 sequences that exceeded the RDP Classifier confidence threshold of 0.8 at the genus level (i.e., the default threshold of RDP Classifier), to ensure a confident taxonomic classification. The selected sequences were then further clustered into 1,607 genera based on their RDP classification. Next, we randomly sampled one sequence from each genus cluster to create a test dataset. This random sampling was repeated ten times, thus, ten total test datasets were created for evaluating the accuracy of the RDP Classifier. For each 16S rRNA sequence in the test datasets, we extracted its V1–V3, V3–V5, V6–V9, V2–V3, and V2–V4 regions as well as the 9 individual regions (i.e., V1 through V9 individually) based on the aligned coordinates of the reference *E. coli* 16S rRNA gene [1, 2]. The RDP Classifier results of the extracted gene fragments were compared to those of the corresponding full-length gene sequences from the phylum to genus levels, at the confidence threshold scores of 0.7, 0.8, and 0.9. Similar analysis was also done by grouping sequences at the putative species level (data not shown).

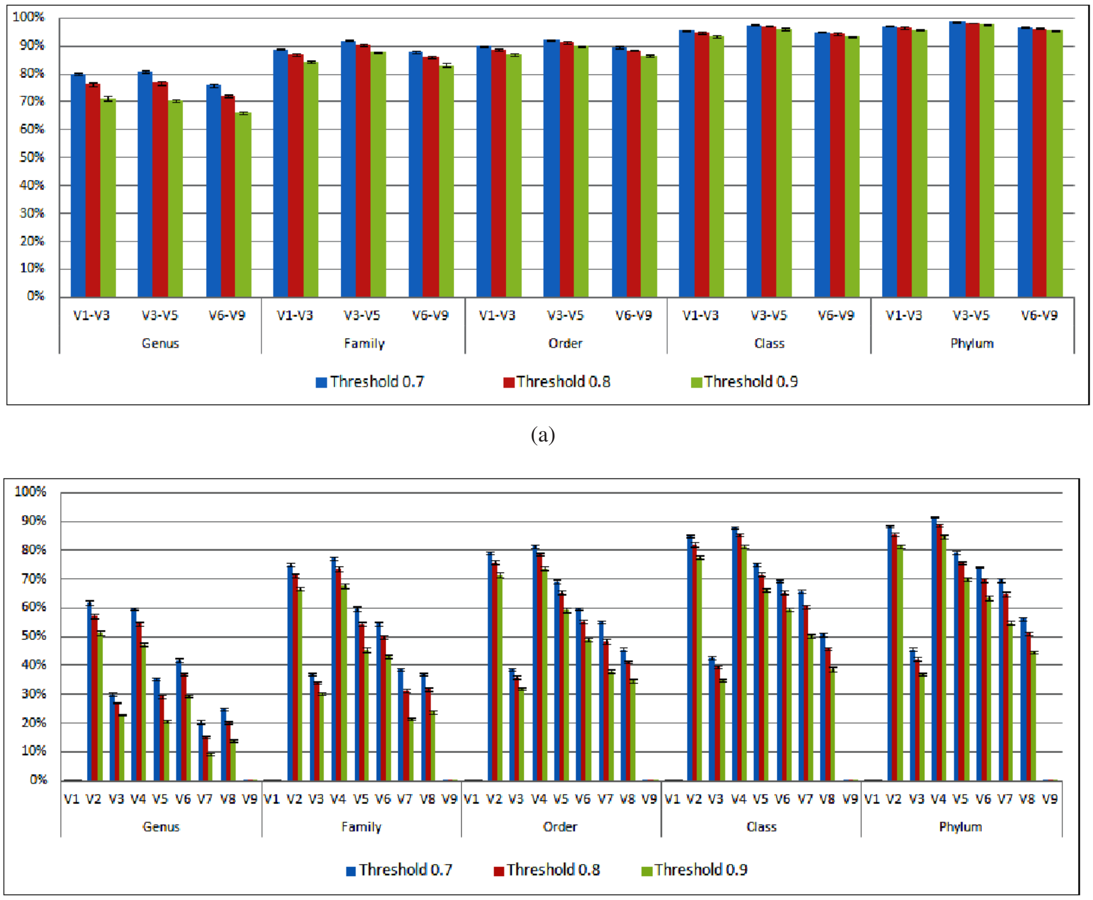
3 Results

Our results showed that the classification accuracies of the V1–V3, V3–V5, and V6–V9 combined variable regions were highly similar, providing sufficiently accurate classification using the RDP Classifier program (Figure 1(a)). For example, at the RDP Classifier confidence threshold score of 0.8 (i.e., the default threshold used at the RDP Classifier web server), the V1–V3 region accurately classified 96.83% of the phyla, 95.06% of the classes, 88.91% of the orders, 87.48% of the families, and 76.90% of the genera. Similarly, the V3–V5 region accurately classified 98.44% of the phyla, 97.22% of the classes, 91.41% of the orders, 90.61% of the families, and 77.15% of the genera. Also, the V6–V9 region accurately classified 96.43% of the phyla, 94.54% of the classes, 88.64% of the orders, 86.32% of the families, and 72.65% of the genera. Similar results were obtained either using the RDP Classifier confidence thresholds of 0.7 and 0.9 (Figure 1(a)) or using simulated reads containing 0.5% sequencing errors (Figure 1), indicating the robust classification potentials of each of the above regions (Figure 1(a)).

In addition, we evaluated whether the V1–V3, V3–V5, and V6–V9 regions exhibit any classification bias towards certain taxonomic groups. Our findings indicated that the observed biases were minimal, however, they could play an important role in samples with overrepresented taxa (Table 1). For example, V1–V3 seemed to be a better choice to classify *Alteromonas* than the other two regions; on the other hand, V3–V5 and V6–V9 were the best choices for *Erythrobacter* and *Erwinia*, respectively. It is interesting that such biases tend to be enriched in small subsets of higher taxonomic groups such as those that can be observed in the family *Enterobacteriaceae*, the order *Enterobacteriales*, the class *Gammaproteobacteria*, and the phylum *Proteobacteria*.

Table 1: *Potential classification biases of 16S rRNA variable regions were detected for some bacteria genera.* The first column of the table shows the genera names that exhibited the difference in classification accuracy (only the ones with at least 10% difference in their classification accuracy among each other are shown here). The second column shows the number of species clusters belonging to each of the listed genera. The classification accuracy of V1–V3, V3–V5 and V6–V9 are displayed in the subsequent columns. The results are based on the RDP Classifier confidence threshold score 0.8. Similar results can be observed using the threshold score 0.7 and 0.9 (data not shown).

Genus	**Number of species**	**Percentages of classification accuracy for combined variable regions**		
		V1–V3	**V3–V5**	**V6–V9**
Acidovorax	14	97.1	85	85.7
Actinobacillus	21	91	97.6	77.6
Actinoplanes	41	93.9	78	96.8
Alteromonas	12	99.2	86.7	87.5
Arthrobacter	65	95.4	86.5	97.8
Caulobacter	11	100	89.1	99.1
Chlorobium	10	100	100	85
Citrobacter	11	77.3	61.8	59.1
Enterobacter	17	67.6	58.8	54.1
Erwinia	13	73.1	53.8	92.3
Erythrobacter	11	79.1	100	80.9
Haemophilus	12	85.8	92.5	67.5
Leifsonia	15	82	79.3	95.3
Lysobacter	20	98.5	100	78
Massilia	11	100	80.9	97.3
Nonomuraea	28	85.7	100	86.1
Pasteurella	12	100	96.7	70.8
Pseudoxanthomonas	16	100	100	81.3
Rhodovulum	12	100	85	100
Selenomonas	12	96.7	87.5	98.3
Serratia	15	93.3	74	78
Streptosporangium	21	93.8	100	86.2
Thiomicrospira	11	98.2	81.8	100



(a)

(b)

Figure 1: *RDP Classifier accuracy evaluated with different 16S rRNA variable regions by using 0.5% simulated sequencing errors.* The sequences from the SILVA database were clustered into 1,607 genera; then one sequence was randomly selected from each genus cluster in order to create test datasets. For each test dataset, we extracted its V1–V3, V3–V5, and V6–V9 combined regions, and V1 to V9 individuals regions. Then, we simulated sequencing errors by replacing a 0.5% of the nucleotides by an "N", from all fragments in the dataset, randomly. The taxonomic classification of the selected full-length gene sequences were then compared to the classification of its V1–V3, V3–V5, and V6–V9 combined regions, and V1 to V9 individuals regions. The above procedure was repeated ten times by randomly sampling with replacements from each species cluster. The x-axis displays the variable regions evaluated at each of the different taxonomic levels (i.e., genus, family, order, class, and phylum). The y-axis shows the average percentages of classification accuracy. The accuracies were evaluated by using the RDP Classifier confidence threshold scores of 0.7 (blue), 0.8 (red), and 0.9 (green), with the 95% confidence interval being plotted. (a) Results for the combined variable regions V1–V3, V3–V5, and V6–V9. Overall, they provide very similar accuracy. (b) Results for each individual variable region V1 to V9.

We were also interested in comparing the classification accuracy among individual variable regions. Among the individual regions, the V2 and V4 regions showed the most accurate results at every taxonomical level. However, the V5, V6, V7, and V8 regions also increased their accuracies at the class and phylum levels to a point that they provided comparable classification results to those of the V2 and V4 regions (Figure 1(b)). At the RDP Classifier confidence threshold of 0.8, the results showed that the V2 region accurately classified 85.99% of the phyla, 82.52% of the classes, 76.50% of the orders, 72.07% of the families, and 58.13% of the genera. The V4 region accurately classified 89.28% of the phyla, 86.12% of the classes, 79.45% of the orders, 74.56% of the families, and 55.69% of the genera.

Additionally, we evaluated the V2–V3 and V2–V4 regions. The results of the V2–V3 region showed similar accuracy as the V1–V3 region. At the RDP Classifier confidence threshold of 0.8, the V2–V3 region accurately classified 96.52% of the phyla, 94.68% of the classes, 88.48% of the orders, 86.24% of the families, and 74.38% of the genera. In addition, the results showed that V2–V4 region provided slightly better accuracy at genus level compared to the other three combined regions (i.e., V1–V3, V3–V5, and V6–V9). For example, at RDP Classifier confidence threshold of 0.8, V2–V4 region accurately classified 98.2% of the phyla, 96.92% of the classes, 91.12% of the orders, 90.73% of the families, and 81.97% of the genera (data not shown).

4 Discussion

Limitations in budget, and sequencing platforms, often place biologists in the dilemma of deciding which of the 16S rRNA gene fragments they should select for sequencing. We have encountered such situations in our own microbiome projects, in which we have had to select from the V1–V3, V3–V5, and V6–V9 regions for bacteria 16S rRNA gene profiling. In our experience, the V1–V3, V3–V5, and V6–V9 regions have typically showed different microbial community compositions from the same samples. The observed differences are often attributed to the possibility that the designed PCR primers lack the ability to amplify the V1–V3, V3–V5, and V6–V9 regions of all the 16S rRNA genes with equal efficiency [17]. However, even with V1–V3, V3–V5, and V6–V9 perfect PCR amplification, there could exist an intrinsic classification bias in the RDP Classifier program towards those different 16S rRNA gene regions.

Although the V1–V3, V3–V5, and V6–V9 regions have increasingly been used for bacterial 16S gene profiling, no published results have evaluated their classification potentials with the RDP Classifier program. For example, Wang et al. [23] indirectly evaluated the accuracy of the variable regions by using 16S rRNA gene fragments that were 100 bp long. The gene fragments were extracted regardless of the variable regions positions. Specifically, they extracted a 100 bp window moving along the 16S rRNA genes sequences at a 25 bp interval. Using this approach, the most accurate results were obtained with the fragments of the gene that contained the regions V2 or V4. In addition, Liu et. al. [10] studied the accuracies of the variable regions by comparing the accuracies of fragments extracted from the gene sequences of several datasets. The fragments were extracted by using primers named with the start position from which the fragments were extracted. Using this start position, they then moved forward, or backward, through the gene sequences to extract fragments of 100 bp, 250 bp, or 400 bp in length. They showed that the best sets of primers that amplify 100 bp gene fragments were F343 (for V3 region), F517 (for V4 region), F784 (for V5 region), R357 (for region between V2 and V3), R534 (for V3 region), R798 (for region between V4 and V5), and R926 (for V5 region). Also, the primers that gave the best results for amplifying 250 bp gene fragments were R357 and 8F used together (for the region V2 and V3). Jeraldo et al. [8] studied the phylogenetic information contained in fragments of the 16S rRNA gene between 120 and 400 base pairs long. Specifically, they investigated variable regions V3, V6, and V1–V3 by extracting them from the Greengenes 16S rRNA database. Their analysis showed that the V1–V3 region was the best estimator of phylogeny out of those three variable regions. Another study by Chakravorty et al. [2], evaluated the 16S variable region accuracies of 110 pathogenic and environmental bacteria species. Their results showed that the variable regions V2, V3, and V6 were more accurate than V4, V5, V7, and V8 for species classification in their test datasets. However, none of the previous studies specifically evaluated the V1–V3, V3–V5, and V6–V9 regions.

In this study, we simulated the V1–V3, V3–V5, and V6–V9 regions for comparison in order to show their intrinsic classification potential with the RDP Classifier, which is perhaps the most popular 16S rRNA classification program used in this field. Our results indicated that the V1–V3, V3–V5, and V6–V9 regions showed similar accuracy results (Figure 1(a)). These findings are consistent with Liu et al. [10] study, which showed that short fragments of 100 bp, belonging to the V1–V3 and V3–V5 region, could provide accurate classifications. Also, Jeraldo et al. [8] study showed that V1–V3 region was a good estimator of phylogenetic information. In addition, Chakravorty et al. [2] study showed that combining the V2, V3, and V6 variable regions provided accurate bacterial identification. In addition, we only observed few potential biases towards certain taxonomic groups by V1–V3, V3–V5, and V6–V9 region (Table 1).

Although the classification accuracy at higher taxonomic levels was good for several of the individual variable regions, the accuracy at the genus level was, in general, less accurate, depending on the specific variable region (Figure 1(b)). Our results showed that the V2 and V4 regions delivered the best results compared to other individual regions. Wang et al. [23] also reported that the fragments of the gene containing the variable regions V2 or V4 provided better taxonomic classification accuracy. Since the V2 and V4 regions are the longest among other individual variable regions (105 bp and 106 bp respectively, according to the reference *E. coli* 16S rRNA gene), they may potentially provide more phylogenetic signals that would allow for a more accurate taxonomic classification. Our results are consistent with the findings of Nossa et al. [14] who also showed that the longer 16S sequence amplicons could improve the accuracy of the bacterial classification.

Since the V1 variable region by itself provided very low accuracy, we additionally compared the V2–V3 and V1–V3 regions. Our results showed that a slight difference in the accuracies between these two regions existed only at the genus level. On the other hand, since the V2 and V4 regions were the most accurate among individual regions, we compared the V2–V4 to the V1–V3, V3–V5, and V6–V9 regions. Our results showed that V2–V4 provides better accuracy at genus level compared to the other three combined regions (i.e., V1–V3, V3–V5, and V6–V9) (data not shown).

In this study, we have chosen to extract theoretic 16S variable regions, i.e., each region was defined based on its position on the *E. Coli* 16S gene. The regions that are amplified by actual primers used in various microbiome studies may deviate from the simulated regions used in this study. However, instead of trying to select a particular set of primers (different group design their own primers (e.g., [2,5,10]) to target the variable regions), we intend to show the intrinsic classification accuracies of the variable regions without having to be specific to a particular primer choice. The length distribution of our simulated data set (V1–V3, mean = 415 bp, standard deviation (SD) = 26; V3–V5, mean = 437 bp, SD = 12; V6–V6, mean = 452 bp, SD = 37) is highly similar to real data sets that we have worked on (e.g., V1–V3, mean = 471 bp, SD = 51; V3–V5, mean = 497 bp, SD = 60; V6–V6, mean = 477 bp, SD = 51 [13]), indicating that our observations based on the simulated data also apply to the real amplicons.

5 Conclusion

Overall, the RDP Classifier achieved similar and accurate classification results when using the combined variable regions of the 16S rRNA gene, i.e., V1–V3, V3–V5, and V6–V9 in our simulation study. In addition, we only observed few potential biases towards certain taxonomic groups by V1–V3, V3–V5, and V6–V9 region.

Acknowledgments This work was supported in part by NIH/NHLBI U01 HL098960-01 and NIH/NIAID UH3 AI094641. The authors thank Marco Arancibia for his assistance in the design and implementation of the experiments, and in the preparation of the manuscript. The authors would also like to thank Michael Plunkett and Daniel Munro for their critical reading of this manuscript.

References

[1] G. C. Baker, J. J. Smith, and D. A. Cowan, *Review and re-analysis of domain-specific 16S primers*, J Microbiol Methods, 55 (2003), 541–555.

[2] S. Chakravorty, D. Helb, M. Burday, N. Connell, and D. Alland, *A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria*, J Microbiol Methods, 69 (2007), 330–339.

[3] J. E. Clarridge 3rd, *Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases*, Clin Microbiol Rev, 17 (2004), 840–862.

[4] F. E. Dewhirst, T. Chen, J. Izard, B. J. Paster, A. C. Tanner, W. H. Yu, et al., *The human oral microbiome*, J Bacteriol, 192 (2010), 5002–5017.

[5] Q. Dong, D. E. Nelson, E. Toh, L. Diao, X. Gao, J. D. Fortenberry, et al., *The microbial communities in male first catch urine are highly similar to those in paired urethral swab specimens*, PLoS One, 6 (2011), e19709.

[6] C. Humblot and J. P. Guyot, *Pyrosequencing of tagged 16S rRNA gene amplicons for rapid deciphering of the microbiomes of fermented foods such as pearl millet slurries*, Appl Environ Microbiol, 75 (2009), 4354–4361.

[7] Illumina, http://www.illumina.com/company/technology.ilmn.

[8] P. Jeraldo, N. Chia, and N. Goldenfeld, *On the suitability of short reads of 16S rRNA for phylogeny-based analyses in environmental surveys*, Environ Microbiol, 13 (2011), 3000–3009.

[9] V. Lazarevic, K. Whiteson, D. Hernandez, P. François, and J. Schrenzel, *Study of inter- and intra-individual variations in the salivary microbiota*, BMC Genomics, 11 (2010), 523.

[10] Z. Liu, T. Z. DeSantis, G. L. Andersen, and R. Knight, *Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers*, Nucleic Acids Res, 36 (2008), e120.

[11] C. Manichanh, J. Reeder, P. Gibert, E. Varela, M. Llopis, M. Antolin, et al., *Reshaping the gut microbiome with bacterial transplantation and antibiotic intake*, Genome Res, 20 (2010), 1411–1419.

[12] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, et al., *Genome sequencing in microfabricated high-density picolitre reactors*, Nature, 437 (2005), 376–380.

[13] D. E. Nelson, Q. Dong, B. Van der Pol, E. Toh, B. Fan, B. P. Katz, et al., *Bacterial communities of the coronal sulcus and distal urethra of adolescent males*, PLoS One, 7 (2012), e36298.

[14] C. W. Nossa, W. E. Oberdorf, L. Yang, J. A. Aas, B. J. Paster, T. Z. Desantis, et al., *Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome*, World J Gastroenterol, 16 (2010), 4135–4144.

[15] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, et al., *The NIH Human Microbiome Project*, Genome Res, 19 (2009), 2317–2323.

[16] E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, et al., *SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB*, Nucleic Acids Res, 35 (2007), 7188–7196.

[17] J. Rajendhran and P. Gunasekaran, *Microbial phylogeny and diversity: Small subunit ribosomal RNA sequence analysis and beyond*, Microbiol Res, 166 (2011), 99–110.

[18] P. D. Schloss, *A high-throughput DNA sequence aligner for microbial ecology studies*, PLoS One, 4 (2009), e8230.

[19] P. D. Schloss, *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 (2010), e1000844.

[20] Science Citation Index, *ISI Web of Knowledge*. http://apps.webofknowledge.com, 2010.

[21] C. Simon and R. Daniel, *Metagenomic analyses: past and future trends*, Appl Environ Microbiol, 77 (2011), 1153–1161.

[22] P. J. Turnbaugh and J. I. Gordon, *The core gut microbiome, energy balance and obesity*, J Physiol, 587 (2009), 4153–4158.

[23] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, *Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy*, Appl Environ Microbiol, 73 (2007), 5261–5267.

[24] G. D. Wu, J. D. Lewis, C. Hoffmann, Y. Y. Chen, R. Knight, K. Bittinger, et al., *Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags*, BMC Microbiol, 10 (2010), 206.

[25] S. Yildirim, C. J. Yeoman, M. Sipos, M. Torralba, B. A. Wilson, T. L. Goldberg, et al., *Characterization of the fecal microbiome from non-human wild primates reveals species specific microbial communities*, PLoS One, 5 (2010), e13963.