The ISME Journal (2010) 4, 962–974

www.nature.com/ismej

ORIGINAL ARTICLE

Bacterial diversity in the oral cavity of 10 healthy individuals

Elisabeth M Bik[1], Clara Davis Long[1], Gary C Armitage[2], Peter Loomer[2], Joanne Emerson[3,7], Emmanuel F Mongodin[4], Karen E Nelson[3], Steven R Gill[5], Claire M Fraser-Liggett[4] and David A Relman[1,6]

[1]*Departments of Microbiology & Immunology, and of Medicine, Stanford University School of Medicine, Stanford, CA, USA;* [2]*School of Dentistry, University of California, San Francisco, CA, USA;* [3]*J Craig Venter Institute, Rockville, MD, USA;* [4]*Institute For Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA;* [5]*Department of Oral Biology, State University of New York, Buffalo, NY, USA and* [6]*Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA*

The composition of the oral microbiota from 10 individuals with healthy oral tissues was determined using culture-independent techniques. From each individual, 26 specimens, each from different oral sites at a single point in time, were collected and pooled. An 11th pool was constructed using portions of the subgingival specimens from all 10 individuals. The 16S ribosomal RNA gene was amplified using broad-range bacterial primers, and clone libraries from the individual and subgingival pools were constructed. From a total of 11 368 high-quality, nonchimeric, near full-length sequences, 247 species-level phylotypes (using a 99% sequence identity threshold) and 9 bacterial phyla were identified. At least 15 bacterial genera were conserved among all 10 individuals, with significant interindividual differences at the species and strain level. Comparisons of these oral bacterial sequences with near full-length sequences found previously in the large intestines and feces of other healthy individuals suggest that the mouth and intestinal tract harbor distinct sets of bacteria. Co-occurrence analysis showed significant segregation of taxa when community membership was examined at the level of genus, but not at the level of species, suggesting that ecologically significant, competitive interactions are more apparent at a broader taxonomic level than species. This study is one of the more comprehensive, high-resolution analyses of bacterial diversity within the healthy human mouth to date, and highlights the value of tools from macroecology for enhancing our understanding of bacterial ecology in human health.
The ISME Journal (2010) **4,** 962–974; doi:10.1038/ismej.2010.30; published online 25 March 2010
Subject Category: microbial population and community ecology
Keywords: oral microbiota; ribosomal RNA sequences; human microbial ecology

Introduction

The human body is home to many indigenous microorganisms, with distinct communities at different anatomical sites (Dethlefsen *et al.*, 2007). Recent studies have shown the importance of the gut microbiota in digestion, fat storage, angiogenesis, immune system development and response, colonization resistance and epithelial architecture (reviewed in Flint *et al.*, 2007; Tappenden and Deutsch, 2007; Cogen *et al.*, 2008). The oral cavity is also home to microbial communities, with important implications for human health and disease. Chronic periodontitis is one of the most common inflammatory conditions worldwide, and is associated with bacterial community structures that are distinct from those of health.

Efforts to characterize microbial diversity increasingly rely on cultivation-independent molecular techniques (Hugenholtz, 2002; Schloss and Handelsman, 2004), as the vast majority of bacteria have yet to be cultivated. Most of these molecular studies are based on the small subunit (16S) ribosomal RNA (rRNA) gene because of its universal presence in cellular organisms, the presence of conserved regions and its reliability for phylogenetic analysis (Woese and Fox, 1977). Recent molecular surveys of the human distal gut microbiota have shown that each individual gut is home to 500–3000 bacterial species, with a large degree of interindividual variation (Eckburg *et al.*, 2005; Dethlefsen *et al.*, 2007, 2008). Using rRNA gene-based techniques, it is

Correspondence: DA Relman, Departments of Microbiology & Immunology, and of Medicine, Stanford University, VA Palo Alto Health Care System 154T, 3801 Miranda Avenue, Palo Alto, CA 94304, USA.
E-mail: relman@stanford.edu
[7]Current address: Department of Earth and Planetary Science, University of California, Berkeley, CA, USA
Received 30 November 2009; revised 11 February 2010; accepted 11 February 2010; published online 25 March 2010

estimated that the human oral cavity harbors 500–700 different bacterial species (Kroes *et al.*, 1999; Paster *et al.*, 2001; Kazor *et al.*, 2003; Aas *et al.*, 2005; Dewhirst *et al.*, 2008). A recent study based on 14 115 partial 16S rRNA gene sequences obtained from saliva specimens from 120 healthy individuals from 12 different geographic locations around the world found 101 different bacterial genera, with a high level of interindividual variation (Nasidze *et al.*, 2009). Two recent 16S rRNA gene-tag pyrosequencing-based studies have suggested that there are ~250–300 species-level phylotypes in the mouth of any given individual, and that they segregate based on mucosal versus dental surfaces (Keijser *et al.*, 2008; Zaura *et al.*, 2009). All three of these recent studies are limited by their dependence on relatively short (<500 nucleotides) sequences, and hence, by limited phylogenetic resolution.

We analyzed ~1000 near full-length-cloned 16S rRNA gene sequences from each of 10 individuals with healthy oral tissues and gingiva, and examined variation in patterns of diversity between individuals.

Materials and methods

Subjects and specimen collection

Specimens were collected from 10 individuals with healthy oral tissues and gingiva (five women; age range 27–61 years; average age 38.1 years; ethnicity: six Caucasian, one Afro-American, two Chinese and one from India). Oral health status of all individuals was determined by a dentist who performed a full-mouth clinical examination that included inspection of the teeth, oral mucosa and periodontal tissues. All participants had normal oral mucous membranes and were free from nonrestored carious lesions. At most sites, periodontal tissues showed no clinical signs of inflammation, such as redness, swelling or bleeding on probing, and were judged to be free of gingivitis or periodontitis. Details of the periodontal data obtained from sites from which plaque specimens were taken are provided in Table 1 and Supplementary Methods. From each individual, 26 oral specimens were collected. Separate dental plaque specimens were taken with sterile curettes from supragingival and subgingival surfaces of seven target teeth (no. 3, 9, 12, 19, 25, 28 and 30). The 26th sample consisted of whole saliva that was expectorated into a test tube. Healthy human mouths have relatively little bacterial biomass compared with the gastrointestinal tract; therefore, because the ultimate purpose of this project was to obtain community-wide shotgun sequence data, specimens were pooled in order to ensure sufficient DNA. One-third of each of the 26 specimens obtained from each individual was combined to obtain 10 'individual-specific' pools, whereas a separate third of each of the subgingival specimens from all 10 individuals (seven specimens per subject) was pooled to create a single 'subgingival pool'. To study the influence of DNA isolation method in the UniFrac analysis (see below), specimens were also collected from three additional healthy individuals. These specimens were not included in downstream analyses, unless otherwise noted. Further details about inclusion and exclusion criteria, specimen collection and other procedures are provided in Supplementary Methods.

Table 1 Characteristics of subjects who participated in this study

Subject ID	*Gender*	*Age*	*Ethnicity*	*CAL*	*BOP*
1	F	61	Caucasian	0.090	0
2	M	42	Caucasian	0.262	0
3	F	49	Afro-American	0.400	1.3
4	F	33	Chinese	0.173	0
5	M	29	Chinese	0.153	0
6	M	29	Indian-Asian	0.268	0
7	M	27	Caucasian	0.196	0
8	M	42	Caucasian	0.525	0
9	F	37	Caucasian	0.278	0
10	F	32	Caucasian	0.232	0

Abbreviations: BOP, percentage of teeth that displayed bleeding upon probing; CAL, mean clinical attachment loss.
Subjects were considered periodontally healthy if mean CAL was ⩽0.6 mm and if ⩽2% of sites displayed BOP.

DNA extraction

To extract DNA, pooled specimens were washed twice in 1 ml ice-cold phosphate-buffered saline, pelleted by 5 min centrifugation at 16 000 *g* at 4 °C and resuspended in 100 µl phosphate-buffered saline, to reduce the amount of contaminating free human DNA. To this suspension, 10 µl of a 10% Triton-X100 solution and 2.5 µl of a 20 mg ml^{-1} Proteinase K solution (Qiagen, Valencia, CA, USA) were added, and the suspension was incubated at 60 °C for 30 min. A volume of 200 µl of a cell lysis buffer (100 mM Tris-HCl (pH 7.4), 20 mM EDTA, 5 M guanidine isothiocyanate) was added. To obtain maximum bacterial diversity, we split each specimen pool into two equal portions. To one specimen portion, three sizes of baked zirconia beads were added, and the mixture was agitated in a FastPrep FP120 machine (Qbiogene, Carlsbad, CA, USA) at 4.0 m s^{-1} for 30 s. The bead-beaten portion was recombined with the nonbead-beaten portion. The DNA was further purified, precipitated, washed, dried and resuspended in 50 µl of 10 mM Tris (pH 8.0) (details are provided in the Supplementary Methods). Extraction controls were processed in parallel during the DNA extraction procedure to monitor contamination. A second set of pooled oral specimens from three additional healthy mouths was extracted using the QIAamp DNA mini kit (Qiagen).

16S rRNA gene amplification, cloning and sequencing

The 16S rRNA gene was amplified using broad-range bacterial-specific primers 8FM (5′-AGAGT TTGATCMTGGCTCAG-3′) (Edwards *et al.*, 1989;

Palmer *et al*., 2007) and 1391R (5′-GACGGGCGGT GTGTRCA-3′) (Lane *et al*., 1985; Palmer *et al*., 2007). These primers amplify ~90% of the bacterial 16S rRNA coding sequence. PCR was performed as described previously (Eckburg *et al*., 2005), except that PCRs were performed with 5 min at 95 °C, 20 cycles of 30 s at 94 °C, 30 s at 55 °C and 90 s at 72 °C, followed by 8 min at 72 °C. To obtain sufficient PCR product for cloning, the products of four replicate 20-cycle amplification reactions were pooled. No amplification product was observed in the extraction controls and negative PCR controls. Purified PCR products were cloned with the TOPO TA cloning kit (Invitrogen, Carlsbad, CA, USA), and plasmid inserts were sequenced on both strands.

Phylogenetic analysis

A total of 11 447 high-quality, ~1400 bp-length, 16S rRNA gene sequences were aligned with the online Greengenes NAST aligner (DeSantis *et al*., 2006) (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi) and inserted into the Greengenes version of ARB (Ludwig *et al*., 2004). The alignment was further perfected by manual optimization. In total, 79 chimeras (0.7%) were manually identified and removed from the analysis, so that 11 368 sequences were included in the final analysis. Operational taxonomic units (OTUs; phylotypes) were defined using a 99% sequence similarity cutoff, by using similarity matrices and a filter of 1253 nucleotide positions, masking out the hypervariable regions. The 99% cutoff in this setting roughly corresponds to species-level groupings. One representative for each of the 247 OTUs found in this study was deposited in GenBank (accession numbers FJ976202 to FJ976448) (Supplementary Table S1). Sequences with less than 99% similarity to sequences in public databases were considered novel (Supplementary Table S2). Genus names were assigned based on placement of sequences within defined groups, or on a cutoff of 95% sequence identity in the case of unclassifiable sequences. The DOTUR and mothur packages were used to calculate the number of OTUs at different cutoffs, and to calculate collector's curves and the Chao1 species richness (Schloss and Handelsman, 2005; Schloss *et al*., 2009).

Estimates of microbial diversity

Richness estimates and diversity indices were determined (Simpson and Shannon formulae) with EstimateS (Colwell, 2005). The percentage of coverage was calculated by Good's method using the formula $[1-(n/N)] \times 100$, where n is the number of phylotypes in a specimen represented by one clone (singletons) and N is the total number of sequences in that specimen (Good, 1953). The Shannon index of evenness was calculated using the formula $E = e^D/N$, where D is the Shannon diversity index.

UniFrac analysis

After calculating, with an Olsen correction, a neighbor-joining tree containing representatives of all 247 OTUs found in this study, the 11 different oral environments were clustered using principal coordinates analysis, as enabled in UniFrac (Lozupone *et al*., 2006), using weighted, normalized abundance data. To compare sequence data obtained from oral specimens in this study against data obtained from other locations in the human body from subjects in previously published studies, UniFrac principal coordinates analysis was also performed on a second data set. These combined data included data obtained from the 11 oral pools of this study, 3 additional oral pools from healthy human mouths isolated using a different DNA extraction method (QIAamp DNA mini kit from Qiagen; 1034 sequences; unpublished data), 18 colonic biopsy and 3 stool specimens from 3 healthy subjects (Eckburg *et al*., 2005) and 15 stool specimens from 3 healthy subjects in an antibiotic perturbation study (Dethlefsen *et al*., 2008).

Community comparisons

Community composition was examined in two separate ways. First, the communities were compared using shared species estimators. Second, community assembly was examined using taxon co-occurrence. The Chao–Jaccard abundance-based similarity index is a shared species estimator that measures the probability that two individuals chosen from two different specimens are members of species shared by both specimens (Chao *et al*., 2005). This particular test can only be used to examine similarity between two communities at a time. The Chao–Jaccard similarity index was calculated using EstimateS (Colwell, 2005) for all possible pairwise comparisons of the communities from the 10 mouths. Community similarity was compared at two taxonomic levels—OTU and genus. The subgingival pool was not included in this analysis.

In addition to community similarity, we tested for nonrandom patterns of taxon co-occurrence by calculating *C*-scores for this data set (Stone and Roberts, 1990). This measure of community structure calculates the number of checkerboard units (specimens in which two taxa are not found together) between all possible taxon pairs in a matrix, and calculates a single score for the entire data set. The *C*-score is the average for all of the possible pairs in the matrix. This measure is compared with a null distribution of random matrices of the same size. If the observed *C*-score is larger than the score for the null hypothesis, it suggests significant segregation between taxa, and if the observed *C*-score is smaller than the score for the null hypothesis, it suggests significant aggregation between taxa. In this case, we calculated *C*-scores using an abundance matrix of all taxa, organized by mouth (Supplementary Table S1), which was then

converted to a presence/absence matrix. The subgingival pool was not included in this analysis. These scores were compared with those generated from a null model based on 500 randomly generated matrices of the same size using the program EcoSim (Gotelli and Entsminger, 2004). Co-occurrence patterns were examined at three separate taxonomic levels—OTU level ($n = 247$, approximately species level), genus level ($n = 53$) and phylum level ($n = 9$).

Results

Bacterial diversity of the healthy human mouth

From each of 10 individuals with a healthy oral status, 26 specimens from different parts of the mouth were collected. Portions of the specimens were pooled per individual and an 11th pool was constructed with portions of each subgingival specimen from all 10 individuals. Ribosomal RNA gene sequences were amplified using broad-range bacterial primers, cloned and sequenced. The 11 368 near full-length, nonchimeric sequences of the combined data set were manually assigned to 247 OTUs (phylotypes) using a cutoff of 99% sequence identity (Supplementary Table S1). DOTUR and mothur analyses revealed a total of 228 OTUs at this cutoff level, with an expected OTU richness of 236 (Supplementary Figure S1, which also shows the rarefaction curves of each of the 11 clone libraries). A graph displaying the DOTUR-determined number of phylotypes versus the phylogenetic distance displayed the typical 'hockey stick shape' that is found in most animal-associated bacterial communities, with an enriched representation of diversity at the tip (Supplementary Figure S2). Nine bacterial phyla were identified within the combined data set (Figure 1). Of these, Firmicutes (33.2% of all sequences; mean abundance in 11 pools is $32.2 \pm 8.1\%$), Proteobacteria (27.5% in combined set; mean $24.6 \pm 8.1\%$), Bacteroidetes (16.6%; mean $14.6 \pm 8.4\%$) and Actinobacteria (14.5%; mean $11.9 \pm 10.3\%$) were the most abundant. Less abundant phyla included Fusobacteria (6.7%; mean 5.6 ± 4.1), TM7 (1.3%; $0.52 \pm 1.3\%$), as well as Spirochaetes, OD2 and Synergistes (all $<1\%$). Figure 2 displays a phylogenetic tree and relative abundance of all genera found in this study. In the combined data set, the genus *Streptococcus* was the most abundant genus (2180 sequences, 19.2% of total). Other abundant genera include *Haemophilus* (1325; 11.7%), *Neisseria* (1042; 9.2%), *Prevotella* (974; 8.6%), *Veillonella* (973, 8.6%) and *Rothia* (820; 7.2%). However, the genera and species that dominate the mouth vary between individuals (see below).

Novel sequences

Using a 1% sequence identity cutoff, 24 OTUs (10%) were considered novel (Supplementary Table S2). Of these, six had less than 97% sequence identity to published sequences (Table 2). The sequences with the least identity to previously reported sequences were clone 10B928 (phylum Bacteroidetes), which was 92.5% identical to AF371900 (isolated from the intestinal tract of a pig, (Leser *et al*., 2002)), and clone 7BB842 (phylum OD2), which displayed 92.5% sequence similarity to its closest neighbor, AB243989 (detected in a Japanese oil well (unpublished data)).

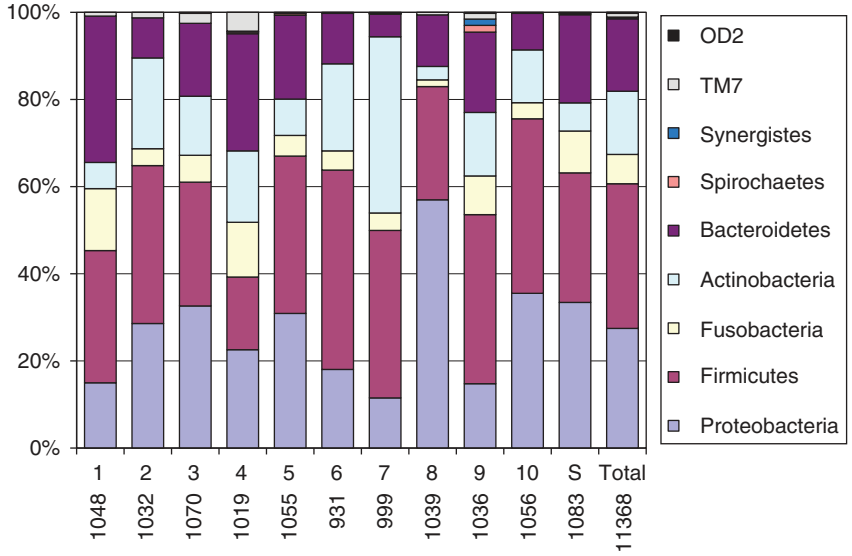

Figure 1 Relative abundance of phylum members of the oral communities from 10 healthy individuals. A total of 11 368 bacterial rRNA gene sequences derived from pools of specimens from different oral habitats, from each of 10 healthy individuals (numbered 1–10), as well as from a pool of all subgingival samples (S), were analyzed and assigned to phyla (color-coded, according to the scheme at the right). 'Total' refers to the combined set of sequences from all pools. The number of clones in each rRNA gene library is given below the name of the pool.

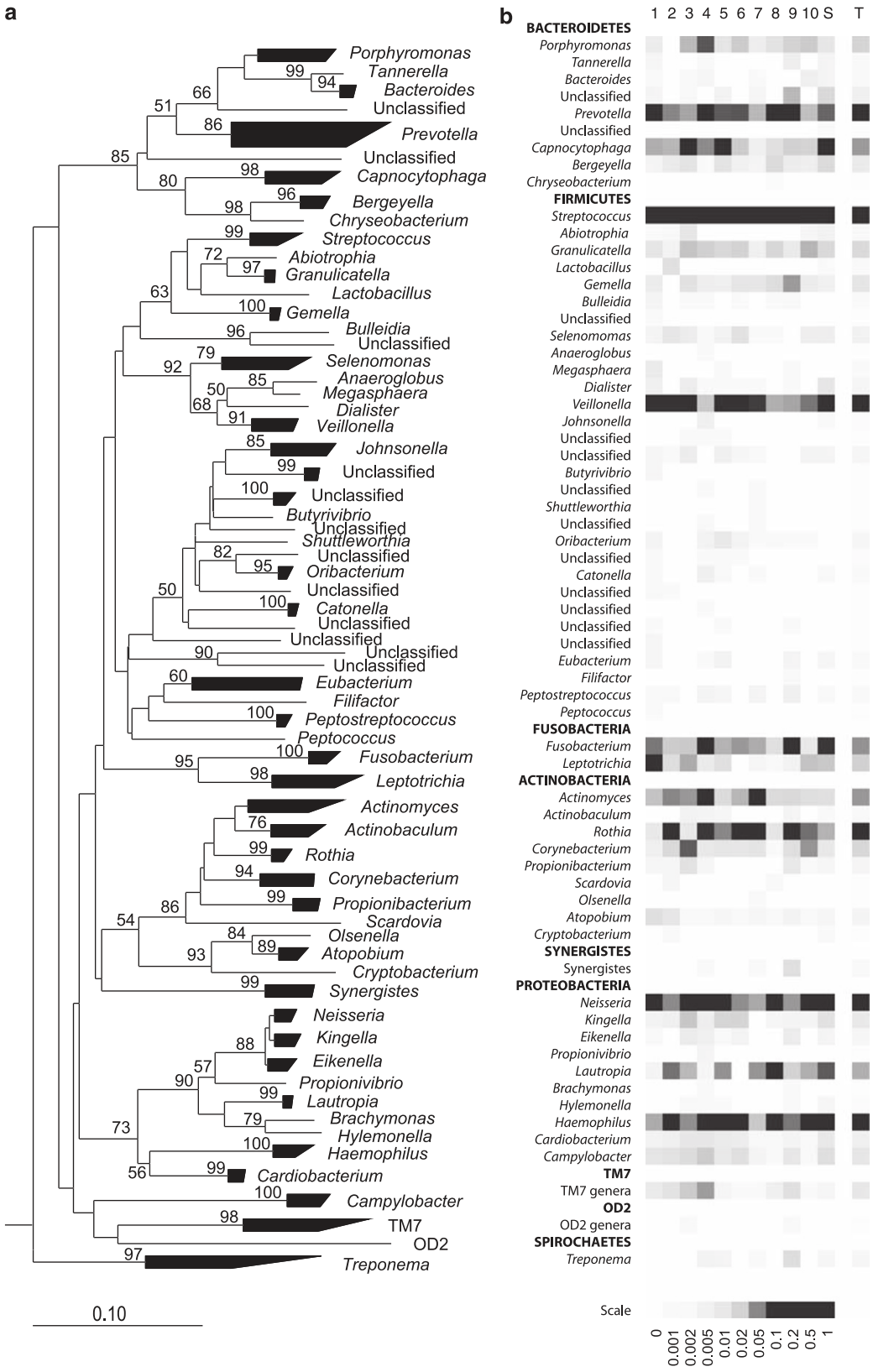

Figure 2 Phylogenetic relationships and relative abundance of the genera found in pools of oral specimens. (**a**) Phylogenetic tree for the 247 OTUs found in this study, grouped by genus. A 95% sequence similarity threshold was used for unclassified groups. The tree was constructed by neighbor-joining analysis with an Olsen correction. Bootstrap values $\geqslant 50$ (expressed as percentages of 1000 replicates) are shown at branch points. The scale bar represents evolutionary distance (10 substitutions per 100 nucleotides). (**b**) Relative abundance of genera in each of the 11 oral specimen pools displayed with gray scale values (white, 0% present; black, 100% of clone library; exact scale shown at the bottom). 1–10, each of the individual subject pools; S, subgingival, T, total. Genera are shown in the same order as in (**a**).

Table 2 Novel OTUs found in this study

Phylum	*Genus*	*OTU representative*	*GenBank accession number*	*Number of clones*	*Sequence identity (%)*[a]	*Closest published relative*
Bacteroidetes	*Prevotella*	10B928	FJ976326	1	92.49	AF371900
OD2	Unclassified	7BB842	FJ976283	4	92.51	AB243989
Fusobacteria	*Leptotrichia*	22B817	FJ976383	1	93.77	AY029807
TM7	Unclassified	7BB623	FJ976276	4	95.77	AY385520
Bacteroidetes	*Prevotella*	22B412	FJ976378	3	96.09	FJ577257
Actinobacteria	*Actinobaculum*	7BB627	FJ976277	2	96.73	AY349363

Abbreviation: OTU, Operational taxonomic unit.
Sequences shown here displayed <97% sequence identity to sequences available in public databases. For each novel OTU, the assigned GenBank accession number, the number of clones in the combined data set (11 368 sequences) and their closest published relative is shown. A list of all novel sequences using a cutoff of 99% is given in Supplementary Table S2.
[a]Sequence identity to closest published sequence longer than 1000 nucleotides.

Table 3 Estimators of sequence library diversity, evenness and coverage

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*S*	*Total*
Number of clones	1048	1032	1070	1019	1055	931	999	1039	1036	1056	1083	11 368
Number of OTUs	111	75	107	126	97	73	68	65	117	67	128	247
Number of singletons	28	17	25	49	24	22	26	24	35	13	47	39
Number of doubletons	13	9	15	15	18	8	9	7	20	9	20	20
Chao1 (richness)	138.0	88.6	125.8	199.5	111.5	98.7	100.5	99.5	145.3	74.8	179.5	282.3
Shannon (Diversity)	3.82	3.31	3.9	3.88	3.61	3.03	2.7	2.76	3.75	3.01	3.88	4.03
Simpson (Diversity)	26.13	15.81	30.51	28.92	21.94	9.31	8.20	7.12	16.95	10.27	27.35	26.24
Evenness	0.41	0.37	0.46	0.38	0.38	0.28	0.22	0.24	0.36	0.30	0.38	0.23
Good's estimator of coverage	97.33	98.35	97.66	95.19	97.73	97.63	97.40	97.69	96.62	98.77	95.66	99.66

Abbreviations: 1–10, individual pools; OTU, Operational taxonomic unit; S, subgingival pool.
Estimators were calculated using EstimateS for each of the 11 clone libraries described in this study, as well as for the total sequence set.
Evenness was calculated using the formula (EXP(Shannon))/Sobs, in which Sobs is the observed number of species (OTUs).
Good's estimator of coverage was calculated using the formula: $(1-(\text{singletons/individuals})) \times 100$.

Comparisons between bacterial communities in the 11 oral pools

Observed bacterial richness was highest in subject 4, in whom the highest number of OTUs, singletons and doubletons was found (Table 3). In contrast, both Shannon and Simpson estimators of bacterial diversity were the highest for subject 3. This subject also showed the highest Shannon estimator of evenness. Good's estimator suggested >95% coverage for each of the 11 libraries, indicating that only an additional five OTUs would be found if 100 additional clones were sequenced. UniFrac analysis showed no clustering of the oral communities from the 10 individuals based on gender, age or ethnic background (Figure 3a). Pairwise comparisons of the oral pools showed that all individuals were equally distinct (Bonferroni-corrected *P*-values all >0.5).

Microbiota from human oral cavity is distinct from that of other human habitats

We compared the oral bacterial communities described in this study with those found in previously published studies of the human colon and stool (Figure 3b). Although these specimens were derived from different studies and different individuals (except for certain stool and colonic specimens that were derived from the same three individuals), specimens from different anatomical sites clustered in a distinct fashion; the corrected UniFrac significance (all environments together) was $\leqslant 0.01$, indicating that the environments were significantly different from each other. Three additional oral communities from QIAamp-extracted specimens (CDL, unpublished results) clustered with the 11 communities from 11 benzyl alcohol-extracted specimen pools described in this study, suggesting that DNA extraction method accounts for less variation in the composition of communities than do differences between individuals.

Shared taxa among the bacterial communities of the healthy human mouth

The different bacterial communities were compared using the Chao–Jaccard abundance-based similarity index. An average of 50.5 (range 29–76) OTUs were found to be shared between any two specimens

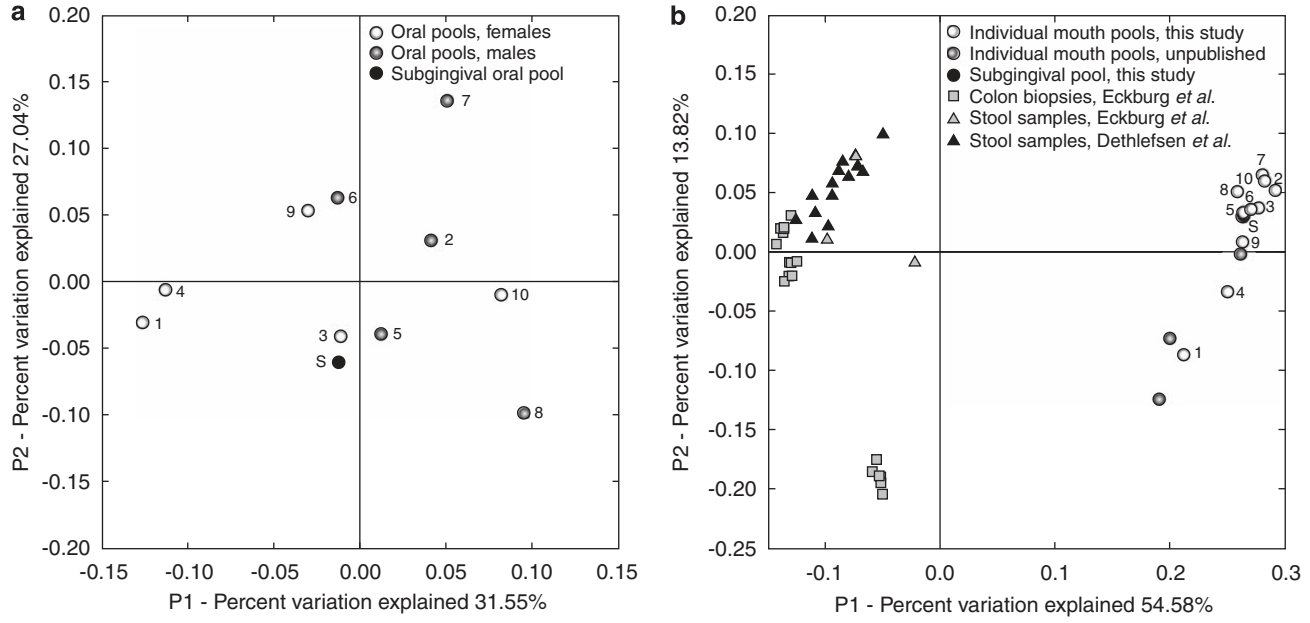

Figure 3 Variation in patterns of diversity. Unifrac principal components analysis was performed using weighted, normalized abundance data (Lozupone *et al.*, 2006). (**a**) Analysis of oral specimen pools from each of 10 healthy subjects (white circles, females, $n=5$; gray circles, males, $n=5$), and one pool of the subgingival specimens from all of these 10 subjects (black circle). (**b**) Analysis of the oral specimen pool data obtained from this study (white circles, $n=11$), data from additional oral specimen pools extracted with a different DNA extraction method (gray circles, $n=3$, unpublished data) and previously published data from human colon samples (gray squares, $n=18$, Eckburg *et al.* 2005) and human stool samples (gray triangles, $n=3$, Eckburg *et al.* 2005; black triangles, $n=15$, Dethlefsen *et al.* 2008). All sequences were compared using the same alignment and 1253-nucleotide filter.

Table 4a Observed, shared OTUs (to the left and below the diagonal, in bold) and genera (above and to the right of the diagonal, italicized) for each subject pair

Subject 1		*23*	*28*	*32*	*29*	*25*	*27*	*25*	*30*	*27*
Subject 2	**47**		*22*	*22*	*21*	*18*	*20*	*21*	*21*	*21*
Subject 3	**67**	**49**		*29*	*26*	*23*	*23*	*23*	*28*	*25*
Subject 4	**70**	**48**	**65**		*31*	*25*	*29*	*25*	*34*	*26*
Subject 5	**62**	**44**	**60**	**69**		*24*	*27*	*24*	*30*	*24*
Subject 6	**54**	**38**	**54**	**59**	**54**		*22*	*23*	*24*	*22*
Subject 7	**50**	**36**	**48**	**51**	**45**	**44**		*23*	*26*	*24*
Subject 8	**47**	**34**	**40**	**44**	**41**	**38**	**29**		*24*	*24*
Subject 9	**71**	**45**	**71**	**76**	**65**	**56**	**49**	**47**		*24*
Subject 10	**50**	**38**	**51**	**50**	**50**	**44**	**35**	**41**	**50**	
	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10

Values to the left and below the diagonal (in bold) are calculated from shared OTUs and those above and to the right of the diagonal (italicized) are calculated from shared genera.

Table 4b Chao–Jaccard abundance-based similarities between each pair of subjects (raw values)

Subject 1		*0.933*	*0.967*	*0.968*	*0.981*	*0.889*	*0.970*	*0.971*	*0.953*	*0.974*
Subject 2	**0.643**		*0.912*	*0.864*	*0.933*	*0.862*	*0.945*	*0.946*	*0.845*	*0.943*
Subject 3	**0.706**	**0.603**		*0.950*	*0.947*	*0.889*	*0.920*	*0.945*	*0.927*	*0.973*
Subject 4	**0.59**	**0.501**	**0.694**		*0.978*	*0.949*	*0.964*	*0.952*	*0.974*	*0.956*
Subject 5	**0.682**	**0.720**	**0.723**	**0.726**		*0.925*	*0.970*	*0.981*	*0.955*	*0.975*
Subject 6	**0.571**	**0.668**	**0.641**	**0.655**	**0.764**		*0.919*	*0.913*	*0.898*	*0.921*
Subject 7	**0.616**	**0.765**	**0.720**	**0.663**	**0.787**	**0.799**		*0.983*	*0.953*	*0.983*
Subject 8	**0.647**	**0.625**	**0.567**	**0.553**	**0.703**	**0.535**	**0.646**		*0.923*	*0.988*
Subject 9	**0.683**	**0.602**	**0.729**	**0.715**	**0.762**	**0.664**	**0.716**	**0.597**		*0.920*
Subject 10	**0.642**	**0.679**	**0.702**	**0.560**	**0.769**	**0.686**	**0.801**	**0.745**	**0.646**	
	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10

Values to the left and below the diagonal (in bold) are calculated from shared OTUs and those above and to the right of the diagonal (italicized) are calculated from shared genera.

(Table 4a). Similarity between communities was typically low, averaging 0.671 (range 0.501–0.801) with the raw index and 0.760 (range 0.533–0.969) with the estimated index (Table 4b). When community similarity was examined on the genus level, observed shared genera averaged 25 (range 18–34) (Table 4a) and the Chao–Jaccard abundance similarity averaged 0.942 (range 0.845–0.988 for raw) and 0.963 (range 0.845–1 for estimated) (Table 4b). A value of 1 indicates that all genera are shared between the two specimens examined. A total of 15 bacterial genera were observed in all 10 healthy individuals: *Neisseria*, *Cardiobacterium*, *Haemophilus* and *Campylobacter* (Proteobacteria); *Streptococcus*, *Granulicatella* and *Veillonella* (Firmicutes); *Fusobacterium* (Fusobacteria); *Rothia*, *Actinomyces*, *Corynebacterium* and *Atopobium* (Actinobacteria); and *Prevotella*, *Capnocytophaga* and *Bergeyella* (Bacteroidetes). Every individual also contained TM7 sequences (Figure 4). All of these bacterial taxa were also present in the pooled subgingival library. Of these shared genera, species belonging to eight were present in all 10 individuals, leading to eleven shared bacterial species: *Haemophilus parainfluenzae*, *Streptococcus oralis*, *Streptococcus sanguinis*, *Granulicatella adiacens*, *Veillonella parvula*,

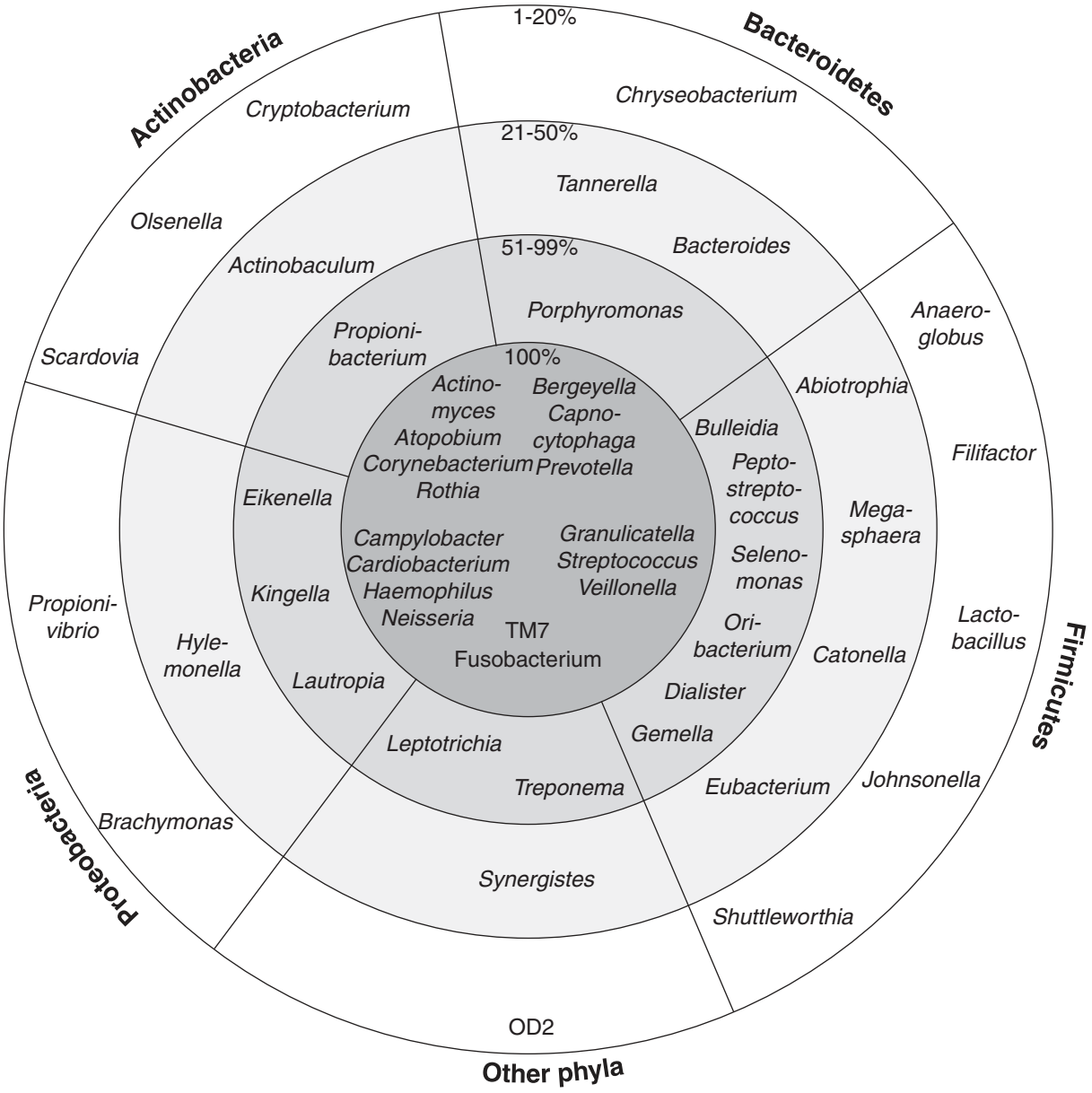

Figure 4 Schematic depiction of oral community membership among 10 healthy individuals. Inner circle, bacterial genera found in all 10 individuals (100%); second circle, present in 6–9 out of 10 individuals (51–99%); third circle, present in 3–5 individuals (21–50%); outer circle, present in 1–2 individuals (1–20%). Genera are grouped according to phylum.

Veillonella dispar, *Rothia aeria*, *Actinomyces naeslundii*, *Actinomyces odontolyticus*, *Prevotella melaninogenica* and *Capnocytophaga gingivalis*.

Interindividual differences among the bacterial communities of the human mouth

Despite conserved oral bacterial community composition at the genus level, there were also interindividual differences. Several different patterns of genus dominance were found in the 10 healthy mouths. Of the 10 mouths, 5 were dominated by *Streptococcus* species (nos. 2, 5, 7, 9 and 10). Two mouths were dominated by *Prevotella* (nos. 1, 4), and one each was dominated by *Neisseria* (no. 3), *Haemophilus* (no. 8) and *Veillonella* (no. 6) (Supplementary Figure S3). In addition, even among the genera present in all 10 healthy individuals, the presence of particular species within that genus was variable between individuals. For example, although every subject had sequences belonging to the genus *Neisseria*, no single *Neisseria* species was shared across all subjects. The same was true for species in the genera *Fusobacterium* and *Corynebacterium*.

Co-occurrence of bacterial taxa

Co-occurrence analysis was performed on the data obtained from the 10 individual subjects, using the *C*-score of Stone and Roberts (1990), which compares the taxon distribution of a data set to a randomized distribution of the same number of taxa. This method calculates the checkerboard units for each taxon pair (how often those two taxa are found together). When analyzed at the level of OTU, the observed *C*-score was not significantly different from the null hypothesis (random distribution). When the same data were analyzed at the genus level, the *C*-score indicated that the communities display co-occurrence patterns significantly different from the null hypothesis (observed $C = 0.99184$, expected $C = 0.95366$; $P = 0.02860$). These scores (higher than expected) suggested segregation or competition among taxa. Examination of the matrix of checkerboard units between each taxon pair can pinpoint taxa that are more or less likely to be found together. Figure 5 displays the taxa pairs as a matrix of *C*-scores. Taxa with low *C*-scores (found together frequently) are colored white, whereas those with high *C*-scores (rarely or never found together) are colored black. Genus pairs in which both genera are found in all mouths, such as *Streptococcus*, *Neisseria* and *Haemophilus* have zero checkerboard units, as expected. When examining the genus pairs with high checkerboard units, the genus *Abiotrophia* was identified as unlikely to be found together with the genera *Dialister*, *Oribacterium*, *Eubacterium* and *Treponema*. In addition, the genus *Scardovia* was unlikely to be found with *Eikenella* or *Dialister*. Because it may be inappropriate to compare this broad range of bacterial taxa in a single analysis (owing to the fact that members of different phyla may not be in competition), we re-analyzed the OTU-level data, but in this case, comparing the patterns only within a given phyla. In this case, we also calculated the *C*-scores based on presence/absence for all OTUs (but only within a given phylum). This was repeated for each phylum, except for OD2 and Synergistes, owing to the few observations in each of these two groups. This OTU-level, within-phylum analysis revealed that only the taxa within Firmicutes showed a *C*-score significantly different from the null hypothesis (observed = 2.1370, expected = 2.08243; $P = 0.03460$), suggesting segregation of species and evidence of possible competitive species interactions.

Discussion

The composition of the microbial communities on and within the human body varies between individuals. Interindividual variation has been shown in a variety of studies for the healthy intestinal tract (Eckburg *et al.*, 2005; Dethlefsen *et al.*, 2006; Ley *et al.*, 2006; Palmer *et al.*, 2007). In contrast, knowledge about the interindividual differences in the healthy human mouth microbiota and the uniqueness of the oral microbiota compared with other microbial communities in our bodies is still somewhat sparse. Several molecular studies have been carried out regarding the composition of the oral microbiota, but these studies used limited numbers of sequences per individual or only looked at short regions of the 16S rRNA gene (Kroes *et al.*, 1999; Paster *et al.*, 2001; Kazor *et al.*, 2003; Aas *et al.*, 2005). A study by Diaz *et al.*, 2006 in three individuals showed that early colonization of enamel is subject specific. The distinctness of the phylogenetic structure of the human oral microbiota in relation to the microbiota of the skin and feces in nine individuals was revealed in a recent study (Costello *et al.*, 2009). Although other studies have considered the oral microbiota of a larger number of individuals, our study was based on one of the largest sets of near full-length sequences per individual to date for the human oral cavity. The most important contributions of this work are the combination of depth of coverage and degree of phylogenetic resolution for the human mouth, the features of a human oral core microbiota and previously unrecognized patterns of taxon co-occurrence.

In this study, we amplified and analyzed an average number of 1029 near full-length, well-aligned oral 16S rRNA gene sequences (range 931–1070) per subject from each of 10 healthy individuals, as well as an additional 1083 clones from the pooled subgingival specimens, bringing the total number of sequences analyzed in this study to 11 368. The advantage of near full-length 16S rRNA gene sequences in providing greater phylogenetic

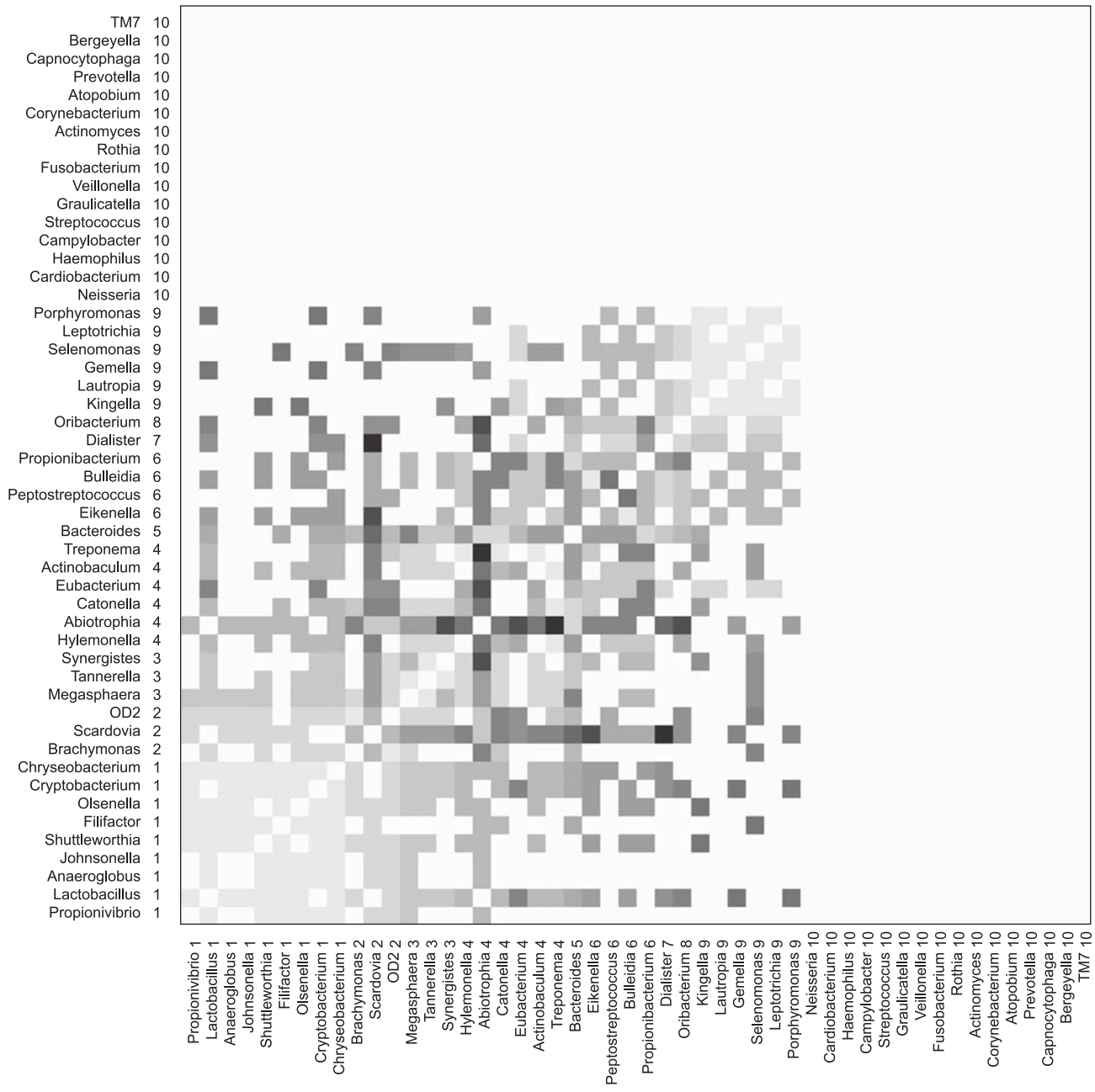

Figure 5 Checkerboard (*C*) scores for each possible combination of two genera. The *C*-scores are shown in gray scale. White depicts a *C*-score of 0 for genera always found together. Darker colors show higher *C*-scores for genera that co-occur less frequently than expected. The highest *C*-score in this data set, 16, was found for the *Abiotrophia*–*Treponema* genus pair, two fairly abundant genera never found together. Genera are ordered according to their overall abundance in the 10 individual mouth pools. The numbers after the genus names indicate the number of individuals (out of 10) in which that genus was found. The 16 taxa that were found in all 10 individuals, as expected, show a *C*-score of 0 (white). Data obtained from the subgingival pool were not included in this analysis.

resolution than hypervariable region 'tag' sequences was highlighted in a comparative analysis of these two types of sequence data (Huse *et al.*, 2008). In this data set, we identified a total of 247 different OTUs at the level of species, of which 24 were less than 99% identical than previously published sequences. Approximately 10% of the OTUs found in this study were previously uncharacterized.

The abundant bacterial groups found in our study are similar to those found in most other studies. For example, 20% of our sequences belonged to the genus *Streptococcus*, confirming the preponderance of *Streptococcus* species within a healthy mouth by microscopy and culture (Socransky, 1963) and by molecular methods (Kroes *et al.*, 1999). In a recent molecular study, the most predominant bacterial genera in the oral cavity were *Streptococcus*, *Gemella*, *Abiotrophia*, *Granulicatella*, *Rothia*, *Neisseria* and *Prevotella* (Aas *et al.*, 2005). We found those same groups to be prevalent as well, but, in addition,

we found many Proteobacteria (for example, *Haemophilus* and *Lautropia*) to be abundant. This difference may be the result of a deeper sequencing effort per individual in the current study (average 57.5 clones per subject in the Aas *et al*. study for a total of 2589 clones, in contrast to an average 1029 clones per individual in this study). In addition, different DNA extraction methods and different broad-range PCR primers could also explain the divergent results.

Despite the evidence for a conserved healthy oral community at the genus level in all 10 healthy mouths, there was also evidence in this study for large interindividual differences. Our study confirms results by Nasidze *et al*. (2009), suggesting high variability in the oral microbiome between individuals, although in the latter study, saliva was the only specimen-type examined. In addition to *Streptococcus,* which was the most abundant genus in the combined data set and in five of the individual mouths, we identified four additional genera that may dominate the oral ecosystem of a healthy subject. Our data indicate that there are various alternative oral bacterial community structures and a greater degree of variation in patterns of diversity associated with oral health than previously thought. It remains to be seen what factors, for example, human genetics or lifestyle, correlate with oral bacterial community structure. Clearly, the concept of a core oral microbiome may be better defined with measurements of community function rather than community membership. Such analyses will need to include community-wide assessments of gene content, gene transcript abundance and protein products.

The role of bacteria in periodontal disease is complex, and likely involves polymicrobial consortia (Lepp *et al*., 2004). Socransky and Haffajee have proposed that the presence of a high proportion of so-called 'red complex' bacteria, that is, *Porphyromonas gingivalis*, *Tannerella forsythia* and *Treponema denticola,* is associated with periodontal disease (Socransky *et al*., 1998; Haffajee *et al*., 2008). In a survey of five healthy mouths, Aas *et al*. (2005) did not find any representatives of the 'red complex'. Other studies have, however, identified members of this complex in healthy mouths (Ximenez-Fyvie *et al*., 2000). In our study, all three species were found in subjects with healthy gingival tissues, although in low numbers and limited to subjects 1, 4 and 9. Taken together with previous studies, this study confirms that the 'red complex' group may be found in small numbers in healthy individuals. Other bacterial species such as *Filifactor alocis*, *Selenomonas* species and *Dialister* species have been associated with a worsening periodontal status (Kumar *et al*., 2005). A bacterial species previously shown to be associated with periodontal health (*Veillonella parvula*, *Veillonella* X042, Genbank accession number AF287781) (Kumar *et al*., 2005) was found in all specimens in this study, and was the third most abundant OTU in our combined sequence data set.

UniFrac principal coordinates analysis showed no apparent clustering of oral microbial communities based on gender, age or ethnicity. In addition, UniFrac analysis showed no apparent effect of DNA extraction method of oral specimens. No individual pool was found to be more significantly different than others in pairwise comparisons, and the subgingival library was not significantly different from the individual pools. This may be indicative of the fact that (1) despite the many different habitats in the human mouth, many bacterial species are shared among those habitats or (2) that the individual pools are dominated by the subgingival specimens. However, the number of subjects in this study was relatively small, and interindividual differences associated with gender, age or ethnicity might become apparent when larger numbers of subjects are studied. Because specimens from multiple sites within an individual were pooled, bacterial community differences between anatomical sites could not be examined.

When the oral sequence libraries were compared with similar sequence libraries from the human colon and stool, a clear clustering according to anatomical site was observed. These results need to be interpreted with caution, as data were obtained from different individuals and differences between study groups might drive some of the findings. But it is appealing to assume that each anatomical location within a healthy human has specific physiochemical conditions that shape the composition of a microbial community specifically adapted to that site. Our finding of human habitat-specific microbial community structure is supported by recently published data (Costello *et al*., 2009).

Tests for significant segregation patterns of taxa were originally developed as a means of assessing whether competition between taxa is a driving force behind community assembly. *C*-scores higher than expected are consistent with inter-species competition, as well as with habitat differences that cross over the sampling scheme and historical processes. We feel that habitat differences (other than host genotype) were minimized in our study owing to the fact that the pools presumably represented multiple intra-oral sites in a consistent manner across individuals. However, successional or early historical differences between subjects cannot be eliminated as a possible explanation of the observed segregation patterns. It has been previously suggested that as taxonomic level is refined, *C*-scores become more statistically significant (Horner-Devine *et al*., 2007). The fact that significant segregation was found at the genus level in our study, but not at a level equivalent to species, has several possible interpretations. One possibility is that taxonomic levels are not the relevant biological units of measure. Another possibility is that the level of ecological interest and interaction in the mouth is

the level that humans have chosen to label as genus, rather than species.

Co-occurrence analysis not only addresses the forces structuring a community but also draws attention to specific taxa that have apparent interactions and may be worthy of further investigation. For instance, in this study, *Abiotrophia* was found to have a high number of checkerboard units with the genera *Dialister*, *Oribacterium*, *Eubacterium* and *Treponema*, and the genus *Scardovia* had a high number of checkerboard units with *Eikenella* and *Dialister*. Interactions among these genera have not been the focus of research so far, but such research may lead us to understand whether and why these taxa compete. Each of these genera (except *Treponema*) is represented in this data set by a single species, each of which has been implicated in human disease; recognition of competitive partners may prove useful in preventive medicine. For instance, it has been suggested that known competitive interactions between *Streptococcus mutans* and other species may be exploited to develop preventive treatments for dental caries by encouraging growth of species with lower cariogenicity (Kreth *et al.*, 2005).

This study shows that each person's mouth harbors a unique community of bacterial species, but that these communities tend to be more similar when classified at the level of genus. Ecological tools initially developed for larger organisms, such as co-occurrence analysis, will greatly facilitate the analysis of complex bacterial communities such as those found in the human body and will enhance our understanding of the role of the microbiota in health and disease.

Acknowledgements

We thank Karla Lightfield for technical assistance and Katie Shelef for help with pooling and extracting the oral specimens. This work was funded by NIH R01-DE014868 (CFL), NIH R01-DE13541 (DAR) and NIH Pioneer Award DP1-OD000964 (DAR). DAR is supported by the Thomas C and Joan M Merigan Endowment at Stanford University.

References

Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. (2005). Defining the normal bacterial flora of the oral cavity. *J Clin Microbiol* **43**: 5721–5732.

Chao A, Chazdon RL, Colwell RK, Shen T-J. (2005). A new statistical approach for assessing compositional similarity based on incidence and abundance data. *Ecol Lett* **8**: 148–159.

Cogen AL, Nizet V, Gallo RL. (2008). Skin microbiota: a source of disease or defence? *Br J Dermatol* **158**: 442–455.

Colwell RK. (2005). *EstimateS: Statistical Estimation of Species Richness and Shared Species from Specimens*. Version 7. User's Guide and application published at:http://purl.oclc.org/estimates.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. (2009). Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694–1697.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Dethlefsen L, Eckburg PB, Bik EM, Relman DA. (2006). Assembly of the human intestinal microbiota. *Trends Ecol Evol* **21**: 517–523.

Dethlefsen L, Huse S, Sogin ML, Relman DA. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* **6**: e280.

Dethlefsen L, McFall-Ngai M, Relman DA. (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* **449**: 811–818.

Dewhirst FE, Izard J, Paster BJ, Tanner AC, Wade WG, Yu W-H *et al.* (2008). *The Human Oral Microbiome Database*. http://www.HOMD.org.

Diaz PI, Chalmers NI, Rickard AH, Kong C, Milburn CL, Palmer Jr RJ *et al.* (2006). Molecular characterization of subject-specific oral microflora during initial colonization of enamel. *Appl Environ Microbiol* **72**: 2837–2848.

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M *et al.* (2005). Diversity of the human intestinal microbial flora. *Science* **308**: 1635–1638.

Edwards U, Rogall T, Blocker H, Emde M, Bottger EC. (1989). Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucleic Acids Res* **17**: 7843–7853.

Flint HJ, Duncan SH, Scott KP, Louis P. (2007). Interactions and competition within the microbial community of the human colon: links between diet and health. *Environ Microbiol* **9**: 1101–1111.

Good IJ. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**: 237–264.

Gotelli NJ, Entsminger GL. (2004). *EcoSim: Null Models Software for Ecology* Version 7. Acquired Intelligence Inc. & Kesey-Bear. Jericho, VT 05465. http://garyentsminger.-com/ecosim/index.htm.

Haffajee AD, Socransky SS, Patel MR, Song X. (2008). Microbial complexes in supragingival plaque. *Oral Microbiol Immunol* **23**: 196–205.

Horner-Devine MC, Silver JM, Leibold MA, Bohannan BJ, Colwell RK, Fuhrman JA *et al.* (2007). A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology* **88**: 1345–1353.

Hugenholtz P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3**: reviews0003.1–0003.8.

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

Kazor CE, Mitchell PM, Lee AM, Stokes LN, Loesche WJ, Dewhirst FE *et al.* (2003). Diversity of bacterial populations on the tongue dorsa of patients with halitosis and healthy patients. *J Clin Microbiol* **41**: 558–563.

Keijser BJ, Zaura E, Huse SM, van der Vossen JM, Schuren FH, Montijn RC *et al.* (2008). Pyrosequencing analysis of the oral microflora of healthy adults. *J Dent Res* **87**: 1016–1020.

Kreth J, Merritt J, Shi W, Qi F. (2005). Competition and coexistence between Streptococcus mutans and Streptococcus sanguinis in the dental biofilm. *J Bacteriol* **187**: 7193–7203.

Kroes I, Lepp PW, Relman DA. (1999). Bacterial diversity within the human subgingival crevice. *Proc Natl Acad Sci USA* **96**: 14547–14552.

Kumar PS, Griffen AL, Moeschberger ML, Leys EJ. (2005). Identification of candidate periodontal pathogens and beneficial species by quantitative 16S clonal analysis. *J Clin Microbiol* **43**: 3944–3955.

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* **82**: 6955–6959.

Lepp PW, Brinig MM, Ouverney CC, Palm K, Armitage GC, Relman DA. (2004). Methanogenic Archaea and human periodontal disease. *Proc Natl Acad Sci USA* **101**: 6176–6181.

Leser TD, Amenuvor JZ, Jensen TK, Lindecrona RH, Boye M, Moller K. (2002). Culture-independent analysis of gut bacteria: the pig gastrointestinal tract microbiota revisited. *Appl Environ Microbiol* **68**: 673–690.

Ley RE, Peterson DA, Gordon JI. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**: 837–848.

Lozupone C, Hamady M, Knight R. (2006). UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7**: 371.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.

Nasidze I, Li J, Quinque D, Tang K, Stoneking M. (2009). Global diversity in the human salivary microbiome. *Genome Res* **19**: 636–643.

Palmer C, Bik EM, Digiulio DB, Relman DA, Brown PO. (2007). Development of the human infant intestinal microbiota. *PLoS Biol* **5**: e177.

Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, Levanos VA *et al.* (2001). Bacterial diversity in human subgingival plaque. *J Bacteriol* **183**: 3770–3783.

Schloss PD, Handelsman J. (2004). Status of the microbial census. *Microbiol Mol Biol Rev* **68**: 686–691.

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.

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

Socransky SS. (1963). The microbiota of the gingival crevice area of man. I. Total microscopic and viable counts and counts of specific organisms. *Arch Oral Biol* **8**: 275–280.

Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent Jr RL. (1998). Microbial complexes in subgingival plaque. *J Clin Periodontol* **25**: 134–144.

Stone L, Roberts A. (1990). The checkerboard score and species distributions. *Oecologia* **85**: 74–79.

Tappenden KA, Deutsch AS. (2007). The physiological relevance of the intestinal microbiota–contributions to human health. *J Am Coll Nutr* **26**: 679S–683S.

Woese CR, Fox GE. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* **74**: 5088–5090.

Ximenez-Fyvie LA, Haffajee AD, Socransky SS. (2000). Comparison of the microbiota of supra- and subgingival plaque in health and periodontitis. *J Clin Periodontol* **27**: 648–657.

Zaura E, Keijser BJ, Huse SM, Crielaard W. (2009). Defining the healthy 'core microbiome' of oral microbial communities. *BMC Microbiol* **9**: 259.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)