Metagenomic study of the oral microbiota by Illumina high-throughput sequencing

Vladimir Lazarevic, Katrine Whiteson, Susan Huse, David Hernandez, Laurent Farinelli, Magne Østerås, Jacques Schrenzel, Patrice François

PII: S0167-7012(09)00296-6
DOI: doi: 10.1016/j.mimet.2009.09.012
Reference: MIMET 3257

To appear in: *Journal of Microbiological Methods*

Received date: 11 September 2009
Accepted date: 14 September 2009

Please cite this article as: Lazarevic, Vladimir, Whiteson, Katrine, Huse, Susan, Hernandez, David, Farinelli, Laurent, Østerås, Magne, Schrenzel, Jacques, François, Patrice, Metagenomic study of the oral microbiota by Illumina high-throughput sequencing, *Journal of Microbiological Methods* (2009), doi: 10.1016/j.mimet.2009.09.012

Metagenomic study of the oral microbiota by Illumina high-throughput sequencing

Vladimir Lazarevic[a*Δ], Katrine Whiteson[a*Δ], Susan Huse[b], David Hernandez[a], Laurent Farinelli[c], Magne Østerås[c], Jacques Schrenzel[a], Patrice François[a]

[a] Genomic Research Laboratory, Geneva University Hospitals, Rue Gabrielle-Perret-Gentil 4, CH-1211 Geneva 14, Switzerland

[b] Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA 02543, USA

[c] Fasteris, Ch. du Pont-du-Centenaire 109, Case postale 28, CH-1228 Plan-les-Ouates, Switzerland

Short title: metagenomic survey using oral bacteria

Keywords: metagenomics, oral cavity, flora composition, microbiome, high-throughput sequencer

[Δ] both authors contributed equally to this work

* Corresponding authors. Mailing address: Genomic Research Laboratory, Geneva University Hospitals, Rue Gabrielle-Perret-Gentil 4, CH-1211 Geneva 14, Switzerland. Tel.: +41 22 372 93 38; fax: +41 22 372 98 30.

E-mail address: vladimir.lazarevic@genomic.ch and katrine.whiteson@genomic.ch

ABSTRACT

To date, metagenomic studies have relied on the utilization and analysis of reads obtained using 454 pyrosequencing to replace conventional Sanger sequencing. After extensively scanning the 16S ribosomal RNA (rRNA) gene, we identified the V5 hypervariable region as a short region providing reliable identification of bacterial sequences available in public databases such as the Human Oral Microbiome Database. We amplified samples from the oral cavity of three healthy individuals using primers covering an ~82-base segment of the V5 loop, and sequenced using the Illumina technology in a single orientation. We identified 135 genera or higher taxonomic ranks from the resulting 1,373,824 sequences. While the abundances of the most common phyla (*Firmicutes, Proteobacteria, Actinobacteria, Fusobacteria* and *TM7*) are largely comparable to previous studies, *Bacteroidetes* were less present. Potential sources for this difference include classification bias in this region of the 16S rRNA gene, human sample variation, sample preparation and primer bias. Using an Illumina sequencing approach, we achieved a much greater depth of coverage than previous oral microbiota studies, allowing us to identify several taxa not yet discovered in these types of samples, and to assess that at least 30,000 additional reads would be required to identify only one additional phylotype. The evolution of high-throughput sequencing technologies, and their subsequent improvements in read length enable the utilization of different platforms for studying communities of complex flora. Access to large amounts of data is already leading to a better representation of sample diversity at a reasonable cost.

1. Introduction

Oral health, which is strongly influenced by oral microbiota, has a significant impact on general health. The bacterial community in the mouth contains species that promote health, and others that contribute to illness. Recent studies have shown that poor oral hygiene and/or the presence of particular species in the mouth may be associated with periodontitis, respiratory infection and intestinal disease (Avila et al., 2009; Kuehbacher et al., 2008; Raghavendran et al., 2007). In addition, the salivary microbiota may be used as a diagnostic marker for cancer (Mager et al., 2005) and periodontal disease (Faveri et al., 2008) as well as to provide insights into human population studies (Nasidze et al., 2009a). Understanding which species are present and how the community is composed in healthy adults is the first step towards understanding how changes can lead to disease.

Experts have recently raised the hypothesis that in some chronic diseases, the "pathogen" might be a disturbed microbial community rather than a single organism (Friedrich, 2008). Understanding the contribution of “behind-the-scenes” species which influence the pathogenicity of other species has already led to important changes in treatment strategies (Sibley et al., 2006). These unexpected interactions are changing how microbiologists think about causation of infection and disease (Lipkin, 2009).

Until recently, knowledge of the bacteria that reside in the human oral cavity was limited to those species that could be cultured in the laboratory. New sequencing technologies have brought tremendous improvements in automated sequencing and analysis of genome features. Today around 900 complete prokaryotic genomes are publicly available (www.ncbi.nlm.nih.gov/genomes/lproks.cgi) as well as more than a million 16s rRNA gene sequences, and several hundred metagenomic datasets. The Human Oral Microbiome Project (www.homd.org) now contains close to 1000 species which have been found in the mouth, while a metagenomic-based estimate of the diversity is one order of magnitude higher (Keijser et al., 2008). The availability of these extensive and varied sequences has opened the way for comparative genomics techniques (Fraser et al., 2000) for evaluating relatedness and diversity as well as studying whole viral or bacterial content of various media (Venter et al., 2004; Williamson et al., 2008) or bacterial infections (Cox-Foster et al., 2007; Nakamura et al., 2008; Turnbaugh et al., 2009).

Here, we evaluate the potential of Illumina high-throughput sequencing with an unprecedented depth of sequence coverage for the study of human oral microbiota diversity. We use partial sequences from the well-characterized and conserved 16S rRNA gene, to enable classification of bacteria from human oral samples.

2. Materials and methods

2.1. Sampling and DNA extraction

We collected saliva and oropharyngeal samples over a one-week period from three adult individuals with informed consent. Saliva samples were collected by expectoration into a sterile plastic 50-ml tube and kept frozen at -20°C until processing. We mixed 100 µL of each saliva sample with the same volume of 2x lysis buffer [Tris 20 mM, EDTA 2 mM (pH 8), Tween 1%, proteinase K (Fermentas) 400 µg/ml] and incubated them for 2 hours at 55°C (Faveri et al., 2008). Proteinase K was inactivated by a 10 min incubation at 95°C and the samples were frozen at -20°C.

Dry cotton swabs (Copan) were used to gently swab the posterior wall of the oropharynx. They were directly suspended in a microtube containing 200 µL of lysis buffer and processed in the same way as the saliva samples. The saliva and oropharyngeal lysates from all three subjects were mixed in a 1:10 ratio with roughly equal contributions from the two sampling sites according to PCR yield.

2.2. PCR primers and conditions

We aligned 753 16S rDNA sequences from the Human Oral Microbiome Database (HOMD, October 2008) using MAFFT (-FFT-NS-2, v6.531b) (Katoh et al., 2002). We chose primers from the conserved areas of the alignment flanking the V5 region so as to match most sequences. With a 100% match, primers 784DEG (5'-GGMTTAGATACCC) and 880RDEG (5'-CRTACTHCHCAGGYG) sequences produced 740 and 745 hits, respectively, or 732 (97.2%) of the HOMD sequences. Species coverage was within the 91-100% range for all HOMD bacterial phyla except *Chloroflexi* which is very rare in oral microbiota (Keijser et al., 2008) and has a single representative in the HOMD.

PCR amplification was carried out in a 50 µL PrimeStar HS Premix (Takara) containing 5 µL of lysate and 0.5 µM of each forward (784DEG) and reverse (880RDEG) primer. The samples were run in two separate PCRs for 15 cycles using the following parameters: 98°C for 10 s, 46°C for 15 s, and 72°C for 1 min. The two PCRs were then pooled and phosphorylated with polynucleotide kinase and the Illumina paired-ends adapters were ligated with T4 DNA

ligase. After PCR amplification with Phusion for 10 cycles using Illumina paired-ends PCR primers, the library was quality controlled by cloning an aliquot into a TOPO plasmid and capillary sequencing 16 clones. The library was sequenced from the forward end for 76 cycles on the Illumina Genome Analyzer system GAII using sequencing kits version 3.0. The 16S V5 amplicons correspond to *E. coli* positions 785 to 894 including primer sequence and to positions 798 to 879 excluding primers.

2.3. Sequence analysis

Base-calling was performed with the GAPipeline 1.3.2 using standard parameters, which include purity filtering with "chastity 0.6". We removed sequences containing uncalled bases, incorrect primer sequence or runs of ≥12 identical nucleotides. Seventy-two-base sequence reads were trimmed to remove the 13-base forward primer sequence, yielding 59-base sequences.

We assigned taxonomy to sequences with GAST (Huse et al., 2008), using a database of reference V5 rDNA sequences (RefHVR_V5) from SILVA (version 98) (Pruesse et al., 2007), and taxonomy from known cultured isolates, Entrez Genome projects, the Ribosomal Database Project [RDP; (Cole et al., 2005)], Greengenes (DeSantis et al., 2006) and hand curation. GAST compares each tag to the RefHVR_V5 and aligns it to its nearest neighbors in the database and then selects the closest reference(s). The taxonomy for the tag is the lowest common ancestor for a two-thirds majority of all 16S rDNA sequences associated with the nearest V5 reference sequences.

Before generating clusters of phylotypes, we filtered out all sequences that occurred fewer than 3 times. This reduced the number of unique sequences to a computationally manageable level, and potentially reduced the number of errors from sequencing and contamination. We created a multiple sequences alignment of the remaining data using MUSCLE (Edgar, 2004) with parameters –maxiters 2 and -diags, and generated phylotype clusters and diversity estimates using MOTHUR (Schloss and Handelsman, 2005).

3. Results and discussion

3.1 Evaluation of the oral microbiota diversity using the V5 region of the 16S rRNA gene

To examine which region of the 16S rRNA gene would be possible to target with the short Illumina sequencing reads, we extracted various sections of aligned 16S rDNA sequences available for 753 species in the Human Oral Microbiome Database and submitted them to the RDP classifier with a 80% confidence cutoff. The entire V5 120-base region as well as the 59-base segments from its forward end lead to many fewer unclassified sequences than their V6-region counterparts. (Table 1). Therefore, the paired-end data from the ~82-base V5 region we amplified in the current study would provide a means to capture taxonomic information suitable for studying the microbial diversity with the Illumina technology, similar to that of the favored V1-3 and V6 regions which are used when longer sequence reads are possible.

We explored the microbial diversity of the pooled saliva and oropharyngeal swab samples from three individuals by targeting the 16S rDNA hypervariable V5 regions. Of 1,373,824 obtained reads, 1,237,319 [publicly available at the RAST repository (Meyer et al., 2008) under ID:4444448.3] passed the quality control. They were clustered in 377,275 distinct sequences most of which (330,815) were unique.

3.2. Taxonomic analysis of the oral microbiota

We analyzed the taxonomic composition and abundance of the oral bacterial community using GAST (Huse et al., 2008), the MG-RAST server (Meyer et al., 2008) and the RDP classifier (Wang et al., 2007). RDP's Seqmatch program may also eventually be useful for determining which sequences in the RDP database are most closely related to our sequences, it works for sequences as short as 7 bases but only for 2000 sequences at a time.

The mean RDP confidence level for the six taxonomic levels from domain to genus was calculated as a function of the sequence abundance. The confidence decreases as the sequence copy number decreases (Fig. 1). This general trend is most likely due to the fact that the most frequent sequences correspond to known species whose 16S rDNA sequences

are available. Conversely, the rare species include a higher proportion of 16S rDNA sequences absent or distant from those in the RDP reference database. In addition, the probability that a sequence contains an error is expected to be higher in low frequency sequences (Andersson et al., 2008).

To limit the impact of sequencing errors, we removed all sequences that occurred less than three times. This new dataset contains 865,540 reads representing 25,978 distinct sequences. We discarded 381 reads (< 0.05%) with a GAST distance that diverged more than 30% from their nearest reference sequence, leaving 865,159 sequences. For the MG-RAST analysis with a minimum alignment length of 50 and a maximum BLAST e-value of 10^{-10} 21,713 sequences (2.5%) were removed, leaving 843,827 sequences. The phylogenetic assignments using the RDP classifier were performed after two additional filtering steps. They included the removal of sequences that were better classified when considered as reverse complements and those that had <80% confidence at the domain level. In this way the number of reads was reduced to 854,968 (24,757 distinct sequences).

The combined saliva and oropharyngeal swab samples were dominated by the phyla *Firmicutes*, *Proteobacteria*, *Actinobacteria*, *Fusobacteria*, *TM7* and *Spirochaetes* (Table 2), that are also abundant in other oral samples assessed by means of phyloarrays (Huyghe et al., 2008) or massively parallel pyrosequencing of the 16S rDNA clones or amplicons (Keijser et al., 2008; Nasidze et al., 2009a). Their proportions, however, differ in different studies (Table 2). Removing the 80% confidence cut-off in the RDP classification results in phyla breakdown that are very similar between RDP and GAST (data not shown); however, bootstrap values of less than 80% cannot be trusted. The MG-RAST server and RDP classifer returned a higher fraction of unclassified bacteria, likely because they are not well designed for such short sequences. This may explain the lower content of major phyla relative to that generated by GAST which was designed specifically for short tag sequences. *Proteobacteria* is an exception since their abundances were similar with the three classification tools. Therefore, the RDP- and MG-RAST-based classification of V5 rDNA sequences appeared to be more sensitive for *Proteobacteria* than for other major phyla. Indeed, the RDP classification of the HOMD 16S rDNA sequences showed that the relative abundance of *Proteobacteria*, in contrast to those of other major phyla, was not reduced when using 59-base V5 sequences instead of their full length counterparts (Table 1).

The ability to identify taxa from class down to the genus level varied between phyla and was dependent on the classification approach (Fig. 2). For the six major phyla, GAST generated the highest proportion of reads placed at these levels of taxonomy. Fusobacteria and Spirochaetes had the largest proportion of reads that can be confidently placed at the genus level using all three classification approaches. This proportion was the lowest for Proteobacteria despite their robust classification at the phylum level (see above).

Some consider organisms with more than 1.3% sequence difference in 16S rDNA sequence (based on the full-length) to belong to different species (Stackebrandt and Ebers, 2006). Since a single nucleotide difference in a 59-base-long sequence corresponds to a 1.7% resolution, there may be more than 25,000 species-level phylotypes in our dataset (Fig. 3). For a more conservative estimate of species-level phylotypes, we used a cutoff of 3% corresponding to a 2-base resolution (Stackebrandt and Goebel, 1994) to create clusters of sequences. There are at least 8,000 different phylotypes at the 3% level. This will be an underestimate since we removed all sequences occurring less than three times prior to analysis. These filtered sequences would include valid but rare organisms as well as many low-quality sequences.

We used rarefaction analysis to determine the microbial diversity recovery in the filtered dataset. The rarefaction curve is very stable at ~8,000 (Fig. 3), suggesting that the sampling completeness is high – at least 30,000 additional reads would be required to discover a new unique phylotype, and more than 120,000 additional reads would be required to discover a new 3% phylotype. The removal of unique sequences impacts the rarefaction curve, and may underestimate the potential for new species detection in human saliva samples. However, sampling is sufficient among the sequences likely to be prevalent in human saliva because they were found at least 3 times.

A total of 135 genera or next higher taxonomic ranks were identified by GAST (Supplementary material). The most frequent genera were *Neisseria* and *Streptococcus*, constituting about 70% of the sequences. Thirty-four taxa have not been identified in previous studies of oral microbiotas (Keijser et al., 2008; Nasidze et al., 2009a; Nasidze et al., 2009b) and are not listed in the Human Oral Microbiome Database. They include some low-

abundance genera as well as putative members of the candidate divisions BRC1, OP10, OP3. The MG-RAST server also identified BRC1 and OP10 sequences.

The observed relative low abundance in *Bacteroidetes* in our data compared to previous studies may be accounted for by many factors including sampling from different anatomical sites, individual variation, sample size, as well as potential bias in lysis, amplification and classification. Indeed, it has been shown that some of the "frequent" species are absent in some individuals (Aas et al., 2005). Good oral hygiene is known to decrease the proportion of Gram-negative bacteria including some *Bacteroidetes* species . The amplification bias has been invoked to explain a decline of *Bacteroidetes* in a metagenomic study of a series of fecal samples (Andersson et al., 2008). This is unlikely to be the case in our study since the PCR primers used cover 104 of 107 (97%) *Bacteroidetes* species listed in the HOMD.

To the best of our knowledge this is the first metagenomic study based on the utilization of the Illumina high-throughput sequencing technology. Illumina sequencing provides more sequence reads per run, allowing for more in-depth coverage than the competing technologies. This enables analysis of larger sample sizes, inclusion of more bar-coded time-points and samples, and better assessment of total diversity in microbial communities. Metagenomic studies of other human microbial communities in the gut, stomach and skin have shown that it is not clear whether a core community of bacteria is common to most humans, making the less common species important to understanding human health and disease (Hamady and Knight, 2009).

The advantage of generating and sequencing short 16S rDNA amplicons for bacterial community analysis is that it reduces the likelihood of generating chimera and increases the likelihood of detecting low-abundance taxa (Huber et al., 2009). Moreover, reads of 100-200 bases obtained with carefully chosen amplification primers can yield the same clustering as long 16S rDNA sequences (Liu et al., 2007).

There is a concern that short read length may compromise the classification quality. However, for the current Illumina read length using the V5 region of the 16s rDNA, taxonomic assessment at the phylum level is sufficient to effectively compare samples. The taxonomic analysis based on the Illumina technology will be improved by paired-end reads

(Table 1) which are expected not only to generate longer sequences but also to increase the sequence quality. Since the probability of a sequencing error increases with the read length (Qu et al., 2009), partially overlapping complementary reads of the same amplicon may help to predict sequencing errors and aid the removal of ambiguous reads or parts of reads.

Acknowledgements

This work was supported by grants from the Swiss National Science Foundation 3100A0-112370/1 (JS) and 3100A0-116075 (PF), and United States National Institutes of Health grant UH2DK083993 (SH).

References

Aas, J.A., Paster, B.J., Stokes, L.N., Olsen, I., Dewhirst, F.E., 2005. Defining the normal bacterial flora of the oral cavity. J. Clin. Microbiol. 43, 5721-5732.

Andersson, A.F., Lindberg, M., Jakobsson, H., Backhed, F., Nyren, P., Engstrand, L., 2008. Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLoS ONE 3, e2836.

Avila, M., Ojcius, D.M., Yilmaz, O., 2009. The oral microbiota: living with a permanent guest. DNA Cell Biol. 28, 405-411.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., McGarrell, D.M., Garrity, G.M., Tiedje, J.M., 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res. 33, D294-296.

Cox-Foster, D.L., Conlan, S., Holmes, E.C., Palacios, G., Evans, J.D., Moran, N.A., Quan, P.L., Briese, T., Hornig, M., Geiser, D.M., Martinson, V., vanEngelsdorp, D., Kalkstein, A.L., Drysdale, A., Hui, J., Zhai, J., Cui, L., Hutchison, S.K., Simons, J.F., Egholm, M., Pettis, J.S., Lipkin, W.I., 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318, 283-287.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L., 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72, 5069-5072.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792-1797.

Faveri, M., Mayer, M.P., Feres, M., de Figueiredo, L.C., Dewhirst, F.E., Paster, B.J., 2008. Microbiological diversity of generalized aggressive periodontitis by 16S rRNA clonal analysis. Oral Microbiol. Immunol. 23, 112-118.

Fraser, C.M., Eisen, J., Fleischmann, R.D., Ketchum, K.A., Peterson, S., 2000. Comparative genomics and understanding of microbial biology. Emerg. Infect. Dis. 6, 505-512.

Friedrich, M.J., 2008. Microbiome project seeks to understand human body's microscopic residents. JAMA 300, 777-778.

Hamady, M., Knight, R., 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res. 19, 1141-1152.

Huber, J.A., Morrison, H.G., Huse, S.M., Neal, P.R., Sogin, M.L., Mark Welch, D.B., 2009. Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. Environ. Microbiol. 11, 1292-1302.

Huse, S.M., Dethlefsen, L., Huber, J.A., Welch, D.M., Relman, D.A., Sogin, M.L., 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet. 4, e1000255.

Huyghe, A., Francois, P., Charbonnier, Y., Tangomo-Bento, M., Bonetti, E.J., Paster, B.J., Bolivar, I., Baratti-Mayer, D., Pittet, D., Schrenzel, J., 2008. Novel microarray design strategy to study complex bacterial communities. Appl. Environ. Microbiol. 74, 1876-1885.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059-3066.

Keijser, B.J., Zaura, E., Huse, S.M., van der Vossen, J.M., Schuren, F.H., Montijn, R.C., ten Cate, J.M., Crielaard, W., 2008. Pyrosequencing analysis of the oral microflora of healthy adults. J. Dent. Res. 87, 1016-1020.

Kuehbacher, T., Rehman, A., Lepage, P., Hellmig, S., Folsch, U.R., Schreiber, S., Ott, S.J., 2008. Intestinal TM7 bacterial phylogenies in active inflammatory bowel disease. J. Med. Microbiol. 57, 1569-1576.

Lipkin, W.I., 2009. Microbe hunting in the 21st century. Proc. Natl. Acad. Sci. USA 106, 6-7.

Liu, Z., Lozupone, C., Hamady, M., Bushman, F.D., Knight, R., 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. 35, e120.

Mager, D.L., Haffajee, A.D., Devlin, P.M., Norris, C.M., Posner, M.R., Goodson, J.M., 2005. The salivary microbiota as a diagnostic indicator of oral cancer: a descriptive, non-randomized study of cancer-free and oral squamous cell carcinoma subjects. J. Transl. Med. 3, 27.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., Edwards, R.A., 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9, 386.

Nakamura, S., Maeda, N., Miron, I.M., Yoh, M., Izutsu, K., Kataoka, C., Honda, T., Yasunaga, T., Nakaya, T., Kawai, J., Hayashizaki, Y., Horii, T., Iida, T., 2008. Metagenomic diagnosis of bacterial infections. Emerg. Infect. Dis. 14, 1784-1786.

Nasidze, I., Li, J., Quinque, D., Tang, K., Stoneking, M., 2009a. Global diversity in the human salivary microbiome. Genome Res. 19, 636-643.

Nasidze, I., Quinque, D., Li, J., Li, M., Tang, K., Stoneking, M., 2009b. Comparative analysis of human saliva microbiome diversity by barcoded pyrosequencing and cloning approaches. Anal. Biochem. 391, 64-68.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., Glockner, F.O., 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35, 7188-7196.

Qu, W., Hashimoto, S., Morishita, S., 2009. Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. Genome Res. 19, 1309-1315.

Raghavendran, K., Mylotte, J.M., Scannapieco, F.A., 2007. Nursing home-associated pneumonia, hospital-acquired pneumonia and ventilator-associated pneumonia: the contribution of dental biofilms and periodontal inflammation. Periodontol. 2000 44, 164-177.

Schloss, P.D., Handelsman, J., 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl. Environ. Microbiol. 71, 1501-1506.

Sibley, C.D., Rabin, H., Surette, M.G., 2006. Cystic fibrosis: a polymicrobial infectious disease. Future Microbiol. 1, 53-61.

Stackebrandt, E., Ebers, J., 2006. Taxonomic parameters revisited: tarnished gold standards. Microbiol. Today 33, 152-155.

Stackebrandt, E., Goebel, B.M., 1994. Taxonomic note: A place for DNA-DNA reassociation and 16s rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. 44, 846-849.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., Egholm, M., Henrissat, B., Heath, A.C., Knight, R., Gordon, J.I., 2009. A core gut microbiome in obese and lean twins. Nature 457, 480-484.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H., Smith, H.O., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66-74.

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261-5267.

Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C.S., Sutton, G., Frazier, M., Venter, J.C., 2008. The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 3, e1456.

Fig. 1

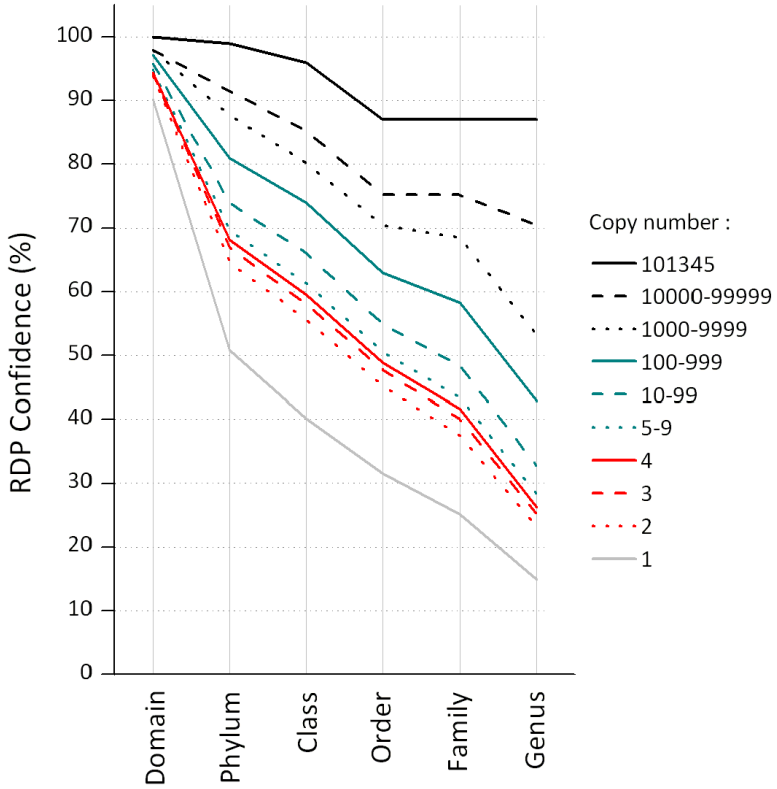

Fig. 2.

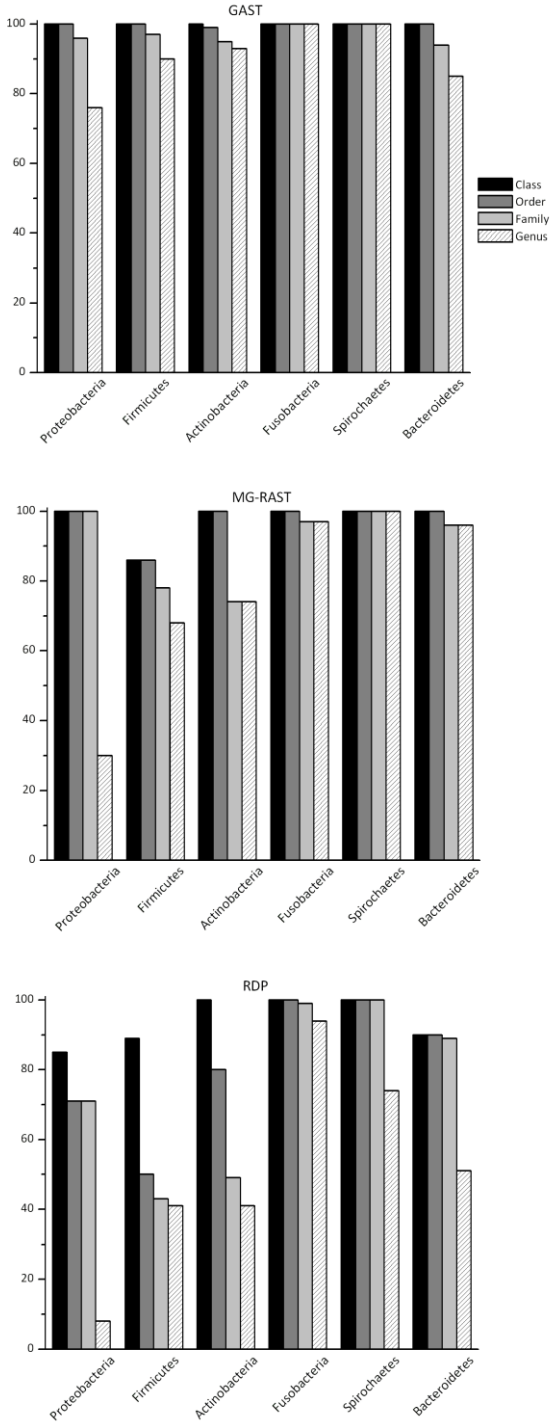

Fig. 3.

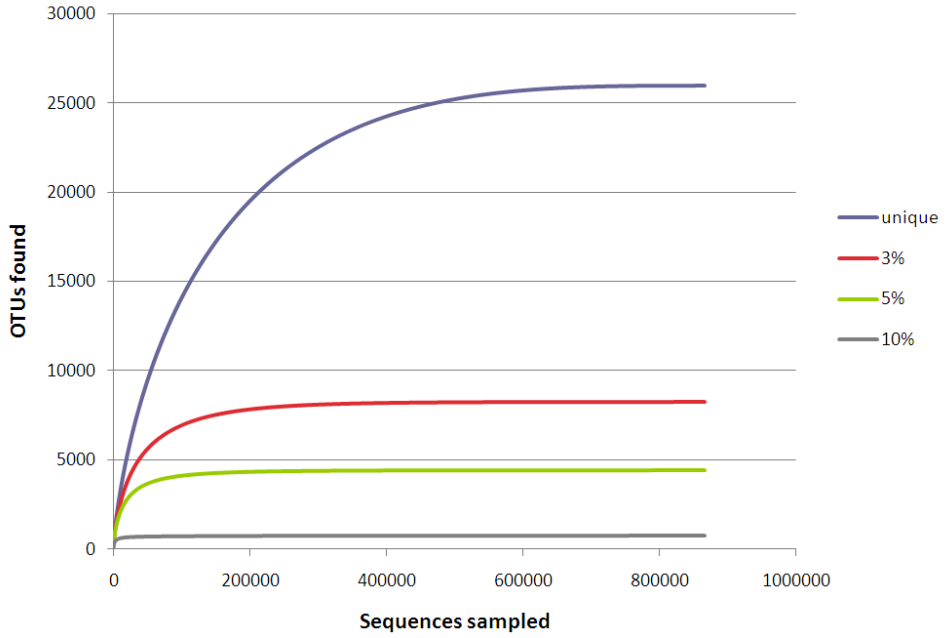

Table 1

RDP Classification of aligned segments of the 16S rRNA genes from the 753 sequences in the Human Oral Microbiome Database, using 80% confidence level cutoffs with the RDP Classifier software

		Percentage of sequences classified for different 16S rDNA regions						
	Position in 16S rDNA[a]	**8-1520**	**28-512**	**798-856**	**798-879**	**798-917**	**993-1051**	**932-1051**
	Variable region(s)	**All**	**V123**	**V5**	**V5**	**V5**	**V6**	**V6**
Phylum	**Sequence length (nt)**	**1513**	**485**	**59 F**[b]	**82**	**120**	**59 R**[b]	**120**
Firmicutes		38.6	36.8	28.6	32.9	35.7	25.3	32.3
Proteobacteria		15.1	15.1	15.1	15.0	15.1	14.8	14.7
Actinobacteria		10.8	10.8	9.7	10.4	10.6	7.1	10.4
Fusobacteria		5.7	5.7	4.7	5.7	5.7	3.1	4.6
Bacteroidetes		17.8	17.8	9.9	14.5	16.6	7.8	16.9
Spirochaetes		8.6	8.2	3.5	5.8	8.1	3.9	5.2
unclassified		0.1	2.4	28	13.8	5.6	35.8	13.7

[a] Numbering corresponds to the *E. coli* 16S rRNA gene sequence.
[b] F, from the forward end; R, from the reverse end

Table 2

Comparison of oral phyla abundance obtained using different classification tools

Phylum	Percentage of total sequences classified			
	Saliva + throat swab			**Saliva** (Keijser et al., 2008)
	GAST[a]	**RAST**[b]	**RDP**[c]	
Firmicutes	61.2	28.6	39.3	40.7
Proteobacteria	29.3	29.9	27.8	21
Actinobacteria	4.9	4.1	2.8	6.3
Fusobacteria	2.9	0.93	1.6	2.9
TM7	1.2	1.2	0.0063	1.9
Bacteroidetes	0.027	0.026	0.021	27.2
BRC1	0.07	0.076	0	0
OP10	0.12	0.040	0	0
Spirochaetes	0.20	0.20	0.064	0.2
Cyanobacteria	0.0007	0.0020	0.0016	0.020
SR1	0.013	0	0.0013	0.014
Acidobacteria	0.0018	0	0	0.049
OP3	0.0006	0	0	0
Unclassified Bacteria	0.023	35	28.5	0.2

[a] $\leq$ 30% divergent from their nearest reference sequence

[b] Maximum e-value 10^{-10} and minimum alignment length of 50 nucleotides

[c] 80% confidence cutoff applied at the phylum level

[d] NR, not reported

Supplementary Material

Relative abundance of taxa identified by GAST

Phylum	Taxon	Rank if higher than genus[a]	Relative amount of total (%)
Firmicutes	*Streptococcus*		48.613
Proteobacteria	*Neisseria*		12.471
Proteobacteria	*Haemophilus*		9.180
Proteobacteria	*Pasteurellaceae*	Family	5.748
Firmicutes	*Enterococcaceae*	Family	2.564
Fusobacteria	*Leptotrichia*		2.056
Firmicutes	*Veillonellaceae*	Family	1.961
Actinobacteria	*Actinomyces*		1.921
Firmicutes	*Abiotrophia*		1.799
Actinobacteria	*Atopobium*		1.624
Firmicutes	*Gemella*		1.557
Proteobacteria	*Burkholderiales*	Order	1.165
TM7	*TM7*	Phylum	1.152
Firmicutes	*Lactobacillales*	Order	1.076
Actinobacteria	*Arthrobacter*		1.012
Fusobacteria	*Fusobacterium*		0.838
Firmicutes	*Granulicatella*		0.793
Firmicutes	*Enterococcus*		0.562
Firmicutes	*Selenomonas*		0.511
Firmicutes	*Bulleidia*		0.417
Proteobacteria	*Campylobacter*		0.309
Firmicutes	*Clostridiales*	Order	0.253
Actinobacteria	*Actinomycetales*	Order	0.211
Spirochaetes	*Treponema*		0.197
Firmicutes	*Caryophanon**		0.176
Firmicutes	*Anaerovorax*		0.163
Proteobacteria	*Neisseriaceae*	Family	0.133
Proteobacteria	*Pasteurella*		0.127
OP10	*OP10**	Phylum	0.123
Firmicutes	*Clostridia*	Class	0.109
BRC1	*BRC1**	Phylum	0.070
Firmicutes	*Bacilli*	Class	0.066
Firmicutes	*Bacillus*		0.065
Actinobacteria	*Micrococcaceae*	Family	0.064
Firmicutes	*Parvimonas*		0.061
Actinobacteria	*Actinobacteria*	Class	0.047
Proteobacteria	*Actinobacillus*		0.046
Firmicutes	*Facklamia**		0.045

Proteobacteria	*Kingella*		0.043
Firmicutes	*Mogibacterium*		0.042
Firmicutes	*Exiguobacterium**		0.040
Proteobacteria	*Cardiobacterium*		0.036
Proteobacteria	*Aggregatibacter*		0.034
Firmicutes	*Firmicutes*	Phylum	0.026
Firmicutes	*Carnobacteriaceae*	Family	0.026
Firmicutes	*Bacillales*	Order	0.026
Firmicutes	*Mitsuokella*		0.025
Unclassified	*Bacteria*	Domain	0.023
Firmicutes	*Veillonella*		0.023
Firmicutes	*Trichococcus**		0.021
Firmicutes	*Acidaminococcus*		0.020
Proteobacteria	*Mannheimia*		0.020
Firmicutes	*Carnobacterium*		0.017
Fusobacteria	*Cetobacterium*		0.016
Firmicutes	*Tetragenococcus*		0.015
Actinobacteria	*Propionibacterium*		0.015
Bacteroidetes	*Porphyromonas*		0.014
Firmicutes	*Megamonas**		0.014
Firmicutes	*Syntrophomonadaceae**	Family	0.014
Actinobacteria	*Coriobacteriaceae*	Family	0.013
Firmicutes	*Lactobacillus*		0.013
SR1	*SR1*	Phylum	0.013
Firmicutes	*Seinonella**		0.012
Proteobacteria	*Comamonadaceae*	Family	0.010
Firmicutes	*Filifactor*		0.009
Firmicutes	*Eubacteriaceae*	Family	0.009
Actinobacteria	*Renibacterium**		0.008
Bacteroidetes	*Prevotella*		0.006
Actinobacteria	*Propionibacteriaceae*	Family	0.006
Firmicutes	*Leuconostoc**		0.005
Actinobacteria	*Corynebacterium*		0.005
Firmicutes	*Alloiococcus*		0.005
Actinobacteria	*Micromonosporaceae**	Family	0.005
Actinobacteria	*Kocuria*		0.005
Firmicutes	*Staphylococcus*		0.004
Firmicutes	*Peptostreptococcaceae*	Family	0.004
Proteobacteria	*Nicoletella**		0.004
Firmicutes	*Halobacillus**		0.004
Firmicutes	*Lachnospiraceae*	Family	0.003
Firmicutes	*Moryella*		0.003
Firmicutes	*Aerococcaceae*	Family	0.003
Firmicutes	*Jeotgalicoccus**		0.003
Firmicutes	*Succiniclasticum**		0.003
Firmicutes	*Megasphaera*		0.002
Actinobacteria	*Nesterenkonia*		0.002

Firmicutes	*Peptoniphilus*		0.002
Bacteroidetes	*Flavobacteriaceae*	Family	0.002
Fusobacteria	*Sneathia*		0.002
Acidobacteria	*Acidobacteria*	Phylum	0.002
Proteobacteria	*Acinetobacter*		0.002
Proteobacteria	*Tetrathiobacter**		0.002
Firmicutes	*Melissococcus**		0.002
Actinobacteria	*Rothia*		0.002
Bacteroidetes	*Bacteroidales*	Order	0.002
Firmicutes	*Ruminococcaceae*	Family	0.001
Firmicutes	*Alkalibacterium*		0.001
Fusobacteria	*Fusobacteriaceae*	Family	0.001
Firmicutes	*Lactococcus*		0.001
Proteobacteria	*Proteobacteria*	Phylum	0.001
Proteobacteria	*Alysiella*		0.001
Firmicutes	*Butyrivibrio*		<0.001
Proteobacteria	*Volucribacter**		<0.001
Bacteroidetes	*Chryseobacterium*		<0.001
Bacteroidetes	*Flavobacterium**		<0.001
Firmicutes	*Globicatella**		<0.001
Proteobacteria	*Moraxella*		<0.001
Actinobacteria	*Actinobaculum*		<0.001
Firmicutes	*Aerococcus**		<0.001
Cyanobacteria	*Cyanobacteria*	Phylum	<0.001
Proteobacteria	*Gammaproteobacteria*	Class	<0.001
Proteobacteria	*Pseudomonadales*	Order	<0.001
Bacteroidetes	*Capnocytophaga*		<0.001
Proteobacteria	*Desulfobulbus*		<0.001
Proteobacteria	*Marinomonas**		<0.001
OP3	*OP3**	Phylum	<0.001
Firmicutes	*Oribacterium*		<0.001
Actinobacteria	*Rubrobacteraceae**	Family	<0.001
Proteobacteria	*Avibacterium**		<0.000
Actinobacteria	*Brevibacteriaceae**	Family	<0.001
Actinobacteria	*Cryptobacterium*		<0.001
Bacteroidetes	*Prevotellaceae*	Family	<0.001
Actinobacteria	*Streptosporangiaceae**	Family	<0.001
Actinobacteria	*Actinomycetaceae*	Family	<0.001
Firmicutes	*Bacillaceae*	Family	<0.001
Proteobacteria	*Betaproteobacteria*	Class	<0.001
Firmicutes	*Desulfotomaculum**		<0.001
Firmicutes	*Kurthia**		<0.001
Actinobacteria	*Microbispora**		<0.001
Proteobacteria	*Nitrosomonas**		<0.001
Actinobacteria	*Olsenella*		<0.001
Proteobacteria	*Pseudomonas*		<0.001
Actinobacteria	*Pseudonocardia**		<0.001

Proteobacteria	*Simonsiella*		<0.001
Proteobacteria	*Sphingomonadaceae*	Family	<0.001
Firmicutes	*Turicibacter**		<0.001

Blank corresponds to genus. Taxa that have not been listed in previous large-scale bacterial oral community studies (Keijser et al., 2008; Nasidze et al., 2009a; Nasidze et al., 2009b) and the Human Oral Microbiome database are marked by an asterisk.

[a] The lowest taxonomic rank to which the sequence was assigned.

Figure 1. Average confidence level for the six taxonomic levels as a function of sequence counts.

Figure 2. Proportions of taxonomic assignments under the phylum level. Reads assigned to each of the four taxonomic levels for each major phylum are represented by bars. Their height represent the percentage of reads that can be placed at a given level of taxonomy using GAST, the MG-RAST server and the RDP Classifier.

Figure 3. Rarefaction analysis of the oral metagenome. The curves include only sequences which occur 3 or more times. Number of OTUs with different cutoff values was plotted as a function of the number of sequences sampled. OTUs with ≥97%, ≥95% and ≥90% pairwise sequence identity are arbitrarily assumed to form the same species, genus and family, respectively.