Identification and Quantification of Abundant Species from Pyrosequences of 16S rRNA by Consensus Alignment

Yuzhen Ye
School of Informatics and Computing, Bloomington, IN 47408, U.S.A
E-mail: yye@indiana.edu

***Abstract*—16S rRNA gene profiling has recently been boosted by the development of pyrosequencing methods. A common analysis is to group pyrosequences into Operational Taxonomic Units (OTUs), such that reads in an OTU are likely sampled from the same species. However, species diversity estimated from error-prone 16S rRNA pyrosequences may be inflated because the reads sampled from the same 16S rRNA gene may appear different, and current OTU inference approaches typically involve time-consuming pairwise/multiple distance calculation and clustering. I propose a novel approach AbundantOTU based on a Consensus Alignment (CA) algorithm, which infers consensus sequences, each representing an OTU, taking advantage of the sequence redundancy for abundant species. Pyrosequencing reads can then be recruited to the consensus sequences to give quantitative information for the corresponding species. As tested on 16S rRNA pyrosequence datasets from mock communities with known species, AbundantOTU rapidly reported identified sequences of the source 16S rRNAs and the abundances of the corresponding species. AbundantOTU was also applied to 16S rRNA pyrosequence datasets derived from real microbial communities and the results are in general agreement with previous studies.**

***keywords*—16S rRNA gene; pyrosequencing; Operational Taxonomic Unit (OTU); abundant species**

I. Introduction

16S rRNA gene profiling has been applied to the analysis of complex microbial populations since the middle 1990s [1], but has recently been boosted by advances in sequencing techniques that can produce large 16S rRNA datasets containing hundreds thousands of 16S RNAs fragments spanning the hypervariable regions of 16S rRNA genes, enabling deep views into hundreds of microbial communities [2][3]. 16S rRNA pyrosequencing studies have revealed much greater species diversity in many environments (e.g., soils, ocean water, and human bodies) than previously anticipated—although it was shown in one study [4] that some of the projects may overestimate the species diversity—and have great potential in a variety of applications.

16S rRNA pyrosequences can be mapped onto the phylogenetic tree of known 16S rRNA sequences to provide a view of the taxonomic distribution of the species they represent. In this type of approach, however, 16S rRNA pyrosequences can only be mapped to known species or branches. Alternatively, taxonomy independent analysis can be applied when sequences are classified into Operational Taxonomic Units (OTUs) of specified sequence variations, although it is still arguable which similarity cutoff should be used to define species (or genus) [5]. Typically, sequences with $< 3\%$ dissimilarity are assigned to the same species, while those with $< 5\%$ dissimilarity are assigned to the same genus [5].

Typical methods of OTU inference cluster sequences into groups—either through pairwise comparison or multiple alignment, followed by sequence clustering, as in mothur [6] and ESPRIT [7], or derived by heuristic clustering of sequencing reads as in FastGroupII [8] and CD-HIT [9][10]—and then a consensus sequence may be derived for each group of sequences. The pitfalls of these approaches are: 1) without knowing the consensus sequence (central sequence) of a group, sequences may be mistakenly classified into the group; 2) deriving consensus sequence from a group of sequences is nontrivial, often requiring a multiple sequence alignment; and 3) clustering algorithms often involve all-against-all pairwise comparison of the reads, which requires large memory and long computational time [7].

A fast approach, AbundantOTU, is proposed that first finds consensus sequences, without any clustering, followed by recruitment of reads to the consensus sequences. A consensus sequence may represent an OTU, and thus the reads recruited to the same consensus sequence can contain sequences of different strains of the same species, or error-prone reads from the same strain. This approach was inspired by the fact that if one knew the species composition of a microbial community, it would be easy to assign error-prone reads to their *source* 16S rRNAs. AbundantOTU utilizes the redundant sequence information (even though the reads may contain sequencing errors), so that consensus sequences can be derived by a Consensus Alignment (CA) algorithm. For low abundant sequences, it is difficult to determine if they are sampled from rare species (they represent rare species), or if they cannot be recruited due to high sequencing errors. Although characterizing rare species is not the focus of the paper, AbundantOTU can be applied first to group the abundant sequences, and then the remaining sequences (typically a small proportion of the original sequences) can be further analyzed by using other tools, such as mothur [6] and PyroNoise [4].

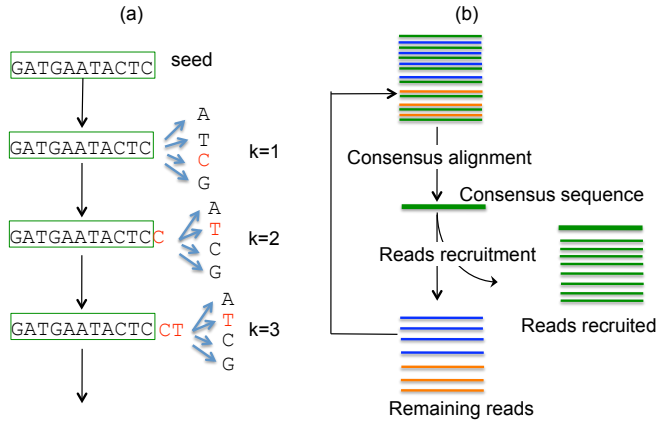

Figure 1. A schematic demonstration of AbundantOTU algorithm by using consensus alignment. (a) Consensus alignment by using a dynamic programming algorithm, adding one nucleotide at a time. (b) Abundant OTU inference by deriving consensus sequence and recruiting reads to the consensus sequence iteratively.

II. Methods

AbundantOTU starts by finding the consensus sequence of a group of sequences by a consensus alignment algorithm, followed by assigning sequences to the consensus sequence (Fig. 1). It may also be used as a generic tool for consensus sequence inference for a predefined group of DNA sequences. The consensus alignment algorithm is similar to the algorithm that was proposed by Li and colleagues for finding similar regions in many strings [11] and an algorithm that was developed for repeat detection in genomic sequences [12].

A. Consensus alignment algorithm

Given a collection of m pyrosequences $S = s^1, s^2, ..., s^m$, the consensus alignment problem is to find the optimal consensus sequence that is similar to the most input sequences. Denote a consensus sequence as s^c. The consensus alignment evaluates the similarity between the consensus sequence and all of the input sequences by an objective similarity function as,

$$Sim(s^c, S) = \sum_{i=1}^{m} Sim(s^c, s^i) \quad (1)$$

where $Sim(s^c, s^i)$ is the sequence similarity between the consensus sequence s^c, and an input sequence s^i.

To obtain a consensus sequence with the optimal similarity function, a frequent l-mer ($l = 40$ by default) is first identified in the input sequences, using the hashing technique. The frequent l-mer, which serves as a seed, can then be extended forward and backward to obtain the entire consensus sequence. A greedy strategy is adopted that adds one nucleotide at a time with the highest similarity between the growing consensus sequence with the input sequences that share the same seed (Fig. 1a). The forward and backward extension can be achieved by using the same algorithm. Here I use the forward extension as an example to illustrate the algorithm.

Assume $k-1$ nucleotides has been extended to the consensus sequence, denoted as $s^c_{k-1} = n_1 n_2 ... n_{k-1}$ (where n_i is the nucleotide at position i; note here the position index is relative to the end of the seed). The optimal nucleotide (either A, T, C or G) to be added at position k in the consensus sequence, n_k, can be computed as,

$$n_k = \underset{n \in \{A,T,C,G\}}{argmax} \sum_{i}^{m} Sim(s^c_{k-1} n, s^i) \quad (2)$$

where

$$Sim(s^c_{k-1} n, s^i) = \underset{k-\epsilon \le j \le k+\epsilon}{max} Sim(s^c_{k-1} n, s^i_j) \quad (3)$$

Here ϵ is the alignment bandwidth that is used to speed up the alignment between the consensus sequence $s^c_{k-1} n$ and pyrosequence s_i (i.e., position k of the consensus sequence will be allowed to align to a limited number of positions of sequence i). We used $\epsilon = 5$ by default, considering that sequences grouped in the same species-level OTU can not differ too much from each other [1]. Also, only the sequences that have the seed are compared to the consensus sequence. The best alignment score between the consensus sequence and sequence i with aligned positions of k (in the consensus sequence) and j (in sequence i) can be computed by a dynamic programming algorithm as follows.

$$Sim(s^c_{k-1} n, s^i_j) = max \begin{cases} Sim(s^c_{k-1}, s^i_{j-1}) + Score(n, s^i_j) \\ Sim(s^c_{k-1}, s^i_j) - g \\ Sim(s^c_{k-1} n, s^i_{j-1}) - g \end{cases} \quad (4)$$

where $Score(n, s^i_j)$ is the similarity score between two nucleotides n and s^i_j, and g is the gap penalty. Here we consider a simple scoring function: $Score(n, s^i_j) = 1$ if $n = s^i_j$, and 0 otherwise.

The initialization of the alignment is as follows.

$$Sim(s^c_0, s^i_j) = -g * j \ \forall \ 0 \le j \le \epsilon \quad (5)$$

$$Sim(s^c_{k-1} n, s^i_{k-\epsilon-1}) = -\infty \ \forall \ k > 1 \quad (6)$$

B. OTU inference

OTU inference can be achieved by applying iteratively the consensus alignment algorithm until no more consensus sequences can be found that recruit a minimum number of reads (here 5 is used, considering that fewer reads will be insufficient for the inference of their consensus sequence),

[1]Note that it takes exponential time to explore all possible sequences if the seeds are extended rigorously.

as shown in Fig. 1b. Once a new consensus sequence is found, all the reads that are nearly identical (e.g., with sequence difference $\leq 3\%$) to the consensus sequence are recruited to the consensus sequence, and assigned to the same species-level OTU. Note a read that does not share the same seed as the consensus sequence (so was not used for defining the consensus sequence) can still be recruited to the same consensus sequence as long as their overall sequence difference is $\leq 3\%$. Finally, the consensus sequences that recruit < 5 reads each are discarded.

C. Computational time of AbundantOTU

For a single step of consensus alignment, the computational complexity is $O(\epsilon L m_{seed})$, where ϵ is the bandwidth, L is the average length of a consensus sequence (which is approximately the average length of the input sequences), and m_{seed} is the total number of input sequences that contain the seed to be extended ($m_{seed} \leq m$; m is the total number of input sequences). As ϵ is a small constant (5 by default), the computational complexity of a single step of consensus alignment is equivalent to $O(L m_{seed})$. Since AbundantOTU involves iterative consensus alignment until no more abundant OTU can be inferred, the total computational complexity of AbundantOTU is $O(k L m_{seed})$, where k is the total number of consensus sequences that can be inferred, a number that is typically much smaller than m, the total number of input sequences.

D. Taxonomic analysis of the consensus sequences

Once the consensus sequences are derived, they can be passed to other tools for further taxonomic analysis. One of the advantages of using consensus sequences is that the number of consensus sequences is significantly smaller than the original pyrosequences, and thus dramatically decreases the computational time of downstream analysis, such as phylogenetic mapping using megablast or other rigorous phylogentic mapping methods ([13]; [14]). Here we used the RDP online classifier (http://rdp.cme.msu.edu/) [15], and BLAST search against Greengene 16S rRNA gene database [16] downloaded from http://greengenes.lbl.gov/cgi-bin/nph-index.cgi (as of Jan 20, 2010) for taxonomic assignments.

E. Benchmarks and tests

We tested AbundantOTU on two mock datasets, for which the microbial composition and reference sequences are known, and three metagenomic datasets derived from real communities (see Table I for the summary of the datasets). The first mock dataset (designated as Priest09) is the 'divergent sequence' dataset from [4] that contains amplified and pyrosequenced sequences from 23 divergent 16S rRNA fragments spanning V5 (the pyrosequences dataset and reference sequences were downloaded from http://people.civil.gla.ac.uk/∼quince/Data/PyroNoise.html). The second mock dataset (designated as Mock07) contains short sequences generated by pyrosequencing PCR amplicon libraries of 43 known 16S rRNA gene fragments spanning V6 using the Roche GS20 system, generated in a study of sequencing errors [17]. This dataset was downloaded from http://genomebiology.com/2007/8/7/R143. The three real metagneomic datasets contain reads from oral [18] and skin [2] samples, respectively, downloaded from the NCBI Short Read Archive (SRA) with accession numbers SRR002260 (oral/plaque), SRR002259 (oral/saliva), and SRR00606 (skin).

F. Availability

The sources codes of AbundantOTU are available at http://omics.informatics.indiana.edu/AbundantOTU.

III. RESULTS

The AbundantOTU results are summarized in Table I. The results show that AbundantOTU can generate consensus sequences that are identical or very similar to the known reference sequences in the mock communities. For the 16S rRNA datasets derived from environmental samples, the results are in general agreement with previous studies. Note that we focused on comparisons with methods that do not require computationally expensive pairwise/multiple alignments and/or clustering algorithms. For both CD-HIT and FastGroupII, we used 97% similarity (i.e., 3% dissimilarity) threshold. And our results show that AbundantOTU tolerates sequencing errors and gives reliable estimations of abundance of the species represented in the dataset. Due to the page limit, we only show detailed analysis of two datasets below.

A. Evaluation on mock community Priest09

Priest09 dataset contains reads sampled from the V5 regions of 23 divergent 16S rRNA genes. Since the reference sequences are known, we mapped each of the reads to the reference sequence with lowest distance to get the expected abundance level for each of the reference sequences, using the crossmatch program from the phrap package (http://www.phrap.org/). AbundantOTU reported 24 abundant OTUs (the least abundant OTU contains only 14 reads), which recruited most of the sequences included in the dataset (99.9%). We also compared the representative sequences of these abundant OTUs (their consensus sequences) to the known reference sequences. Overall, the consensus sequences inferred by AbundantOTU are identical or very similar to the known reference sequences: 5 are identical to the reference sequences and 13 differ from the reference sequences by one indel or mismatch (Fig. 2). By contrast, the representative sequences derived by CD-HIT are less similar to the known reference sequences. Further, the abundance-rank curves of this database derived by three different methods and the expected abundance-curve (Fig. 3) show that AbundantOTU produced the abundance rank

Table I
SUMMARY OF THE ABUNDANTOTU RESULTS OF FIVE DATASETS.

Datasets	Number of reads	Number of abundant OTUs	Reads recruited to		Running time
			The most abundant OTU	All abundant OTUs	
Priest09	38,351	24	2,480 (6.5%)	38,299 (99.9%)	1 min
Mock07	340,150	75	56,116 (16.5%)	334,136 (98.2%)	2 min
Oral/saliva	100,537	109	14,166 (14.1%)	86,973 (86.5%)	5 min
Oral/plaque	172,659	149	16,312 (9.4%)	144,859 (83.9%)	11 min
Skin	496,499	234	143,423 (28.9%)	434,617 (87.5%)	146 min

OTUs that recruit at least 5 reads are considered as abundant; all the calculations were carried out on a linux computer (Intel Xeon 2.93GHz)

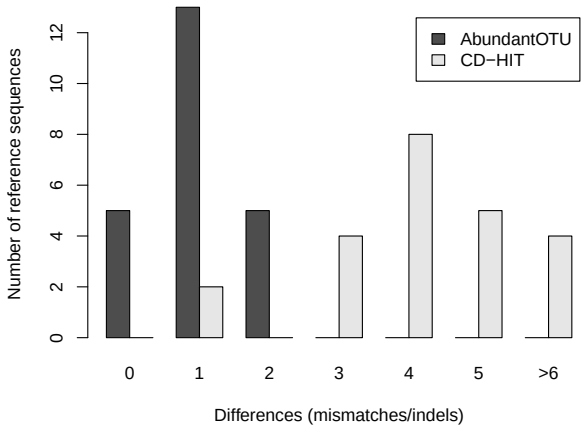

Figure 2. Comparison of the differences between the inferred and known reference sequences. The differences are measured as the total number of mismatchs and indels involved in aligning a reference sequence with the inferred sequence. The difference of 0 means that the inferred sequence is identical to the corresponding reference sequence.

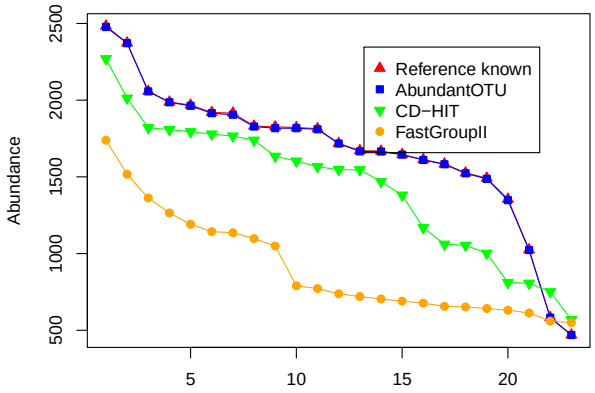

Figure 3. The abundance-rank curves of the Priest09 dataset using different methods. OTUs/clusters are plotted from most to least abundant along the x-axis, with their abundances displayed on the y-axis. The curves only show the high abundant OTUs/clusters. The reference curve shows the best result that any method can achieve, in that the reference sequences are known so that sequencing reads can be mapped to the references directly. The AbundantOTU curve overlaps nicely with the reference curve.

that is closest to the expected curve, whereas FastGroupII tends to produce more but smaller OTUs.

B. Evaluation on a skin-associated microbial community

AbundantOTU analysis of the skin dataset revealed several very abundant species (see Table II). The top species identified by AbundantOTU are in general agreement with the abundant species in the original report ([2]) (the top three species are the same), with some exceptions. Consensus sequence 8 and 9 were classified as from the same genus but represent different species—these two sequences only share 91% sequence identify. BLAST search shows that consensus sequence 8 is identical to *Streptococcus salivarius*, and consensus sequence 9 is identical to *Streptococcus sanguinis* strain GumJ19. Interestingly, consensus sequence 6 is identical to a fragment in the *Arachis hypogaea* (peanut) chloroplast rRNA gene, but there is no discussion about whether reads sampled from chroloplast rRNA are present or if these large number of sequences (more than 7000 sequences are recruited to this consensus sequence by AbundantOTU) were filtered out in [2].

Note that the dataset we downloaded from NCBI SRA contains 496,499 sequences, while the analysis presented in [2] was based on a filtered dataset that contains only 351,630 sequences, i.e., 71% of the original sequences (others did not pass the quality control [2]). AbundantOTU revealed 144 abundant OTUs, which recruited 434,617 (87.5%) of the pyrosequences. This suggests that AbundantOTU is sequencing error tolerant, and can utilize sequences that otherwise will be discarded due to sequencing errors.

Table II
SUMMARY OF THE TOP 10 MOST ABUNDANT OTUS IDENTIFIED BY ABUNDANTOTU FROM THE SKIN DATASET

ID	Reads	Taxonomic assignment	
		Phylum	Family
1	143,423 (28.9%)	Actinobacteria	Propionibacteriaceae
2	41,861 (8.4%)	Firmicutes	Streptococcaceae
3	28,299 (5.7%)	Firmicutes	Staphylococcaceae
4	10200 (2.1%)	Actinobacteria	Micrococcaceae
5	10176 (2.0%)	Firmicutes	Streptococcaceae
6	8,664 (1.7%)	Cyanobacteria	Chloroplast
7	7,631 (1.5%)	Firmicutes	Staphylococcaceae
8	7,081 (1.4%)	Firmicutes	Streptococcaceae
9	7,046 (1.4%)	Firmicutes	Streptococcaceae
10	6,760 (1.4%)	Actinobacteria	Corynebacteriaceae

Note the consensus sequences were taxonomically assigned using the online RDP classifier (http://rdp.cme.msu.edu/).

IV. Discussion

AbundantOTU takes advantage of the redundancy of the pyrosequences of 16S rRNA to infer the OTUs and their representative sequences, using a consensus alignment algorithm, assuming that there should be more reads that have the correct nucleotide than reads with sequencing errors at a particular position. As such, AbundantOTU is robust, and less affected by sequencing errors with sequencing errors. Since it relies on sequence redundancy, it cannot be used to identify rare species, which are only represented by one or very few sequences. It is challenging to infer rare species correctly, since these species are barely represented by sequencing reads and sequencing errors may be difficult to spot (if not impossible). AbundantOTU can report the sequences that are not recruited to the abundant OTUs, so that further analysis can be carried out on these 'singleton" sequences. But we believe that these rare sequences (sampled from rare species or sequences from abundant species but with high sequencing errors) should be treated cautiously; otherwise, they may cause inflated species diversity estimations ([4]). In this paper, the algorithm has been tested on pyrosequences of 16S rRNA genes. But it can also be applied to sequences of 16S rRNA genes derived from any sequencing machine, as long as redundant sequences exist.

AbundantOTU can be combined with other taxonomy-based analysis of pyrosequences of 16S rRNAs. For example, as we demonstrated in the analysis of skin dataset, taxonomic assignment of the representative sequences of the abundant OTUs can be achieved using the RDP classifier. Using AbundantOTU can reduce the computational time, since often the pyrosequence datasets are dominated by a few abundant species, and deriving these sequences with AbundantOTU is fast in comparison to taxonomic assignment on the entire dataset. Rigorous but time-consuming taxonomic assignment may then be pursued on the few consensus sequences derived from AbundantOTU. Another potential application of AbundantOTU is to combine it with more time-consuming OTU inference methods that rely on clustering of reads, based on their pairwise distances. Abundant OTUs can be inferred by AbundantOTU (taking advantage of the fact that it achieves fast and accurate inference of abundant OTUs), and the remaining sequences can then be analyzed by those methods.

V. Acknowledgement

The author would like to thank Drs. Haixu Tang and Thomas G. Doak for helpful discussions and reading the manuscript. This work was supported by National Institutes of Health grants (1R01HG004908-02 and 1U01HL098960-01).

References

[1] G. Muyzer, E. C. de Waal, and A. G. Uitterlinden, "Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA," *Appl. Environ. Microbiol.*, vol. 59, pp. 695–700, Mar 1993.

[2] N. Fierer, M. Hamady, C. L. Lauber, and R. Knight, "The influence of sex, handedness, and washing on the diversity of hand surface bacteria," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, pp. 17 994–17 999, Nov 2008.

[3] M. Hamady, J. J. Walker, J. K. Harris, N. J. Gold, and R. Knight, "Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex," *Nat. Methods*, vol. 5, pp. 235–237, Mar 2008.

[4] C. Quince, and et al, "Accurate determination of microbial diversity from 454 pyrosequencing data," *Nat. Methods*, vol. 6, pp. 639–641, Sep 2009.

[5] S. A. Breglia, N. Yubuki, M. Hoppenrath, and B. S. Leander, "Ultrastructure and molecular phylogenetic position of a novel euglenozoan with extrusive episymbiotic bacteria: Bihospites bacati n. gen. et sp. (Symbiontida)," *BMC Microbiol.*, vol. 10, p. 145, 2010.

[6] P. D. Schloss, and et al, "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Appl. Environ. Microbiol.*, vol. 75, pp. 7537–7541, Dec 2009.

[7] Y. Sun, and et al, "ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences," *Nucleic Acids Res.*, vol. 37, p. e76, Jun 2009.

[8] Y. Yu, M. Breitbart, P. McNairnie, and F. Rohwer, "FastGroupII: a web-based bioinformatics platform for analyses of large 16S rDNA libraries," *BMC Bioinformatics*, vol. 7, p. 57, 2006.

[9] W. Li, L. Jaroszewski, and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases," *Bioinformatics*, vol. 17, pp. 282–283, Mar 2001.

[10] E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight, "Bacterial community variation in human body habitats across space and time," *Science*, vol. 326, pp. 1694–1697, Dec 2009.

[11] M. Li, B. Ma, and L. Wang, "Finding similar regions in many strings," in *STOC '99: Proceedings of the thirty-first annual ACM symposium on Theory of computing*. New York, NY, USA: ACM, 1999, pp. 473–482.

[12] A. L. Price, N. C. Jones, and P. A. Pevzner, "De novo identification of repeat families in large genomes," *Bioinformatics*, vol. 21 Suppl 1, pp. i351–358, Jun 2005.

[13] C. von Mering, and et al, "Quantitative phylogenetic assessment of microbial communities in diverse environments," *Science*, vol. 315, pp. 1126–1130, Feb 2007.

[14] M. Wu and J. A. Eisen, "A simple, fast, and accurate method of phylogenomic inference," *Genome Biol.*, vol. 9, p. R151, 2008.

[15] J. R. Cole, and et al, "The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis," *Nucleic Acids Res.*, vol. 33, pp. D294–296, Jan 2005.

[16] T. Z. DeSantis, and et al, "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB," *Appl. Environ. Microbiol.*, vol. 72, pp. 5069–5072, Jul 2006.

[17] S. M. Huse, J. A. Huber, H. G. Morrison, M. L. Sogin, and D. M. Welch, "Accuracy and quality of massively parallel DNA pyrosequencing," *Genome Biol.*, vol. 8, p. R143, 2007.

[18] B. J. Keijser, and et al, "Pyrosequencing analysis of the oral microflora of healthy adults," *J. Dent. Res.*, vol. 87, pp. 1016–1020, Nov 2008.