Hindawi
BioMed Research International
Volume 2018, Article ID 3130607, 14 pages
https://doi.org/10.1155/2018/3130607

Research Article

Composition Analysis and Feature Selection of the Oral Microbiota Associated with Periodontal Disease

Wen-Pei Chen,[1] Shih-Hao Chang,[2,3] Chuan-Yi Tang,[4] Ming-Li Liou,[5] Suh-Jen Jane Tsai,[1] and Yaw-Ling Lin[1,4]

[1]*Department of Applied Chemistry, Providence University, Taichung City, Taiwan*
[2]*Department of Periodontics, Linkou Medical Center, Chang Gung Memorial Hospital, Taoyuan, Taiwan*
[3]*Graduate Institute of Dental and Craniofacial Science, Chang Gung University, Taoyuan, Taiwan*
[4]*Department of Computer Science and Information Engineering, Providence University, Taichung City, Taiwan*
[5]*Department of Medical Laboratory Science and Biotechnology, Yuanpei University of Medical Technology, Hsin-Chu City, Taiwan*

Correspondence should be addressed to Yaw-Ling Lin; yllin@pu.edu.tw

Received 30 May 2018; Revised 10 October 2018; Accepted 4 November 2018; Published 15 November 2018

Academic Editor: Momiao Xiong

Periodontitis is an inflammatory disease involving complex interactions between oral microorganisms and the host immune response. Understanding the structure of the microbiota community associated with periodontitis is essential for improving classifications and diagnoses of various types of periodontal diseases and will facilitate clinical decision-making. In this study, we used a 16S rRNA metagenomics approach to investigate and compare the compositions of the microbiota communities from 76 subgingival plagues samples, including 26 from healthy individuals and 50 from patients with periodontitis. Furthermore, we propose a novel feature selection algorithm for selecting features with more information from many variables with a combination of these features and machine learning methods were used to construct prediction models for predicting the health status of patients with periodontal disease. We identified a total of 12 phyla, 124 genera, and 355 species and observed differences between health- and periodontitis-associated bacterial communities at all phylogenetic levels. We discovered that the genera *Porphyromonas*, *Treponema*, *Tannerella*, *Filifactor*, and *Aggregatibacter* were more abundant in patients with periodontal disease, whereas *Streptococcus*, *Haemophilus*, *Capnocytophaga*, *Gemella*, *Campylobacter*, and *Granulicatella* were found at higher levels in healthy controls. Using our feature selection algorithm, random forests performed better in terms of predictive power than other methods and consumed the least amount of computational time.

1. Introduction

The human mouth harbors a complex microbial community, with estimates of up to 700 or more different bacterial species, most of which are commensal and required to maintain the balance of the mouth ecosystem [1]. However, some of the bacteria in the mouth microbiota play important roles in the development of oral diseases, including dental caries and periodontal disease [2]. Periodontal disease and dental caries initiate with the growth of the dental plaque, a biofilm formed by the accumulation of bacteria together with various human salivary glycoproteins and polysaccharides secreted by the microbes [3]. The subgingival plaque, located within the neutral or alkaline subgingival sulcus, is typically inhabited by anaerobic gram-negative bacteria and is responsible for the development of gingivitis and periodontitis. The composition of oral microorganisms depends on multiple factors, including lifestyle (e.g., diet, oral care habits), health (e.g., oral diseases, host immune responses, and genetic susceptibility), and physical location in the oral cavity (tongue or tooth surfaces, as well as supragingival or subgingival sites) [4]. Periodontitis is an inflammatory disease involving a complex interaction between oral microorganisms organized in a biofilm structure and the host immune response. Clinically, periodontitis results in the destruction of tissues that support and protect the tooth and is a major cause of tooth loss in adults [5]. Moreover, periodontitis can also affect systemic health by increasing the risk of atherosclerosis, adverse

pregnancy outcomes, rheumatoid arthritis, aspiration pneumonia, and cancer [6–11].

In the past half century, numerous studies have characterized the community composition of the oral microbiota and described the association between periodontitis and pathogenic microorganisms. For example, *Aggregatibacter actinomycetemcomitans*, *Porphyromonas gingivalis*, *Tannerella forsythia*, *Treponema denticola*, *Fusobacterium nucleatum*, and *Prevotella intermedia* have traditionally been considered pathogenic bacteria contributing to periodontitis [5, 12, 13]. Socransky et al. [14] described the role of 5 main microbial complexes in the subgingival biofilm. They reported that red complex species *Porphyromonas gingivalis*, *Treponema denticola*, and *Tannerella forsythia* exhibited a very strong relationship with periodontitis. Subsequently, other association and elimination studies have confirmed the involvement of the three members of the red complex and some members of the orange complex, such as *Prevotella intermedia*, *Parvimonas micra*, *Fusobacterium nucleatum*, *Eubacterium nodatum*, and *Aggregatibacter actinomycetemcomitans*, in the etiology of different periodontal conditions [15]. Additionally, during the past decade, researchers using culture-independent molecular techniques have shown that some representatives of the genera *Megasphaera*, *Parvimonas*, *Desulfobulbus*, and *Filifactor* are more abundant in patients with periodontal diseases, whereas members of *Aggregatibacter*, *Prevotella*, *Selenomonas*, *Streptococcus*, *Actinomyces*, and *Rothia* are more abundant in healthy patients [16–19].

Machine learning is data method that involves finding patterns and making predictions from data based on multivariate statistics, data mining, and pattern recognition. This technology had been used to solved many metagenomic problems, such as operational taxonomic unit (out) clustering [20–24], binning [25–30], taxonomic profiling and assignment [31–35], comparative metagenomics [36–38], and gene prediction [39–42]. In addition to the learning algorithm and the model, the most important component of a learning system is how features are extracted from the domain data, a process known as feature selection. The purposes of feature selection include improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data [43–45]. Feature selection methodology can be categorized into three classes (filter, wrapper, and embedded methods) according to how the feature selection search is combined with the construction of the classification mode. Filter methods estimate the relevance of features by analysis of the intrinsic properties of the data. These methods are computationally simple and fast, can scale to very high-dimensional datasets easily, and are independent of the classification algorithm.

Although much is known about individual species associated with pathogenesis, the global structure of the bacterial community and the microbial signatures of periodontal disease are still poorly understood. In this study, we explored the microbial diversity in the subgingival plaque of healthy patients and patients with periodontal disease using culture-independent molecular methods based on 16S ribosomal DNA cloning. We also compared the bacterial community compositions between healthy patients and patients with periodontal disease and determined the core microbiomes present in these patients. Furthermore, we proposed a novel algorithm for feature selection, and microbes with significant differences were extracted as features and provided to generate feature combinations by applying our algorithm. Using machine learning methods, we built prediction models and found that the health status of patients with periodontal disease could be identified accurately using only a few features.

2. Materials and Methods

2.1. 16S rRNA Sequence Dataset. In total, 76 samples used for this study were collected from subgingival plaques of 76 unrelated individuals, including 10 patients with severe periodontal disease, 40 patients with moderate periodontal disease, and 26 healthy controls. This study was approved by the Institutional Review Board of Chang Gung Memorial Hospital, Taiwan (approval no. 102-4239B). All patients provided informed consent prior to their enrolment in the study. The oral health statuses of all individuals were determined by a dentist who performed a full-mouth clinical examination that included clinical parameters of periodontal pocket depths, gingival recession, clinical attachment loss, bleeding on probing, tooth mobility, and furcation involvement. These clinical parameters were measured at 6 sites per tooth (mesiobuccal, buccal, distobuccal, distolingual, lingual, and mesiolingual) at all teeth. Table 1 summarizes the parameters of periodontal pocket depths, bleeding on probing and clinical attachment loss for all of the samples. The classification of periodontitis as slight, moderate, or severe was based on the guidelines of the American Academy of Periodontology [46]. Subjects who had received previous periodontal therapy within two years and recent history of antibiotics taking within last 6 months were excluded.

After sampling, DNA extraction and polymerase chain reaction (PCR) were performed based on methods described by Tang et al. [47]. Following extraction, barcoded PCR amplification was performed with 382-bp amplicons flanking the highly variable V1-V2 region of the 16S rRNA gene sequence [48]. Next-generation sequencing evaluation of oral microbial communities was carried out using an *Illumina MiSeq* Desktop Sequencer after 30 cycles of PCR to enrich the adapter-modified DNA fragments.

2.2. Sequence Processing. Paired-end reads sequenced by the Illumina Sequencer were assembled with PEAR software [49]. Using split_libraries.py in QIIME with default parameters [50], assembled reads were demultiplexed, and low-quality reads were filtered. The GoldG database containing the ChimeraSlayer reference database in the Broad Microbiome Utilities [51] was used with UCHIME software [52] for chimera detection and removal. The remaining reads were clustered into OTUs using a de novo OTU selection protocol at the 97% identity level with a USEARCH algorithm [21]. Before clustering sequences, we filtered out all reads that occurred fewer than three times. This reduced the number of unique sequences to a computationally manageable level and

Table 1: Clinical characteristics of studied subjects. Clinical attachment loss and probing depth were measured in mm and represent the mean for all collected sites in the oral cavity of studied subjects.

Characteristics	Healthy	Moderate periodontitis	Severe periodontitis
Probing depth (mean ± s.d.)	1.3 ± 0.6	5.0 ± 1.3	7.9 ± 0.7
Clinical attachment loss (mean ± s.d.)	1.6 ± 0.7	5.7 ± 1.5	8.6 ± 1.1
% sites with bleeding on probing (mean ± s.d.)	2.8 ± 1.8	68.3 ± 23.2	79.7 ± 17.5

potentially reduced the number of errors from sequencing and contamination. The taxonomy associated with each OTU was assigned by blasting a representative sequence of each OTU against the Human Oral Microbiome Database [53] (HOMD). The sequence processing was carried out using our metagenomic analysis platforms [45].

2.3. Diversity and Significance Analysis. Sample data stored in the biological observation matrix format were subjected to statistical analysis using R language. We analyzed the sequencing depth of samples prior to downstream analysis using the Shannon index. The main microbes and taxonomic composition of the microbiota in each sample were also estimated. Abundance differences of microbes between sample groups were evaluated using the Kruskal–Wallis test. Four non-phylogeny-based metrics, namely, the observer species, chao 1 metric [54], Ace richness, and Shannon index, were used to evaluate alpha diversity, which represented the amount of diversity contained within communities, by applying the phyloseq R package. UniFrac is a distance metric used for comparing biological communities. Principal coordinate analysis (PCoA) with weighted UniFrac distances was applied to evaluate beta diversity, which represented the amount of diversity shared among communities. Principal component analysis (PCA) was used to characterize the primary microbes contained within communities.

2.4. Feature Selection and Machine Learning. In this study, we proposed a method of feature selection for selecting the informative microbes to predict whether an individual suffered from periodontal disease. First, the microbes present at less than 0.5% relative abundance in all samples were ignored, and nonparametric Kruskal–Wallis tests were used to detect microorganisms with significantly differential abundance between healthy patients and patients with periodontal disease. Microbes with more significant differential scores were considered features with more information. Then, the prioritized feature combination-generated algorithm shown in Algorithm 1 was adopted to produce the feature combinations composed by these more informative features.

In prioritized order, the feature combinations were applied to build classifiers with machine learning algorithms, such as deep learning, support vector machine (SVM), random forests, and logistic regression. We picked 80% of samples from both healthy and disease cases to train the prediction model, and the remaining cases were used for testing. The prediction ability of each feature combination was evaluated by calculating the average accuracy from 10 predictions with different training and testing sample sets. Here, we selected 10 of the most significant features having p values between 3.27E-11 and 7.77E-9. In total, 1,023 feature combinations were evaluated for their prediction ability using deep learning, SVM, random forest, and logistic regression methods. These machine learning algorithms were supported by the R packages H_2O, e1071, randomForest, and stats, respectively. We considered the radial basis function kernel for SVM. Parameters for each machine learning algorithm were tuned using grid search, and the parameters that obtained better accuracy were adopted for training prediction models.

3. Results and Discussion

3.1. Sample Sequencing and Identification. In total, 76 subgingival plaque samples from 76 unrelated individuals were divided into three classes according to their periodontal health status, i.e., healthy (H), severe periodontitis (SP), and moderate periodontitis (MP). Following DNA extraction and barcoded PCR amplification, these samples were sequenced, generating a total of 7,530,767 sequences. After filtering and trimming, 6,170,984 sequences remained, and there were 481 OTUs in all samples (481 and 429 in diseased and healthy samples, respectively). Due to variations in the number of sequences among samples, the total sequence reads within a sample was normalized to the relative abundance for subsequent analyses.

3.2. Taxonomic Composition of the Human Oral Microbiota. Table 2 summarizes the dominant microbes in the human oral microbial communities. In the experimental results, the microbial communities included 12 different phyla: *Bacteroidetes*, *Firmicutes*, *Fusobacteria*, *Proteobacteria*, *Spirochaetes*, *Actinobacteria*, *Candidate division TM7*, *Synergistetes*, *Fusobacteria*, *Candidate division SR1*, *Gracilibacteria*, and *Chloroflexi*. *Bacteroidetes* (37%) was the most abundant phylum in the human oral microbiota. The major genera consisted of previously characterized oral bacteria, including *Prevotella* (13.56%), *Fusobacterium* (11.30%), *Porphyromonas* (10.94%), *Treponema* (8.86%), *Streptococcus* (6.52%), *Leptotrichia* (4.76%), and *Capnocytophaga* (3.64%). In summary, there were 25 classes, 40 orders, 66 families, 124 genera, and 355 species at each taxonomic level.

In comparison of the compositions of microbial communities between healthy patients and patients with periodontitis, we found that the spectra of microbial communities differed. In healthy samples, the dominant genera were *Streptococcus* (13.09%), *Prevotella* (12.43%), *Fusobacterium* (11.70%), *Capnocytophaga* (6.25%), *Leptotrichia* (5.60%), *Alloprevotella* (4.26%), *Campylobacter* (3.94%), *Porphyromonas* (3.78%),

```
GenPFC((a_1, a_2, ... a_n))      C Generate prioritized feature combinations.
Input: (a_1, a_2, ... a_n) a list with n features in prioritized order.
Output: a queue Q used to store 2^n − 1 feature combinations.
1 Q ⟵ (∅)      C Enqueue empty set ∅ into queue Q
2 for i ⟵ 1 to n do      C Generate attribute combinations according to features in the list.
3     T ⟵ Q      C Copy Q into T which is a temporary queue.
4     for each s in T do
5         Enqueue(Q, s ∪ {a_i})
6 Dequeue(Q)      C Delete first empty set ∅ from queue Q
7 return Q
```

Algorithm 1: The prioritized feature combination-generated algorithm was used to generate all combinations of selected features in prioritized order. As an example, when n equals four, the generated list will be *(*1000, 0100, 1100, 0010, 1010, 0110, 1110, 0001, 1001, 0101, 1101, 0011, 1011, 0111, 1111*)*. Each element is a combination and denotes whether the four features were selected in that combination (e.g., the combination containing the first and third features is represented as 1010).

Table 2: Dominant microbes of the human oral microbiota at each taxonomic level.

Phylum	Class	Order	
Bacteroidetes	37.41% *Bacteroidia*	31.71% *Bacteroidales*	31.71%
Firmicutes	20.82% *Fusobacteria*	16.06% *Fusobacteriales*	16.06%
Fusobacteria	16.06% *Spirochaetia*	8.86% *Spirochaetales*	8.86%
Proteobacteria	9.30% *Bacilli*	7.83% *Lactobacillales*	7.06%
Spirochaetes	8.86% *Clostridia*	6.78% *Clostridiales*	6.78%
Actinobacteria	2.38% *Negativicutes*	5.21% *Selenomonadales*	5.21%
Family	Genus	Species	
Prevotellaceae	16.39% *Prevotella*	13.56% *Porphyromonas gingivalis*	7.30%
Porphyromonadaceae	12.96% *Fusobacterium*	11.30% *Fusobacterium nucleatum_subsp._vincentii*	5.23%
Fusobacteriaceae	11.30% *Porphyromonas*	10.94% *Prevotella intermedia*	4.62%
Spirochaetaceae	8.86% *Treponema*	8.86% *Streptococcus sp._oral_taxon_423*	2.62%
Streptococcaceae	6.52% *Streptococcus*	6.52% *Bacteroidales sp._oral_taxon_274*	2.18%
Veillonellaceae	5.21% *Leptotrichia*	4.76% *Prevotella loescheii*	2.15%

Veillonella (3.49%), and *Neisseria* (3.27%); however, in patients with periodontal disease, the dominant genera were *Porphyromonas* (14.67%), *Prevotella* (14.16%), *Treponema* (11.90%), *Fusobacterium* (11.09%), *Leptotrichia* (4.32%), and *Streptococcus* (3.10%). At the species level, *Streptococcus* sp. *oral taxon 423* (0.2-36%) was the most abundant species in healthy patients, whereas *Porphyromonas gingivalis* (0-31%) was the most abundant species in patients with periodontitis. Table 3 compares the dominant microbes between healthy patients and patients with periodontitis at each taxonomic level. The genus and species level taxonomic compositions between healthy patients and patients with periodontitis are shown in Figures 1 and 2. *Streptococcus* was more abundant in samples from all healthy individuals but decreased in samples from patients with periodontitis. Additionally, *Porphyromonas* and *Treponema* were more abundant in patients with periodontitis but decreased significantly in samples from healthy individuals. In total, 25 species were identified with significantly different abundances between sample groups; *Porphyromonas gingivalis* was the species with the most significantly differential abundance between samples from healthy patients and patients with periodontitis (p value = 2.41E-9).

Overall, our findings were largely comparable to those of previous studies [14, 55–61], indicating that species such as *Porphyromonas gingivalis*, *Treponema denticola*, *Tannerella forsythia*, *Filifactor alocis*, *Treponema socranskii*, *Aggregatibacter actinomycetemcomitans*, *Treponema vincentii*, and *Mycoplasma faucium* were significantly enriched in samples from patients with periodontitis. Furthermore, we found a set of species, including *Streptococcus sanguinis*, *Haemophilus parainfluenzae*, *Capnocytophaga granulosa*, *Gemella morbillorum*, *Campylobacter showae*, and *Granulicatella adiacens*, were significantly enriched in samples from healthy individuals.

Several studies have described the bacterial communities in patients with periodontitis and healthy control participants using metagenomics [16–19, 61–63]. The dominant microorganisms associated with periodontitis and the healthy state were largely consistent in those studies; however, we observed several discrepancies. First, in addition to common diseased-associated microorganisms, such as *Porphyromonas gingivalis*, *Treponema denticola*, *Tannerella forsythia*, *Filifactor alocis*, and *Aggregatibacter actinomycetemcomitans*, we also found that the species *Mycoplasma faucium* was significantly enriched in samples from patients with periodontal disease.

Table 3: Dominant microbes of the oral microbiota between healthy patients and patients with periodontitis at each taxonomic level.

	Healthy patients	Patients with periodontitis	
Phylum			
	Bacteroidetes	31.93% *Bacteroidetes*	40.26%
	Firmicutes	26.90% *Firmicutes*	17.66%
	Fusobacteria	17.31% *Fusobacteria*	15.42%
	Proteobacteria	11.81% *Spirochaetes*	11.90%
	Actinobacteria	3.36% *Proteobacteria*	7.99%
	Saccharibacteria	3.20% *Synergistetes*	2.50%
Class			
	Bacteroidia	24.76% *Bacteroidia*	35.32%
	Fusobacteria	17.31% *Fusobacteria*	15.42%
	Bacilli	15.23% *Spirochaetia*	11.90%
	Negativicutes	6.71% *Clostridia*	7.93%
	Flavobacteriia	6.67% *Negativicutes*	4.43%
	Clostridia	4.57% *Bacilli*	3.98%
Order			
	Bacteroidales	24.76% *Bacteroidales*	35.32%
	Fusobacteriales	17.31% *Fusobacteriales*	15.42%
	Lactobacillales	13.94% *Spirochaetales*	11.90%
	Selenomonadales	6.71% *Clostridiales*	7.93%
	Flavobacteriales	6.67% *Selenomonadales*	4.43%
	Clostridiales	4.57% *Lactobacillales*	3.48%
Family			
	Prevotellaceae	16.69% *Porphyromonadaceae*	17.19%
	Streptococcaceae	13.09% *Prevotellaceae*	16.24%
	Fusobacteriaceae	11.70% *Spirochaetaceae*	11.90%
	Veillonellaceae	6.71% *Fusobacteriaceae*	11.09%
	Flavobacteriaceae	6.67% *Veillonellaceae*	4.43%
	Leptotrichiaceae	5.61% *Leptotrichiaceae*	4.32%
Genus			
	Streptococcus	13.09% *Porphyromonas*	14.67%
	Prevotella	12.43% *Prevotella*	14.16%
	Fusobacterium	11.70% *Treponema*	11.90%
	Capnocytophaga	6.25% *Fusobacterium*	11.09%
	Leptotrichia	5.60% *Leptotrichia*	4.32%
	Alloprevotella	4.26% *Streptococcus*	3.10%
Species			
	Streptococcus sp._oral_taxon_423	5.88% *Porphyromonas gingivalis*	11.01%
	Fusobacterium nucleatum_subsp._vincentii	4.22% *Prevotella intermedia*	6.02%
	Fusobacterium nucleatum_subsp._polymorphum	3.52% *Fusobacterium nucleatum_subsp._vincentii*	5.76%
	Veillonella parvula	3.33% *Treponema denticola*	2.68%
	Bacteroidales sp._oral_taxon_274	3.11% *Fusobacterium nucleatum_subsp._nucleatum*	2.34%
	Fusobacterium nucleatum_subsp._animalis	3.09% *Tannerella forsythia*	2.32%

There were 26 samples that contained this species at greater than 0.5% abundance, and only one of these samples was derived from a healthy patient. The average relative abundance of *Mycoplasma faucium* was 0.59% in all samples (0.04% and 0.87% in samples from healthy patients and patients with periodontal disease, respectively) and was up to 4.85% in one diseased sample. Although this is a rare bacterium in the normal microbiota of the human oropharynx, some reports have identified this pathogen in brain abscesses [64, 65]. Additionally, Liu et al. [61] characterized the genomes of key players in the subgingival microbiota in patients with periodontitis, including an unculturable *TM7* organism. They also demonstrated that *TM7* organisms were significantly enriched in samples from patients with

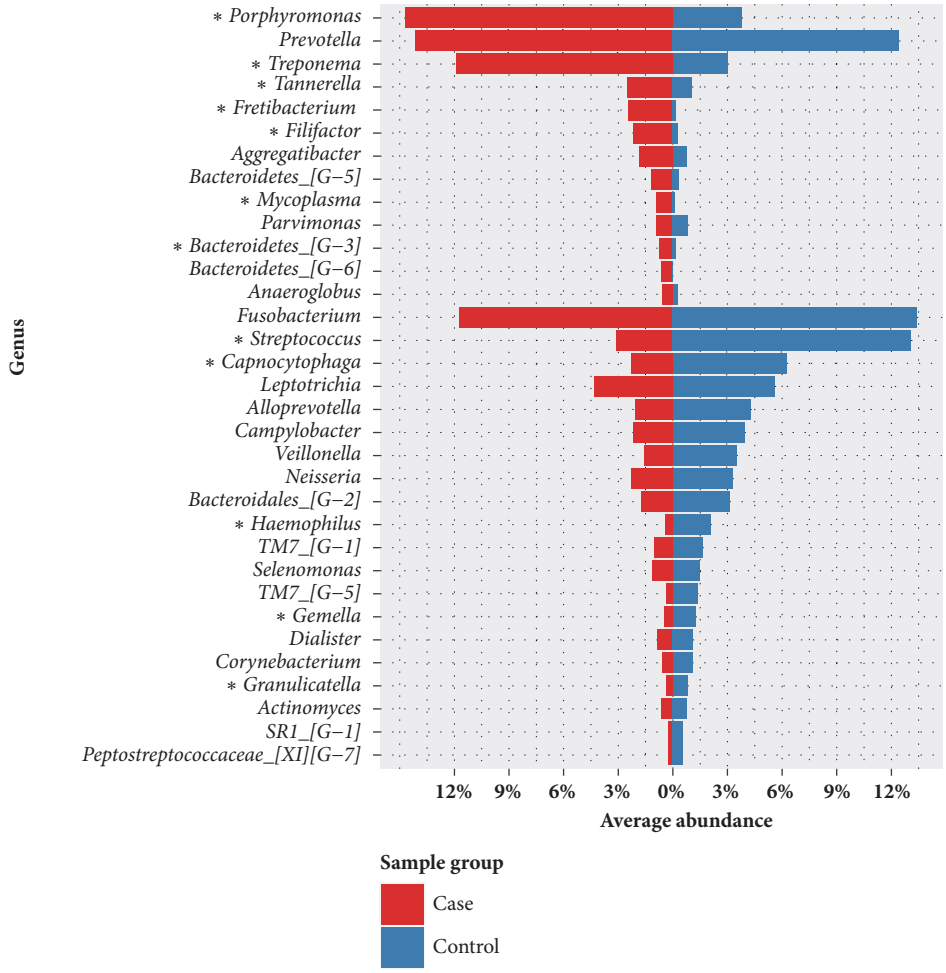

Figure 1: Microbial compositions of samples from healthy patients and patients with periodontitis at the genus level. The abundances were calculated by averaging the relative abundances in samples from healthy patients and patients with periodontitis. Only genera with > 0.5% abundance in at least one sample were included. Genera with significant differences in abundance between sample groups are indicated with asterisks (∗) (p value < 0.0001).

periodontitis. In our study, 49 of 76 samples contained *TM7* bacteria at greater than 1% abundance (average abundance of 2.1% in all samples). In samples from healthy patients and patients with periodontitis, the average abundances were 3.2% and 1.49%, respectively. However, significant enrichment was not observed in samples from patients with periodontitis. Furthermore, we found that the subspecies *Fusobacterium nucleatum* subsp. *polymorphum*, which is related to periodontal disease and is the member of the orange cluster described by Socransky et al. [14], is more abundant in healthy patients. In our results, the average abundances were 3.52% and 1.13% in samples from healthy patients and patients with periodontitis, respectively. This situation also can be observed in other three species, including *Campylobacter gracilis*, *Campylobacter rectus,* and *Campylobacter showae*. This discrepancy could be explained by geographic variability [66] or by differences in the depths of the pockets sampled [14], as well as the sample size and the DNA analytic bias [67]. Finally, Spearman's rank correlation coefficient was computed to assess association between each pair of species associated with periodontal disease. Figure 3 shows that a very strong relationship exhibited among species *Porphyromonas gingivalis*, *Treponema denticola*, and *Tannerella forsythia*.

In our study, there are 25 bacterial species with significantly different abundances between healthy patients and patients with periodontitis. The relationships of these species to pocket depth and clinical attachment loss were examined. Figure 4 shows that three species, *Porphyromonas gingivalis*, *Treponema denticola*, and *Tannerella forsythia,* exhibited a very strong relationship with pocket depth and clinical attachment loss. For instance, the three species increased in abundance with increasing pocket depth and clinical attachment loss. The abundances of those species among different level of pocket depth and clinical attachment loss were different significantly. However, it should be noted that not only oral microorganisms but also others factors, such as supragingival plaque, would affect the pocket depth and clinical attachment loss [68].

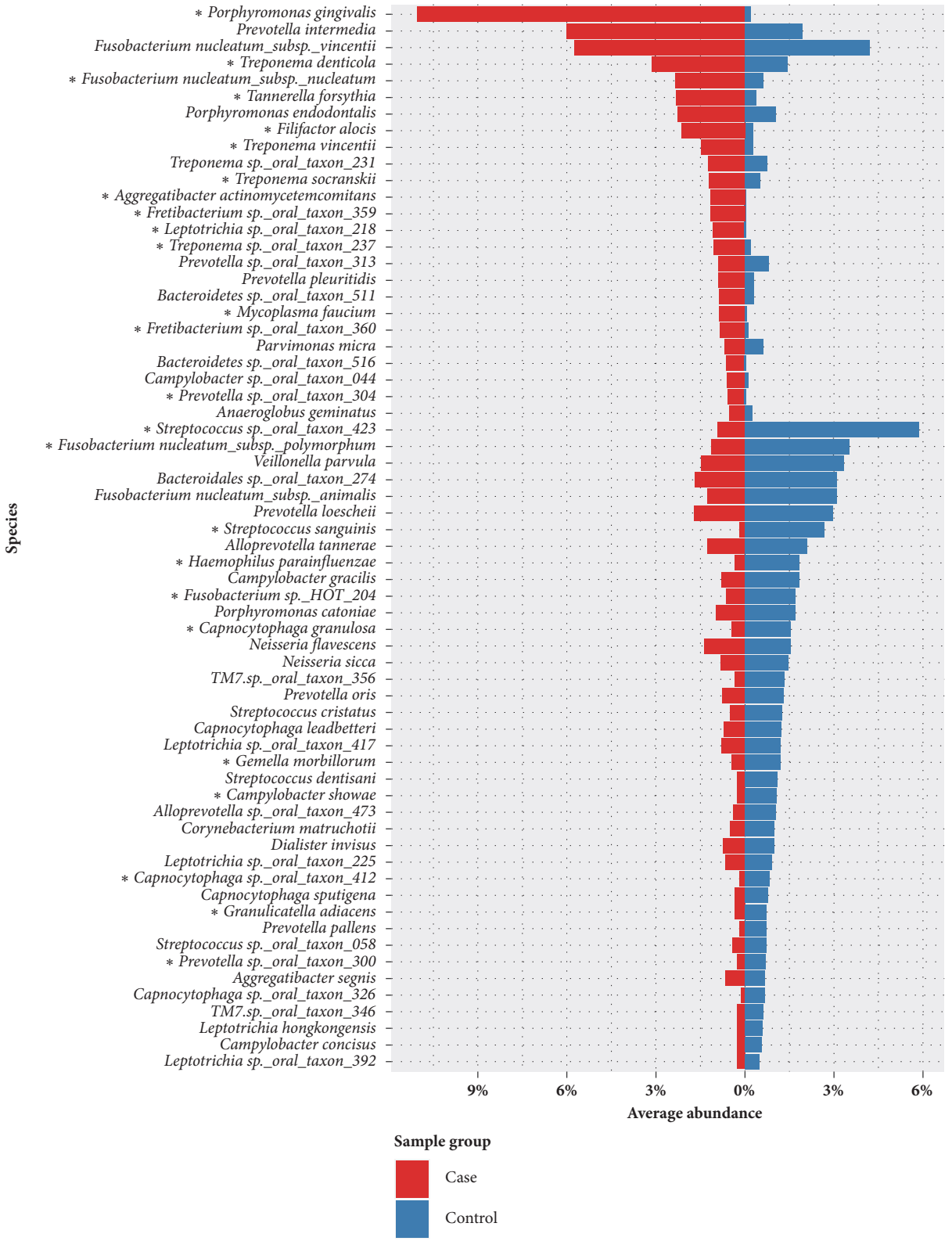

Figure 2: Microbial compositions of samples from healthy patients and patients with periodontitis at the species level. The abundances were calculated by averaging the relative abundances in samples from healthy patients and patients with periodontitis. Only species with > 0.5% abundance in at least one sample are shown. Species with significant differences in abundance between sample groups are indicated with asterisks (∗) (p value < 0.0001).

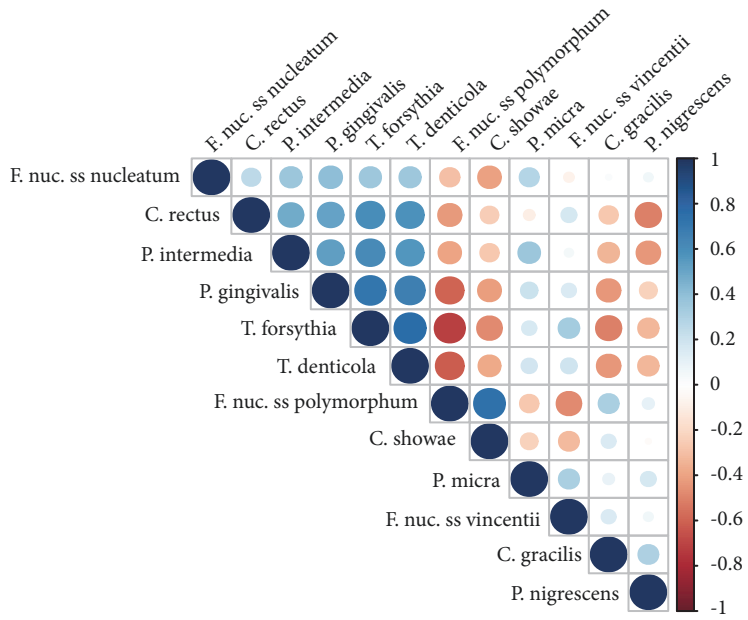

Figure 3: The relationships among species were evaluated using Spearman's rank correlation coefficient.

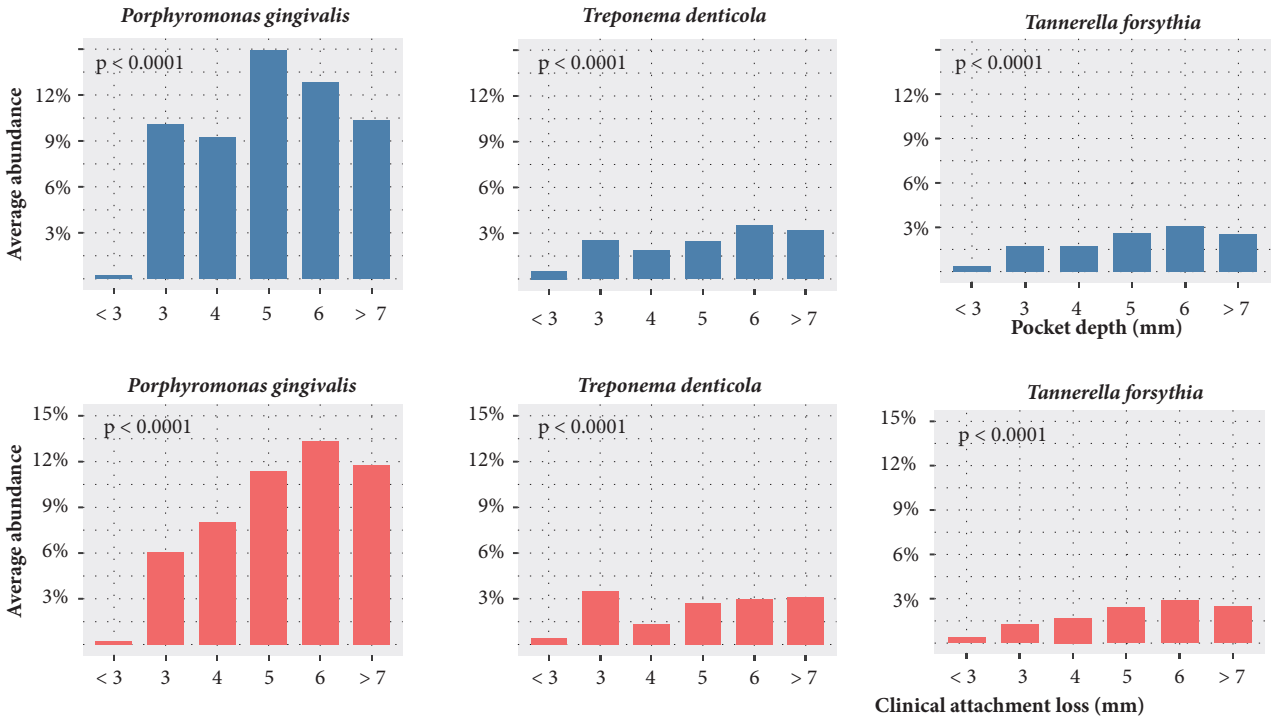

Figure 4: Relationships of the average abundance of three species to selected pocket depths and clinical attachment loss levels. Significance of differences among pocket depth levels was tested using the Kruskal-Wallis test.

3.3. Diversity of Bacterial Community Profiles. To evaluate the alpha diversity of the microbial communities, Shannon index curves scores and richness metrics (Observed, Chao1, and ACE) were applied, as shown in Figure 5. As depicted in Figure 5(a), the Shannon diversity index curves clearly reached plateau levels after the sequence number exceeded 5,000 in all three health statuses, indicating that the microbial composition for each health status was well represented by the sequencing depth. As shown in Figure 5(b), the average richness measured by Observed, Chao1, and Ace indexes was higher in samples from patients with periodontitis than in samples from healthy individuals; however, these results were in contrast to the results from the Shannon diversity index. Thus, the relative abundance of each microbe was

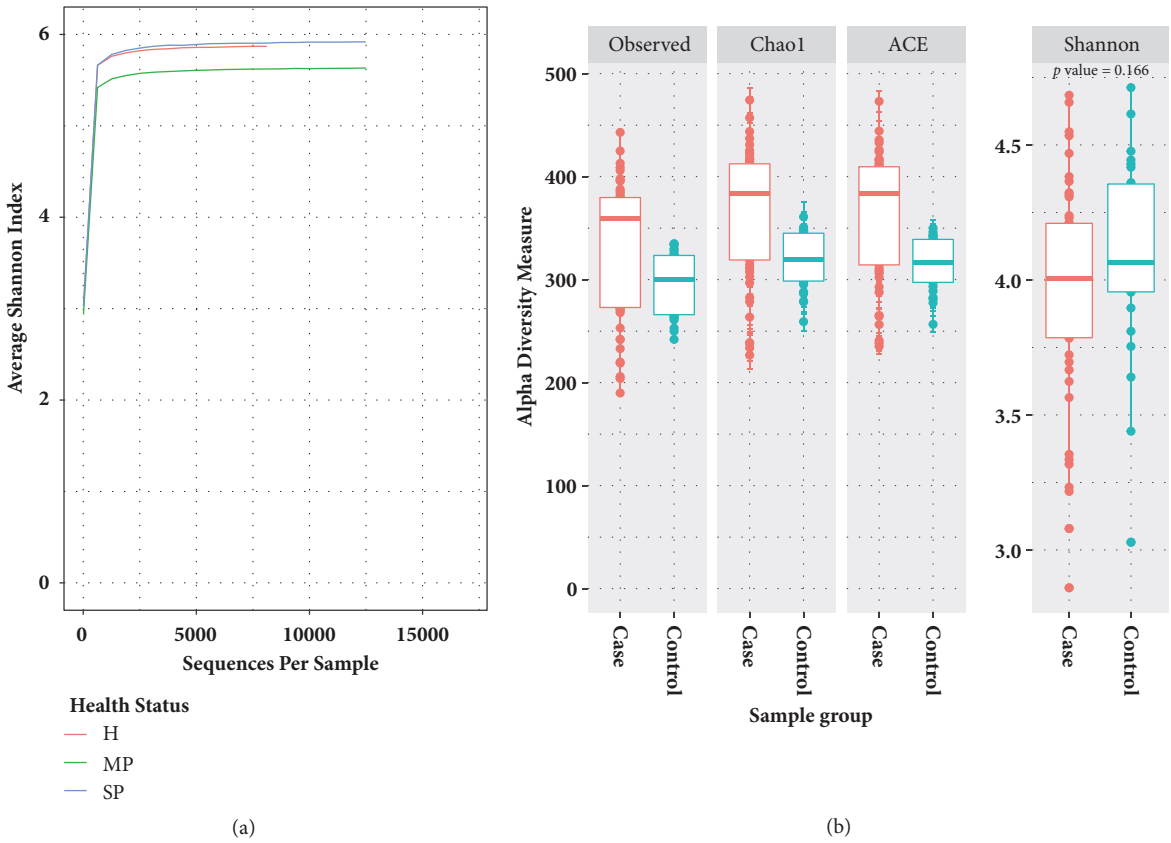

FIGURE 5: (a) The sequencing depths measured by average scores from the Shannon index reached a plateau when the sequence number exceeded 5,000. (b) Alpha-diversity metrics (richness and Shannon index) were employed to measure the microbial communities of samples from healthy patients and patients with periodontitis. The average richness of microbes was higher in patients with periodontal disease than in healthy patients; however, the microbial communities of healthy patients exhibited higher Shannon indexes.

more balanced in samples from healthy individuals than in samples from patients with periodontal disease, and there were more microbes with low relative abundance in samples from patients with periodontitis.

To further explore the relationships between bacterial communities in healthy patients and patients with periodontal disease, PCoA was performed (Figure 6(a)). Analysis of beta diversity based on the weighted UniFrac distances showed greater concentration in diseased samples than in healthy samples. In other words, the microbial compositions of diseased samples were more similar to each other. As shown in Figure 6(b), PCA of microbial communities revealed that the core genera in healthy samples included *Streptococcus*, *Capnocytophaga*, *Campylobacter*, *Veillonella*, *Alloprevotella*, *TM7_[G-1]*, *Leptotrichia*, and *Selenomonas*, whereas those in samples from patients with periodontitis were *Filifactor*, *Treponema*, *Fretibacterium*, *Porphyromonas*, and *Tannerella*.

3.4. Machine Learning and Feature Selection. Before applying the machine learning algorithm to classify samples, it is necessary to select the features from the samples and train prediction models. Table 4 lists features with difference scores $p < 1.E\text{-}07$. Based on significant differences between healthy patients and patients with periodontitis, we selected the top 10 microbes with more information as features. In total, 1,023 combinations of selected features were generated by our algorithm. All feature combinations were evaluated by SVM, random forest, logical regression, and deep learning machine learning methods, and the average accuracies were 0.88, 0.93, 0.85, and 0.90, respectively. Figure 7 shows the performance of each machine learning method. In general, the accuracy of prediction increased slightly with the number of features used, except in logistic regression. From our results, we found that random forests had better predictive ability than the other methods. Applying combinations consisting of *Peptoniphilaceae* sp. *oral taxon 113*, *Streptococcus sanguinis*, *Mollicutes* sp. *oral taxon 906*, *Aggregatibacter actinomycetemcomitans*, *Porphyromonas gingivalis*, *Peptostreptococcaceae* sp. *oral taxon 950*, and *Lachnospiraceae* sp. *oral taxon 500* or *Stomatobaculum* sp. *oral taxon 373*, *Desulfobulbus* sp. *oral taxon 041*, *Peptoniphilaceae* sp. *oral taxon 113*, *Streptococcus sanguinis*, *Aggregatibacter actinomycetemcomitans*, *Porphyromonas gingivalis*, and *Leptotrichia* sp. *oral taxon 218* showed that random forests could predict the health status of samples accurately. The feature combinations having average accuracies of more than 0.94 are reported in Table 5.

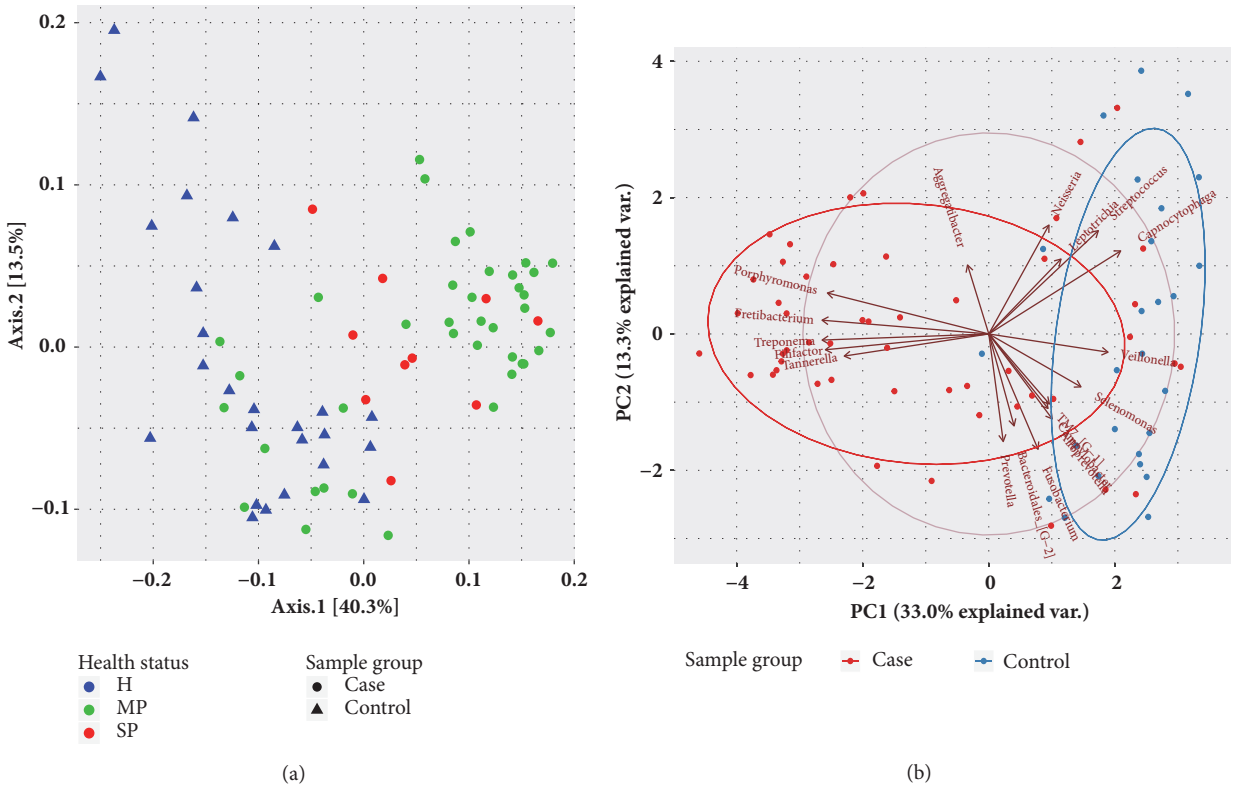

Figure 6: (a) Principal coordinate analysis (PCoA) with weighted UniFrac distance matrixes for bacterial communities associated with the three health statuses. (b) Principal component analysis (PCA) of the dominant genera between samples from healthy patients and patients with periodontitis. Only genera with ≥ 1% mean relative abundance across all samples are shown.

Table 4: Features with significant differences between healthy patients and patients with periodontitis. Correlation coefficients and p values were determined by Spearman's rank correlation coefficient and Kruskal–Wallis tests, respectively. Negative correlations indicated that the features were observed more often in patients with periodontitis than in healthy patients.

No	Feature (Species)	Correlation coefficient	p
1	*Stomatobaculum sp._oral_taxon_373*	-0.766029754	3.27E-11
2	*Desulfobulbus sp._oral_taxon_041*	-0.74877058	8.90E-11
3	*Peptoniphilaceae sp._oral_taxon_113*	-0.723418056	3.73E-10
4	*Streptococcus sanguinis*	0.71684624	5.36E-10
5	*Mollicutes sp._oral_taxon_906*	-0.709369416	8.08E-10
6	*Aggregatibacter actinomycetemcomitans*	-0.686608198	2.74E-09
7	*Porphyromonas gingivalis*	-0.683993685	3.15E-09
8	*Peptostreptococcaceae sp._oral_taxon_950*	-0.681489164	3.59E-09
9	*Lachnospiraceae sp._oral_taxon_500*	-0.670324546	6.43E-09
10	*Leptotrichia sp._oral_taxon_218*	-0.666642231	7.77E-09
11	*Bosea vestrisii*	0.665468802	8.26E-09
12	*Filifactor alocis*	-0.656797473	1.29E-08
13	*Mycoplasma faucium*	-0.641322841	2.79E-08
14	*Prevotella sp._oral_taxon_304*	-0.638587976	3.20E-08
15	*Fretibacterium sp._oral_taxon_359*	-0.632290825	4.36E-08
16	*Bergeyella sp._oral_taxon_322*	0.630961524	4.65E-08
17	*Tannerella forsythia*	-0.628346704	5.28E-08
18	*Peptostreptococcus indolicus*	-0.626504998	5.77E-08
19	*Johnsonella sp._oral_taxon_166*	-0.622396393	7.04E-08
20	*Peptostreptococcaceae [Eubacterium]_saphenum*	-0.616735679	9.24E-08

Table 5: Feature combinations and their predictive accuracies with different machine learning methods. Only feature combinations with more than 0.94 average accuracy are shown. DL, RF, and LR represent deep learning, random forests, and logistic regression, respectively.

Feature combination	DL	RF	SVM	LR	Average accuracy
Stomatobaculum sp._oral_taxon_373 *Peptoniphilaceae sp._oral_taxon_113*	0.967	0.973	0.960	0.933	0.958
Desulfobulbaceae sp._oral_taxon_041 *Peptoniphilaceae sp._oral_taxon_113* *Aggregatibacter actinomycetemcomitans* *Lachnospiraceae sp._oral_taxon_500* *Leptotrichia sp._oral_taxon_218*	0.933	0.960	0.973	0.947	0.953
Stomatobaculum sp._oral_taxon_373 *Streptococcus sanguinis* *Aggregatibacter actinomycetemcomitans*	0.933	0.973	0.960	0.947	0.953
Desulfobulbaceae sp._oral_taxon_041 *Mollicutes sp._oral_taxon_906* *Porphyromonas gingivalis* *Aggregatibacter actinomycetemcomitans* *Peptostreptococcaceae sp._oral_taxon_950*	0.973	0.967	0.933	0.927	0.950
Stomatobaculum sp._oral_taxon_373 *Streptococcus sanguinis* *Mollicutes sp._oral_taxon_906* *Porphyromonas gingivalis* *Aggregatibacter actinomycetemcomitans*	0.947	0.953	0.907	0.987	0.948
Stomatobaculum sp._oral_taxon_373 *Peptoniphilaceae sp._oral_taxon_113* *Aggregatibacter actinomycetemcomitans* *Leptotrichia sp._oral_taxon_218*	0.960	0.967	0.947	0.913	0.947
Desulfobulbaceae sp._oral_taxon_041 *Peptoniphilaceae sp._oral_taxon_113* *Aggregatibacter actinomycetemcomitans* *Leptotrichia sp._oral_taxon_218*	0.933	0.973	0.933	0.947	0.947
Stomatobaculum sp._oral_taxon_373 *Peptoniphilaceae sp._oral_taxon_113* *Mollicutes sp._oral_taxon_906*	0.967	0.933	0.953	0.933	0.947
Peptoniphilaceae sp._oral_taxon_113 *Streptococcus sanguinis* *Aggregatibacter actinomycetemcomitans*	0.960	0.987	0.867	0.967	0.945
Stomatobaculum sp._oral_taxon_373 *Aggregatibacter actinomycetemcomitans* *Peptostreptococcaceae sp._oral_taxon_950*	0.920	0.947	0.967	0.947	0.945
Stomatobaculum sp._oral_taxon_373 *Peptoniphilaceae sp._oral_taxon_113* *Porphyromonas gingivalis* *Aggregatibacter actinomycetemcomitans*	0.967	0.967	0.953	0.893	0.945

According to previous studies, Caruana et al. [69, 70] proposed that the random forest method showed better accuracy in high-dimensional and large-scale data than neural nets, SVM, and logistic regression. In this study, we found that the random forest method was more suitable for small-scale data than other methods. In contrast, deep learning approaches led to good performance, but required long computation times and large amounts of memory, particularly when the hidden layer size was increased.

4. Conclusions

With the development of high-throughput DNA sequencing technology, the limitations associated with difficult culture of

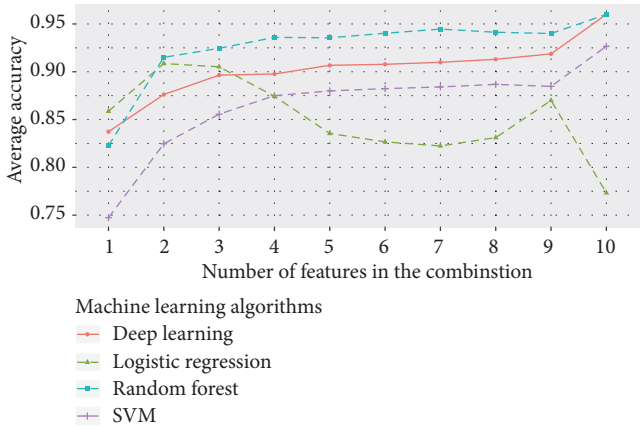

Figure 7: Average accuracies of different numbers of features.

many microbes that populate the oral cavity can be overcome, facilitating the analysis of bacterial community composition. Using 16S rRNA sequencing of subgingival samples from 50 individuals with periodontitis and 26 periodontally healthy controls, we determined the diversity of and differences in community compositions. Moreover, we identified microbes associated with good health and periodontal disease and provided a machine learning method for finding patterns and making predictions for oral microbiota associated with periodontal disease.

Our results showed that there was a higher diversity of microbes in samples from patients with periodontal disease than in samples from healthy patients. Importantly, the core microbes in healthy patients were different significantly from those in patients with periodontitis. We also found that bacterial communities associated with healthy and diseased states were highly different in PCA and PCoA, and the compositions of microorganisms were more similar to each other in samples from patients with periodontal disease than in samples from healthy individuals.

We proposed a novel feature selection method and investigated the potential of machine learning approaches for determination of health status based on oral metagenomics data. By using nonparametric Kruskal–Wallis tests to assess the significance of each microorganism, we selected significant microbes to generate prioritized feature combinations by our algorithm. The performances of four machine learning approaches were evaluated with these feature combinations, and random forests showed the best performance (average accuracy of 0.93 from 1,023 feature combinations), followed by deep learning, SVM, and logistic regression. Using machine learning methods, training models could accurately predict the health status of samples by examining fewer features. According to our observations, the accuracy of prediction generally increased slightly with the number of features used, except for logistic regression. Notably, certain combinations composed of fewer features showed better accuracy than combinations composed of all selected features. These combinations of features may only apply to our dataset. However, the results implied that a few related features may have better predictive ability than multiple independent features. Therefore, in order to improve the prediction accuracy of the model, it is essential to identify the most informative features. Due to limitations in funding, time, and ethical considerations, it is not easy to obtain large numbers of oral samples from patients with periodontitis. Although insufficient and incomplete samples could easily lead to bias and variance in training models, our study still provided an important basis for further studies.

Periodontitis is a chronic inflammatory disease involving complex interactions between the oral microorganisms and the host immune response. In addition to the individual species associated with pathogenesis, the system-level mechanisms underlying the transition from a healthy state to a diseased state are key points for studying periodontal disease. Thus, in our future studies, we aim to elucidate the global genetic, metabolic, and ecological changes associated with periodontitis and identify the pathogenic features of constructing machine learning models. Rapid molecular techniques and machine learning methods capable of identifying periodontal bacteria with great accuracy may eventually provide improved classification and diagnosis of various types of periodontal diseases and aid significantly in clinical decision-making.

Data Availability

The raw sequences of human oral subgingival plaque samples were deposited at the NCBI Sequence Read Archive under the Bioproject Accession no. PRJNA437129.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Wen-Pei Chen and Shih-Hao Chang contributed equally to this work.

Funding

The present work was partially supported by a grant from the Ministry of Science and Technology [grant number MOST 107-2218-E-126-001-] and [grant number NSC 102-2622-E-126-002 CC1].

References

[1] L. Gao, T. Xu, G. Huang, S. Jiang, Y. Gu, and F. Chen, "Oral microbiomes: more and more importance in oral cavity and whole body," *Protein & Cell*, pp. 1–13, 2018.

[2] P. D. Marsh, "Microbiology of dental plaque biofilms and their role in oral health and caries," *Dental Clinics of North America*, vol. 54, no. 3, pp. 441–454, 2010.

[3] P. D. Marsh, "Dental plaque as a biofilm and a microbial community - Implications for health and disease," *BMC Oral Health*, vol. 6, no. 1, 2006.

[4] R. J. Palmer Jr., "Composition and development of oral bacterial communities," *Periodontology 2000*, vol. 64, no. 1, pp. 20–39, 2014.

[5] B. L. Pihlstrom, B. S. Michalowicz, and N. W. Johnson, "Periodontal diseases," *The Lancet*, vol. 366, no. 9499, pp. 1809–1820, 2005.

[6] R. J. Genco and T. E. Van Dyke, "Reducing the risk of CVD in patients with periodontitis," *Nature Reviews Cardiology*, vol. 7, no. 9, pp. 479-480, 2010.

[7] K. Lundberg, N. Wegner, T. Yucel-Lindberg, and P. J. Venables, "Periodontitis in RA-the citrullinated enolase connection," *Nature Reviews Rheumatology*, vol. 6, no. 12, pp. 727–730, 2010.

[8] M. Kebschull, R. T. Demmer, and P. N. Papapanou, ""Gum bug, leave my heart alone!"—epidemiologic and mechanistic evidence linking periodontal infections and atherosclerosis," *Journal of Dental Research*, vol. 89, no. 9, pp. 879–902, 2010.

[9] S. E. Whitmore, R. J. Lamont, and W. E. Goldman, "Oral Bacteria and Cancer," *PLoS Pathogens*, vol. 10, no. 3, p. e1003933, 2014.

[10] Y. W. Han and X. Wang, "Mobile microbiome: oral bacteria in extra-oral infections and inflammation," *Journal of Dental Research*, vol. 92, no. 6, pp. 485–491, 2013.

[11] P. N. Madianos, Y. A. Bobetsis, and S. Offenbacher, "Adverse pregnancy outcomes (APOs) and periodontal disease: Pathogenic mechanisms," *Journal of Clinical Periodontology*, vol. 40, no. 14, pp. S170–S180, 2013.

[12] S. Témoin, K. L. Wu, V. Wu, M. Shoham, and Y. W. Han, "Signal peptide of FadA adhesin from Fusobacterium nucleatum plays a novel structural role by modulating the filament's length and width," *FEBS Letters*, vol. 586, no. 1, pp. 1–6, 2012.

[13] S. Malm, M. Jusko, S. Eick, J. Potempa, K. Riesbeck, and A. M. Blom, "Acquisition of complement inhibitor serine protease factor i and its cofactors C4b-binding protein and factor H by prevotella intermedia," *PLoS ONE*, vol. 7, no. 4, 2012.

[14] S. S. Socransky, A. D. Haffajee, M. A. Cugini, C. Smith, and R. L. Kent Jr., "Microbial complexes in subgingival plaque," *Journal of Clinical Periodontology*, vol. 25, no. 2, pp. 134–144, 1998.

[15] R. Teles, F. Teles, J. Frias-Lopez, B. Paster, and A. Haffajee, "Lessons learned and unlearned in periodontal microbiology," *Periodontology 2000*, vol. 62, no. 1, pp. 95–162, 2013.

[16] P. Belda-Ferre, L. D. Alcaraz, R. Cabrera-Rubio et al., "The oral metagenome in health and disease," *The ISME Journal*, vol. 6, no. 1, pp. 46–56, 2012.

[17] P. S. Kumar, A. L. Griffen, M. L. Moeschberger, and E. J. Leys, "Identification of candidate periodontal pathogens and beneficial species by quantitative 16S clonal analysis," *Journal of Clinical Microbiology*, vol. 43, no. 8, pp. 3944–3955, 2005.

[18] A. P. V. Colombo, S. K. Boches, S. L. Cotton et al., "Comparisons of subgingival microbial profiles of refractory periodontitis, severe periodontitis, and periodontal health using the human oral microbe identification microarray," *Journal of Periodontology*, vol. 80, no. 9, pp. 1421–1432, 2009.

[19] S. Jünemann, K. Prior, R. Szczepanowski et al., "Bacterial community shift in treated periodontitis patients revealed by Ion Torrent 16S rRNA gene amplicon sequencing," *PLoS ONE*, vol. 7, no. 8, Article ID e41606, 2012.

[20] P. D. Schloss, S. L. Westcott, T. Ryabin et al., "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and Environmental Microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[21] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460-2461, 2010.

[22] F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn, "Swarm: Robust and fast clustering method for amplicon-based studies," *PeerJ*, vol. 2014, no. 1, 2014.

[23] M. Ghodsi, B. Liu, and M. Pop, "DNACLUST: Accurate and efficient clustering of phylogenetic marker genes," *BMC Bioinformatics*, vol. 12, article no. 271, 2011.

[24] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[25] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner, "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," *BMC Bioinformatics*, vol. 5, article 163, 2004.

[26] Y. Wang, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "MetaCluster 4.0: A novel binning algorithm for NGS reads and huge number of species," *Journal of Computational Biology*, vol. 19, no. 2, pp. 241–249, 2012.

[27] Y. Wang, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "Metacluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample," *Bioinformatics*, vol. 28, no. 18, Article ID bts397, pp. i356–i362, 2012.

[28] D. D. Kang, J. Froula, R. Egan, and Z. Wang, "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities," *PeerJ*, vol. 2015, no. 8, 2015.

[29] J. Alneberg, B. S. Bjarnason, I. De Bruijn et al., "Binning metagenomic contigs by coverage and composition," *Nature Methods*, vol. 11, no. 11, pp. 1144–1146, 2014.

[30] Y.-W. Wu, Y.-H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer, "MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm," *Microbiome*, vol. 2, no. 1, 2014.

[31] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Applied and Environmental Microbiology*, vol. 73, no. 16, pp. 5261–5267, 2007.

[32] N. Chaudhary, A. K. Sharma, P. Agarwal, A. Gupta, and V. K. Sharma, "16S classifier: A tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets," *PLoS ONE*, vol. 10, no. 2, 2015.

[33] S. P. Szafranski, M. L. Wos-Oxley, R. Vilchez-Vargas et al., "High-resolution taxonomic profiling of the subgingival microbiome for biomarker discovery and periodontitis diagnosis," *Applied and Environmental Microbiology*, vol. 81, no. 3, pp. 1047–1058, 2015.

[34] K. Vervier, P. Mahé, M. Tournoud, J.-B. Veyrieras, and J.-P. Vert, "Large-scale machine learning for metagenomics sequence classification," *Bioinformatics*, vol. 32, no. 7, pp. 1023–1032, 2016.

[35] A. E. Darling, G. Jospin, E. Lowe, F. A. Matsen, H. M. Bik, and J. A. Eisen, "PhyloSift: Phylogenetic analysis of genomes and metagenomes," *PeerJ*, vol. 2013, no. 1, 2014.

[36] Z. Liu, W. Hsiao, B. L. Cantarel, E. F. Drábek, and C. Fraser-Liggett, "Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data," *Bioinformatics*, vol. 27, no. 23, pp. 3242–3249, 2011.

[37] O. Tanaseichuk, J. Borneman, and T. Jiang, "Phylogeny-based classification of microbial communities," *Bioinformatics*, vol. 30, no. 4, pp. 449–456, 2014.

[38] H. Cui and X. Zhang, "Alignment-free supervised classification of metagenomes by recursive SVM," *BMC Genomics*, vol. 14, no. 1, article no. 641, 2013.

[39] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights," *PLoS Computational Biology*, vol. 12, no. 7, 2016.

[40] M. Arumugam, E. D. Harrington, K. U. Foerstner, J. Raes, and P. Bork, "SmashCommunity: A metagenomic annotation and analysis tool," *Bioinformatics*, vol. 26, no. 23, pp. 2977-2978, 2010.

[41] K. J. Hoff, T. Lingner, P. Meinicke, and M. Tech, "Orphelia: Predicting genes in metagenomic sequencing reads," *Nucleic Acids Research*, vol. 37, no. 2, pp. W101–W105, 2009.

[42] K. J. Hoff, M. Tech, T. Lingner, R. Daniel, B. Morgenstern, and P. Meinicke, "Gene prediction in metagenomic fragments: A large scale machine learning approach," *BMC Bioinformatics*, vol. 9, article no. 217, 2008.

[43] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[44] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[45] W.-P. Chen, S.-J. J. Tsai, Y.-C. Hu, and Y.-L. Lin, "Metagenomic analysis and features selection in human oral microbiota associated with periodontal disease," in *Proceedings of the 33rd Workshop on Combinatorial Mathematics and Computation Theory*, pp. 58–64, 2016.

[46] G. C. Armitage, "Development of a classification system for periodontal diseases and conditions.," *Annals of Periodontology*, vol. 4, no. 1, pp. 1–6, 1999.

[47] C. Y. Tang, S.-M. Yiu, H.-Y. Kuo et al., "Application of 16S rRNA metagenomics to analyze bacterial communities at a respiratory care centre in Taiwan," *Applied Microbiology and Biotechnology*, vol. 99, no. 6, pp. 2871–2881, 2015.

[48] L. Cai, L. Ye, A. H. Y. Tong, S. Lok, and T. Zhang, "Biased Diversity Metrics Revealed by Bacterial 16S Pyrotags Derived from Different Primer Sets," *PLoS ONE*, vol. 8, no. 1, 2013.

[49] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis, "PEAR: A fast and accurate Illumina Paired-End reAd mergeR," *Bioinformatics*, vol. 30, no. 5, pp. 614–620, 2014.

[50] J. G. Caporaso, J. Kuczynski, J. Stombaugh et al., "QIIME allows analysis of high-throughput community sequencing data," *Nature Methods*, vol. 7, no. 5, pp. 335-336, 2010.

[51] B. J. Haas, D. Gevers, A. M. Earl et al., "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons," *Genome Research*, vol. 21, no. 3, pp. 494–504, 2011.

[52] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight, "UCHIME improves sensitivity and speed of chimera detection," *Bioinformatics*, vol. 27, no. 16, pp. 2194–2200, 2011.

[53] T. Chen, W.-H. Yu, J. Izard, O. V. Baranova, A. Lakshmanan, and F. E. Dewhirst, "The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information," *Database : the journal of biological databases and curation*, vol. 2010, p. baq013, 2010.

[54] T. C. J. Hill, K. A. Walsh, J. A. Harris, and B. F. Moffett, "Using ecological diversity measures with bacterial communities," *FEMS Microbiology Ecology*, vol. 43, no. 1, pp. 1–11, 2003.

[55] R. P. Darveau, "Periodontitis: a polymicrobial disruption of host homeostasis," *Nature Reviews Microbiology*, vol. 8, no. 7, pp. 481–490, 2010.

[56] J. M. Albandar, L. J. Brown, and H. Löe, "Putative Periodontal Pathogens in Subgingival Plaque of Young Adults with and Without Early-Onset Periodontitis," *Journal of Periodontology*, vol. 68, no. 10, pp. 973–981, 1997.

[57] Y. Takeuchi, M. Umeda, M. Sakamoto, Y. Benno, Y. Huang, and I. Ishikawa, "Treponema socranskii, Treponema denticola, and Porphyromonas gingivalis are associated with severity of periodontal tissue destruction," *Journal of Periodontology*, vol. 72, no. 10, pp. 1354–1363, 2001.

[58] J. J. Zambon, "Actinobacillus actinomycetemcomitans in human periodontal disease," *Journal of Clinical Periodontology*, vol. 12, no. 1, pp. 1–20, 1985.

[59] J. Slots and M. Ting, "Actinobacillus actinomycetemcomitans and Porphyromonas gingivalis in human periodontal disease: Occurrence and treatment," *Periodontology 2000*, vol. 20, no. 1, pp. 82–121, 1999.

[60] B. K. Choi, B. J. Paster, F. E. Dewhirst, and U. B. Gobel, "Diversity of cultivable and uncultivable oral spirochetes from a patient with severe destructive periodontitis," *Infection and Immunity*, vol. 62, no. 5, pp. 1889–1895, 1994.

[61] B. Liu, L. L. Faller, N. Klitgord et al., "Deep sequencing of the oral microbiome reveals signatures of periodontal disease," *PLoS ONE*, vol. 7, no. 6, Article ID e37919, 2012.

[62] J. Wang, J. Qi, H. Zhao et al., "Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease," *Scientific Reports*, vol. 3, 2013.

[63] A. L. Griffen, C. J. Beall, J. H. Campbell et al., "Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing," *The ISME Journal*, vol. 6, no. 6, pp. 1176–1185, 2012.

[64] M. A. Masalma, F. Armougom, W. Michael Scheld et al., "The expansion of the microbiological spectrum of brain abscesses With use of multiple 16S ribosomal DNA Sequencing," *Clinical Infectious Diseases*, vol. 48, no. 9, pp. 1169–1178, 2009.

[65] M. Al Masalma, M. Lonjon, H. Richet et al., "Metagenomic analysis of brain abscesses identifies specific bacterial associations," *Clinical Infectious Diseases*, vol. 54, no. 2, pp. 202–210, 2012.

[66] I. Nasidze, J. Li, D. Quinque, K. Tang, and M. Stoneking, "Global diversity in the human salivary microbiome," *Genome Research*, vol. 19, no. 4, pp. 636–643, 2009.

[67] F. V. Wintzingerode, U. B. Göbel, and E. Stackebrandt, "Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis," *FEMS Microbiology Reviews*, vol. 21, no. 3, pp. 213–229, 1997.

[68] M. Tezal, F. A. Scannapieco, J. Wactawski-Wende, S. G. Grossi, and R. J. Genco, "Supragingival plaque may modify the effects of subgingival bacteria on attachment loss," *Journal of Periodontology*, vol. 77, no. 5, pp. 808–813, 2006.

[69] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the ICML 2006: 23rd International Conference on Machine Learning*, pp. 161–168, USA, June 2006.

[70] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 96–103, Finland, July 2008.