The ISME Journal (2012) 6, 46–56

www.nature.com/ismej

ORIGINAL ARTICLE

The oral metagenome in health and disease

Pedro Belda-Ferre[1], Luis David Alcaraz[1], Raúl Cabrera-Rubio[1], Héctor Romero[2], Aurea Simón-Soro[1], Miguel Pignatelli[1] and Alex Mira[1]

[1]*Department of Genomics and Health, Center for Advanced Research in Public Health, Valencia, Spain and* [2]*Laboratorio de Organización y Evolución del Genoma, Facultad de Ciencias/C.U.R.E., Universidad de la República, Montevideo, Uruguay*

The oral cavity of humans is inhabited by hundreds of bacterial species and some of them have a key role in the development of oral diseases, mainly dental caries and periodontitis. We describe for the first time the metagenome of the human oral cavity under health and diseased conditions, with a focus on supragingival dental plaque and cavities. Direct pyrosequencing of eight samples with different oral-health status produced 1 Gbp of sequence without the biases imposed by PCR or cloning. These data show that cavities are not dominated by *Streptococcus mutans* (the species originally identified as the ethiological agent of dental caries) but are in fact a complex community formed by tens of bacterial species, in agreement with the view that caries is a polymicrobial disease. The analysis of the reads indicated that the oral cavity is functionally a different environment from the gut, with many functional categories enriched in one of the two environments and depleted in the other. Individuals who had never suffered from dental caries showed an over-representation of several functional categories, like genes for antimicrobial peptides and quorum sensing. In addition, they did not have *mutans* streptococci but displayed high recruitment of other species. Several isolates belonging to these dominant bacteria in healthy individuals were cultured and shown to inhibit the growth of cariogenic bacteria, suggesting the use of these commensal bacterial strains as probiotics to promote oral health and prevent dental caries.

The ISME Journal (2012) **6,** 46–56; doi:10.1038/ismej.2011.85; published online 30 June 2011
Subject Category: microbe–microbe and microbe–host interactions
Keywords: metagenomics; human microbiome; dental caries; *Streptococcus mutans*; pyrosequencing; probiotics

Introduction

The oral cavity of humans is inhabited by hundreds of bacterial species, most of which are commensal and required to keep equilibrium in the mouth ecosystem. However, some of them have a key role in the development of oral diseases, mainly dental caries and periodontal disease (Marsh, 2010). Oral diseases initiate with the growth of the dental plaque, a biofilm formed by the accumulation of bacteria in a timely manner together with the human salivary glycoproteins and polysaccharides secreted by the microbes (Marsh, 2006). The subgingival plaque, located within the neutral or alkaline subgingival sulcus, is typically inhabited by anaerobic Gram negatives and is responsible for the development of gingivitis and periodontitis. The supragingival dental plaque is formed on the teeth surfaces by acidogenic and acidophilic bacteria, which are responsible for dental caries. This is considered the most extended infectious disease in the world, affecting over 80% of the human population (Petersen, 2004). A poor oral health has also been related to the stomach ulcers, gastric cancer or cardiovascular disease, among others (Watabe *et al.*, 1998; Wu *et al.*, 2000). It is therefore surprising that no efficient strategies to combat oral diseases have been developed, despite their dramatic impact on human health. Some of the main reasons that oral pathogens have not been eradicated are related to the difficulty of studying the microbial communities inhabiting the oral cavity: First, the complexity of the ecosystem (several hundreds of species have been reported with multiple interaction levels) makes the potential pathogenical species difficult to target (Socransky *et al.*, 1998); second, not a single ethiological agent can be identified as in classical, Koch's postulates diseases. This has been clearly shown in periodontal disease, where at least three bacterial species that belong to very different taxonomic groups (the so-called 'red complex' of periodontal pathogens) are known to be involved in the illness (Darveau, 2010); and third, a large proportion of oral bacteria cannot be cultured (Paster *et al.*, 2001), and therefore traditional microbiological approaches give an incomplete picture of the natural communities inhabiting the

Correspondence: A Mira, Department of Genomics and Health, Center for Advanced Research in Public Health (CSISP), Avda. Cataluña 21, Valencia 46020, Spain.
E-mail: mira_ale@gva.es
Received 17 March 2011; revised 10 May 2011; accepted 12 May 2011; published online 30 June 2011

dental plaque. However, the development of metagenomic techniques and next-generation sequencing technology now allows the study of whole bacterial communities by analysing the total DNA pool from complex microbial samples.

Pioneering metagenomic studies in the human microbiome centred in the gut ecosystem, initially through a shot-gun approach, in which DNA was cloned in small-size plasmids followed by traditional Sanger sequencing method (Gill *et al.*, 2006; Kurokawa *et al.*, 2007), obtaining reads of about 800–1000-bp long. Recent approaches include the end sequencing of large-size fosmids (Vaishampayan *et al.*, 2010) and the use of Illumina sequencing technology to deliver vast amounts of small-size reads that could be later assembled (Qin *et al.*, 2010). Studies of the oral cavity microbiota, as well as other body habitats within the human microbiome such as the skin, the vagina or the respiratory tract, have mainly focused on the sequencing of PCR-amplified rRNA genes (Aas *et al.*, 2005; Grice *et al.*, 2008). These PCR-based studies have provided a substantial improvement of our knowledge of oral bacterial communities compared with past culture-based research, but the estimates of microbial diversity are hampered by biases in PCR amplification (de Lillo *et al.*, 2006), cloning bias (Ghai *et al.*, 2010) and when short pyrosequencing reads of the 16S rRNA gene were used, uncertainties in taxonomic assignment (Keijser *et al.*, 2008; Lazarevic *et al.*, 2009) and inflated diversity due to pyrosequencing errors (Quince *et al.*, 2009). Recently, the first study of the oral metagenome has been carried out by directly applying next-generation sequencing to a single sample from a healthy individual (Xie *et al.*, 2010), thus removing potential biases imposed by cloning and PCR. We have applied a similar approach to several samples varying in health status, directly sequencing the metagenomic DNA by 454 pyrosequencing, which has allowed us to compare the total genetic repertoire of the bacterial community under different health conditions.

Materials and methods

Sample collection

Supragingival dental plaque was obtained from 25 volunteers after signing an informed consent. The sampling procedure was approved by the Ethical Committee for Clinical Research from the DGSP-CSISP (Valencian Health Authority, Spain). The oral health status of each individual was evaluated by a dentist following recommendations and nomenclature from the Oral Health Surveys from the WHO, taking samples with sterile curettes. Plaque material from all teeth surfaces from each individual was pooled. In volunteers with active caries, the dental plaque samples were taken without touching cavities. In those cases, material from individual cavities was also extracted and kept separately. The volunteers were asked not to brush their teeth 24 h before the sampling. Information was obtained regarding oral hygiene, diet and signs of periodontal disease. DNA was extracted using the MasterPure Complete DNA and RNA Purification Kit (Epicentre Biotechnologies, Madison, WI, USA), following the manufacturer's instructions, adding a lysozyme treatment (5 mg ml^{-1}, at 37 °C for 30 min). For this study, eight samples were used for subsequent pyrosequencing, selected on the basis of homogeneity in their clinical features, including similar age, periodontal status, smoking habits and mucosal health. Supragingival dental plaque samples were taken from six individuals that were divided in three groups according to the number of caries they had suffered and that represented different degrees of oral health: two individuals had never developed caries in their lives (healthy controls), another two individuals had been regularly treated for caries in the past and had a low number of active caries at the moment of sampling (one and four cavities, respectively); and the last two individuals had a high number of active caries (8 and 15) and poor oral hygiene. In addition, samples from individual cavities were collected, and for two of them enough DNA for pyrosequencing was obtained: one at an intermediate stage and the other one at an advanced stage of caries development (dentin lesion), corresponding to teeth 1.6 and 4.6 following WHO nomenclature. The sequencing was performed at Macrogen Inc. (Seoul, South Korea) using the GS-FLX sequencer (Roche, Basel, Switzerland) with Titanium chemistry. After quality checking, average read length was 425 ± 117 bp. Sequences were deposited, and are publicly available in the MG-RAST server with the following accesions: 4447192.3, 4447102.3, 4447103.3, 4447101.3, 4447943.3, 4447903.3, 4447971.3 and 4447970.3.

Sequence analysis

Artificially replicated sequences (accounting for 1.2–4.54% of the raw reads) were removed from the data set using the '454 replicate filter' (Gomez-Alvarez *et al.*, 2009). The human sequences were identified by MegaBlast (Altschul *et al.*, 1990) against the human genome (e-value cutoff 1e−10) and were removed from the final data set. They accounted for 2.23–74.99% of the replicate-filtered reads (Supplementary Table 1). The metagenomic reads were mapped against 1117 sequenced reference genomes using the Nucmer and Promer v3.06 alignment algorithms, with the default parameters (Kurtz *et al.*, 2004). The nucleotide identity values of each read against its hit in the genome were used to generate frequency histograms. If the mode was 94% or higher the plot was considered to represent sequence identity against the same species (Konstantinidis and Tiedje, 2005). Stand-alone RPSBlast was used to align reads (translated into

all six possible reading frames) to protein profiles (represented by position-specific scoring matrices). Queries were performed against the complete conserved domains database (Marchler-Bauer *et al.*, 2009) and against the COGs (Tatusov *et al.*, 2003) and Tigrfams (Selengut *et al.*, 2007) databases. Fractions of sequences assigned in each case are shown in Supplementary Table 2. TFams classification assignments were integrated into higher hierarchical levels, according to the Tigrfam classification scheme, in subroles and main roles. COGs assignments were also integrated into the higher level of COG's functional categories. In addition, samples were uploaded to the MGRAST server (Meyer *et al.*, 2008) and the functional assignment based on SEED subsystems was retrieved for the three hierarchical levels used: Subsystem, subsystem hierarchy 2 and subsystem hierarchy 1 (bottom up). In all cases, a table containing the counts of functional categories per sample was generated and used for subsequent analysis. All statistical analyses were conducted on *R* (2.6.2). Heat maps of taxonomic composition were generated using the gplots library of *R* (Warnes *et al.*, 2009) with relative frequencies per sample, as well as Euclidean distance, or normal medians. The relative rates of over-represented features present in the people without caries were estimated using a control of the false discovery rate, for testing the amount of false positive predictions (*q*-values) for a given *P*-value of significance, with the algorithm described by White *et al.* (2009).

Taxonomic assignment

16S rRNA sequences were extracted from the reads of each metagenome by similarity search using BLASTn (Altschul *et al.*, 1990) against the RDP database, with an e-value cutoff of 1e−10. Sequences <200 bp were removed. Phylogenetic assignment of the sequences was made using the RDP Classifier (Wang *et al.*, 2007), using an 80% confidence threshold. New operational taxonomic units were proposed if the reads were over 400 bp in length and had a nucleotide identity between 80–95% to known 16S sequences. Taxonomic assignments of all open reading frames were carried out based on a lowest common ancestor (LCA) algorithm (Alstrup *et al.*, 2004) with the characteristics described in the MEGAN software (Huson *et al.*, 2007). We implemented the algorithm in a multi-threaded command-line oriented in-house software in order to obtain faster analysis and simplify its integration in pipelines and downstream analysis. To obtain the LCA of each sequence, we carried out BLASTx homology searches against a custom database comprising the non-eukaryotic sequences of the NCBI's non-redundant database. For each query sequence (read), only hits with a bit score at least 90% of the best matches were considered in the LCA computation. We also made use of the script phymmBL (Brady and Salzberg, 2009) that combines the assignment of sequences both by homology and by nucleotide composition using hidden Markov Models. All the available complete and WGS genomes were retrieved from the human oral microbiome database (Chen *et al.*, 2010), as well as the RefSeq of NCBI containing all bacterial and archaea genomes (june 2010), and were used to build a local database to perform taxonomic model constructions and homology searches, using sequences larger than 200 bp to predict taxonomic affiliation. At this read length, phymmBL's performance at the class level has been estimated to be over 75%. All the taxonomic and functional results were parsed into a MySQL database for further analysis.

Results and discussion

The oral microbiome by pyrosequencing

Supragingival dental plaque samples were taken from six individuals that were divided in three groups according to the number of caries they had suffered and that represented different degrees of oral health: two individuals had never developed caries in their lives (healthy controls), another two individuals had been regularly treated for caries in the past and had a low number of active caries at the moment of sampling; and the last two individuals had a high number of active caries and poor oral hygiene. In addition, samples from individual cavities were collected, and for two of them enough DNA for pyrosequencing was obtained. A total of 1 Gbp of DNA sequence was obtained from the eight samples selected. The amount of human DNA in the metagenomes varied from 0.5–40% in supragingival dental plaque samples (Supplementary Table 1), thus the total size of the studied metagenome was reduced to 842 Mbp of sequence. We obtained an average read length of 425 ± 117 bp, which allowed a functional assignment in a significant fraction of the metagenome (Supplementary Table 2). In addition, assembly of those reads produced 1103 contigs larger than 5 Kb and 354 longer than 10 Kb. Success in the assembly of large contigs was dependent on sequencing effort. We obtained an average of 129.5 Mbp of filtered, high-quality sequences for each of the six oral samples. In the two cavity samples, around 70% of the reads corresponded to human DNA, and an average of 32.5 Mbp of filtered, high-quality reads were obtained.

Estimating diversity in the oral metagenome

We estimated microbial diversity in all samples by three different methods. First, we selected the reads matching 16S rRNA genes, assigning them to different taxonomic levels. A total of 4254 16S rRNA sequences were obtained (Supplementary Table 1), giving a similar picture of diversity to that obtained through 16S rRNA PCR-dependent procedures (Bik *et al.*, 2010), although the relative proportions of each taxonomic group were different (Figure 1). These

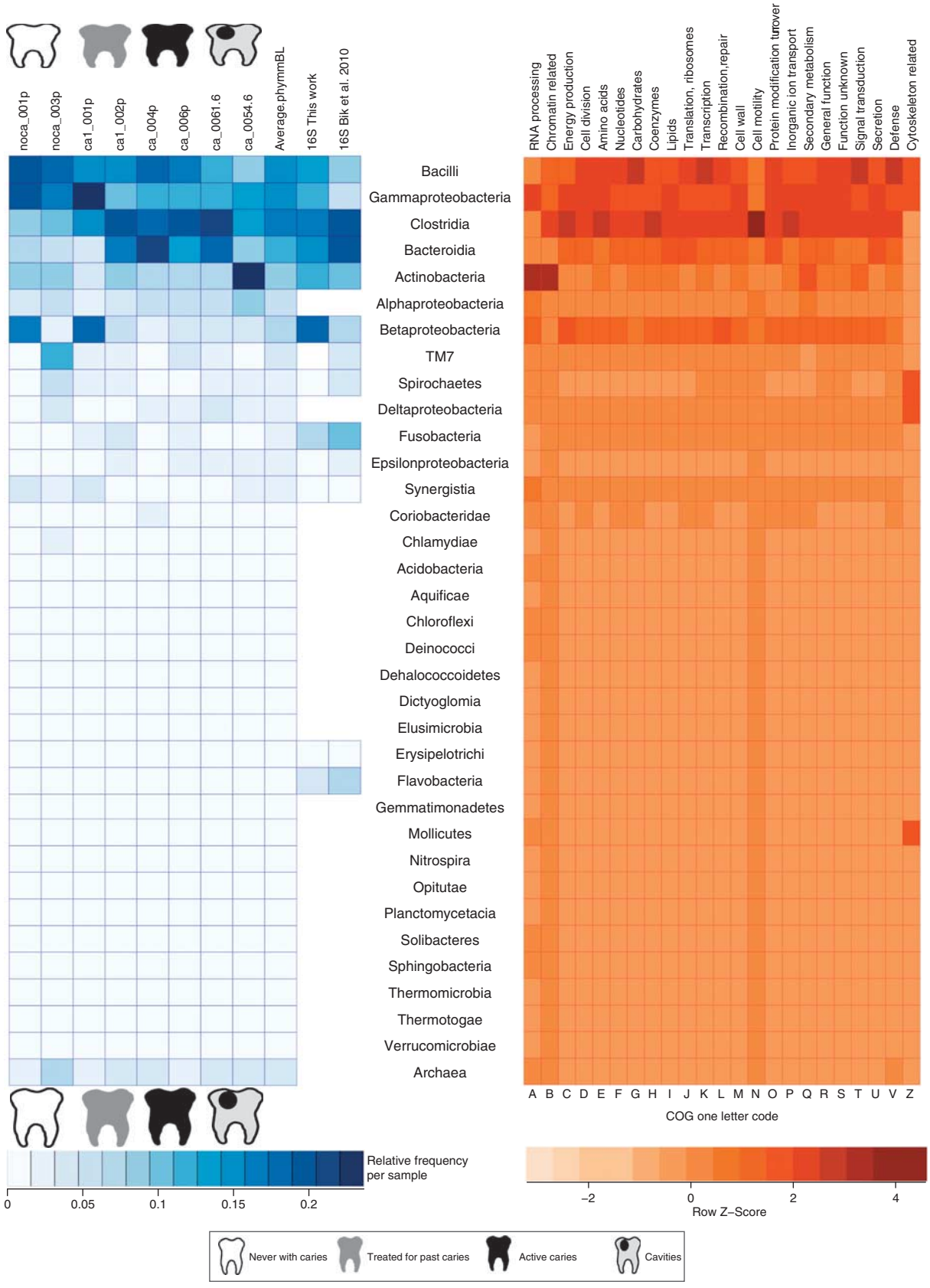

Figure 1 Bacterial diversity in the oral cavity. The graph on the left shows the relative frequency of different bacterial taxa, based on the assignment of the DNA reads by the PhymmBL software and by 16S rRNA reads extracted from the metagenome, and compared with the PCR results obtained by Bik *et al.* (2010). The graph on the right indicates the relative contribution of each taxonomic group to the coding potential of the ecosystem, based on the COGs functional classification system. It can be observed that the functional contribution is not equal among taxa.

16S rRNA reads identified 186 sequences representing novel operational taxonomic units previously undetected by PCR amplification (Supplementary Table 3). Rarefaction curves and different diversity indexes based on the rRNA sequences obtained from the metagenomic reads indicate an estimate of 73–120 genera for dental plaque samples (Supplementary Table 1 and Supplementary Figure 2). A second approach to estimate diversity was the use of a LCA algorithm to classify all reads giving a hit in public databases at the taxonomic level for which the assignment was unambiguous (Huson *et al.*, 2007). Over 1.5 million reads were assigned by this procedure, confirming the presence of bacterial groups detected by 16S rRNA genes, but suggesting that a wider range of taxonomic groups was present (Supplementary Figure 1). Finally, the recently developed phymmBL binning procedure (Brady and Salzberg, 2009) was used to taxonomically assign 1.94 million reads from our data set. The results agreed again with the taxonomic distribution described by the 16S rRNA and the LCA approaches, but with further implication of other bacterial taxa. The results from these three methods show that the relatively small numbers of 16S genes in directly sequenced metagenomes are enough to describe the main taxonomic groups present without cloning or PCR-based biases, although at the expense of lower sequence depth. Some of the taxa found at low proportions in our data set were also detected by large-scale 16S rRNA cloning studies (Paster *et al.*, 2001; Bik *et al.*, 2010) but others were not (Figure 1). This could be not only due to lower amplification efficiency of these bacteria by universal primers, but also due to the detection of false positive hits by the LCA and phymmBL approaches.

Despite the low number of samples examined, interesting differences in diversity can be seen between healthy and diseased individuals. All three methods showed a tendency for Bacilli and Gamma-Proteobacteria to be more common in healthy individuals, whereas typically anaerobic taxa like Clostridiales and Bacteroidetes are more frequent in diseased samples (Figure 1, Supplementary Figure 1). Bacilli are particularly depleted in the two samples from within cavities, and one of them showed a high proportion of Actinobacteria. Reads assigned to beta-Proteobacteria (mainly Neisseriales) and TM7 were at very low proportions in diseased samples, and studies based on a larger number of individuals should test whether their presence could be associated to healthy conditions. Correspondence analysis between the metagenomes based on the taxonomic assignation by 16S rRNA reads showed that samples with poor oral health tended to cluster together, whereas different consortia of bacteria can be found in healthy individuals (Figure 2). Some genera, like *Rothia* or *Aggregatibacter* appear to be specifically associated to healthy samples, in agreement with PCR-based studies that compared bacterial diversity in healthy

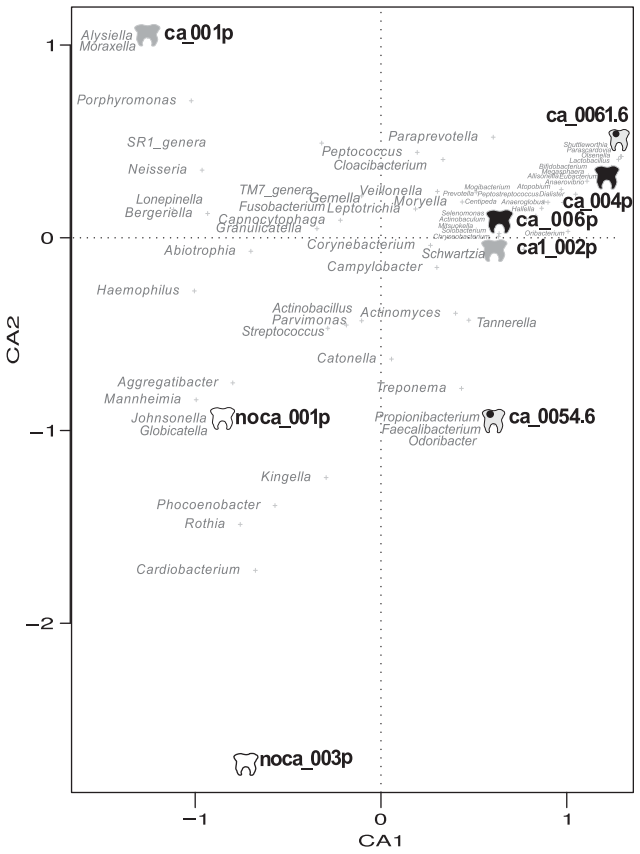

Figure 2 Correspondence analysis (CoA) of the bacterial diversity in oral samples based on 16S rRNA reads extracted from the metagenomes. The first axis successfully separates healthy from diseased individuals. The graph suggests bacterial genera which are potentially associated with absence of caries.

controls and diseased volunteers (Aas *et al.*, 2005, 2008; Corby *et al.*, 2005). The metagenomic recruitments also showed *Aggregatibacter* as one of the prevalent species in individuals without caries (see below).

Sequence similarity searches against 18S rRNA databases revealed very few significant hits against eukaryotic species. No rRNA reads were identified from *Candida* or other fungi that are regular inhabitants of the oral cavity, indicating that although these organisms are frequently detected by PCR amplification (Ghannoum *et al.*, 2010), they are probably present at low proportions. In sample CA-04, significant hits to the rRNA ITS region of the protozoan *Trichomonas tenax* were found. *Trichomonas tenax* is found particularly in the oral cavity of patients with poor oral hygiene and advanced periodontal disease (Kleinberg, 2002), and it has been shown to be involved in bronchopulmonary infections.

An effective tool to quantify the presence of selected species in metagenomes is provided by sequence recruitments (Rodriguez-Valera *et al.*, 2009). Individual metagenomic reads that give a hit over a certain identity threshold against a

reference bacterial genome are 'recruited' to plot a graph, which will vary in density depending on the abundance of that organism in the sample. If the average nucleotide identity displayed is above 94%, the recruitment is very likely made against reads of the same species (Konstantinidis and Tiedje, 2005). By comparing our metagenomes against the genomes of 1117 fully sequenced genomes available in databases, we were able to estimate the abundance of close relatives of these reference species in our samples (Supplementary Figure 3A). Interestingly, bacteria closely related to *Aggregatibacter* and *Streptococcus sanguis* were among the three with the highest level of recruitment in individuals without caries, in agreement with these species being more frequently amplified from the oral cavity of healthy individuals (Aas *et al.*, 2005; Corby *et al.*, 2005). On the other hand, *Streptococcus gordonii* and *Leptotrichia buccalis* were abundant in individuals with caries. Strains of *Veillonella parvula* were the most abundant in all individuals with caries and appeared to be common to all samples, but interestingly the recruitment plots show differences between strains (Supplementary Figure 4). For instance, the *Veillonella* present in the two healthy individuals shows a genomic island without recruitment, even at the protein level, between positions 2066–2094 Kb of the reference genome. Individuals with caries CA-04 and CA1-01 do contain this region, which includes CRISPR-associated genes, hypothetical proteins, a protein involved in DNA uptake and an amidophosphoribosyltransferase. This way, differences between strains of the same species can be identified which would pass unnoticed by 16S rRNA studies, and future work should identify whether those differential genes might be involved in pathogenesis. In addition, recruitment plots indicate that few taxa are normally dominant in each metagenome (Supplementary Figure 3B). This suggests that although bacterial diversity is indeed very large in the oral cavity, very few taxa account for most of the bacterial cells, and a big portion of the identified species are present at very low densities.

Functional diversity in the oral ecosystem

One of the powerful applications of LCA and phymmBL approaches is that each read with a significant hit can be assigned a taxonomic origin, and at the same time can also be related in many cases to a putative function. By relating taxonomy to function we have been able to predict what ecological or metabolic role each bacterial group can have. An example of this 'who can do what' approach can be seen in Figure 1 by using the COGs function classification system. It shows that categories are not equally distributed, and that some taxonomic groups are especially endowed for performing concrete functions. For example, a large portion of genes involved in defence mechanisms (that is, restriction endonucleases and drug efflux pumps) appear to be encoded by Bacilli. Other functions unequally distributed were cell motility genes in Clostridiales (mainly flagellar proteins) or signal transduction and carbohydrate metabolism in Bacilli (Figure 1, right). A more detailed functional analysis of the metagenome was performed using several systems for gene classification at different hierarchical levels. All pyrosequencing reads were compared against the conserved domains database, the Subsystems annotation environment (SEED) and the Tigrfams profiles (see Materials and methods section). Correspondence analysis (CoA) of the eight samples according to the functional assignment of the reads gave similar clustering patterns for the three function classification systems (Supplementary Figure 5). Samples from diseased individuals tended to cluster together, indicating that a similar set of functions were encoded in their metagenomes, and the two samples from individuals that had never suffered from caries, together with sample CA1-01 (with only one cavity at the moment of sampling), could be separated from the rest by the principal component. When the functional assignment of the oral microbiome was compared with that of the adult gut microbiome (Kurokawa *et al.*, 2007) a χ^2-test of independence revealed that the overall gut and oral functional roles depicted in the RAST subsystems are significantly different ($\chi^2_{(df=158)} = 17\,057.42$, $P < 2.2e{-}16$, $\phi = 0.123$), and this was supported also by clustering analysis where the oral samples clustered together (Figure 3), indicating that the gut and the mouth are two different ecosystems in terms of the relative frequencies of functions encoded in their metagenomes. It had previously been shown that the taxonomic diversity of the gut and oral ecosystems is clearly distinct (Bik *et al.*, 2010), despite the fact that clear examples of horizontal gene transfer have been shown between these two interconnected niches (Mira, 2007). Our data show large blocks of over-represented functions in the gut microbiome, while others appear over-represented in the oral samples (a detailed list of these functional categories is represented in Supplementary Figure 6). It is interesting to note that metabolic genes, like those involved in sugar uptake and assimilation, are enriched in gut bacteria together with adhesion proteins and prophage genes, whereas gene families related to oxidative and osmotic stress or iron scavenging are more frequent in the oral microbiome (Figure 3). Thus, the relative proportion of these functional categories provides important insights into the ecology of each ecosystem and the potential role of the corresponding microbiotas for human health.

Within the oral samples, individuals are clustered according to their health status (Figure 3). From an applied viewpoint, it is interesting that several functional categories are over-represented in samples from individuals without caries. Remarkable

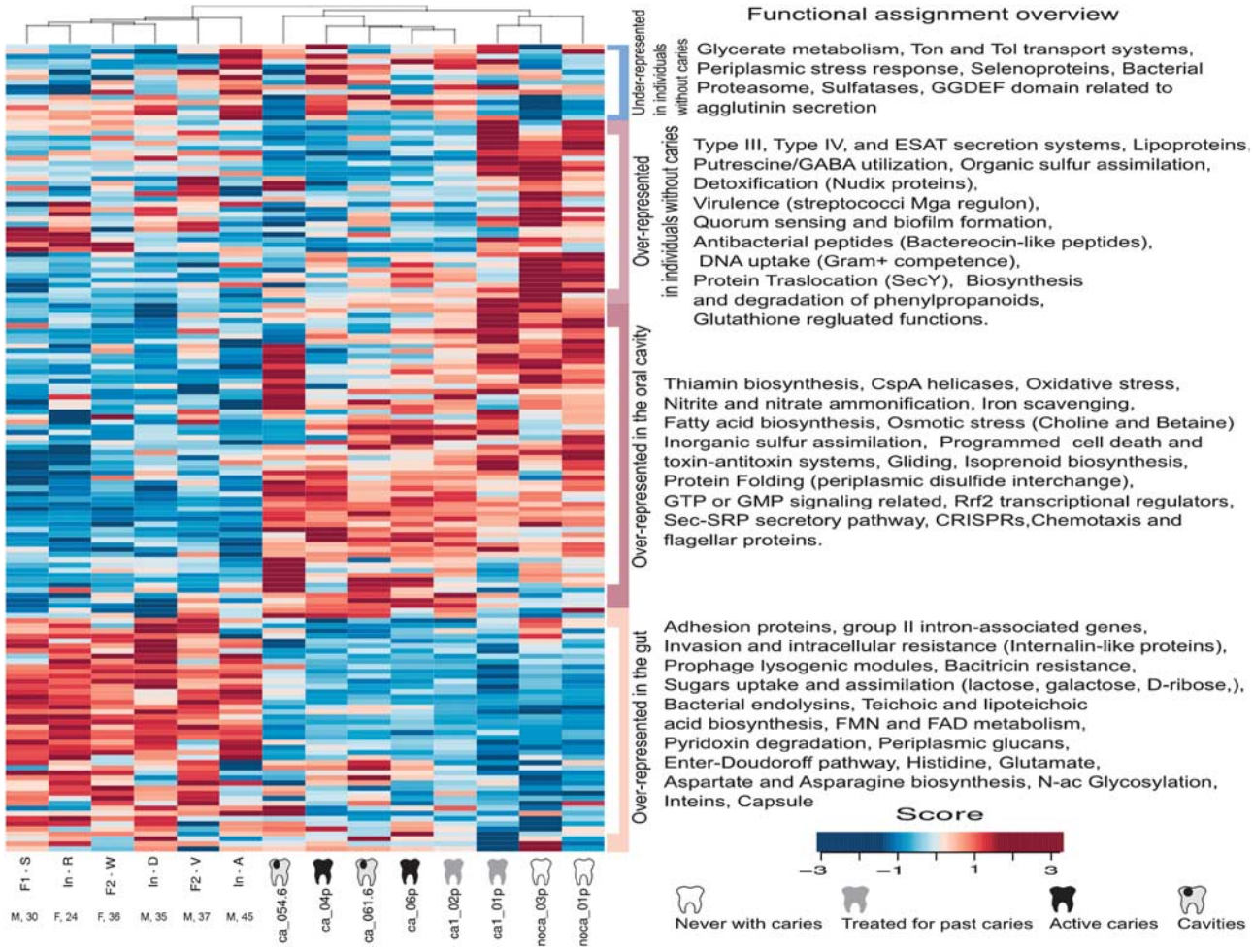

Figure 3 Functional profiles from oral and adult-gut metagenomic samples. Classification was based on Subsystem hierarchy 2 of MG-RAST. Counts were normalized to the total number of reads per sample and then normalized by function. Blue to red gradient indicates levels of under/over-representation. Large blocks of gene categories are over-represented in each of the two microbiotas, indicating that the gut and the oral cavity are two functionally distinct ecosystems. Within the oral microbiome, some functional roles are over-represented in individuals without caries. A full version of this figure indicating all 101 functional categories is included in Supplementary Figure 6. Sequences from the healthy adult-gut metagenomes were taken from Kurokawa *et al.* (2007). The age and sex of each individual are indicated below each label.

uprepresented genes in healthy individuals are involved in antibacterial peptides like bacteriocins (P-value = 2.95 e−7; q-value = 4.63 e−8), periplasmic stress response genes like *degS, degQ* (P = 2.46 e−46; q = 3.22 e−46), capsular and extracellular polysaccharides (P = 7.04 e−5; q = 8.5 e−6) and bacitracin stress response genes (P = 3.4 e−3; q = 3.24 e−4). Other functional categories were also over-represented but the difference was not statistically significant, like genes involved in quorum sensing and phospholipid metabolism. The higher presence of bacteriocin-related genes points at these bioactive compounds as promising potential anti-caries agents. Some gene features over-represented in individuals with active caries are involved in mixed-acid fermentation (P = 2.85 e−260; q = 2.65 e−259) and DNA uptake and competence (P = 6.29 e−8; q = 1.13 e−8). Finally, it must be underlined that some over-represented genes in healthy individuals have an unknown function, and future studies should elucidate whether they are involved in the protection of the teeth against cariogenic conditions.

Cavities are complex ecosystems

We were able to extract sufficient DNA for 454 pyrosequencig in two samples from individual teeth, one at an intermediate stage and the other one at an advanced stage of caries development (dentin lesion). Given that mutans streptococci initially were considered to be the main ethiological agents of dental caries (Loesche, 1986), it is not surprising that most strategies against this disease have aimed at targeting *Streptococcus mutans*. These include the development of a vaccine using known surface antigens, passive immunization strategies that could neutralize the bacterium, the co-aggregation of *S. mutans* to probiotic strains or the use of specific inhibitors of *S. mutans* proteins, among others (Russell *et al.*, 2004). In addition, the

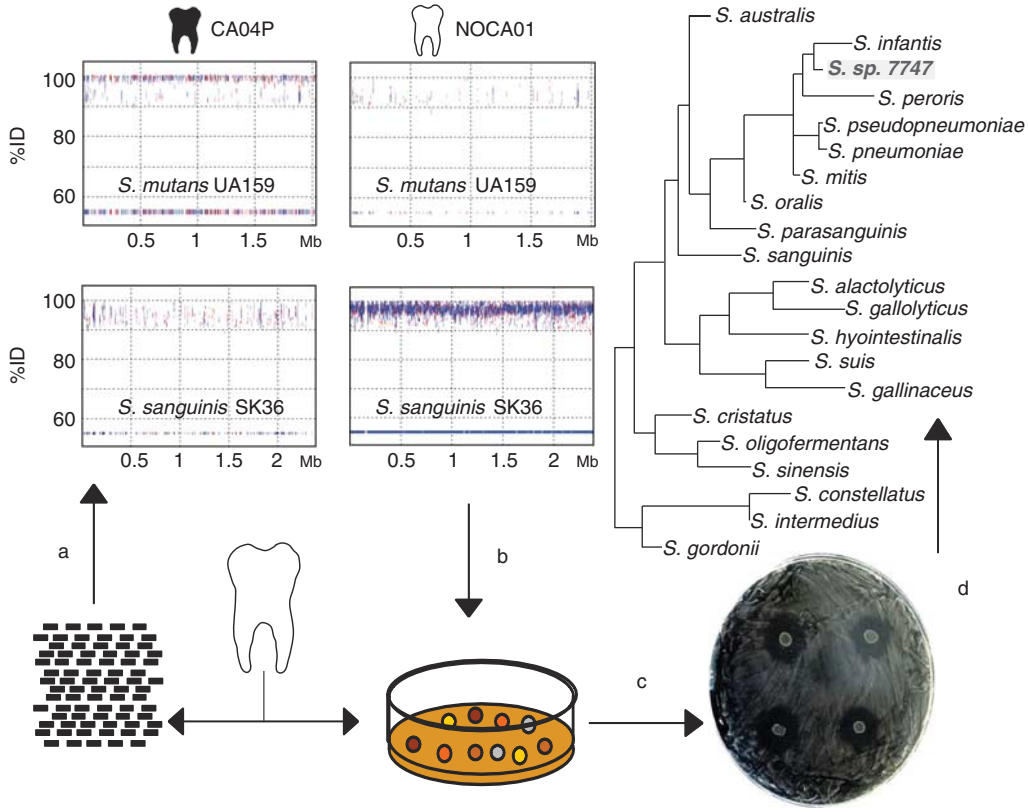

Figure 4 Searching of bacterial strains with a potential antagonistic effect against cariogenic bacteria. Metagenomic recruitment plots are used to detect the species (**a**), which are at low frequencies in individuals with caries but are among the most common in caries-free subjects. These species are then selected based on culture conditions and microscopic examination (**b**). The isolates are grown in solid media to provide an inhibition screening against caries-producing bacteria (**c**), selecting the strains that display inhibition rings (**d**), such as the *Streptococcus* strain 7747. Sequencing the genome of these inhibitory strains and comparing it against the metagenome of caried individuals must confirm that these strains are absent under diseased conditions.

presence of mutans streptococci in children is typically associated to caries risk in oral-health evaluation protocols (Ge *et al.*, 2008). However, pioneering molecular-based studies of cavities have failed to amplify mutans streptococci by PCR or hybridization in a significant proportion of cavities, suggesting that other bacterial genera like *Lactobacillus*, *Actinomyces* or *Bifidobacterium* could be involved in the disease (Aas *et al.*, 2008; Becker *et al.*, 2002). Recent molecular work has confirmed this finding and expanded the list of potential cariogenic bacteria to other species like *Veillonella*, *Propionibacterium* and *Atopobium* (Aas *et al.*, 2008), most of them are poorly studied bacteria. The metagenomes of cavities studied here showed an almost complete absence of *S. mutans*. However, they displayed a large taxonomic diversity, which are included among the most common genera, *Veillonella, Corynebacterium* or *Leptotrichia* (Supplementary Table 4). Some of these bacteria, particularly *Veillonella*, have been shown to be predominant at all stages of caries progression (Aas *et al.*, 2008) and under high-glucose conditions, and appear to be implied in acid production (Bradshaw and Marsh, 1998). Interestingly, consortia between *Veillonella alcalescens* and *S. mutans* were shown to produce more acid than any one of these species separately (Noorda *et al.*, 1988), suggesting that synergistic effects probably take place, as it has been demonstrated in other complex microbial communities. Thus, although these data are based on the metagenomes from only two cavities, they favour a nonspecific plaque hypothesis for the development of dental caries (Marsh, 1994; Kleinberg, 2002). Further work should elucidate the potential role these bacteria had other than mutans streptococci in the progression of caries, as well as their synergistic and antagonistic interactions. The forecoming improvements in the amount of DNA required for next-generation sequencing techniques will allow a metagenomic study of cavities at different stages of development, including initial, white-spot lessions. This is important because mutans streptococci could be instrumental at initial stages of caries, after which other species could colonize the niche. If caries is confirmed to be a polymicrobial disease, this should be taken into account for future therapeutic strategies. For instance, a potential solution for immunization strategies could pass through the selection of vaccine targets shared by different pathogens involved in the process of tooth decay (Mira *et al.*, 2004; Mira, 2007).

Search for potential probiotics through metagenomics
The existence of a small proportion of the human adult population that has never suffered from dental caries has led some authors to suggest the presence of some bacterial species with a potential antagonistic effect against cariogenic bacteria (Corby *et al*., 2005). Bacterial replacement of pathogenic strains by innocuous isolates obtained from healthy individuals has been successfully shown to prevent pharynx infections and is the basis for probioticts preventing infectious disease in the gut and other human niches (Tagg and Dierksen, 2003). Metagenomic recruitment of cariogenic bacteria against the oral microbiome of healthy individuals shows a complete absence of *S. mutans* and *S. sobrinus*. Interestingly, the lack of detection of the cariogenic bacteria is accompanied by an intense recruitment of other streptococci (mainly those related to *S. sanguis*) and *Neisseria*, which comprise the most abundant genera in these individuals (Supplementary Figure 3B). Given the possibility that isolates of these dominant genera could be involved in antagonistic interactions with cariogenic bacteria, fresh dental plaque samples from 10 healthy individuals (including those from which the metagenomic sequences were obtained) were collected and used for culturing under conditions optimal for the growth of neisserial and streptococcal species. After microscopic examination, diplococci and streptococci were selected, providing a collection of 249 isolates. Those that could be grown on the same culture medium as *S. mutans* and *S. sobrinus* were transferred to a loan culture of these cariogenic bacteria. This simple screening identified 16 strains that displayed inhibition rings (Figure 4). PCR amplification of the 16S rRNA gene identified most of them as streptococci, with a 96–99% sequence identity to *S. oralis*, *S. mitis* and *S. sanguis*. Thus, this metagenomic approach allowed us to quantify the most abundant bacteria and confirms the previously hypothesized presence of bacteria with a protective effect against cariogenic species. This effect appears to be direct (that is, inhibitory), but other indirect effects such as stimulation of the immune response or direct competition for the same substrate or niche cannot be ruled out. Future research on these isolates should aim at identifying the secreted compounds responsible for the inhibition of caries-producing bacteria, and metagenomic libraries of dental plaque DNA may prove useful in this respect (Seville *et al*., 2009). Our own inhibition screenings performed on metagenomic fosmid libraries from dental plaque of healthy individuals against cariogenic bacteria suggest that antimicrobial peptides are among the products causing the inhibition. We propose the probiotic use of these anti-cariogenic bacteria or the utilization of the antibiotics they encode as promising new therapies against dental caries and other oral diseases (Devine and Marsh, 2009).

Conclusion

We have shown that the direct pyrosequencing of human samples is a feasible approach to study the human microbiome, which would obviate the biases imposed by cloning and PCR and that would provide a more complete view of human-related bacterial communities beyond their composition inferred from the 16S rRNA gene (Ghai *et al*., 2010; Xie *et al*., 2010). Even in samples with a large proportion of human DNA such as cavities, the large throughput of next-generation sequencing has provided enough sequences to gain insights into the microbiology of caries, suggesting that it is the outcome of a complex bacterial community. Despite the limited number of samples analyzed in this first study, important differences between healthy and diseased sites and individuals can be observed at the taxonomic and functional level, suggesting that the dental plaque of individuals that have never suffered from caries can be a genetic reservoir of new anticaries compounds and probiotics. Future population-based studies must evaluate whether the trends described in this study hold when higher sample sizes are used. We hope that these results stimulate further sequencing of the oral metagenome and metatranscriptome in the future as a tool to understand and combat the development of oral diseases.

Acknowledgements

This work was funded by the following projects from the Spanish MICINN: SAF2009-13032-C02-02 from the I+D program, BIO2008-03419-E from the EXPLORA program and MICROGEN CSD2009-00006 from the Consolider-Ingenio program. We thank Professor F Rodriguez-Valera and Professor Siv G Andersson for their advice and comments, and three anonymous referees for their constructive comments that significantly improved the manuscript.

References

Aas JA, Griffen AL, Dardis SR, Lee AM, Olsen I, Dewhirst FE *et al.* (2008). Bacteria of dental caries in primary and permanent teeth in children and young adults. *J Clin Microbiol* **46**: 1407–1417.

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.

Alstrup S, Gavoille C, Kaplan HRT. (2004). Nearest common ancestors: a survey and a new Algorithm for a distributed environment. *Theory Comp Syst* **37**: 441–456.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Becker MR, Paster BJ, Leys EJ, Moeschberger ML, Kenyon SG, Galvin JL *et al.* (2002). Molecular analysis of bacterial species associated with childhood caries. *J Clin Microbiol* **40**: 1001–1009.

Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, Mongodin EF *et al.* (2010). Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J* **4**: 962–974.

Bradshaw DJ, Marsh PD. (1998). Analysis of pH-driven disruption of oral microbial communities *in vitro*. *Caries Res* **32**: 456–462.

Brady A, Salzberg SL. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* **6**: 673–676.

Chen T, Yu W-H, 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: J Biol Databases Curation* **2010**: baq013.

Corby PM, Lyons-Weiler J, Bretz WA, Hart TC, Aas JA, Boumenna T *et al.* (2005). Microbial risk indicators of early childhood caries. *J Clin Microbiol* **43**: 5753–5759.

Darveau RP. (2010). Periodontitis: a polymicrobial disruption of host homeostasis. Nature reviews. *Microbiology* **8**: 481–490.

de Lillo A, Ashley FP, Palmer RM, Munson MA, Kyriacou L, Weightman AJ *et al.* (2006). Novel subgingival bacterial phylotypes detected using multiple universal polymerase chain reaction primer sets. *Oral Microbiol Immunol* **21**: 61–68.

Devine DA, Marsh PD. (2009). Prospects for the development of probiotics and prebiotics for oral applications. *J Oral Microbiol* **1**: 1–11.

Ge Y, Caufield PW, Fisch GS, Li Y. (2008). *Streptococcus mutans* and *Streptococcus sanguinis* colonization correlated with caries experience in children. *Caries Res* **42**: 444–448.

Ghai R, Martin-Cuadrado AB, Molto AG, Heredia IG, Cabrera R, Martin J *et al.* (2010). Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J* **4**: 1154–1166.

Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, Naqvi A *et al.* (2010). Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathogens* **6**: e1000713.

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS *et al.* (2006). Metagenomic analysis of the human distal gut microbiome. *Science (New York, NY)* **312**: 1355–1359.

Gomez-Alvarez V, Teal TK, Schmidt TM. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J* **3**: 1314–1317.

Grice EA, Kong HH, Renaud G, Young AC, Bouffard GG, Blakesley RW *et al.* (2008). A diversity profile of the human skin microbiota. *Genome Res* **18**: 1043–1050.

Huson DH, Auch AF, Qi J, Schuster SC. (2007). MEGAN analysis of metagenomic data. *Genome Res* **17**: 377–386.

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

Kleinberg I. (2002). A mixed-bacteria ecological approach to understanding the role of the oral bacteria in dental caries causation: an alternative to Streptococcus mutans and the specific-plaque hypothesis. *Crit Rev Oral Biol Med* **13**: 108–125.

Konstantinidis KT, Tiedje JM. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102**: 2567–2572.

Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A *et al.* (2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14**: 169–181.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C *et al.* (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.

Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Osterås M *et al.* (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods* **79**: 266–271.

Loesche WJ. (1986). Role of Streptococcus mutans in human dental decay. *Microbiol Rev* **50**: 353–380.

Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH *et al.* (2009). CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* **37**: D205–D210.

Marsh PD. (1994). Microbial ecology of dental plaque and its significance in health and disease. *Adv Dental Res* **8**: 263–271.

Marsh PD. (2006). Dental plaque as a biofilm and a microbial community—implications for health and disease. *BMC Oral Health* **6**(Suppl 1): S14.

Marsh PD. (2010). Microbiology of dental plaque biofilms and their role in oral health and caries. *Dental clin North Am* **54**: 441–454.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* (2008). The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinfo* **9**: 386.

Mira A. (2007). Horizontal gene transfer in oral bacteria. In: Rogers AH (ed). *Oral Molecular Microbiology*. Horizon Scientific Press: Norfolk, UK, pp 65–85.

Mira A, Pushker R, Legault BA, Moreira D, Rodríguez-Valera F. (2004). Evolutionary relationships of Fusobacterium nucleatum based on phylogenetic analysis and comparative genomics. *BMC Evol Biol* **4**: 50.

Noorda WD, Purdell-Lewis DJ, van Montfort AM, Weerkamp AH. (1988). Monobacterial and mixed bacterial plaques of Streptococcus mutans and Veillonella alcalescens in an artificial mouth: development, metabolism, and effect on human dental enamel. *Caries Res* **22**: 342–347.

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.

Petersen PE. (2004). [Continuous improvement of oral health in the 21st century: the approach of the WHO Global Oral Health Programme]. *Zhonghua kou qiang yi xue za zhi = Zhonghua kouqiang yixue zazhi = Chin J Stomatol* **39**: 441–444.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**: 59–65.

Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM *et al.* (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6**: 639–641.

Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F *et al.* (2009). Explaining microbial population genomics through phage predation. Nature reviews. *Microbiology* **7**: 828–836.

Russell MW, Childers NK, Michalek SM, Smith DJ, Taubman MA. (2004). A Caries Vaccine? The state of

the science of immunization against dental caries. *Caries Res* **38**: 230–235.

Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC *et al.* (2007). TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res* **35**: D260–D264.

Seville LA, Patterson AJ, Scott KP, Mullany P, Quail MA, Parkhill J *et al.* (2009). Distribution of tetracycline and erythromycin resistance genes among human oral and fecal metagenomic DNA. *Microbial Drug Resist (Larchmont, NY)* **15**: 159–166.

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

Tagg JR, Dierksen KP. (2003). Bacterial replacement therapy: adapting 'germ warfare' to infection prevention. *Trends Biotechnol* **21**: 217–223.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV *et al.* (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinfo* **4**: 41.

Vaishampayan PA, Kuehl JV, Froula JL, Morgan JL, Ochman H, Francino MP. (2010). Comparative metagenomics and population dynamics of the gut microbiota in mother and infant. *Genome Biol Evol* **2010**: 53–66.

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.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T *et al.* (2009). gplots: Various R programming tools for plotting data. The Comprehensive R Archive Network. http://cran.r-project.org/package=gplots.

Watabe K, Nishi M, Miyake H, Hirata K. (1998). Lifestyle and gastric cancer: a case-control study. *Oncol Rep* **5**: 1191–1194.

White JR, Nagarajan N, Pop M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples (CA Ouzounis, Ed.) *PLoS Comput Biol* **5**: e1000352.

Wu T, Trevisan M, Genco RJ, Dorn JP, Falkner KL, Sempos CT. (2000). Periodontal disease and risk of cerebrovascular disease: the first national health and nutrition examination survey and its follow-up study. *Arch Int Med* **160**: 2749–2755.

Xie G, Chain PSG, Lo C-C, Liu K-L, Gans J, Merritt J *et al.* (2010). Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Mol Oral Microbiol* **25**: 391–405.

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