molecular oral microbiology

molecular oral microbiology

Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing

G. Xie[1,2], P.S.G. Chain[1,2], C.-C. Lo[1,2], K.-L. Liu[1], J. Gans[1], J. Merritt[3] and F. Qi[3]

1 Oralgen, Genome Science Group, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA
2 Metagenomics Program, Joint Genome Institute, Walnut Creek, CA, USA
3 University of Oklahoma Health Sciences Center, College of Dentistry, Oklahoma City, OK, USA

Correspondence: Gary Xie, Oralgen Database, Genome Science Group (B-6), Bioscience Division, MS-M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA Tel.: 5056672310; fax: 5056653024; E-mails: Gary Xie (xie@lanl.gov), Patrick Chain (pchain@lanl.gov), Felicia Qi (feliciaqi@ouhsc.edu)

Keywords: dental plaque; fragment recruitment; human oral microbiome; metagenomic sequencing; next generation sequencing; oralgen
Accepted 16 June 2010

SUMMARY

Human dental plaque is a complex microbial community containing an estimated 700 to 19,000 species/phylotypes. Despite numerous studies analysing species richness in healthy and diseased human subjects, the true genomic composition of the human dental plaque microbiota remains unknown. Here we report a metagenomic analysis of a healthy human plaque sample using a combination of second-generation sequencing platforms. A total of 860 million base pairs of non-human sequences were generated. Various analysis tools revealed the presence of 12 well-characterized phyla, members of the TM-7 and BRC1 clade, and sequences that could not be classified. Both pathogens and opportunistic pathogens were identified, supporting the ecological plaque hypothesis for oral diseases. Mapping the metagenomic reads to sequenced reference genomes demonstrated that 4% of the reads could be assigned to the sequenced species. Preliminary annotation identified genes belonging to all known functional categories. Interestingly, although 73% of the total assembled contig sequences were predicted to code for proteins, only 51% of them could be assigned a functional role. Furthermore, $\sim$ 2.8% of the total predicted genes coded for proteins involved in resistance to antibiotics and toxic compounds, suggesting that the oral cavity is an important reservoir for antimicrobial resistance.

INTRODUCTION

The human oral cavity is colonized by a complex microbial community that plays an important role in dictating the oral health status of the host (for review, see Marsh, 1994, 2006; Haffajee & Socransky, 2005, 2006; Socransky & Haffajee, 2005; Paster *et al.*, 2006). Oral diseases (such as dental caries, periodontitis, halitosis) develop as a result of major disruptions of the ecological balance in the oral microbial community as the result of environmental changes in the oral cavity. Consequently, it is of paramount importance to understand the genomic composition of the oral microbial community and the forces that shape this ecological balance to prevent and manage the progression of disease. In the past 50 years, numerous studies have characterized the community composition of the oral microbiota (Kroes *et al.*, 1999; Paster *et al.*, 2001, 2006; Becker *et al.*, 2002; Kumar *et al.*, 2003; Aas *et al.*, 2005, 2008; Preza *et al.*, 2008). Using culture-dependent and independent methods, estimates of oral biodiversity have implicated >700 different microbial species (Socransky *et al.*, 1998; Kroes *et al.*, 1999; Paster *et al.*, 2001, 2006; Aas *et al.*, 2008). Recently, several studies have employed next-generation sequencing technologies to analyse the species richness of the oral microbiota (Keijser *et al.*, 2008; Lazarevic *et al.*, 2009; Zaura *et al.*, 2009). Estimates from one of these studies suggested that up to 19,000 phylotypes may exist in the human oral cavity (Keijser *et al.*, 2008). Despite these tremendous advancements in our understanding of community structure, only a minute fraction of the genomic content within the plaque community is known. As a result, even less is known about the ecological roles of most of these species/phylotypes in mediating plaque homeostasis. In this study, we conducted a shotgun metagenomic analysis of dental plaque from a healthy human volunteer using a combination of 454 and Illumina sequencing platforms. Using this approach, we were able to successfully assemble the first gene catalog of the dental plaque microbiota. In the process, we also developed new strategies for metagenome sequence assembly and data analysis. With these data, we were able to obtain the first glimpse of the genomic contents of a human plaque microbiota.

METHODS

Plaque collection and DNA isolation

Upon Institutional review Board approval (#14107), supragingival and subgingival plaques were collected from a caries-free and periodontally healthy volunteer using sterile toothpicks for supragingival plaque, sterile curettes for subgingival plaque, and dental floss for interproximal regions. To increase plaque accumulation, brushing and flossing were restricted for 24 h before the plaque samples were taken. Plaque from eight teeth (four anterior and four posterior) were collected, combined, and suspended in an Eppendorf tube containing 480 μl 50 mM ethylenediaminetetraacetic acid. Freshly prepared lysozyme was added to a final concentration of 10 mg ml^{-1}, and the tube was incubated at 37°C for 3 h. For total chromosomal DNA isolation, the Wizard Genomic DNA Purification kit (Promega, Madison, WI) for bacteria was used. DNA was isolated following the manufacturer's instructions. We were able to obtain 15 μg high-quality DNA, which was sufficient for sequencing.

Sequencing and quality control

Metagenomic DNA sequence data were generated using a combination of two sequencing technologies, the Roche 454 FLX system using titanium kits and version 2.3 software, and the Illumina Genome Analyzer IIx (76 cycles) using sequencing control software version 2.5 and version 3.0 cluster generation and sequencing kits. The resulting 454 and Illumina reads were subjected to quality filtering using the LUCY program (http://lucy.sourceforge.net/) (Chou & Holmes, 2001), which discarded reads with poor quality and trimmed low-quality regions. Contaminating host sequences were removed after detecting top significant hits to human sequences using a BLASTN search (Altschul *et al.*, 1990, 1997) of the GenBank non-redundant sequence (NR) database.

Metagenome assembly, mapping, and annotation

The curated 454 and Illumina data were assembled using the NEWBLER (http://www.454.com/products-solutions/analysis-tools/gs-de-novo-assembler.asp) and VELVET (http://www.ebi.ac.uk/~zerbino/velvet/) programs, respectively. A number of different hybrid assemblies of combined 454 and Illumina reads were performed varying parameters for fragment length and estimated coverage, and the best assemblies, based on contig sizes and total number of base pairs assembled into large contigs, were selected as the final combined assembly.

Metagenome Rapid Annotation using Subsystem Technology (MG-RAST; http://metagenomics.nmpdr.org/) (Aziz *et al.*, 2008) and the Integrated Microbial Genomes (IMG) system with Microbiome Samples Expert Review (IMG-M ER) (http://img.jgi.doe.gov/mer/) (Markowitz *et al.*, 2008) served as a base annotation. The read-based MG-RAST annotation used BLASTX (ver. 2.0.11 (Altschul *et al.*, 1990)) similarity search against the SEED subsystems (an annotation/analysis tool provided by FIG (The Fellowship for Interpretation of Genomes); http://www.theseed.org/wiki/index.php/Home_of_the_SEED) (Overbeek *et al.*, 2005). A histogram of the distribution of 454 reads GC% and the distribution among five major phylotypes assigned by using the MG-RAST annotation is shown in Fig. S1. The species/phylotype level taxa were estimated by counting how many reference genomes all 454 reads matched using the MG-RAST

'Phylogenetic Profile' feature. Ribosomal RNA (rRNA) and putative virulence-related genes were flagged using the MG-RAST program (Aziz *et al.*, 2008).

The IMG-M annotation is based on the combined approach of the BLASTX similarity search and *de novo* gene prediction. Basically, metagenomic sequences were split into three bins: 80–299 base pairs (bp), 300–699 bp, and ≥700 bp. For the shortest bin, MULTIBLASTX was used against the IMG version of the non-redundant database (IMG-NR) with an out-of-frame penalty of 25, which detects frame-shifted genes. All frameshift fragments are joined afterwards. For the mid-length bin, MULTIBLASTX (Carter *et al.*, 2001), metagenomic versions of METAGENE (http://metagene.cb.k.u-tokyo.ac.jp/) (Noguchi *et al.*, 2006) and GENEMARK (http://exon.biology.gatech.edu/) (Zhu *et al.*, 2010), were used with the preference given to the genes predicted by MULTIBLASTX, then by GENEMARK (in the spaces between MULTIBLASTX genes), then by METAGENE. For the longest bin, only METAGENE and GENEMARK were used with the same order of preference (GENEMARK > METAGENE).

In-house GenBank similarity searches for 454 and Illumina reads

Both BLASTN and BLASTX similarities were performed in parallel on an Intel-based cluster using the National Center for Biotechnology Information (NCBI) BLASTALL program (version 2.2.21). BLASTN was used to find similarities of all Illumina reads (partitioned into 991 files of approximately 15,000 reads each) with sequences in the GenBank NT (non-redundant nucleotide database) database (i.e. all GenBank, European Molecular Biology Laboratory, DNA Data Bank of Japan and Protein Data Bank sequences, but no expressed sequence tags, sequence tagged sites, Genome Survey Sequences, environmental samples or phase 0, 1, or 2 high throughput genomic sequences), downloaded on 7 March 2010. BLASTX was used to find more distant similarities between 454 reads (partitioned into 114 files of approximately 1000 reads each) and sequences in the GenBank NR database (i.e. all non-redundant GenBank coding region sequence translations, Protein Data Bank, SwissProt, Protein Information Resource and Protein Research Foundation sequences, but no environmental samples from Whole Genome Shotgun projects), downloaded on 5 April 2010. The size of the dataset made the BLASTX of Illumina reads computationally prohibitive.

Community composition profiling

Multiple complimentary methods were used to assess the Community Composition.

1 16S rRNA-based approach: 454 reads with similarity to rRNA were first identified in MG-RAST (http://metagenomics.nmpdr.org/), and then searched against the ribosomal RNA databases [Ribosomal Database Project (RDP), Silva ssu rRNA, and Greengene] using BLASTN with an e-value cut-off of $1e^{-5}$ and a minimum alignment length of 50 bp. Similarly, BLASTN comparisons of these reads were made against the Human Oral Microbiome Database (HOMD; http://www.homd.org) 16s rRNA sequences.

2 Phylogenetic marker protein-based approach: Protein files from both MG-RAST and IMG-M ER annotations were used as input for the AMPHORA program (Wu & Eisen, 2008). Homologs of the 31 pre-built phylogenetic marker genes were extracted. Each marker gene sequence identified from this analysis was individually aligned to the corresponding reference sequences, trimmed using a pre-built mask, and inserted into the reference tree using the RAxML (Stamatakis, 2006) maximum parsimony method with 100 bootstrap replicates to assess the confidence of the branching order. A tree-based bracketing algorithm was then employed as described in Stamatakis (2006) to assign a phylotype to each query sequence. Starting from the immediate ancestor of the query sequence and moving toward the root of the tree, the first internal node (N1) whose bootstrap support exceeded a cut-off of 70% was identified, and the common NCBI taxonomic level, shared by all descendants of this node, represents the most conservative taxonomic prediction for the query sequence. The taxonomic rank assignment for each sequence is summed to assess both organism identity and relative abundance.

3 Gene-based approach: all 113,000 454 reads were searched against the GenBank NR database using BLASTX, followed by MEGAN (Huson *et al.*, 2007) analysis. This software reads the results of a BLAST comparison as input and attempts to place each read on a node in the NCBI taxonomy. This is performed by the Lowest Common Ancestor

algorithm, which assigns each read to the lowest common ancestor in the taxonomy from a subset of the best scoring matches in the BLAST result with default value settings. The 454 reads that have no BLAST matches are assigned to the special node 'no hits' and those unassigned for algorithmic reasons (e.g. below an applied threshold) are placed on the special node 'unassigned'. The result of the analysis is displayed as a tree representation of the NCBI taxonomy. Meanwhile, all 454 reads and all contigs obtained from the 454 plus Illumina hybrid assembly were searched against the SEED and IMG databases, respectively, using BLASTX and BLASTP, and the top BLAST-based taxonomy assignment was obtained through both MG-RAST and IMG-M ER servers.

Metagenome sequence recruitment

An in-house sequence recruitment program was used to align each read to reference genomes or genomic fragments. The mapping of 454/Illumina reads against 454 contigs was performed by the MOSAIK aligner (http://bioinformatics.bc.edu/marthlab/Mosaik) with 0.05% mismatch and the number of aligned reads was used to estimate the sequence coverage and abundance profile of that contig in the sampled community. Furthermore, the Human Microbiome Project (HMP) oral reference genomes were downloaded from the HMP DACC website (http://www.hmpdacc-resources.org/) and concatenated as a large reference sequence for alignment with all 454 and Illumina reads using MUMmer (Kurtz *et al.*, 2004). Coordinate files produced from MUMmer alignments were parsed using an in-house developed JAVA program and alignment plots of the 454 and Illumina reads against the reference sequences were created using an R script (http://www.r-project.org/).

Ecosystems comparison

Different ecosystem datasets were downloaded from the MEGAN website (http://www-ab.informatik.uni-tuebingen.de/software/megan/comparative): the selected marine metagenome data are based on ~ 145,000 Sanger reads that were randomly sampled from the Global Ocean Survey project (Yooseph *et al.*, 2007); data of the soil metagenome are based on ~ 140,000 Sanger reads from the Iowa soil sample (Tringe & Rubin, 2005); the mouse gut summary dataset (obese1) is based on ~ 675,000 454 reads (Turnbaugh *et al.*, 2006) and the human gut metagenome is based on ~ 145,000 Sanger reads from (Gill *et al.*, 2006). After multiple datasets were loaded, MEGAN was used to compare the number of reads that have been assigned to each node (normalized based on the sample size) from different datasets. For phylogenetic diversity comparison, the NCBI taxonomy tree was collapsed at phylum level and a bar chart summarizing the number of reads assigned at the desired rank of the NCBI taxonomy was generated. Meanwhile, the prokaryotic attributes were also obtained using MEGAN. The NCBI 'Prokaryotic Attributes Table' that lists the attributes of microbes, such as their cellular features, environment, temperature, pathogenicity, and relevance for diseases, was downloaded and represented as nodes in tree view. If a taxon had been detected at the species level by MEGAN and this organism was known to have a certain attribute, it would be inserted as a child node beneath this property node. A broad overview about the physiological and environmental features of microbial organisms within metagenome samples were obtained by using this microbial attributes feature of MEGAN.

The functional comparison of three ecosystems (human oral, human gut, and mouse gut) was conducted using the MG-RAST Metagenome Heat Map feature, which computes the metabolic profiles based on SEED subsystem classifications of all 454 reads. A minimum e-value of $1e^{-5}$ was used as the cut-off to identify unique genes for each ecosystem and the intersection of genes among them. Meanwhile, both Function Comparisons and the Functional Category Comparison feature of IMG-M ER were used to compare all predicted genes from dental plaque with two human gut samples (Gill *et al.*, 2006), in terms of the relative abundance of the protein families (Clusters of Orthologous Groups of proteins; COGs) and the genes assigned to different functional categories (COG Pathway, Pfam Category, TIGRfam subroles), with estimates of the statistical significance of the observed differences. The comparison result includes an assessment of statistical significance of the relative frequencies of the genes assigned to different functional categories.

Data sharing

The metagenome data have been deposited in the MG-RAST database http://mg-rast.mcs.anl.gov/mg-rast/FIG/linkin.cgi?metagenome=4446622.3, and can be accessed after registration with the web server. Raw sequence reads can also be downloaded from the Oralgen site at http://www.oralgen.lanl.gov/oralgen/downloads/supplemental_files/supplemental_files.html.

RESULTS AND DISCUSSION

Metagenomic sequencing of a human dental plaque microbiome

To obtain a first glimpse of the metagenomic composition of the human dental plaque microbiome, we sequenced the plaque sample of a caries-free and periodontally healthy human volunteer using the massively parallel sequencing platforms 454 Titanium and Illumina GA iiX. To ensure enough DNA was obtained for sequencing (each sequence platform requires 5–10 μg high-molecular-weight DNA), both supragingival and subgingival plaques were taken from eight teeth and combined. A total of 15 μg DNA was obtained. This is probably the maximal amount of DNA one could obtain from a healthy subject without the volunteer suffering more than 1 day of no oral hygiene.

To obtain the sequence, one quarter-channel of a 454 and one lane of Illumina were used for this sample. The 454 run yielded ~1 77 K reads compared with ~ 16 M reads from Illumina. These reads were first checked for quality, which showed that ~ 176 K 454 reads (99%) and ~ 15 M Illumina reads (91%) were of high quality (quality score >20), indicating that our sequencing protocols were highly effective. Next, the reads were analysed for host contamination using a BLASTN search of the GenBank NT database, followed by parsing any top BLAST hits to human chromosomes. Our results indicated that approximately one-third of the reads were of host (human) origin (see Table S1). This manageable level of human contamination suggests that plaque DNA sample collection and preparation procedures were appropriate. It should be noted however that if samples are to be taken from deep periodontal pockets, fluid in the pocket should be removed before plaque on the tooth surface is taken to avoid high human cell contamination.

After eliminating the human-like sequence reads, the remaining sequences of each technology were assembled separately using an optimal assembler (NEWBLER for 454 data and VELVET for Illumina plus 454 data). We used the VELVET assembler to combine Illumina and 454 reads and used a number of assembly parameters for each data input, which resulted in contigs that differed quantitatively from one another (see Table S2). This combined strategy yielded around 3.2–15.2X more total assembled base pairs and 9.2–86.9X more total number of contigs, as well as up to 6.9X longer contigs when compared with the 454 alone contigs, depending on which parameter was used. The VELVET hash size 35 was considered to have the best assembly based on the amount of cumulative data assembled into the largest contigs. This study represents one of the first to combine two of the most recent complimentary platforms, without the use of traditional (and longer) Sanger sequencing data, to generate and assemble a metagenome. We have shown that the deep sequencing using short reads from Illumina complement the longer, and therefore easier-to-assemble, 454 reads to generate longer contigs. It appears that for metagenomes such as this one, such a 'hybrid' approach yields the best results, although further study is required to see if more coverage using only one platform would be sufficient.

Community composition of the plaque microbiome

With a combination of complementary strategies (see Methods), we were able to obtain a largely unbiased assessment of the community composition of a dental plaque microbiome. The assessment of both organism identity and relative abundance from read-based methods are summarized in Fig. 1 and Table 1. Although some differences exist among different analysis methods in terms of proportion of the predicted phyla within the combined sequencing pool, the relative proportions of major phyla, (i e. Firmicutes, Proteobacteria, Actinobacteria, Fusobacteria, and Bacteroidetes) are similar between different methods. Moreover, at the phylum level, these data were also consistent with previous 16S rRNA-based community profiling surveys (Keijser *et al.*, 2008; Zaura *et al.*, 2009). The only exception is a recent Illumina 16S rRNA survey that targeted the variable

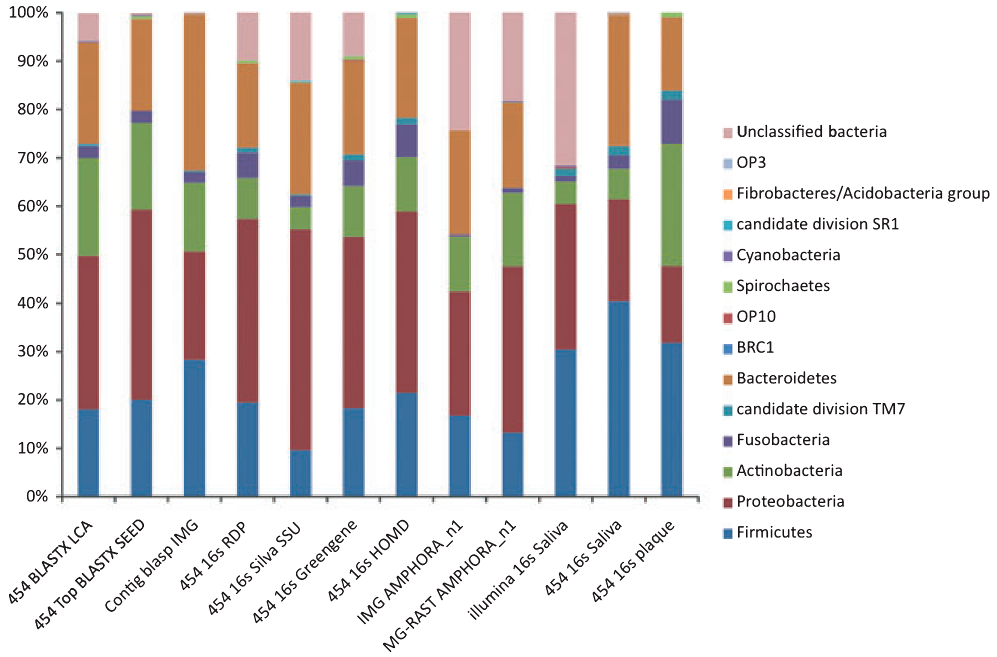

Figure 1 Proportions of taxonomic assignments at the phylum level. Four hundred and fifty-four reads assigned to each major phylum are represented by bars in the histogram. Their relative height represents the percentage of reads that can be placed at phylum level of taxonomy using 454 reads with a BLASTX search of the SEED database (cut-off 1e^{-5}), 454 BLASTN against RDP, Silva SSU, Greengene (cut-off 1e^{-5} and minimum alignment length 50 bp), Forsyth HOMD 16S rRNA RefSeq Version 10.1 (cut-off 0.0001), and MEGAN analysis megablast against GenBank NT (minscore = 35.0 minscorebylength = 0.0 toppercent = 10.0 winscore = 0.0 minsupport = 5) and Amphora analysis N0 (the immediate ancestor) and N1 (the first internal node) values. The three columns on the right (v5 16s illumina, 16s saliva, 16s plaque) were taken from references (Keijser *et al.*, 2008; Lazarevic *et al.*, 2009) for comparison.

region V5 (Lazarevic *et al.*, 2009), which showed an extremely low representation of the Bacteroides phylum. However, this was also noted by the authors of the study and was suggested to be the result of classification bias for the specific region or methods used in that study (Lazarevic *et al.*, 2009).

It is also worth noting that some differences exist among the different databases used in the homology-based taxonomy assignment of 16S reads. For example, using Silva SSU, the Firmicutes represent 9.62% of the total population, whereas the other databases provided estimates ranging from 16.83 to 21.47% (Table 1). The reason is that Silva SSU has over 600 K sequences and is six times larger than RDP and Greengene. Typically, a larger database will result in increased e-values, which reduces the number of reads that pass the blast cut-off (1e^{-5} and minimum alignment = 50 bp). Silva results also assigned more reads as unclassified bacteria for the same reason. Despite these few differences, the majority of phylum-level classifications are similar regardless of the database used. In addition, predicted gene-based taxonomy assignments, such as BLASTX vs. NR or SEED, and BLASTP vs. IMG database, yielded similar results for all major phyla except TM7, which was not found with SEED. This is because only finished and draft genome sequences were deposited in the SEED subsystem database and protein sequences are not available for TM7, as opposed to the 16S RDP, Silva SSU, and Greengene installed at the SEED database. From these data, we conclude that, despite mostly congruent taxonomic predictions for major phyla within the oral cavity, different approaches (methods and databases) need to be tested to obtain an accurate estimation, particularly for the minor phyla.

It is also noted that the 16S rRNA-based analysis (RDP, Silva, SSU, Greengene, and HOMD) gives similar estimates of microbial composition as marker gene-based assays (AMPHOA; see Fig. S2). The slight differences between these two strategies are likely caused by large variations in the rRNA gene copy numbers among different species. The phylogenetic markers used in this study are all single-copy

Table 1 Distribution of major phyla using different analysis programs in comparison with known datasets

Phyla	Open reading frame-based approach			16s rRNA-based approach				Phylogenetic marker protein-based approach		Data from previous reports		
	454 blastx LCA	454 top blastx SEED	Contig blasp IMG	454 16s RDP	454 16s Silva SSU	454 16s Greengene	454 16s HOMD	IMG AMPHORA_n1	MG-RAST AMPHORA_n1	V5–16s Illumina	V5–V6 16s 454 Saliva	V5–V6 16s 454 Plaque
Firmicutes	17144 (18.05)	14989 (20.02)	23346 (28.53%)	95 (19.47)	98 (9.62)	72 (18.83)	82 (21.47)	189 (16.8%)	131 (13.25)	322683 (30.48)	(40.7)	(30.93)
Proteobacteria	30034 (31.62)	29874 (39.9)	18338 (22.27%)	185 (37.91)	465 (45.63)	139 (35.37)	143 (37.43)	289 (25.69%)	339 (34.28)	317231 (29.97)	21	15.24
Actinobacteria	19240 (20.26)	12966 (17.32)	11694 (14.20%)	41 (8.4)	46 (4.51)	41 (10.43)	43 (11.26)	126 (11.20%)	150 (15.17)	49131 (4.64)	6.3	24.55
Fusobacteria	2408 (2.54)	1908 (2.55)	2049 (2.49%)	26 (5.33)	25 (2.45)	21 (5.34)	26 (6.81)	5 (0.44%)	11 (1.11)	11765 (1.11)	2.9	8.88
Candidate division TM7	371 (0.39)	0 (0.00)	23 (0.03%)	5 (1.02)	2 (0.2)	5 (1.27)	5 (1.31)	0 (0.00)	0 (0.00)	17691 (1.67)	1.9	1.86
Bacteroidetes	19955 (21.01)	14201 (18.97)	26711 (32.44%)	85 (17.42)	236 (23.16)	76 (19.34)	79 (20.68)	243 (21.60%)	174 (17.59)	693 (0.07)	27.2	14.72
BRC1	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1286 (0.12)	0	0
OP10	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.25)	0 (0.00)	0 (0.00)	0 (0.00)	649 (0.06)	0	0
Spirochaetes	308 (0.32)	481 (0.64)	65 (0.08%)	3 (0.61)	3 (0.29)	3 (0.76)	3 (0.79)	0 (0.00)	3 (0.30)	2758 (0.26)	0.2	0.86
Cyanobacteria	7 (0.01)	340 (0.45)	88 (0.11%)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.10)	130 (0.01)	0.02	0
Candidate division SR1	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.10)	0 (0.00)	1 (0.26)	0 (0.00)	0 (0.00)	0 (0.00)	0.014	0
Fibrobacteres/ Acidobacteria group	0 (0.00)	86 (0.11)	16 (0.02%)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0.049	0
OP3	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0	0
Unclassified bacteria	5505 (5.80)	28 (0.04)	8 (0.01%)	48 (9.84)	143 (14.03)	35 (8.91)	0 (0.00)	273 (24.27%)	180 (18.20)	334532 (31.60)	0.2	0
Total reads	94972	74873	82338	488	1019	393	382	1125	989	1058549		

genes. Therefore it should theoretically give a more accurate estimation of the microbial composition. Another factor affecting both 16S rRNA-based and phylogenetic marker gene-based analyses using the 454 reads is that <1% of the reads encode the 16S and marker genes. Considering the low sequencing coverage of 454 reads, these methods only represent a snapshot of the community and might underestimate the less abundant organisms through undersampling. To overcome this limitation, multiple complimentary methods should be employed when assessing community composition.

Overall, we were able to detect 668 bacterial phylotypes in the metagenome sequence of this dental plaque microbiome (see http://mg-rast.mcs.anl.gov/mg-rast/FIG/linkin.cgi?metagenome=4446622.3 and Table S3 for detailed assignments). Of these, 382 16S rDNA reads had significant similarity to the HOMD 16S reference sequences (Table S3) and the remaining 58 reads could not be assigned to any species/phylotypes, suggesting novel species/phylotypes. This level of diversity is substantially higher than previous estimates of ~ 100–200 species/phylotypes per person (Aas *et al.*, 2005; Paster *et al.*, 2006; Nasidze *et al.*, 2009), but is within the range reported by a recent 16S pyrosequencing study (Zaura *et al.*, 2009). Taken together, these results suggest that a combination of 454 and Illumina random shotgun sequencing is sufficient to achieve comparable community diversity coverage, and possibly with less bias than targeted 16S-based community profiling surveys.

It is important to note that, as with other community sequencing efforts, the oral metagenome determined here is a collection of genomic fragments and not all members of the community are equally represented, nor do they necessarily have large portions of their genome represented, particularly if they are rare community members. In fact, despite the exceptional depth of our sequence coverage, some species are still probably represented by only a handful of reads. Therefore, although this study has generated many thousands of contigs that range in size from hundreds of bp to >29 kb with ‘sufficient’ sequence average coverage (from 3.5-fold to 27.5-fold) for adequate functional and phylogenetic interpretation, using contig data alone is difficult for obtaining an accurate genomic abundance profile for the entire community.

Mapping the metagenome reads to reference genomes

Because a number of oral reference genomes are currently available, we mapped all of our sequencing reads against the HMP oral reference genomes to assess the coverage and abundance of these sequenced references or close neighbors within the plaque community. The number of 454 and Illumina reads recruited by the 50 oral reference genomes was ~ 500,000, or roughly 4% of the total (12 million) non-human reads (Fig. 2A,B). The top 10 species that matched to the reference genomes are from five major phyla (Bacteroidetes, Actinobacteria, Firmicutes, Proteobacteria, Fusobacteria), which is consistent with our community profile data shown in Fig. 1, Tables 1 and S3. The top two *Streptococcus* recruits were from *Streptococcus mitis* NCTC 12261 (1.7 Mb recruited size and 40% genome coverage) and *Streptococcus sanguinis* SK36 (1.7 MB recruited size and 32% genome coverage) with 22 K reads of 97% identity at the DNA level. This result is consistent with these two species being prominent members of the pioneer plaque community. Interestingly, the number of reads (12 K) recruited by *Streptococcus gordonii* Challis substr. CH1 (0.8 Mb recruited size and 20% coverage), another prominent member of the pioneer dental plaque community, is lower than the recruitment (17 K) by *Streptococcus pneumoniae* TIGR4 (1.0 Mb recruited size and 19% coverage), a typical member of the human nasopharyngeal and oral flora. Surprisingly, the top five recruits were *Capnocytophaga gingivalis* JCVIHMP016 (87 K reads, 6.8 Mb, and 68% genome coverage), *Corynebacterium matruchotii* ATCC 33806 (49 K reads, 4 Mb, and 54% genome coverage) and ATCC 14266 (~ 49 K reads, 4 Mb, and 54% coverage), *Capnocytophaga sputigena* (~ 37 K reads, 2.7 Mb, and 44%), and *Capnocytophaga ochracae* (~ 25 K reads, 2.2 Mb, and 41%). This is contradictory to the common belief that the streptococci are the predominant species in the plaque community. A likely explanation for this discrepancy is that this individual uniquely harbors more of these species. Future studies of sequencing more samples from a large number of individuals will help to resolve this issue. Another possible reason for this discrepancy is that the streptococcal strains in this plaque are divergent from the streptococcal strains represented in the reference genome database.

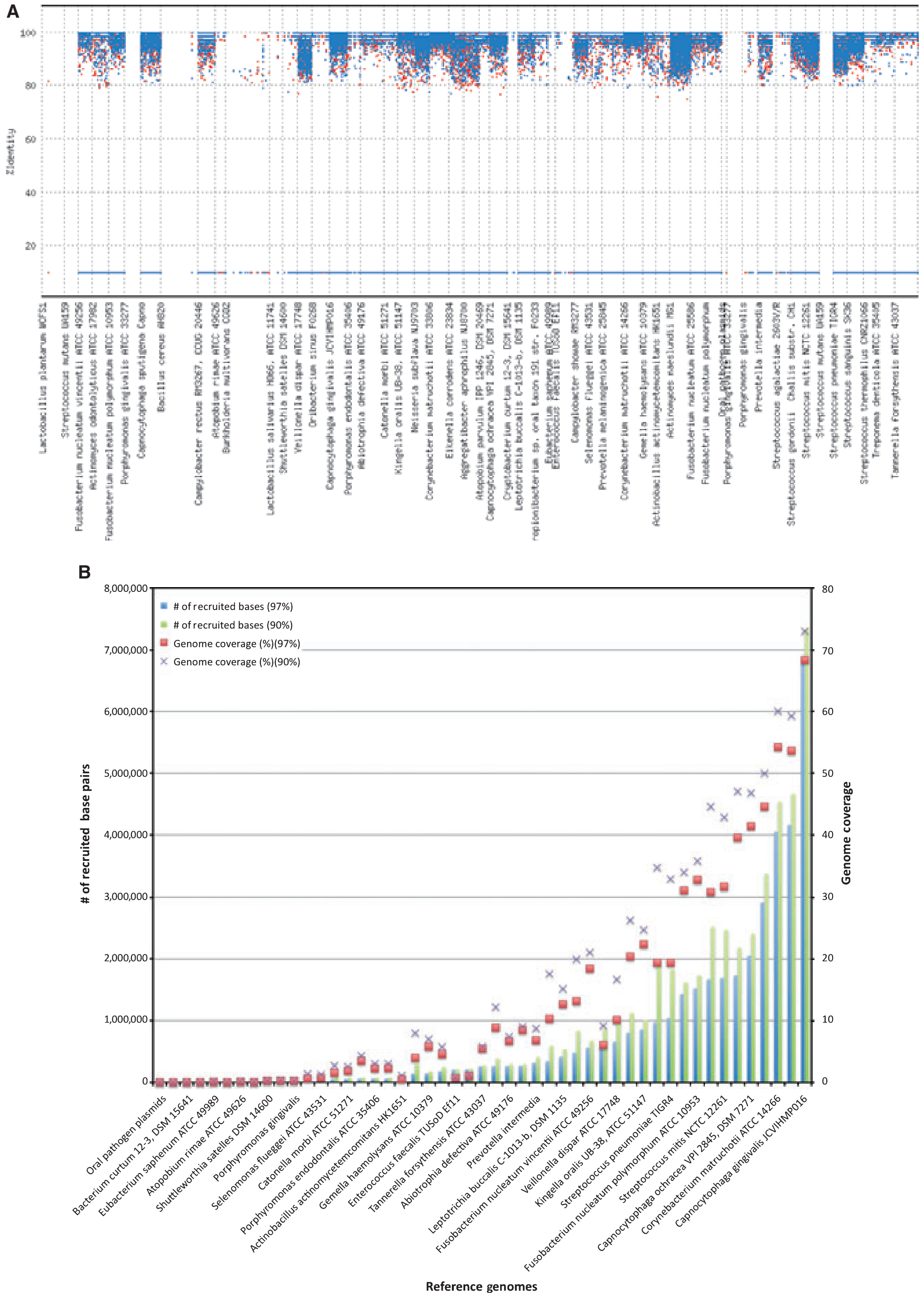

Figure 2 (A) Metagenome fragment recruitment using 454 plus Illumina reads against the HMP and Oralgen reference genomes. Fragment recruitment was performed with MUMMER. For all reference genomes, the contigs are concatenated and arranged by length along the *x*-axis, as indicated by the tick marks. Percentage identity is shown along the *y*-axis. The red and blue dots denote the forward and reverse direction matches, respectively. (B) 454 and Illumina reads recruitment by 50 HMP reference genomes. Percentage of reference genome coverage is shown along the *y*-axis on the right. Number of sequence reads recruited by reference genome is shown along the *y*-axis on the left.

Table 2 Comparison of MG-RAST and IMG-M ER annotation of 454 reads and contigs obtained from the 454 plus Illumina hybrid assembly

Annotation submissions	MG-RAST 454 reads	MG-RAST 454 + illumina contigs	IMG M 454 + illumina contigs
Total no. of sequences	109,708	128,556	113652
Total sequence size (bp)	43,613,321	29,276,210	27,411,856
Shortest sequence length (bp)	51	69	69
Longest sequence length (bp)	746	39586	39586
Average sequence length (bp)	397.54	227.73	227.73
Phylogenetic profile	Blastx against SEED ($1e^{-5}$)	Blastx against SEED ($1e^{-5}$)	Blastp against IMG (30% identities)
Classified	70.44% (77278)	54.23% (69715)	73.11% (83099)
Non-classified	29.56% (32430)	45.77% (58841)	26.88% (30553)
Total	100% (109708)	100% (128556)	100% (113652)
Domain level			
Archaea	0.38% (295)	0.24% (169)	0.08% (67)
Bacteria	88.83% (68644)	55.63% (38783)	99.47% (82662)
Eukaryota	0.89% (685)	0.36% (252)	0.12% (106)
Virus	0.00%	0.00%	0.22% (189)
Other	9.9% (7654)	43.77% (30511)	0%
Total	100% (77278)	100% (69715)	100% (83099)
Function annotation	Blastx aginst SEED subsystem	Blastx aginst SEED subsystem	IMG gene prediction
Coding	50.58% (55488)	37.80% (48600)	72.72% (19934878 bp)
Non-coding	49.42% (54220)	62.20% (79956)	27.28% (7476978 bp)
Total	100% (109708)	100% (128556)	100% (27411856 bp)
Protein coding with function prediction	96.77% (53640)	96.79 (47042)	50.60% (43956)
Protein coding without function prediction	3.33% (1848)	3.21% (1558)	48.36% (42006)

Indeed, most of the sequence reads (96%) in this study cannot be mapped with high confidence to any of the 50 reference genomes based on high similarities at the DNA level. This speculation is consistent with our other observations. For example, Fig. 2A shows that 60% or 30 reference genomes have <10% genome coverage from the recruited fragments. Among the 40% or 20 remaining references, the recruitments are not evenly distributed, with about seven genomes having more recruitments at 90–97% identity levels. When 90% identity cut-off was tested, four streptococci showed a 26–75% increase in recruitment size (bp) and 15–41% increase in genome coverage (see Table S4). Two other species, *Neisseria subflava* NJ9703 and *Actinomyces naeslundii* MG1, also showed a 51 and a 106% increase in recruitment size, respectively. This observation suggests that the current HMP reference genomes could serve as a starting point for reconstruction of microbial genomes from metagenomic sequences; however, more strains of each species need to be sequenced to cover the intra-species diversity. It should also be noted that the percentage of species matched to the reference genomes by no means reflects the percentage of these species in the metagenome, because only 4% of the total reads could be matched to the reference genomes.

Functions encoded by the plaque microbiome

A preliminary assessment of the functional capacity of the plaque microbiome was determined by subjecting 454 reads, as well as the contigs obtained from the 454 plus Illumina hybrid assembly, to automated annotation using publicly available pipelines (MG-RAST and IMG-M ER). The high-level results are summarized in Table 2. As a result of the size of the dataset, the annotation of Illumina reads is computationally prohibitive, except mapping them to HMP oral reference genomes at the DNA level using MUMMER. Among all 454 reads submitted to MG-RAST, ~ 50% could be assigned to metabolic subsystems based on top BLASTX hits to SEED and were sorted into functional categories (see http://mg-rast.mcs.anl.gov/mg-rast/FIG/linkin.cgi?metagenome=4446622.3 for detailed information). In contrast, among all 454-Illumina contigs submitted to IMG-M ER, ~ 73% of the total sequences were predicted to code for proteins, among them 50.6% (or 43,956 genes) had predicted functions. The slightly more and better gene prediction for the hybrid 454-Illumina contigs (73% of coding region vs. 50%) is most likely the result of the longer contig sequence. Despite the fact that IMG-M ER and MG-RAST used different annotation approaches, the overall COG category or Subsystem function assignment is about the same. The predominant functional categories included carbohydrate metabolism (11.88% of the assigned reads), amino acids and derivatives (7.89%), proteins (9.34%), cofactors, vitamins, prosthetic groups, and pigments (6.26%); cell wall and capsules (5.24%), RNA metabolism (4.53%), DNA metabolism (6.07%), nucleoside and nucleotide metabolism (3.55%), membrane transport (3.16%), cell division (2.1%), respiration (3.53%), regulation and cell signaling (1.31%), fatty acid and lipid metabolism (1.27%), motility and chemotaxis (1.11%), phosphorus metabolism (1.07%), and sulfur metabolism (1%). The relative abundances of the different COG categories and pathways based on IMG-M ER annotation by extracting all COG identifiers from the BLAST output is summarized in Table S5.

Interestingly, the fourth largest percentage of reads was assigned to the functional category of virulence (6.46%), with an additional 2.35% of the reads assigned to functions involved in stress responses. Furthermore, 42% of the reads belonging to the virulence gene category (or 2.79% of total reads) encode proteins with putative functions related to antibiotic and toxin resistance, and a further 25.59% (or 1.69% of total reads) were related to iron scavenging. In light of the recent finding that the human microbiota may be a reservoir for antibiotic resistance genes (Sommer *et al.*, 2009), the abundance of this functional category is of great interest. These include functions involved in resistance to the major classes of antibiotics, such as β-lactams, aminoglycosides, fluoroquinolones, and the peptide antibiotic bacitracin, as well as general multidrug or heavy-metal resistance functions such as efflux pumps. These findings could have significant implications for the spread of drug resistance to human pathogens, because the oral cavity is a portal of entry for numerous pathogens that cause systemic infections. Further investigations are needed to determine the relationship between antibiotic resistance genes in the oral microbiome and those found in antibiotic-resistant pathogens.

Overall, 660 functional gene groups were detected, with multiple sequences in each group. These data will serve as an important resource for the dental research community, both in terms of its use as a benchmark for further detailed analysis of the plaque microbiome, including refined analysis of this current dataset, or used for other oral community investigations such as gene expression (metatranscriptome) studies in health or disease.

Comparison of the dental plaque microbiome with other microbiomes

A number of metagenomic datasets from other ecosystems are now available in the public domain, so we compared the plaque microbiome with other microbiomes from scientific curiosity. Four microbiomes were used for the comparison: (i) the human gut (Gill *et al.*, 2006); (ii) the obese mouse gut (Turnbaugh *et al.*, 2006); (iii) the soil (Tringe & Rubin, 2005); and (iv) the ocean (Rusch *et al.*, 2007). Three aspects of these datasets were compared: (i) taxonomic distribution, (ii) physiological properties, and (iii) array of biochemical functions predicted.

At the taxonomic level, dramatic differences were observed in a number of taxa among the different ecosystems. For example, the marine environment harbors the largest proportion of Proteobacteria, followed by the dental and the soil microbiomes, whereas the human gut harbors the lowest level (Fig. 3A). In contrast, the human gut harbors the highest proportion of Firmicutes, followed by the dental plaque and the obese mouse gut, whereas few Firmicutes are found in the soil and the marine environments. Some taxa such as Fusobacteria and the TM7 Division are most often seen in the dental plaque microbiota, whereas the Fibrobacteres and the Acidobacteria groups are predominant in the soil microbiome. Interestingly, the Actinobacteria is found to be highly abundant in the human dental plaque and the human gut, but not in the obese mouse gut.

At the physiological level, dramatic differences were also observed among the different microbiomes. For example, the human gut harbors more predicted anaerobic and host-associated microorganisms than the dental plaque, whereas the latter harbors more organisms with ambiguity in a number of defined parameters, such as oxygen or temperature requirements, gram-positive or gram-negative, and

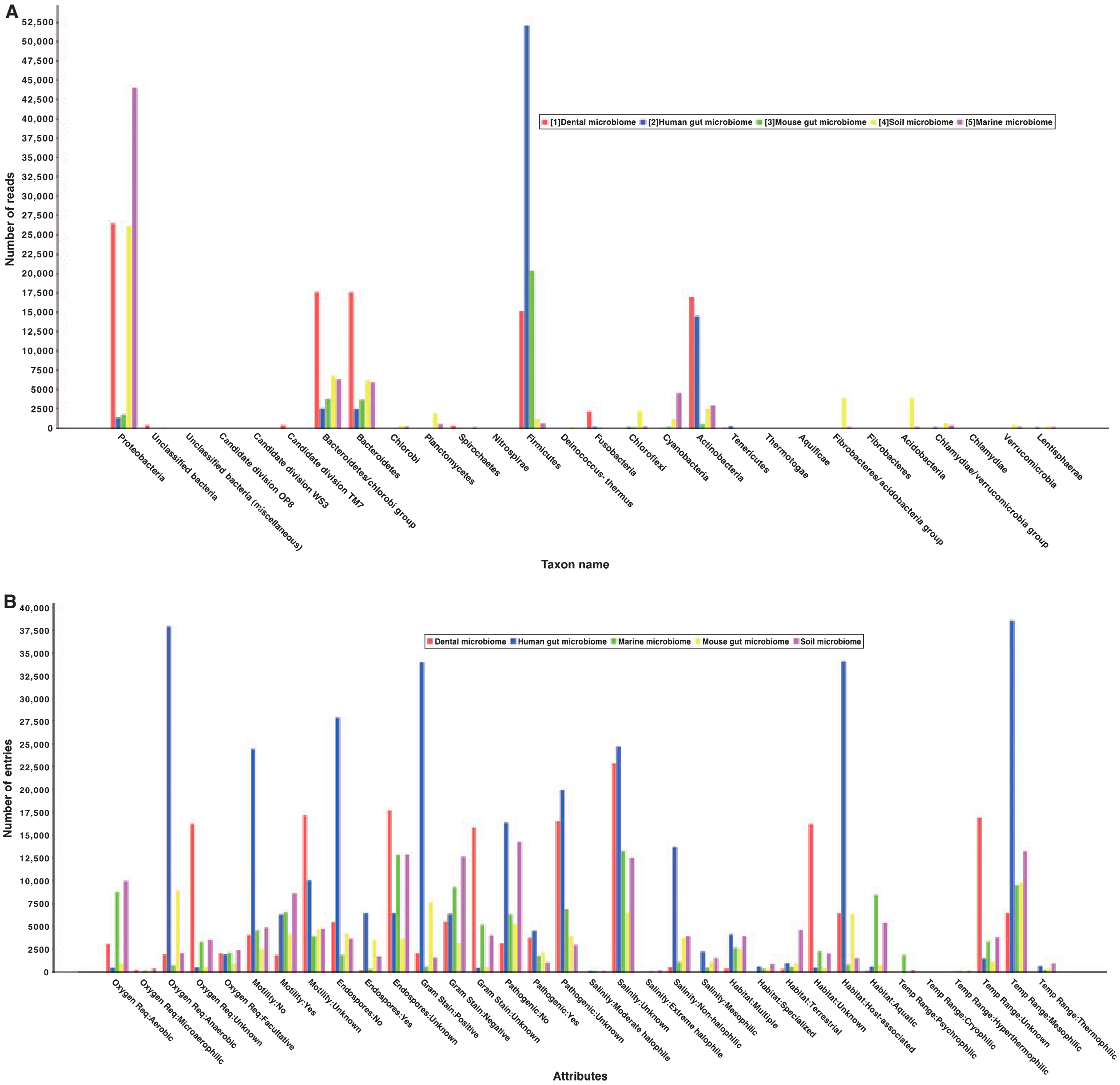

Figure 3 (A) Summary of the comparison of the dental plaque (red), human gut (blue), mouse gut (green), soil (yellow), and marine (magenta) datasets, generated at phylum level ranks. (B) Summary of the comparison of the microbial attributes of dental plaque (red), human gut (blue), mouse gut (yellow), soil (magenta) and marine (green) datasets based on the NCBI's 'Prokaryotic Attributes Table'. In the bar chart, the number of classified species having the indicated property is displayed.

specialized or mixed habitats (e.g. aquatic, terrestrial, or host-associated) (Fig. 3B). These differences may reflect the variable environmental conditions within the oral cavity (i.e. large variability in oxygen levels and temperature as a result of the opening and closing of the mouth during the day and night), compared with the constant body temperature and general anaerobic environment within the gut. In addition, the gut microflora and the gut epithelial cells constantly interact with each other, whereas on the tooth's surface such interactions are rare or non-existent, except within the deep periodontal pockets.

Interestingly, when the functional genes were compared at Subsystem hierarchy level 1 (group level of subsystems such as amino acid and derivatives), the two human microbiomes and mouse microbiome had almost identical distribution and abundance of functional groups (see Fig. S3). At Subsystem hierarchy level 2 (subgroup level of subsystems such as alanine, serine, and glycine), the two human microbiomes start to show some minor differences. At subsystem level (for example, alanine biosynthesis), more significant differences are observed (see Fig. S4). There are 43 subsystems that appear to be unique in dental plaque, which are encoded by 370 reads (see Table S6). Similarly, COG category and pathway, Pfam and TIGRfam were compared using the IMG-M Abundance Profiles Tool. Despite a few differences in COG and Pfam functional categories between dental plaque and human gut samples (see Fig. S3), the two microbiomes are very similar to each other in the TIGRfam category (see Table S7). At the lower level, about 10% COG (263) and Pfam (166) showed significantly different abundance profiles between the human oral and gut samples. This finding provides further support for the notion that functional redundancy exists in high levels of metabolism-based subsystems, but at the lower level, different organisms may harbor genes for its niche-specific function. Hence, only by analysing both species richness and gene expression can we identify the microbial attributes for oral health or disease.

CONCLUSION

During this pilot short-gun metagenomic sequencing and data analysis of the human dental plaque microbiome, we have learned the following. Pooled plaque samples from each individual are required to obtain sufficient DNA for sequencing, especially from patients with periodontal health. Special care should be taken to avoid human cell contamination during sampling. A hybrid assembly of 454 pyrosequencing and Illumina reads may be a more cost-effective way to sequence and assemble the metagenome of human microbiomes. Using the 31 phylogenetic marker genes for community profiling may yield more accurate estimates than 16S rRNA-based assays because of the presence of a single copy for each marker per microbial genome. Each individual may harbor a unique microbiome, and only by analysing a large number of microbiomes can we obtain a general picture of the microbiomes of health and disease. The flexibility in nutrient and oxygen requirements may be important in allowing the oral microbes to reside in the oral cavity. It is our hope that this information will prove useful for other investigators in the oral microbiome research community. It should also be noted that the primary purpose of this pilot study was to resolve a number of technical issues in metagenomic sequencing and data analysis, such as how much plaque sample is needed to obtain a sufficient amount of DNA for sequencing, how to avoid host cell contamination in samples, what is the most cost-effective way to achieve sequence coverage and depth, what is the best way to assemble the sequence reads, and what software or database should be used for community profiling or functional assignment etc. The results obtained are interesting but are secondary to the technical issues resolved during this pilot study. We expect that in the future, upon availability of funding, more biology-oriented studies will be conducted, which will provide a true estimate of the functional repertoire of the human oral microbiome.

ACKNOWLEDGEMENTS

We thank Floyd Dewhirst for providing us the HOMD taxonomy table; Stephanie Eichorst for her assistance in curating 16s rRNA tree; Chris Detter, Lance Green, and Yvonne Rogers for their assistance in metagenome sequencing; Pavel Senin for his assistance in trimming sequences and working on sequence recruitment algorithms; the JGI IMG-M ER and MG-RAST teams for providing annotation service; and Natalia Ivanova and Mark D'Souza for addressing many questions regarding their annotation. This study was supported in part by grants from the National Institutes of Health (NIH Y1-DE-6006-02) and from the Los Alamos National Laboratory (LANL; 20080662DR) for LANL participants. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

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

Aas, J.A., Griffen, A.L., Dardis, S.R. *et al.* (2008) Bacteria of dental caries in primary and permanent teeth in children and young adults. *J Clin Microbiol* **46**: 1407–1417.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Aziz, R.K., Bartels, D., Best, A.A. *et al.* (2008) The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**: 75.

Becker, M.R., Paster, B.J., Leys, E.J. *et al.* (2002) Molecular analysis of bacterial species associated with childhood caries. *J Clin Microbiol* **40**: 1001–1009.

Carter, K., Oka, A., Tamiya, G. and Bellgard, M.I. (2001) Bioinformatics issues for automating the annotation of genomic sequences. *Genome Inform* **12**: 204–211.

Chou, H.H. and Holmes, M.H. (2001) DNA sequence quality trimming and vector removal. *Bioinformatics* **17**: 1093–1104.

Gill, S.R., Pop, M., Deboy, R.T. *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312**: 1355–1359.

Haffajee, A.D. and Socransky, S.S. (2005) Microbiology of periodontal diseases: introduction. *Periodontol 2000* **38**: 9–12.

Haffajee, A.D. and Socransky, S.S. (2006) Introduction to microbial aspects of periodontal biofilm communities, development and treatment. *Periodontol 2000* **42**: 7–12.

Huson, D.H., Auch, A.F., Qi, J. and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res* **17**: 377–386.

Keijser, B.J., Zaura, E., Huse, S.M. *et al.* (2008) Pyrosequencing analysis of the oral microflora of healthy adults. *J Dent Res* **87**: 1016–1020.

Kroes, I., Lepp, P.W. and Relman, D.A. (1999) Bacterial diversity within the human subgingival crevice. *Proc Natl Acad Sci U S A* **96**: 14547–14552.

Kumar, P.S., Griffen, A.L., Barton, J.A., Paster, B.J., Moeschberger, M.L. and Leys, E.J. (2003) New bacterial species associated with chronic periodontitis. *J Dent Res* **82**: 338–344.

Kurtz, S., Phillippy, A., Delcher, A.L. *et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.

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

Markowitz, V.M., Ivanova, N.N. and Szeto, E. and other authors (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* **36**: D534–D538.

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

Marsh, P.D. (2006) Dental diseases – are these examples of ecological catastrophes? *Int J Dent Hyg* **4**(Suppl 1): 3–10; discussion 50–52.

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

Noguchi, H., Park, J. and Takagi, T. (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* **34**: 5623–5630.

Overbeek, R., Begley, T., Butler, R.M. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33**: 5691–5702.

Paster, B.J., Boches, S.K., Galvin, J.L. *et al.* (2001) Bacterial diversity in human subgingival plaque. *J Bacteriol* **183**: 3770–3783.

Paster, B.J., Olsen, I., Aas, J.A. and Dewhirst, F.E. (2006) The breadth of bacterial diversity in the human periodontal pocket and other oral sites. *Periodontol 2000* **42**: 80–87.

Preza, D., Olsen, I., Aas, J.A., Willumsen, T., Grinde, B. and Paster, B.J. (2008) Bacterial profiles of root caries in elderly patients. *J Clin Microbiol* **46**: 2015–2021.

Rusch, D.B., Halpern, A.L., Sutton, G. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.

Socransky, S.S. and Haffajee, A.D. (2005) Periodontal microbial ecology. *Periodontol 2000* **38**: 135–187.

Socransky, S.S., Haffajee, A.D., Cugini, M.A., Smith, C. and Kent, R.L. Jr (1998) Microbial complexes in subgingival plaque. *J Clin Periodontol* **25**: 134–144.

Sommer, M.O., Dantas, G. and Church, G.M. (2009) Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science (New York, NY)* **325**: 1128–1131.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, UK)* **22**: 2688–2690.

Tringe, S.G. and Rubin, E.M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev* **6**: 805–814.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R. and Gordon, J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**: 1027–1031.

Wu, M. and Eisen, J.A. (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**: R151.

Yooseph, S., Sutton, G., Rusch, D.B. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**: e16.

Zaura, E., Keijser, B.J., Huse, S.M. and Crielaard, W. (2009) Defining the healthy 'core microbiome' of oral microbial communities. *BMC Microbiol* **9**: 259.

Zerbino, D.R. and Birney, E. (2008) Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.

Zhu, W., Lomsadze, A. and Borodovsky, M. (2010) *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res* **38**: e132. Doi: 10.1093/nar/gkq275.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Histogram of GC% distribution of 454 reads showing the frequency of shotgun metagenome reads that group with four major phylotypes detected in the MG-RAST annotation.

Figure S2. Major phylotypes identified in dental plaque data.

Figure S3. Comparing classification of all COGs determined in the dental plaque sample to other ecosystem.

Figure S4. Subsystem level comparison between dental plaque metagenome and human gut TS1 (MG-RAST id 4440452.7) and mouse gut (obese) metagenomes (id 4440464.3).

Table S1. Summary of sequence quality control.

Table S2. Comparison of metagenome sequence assembly using different parameters.

Table S3. Percentage of major genera in predominant phyla among total reads.

Table S4. The summary of Oral 454 and Illumina read recruitment by 50 HMP reference genomes at 97 and 90% identities cut-off.

Table S5. Summary statistics of COG functional categories and pathways associated with the metagenome genes across all phyla/classes that have best BLASTp hits with 30, 60–90, and 90% identity.

Table S6. Comparison of dental plaque with two human gut samples, in terms of their relative abundance of protein families (COGs, Pfams, TIGRfams) and functional categories (COG Pathway, COG category, Pfam Category, TIGRfam sub-roles), with estimates of the *statistical significance* of the observed differences.

Table S7. Subsystem comparison between dental plaque metagenome and human gut TS1 (MG-RAST id 4440452.7) and mouse gut (obese) metagenomes (id 4440464.3).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.