BMB Reports - Manuscript Submission

Manuscript Draft

Manuscript Number: BMB-16-164

Title: Probing the diversity of healthy oral microbiome with bioinformatics approaches

Article Type: Mini Review

Keywords: Oral microbiome; Diversity; Bioinformatics; Next-generation Sequencing (NGS); Human Microbiome Project (HMP)

Corresponding Author: Jae-Hyung Lee

Authors: Ji-Hoi Moon[1], Jae-Hyung Lee[1,*]

Institution: [1]Department of Maxillofacial Biomedical Engineering, School of Dentistry, and Department of Life and Nanopharmaceutical Sciences, Kyung Hee University, Seoul 02447, Republic of Korea,

Manuscript Type: Mini Review

Title: Probing the diversity of healthy oral microbiome with bioinformatics approaches

Author's name: Ji-Hoi Moon* & Jae-Hyung Lee*

*Corresponding authors.

Affiliation: Department of Maxillofacial Biomedical Engineering, School of Dentistry, and Department of Life and Nanopharmaceutical Sciences, Kyung Hee University, Seoul 02447, Republic of Korea.

Running Title: Analysis of oral microbiome

Keywords: Oral microbiome, Diversity, Bioinformatics, Next-generation Sequencing (NGS), Human Microbiome Project (HMP)

Corresponding Author's Information:

Tel: +82-2-961-9290; Fax: +82-2-962-0598; E-mail: jaehlee@khu.ac.kr (J.-H. Lee)

Tel: +82-2-961-0795; Fax: +82-2-962-0598; E-mail: prudence75@khu.ac.kr (J.-H. Moon)

ABSTRACT

Human oral cavity contains a highly personalized microbiome that is essential to maintaining health but capable of causing oral and systemic diseases. Thus, an in-depth definition of "healthy oral microbiome" is critical to understanding variations in disease states from preclinical conditions and disease onset through progressive states of disease. With rapid advances in DNA sequencing and analytical technologies, population-based studies have documented the ranges and diversity of both taxonomic compositions and functional potentials observed in the oral microbiome in healthy individuals. Besides factors specific to the host, such as age and race/ethnicity, environmental factors also appear to contribute to the variability of the healthy oral microbiome. Here, we review bioinformatic techniques for metagenomic dataset, with some comments on their strengths and limitations. We also summarize our knowledge on the interpersonal and intrapersonal diversity of the oral microbiome, in the light of recent large-scale and longitudinal studies including Human Microbiome Project.

INTRODUCTION

The human microbiota (the collection of microbes that live on and inside us) consists of a wide range of microorganisms whose aggregate membership exceeds human somatic and germ cells by at least an order of magnitude (1,2). The collection of genes in the microbiota is called the human microbiome (2) but “microbiota” and “microbiome” are often used interchangeably (3). As one of the most clinically relevant microbial habitats, the human oral cavity is colonized by a personalized set of microorganisms, including bacteria, archaea, fungi, and viruses (4). During health, the oral microbiota lives in harmony with the host, as found at other body sites. The host is providing its microbiome with an environment, in which they can flourish and keep their host healthy (5). On the other hand, the oral microbiome is also considered a key source in the etiology of oral diseases, including dental caries and the periodontal diseases, as well as many systemic diseases such as diabetes and cardiovascular diseases (5,6). Because of its crucial role in oral and systemic health, the oral microbiome has become an essential part of microbiomics.

An in-depth definition of healthy microbiome is indispensable step toward detecting significant variations both in disease states and in pre-clinical conditions as well as understanding disease onset and progression (7). The advent of next generation sequencing (NGS) or high-throughput sequencing has revolutionized the field of microbiome analysis, providing the tools necessary to address the issue (8). This led to the launch of the NIH's Human Microbiome Project (HMP), constructed as a large, genome-scale community research project (NIH HMP Working Group, 2009). This project enrolled over 200 healthy adults and collected samples from 15 to 18 body

habitats, including oral, stool, skin, nasal, and vaginal areas, over one to three visits (9). Besides two major scientific reports (9,10) several companion papers have analyzed HMP oral datasets (7, 11-13), revealing great variability of the oral microbiome among and within healthy individuals. Furthermore, other recent large-scale and longitudinal studies have expanded our view of the oral microbiome beyond that of the HMP.

In this paper, we review bioinformatic techniques for metagenomic dataset including microbial community profiling, and highlight strengths and weaknesses of the experimental approaches. We also summarize important findings that lead to the current understanding of the ranges of healthy microbial diveristy. While viruses, fungi, archaea and protozoa form a part of a normal microbiome (4) the majority of the research is concentrated on the domain Bacteria. Therefore, we will focus exclusively on the oral bacteria in this review.

BIOINFORMATIC ANALYSIS OF MICROBIOME SEQUENCE DATA

Two distinct metagenomics approaches are commonly used: marker gene metagenomics and full shotgun metagenomics. Marker gene metagenomics is a fast and cost-effective way to obtain a taxonomic distribution profile. In this approach, specific regions of evolutionarily conserved marker genes are firstly amplified by PCR and subsequently sequenced (14). In the case of bacterial (and/or archaeal) community analysis, the target region usually contains the 16S ribosomal RNA (rRNA) gene (15), hence herein the approach is referred to as 16S rRNA profiling. Meanwhile full shotgun metagenomics, also referred as metagenomic whole genome sequencing (WGS), does not target a specific locus or marker genes, but instead involves breaking the isolated metagenomic

DNA into small pieces and subsequent sequencing the individual pieces (14). The sequenced small fragments (i.e., sequencing raw reads) can be used not only for taxonomy profiling (who is there?) as well as for functional profiling (what are they doing?) (14). In this section, we briefly describe the scheme of the techniques and the bioinformatic pipelines to analyze microbiome sequence data obtained from the both methods.

16S rRNA profiling

Ever since their introduction as markers for the bacterial phylogeny by Woese et al (16), the 16S rRNA gene has been considered the gold standard for phylogenetic studies of microbial communities and for assigning taxonomic names to bacteria (11). Bacterial 16S rRNA genes generally contain nine hypervariable regions (V1-V9) that demonstrate considerable sequence diversity among different bacterial species (17). Numerous studies have assessed the 16S rRNA gene regions to choose most appropriate conserved regions that can be used generate amplicons using universal primers as well as most effective hypervariable regions to target (17-23): unfortunately, no single hypervariable region is able to distinguish among all bacteria and a bias can be introduced by primer specificity as well as efficiency. Basically, the 16S rRNA profiling can be summarized into three steps; **(1)** Preprocessing and denoising of raw reads, **(2)** Taxonomic assignment, and **(3)** Evaluation of microbial diversity.

(1) Preprocessing and denoising of raw sequencing reads

Although there are standard operations and protocols to generate the sequencing library

in NGS, stochastic errors in the biological processes for the library creation and/or incomplete chemical reactions in sequencing could affect the overall quality of the sequencing library and sequencing datasets. Therefore, raw sequencing reads generated from sequencing machine should be carefully checked for the successful downstream analysis in the preprocessing step. A number of computational tools have been used for the preprocessing: for example, FastQC (bioinformatics.babraham.ac.uk/projects/fastqc/) provides a quick quality check by running a modular set of analyses such as “per base sequence quality”, “per sequence quality score”, “sequence length distribution”, “adapter content”, etc.; FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) allows detecting and trimming the low quality region of the individual read (especially 3’-end of the reads); DUST is used to remove low-complexity regions in the sequencing read (24). Intrinsically, the NGS techniques can harbor various errors in the sequencing reads such as imprecise signals from longer homopolymer runs and chimera sequences. In the denoising step, those errors were identified and corrected for the accurate taxonomic assignments of the sequencing reads. Many popular software, such as QIIME (25) and mothur (26), have implemented the denoising algorithms. In particular, UCHIME is designed to detect chimeric sequences by comparing reference sequences to a database or by performing *de novo* classification (clustering) (27). Preprocessed and denoised raw sequencing reads are subsequently subject to taxonomic assignment process.

(2) Taxonomic assignment

As NGS allows investigators to detect and identify novel bacteria that have previously

gone undetected, assignment of 16S rRNA gene sequences from uncultured bacteria into a bacterial taxonomy is even more challenging. Two frequently used methods assign reads into bins based on either their similarity to reference sequences (i.e., phylotyping) or their similarity to other sequences in the community (i.e., operational taxonomic units [OTUs]) (28). First method relies upon aligning reads with the reference 16S rRNA database using sequence alignment algorithms, such as BLAST (29). Besides NCBI Genbank, a number of rRNA databases have been constructed and used for the taxonomic assignment (Table 1). Each database has own criteria for the curation of data from the original resources. For example, Human Oral Microbiome Database (HOMD) (30) and CORE (31) database have been constructed using 16S rRNA sequences exclusively from human oral bacteria. The other approach is to group 16S rRNA sequencing reads into bins called OTUs with distance-based agglomerative clustering methods, such as CD-HIT (32) and UCLUST (33). Defining species by 97% identity in 16S rRNA gene sequence is a commonly used criterion, but these distinctions are still controversial (11,34).

Current NGS platforms produce vastly greater numbers of reads than Sanger sequencing while the reads are relatively much shorter. Unfortunately, existing tools are generally not sufficient to provide species names or phylogenetic information for the millions of short sequence reads (11). For example, the most commonly used tool for assigning taxonomy, the Ribosomal Database Project (RDP) Classifier (35), does not assign taxonomic names below the genus level. (11,36). Moreover, RDP shows insufficient resolution for classifying the GN02 and *Synergistetes*, as revealed in our previous study (37). To complement analyses relying on limited taxonomic names, 16S rRNA sequences are usually grouped using the OTU approach described above. Huse et

al. (11) explored the HMP oral microbiota from over 200 individuals and identified between 857 and 4,216 OTUs (Table 2). In terms of community membership, oral communities were especially diverse, showing the highest estimate of total richness after the stool. Notably, richness as measured by the V1–V3 primers was consistently higher than richness measured by V3–V5 (11). In addition, some taxa (e.g. Lactobacilli OTUs) are resolved better with V1–V3 while others (e.g. Bifidobacteriaceae OTUs) with the V3–V5 (11). These differences may be due a mismatch of the primers for amplification or an inability to differentiate the taxon in that region of the rRNA gene (11). Therefore, as with all 16S rRNA sequencing projects, the specific richness and diversity results should be compared with other results using the same 16S rRNA region, and the presence of primer bias should not be discounted (11). Furthermore, platform-dependent sequencing errors will also affect the taxonomic classification of reads, potentially leading to spurious OTUs and inflated measurements of diversity, thus making direct comparisons between studies difficult (12).

(3) Evaluation of microbial diversity

Diversity measurement is important for understanding community structure and dynamics. Two diversity measurements are frequently used to assess and compare microbial communities; alpha (or within-sample) diversity and beta (or between-sample) diversity. Alpha diversity is usually characterized using the total number of organisms within a sample (richness, might be measured as the number of OTUs), the relative abundances of the organisms (evenness), or indices that combine these two dimensions. Beta diversity, on the other hand, is often characterized using the number

of species (or OTUs) shared between two communities. In particular, UniFrac, a robust method for comparing the differences between microbial communities between samples, measures the proportion of shared branch lengths on a phylogenetic tree between samples (3,38). Principal Coordinates Analysis (PCoA) can summarize and visualize the UniFrac distances between samples in a scatterplot where points (representing samples) that are more distant from one another are more dissimilar.

Metagenomic WGS data analysis

The 16S rRNA profiling is powerful, effective and straightforward techniques to study microbial communities, but it only provides the taxonomic composition. Meanwhile, metagenomic WGS data can provide not only taxonomy but also the biological functional profiles for the microbial communities. The principles of taxonomy profiling process employing WGS data is similar to those described above, hence, in this section, we will focus on the functional profiling of microbial community. The analysis pipeline can be divided into four stages, **(1)** Preprocessing, **(2)** Reconstruction of raw sequencing reads (assembly), **(3)** Gene prediction, and **(4)** Functional annotations.

(1) Preprocessing

Preprocessing is to assess the overall quality of WGS data and most steps are similar to those in 16S rRNA profiling. Additionally, raw metagenomic NGS reads associated with a host (e.g. human) should be checked for the host DNA contamination and the contaminated sequencing reads should be removed. Fast short read mapping tools such

as BWA (39) and Bowtie 2 (40) are used to detect the contaminated sequencing reads by aligning raw sequencing reads against host genome (e.g. human genome).

(2) Reconstruction of raw sequencing reads (assembly)

The metagenomic WGS technique generates raw sequencing reads from the whole microbial genomes in the microbial community. Thus, to identify the specific genomes and/or complete protein coding genes in the genomes accurately, it is helpful to reconstruct the microbial genomes from raw sequencing reads. However, obtaining complete genomes has been challenging not only because of highly repetitive DNA sequences abundant in a broad range of species (from bacteria to mammals) but also because of short reads and high data volumes produced by NGS technology. Therefore, an assembly of shorter reads into genomic contigs and orientation of these into scaffolds is often performed. Most of the metagenomic WGS read assembly tools are designed and implemented based on the graph theory algorithm, de Bruijn graph. Initially, the method fragments all sequencing reads into k-mers and then, the generated k-mers can be used as the edges in the de Bruijn graph. The nodes of (k-1)-mer prefix and suffix are linked by the edges of k-mers for the graph. Finally, the assembler identifies Eulerian paths that go across all edges just once in the graph (41). Velvet (42), ABySS (41) and SOAPdenovo (44) use the de Bruijn graph to assemble whole metagenomes from raw sequencing reads. In HMP, the raw sequencing reads from 749 metagenomic samples were successfully used to assembly of contigs using an optimized SOAPdenovo protocol (8). Recently, more sophisticated algorithms have been developed and applied to the next-generation assemblers such as Meta-IDBA (45), MetaVelvet-SL (46) and

IDBA-UD (47).

(3) Gene prediction

The next stage of the analysis pipeline is to identify genes in the reads or assembled contigs and/or scaffolds. The prediction of genes in metagenomic contents is still a fairly difficult problem, although several gene prediction algorithms have been successfully employed for prokaryotic genomes. To predict genes in metagenomic study, especially for *de novo* genes, several computational methods have been developed, including MetaGeneMark (48), MetaProdigal (49), Glimmer-MG (50), and FragGeneScan (51). Notably, the performance of gene-predicting tools varies considerably: for example, in a comparison of five widely used *ab initio* gene-calling algorithms including FragGeneScan and MetaGeneMark, FragGeneScan is rather accurate for predicting reading frames on short raw reads (75–1000 bp) while other tools, such as MetaGeneMark, are better suited for higher-quality sequences such as assembled contigs (52). Moreover, it has been reported that combining various programs' predictions can improve the accuracy of prediction and annotation of metagenomic reads (53). Accordingly, researchers should carefully decide what tools to use in their metagenomic study, potentially impacting the results and conclusion.

(4) Functional annotations

After gene prediction, the identified genes are functionally annotated by comparing the

known genes in the functional annotation databases such as PFAM (54), IMG/M (55), COG (56) and MetaRef (57). Further analysis of the relationship between the microbiome and the host phenotype is performed using metabolic pathway information database, i.e., KEGG (58), eggNOG (59) and MinPath (60). In the part of HMP, Abubucker et al. devised HMP Unified Metabolic Analysis Network (HUMAnN) to construct metabolic networks of the microbial communities (61). In this study, raw sequencing reads were searched against a protein sequence databases and HUMAnN recovers the abundances of individual orthologues gene families and pathway. More specifically, MBLASTX, KEGG orthology and MinPath have been used to assign genes and available pathways. Recently, several metagenomic analysis pipeline software, such as MG-RAST (62) and IMG/MER (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi) has been developed. The pipelines provide the functional annotation modules in their fully automated pipeline web-server and thus, researchers can easily perform functional annotation tasks using their own data in the web (15).

COMPOSITION AND DIVERSITY OF ORAL MICROBIOME

The HMP assessed oral microbiome composition of seven intra oral sites (buccal mucosa, hard palate, keratinized gingiva, saliva, sub- and supra gingival plaque, and tongue dorsum) and two oropharyngeal sites (throat and palatine tonsils) from 182~206 healthy subjects (18 to 40 years old) and found 185-322 genera, belonging to 13-19 bacterial phyla (13). Dominating phyla were *Firmicutes*, *Bacteroidetes*, *Proteobacteria*, *Fusobacteria* and *Actinobacteria*, accounting for over 95% of the entire oral

279 microbiome. An individual sample from a single site of a single subject contained 23-50
280 genera from 6-9 phyla (13). Among all body habitats, the oral habitats have the highest
281 alpha diversity showing the highest OTU level richness after the stool (Table 2), while
282 the skin and vaginal microbiota show lower alpha diversity (11,13). In comparisons
283 between samples from the same habitat among subjects (beta diversity), oral sites have
284 the lowest beta diversities, which means that members of the population shared
285 relatively similar organisms in oral sites than in other body sites (10). However, HMP
286 oral datasets also emphasize the highly variable nature across individuals, especially at
287 the sub-genus level: even OTUs present in nearly every subject, or that dominate in
288 some samples, showed orders of magnitude variation in relative abundance (11). In the
289 following sections we discuss in more detail about the specific factors that contribute to
290 the variability of the healthy oral microbiome.
291
292 **Different oral biogeographic niches**
293 The oral cavity is a moist environment which is kept at a relatively constant temperature
294 (34 to 36°C) and a pH close to neutrality in most areas and thus supports the growth of a
295 wide variety of microorganisms (63). The oral cavity is composed of diverse habitats
296 with different anatomical structures and physicochemical factors: the oral mucosa
297 covers the cheek, tongue, gingiva, palate, and floor of the mouth and allows rapid
298 elimination of adhering bacteria due to a continuous desquamation of its surface
299 epithelial cells (63). On the other hand, papillary surface of the tongue provides sites of
300 colonization that are protected from mechanical removal. The hard surfaces of teeth
301 offer many different sites for colonization by bacteria below (subgingival) and above
302 (supragingival) the gingival margin. The gingival crevice, the area between the

junctional epithelium of the gingiva and teeth, provides a unique colonization site that include both hard and soft tissues (63). The epithelium may be keratinized (palate) or nonkeratinized (gingival crevice). Hence, the oral cavity is not considered a uniform environment.

HMP revealed a substantial divergence in the species richness and evenness among different oral habitats as well as identified microorganisms with specific niche preferences. Hard palate showed the lowest estimate of total richness, while the gingival plaque showed the high estimate of total richness (11) (Table 2). Oral sites, particularly saliva, have the highest evenness while buccal mucosa and keratinized gingiva have lower alpha diversity than the other oral sites (10,13). Each oral habitat in almost every subject was characterized by one or a few signature taxa making up the plurality of the community with highly variable relative abundance both among individuals and oral habitats. Most oral habitats are dominated by *Streptococcus*, but these are followed in abundance by *Haemophilus* in the buccal mucosa, *Actinomyces* in the supragingival plaque, and *Prevotella* in the subgingival plaque (10,13). There is overlap of species detected in almost all oral sites, such as certain species of *Streptococcus* (OTUs #2, 5 and 6), *Gemella* (OTUs #7 and 8), *Granulicatella* (OTU #13), *Fusobacterium* (OTUs #9 and 27), and *Veillonella* (OTUs #4 and 7) (11). However, several abundant genera had multiple OTUs with distinct preferences for often only one or two of the nine oral sites, such as *Bacteroides*, *Prevotella*, *Corynebacterium*, *Fusobacterium*, *Pasteurella*, and *Neisseria* (11). For example, *Corynebacterium matruchotii* (OTU #15) was present almost exclusively in the supragingival plaque, while *Corynebacterium argentoratense* (OTU #188) mostly in saliva and to a lesser extent on the hard palate (11). It may be due to the shedding of the epithelial cells and the shear forces from chewing in the buccal

fold and the hard palate (64). In an analysis of oral samples collected from the elderly (range 73–93), *Lautropia mirabilis* was significantly associated with the supragingival plaque while *Treponema socranskii* was found only in the subgingival plaque (65), which may be explained by the low oxidation-reduction potential of subgingival plaque. In the oropharynx, the distribution of *Firmicutes*, *Proteobacteria*, and *Bacteroidetes* was similar to that in saliva, but more *Proteobacteria* than in the mouth (66).

Influence of geography, climate and ethnicity

Although the HMP generated an incredible volume of data, the resulting 16S rRNA datasets are composed of samples from medical students in the USA and host information is nearly prohibitive to access, which lead to removal of the potential to observe any systematic patterns and regional or ethnic differences (67). A population-scale study of 120 healthy individuals from 12 worldwide locations found a significant association between variation in the saliva microbiome and the distance of each location from the equator (68). Notably, the saliva microbiome of Batwa Pygmies, a former hunter-gatherer group from Africa, was found to be much more diverse than the saliva microbiome of two agricultural African groups, most likely because of their different lifestyle and diet (69). Another study of 3 human groups from different geographic and climatic areas (76 native Alaskans, 10 Germans and 66 Africans) showed the distinctiveness of the saliva microbiome, the reasons of which (e.g. differential lifestyles including diet and/or host genetics and physiology including the immune system) remain to be elucidated (70). Alpha diversity was highest for the German group and lowest for the African group, while the opposite was true for beta diversity. It is intriguing to speculate that higher population density of Germany may provide more

opportunities for bacteria to be spread among individuals (71).

Ethnicity is likely to exert a selection pressure on the oral microbiome: Mason et al. (71), analyzed dental plaque and saliva samples collected from 192 subjects belonging to four ethnic affiliations (non-Hispanic blacks, non-Hispanic whites, Chinese, and Latinos) and found obvious ethnicity-specific clustering of microbial communities, expanding prior observations (72-74). This selection pressure seems genetic rather than environmental, since the two ethnicities that shared a common food, nutritional and lifestyle heritage (Caucasians and African Americans) demonstrated significant microbial divergence (71). It is known that not only innate immune responses to infectious agents but also tooth morphologies vary according to ethnic affiliation (75-78). Hence, it is possible that ethnicity plays a role in bacterial selection by defining the environment for bacterial colonization (71).

Vertical and horizontal transmission

Vertical transmission from mother to child starts at birth (79). Depending on the delivery mode (vaginal or Caesarian), infants acquire bacterial communities resembling their own mother's vaginal microbiota or similar to those found on the skin surface (80). A study of healthy three-month-old infants delivered vaginally (25 infants) and born by C-section (38 infants) found differences in the infant’s oral microbiota due to mode of delivery, with vaginally delivered infants having a higher taxonomic diversity (81). The method of feeding (breast-feeding or infant formula) also affects the infant’s microbiome as well: oral lactobacilli with antimicrobial properties were found in breast-fed infants but not found in formula-fed infants (82,83). Horizontal transmission of oral microbiota among siblings and other individuals sharing the same environment also

375 contributes to oral microbiome diversity. In a study of 264 saliva samples collected from
376 107 individuals including 45 twin pairs, at up to three time-points during 10-year
377 spanning adolescence, twins resembled each other more closely than the whole
378 population at all time-points, but became less similar to each other when they aged and
379 no longer cohabited (84).
380
381 **Temporal variation**
382 Studies looking at the temporal variation of the oral microbiome have found conflicting
383 results: in a longitudinal study of five adults at three time-points (from 5 to 29 days),
384 salivary microbial community appeared to be stable at different time points (85). HMP
385 consortium (10) and Zhou et al. (13) reported that, among 22 HMP habitats of human
386 body, the oral habitat has the most stable microbiota, showing the highest community
387 similarity between two visits (mean time interval between visits is 212 days) while the
388 skin and vaginal microbiota are less stable. In contrast, a reanalysis of the HMP datasets
389 by a method for quantifying the difference between two cohorts revealed that the
390 relative abundances of core OTUs in individual sample showed significantly greater
391 changes from 1st to 2nd visit at oral and stool body regions compared with vaginal body
392 region (12). More recently, a longitudinal study of 85 adults weekly over 3 months
393 showed high levels of temporal variability in both diversity and community structure in
394 tongue microbiome, as in other body habitats studied (86). Furthermore, this study
395 found that both the composition of an individual's microbiome and their degree of
396 temporal variability shows a personalized feature. Collectively, although intrapersonal
397 variation over time is lower than interpersonal variation, intrapersonal temporal
398 dynamics are need to be considered when attempting to link changes in microbiome

399 structure to changes in health status (86).

400

401 **Age-related changes**

402 Along with a variety of physiological changes which accompany the aging process,
403 microbial habitats also greatly change in the oral cavity: the eruption of primary teeth
404 and replacement of the primary dentition with permanent dentition may lead to shifts in
405 microbial community composition at different phases in people's lives (87). Edentulous
406 infants have been found to have lower diversity than their mothers or primary care
407 givers in the oral microbial composition (88). In the deciduous dentition, a higher
408 proportion of Proteobacteria (*Gammaproteobacteria, Moraxellaceae*) was found than
409 that of Bacteroidetes. With increasing age, Bacteroidetes (mainly genus *Prevotella*),
410 *Veillonellaceae* family, Spirochaetes, and candidate divisionTM7 increased (89).
411 Several organisms, including members of the genera *Veillonella, Actinomyces* and
412 *Streptococcus*, were reported to have age-specific abundance profiles during
413 adolescence (84). Xu et al., (87) analyzed of the oral (saliva, supragingiva and mucosa)
414 microbiome across a wide age range (3 days–76 years), in which only a very small
415 overlap of shared OTU was observed. In this study, a distinct temporal shift was
416 observed in the relative abundance of most genera. The average relative abundance of
417 the dominant bacterial phyla, *Actinobacteria, Bacteroides, Firmicutes, Fusobacteria,*
418 *Proteobacteria, Spirochetes* and candidate division TM7 varied by age/dentition stage
419 (87).

420

421 **CONCLUDING REMARKS**

422 The tremendous diversity of oral microbiome has only begun to be realized and a

number of challenges, such as the vast uncultivated species and the lack of reference genomes, currently remain (90). Until recently, about half of all known bacterial phyla were identified only from their 16S rRNA gene sequences (91). In fact, the bacteria that can be grown in the laboratory are only a portion of the total diversity that exists in the oral cavity (92). One method to address this challenge is single-cell genomics, which is a powerful tool for accessing genetic information from uncultivated microorganisms (93). Future work combining metagenomics and single cell genomics, as well as advances in each separate method, should help to overcome these issues, providing new insights into uncultivated lineages (94).

Rapidly developing sequencing methods and analytical techniques are enhancing our ability to understand the human microbiome, leading to the concept of a 'personal microbiome'. The focus now shifts from characterizing oral microbiota to functional studies encompassing genomics, transcriptomics, and metabolomics of both host and microbes. Future investigations will inevitably be personal omics profiling in order to probe the temporal patterns associated with both molecular changes and related physiological health and disease. This knowledge is vital for the development of efficacious prevention and treatment protocols for oral diseases and, ultimately, contributes to the development of personalized medicine and personalized dental medicine.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science, ICT & Future Planning (NRF-2015R1C1A2A01054588). This work was also supported in part by 2016 Sabbatical Research Program of Kyung Hee University.

REFERENCES

1. Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res 19, 1141–1152
2. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI (2007) The human microbiome project. Nature 449, 804–810
3. Ursell LK, Metcalf JL, Parfrey LW, Knight R (2012) Defining the human microbiome. Nutr Rev Suppl 1, S38-44
4. Wade WG (2013) The oral microbiome in health and disease. Pharmacol Res 69, 137-143
5. Zarco MF, Vess TJ, Ginsburg GS (2012) The oral microbiome in health and disease and the potential impact on personalized dental medicine. Oral Dis 18, 109-120
6. He J, Li Y, Cao Y, Xue J, Zhou X (2015) The oral microbiome diversity and its relation to human diseases. Folia Microbiol (Praha) 60, 69-80
7. Segata N, Haake SK, Mannon P et al (2012) Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol 13, R42
8. Gevers D, Pop M, Schloss PD, Huttenhower C (2012) Bioinformatics for the Human Microbiome Project. PLoS Comput Biol 8, e1002779

9. Human Microbiome Project Consortium (2012b) A framework for human microbiome research. Nature 486, 215-221
10. Human Microbiome Project Consortium (2012a) Structure, function and diversity of the healthy human microbiome. Nature 486, 207-214
11. Huse SM, Ye Y, Zhou Y, Fodor AA (2012) A core human microbiome as viewed through 16S rRNA sequences clusters. PLoS One 7, e34242
12. Li K, Bihan M, Methé BA (2013) Analyses of the stability and core taxonomic memberships of the human microbiome. PLoS One 8, e63139
13. Zhou Y, Gao H, Mihindukulasuriya KA et al (2013) Biogeography of the ecosystems of the healthy human body. Genome Biol 14, R1
14. Sharpton TJ (2014) An introduction to the analysis of shotgun metagenomic data. Front Plant Sci 5, 209
15. Oulas A, Pavloudi C, Polymenakou P et al (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform Biol Insights 9, 75-88
16. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74, 5088-5090
17. Mizrahi-Man O, Davenport ER, Gilad Y (2013) Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. PLoS O 8, e53608
18. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73, 5261–5267

19. Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Relman DA, Sogin ML. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet 4, e1000255

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

21. Nossa CW, Oberdorf WE, Yang L et al (2010) Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. World J Gastroenterol 16, 4135–4144

22. Claesson MJ, Wang Q, O'Sullivan O et al (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res 38, e200

23. Tremblay J, Singh K, Fern A et al (2015) Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol 6, 771

24. Morgulis A, Gertz EM, Schäffer AA, Agarwala R (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol 13, 1028-1040

25. Sevinsky JR, Turnbaugh PJ, Walters WA et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7, 335-336

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

27. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27, 2194-2200

28. Schloss PD, Westcott SL (2011) Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl Environ Microbiol 77, 3219–3226

29. Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402

30. Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE (2010) The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database (Oxford) 2010:baq013

31. Griffen AL, Beall CJ, Firestone ND et al (2011) CORE: a phylogenetically-curated 16S rDNA database of the core oral microbiome. PLoS One 6, e19051

32. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150-3152

33. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460-2461

34. Schloss PD, Handelsman J (2005) Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. Appl Environ Microbiol 71,1501-1506

35. Cole JR, Wang Q, Fish JA et al (2013) Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42, D633-642

538 36. Kotamarti RM, Hahsler M, Raiford D, McGee M, Dunham MH (2010)
539 Analyzing taxonomic classification using extensible Markov models.
540 Bioinformatics 26, 2235-2241
541 37. Moon JH, Lee JH, Lee JY (2015) Subgingival microbiome in smokers and non-
542 smokers in Korean chronic periodontitis patients. Mol Oral Microbiol 30, 227-
543 241
544 38. Lozupone C, Knight R. (2005) UniFrac: a new phylogenetic method for
545 comparing microbial communities. Appl Environ Microbiol 71, 8228-35
546 39. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-
547 Wheeler transform. Bioinformatics 25, 1754-1760
548 40. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2.
549 Nat Methods 9, 357-359
550 41. Kim Y, Koh I, Rho M (2014) Deciphering the human microbiome using next-
551 generation sequencing data and bioinformatics approaches. Methods 79-80, 52-
552 59
553 42. Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read
554 assembly using de Bruijn graphs. Genome Res 18, 821-829
555 43. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS:
556 a parallel assembler for short read sequence data. Genome Res 19, 1117-1123
557 44. Luo R, Liu B, Xie Y et al (2012) SOAPdenovo2: an empirically improved
558 memory-efficient short-read *de novo* assembler. Gigascience 1, 18
559 45. Peng Y, Leung HC, Yiu SM, Chin FY (2011) Meta-IDBA: a *de Novo* assembler
560 for metagenomic data. Bioinformatics 27, i94-101

46. Afiahayati, Sato K, Sakakibara Y (2015) MetaVelvet-SL: an extension of the Velvet assembler to a *de novo* metagenomic assembler utilizing supervised learning. DNA Res 22, 69-77

47. Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420-1428

48. Zhu W, Lomsadze A, Borodovsky M (2010) *Ab initio* gene identification in metagenomic sequences. Nucleic Acids Res 38, e132

49. Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC (2012) Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 28, 2223-30

50. Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res 40, e9

51. Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res 38, e191

52. Trimble WL, Keegan KP, D'Souza M et al (2012) Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. BMC Bioinformatics 13, 183

53. Yok NG, Rosen GL (2011) Combining gene prediction methods to improve metagenomic gene annotation. BMC Bioinformatics 12, 20

54. Finn RD, Coggill P, Eberhardt RY et al (2015) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44, D279-285

583 55. Markowitz VM, Chen IM, Chu K et al (2014) IMG/M 4 version of the
584 integrated metagenome comparative analysis system. Nucleic Acids Res 42,
585 D568-573
586 56. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a
587 tool for genome-scale analysis of protein functions and evolution. Nucleic Acids
588 Res 28, 33-36.
589 57. Huang K, Brady A, Mahurkar A et al (2014) MetaRef: a pan-genomic database
590 for comparative and community microbial genomics. Nucleic Acids Res 42,
591 D617-624.
592 58. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes.
593 Nucleic Acids Res 28, 27-30
594 59. Huerta-Cepas J, Szklarczyk D, Forslund K (2016) eggNOG 4.5: a hierarchical
595 orthology framework with improved functional annotations for eukaryotic,
596 prokaryotic and viral sequences. Nucleic Acids Res 44, D286-293.
597 60. Ye Y, Doak TG (2009) A parsimony approach to biological pathway
598 reconstruction/inference for genomes and metagenomes. PLoS Comput Biol 5,
599 e1000465
600 61. Abubucker S, Segata N, Goll J et al (2012) Metabolic reconstruction for
601 metagenomic data and its application to the human microbiome. PLoS Comput
602 Biol 8, e1002358
603 62. Keegan KP, Glass EM, Meyer F (2016) MG-RAST, a Metagenomics Service for
604 Analysis of Microbial Community Structure and Function. Methods Mol Biol
605 1399, 207-233

606 63. Marcotte H, Lavoie MC (1998) Oral Microbial Ecology and the Role of
607 Salivary Immunoglobulin A. Microbiol Mol Biol Rev 62, 71–109
608 64. Chen H, Jiang W (2014) Application of high-throughput sequencing in
609 understanding human oral microbiome related with health and disease. Front
610 Microbiol 5, 508
611 65. Preza D, Olsen I, Willumsen T, Grinde B, Paster BJ (2009) Diversity and site-
612 specificity of the oral microflora in the elderly. Eur J Clin Microbiol Infect Dis
613 28, 1033-1040
614 66. Lemon KP, Klepac-Ceraj V, Schiffer HK, Brodie EL, Lynch SV, Kolter R
615 (2014) Comparative analyses of the bacterial microbiota of the human nostril
616 and oropharynx. MBio 1, e00129-10
617 67. McDonald D, Birmingham A, Knight R (2015) Context and the human
618 microbiome. Microbiome 3, 52
619 68. Nasidze I, Li J, Quinque D, Tang K, Stoneking M (2009) Global diversity in the
620 human salivary microbiome. Genome Res 19, 636–643
621 69. Nasidze I, Li J, Schroeder R, Creasey JL, Li M, Stoneking M (2011) High
622 Diversity of the Saliva Microbiome in Batwa Pygmies. PLoS One. 6, e23352
623 70. Li J, Quinque D, Horz HP et al (2014) Comparative analysis of the human
624 saliva microbiome from different climate zones: Alaska, Germany, and Africa.
625 BMC Microbiol 14, 316
626 71. Mason MR, Nagaraja HN, Camerlengo T, Joshi V, Kumar PS (2013) Deep
627 sequencing identifies ethnicity-specific bacterial signatures in the oral
628 microbiome. PLoS One 8, e77287

629 72. Rylev M, Kilian M (2008) Prevalence and distribution of principal periodontal
630 pathogens worldwide. J Clin Periodontol 35(8 Suppl), 346-361
631 73. Haffajee AD, Bogren A, Hasturk H, Feres M, Lopez NJ, Socransky SS (2004)
632 Subgingival microbiota of chronic periodontitis subjects from different
633 geographic locations. J Clin Periodontol 31, 996-1002
634 74. Kim TS, Kang NW, Lee SB, Eickholz P, Pretzl B, Kim CK (2009) Differences
635 in subgingival microflora of Korean and German periodontal patients. Arch Oral
636 Biol 54, 223-229
637 75. Miller MA, Cappuccio FP (2007) Ethnicity and inflammatory pathways-
638 implications for vascular disease, vascular risk and therapeutic intervention.
639 Current medicinal chemistry 14, 1409–1425
640 76. Nguyen DP, Genc M, Vardhana S, Babula O, Onderdonk A, Witkin SS (2004)
641 Ethnic differences of polymorphisms in cytokine and innate immune system
642 genes in pregnant women. Obstetrics and gynecology 104, 293–300
643 77. Lavelle CL (1970) Crowding and spacing within the human dental arch of
644 different racial groups. Archives of oral biology 15, 1101–1103
645 78. Lavelle CL (1971) Mandibular molar tooth configurations in different racial
646 groups. Journal of Dental Research 50, 1353
647 79. Zaura E, Nicu EA, Krom BP, Keijser BJ (2014) Acquiring and maintaining a
648 normal oral microbiome: current perspective. Front Cell Infect Microbiol 4, 85
649 80. Dominguez-Bello MG, Costello EK, Contreras M et al (2010) Delivery mode
650 shapes the acquisition and structure of the initial microbiota across multiple
651 body habitats in newborns. Proc Natl Acad Sci USA 107, 11971-11975

81. Lif Holgerson P, Harnevik L, Hernell O, Tanner AC, Johansson I (2011) Mode of birth delivery affects oral microbiota in infants. J Dent Res 90, 1183-1188

82. Holgerson PL, Vestman NR, Claesson R et al (2013) Oral microbial profile discriminates breast-fed from formula-fed infants. J Pediatr Gastroenterol Nutr 56, 127-136

83. Vestman NR, Timby N, Holgerson PL et al (2013) Characterization and in vitro properties of oral lactobacilli in breastfed infants. BMC Microbiol 13, 193

84. Stahringer SS, Clemente JC, Corley RP et al (2012) Nurture trumps nature in a longitudinal survey of salivary bacterial communities in twins from early adolescence to early adulthood. Genome Res 22, 2146-2152

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

86. Flores GE, Caporaso JG, Henley JB et al (2014) Temporal variability is a personalized feature of the human microbiome. Genome Biol 15, 531

87. Xu X, He J, Xue J et al (2015) Oral cavity contains distinct niches with dynamic microbial communities. Environ Microbiol 17, 699-710

88. Cephas KD, Kim J, Mathai RA et al (2011) Comparative analysis of salivary bacterial microbiome diversity in edentulous infants and their mothers or primary care givers using pyrosequencing. PLoS One 6, e23503

89. Crielaard W, Zaura E, Schuller AA, Huse SM, Montijn RC, Keijser BJ (2011) Exploring the oral microbiota of children at various developmental stages of their dentition in the relation to their oral health. BMC Med Genomics 4, 22

90. McLean JS (2014) Advancements toward a systems level understanding of the human oral microbiome. Front Cell Infect Microbiol 4, 98

91. Lasken RS, McLean JS (2014) Recent advances in genomic DNA sequencing of microbial species from single cells. Nature Reviews Genetics 15, 577–584

92. Dewhirst FE, ChenT, Izard J et al (2010). The human oral microbiome. J Bacteriol 192, 5002–5017

93. Clingenpeel S, Schwientek P, Hugenholtz P, Woyke T (2014). Effects of sample treatments on genome recovery via single-cell genomics. ISME J 8, 2546-9

94. Solden L, Lloyd K, Wrighton K (2016) The bright side of microbial dark matter: lessons learned from the uncultivated majority. Current Opinion in Microbiology 31, 217-226

95. Kim OS1, Cho YJ, Lee K et al (2012) Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. Int J Syst Evol Microbiol 62, 716-721

96. C, Pruesse E, Yilmaz P, Gerken J et al (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res 41, D590-596.

97. DeSantis TZ, Hugenholtz P, Larsen N et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72, 5069-5072

698 **Table 1.** A list of 16S ribosomal RNA database

Name	16S rRNA coverage	Database URL (reference)
CORE	Human Oral Bacteria	http://microbiome.osu.edu/ (32)
RDP	Archaea and Bacteria	https://rdp.cme.msu.edu/ (33)
Human Oral Microbiolome Database	Human Oral Bacteria	http://www.homd.org/index.php (65)
EzTaxon-e	Archaea and Bacteria	http://www.ezbiocloud.net/eztaxon (95)
SILVA	Archaea and Bacteria	https://www.arb-silva.de/ (96)
Greengenes	Archaea and Bacteria	http://greengenes.secondgenome.com/ (97)

699

700

701 **Table 2.** Counts of patients included, OTUs and estimated richness (number of species)

702 found for both the V1–V3 and the V3–V5 regions (11).

Body Site	V1-V3			V3-V5		
	Patients	OTUs	[a] Estimated richness	Patients	OTUs	[a] Estimated richness
Buccal mucosa	114	2025	6635	198	898	4650
Hard palate	112	1741	3793	190	912	3125
Keratinized gingiva	117	1545	4387	206	857	3352
Palatine Tonsils	119	3683	10023	204	1633	9020
Saliva	99	2341	6546	181	1399	6801
Subgingival plaque	119	4216	14410	204	1672	11501
Supragingival plaque	121	3851	11154	205	1587	8254
Throat	110	2343	5601	192	1136	4154
Tongue dorsum	119	3651	7910	205	1503	7947
[b] Posterior fornix	59	428	1151	95	400	1466
[b] Stool	118	6050	23665	209	5391	33627

703 [a] Upper and lower confidence limits are not included in this table.

704 [b] Example of extraoral sites. The stool samples have the highest estimate of total

705 richness, followed by the oral sites, particularly the plaque and tonsils. The skin sites,

706 such as posterior fornix, have the lowest estimated richness.