学位論文

Doctoral Dissertation

A study on human gut microbiota using improved methods of 16S ribosomal RNA gene analysis

(改良した 16S リボソーム RNA 遺伝子解析法を用いたヒト腸内細菌叢に関する研究)

金　錫元

KIM SEOK-WON

Acknowledgements

In writing my dissertation, I have contracted many debts. First of all, my most sincere thanks are due to my supervisor, Prof. Masahira Hattori, whose perceptive criticism, kind encouragement, and willing support helped bring my research to a conclusion. I also appreciate my dissertation committee members, Prof. Takeshi Itoh, Assoc. Prof. Kumiko Ui-Tei, Assoc. Prof. Hiroyasu Satoh, and Asst. Prof. Wataru Iwasaki reviewing my dissertation and giving unstinted suggestions for improvement. I owe a debt of thanks to Prof. Hidetoshi Morita at Azabu Univ. for offering high quality samples and encouraging me. I would like to thank Proj. Asst. Prof. Kenshiro Oshima, Dr. Wataru Suda, Dr. Sangwan Kim, and all Hattori laboratory family too. Without their wholehearted support from excellent experimental support to life in Japan, it would have been difficult to finish my dissertation. I would like to thank to Dr. Hideki Hirakawa at Kazusa DNA Res. Inst. and Dr. Hidehiro Toh at Kyushu University for always encouraging me. I would like to express many thanks to my former laboratory, GRC, KRIBB in Korea. Especially, I would like to thank Dr. Hong-Seog Park, Dr. Sang-Haeng Choi, Dr. Haeyoung Jeong, Dr. Sung-Hwa Chae, Dr. Dae-soo Kim, Dr. Daewon Kim (current at Korea NIH), and Prof. Yong-Seok Lee (current at Soonchunhyang University). I would like to thank alumni of the University of Tokyo, Prof. Jung-Kook Hong at KSEAJ, Dr. Daeyun Shin at KMA, Dr. JiEun Jung at Inje Univ., Dr. Seung-Jin Cho at Sharp, Dr. Naesun Park at KIOST, Dr. SoYeon LIM at Samsung Elec. Mech., Dr. Changsuk Song, and Dr. Soonil Kwon. I would like to thank to Prof. Barbara Methé at JCVI and Asst. Prof. Jose C. Clemente at Icahn School of Med. at Mount Sinai for giving critical comments and kind encouragement about GRD. I also deeply appreciate Prof. Jongung Chun at Inje Univ. sharing many experiences in research and development. Finally, I would like to thank to my family and friends for endless love and encouraging me throughout whole process.

Table of Contents

1 Introduction

1.1 Human gut microbiota and health

It is known that there are about 10^{14} indigenous bacteria in various sites of human body, of which the cell number roughly accounts for 10 times more than that of the total human cells. Among human body sites, the gastrointestinal tract is the major habitat of gut microbes, which form distinctive ecosystem by integrating with various components of host cells and exogenous ones from foods [1-3].

International Human Microbiome Consortium (IHMC) was launched in 2008 to comprehensively elucidate human microbiome, the collective genomes of human microbes [3]. One of the major goals of IHMC is to construct a reference database of genomes of human microbes. Up to now, 1,147 strains isolated from several human habitats have been sequenced fully (as finished sequences) or incompletely (as draft sequences), and deposited in Human Microbiome Project (HMP) database (http://www.hmpdacc.org). Present status of sequencing of human microbes was summarized in Table 1.1.

In gut microbiota of healthy adults, it was revealed that the major taxonomic groups at the phylum level include *Bacteroidetes*, *Firmicutes*, *Proteobacteria* and *Actinobacteria*, and the minor taxonomic groups include *Fusobacteria*, *Tenericutes*, and *Verrucomicrobia* [2]. Moreover, the major taxonomic groups at genus level are *Bacteroides*, *Ruminococcus*, *Clostridium*, and *Faecalibacterium* [2]. It is also known that the species composition is highly varied, but gene profiles or contents are highly conserved among individuals [2, 4, 5]. The gut microbiota contributes to a variety of biological functions such as “energy production and storage, carbohydrate metabolism, secondary metabolites biosynthesis, amino acid metabolism, and transport and catabolism” [5], therefore, it can be called as “forgotten organ” [6].

In this concern, imbalance of gut microbiota is also known to be closely related with various diseases such as obese and inflammatory bowel disease (IBD), of which the bacterial composition significantly differed from that of healthy subjects [7-12].

Table 1.1 Present status of sequenced strains of human microbes isolated from various sites

Body habitat	No. of isolated strains	No. of sequenced strains
Airways	148	50
Blood	55	43
Gastrointestinal tract	706	375
Oral	435	205
Skin	295	116
Urogenital tract	315	133
others	13	4
unknown	514	221
Total	2,481	1,147

1.2 Advance of DNA sequencing technologies

Modern DNA sequencing technology was started with two different methods in 1977. One is the Sanger method [13, 14] and the other is the Maxam-Gilbert method [15]. Especially, the Sanger method later became most popular in molecular biology fields and was utilized in the Human Genome Project conducted from 1991 to 2004 [16-22].

Since 2005, various types of sequencing methods and sequencers, so-called "Next-Generation Sequencing (NGS)", have been developed [23-28]. Sequencers based on NGS equip with the technologies different from that of the Sanger method. For example, pyrosequencing is based on detection of pyrophosphate molecule released from DNA synthesis by polymerase, and Roche GS FLX+ system based on pyrosequencing is characterized by simultaneous determination of nucleotide sequences for millions of templates using picoliter reactors, which made it possible to perform high-throughput sequencing with thousands to million-fold of the traditional capillary-type DNA sequencers based on the Sanger method [29-31].

1.3 16S ribosomal RNA gene

Ribosome is a large complex particle composed of various ribonucleic acid molecules and proteins [32, 33]. The synthesis of proteins in cell is carried out by the ribosome composed of ribosomal RNAs (rRNAs) and ribosomal proteins [34]. In prokaryote, there are three different 5S, 16S, and 23S rRNAs, which have different sedimentation coefficients, respectively [35, 36]. The genes encoding the rRNA are commonly existed in all prokaryotes. Among them, 16S rRNA gene can be used in microbial community analysis including species assignment and its abundance [37-40]. *Escherichia coli* known as a model species for bacteria has seven copies of 16S rRNA genes of which the average length is about 1.5 kb in its genome. The 16S rRNA molecule forms its secondary structure (Figure 1.1) in combination with several ribosomal proteins to exhibit its biologically function [41]. Due to this structural feature of 16S rRNA molecule, it possesses highly and nine less conserved regions in the molecule (Figure 1.2) [42, 43]. The nine less conserved regions have less effect on the formation of 30S subunit assembly, resulting in highly mutated sequences (hypervariable regions) specific to bacterial species [44, 45]. Therefore, these hypervariable regions can be used as a phylogenetic marker based on their sequence diversity [46-48]. Moreover, the development of the polymerase chain reaction (PCR) method [49, 50] made 16S rRNA gene method more popular to use for the analysis of microbial communities in various hosts and environments including human, soil, and sea [2, 39, 51].

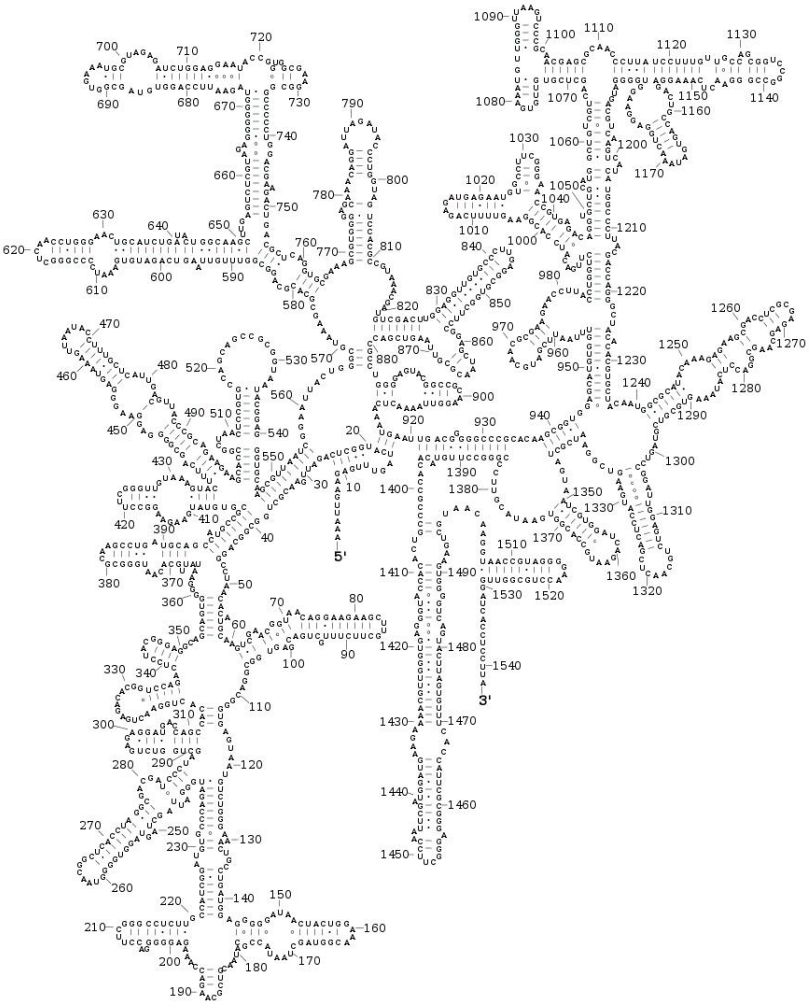

Figure 1.1 Secondary structure of *Escherichia coli* 16S rRNA.
The length of *Escherichia coli* 16S rRNA is about 1.5 kb. 16S forms its secondary structure to function. (Figure taken from The Center for Molecular Biology of RNA website.)

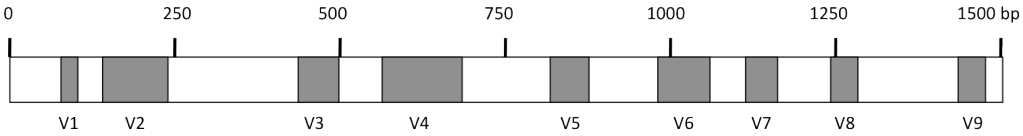

Figure 1.2 Structure and hypervariable regions in 16S rRNA gene.
The 16S gene has highly conserved regions [42] and highly variable (hypervariable) regions [43, 46] and these hypervariable regions can be used as a phylogenetic marker.

1.4 Operational Taxonomic Units

OTU (Operational Taxonomic Unit) is exclusively employed to taxonomically analyze bacterial communities [52-55]. OTU analysis includes computational clustering of 16S sequences obtained from bacterial communities to classify them into the units composed of

sequences having high nucleotide identity, which can be taxonomically assigned to the corresponding species. In early OTU analysis, nearly full-length 16S rRNA gene cloned in plasmid sequenced by the Sanger method was used [56-58]. However, many of recent NGS-based sequencers produce short (35 bp~/read) but massive sequencing data as compared with the Sanger method-based sequencer (~800 bp/read). This NGS performance has changed the OTU analysis of 16S rRNA sequences in which short sequences covering only a few hypervariable regions of 16S rRNA gene are used [59-62]. Current NGS-based OTU analysis consists of mainly three steps, quality filtration of sequenced reads, clustering of high-quality 16S reads (OTUs formation), and various analysis of OTU data using α- or β-diversity metrics including UniFrac distance [63] to examine taxonomical and structural features of bacterial communities (Figure 1.3).

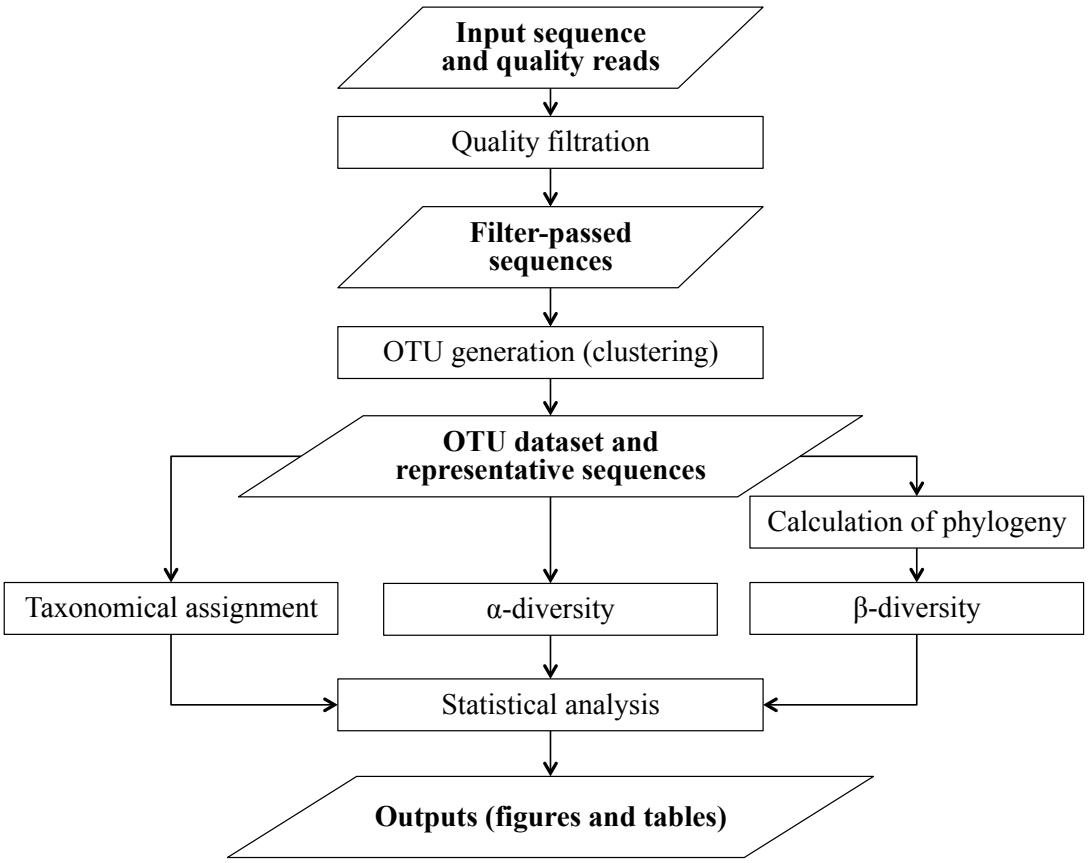

Figure 1.3 Flowchart of 16S-based OTU analysis with quality filtration.
Sequenced reads including quality score are filtered using several filtering methods, and clustered to OTUs. Generated OTU dataset and representative sequences of each OTU are used for various analyses of microbial community.

2 Organization of the dissertation

As described above, NGS-based 16S rRNA gene analysis has become more popular because it provided the culture-independent high-throughput approach to comprehensively evaluate the overall structure and change in relative abundance of species at the OTU level.

This dissertation contains three major chapters. In chapter 3, I described development and construction of genomic-based 16S ribosomal database (GRD). GRD is the highly curated 16S rRNA gene database, and was constructed from full-length 16S rRNA genes in sequenced bacterial genomes using various informatics tools and by my own curated editing. The 16S rRNA sequence data can be more properly assigned to bacterial species by using GRD than other known 16S rRNA gene databases. In addition, I estimated 16S rRNA sequence similarity that determined the boundary between each taxonomic level by using GRD. In chapter 4, I described development of an analytical pipeline for 454 pyrosequencing data of 16S rRNA gene V1-V2 region, by which the quantitative accuracy in 16S-based bacterial composition analysis was greatly improved. First, I estimated the error rate of 454 pyrosequencing data to be ~0.6% from 16S rRNA sequence data obtained from two artificial bacterial communities, Mock 1 and Mock 2, prepared by mixing genomic DNAs of known bacteria with appropriate ratio. Second, I found that the error in the 454 pyrosequencing data was the primary cause of overestimation of species richness based on the number of OTUs generated from clustering of 16S rRNA sequence reads. This overestimation was improved by using clustering with a 96% identity cutoff instead of the conventional 97% identity cutoff. Finally, I developed the modified primer (27Fmod) used for PCR amplification of V1-V2 region in 16S rRNA genes, by which I improved underestimation of the relative abundance of the genus *Bifidobacterium*, one of the major species in human microbiota, observed in the data using the conventional primer 27F. In chapter 5, I analyzed the overall structure of gut microbiota of healthy adults in response to probiotic intervention by using GRD, UniFrac

distance metrics, and the improved analytical pipeline for 454 pyrosequencing data of 16S rRNA V1-V2 region.

3 Development of genomic-based 16S ribosomal RNA gene database

3.1 Background

The 16S rRNA gene data is important and prerequisite for the analysis of microbial communities. It has nine hypervariable regions, which are useful for species assignment and evaluation of the bacterial composition in microbial communities. The copy number of 16S rRNA genes in one bacterium is varied from one to 15. It is known that anti-Shine-Dalgarno sequence (anti-SD) is usually situated at the 3' end of 16S rRNA gene, and is complementary sequence of the Shine-Dalgarno sequence in mRNA [64]. Several variants of 16S rRNA genes have been also found so far. For example, some complete genomes did not have anti-SD in the GenBank annotation [65, 66], and certain bacteria possess intervention sequences within their 16S rRNA genes, some of which were found to be introns [67].

There are publically available 16S rRNA gene databases, Ribosomal Database Project (RDP) [68], SILVA [69], and Greengenes [70]. These databases were constructed on the basis of 16S rRNA genes annotated in genome sequences and those obtained by targeted sequencing of PCR-amplicons from isolated bacteria and bacterial communities in GenBank/DDBJ/EMBL. For construction of these databases, various 16S rRNA gene annotation strategies have been employed, and in some cases inconsistency of the exact annotation of 16S rRNA gene can be seen. In addition, contaminations of unrelated sequences to 16S rRNA gene such as 23S rRNA gene are also observed, and are problematic.

To solve and overcome problems in 16S rRNA gene databases described above, I attempted to construct an improved 16S rRNA gene database, which is called GRD (Genomic-based ribosomal RNA Database) and the website address is http://grd.cb.k.u-tokyo.ac.jp.

3.2 Methods

3.2.1 Detection of 16S ribosomal RNA gene in the genomes publicly available

I downloaded 2,275 completed and 5,664 draft genome sequences including Human Microbiome Project (HMP) data from NCBI and performed similarity search with 16S rRNA gene sequence annotated of *Escherichia coli* str. K-12 substr. W3110 using GLSEARCH (version 36.3.5e) [71]. I extracted 16S rRNA candidate sequences having global/local scores of ≥0 with the *E. coli* 16S rRNA gene sequence. These searches were performed both with combined individual systems and in-house cluster system using Torque job scheduler with dedicated configuration for the massive jobs.

3.2.2 Curation of 16S in complete genome sequences

The 16S rRNA candidate sequences obtained from the complete genomes were compared with the *E. coli* 16S sequence for *Bacteria* and 16S rRNA sequence of *Pyrobaculum aerophilum* str. IM2 for *Archaea*. Additionally, BLASTN (version 2.2.26) [72, 73] search with their own genome sequences was also performed. The results indicated the existence of several annotation errors due to insertion sequences and lacking of part of 16S rRNA gene probably due to the algorithm. Therefore, I manually edited all the sequences using MAFFT [74] with G-INS-i, E-INS-i, and FFT-NS-2 algorithms, ClustalX (version 2.1) [75], and visualized genome mapping results from BLASTN output. I successfully enumerated highly curated 16S rRNA sequences including ones having insertion sequences such as a 16S rRNA gene having a length of 2,933bp in *Micrococcus luteus* NCTC 2665.

3.2.3 Curation and construction of full-length 16S rRNA gene in draft genomes

Several consortia including IHMC and individual research groups have also been publishing bacterial draft genomes, of which the sequence quality is less than that of the finished genomes. Therefore, many of draft genomes contained not only full-length but also partial 16S rRNA gene sequences. To generate full-length or nearly full-length 16S rRNA gene sequence from draft genomes, I performed assembly of partial sequence fragments of 16S rRNA gene in each genome. I developed a pipeline using shell script, Ruby language, and BioRuby (version 1.4.3) API [76] to map partial sequences to reference sequence. Therefore, the selection of reference sequence was important. I collected 16S candidate sequences from scaffold or contig sequences of each draft genome using the same method as that for extracting 16S rRNA gene sequences from the finished genomes. When the finished genomes of species taxonomically same or similar to species to be analyzed were available, I used their full-length 16S rRNA genes as reference sequences for alignment with partial 16S sequences in draft genomes. When appropriate species closely related to species to be analyzed was unavailable, I used full-length 16S sequence of *Escherichia coli* for *Bacteria* and that of *Pyrobaculum aerophilum* for *Archaea* as references. First, all candidate sequences in each draft genome were split by nucleotide code N and assembled to construct one file of multiple FASTA format using Phrap assembler [77]. After aligning the assembled sequences with the reference sequence, distance of each 16S candidate sequence with the reference was calculated. From these distances, most likely 16S sequences were obtained. When candidate sequences still had several gaps between the assembled sequences, I inserted Ns for filling gap according to the length of reference sequence.

3.2.4 Detection of full-length 16S rRNA genes which have 5' and 3' regions in GRD and GenBank

GRD constructed by above-mentioned process was evaluated in comparison with GenBank by investigating the existence of 5' or 3' ends in 16S rRNA genes. I used 8F primer region for the 5' end and the anti-SD site for the 3' end. I extracted the annotated 16S rRNA genes in the GenBank by searching "16S", "ssu", "rrna", "small-subunit", or misannotations like 'LSU" as keywords.

3.2.5 Comparison of microbial compositions using GRD and other public 16S rRNA gene databases

To compare and assess the quality of GRD and other 16S rRNA gene databases, I downloaded 16S rRNA genes from the newest versions of three databases (Greengenes: release 12.10, RDP: release 10 update 32, SILVA: release 111), and also downloaded query sample for the assessment of the accuracy from NCBI (Table 3.1). I used even mixture community (HMP synthetic mock even mixture - 454 sequencing result) as query sample for the assessment because only this sample had qPCR data, enabling to validate the results (Table 3.2). Alignment was performed by BLASTN under the conditions of the aligned length ≥ 200bp, e-value < 1e-5, and top hit.

Table 3.1 Websites for 16S rRNA gene databases and HMP mock community

Database & Sample	URL
Greengenes	http://greengenes.secondgenome.com/downloads/database/12_10
RDP	http://rdp.cme.msu.edu/misc/resources.jsp
SILVA	http://www.arb-silva.de/no_cache/download/archive/current/Exports/
HMP mock	http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=viewer&m=data&s=viewer&run=SRR072233

Table 3.2 HMP Mock even mixture community

Organism	Theoretical		qPCR	
	16S	%	16S	%
Methanobrevibacter smithii ATCC 35061	10,000	4.76	NA	0.000
Actinomyces odontolyticus ATCC 17982	10,000	4.76	1.00E-15	2.092
Propionibacterium acnes DSM16379	10,000	4.76	8.76E-16	4.245
Bacteroides vulgatus ATCC 8482	10,000	4.76	7.57E-16	5.372
Deinococcus radiodurans DSM 20539	10,000	4.76	1.05E-15	44.843
Bacillus cereus ATCC 10987	10,000	4.76	4.47E-16	2.232
Listeria monocytogenes ATCC BAA-679	10,000	4.76	5.03E-16	2.117
Staphylococcus aureus ATCC BAA-1718	10,000	4.76	5.89E-16	3.166
Staphylococcus epidermidis ATCC 12228	10,000	4.76	5.13E-16	6.832
Enterococcus faecalis ATCC 47077	10,000	4.76	8.25E-16	0.720
Lactobacillus gasseri DSM 20243	10,000	4.76	3.25E-16	1.259
Streptococcus agalactiae ATCC BAA-611	10,000	4.76	3.17E-16	1.544
Streptococcus mutans ATCC 700610	10,000	4.76	4.17E-16	3.015
Streptococcus pneumoniae ATCC BAA-334	10,000	4.76	5.54E-16	3.916
Clostridium beijerinckii ATCC 51743	10,000	4.76	4.40E-16	2.317
Rhodobacter sphaeroides ATCC 17023	10,000	4.76	1.41E-15	2.467
Neisseria meningitidis ATCC BAA-335	10,000	4.76	5.83E-16	3.153
Helicobacter pylori ATCC 700392	10,000	4.76	8.55E-16	1.408
Escherichia coli ATCC 700926	10,000	4.76	6.81E-16	1.065
Acinetobacter baumannii ATCC 17978	10,000	4.76	8.16E-16	5.246
Pseudomonas aeruginosa ATCC 47085	10,000	4.76	1.61E-15	2.991
Candida albicans ATCC MY-2876	1,120	0.53	NA	0.000

Percentage of qPCR was converted gDNA per copy of qPCR into 16S copy number.
NA : not analyzed.

3.2.6 Calculation of boundary identity between taxonomic levels

I used highly curated 16S rRNA gene dataset from GRD for calculation of 16S sequence identity at each taxonomic level. Filter-passed 16S sequences having ≥1,400 bp were performed with multiple alignment using MAFFT (v7.029b) with G-INS-i algorithm [74] for accurate alignment. Aligned sequences were performed distance-calculation by Clustal Omega (version 1.1.0) [78]. From the calculated distance matrix, I obtained sequence identity of boundary between each taxonomic level using distance calculation algorithm described below.

$$s_i = representative\ sequence\ of\ \ i\ at\ the\ species\ or\ strain\ level$$

$$c_i = sub\ taxon\ of\ s_i\ in\ the\ calculated\ taxon$$

$$d_{ij} = distance(s_i, s_j)\ if\ c_i \neq c_j$$

All data handling and statistics calculation were done using Ruby and R. In the calculation of higher taxonomic level than species, sequences having the shortest distance from others in the same species were used as representative sequences in each species, respectively. At species level, representative sequences were selected from each strain of the same species. At strain level, all sequences were used for calculation of sequence identity. In this way, I obtained average, minimum, maximum identities of boundary sequences between each taxonomic level with confidence interval of 95%.

3.3 Results

3.3.1 Evaluation and comparison of the quality of full-length 16S rRNA genes between GRD and GenBank

Among 7,038 genomes downloaded, I could totally extract 13,252 16S sequences from 6,519 genomes, of which the average length was 1,520 bp (Table 3.3). However, the multiple alignments of the 16S sequences identified length variations due to the presence or absence of various insertion sequences. Among them, the longest 16S rRNA gene had a length of 3,606 bp of *Pyrobaculum* sp. 1860. Previous study also reported the existence of 3.5kb of 16S rRNA gene of *Thiomargarita namibiensis* [67].

I compared the ratio of strains having the 5’ or 3’ regions in the 16S sequences between GRD and GenBank. In GRD, 100% and 99.9% of total 16S sequences had the 5' and 3' regions, respectively, while 85.2% and 56.7% had 5' and 3’ regions, respectively, in GenBank (Table 3.4). For example, 16S sequence of *Acetobacter pasteurianus* IFO 3283-03 lacked the 5’ and

3’ regions in GenBank, which was annotated from downstream of 27F universal primer region to 13 bp upstream of anti-SD site.

Several differences in copy number and length of 16S rRNA genes were also observed between GRD and GenBank. I found that the average copy numbers of 16S rRNA genes in a genome between GRD and GenBank were almost same. However, several genomes showed differences in 16S copy number between GRD and GenBank. For example, *Staphylococcus aureus* subsp. *aureus* ST398 had no 16S annotation in the GenBank, although other genes were annotated. As another example, *Lactobacillus casei* BD-II had ten copies of 16S sequences in GenBank, while the same strain had five copies of 16S sequences in GRD.

Comparison of the length of 16S rRNA genes between GRD and GenBank showed that the average length of 16S rRNA genes in GRD was 16 bp longer than that in GenBank. However, I found several 16S rRNA genes of which the length difference was more than 1,000 bp between GRD and GenBank. For example, 16S rRNA gene of *Brachyspira pilosicoli* 95/1000 had a length of 3,281 bp in GenBank, which contained non-16S sequences and the accurate length was 1,513 bp in annotation by GRD. For another example, 16S rRNA gene of *Pyrobaculum* sp. 1860 was annotated for a length of 1,499bp in GenBank. However, this annotation excluded intron sequences present in the 16S gene. In GRD, the length of this 16S gene was 3,606 bp including intron sequences.

I also investigated the distribution of 16S copy number in species. For this analysis, I selected and used 1,353 unique representative species from strains in GRD. The result indicated that there is the exponentially inverse correlation between 16S copy number and species number, and that the half of the total species analyzed had three or less 16S copy numbers in their genomes (Figure 3.1). The data also showed that the average 16S copy numbers were 1.65, 3.78, and 3.58 copies of 16S rRNA genes for *Archaea*, *Bacteria*, and all prokaryote, respectively.

I identified 32 strains having difference in 16S copy number with more than one between GRD and GenBank (Appendix Table 1) and 139 strains having difference in the average

length with more than 100 bp between GRD and GenBank (Appendix Table 2). I also found 108 strains having high copy numbers of 16S rRNA genes with more than 10 (Appendix Table 3), and 443 strains having single copy of 16S rRNA gene in GRD, respectively (Appendix Table 4).

Table 3.3 Detection of genomes containing curated 16S gene in GRD

	GenBank (Total)	Detected genome
Complete genome	2,277	2,275
Draft genome	4,761	4,244
Total	**7,038**	**6,519**

Table 3.4 Ratio of full-length 16S rRNA genes in the complete genomes in GRD and GenBank

		Archaea	*Bacteria*	All
No. of 16S gene		147	2,128	2,275
16S sequences extracted by GRD	**Existence of 5'**	147 (100%)	2,128 (100%)	2,275 (100%)
	Existence of 3'	147 (100%)	2,126 (99.9%)	2,273 (99.9%)
	Avg. no. of 16S	1.63	4.12	3.96
	Avg. length(bp)	1,527	1,529	1,529
16S sequences extracted by GenBank	**Existence of 5'**	122 (83%)	1,816 (85.3%)	1,938 (85.2%)
	Existence of 3'	101 (68.7%)	1,188 (55.8%)	1,289 (56.7%)
	Avg. no. of 16S	1.61	4.08	3.92
	Avg. length(bp)	1,484	1,515	1,513

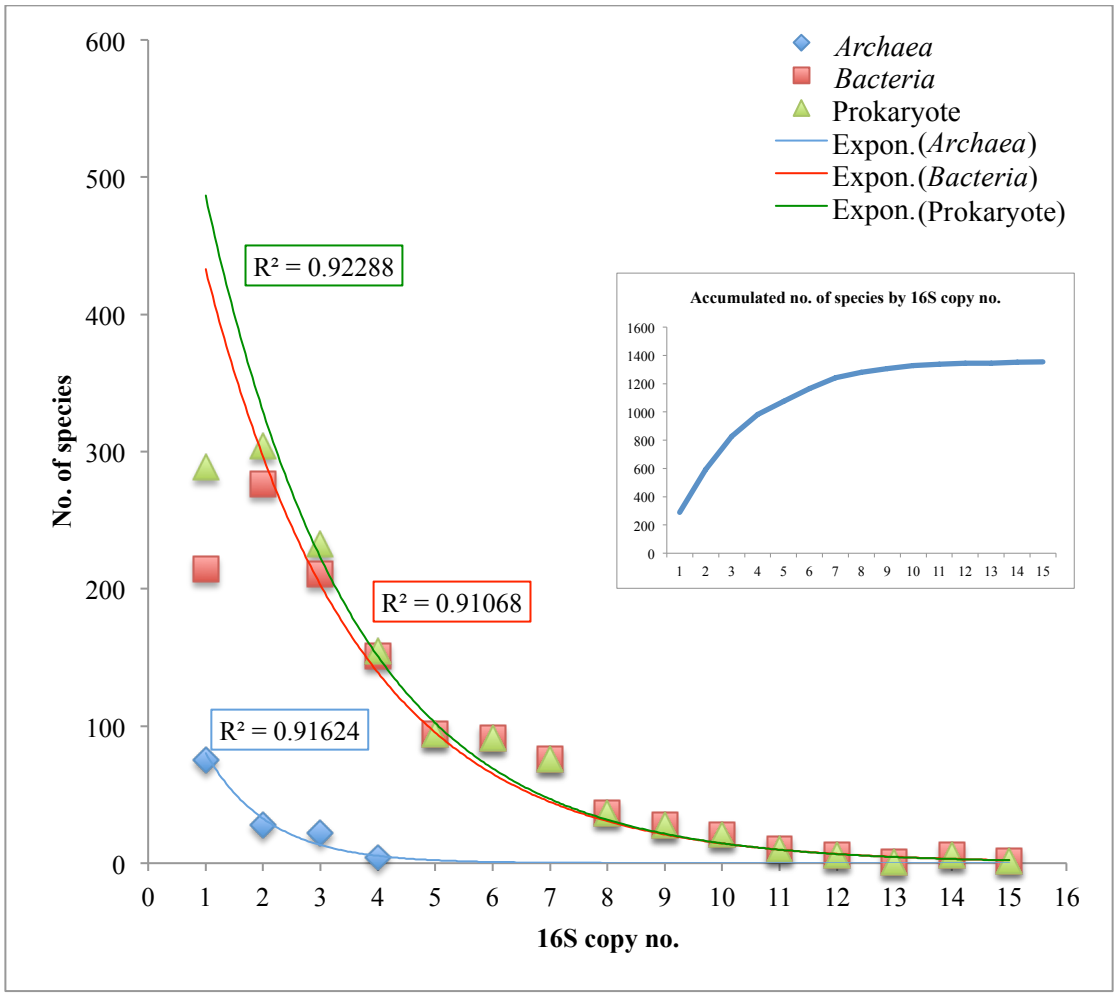

Figure 3.1 Distribution of 16S copy number in species.
The results show the exponentially inverse relationship between 16S copy number and species number.

3.3.2 Comparison of microbial compositions between GRD and other public 16S rRNA gene databases

I performed comparative analysis of GRD with other 16S rRNA gene database for the assessment of accuracy. Commonly used 16S rRNA gene databases were Greengenes conducted by Second Genome Company, RDP by Michigan State University, and SILVA by Max Planck Institute for Marine Microbiology. For assessing the result of assignments, mock even mixture community published by HMP (Table 3.2) was used as query sample. The metagenomic shotgun sequencing data generated by 454 pyrosequencing of HMP Mock community sample had the average read length of about 400 bp. Among reads randomized

generated from the constituted genomes in Mock community, some reads contain part of 16S rRNA gene, its upstream or downstream sequences including 16S-23S internal transcribed spacer (ITS), and part of 23S rRNA gene. GRD was constructed by completely removing unrelated sequences, resulting in 16S rRNA gene database composed only of fully curated 16S sequences. Therefore, one can expect that all metagenomic reads assigned by GRD contain parts of 16S rRNA gene. On the other hand, if database contains sequences other than 16S rRNA gene such as flanking ITS and 23S rRNA gene, metagenomic reads assigned by the database might include ones that do not contain 16S rRNA gene sequence. I performed assignment of metagenomic reads of HMP Mock community sample to GRD and other three databases. The result showed that the number of reads assigned by Greengenes was similar to that by GRD, and SILVA assigned about 1,000 more reads than GRD, but all of them did not contain 16S sequences (Table 3.5). I performed multivariate analysis of the species assignment data of each database to evaluate the quantitative accuracy for estimation of the bacterial composition at the genus level. The results revealed that the bacterial composition estimated by GRD was closest to that by qPCR data, which was thought to be the nearest to the true bacterial composition in HMP Mock community sample. In contrast, the estimation by SILVA showed most different from that by qPCR data (Figure 3.2). These results indicated that GRD gave most quantitatively accurate data for the analysis of the bacterial composition based on the 16S rRNA gene among the databases examined, and that other three databases are contaminated by sequences other than 16S rRNA gene probably due to the incorrect annotation.

Table 3.5 Number of metagenomic reads hit to sequences in GRD and other databases

Database	GRD	Greengenes	RDP	SILVA
No. of assigned reads	3,360	3,362	3,746	4,318

Thresholds were match length ≥ 200bp, e-value < 1e-5, and top hit.

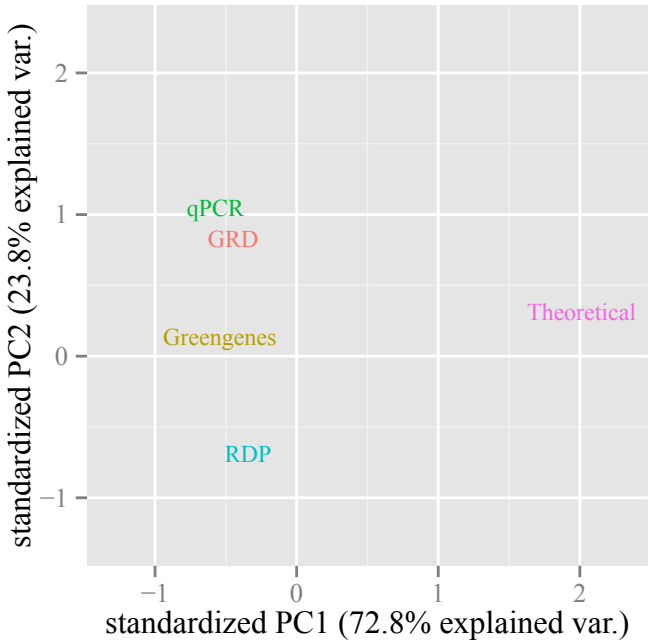

Figure 3.2 Multivariate analysis for evaluation of quantitative accuracy of bacterial compositions estimated from the results using GRD and other databases.
HMP Mock community even mixture sample was used for comparative analysis. PCA was performed at the genus level.

3.3.3 System architectures and features of the web service

For the general purpose to share and make use GRD, I developed the website for GRD. This system was established based on Linux Operating System, and MySQL (version 5.1.69) was used as Database Management System (DBMS). Web interface was developed using HTML5 web standard with Apache HTTPD Server (version 2.2.15-28) and FCGI module for web application. Figure 3.3 explained the architecture of server-side and client-side software for the web service. The communication between JavaScript of front-end layer and applications layer used WebSocket protocol (RFC 6455) that was implemented in HTML5. General method for dynamic (asynchronous) communication was Asynchronous JavaScript and XML (Ajax) technique. However, the advantage of WebSocket is the different connection method compared with conventional asynchronous communication. Once connection is established, its connection can be used again until user leaves the page unlike Ajax technology like XMLHttpRequest. Especially, this method can be totally useful for many requests in the search page in this website, so that overall performance can become up. Currently, there are

several versions of WebSocket protocols including draft version, theoretically I prepared to respond to all version of WebSocket. Furthermore, for users who are using old web browser, Flash version of WebSocket was also applied using web-socket-js (https://github.com/gimite/web-socket-js). Application layers were implemented using Ruby, BioRuby, etc. All application has fault-tolerant structure. This means that even in the critical software level faults, application system will be restored. For massive transactions, I constructed application layer as a distributed system and therefore, when user access the website, front-end layer connects to one of servers. This architecture was pretty simple, however, it could be easy to aggregate or distribute each sub-system. I designed three main tables for database storage and two views for showing the results (Figure 3.4). In the table "bacteria_info" included overall information of its organism like project ID, taxon ID, culture status, etc. In the table "data_source" included accession number and genome size, etc. of each organism. And, the table "sequences" included each 16S sequence and its detailed information like a position and a length. For updating the database, I implemented publishing and obsoleteness concepts. When sequences in NCBI are updated, sequences of GRD can also be updated logically without any harmful of data integrity. Each table has each relationship between tables. However, I did not set up the relationship in the DBMS due to a consistency with GenBank. Therefore, database table has each relationship in the application layer, but not in the system layer.

For web service, I implemented essential functions like a general search, a taxonomy tree, a sequence search, and a dataset download in the website (Figure 3.5). The most important function that I focused was a general search function. For easy to use and simple search, the biggest tip could be obtained from Web search site such as Google. I separated two types of search mode. Default setting is "Taxonomic Name Search". In this mode, user can search using taxonomic name, taxon ID, or accession ID. If user wants to search using complicated search keyword, "Deep Search" mode is able to do for it. In "Deep Search" mode, when user queries one or multiple keywords, the search engine of this site basically finds all contents

and shows the result. Additionally, I prepared several specific search keywords. If user queries just “bacteria”, search results of *Archaea* taxon due to taxa contained such kind of taxon names like “-*bacteria*” in *Archaea* group. In this case, if user search as keyword “taxon:archaea”, the output shows the exact result. Every search is based on case-insensitive. Two or more keywords search in the same time is available using “AND” and “OR” keywords. For example, if users want to search “complete” genome strains concerned with “environmental”, they just input “complete AND environmental”. If “AND” and “OR” are used in the same time, "AND" has higher priority than "OR". For example, "A and B or C and D" means "(A and B) or (C and D)". If users want to exclude some keyword from the result, user can use “-“ prefix. For example, “soil and marine” keyword returns the results that have both “soil” and “marine” keyword. However, if users input “soil and –marine”, the result shows the results that include “soil” keyword excluding the results that include “marine” keyword. Recent GenBank dataset have been included various metadata like sequencing information in “comment” field. More detailed search examples are written in the GRD website. In the research result, when user selects the organism name, the website shows detailed information based on the information of GenBank and its 16S sequences information, and checked organism sequences in the checkbox can be downloaded directly. User can also download all of 16S sequences data including index directly. A downloaded file is zip-archived, and for the convenience, two types of taxonomy information are included. One type includes all rank, and another type includes fixed rank. Theoretically, full taxonomy rank consisted of 28 taxonomic ranks based on the NCBI Taxonomy database [79]. Each bacterium has some of them from all taxonomic rank, respectively. Therefore, the general type using all rank was included taxa of all ranks classified. On the other hand, the fixed rank type used eight taxonomy ranks like “domain” (or “superkingdom”), “phylum”, “class”, “order”, “family”, “genus”, “species”, and “strain”. In the case of fixing taxonomic ranks, some organisms were not classified at some of these eight taxonomic ranks. In this case, I inserted *ad hoc* taxon name like below.

"*@parentknownrankname*_currenttaxonomyrank"

For example, *Acaryochloris marina* MBIC11017 was not classified at class and family levels, and parent taxon names of them were classified as "*Cyanobacteria*" (phylum) and "*Chroococcales*" (order). In this case, I named them as "*@Cyanobacteria*_class" at class level and "*@Chroococcales*_family" at family level. By this, researchers can identify a relationship of its rank and it can be useful to summarize.

Taxonomical lineage-based browser was also implemented in the website. If users want to download several branches of taxonomy, it would be totally useful. NCBI taxonomy team is performing curation tasks of taxonomy information. Therefore, it will be changeable irregularly by update of GRD. According to this reason, taxonomy tree is loaded every time.

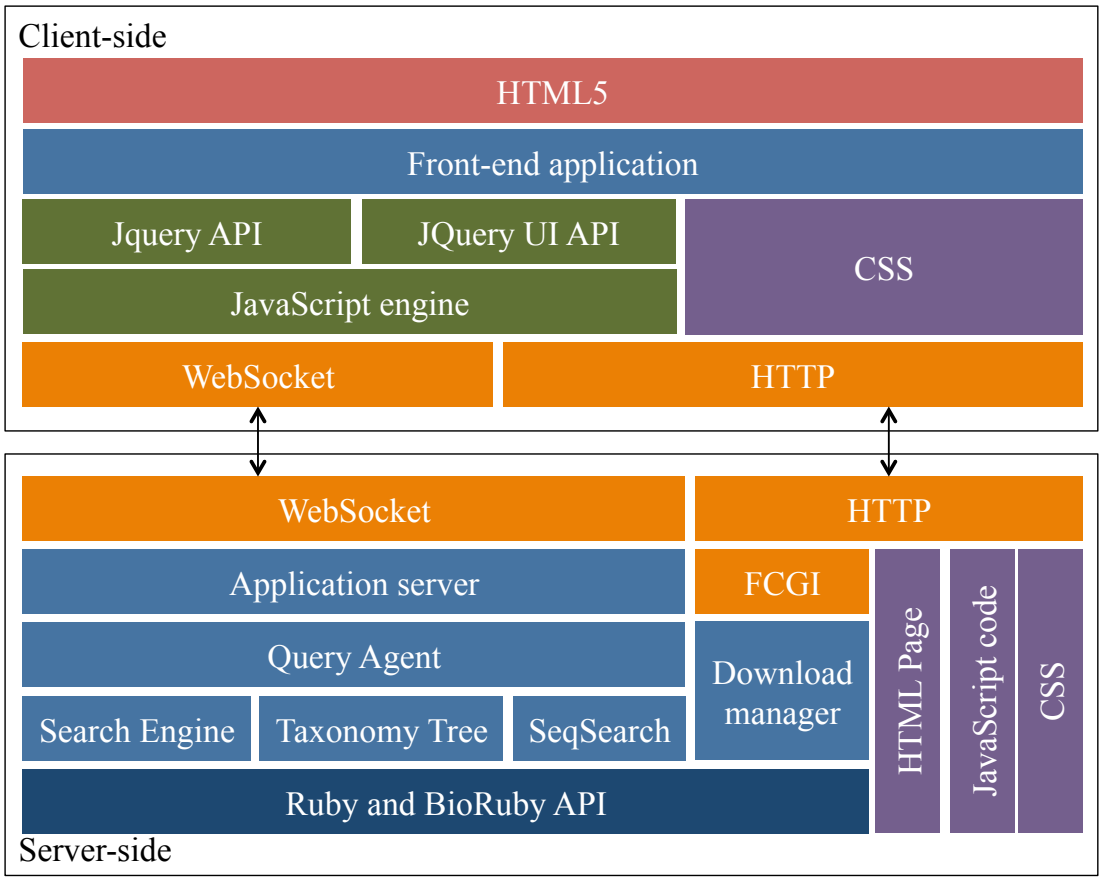

Figure 3.3 Overall application architecture.
This shows all application and communication layer in the web application system. If user in the client-side (web browser) requests a query, it transfers server-side application through each layer, and server-side application performs response for user's request through server-side layer to client-side application, and finally web browser shows the result to user. There are two types of communication in the system. HTTP communication is used for static data like HTML, JavaScript code, and CSS. WebSocket communication is used for dynamic request and response like throwing query and returning results. Query Agent manages transferring data from user's request with non-blocking mode.

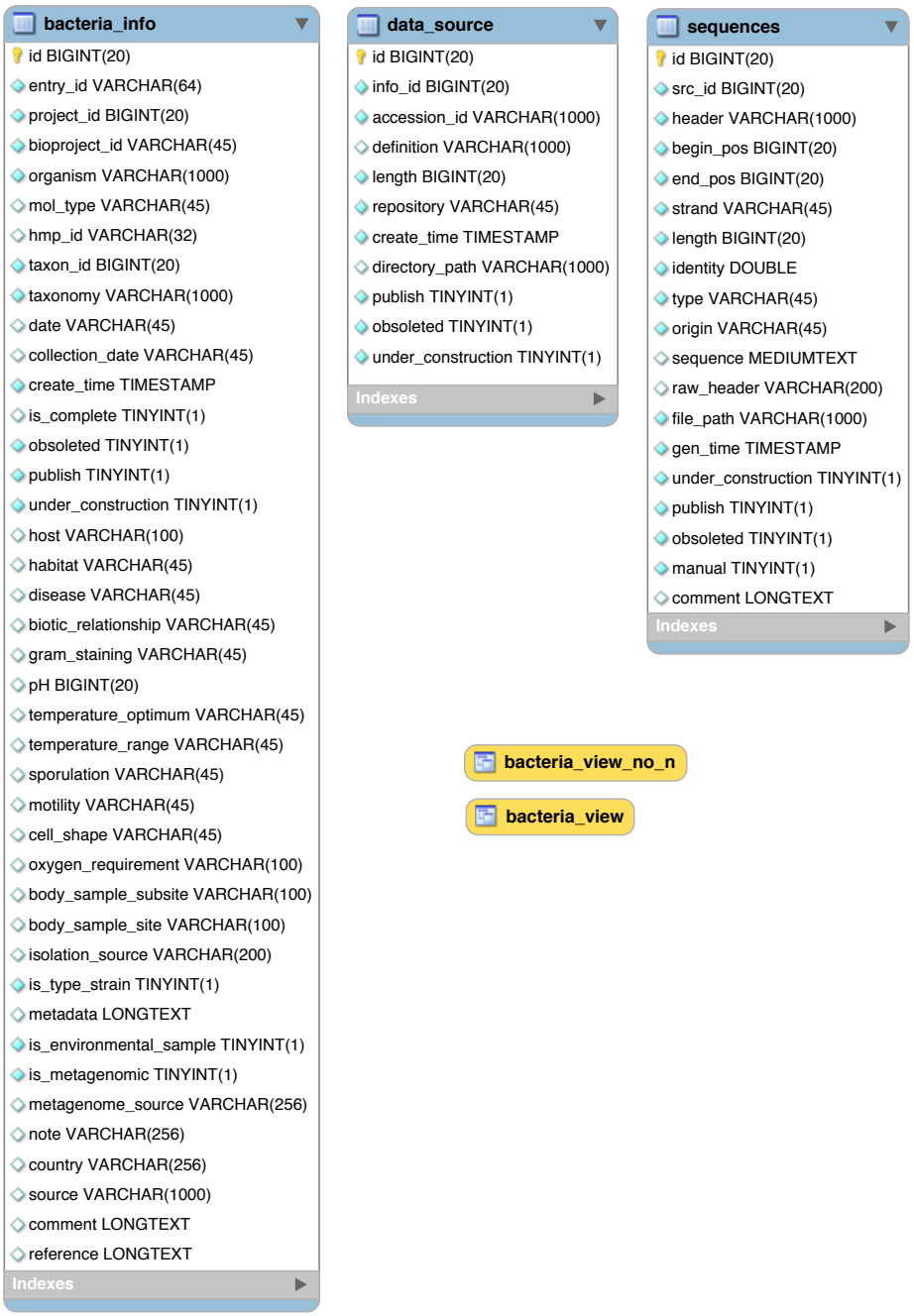

Figure 3.4 Database schema for the website.
All information was stored in each table as necessary dataset. Each table has relationship with another table. However, it has not strict rule in the system level. For the version management, each table has “obsoleted,” “under_construction”, and “publish” fields, and these field values are checked and appropriate results are returned when user requests a query.

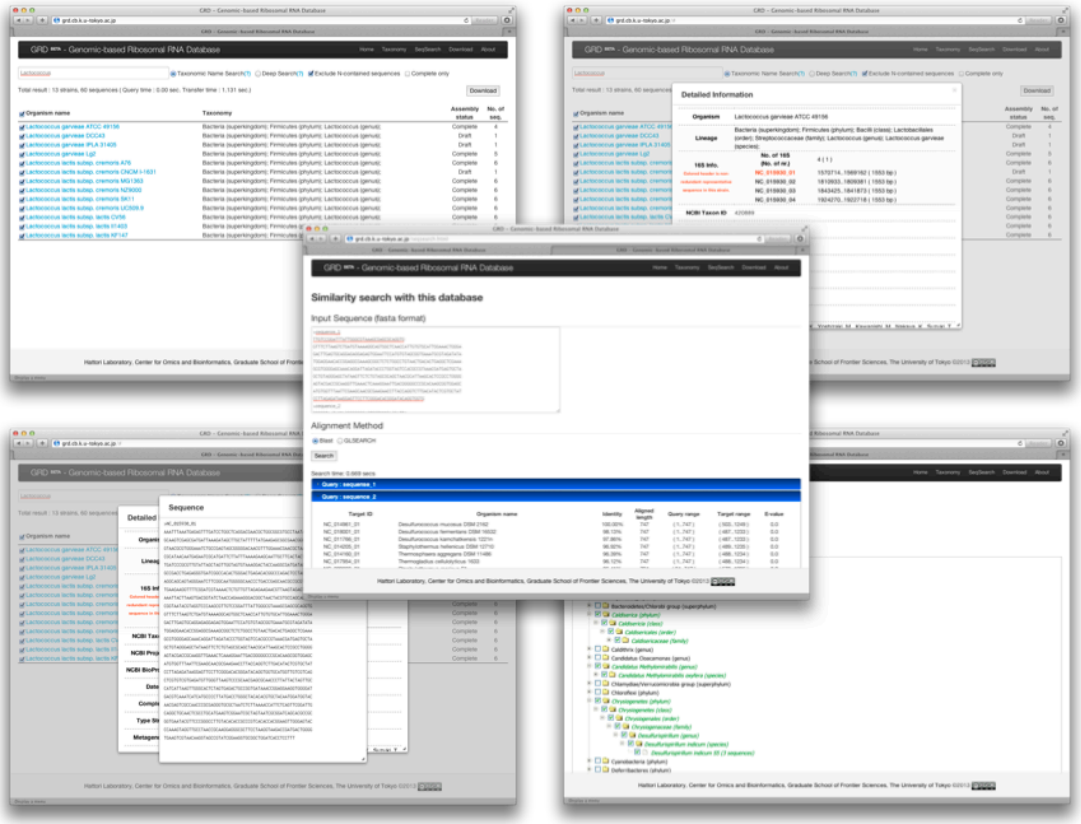

Figure 3.5 Features of the GRD website.
Users can do broad or specific searches using pre-defined search keywords (left upper). In the result, detailed information of the strain including sequence information is shown (right upper). If users click sequence name, it shows its 16S sequence, and users can copy it for another purpose (left lower). Taxonomic tree is also supported (right lower). Users can perform alignment with GRD using their own sequences (center).

3.3.4 Boundary sequence identity of 16S sequences between each taxonomic level

I investigated the sequence identity at boundary between each taxonomic level using GRD. Previous study showed boundary identity at each taxonomic level [80]. However the result was based on type strain and shown only at phylum, family, and genus levels. In this study, I therefore calculated identity boundaries at major taxonomic levels. The results indicated identity boundary at each taxonomic level (Table 3.6). Each identity was calculated with 95% confidence interval, respectively. In this result, minimum identities at species and strain levels were 98.9%±0.3 and 99.7%±0.1. The comparison between the minimum identity at upper taxonomic level and the maximum identity at lower taxonomic level indicated that identity

boundaries were overlapped from domain level to family level. Especially, the result visualized as a boxplot showed overlapped identity between class and order levels (Figure 3.6). However, average identity and median identity at each taxonomic level showed each taxonomic level had gradationally different identity.

Table 3.6 Boundary identity of 16S rRNA genes between each taxonomic level

Category	Prokaryote	Domain	Phylum	Class	Order	Family	Genus	Species	Strain
No. of taxa	1	2	10	21	53	173	340	426	1835
No. of compared sub rank	2	36	45	94	213	776	1921	3993	-
No. of species	2488	2488	1935	2044	1907	2164	1921	-	-
Maximum identity	-	87%±0.5	86.7%±5.1	88.8%±3.4	89.6%±1.7	92.3%±0.8	97.3%±0.4	99.6%±0.2	100%±0
Average identity	42.3%	67.9%±0.2	76.3%±3.1	80.7%±2.6	84.7%±1.3	88.5%±0.9	95.1%±0.5	99.3%±0.2	99.8%±0
Minimum identity	-	47.7%±16	66.9%±6.1	70.6%±5.1	80%±1.8	85.4%±1.1	93.5%±0.7	98.9%±0.3	99.7%±0.1

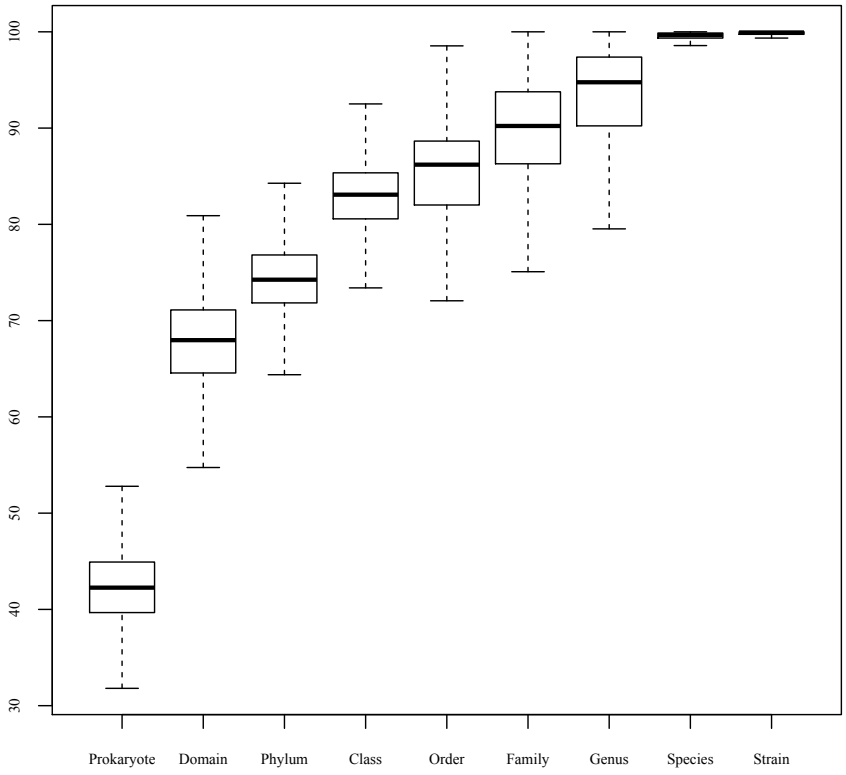

Figure 3.6 Boxplot of boundary identities of full-length 16S rRNA gene at each taxonomic level (without outliers).
This showed each taxonomic level was separated except for between class and order level by identity. The x-axis showed taxonomy rank name and the y-axis showed the identity. Prokaryote result showed the strongly different boundary was existed. The odd boundaries between class and order level indicated that a taxonomical classification was not clarified in the aspect of 16S rRNA.

3.4 Discussion

I constructed well-curated 16S database (GRD) using the method of direct extraction from genome sequences. This GRD would be helpful to analyze bacterial communities including human microbiome samples. And, it can be applied as background database for 16S assignment of novel genomes. Using GRD, one caveat is a taxonomic name like "[*Clostridium*] *bartlettii*". This type of name was changed by NCBI, which might be actually different from taxon name inside brackets. Researcher should consider this kind of name in taxonomical classification using GRD including NCBI taxonomy. In the comparison with GenBank, I showed that 16S of GRD was more accurate than 16S of GenBank. According to this result, GRD would be a good reference for 16S annotation. In comparative analysis with other public 16S databases, I set up the cut-off threshold of minimum 200 bp. It meant that even 200 bp of outside of 16S also was permitted. About 200bp from 3' of 16S might include tRNA or 23S. Therefore, when researchers perform studies using only general public 16S database, the result should be analyzed carefully. The GRD was checked 5' region and 3' region including anti-SD site strictly.

I also investigated the identity boundary at each taxonomic level using GRD. First of all, the result indicated that minimum identity at species level was 98.9%±0.3. This result suggests the standard of the taxonomical classification at species level. In the identity boundaries result, the result showed the overlapping of identity boundary between class and order levels especially. This result implies that it was not clarified in the taxonomical classification by 16S between class and order levels. However, the result of each identity boundary of each bacterium showed its own boundary, respectively (data not shown). This implicates the availability of taxonomical assignment using 16S similarity.

4 Evaluation of improved universal primer and quality of 454 pyrosequencing data using mock communities

4.1 Background

The most of analysis focused on the composition of specific bacterial species or groups by conventional methods such as culturing, quantitative PCR (qPCR), fluorescence in situ hybridization (FISH), terminal-restriction fragment length polymorphism (RFLP), and denaturing gradient gel electrophoresis (DGGE) based on bacterial 16S ribosomal RNA gene (16S). These conventional methodologies may also overlook subtle changes in bacterial community structure and change of species other than targeted species.

Pyrosequencing of PCR amplicons of bacterial 16S rRNA gene hypervariable regions is the most popular and a high-throughput approach to infer and characterize the species composition in bacterial communities [81-84]. The 454 pyrosequencing platform that can produce over 400 bases per read is also superior to shorter read-length sequencers with respect to sequence accuracy in single-end sequencing [85-87]. However, this PCR-based method has a problem particularly in quantification of the composition of the genus *Bifidobacterium*, a dominant species in human gut microbiota because the 16S sequence of *Bifidobacterium* has a few base mismatches with the commonly used PCR primer 27F (or 8F), underestimating this genus in the community [88-91]. To improve this lack, I developed modified 27F primer (27Fmod) by changing the third base A to R (G or A) in 27F-YM [89], by which made the perfect match was made with the annealing site of the 16S gene of *Bifidobacterium*.

Microbial composition analysis using 16S rRNA gene is the OTU-based analysis. The core of this analysis is clustering of 16S sequences by identity. However, sequence error is also a

point to consider clustering conditions of 16S sequences because sequence errors is one of the factors leading to overestimation of OTU number.

In this chapter, I assessed the modified universal PCR primer 27Fmod for the analysis of pyrosequencing data of 16S rRNA gene V1-V2 regions using mock communities, and estimated sequence error rate of 454 pyrosequencing data that may affect OTU number generated from clustering.

4.2 Methods

4.2.1 Construction of mock communities

Two artificial bacterial communities (designated "mock01" and "mock02") were constructed by mixing genomic DNA from ten and eleven different human gut-associated bacterial strains with an appropriate ratio, respectively (Table 4.1). Genome sequences of these microbes were completely sequenced and are publicly available.

Table 4.1 Microbial composition of two mock communities

Mock community	Bacterial species	Strain	Genome size in Mb	16S copy number	Ratio
01	*Bifidobacterium bifidum*	JCM1255	2.0	2	20.0%
	Bifidobacterium sp.	JCM15439	2.3	2	8.0%
	Bacteroides sp.	A_01	5.9	7	30.0%
	Lactobacillus paracasei	JCM 8130T	3.0	5	1.0%
	Lactobacillus zeae	ATCC393T	2.9	5	0.9%
	Clostridium paraputrificum	JCM1293	3.5	9	5.0%
	Clostridium ramosum	JCM1298	3.2	5	15.0%
	Ruminococcus gnavus	TBH11	4.0	5	15.0%
	Escherichia coli	SE11	4.9	7	0.1%
	Raoultella ornithinolytica	AA097	5.5	7	5.0%
02	*Bifidobacterium bifidum*	JCM1255	2.0	2	10.0%
	Bifidobacterium sp.	JCM15439	2.3	2	10.0%
	Bacteroides sp.	A_01	5.9	7	34.0%
	Bacteroides massiliensis	A_03	4.8	7	0.1%
	Lactobacillus paracasei	JCM 8130T	3.0	5	10.0%
	Lactobacillus zeae	ATCC393T	2.9	5	5.0%
	Clostridium paraputrificum	JCM1293	3.5	9	9.0%
	Clostridium ramosum	JCM1298	3.2	5	0.9%
	Ruminococcus gnavus	TBH11	4.0	5	1.0%
	Escherichia coli	SE11	4.9	7	5.0%
	Raoultella ornithinolytica	AA097	5.5	7	15.0%

4.2.2 Pyrosequencing of 16S rRNA gene V1-V2 region using 454 pyrosequencer

The 16S rRNA gene V1-V2 region was amplified using forward primer containing the 454 primer A and 27Fmod (5'-agrgtttgatymtggctcag) in which the third base A in the original primer 27F was changed to R like below.

5'-CCATCTCATCCCTGCGTGTCTCCGACTCAGagrgtttgatymtggctcag-3'

And, reverse primer was contained the 454 primer B and reverse primer 338R (5'-tgctgcctcccgtaggagt) like below.

5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGtgctgcctcccgtaggagt-3'

PCR was performed in 1 × Ex Taq PCR buffer (50 μL), deoxynucleoside triphosphate (2.5 mM), Ex Taq polymerase (Takara Bio, Inc., Shiga), each primer (10 μM), and 40 ng of extracted DNA under conditions of 2 min at 96 °C, 20 cycles of 96 °C for 30 sec, 55 °C for 45 sec, and 72 °C for 1 min, and a final extension of 72 °C for 10 min on a 9700 PCR system (Life Technologies Japan, Ltd., Tokyo). PCR products of approximately 370 bp were confirmed by agarose gel electrophoresis, purified by AMPure XP magnetic purification beads (Beckman Coulter, Inc., Brea, CA), and quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies Japan, Ltd., Tokyo). Mixed samples were prepared by pooling approximately equal amounts of PCR amplicons from each sample and subjected to 454 GS FLX Titanium or 454 GS JUNIOR (Roche Applied Science) sequencing according to the manufacturer's instructions.

qPCR was performed by using Taqman Gene Expression Assays (Life Technologies Japan, Ltd., Tokyo) with each bacteria-specific primers and probe shown in appendix A. Primers and probes for qPCR were designed by Ruby script with NCBI blast tool and Primer3Plus [92]. First, available primer and probes list was generated from Primer3Plus. Using the list, finding that the sequences (forward, probe, reverse) with similarity were performed blast with another

genome sequences. If those three sequences did not have any similarity with another genome sequences, its sequence group was selected (Table 4.2).

Table 4.2 PCR primers and probes used for qPCR in this study

Primer ID	Primer sequence
BA_1_124_F	GTCCCACGGGAAAACCTTAT
BA_1_124_R	AAGAAGAGACCCGAGTGCAA
BA_2_125_F	GTTGTGGTAGGCGTTTCGTT
BA_2_125_R	TGGCCATAGTGACCATCGTA
BB_1_112_F	TTGTGCGATACCACGATGTT
BB_1_112_R	TCAAGGAGCTCACCGTCTCT
BB_2_79_F	AGCCAAGGCAACATCCTCTA
BB_2_79_R	GTGACCGTGACTGTGGAATG
BS_1_98_F	AATCGACTGGTCGGAATCAC
BS_1_98_R	GCATTGACTCGGGAGAAGAG
BS_2_88_F	CGCTGACACACGCATAGAAT
BS_2_88_R	AAGCTTCATCTGCCATTGCT
CP_1_125_F	TTCACGGAAGGGTCCTAATG
CP_1_125_R	AAGCACTGAGGATGCCAGAT
CP_2_89_F	CCAATTGATCCTGGATTGCT
CP_2_89_R	GATGGTCCCTTTGGTGGATA
CR_1_106_F	TCACATAGCGGGCAATTGTA
CR_1_106_R	GTGGCTAAGACCGACCAAAA
CR_2_129_F	GTCGGGAAGACAGGATTTGA
CR_2_129_R	CAAGAGGTTGTGGGTTCGAT
EC_1_123_F	CGTGGTTTTACACGGCTTTT
EC_1_123_R	GCTGGCAATTTTTCGCTAAG
EC_2_124_F	CGGCTGAATATCTGGGAAAA
EC_2_124_R	ACATCACCTAAGCCCATTGC
LP_1_64_F	CAGTCGCTTGTTTCCACAGA
LP_1_64_R	GATCATGCAGCAGTCGAAAA
LP_2_81_F	CAAAACAACCTTGCCCACTT
LP_2_81_R	TTTGACACCATCTCGGTTGA
LZ_1_76_F	TCGCCAAATTACAACCATCA
LZ_1_76_R	AAAAGGCACCTCGCTCAGTA
LZ_2_99_F	TTGGTTGTTGTGCGAATGAT
LZ_2_99_R	CAAGGAGGGGAACCCTTTAG
RG_1_130_F	TGATCTGCCCGTGACAATAA
RG_1_130_R	AGTGCGAGGCATCTCTGAAT
RG_2_132_F	CCGTAATAAGCCAGCGGATA
RG_2_132_R	CGTGCCGGACTGGTATATCT
RO_1_92_F	TTACGGCACCAAACGTAACA
RO_1_92_R	GCGTGACAAGCGCTACAATA
RO_2_133_F	CGATACCTCGATGCCATTCT
RO_2_133_R	GTTCCGTCGACAGGTTTGAT

Probe ID	Probe sequence
BA_1	CAGCAAATCCCTGGGTAGGTCCATA
BA_2	ATTTCCTTGTCAGTTACCGCCTCCC
BB_1	TCACGTTCAGTGACTCCCATGTCCT
BB_2	ATCATTGAGCCAATCGCCCGAC
BS_1	CTACCACCGTAATCCAAGACGGTCC
BS_2	CCGTTTCCATCACGAGAATAGAGCC
CP_1	CCATCAAGAGCAATGGCTAGAGTTCC
CP_2	TCCCAAGGAGCATTTAGGCCTAACC
CR_1	CTTCATCGCCACCAGTACCCTCTTC
CR_2	AAACCAGGTGCACTACCAAGCTGTG
EC_1	CGCGTGATGTTGTCTGCTACTCAAC
EC_2	TTAGCCCGCATTGAATCTATCCGTC
LP_1	CTAGCGGCAGCACCGTAGTCCAT
LP_2	ACCTTCAACTAGAGCGGTCGCAGTC
LZ_1	CGCTCAACCACAGCCTTAACGACT
LZ_2	TGGCTGCAACTCCTGTATACACCCA
RO_1	CTTGTCGCACTTCAGCTCTGAGGAT
RO_2	CAATATCACCAGGGACGCGATAACC
RG_1	TCTGCTCTTCTGACAAATCTGTGCG
RG_2	CAATAGGCGATTCGATATGCCATCA

4.2.3 Quality Filtering of pyrosequencing data

From two mock communities, I amplified the V1-2 region by PCR using 27F-338R and 27Fmod-338R primers, the V5-6 region by 787F-1061R primers, and the V1-9 region by 27F-1492R primers. V1-2 and V5-6 amplicons were subjected to 454 pyrosequencing, and V1-9 amplicons were cloned in *E. coli* and 3,000 clones were sequenced by the Sanger method and the products were analyzed with the ABI3730xl (Life Technologies Japan, Ltd., Tokyo). I also performed duplicate qPCR experiments targeting a specific genomic region of the bacterial strains in the two mock communities. All filter-passed 16S *de novo* sequences and qPCR data were then analyzed by principle component analysis (PCA) to compare and assess the quantitative accuracy. The error rate of filter-passed sequences using 27Fmod-338R primers obtained from two mock communities was estimated by aligning 16S V1-V2 *de novo* with reference 16S sequences in two mock communities. Rarefaction curve of OTU numbers generated from clustering of various qualities of 16S reads was made by using MAFFT [93] and ClustalW (Version 2.0) [75, 94].

4.3 Results and discussion

4.3.1 Assessment of the quantitative accuracy of 16S pyrosequencing data using mock communities

To assess the 16S data using 27Fmod, I compared various 16S sequence and qPCR data obtained from two mock communities (Table 4.1), which are useful to evaluate the quantitative accuracy of 16S-based data and the sequencing error rate [95, 96]. Quantitative accuracy of the overall bacterial composition was evaluated by comparing the similarity of each data to the expected ("Expected") using PCA (Figure 4.1). From the PCA data, Euclidean distance was calculated for evaluation of the similarity of each data with the "Expected". The results revealed that the order of their similarities with the "Expected" was the qPCR data ≥ the V1-2 data using 27Fmod > the V5-6 data > the V1-2 data using 27F >>

the data of Sanger sequencing-based full-length V1-9, indicating that the use of 27Fmod greatly improved the quantitative accuracy for evaluation of the overall bacterial composition (Table 4.3). This improvement was largely dependent on the improved estimation of the *Bifidobacterium* content by the use of 27Fmod. The average relative *Bifidobacterium* content in the two mock communities estimated from the data of V1-2 using 27F was only 1.5% of the "Expected" (100%), while the use of 27Fmod increased the relative content to 61%, which was also better than those estimated from the data of V5-6 and Sanger full-length analyses (Figure 4.2). Since qPCR can be used only when genomes of all bacteria in a given community are known, or only for a limited number of specific known species, I concluded that 454 pyrosequencing of the V1-2 region using 27Fmod-338R provided more quantitatively accurate data for bacterial composition in human gut microbiota than that using the conventional 27F primer.

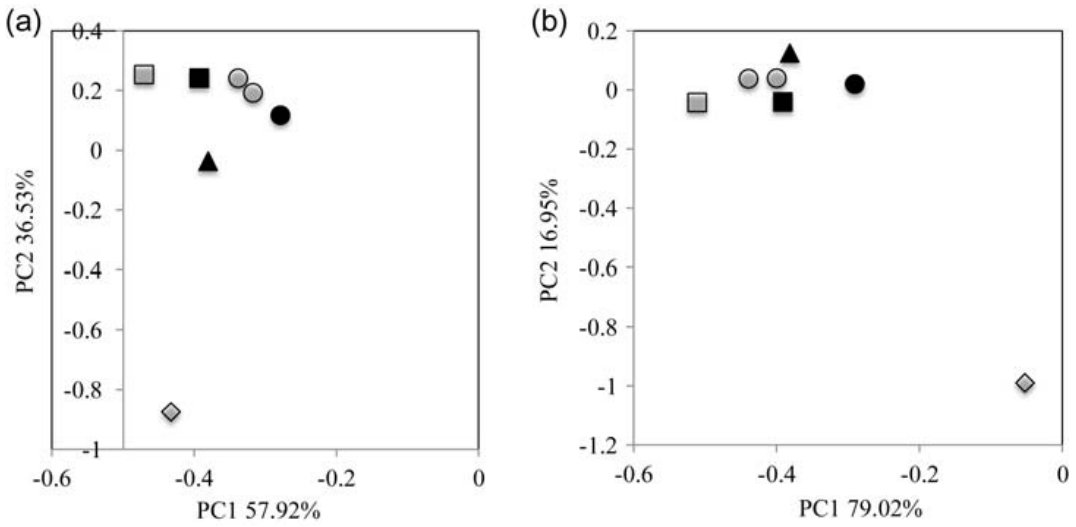

Figure 4.1 Assessment of the quantitative accuracy of the analysis of the bacterial composition of two mock communities by various methods.
PCA analysis of the data was obtained from various methods using mock01 (a) and mock02 (b). Closed circle: Expected, open circle: duplicate qPCR, closed square: pyrosequencing of 16S V1-2 region using 27Fmod, open square: pyrosequencing of 16S V1-2 region using 27F, closed triangle: pyrosequencing of 16S V5-6 region, open diamond: Sanger sequencing of nearly full-length 16S clone.

Table 4.3 Euclidean distance between each data and the "Expected"

Mock01	Euclidean distance	Relative ratio
Expected	0	—
Average qPCR	14.88	1.00
Pyrosequencing of the 16S V1-2 using 27Fmod	21.62	1.45
Pyrosequencing of the 16S V1-2 using 27F	30.21	2.03
Pyrosequencing of the 16S V5-6	24.96	1.68
Sanger sequencing of the full-length 16S	74.23	4.99

Mock02	Euclidean distance	Relative ratio
Expected	0	—
Average qPCR	16.18	1.00
Pyrosequencing of the 16S V1-2 using 27Fmod	13.93	0.86
Pyrosequencing of the 16S V1-2 using 27F	26.92	1.66
Pyrosequencing of the 16S V5-6	20.18	1.25
Sanger sequencing of the full-length 16S	53.07	3.28

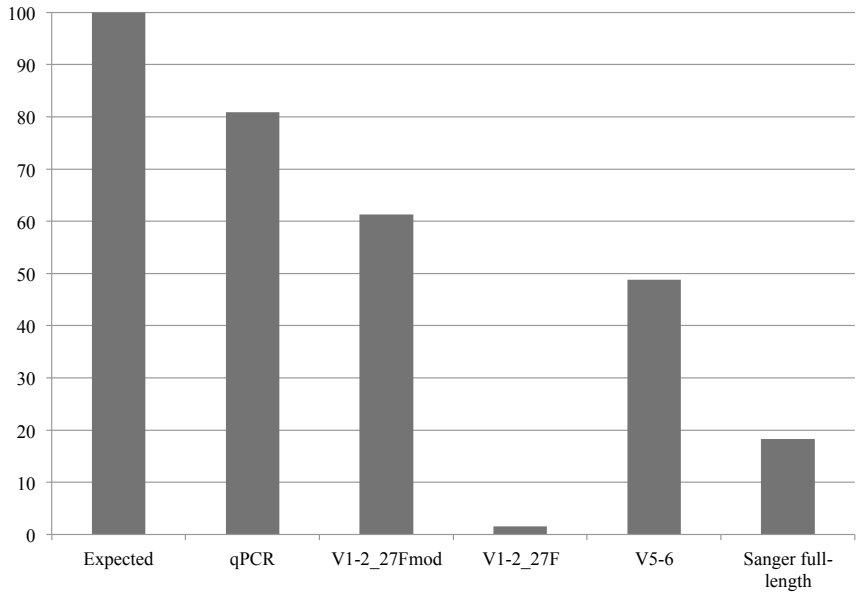

Figure 4.2 Relative ratio of the content of *Bifidobacterium* species estimated from various methods using two mock communities.
The y-axis indicates average values of the *Bifidobacterium* content obtained from each method when the "Expected" was set to 100.

4.3.2 Estimation of sequence error rate in 454 pyrosequencing data and clustering conditions for OTU analysis

I estimated the average error rate of filter-passed V1-V2 data using 27Fmod-338R by aligning the V1-V2 and reference 16S sequences of bacterial strains used in two mock communities. The error rate was estimated to be 0.58% and 0.66% for mock community 1 and mock community 2 in local alignment, respectively (Table 4.4). These error rates were similar to the previously published data [82, 85, 97], however, lower than other result [98]. The later might be due to differences in examined alignment length and between local and global alignments. Errors in 454 pyrosequencing data can be the primary cause for overestimation of OTU number, which is an issue to be improved for accurate estimation of species richness in bacterial community [99, 100].

I also compared OTU numbers generated from clustering of various qualities of 16S reads with a 96% and a 97% pair-wise identity cutoff. For this comparison, we made and used three datasets: only primer check-passed reads having the highest error rates, filter-passed reads, and selected filter-passed reads having the lowest error rates. The results indicated that a 96% cutoff clustering of error-rich reads and a 97% cutoff clustering of filter-passed reads gave the worse results than a 96% cutoff clustering of filter-passed and selected filter-passed reads (Figure 4.3). A 97% cutoff was defined for clustering of highly accurate Sanger full-length 16S sequences [52]. Therefore, in clustering of pyrosequencing data having higher error rate than Sanger data, the use of a cutoff identity lower than 97% and a lower number of reads are reasonable to reduce overestimation of the OTU number. A 96% cutoff clustering of filter-passed reads gave similar OTU numbers up to 30–50 reads to those of filter-passed reads having the lowest error rates. These read numbers are approximately three to five times the number of input strains. After several trials testing the mock communities, I decided to use 3,000–5,000 reads per sample for clustering with a 96% cutoff for the analysis of human gut

microbiota. Indeed, OTU numbers using a 96% cutoff clustering of 3,000 reads decreased about 15% when compared with those using a 97% cutoff clustering.

Table 4.4 Error rate of filter-passed 16S V1-V2 sequences produced by 454 pyrosequencing of mock communities*

Sample	Total aligned length in bases*	Mismatch	Indel	Total error	Error/read	Average read length
Mock 1	340,878	1,475 0.43%	508 0.15%	1,983 0.58%	1.98	340.9 bases
Mock 2	346,702	1,780 0.51%	505 0.15%	2,285 0.66%	2.29	346.7 bases

*One thousand reads were used for calculation of the error rate for each mock community, respectively.

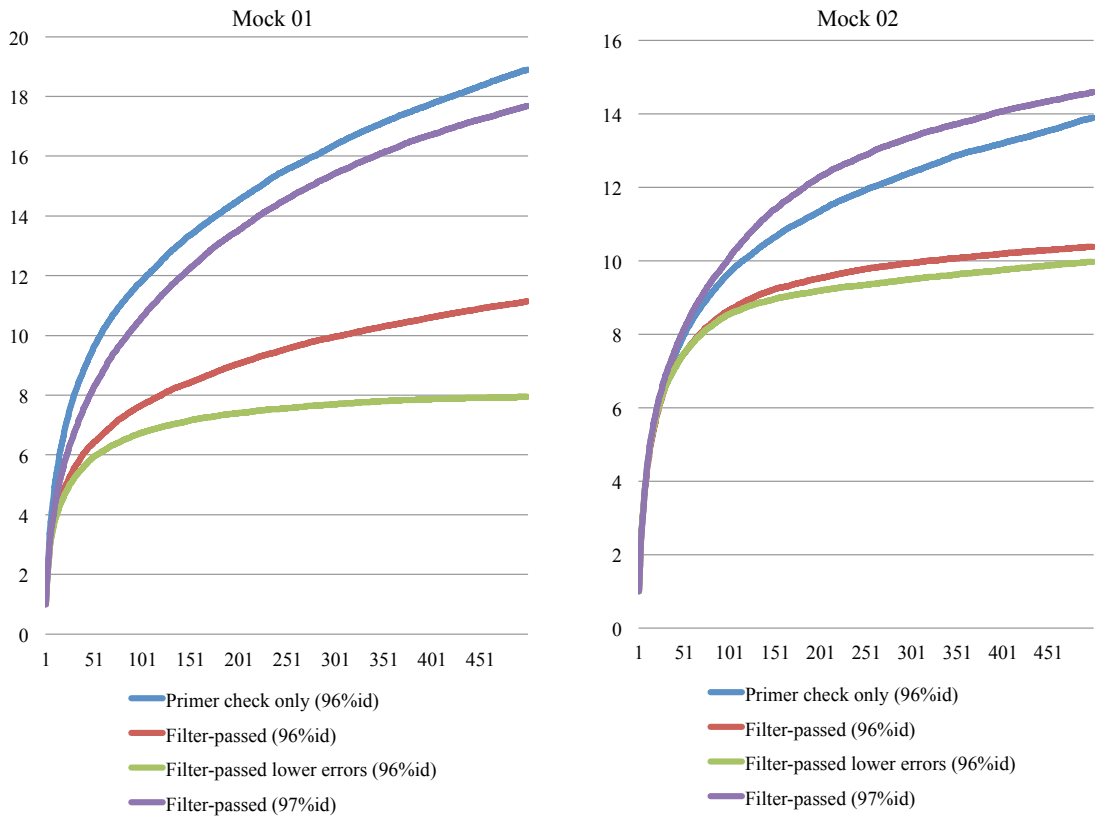

Figure 4.3 Rarefaction curve of OTUs generated by clustering of various qualities of sequences reads obtained from 454 pyrosequencing of mock communities.
The y-axis indicates OTU number generated in clustering. The x-axis indicates read number of 16S sequences. Blue indicates only primer check reads having the highest error rate, red indicates filter-passed reads, and green indicates filter-passed reads having the lowest error rate of clustering with a 96% pair-wise identity cutoff, and purple indicates filter-passed reads of clustering with a 97% pair-wise identity cutoff. The followings are the sequence accuracy and standard deviation of each dataset. For mock community 1, only primer check: 99.1% (SD: 0.96), filter-passed: 99.3% (SD: 0.68), filter-passed having the lowest error rate: 99.5% (SD: 0.26). For mock community 2, primer check only: 99.2% (SD: 0.96), filter-passed: 99.3% (SD: 0.96), filter-passed having the lowest error rate: 99.5% (SD: 0.27).

5 Robustness of gut microbiota of healthy adults in response to probiotic intervention

5.1 Background

Probiotics are defined as live bacterial strains conferring various benefits to the consumer by modulating the intestinal ecosystem, thereby potentially promoting host health and improving host disease risk [101-111]. Various probiotic strains have been industrially developed, and marketed as a variety of products and applications such as fermented foods and supplements including yogurt [112-115]. Most of probiotics taxonomically belong to two genera, *Bifidobacterium* and *Lactobacillus* originated from various environments including the human intestine, and are considered species generally regarded as safe [116-118].

The interaction between administrated probiotics and indigenous microbiota is one of the most attractive and important research areas, particularly because gut microbiota has been shown to be profoundly associated with various host physiology states including diseases, diet, and age through the shift of bacteria composition, and metabolic and nutritional processes [119-123]. The ability of probiotics to survive through the intestine and to modulate gut microbiota is a critical factor in determining their potential for health-related outcomes.

There have been a large number of probiotic intervention studies to assess the impact of probiotics on gut microbiota in healthy adults [124-134], infants/children [135, 136], and its clinical trial to patients with a variety of diseases [137, 138]. Most of these studies were carried out by comparison between probiotics-treated groups and placebo controls and examined only one or two samples from periods before and during intervention or post-intervention for the subject. These experimental designs make the evaluation of results obscure from a viewpoint of statistics and the high interindividual variability of gut microbiota [104].

Recently, a high-throughput sequencing-based analysis was conducted for gut microbiota fed with a probiotic yogurt, and provided new insights into the probiotics research by utilizing the large-scale dataset [139]. Current advance of sequencing technology has enabled us to elucidate complex bacterial community including human gut microbiota [4, 140]. Particularly, 454 pyrosequencing of bacterial 16S gene tags coupled with bioinformatics provides a high-throughput and cost-effective approach for the comprehensive analysis of bacterial community at species level [81-84, 97, 141, 142].

In this study, I developed an analysis pipeline for bacterial communities based on barcoded 454 pyrosequencing of 16S gene tags using modified PCR primers that was described in chapter 4, which improved the quantitative accuracy of inferred species composition in human gut microbiota. Using this pipeline, I analyzed fecal samples longitudinally collected from individuals with and without probiotic administration to evaluate the effect of probiotics on gut microbiota with respect to species richness and diversity. The results revealed that the robustness and stability of gut microbiota of healthy adults in response to probiotic administration.

5.2 Methods

5.2.1 Subjects, fecal sample collection, and probiotic intervention

Eighteen healthy volunteers (Age: 22 ± 3.16 years, 6 male, 12 female) were recruited through Azabu University, Kanagawa, Japan (Table 5.1). All subjects were informed of the purpose of this study. This study was approved by the ethical committee of Azabu University and written consent was obtained from all subjects. No subjects were treated with antibiotics during fecal sample collection. The subjects were divided into six groups (three subjects per group) and each group consumed six different commercially available probiotics supplied from Yakult Honsha Co., Ltd., Kagome Co., Ltd., Morinaga Milk Industory Co., Ltd., Takanashi Milk Products Co., Ltd., Meiji Co., Ltd., and Danone Japan Co., Ltd., respectively (Table 5.1). The

number of each bacterial strain contained in the probiotic products was estimated as the genome equivalent by quantitative PCR of 16S ribosomal RNA genes using 27Fmod-338R followed by pyrosequencing of the 16S amplicons. The genome equivalent per gram or ml and the total genome equivalent of each bacterial strain in one probiotic product was summarized in Table 5.1. Three subjects in each group consumed the same probiotics daily for eight weeks according to the schedule of sampling and probiotic intervention (Figure 5.1). Fecal samples from four weeks before (S00) and eight weeks during probiotic intervention (S01-S04), and eight weeks after cessation of probiotic intervention (S05-S08) were collected every two weeks from each subject. In total, 158 fecal samples from the 18 subjects were collected because one sample each from four of the subjects could not be collect.

Table 5.1 Subjects analyzed and probiotics used in this study

Subject (APr)	Sex	Age	Blood type	BMI	Group	Probiotic strains	Genome equivalent/g or ml[1]	Total genome equivalent/product[2]
1	F	21	B	18.8				
2	F	23	O	18.6	I	*Lactobacillus casei*∗	7.28E+06	4.73E+08
3	F	21	B	19.9				
9	F	20	B	18.5				
11	F	23	A	19.1	II	*Lactobacillus brevis*∗	1.54E+07	1.23E+09
12	M	25	O	19.5				
16	F	20	B	18		*Bifidobacterium longum**	2.63E+07	2.37E+09
17	F	21	O	20.1	III	*Lactococcus lactis*	2.90E+07	2.61E+09
19	M	20	A	19.4		*Streptococcus thermophilus*	6.42E+07	5.78E+09
20	F	21	B	19.8				
22	F	22	A	19.5	IV	*Lactobacillus rhamnosus**	2.01E+07	2.01E+09
23	F	21	B	20.5				
29	F	23	O	20.9		*Lactobacillus delbrueckii**	6.31E+07	5.68E+09
31	M	33	AB	28	V	*Streptococcus thermophilus*	1.51E+08	1.36E+10
32	M	19	AB	21.8				
37	F	21	A	21.2		*Bifidobacterium animals**	1.09E+08	8.72E+09
39	M	23	O	20	VI	*Lactobacillus delbrueckii*	2.40E+07	1.92E+09
40	M	19	B	20.1		*Streptococcus thermophilus*	5.81E+07	4.64E+09

Asterisks indicate probiotic strains contained in commercially available probiotics products
[1]Genome equivalents were average values estimated from triplicate qPCR experiments of 16S genes for each probiotic product.
[2]Total genome equivalents were calculated from genome equivalent/g or ml x total weight or volume of each product.

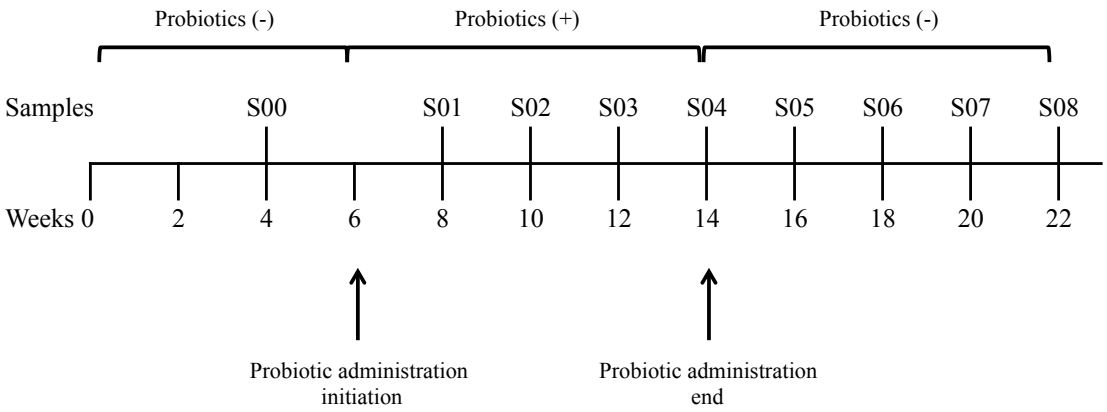

Figure 5.1 Schedule for sampling and probiotic intervention.
Volunteers started at week 0 and first fecal smaple was collected. From week 6 until week 14 (until collecting sample), volunteers took each probiotics product. During 8 weeks, samples were collected for 4 times at intervals of 2 weeks. After 4 times samples were collected, they stopped to have probiotics. Then samples were collected for 4 times at intervals of 2 weeks. Totally 158 samples were collected.

5.2.2 Recovery of bacteria from fecal samples

Freshly collected feces (1.0 g) were suspended in 20% glycerol (Wako Pure Chemical Industries, Ltd.) and phosphate buffer saline (PBS) solution (Life Technologies Japan, Ltd., Tokyo), frozen in liquid nitrogen, and stored at -80 °C until ready for use. Bacterial pellets were prepared from frozen fecal samples as described previously [143].

5.2.3 DNA isolation from bacteria

Fecal DNA was isolated and purified according to the literature, with minor modifications [143]. The bacterial pellet was suspended and incubated with 15 mg/mL lysozyme (Sigma-Aldrich Co. LCC.) at 37 °C for 1 h in TE10. Purified achromopeptidase (Wako Pure Chemical Industries, Ltd.) was added at a final concentration of 2,000 units/mL and then incubated at 37 °C for 30 min. The suspension was treated with 1% (wt/vol) sodium dodecyl sulfate (SDS) and 1 mg/mL proteinase K (Merck Japan) and incubated at 55 °C for 1 h. The lysate was treated with phenol/chloroform/isoamyl alcohol (Life Technologies Japan, Ltd.). DNA was precipitated by adding ethanol, and pelleted by centrifugation at 5,000 rpm at 4 °C for 15 min. The DNA pellet was rinsed with 75% ethanol, dried and dissolved in TE. DNA

samples were purified by treating with 1 mg/mL RNase A (Wako Pure Chemical Industries, Ltd.) at 37 °C for 30 min and precipitated by adding equal volumes of 20% PEG solution (PEG6000-2.5M NaCl). DNA was pelleted by centrifugation at 15,000 rpm at 4 °C, rinsed with 75% ethanol and dissolved in TE.

5.2.4 454 barcoded pyrosequencing of 16S rRNA gene V1-V2 region

The pyrosequencing method was based on the protocol described in chapter 4. In addition, in this study, barcoded-sequencing method was applied for the effectiveness. Using this barcode sequencing, different samples over 100 can be performed in one plate at the same time. The V1-V2 region in the 16S rRNA gene was amplified using forward primer like below.

5'-CCATCTCATCCCTGCGTGTCTCCGACTCAGNNNNNNNNNNagrgtttgatymtggctcag-3'

It contained the 454 primer A, a unique 10-bp barcode sequence for each sample (indicated in N), and 27Fmod (5'-agrgtttgatymtggctcag-3') in which the third base A in original primer 27F was changed to R. And, reverse primer was designed like below.

5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGtgctgcctcccgtaggagt-3'

It contained the 454 primer B and reverse primer 338R (5'-tgctgcctcccgtaggagt-3'). A mixed sample was prepared by pooling approximately equal amounts of PCR amplicons from each sample and subjected to 454 GS FLX Titanium or 454 GS JUNIOR (Roche Applied Science) according to the manufacturer's instructions.

5.2.5 Analysis pipeline for 454 barcoded pyrosequencing data of PCR amplicons of 16S rRNA gene V1-V2 region

I developed the analysis pipeline for 454 barcoded pyrosequencing of PCR amplicons of the V1-V2 region amplified by 27Fmod-338R primers that was described in chapter 4. First, 16S reads were assigned to each sample on the basis of barcode sequence information. Using this data filtering process described in chapter 4 was performed. Reads removed in these

processes accounted for about 35% of all reads, most of which represented reads lacking PCR primer sequences (Table 5.2). Finally, filter-passed reads were obtained for further analysis by trimming both primer sequences.

All 3,000 filter-passed reads of 16S V1-2 sequences obtained from each subject were deposited in DDBJ/GenBank/EMBL and the accession numbers are from DRA000869 to DRA000886.

Table 5.2 Summary of quality filtering of 16S V1-V2 sequences produced by 454 pyrosequencing of 18 healthy adults

	Total	Filter-passed	Reads removed		
			Reads lacking primer sequences	Reads with average Qv<25	Possible chimeric reads
Number of reads	1,801,980	1,165,626	472,570	133,905	29,879
%	100	64.7	26.2	7.4	1.7

5.2.6 Data analysis

For the database preparation, two databases were constructed for the analysis of 16S sequences. Database used mainly was GRD which was made in chapter 3 and highly curated 16S sequences dataset. Another database was the 16S rRNA gene sequence database constructed by collecting 16S sequences of ≥1,200 bp of bacteria isolates in the Ribosomal Database Project (RDP) v. 10.27 [144].

In OTU and UniFrac distance analysis, I used 3,000 filter-passed reads of high-quality 16S sequences for OTU and UniFrac distance analysis of each sample. In OTU analysis, clustering of 16S reads was done by a 96% pairwise-identity cutoff using UCLUST [145]. Representative sequences of each OTU were assigned to bacterial species by BLAST search with a 96% pairwise-identity cutoff against the two databases mentioned above. UniFrac distance analysis was used to determine the dissimilarity (distance) between two communities, based on the fraction of branch length shared between two communities within a phylogenetic tree constructed from 16S sequence datasets [63, 141].

5.3 Results and discussion

5.3.1 Species richness and diversity in human fecal microbiota with probiotic intervention

I randomly selected 3,000 reads of 16S V1-V2 sequences from all filter-passed reads for each sample (Table 5.2), and used 474,000 reads in total from 158 fecal DNA samples of 18 subjects for the analysis of species richness and composition in human gut microbiota. Clustering of all reads with a 96% pairwise-identity cutoff gave a total of 2,758 OTUs. By removing very minority OTUs, 1,175 OTUs having ≥0.1% abundance in at least one sample, accounting for 99.1% of all 16S reads, were used for further analysis.

5.3.1.1 Detection of administrated probiotic strains in fecal sample

I investigated whether administrated strains contained in the probiotic products can be detected in fecal DNA. The 16S V1-2 region of all bacterial strains contained in probiotic products used in this study was performed sequencing. The BLAST search to the databases indicated that except for the *Bifidobacterium longum* strain used in Group III, the 16S sequences of all strains in the probiotic products significantly differed from those of the indigenous species phylogenetically closest to the probiotic strains. The 16S sequence of the *B. longum* strain used in Group III was almost identical to that of an indigenous *Bifidobacterium* species, so that a distinguishable additive *Lactococcus lactis* strain was used in this product for the detection of administrated bacteria in Group III samples. The 16S sequences of these probiotic strains were included in the databases constructed in this study, and the 16S reads assigned to administrated strains had the average similarity between 99.4 and 99.9% identities with the reference sequences (data not shown). The OTUs assigned to the probiotic strains were detected in samples (S01 to S04) during probiotic intervention (designated “Pro(+)”) in various frequencies, however, almost no detected in samples (S00

and S05 to S08) without probiotics administration (designated "Pro(-)") (Table 5.3). Administrated probiotic strains were shown to be more frequently detected in samples during the intervention than in pre- and post-intervention periods by using different detection methods such as culturing, targeted PCR and hybridization[124, 126-128, 130, 132, 133, 146]. In this study, two probiotic *Lactobacillus* and one additive *Lactococcus* strains were detected in post-intervention samples in three subjects with a minimum count, respectively. The similarity of these three 16S sequences was 99.4, 99.7 and 100% identity with those of administrated *Lactobacillus* and *Lactococcus* strains, respectively, indicating that these are administrated strains. The survival of some probiotics in post-intervention period was also reported previously [128, 130]. The data of this study suggested that some probiotic strains seem to be able to persistently colonize the intestine and their survivability may be related to metabolic activity in the intestine [147, 148]. Probiotic *Bifidobacterium* strains were not detected in any Pro(-) samples. However, two distinct 16S sequences both assigned to *B. animals* were found in two subjects APr37 and APr39. One showed a high similarity of >98% identity with the 16S sequence of the administrated *B. animals* and was detected with high frequency only in the Pro(+) samples, while another showed a low similarity of 96.5 to 97.4% identity (a mean of 97.2%) with low frequency in both the Pro(-) and Pro(+) samples. These data suggest the presence of unknown indigenous species phylogenetically close to, but distinct from, probiotic *B. animals* in human gut microbiota. The total number of bacteria contained in each probiotic product was varied between 10^9 and 10^{10}, showing no large difference in quantity among them (Table 5.1). No clear correlation was also observed between the number of bacteria in the products and the frequency in detection of the administrated strains in the Pro(+) samples. From these observations, the frequency of administrated bacteria detected in feces may not be largely affected by their amounts in the products. Therefore, detection of *L. brevis* and *L. delbrueck* at relatively low level in feces cannot be simply explained by the difference in a dose, but could be considered the

association with several factors such as their survivability in the intestine, diet, or physiological conditions of subjects.

Table 5.3 Detection of administrated probiotic strains in fecal sample by OTU assignment of 16S sequence

Group	Subject (APr)	Pro (-) S00	Pro (+) S01	Pro (+) S02	Pro (+) S03	Pro (+) S04	Pro (-) S05	Pro (-) S06	Pro (-) S07	Pro (-) S08	Probiotic strains
Group I	APr01	0	9	25	19	25	0	0	0	0	*Lactobacillus casei*
	APr02	0	17	4	33	15	0	0	0	0	
	APr03	0	3	6	11	15	0	0	0	0	
Group II	APr09	0	0	3	0	3	NA	0	0	0	*Lactobacillus brevis*
	APr11	0	1	1	1	2	0	0	0	0	
	APr12	0	0	2	NA	0	0	0	0	0	
Group III	APr16	0	31	0	10	9	0	0	0	0	*Lactococcus lactis**
	APr17	0	20	12	24	4	0	0	0	0	
	APr19	0	4	7	13	7	1	0	0	0	
Group IV	APr20	0	12	12	23	20	0	0	0	0	*Lactobacillus rhamnosus*
	APr22	0	30	11	16	12	1	0	0	0	
	APr23	0	10	34	14	7	0	0	0	0	
Group V	APr29	0	0	2	0	0	0	0	0	0	*Lactobacillus delbrueckii*
	APr31	0	1	2	7	3	0	0	0	1	
	APr32	0	3	0	6	11	NA	0	0	0	
Group VI	APr37	0	418	629	365	501	NA	0	0	0	*Bifidobacterium animalis*
	APr39	0	142	111	188	52	0	0	0	0	
	APr40	0	64	6	0	0	0	0	0	0	

NA: not analyzed
Numbers indicate frequency of 16S reads contained in OTUs assigned to probiotic strains.
**L. lactis* was used for detection of administrated strains instead of probiotic *Bifidobacterium* almost identical to indigenous species.

5.3.1.2 Change of species richness in samples with and without probiotics

I analyzed species richness (OTU number) in Pro(+) and Pro(-) samples. Figure 5.2 showed the change in OTU numbers for every sample in each subject, indicating that OTU numbers vary dramatically for every sample. Most of the variation can be attributed to single OTUs representing the minority species. I averaged the OTU numbers of the Pro(-) and Pro(+) samples and compared them for subject, group, type of probiotics (*Lactobacillus* and *Bifidobacterium*), and all combined samples, respectively (Figure 5.3). The average OTU numbers in six out of 18 subjects were decreased in the range of the ratio of 0.83 to 0.95 in the Pro(+) samples compared with the Pro(-) samples, while those in other 12 subjects were increased in the range of the ratio of 1.01 to 1.43. For group, only Group IV showed a decrease of the average OTU number in the Pro(+) samples with the ratio of 0.94. For type of

probiotics and all samples, the average OTU numbers in the Pro(+) samples were slightly more abundant (approximately 1.07-fold) than those in the Pro(-) samples, but no statistical significance was observed in any dataset. The increase of OTU number in the Pro(+) samples was largely due to the minority species (Figure 5.2), while the abundance of the majority species (OTUs containing ≥10 reads) was almost constant over time. The same analysis using different sets of 3,000 reads for each subject was also performed. These data argue that administration of probiotics tends to increase species richness in fecal microbiota, which may be beneficial for the consumer because the species richness in fecal microbiota of subjects afflicted with disease such as inflammatory bowl disease is significantly reduced compared with that of healthy subjects [149].

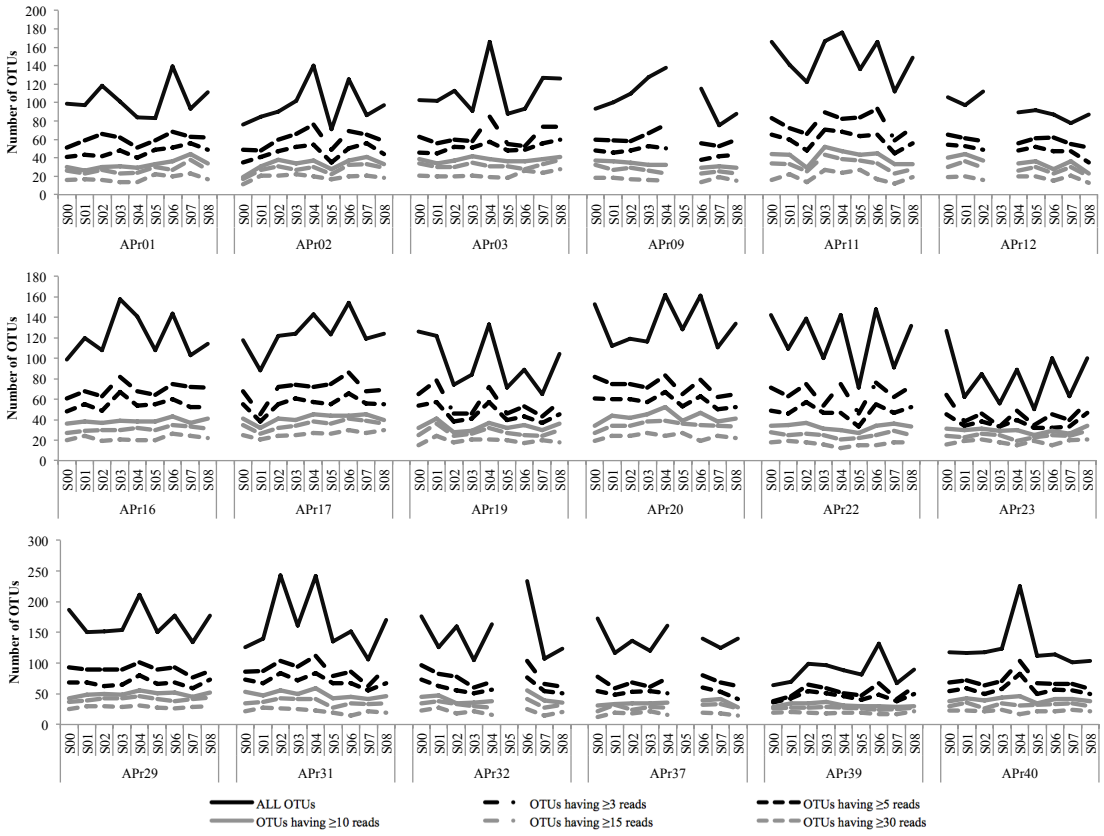

Figure 5.2 Change of OTUs having various numbers of 16S reads for every sample in subjects and the examined period

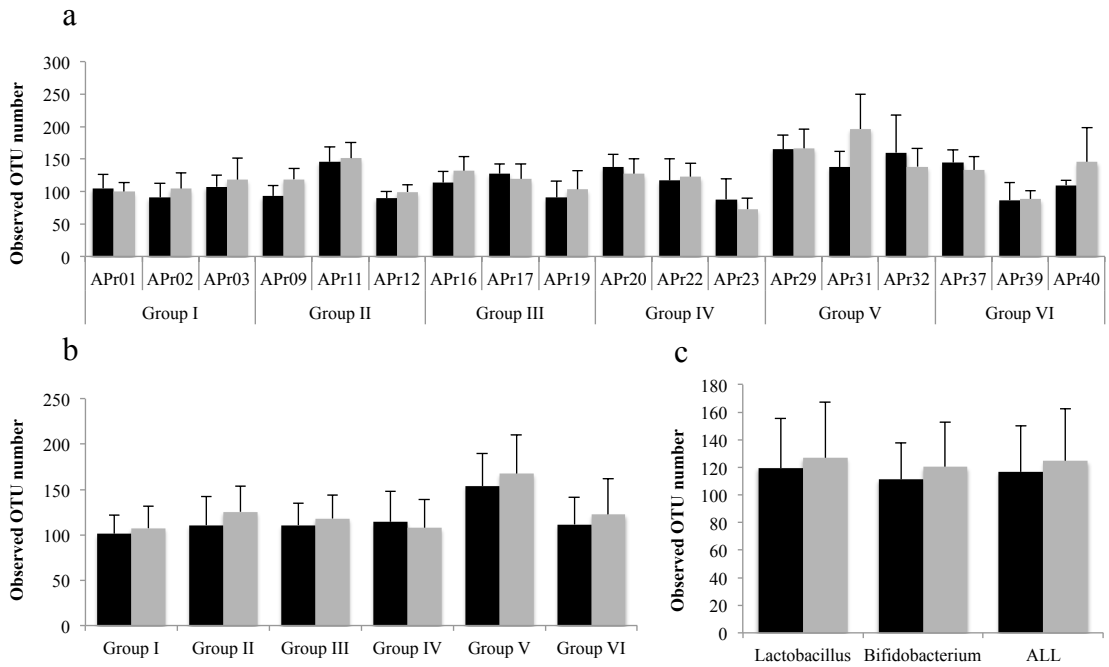

Figure 5.3 Change of OTU number in fecal microbiota with and without probiotics administration.
(a) individual, (b) group, (c) type of probiotics. Black bar indicates Pro(-) samples. Gray bar indicates Pro(+) samples. The error bars represent standard deviation.

5.3.1.3 Change of species composition in samples with and without probiotics

The average weighted and unweighted UniFrac distances within Pro(-) and Pro(+), and between Pro(-) and Pro(+) samples for every group, probiotics types and all subjects, was calculated respectively (Figure 5.4). High UniFrac distance implies the high variability of microbiota structure within and between samples. If difference between any pair of the three distances is statistically significance, it can be considered that probiotics administration significantly affected the overall microbiota composition. It was found that the highest difference in weighted UniFrac distances between Pro(+) and Pro(-) samples in Group VI. However, statistics evaluation of this difference by Student's t test showed no significance (data not shown). These data implied that high stability of gut microbiota to probiotics administration for all subjects examined. The UniFrac distances of intra-subject gut microbiota were also analyzed (Figure 5.5). Although five subjects (APr02, 12, 16, 37 and 39) showed a significant difference in the UniFrac distances between Pro(-) and Pro(+) samples, the results showed that both weighted and unweighted distances between Pro(-) and

Pro(+) of all intra-subjects were significantly lower than the average distance of the 18 unrelated subjects. The Welch's *t* test for these differences showed statistical significance (Table 5.4). Thus, these data suggested that the perturbation of microbiota elicited by probiotics in an intra-subject did not overcome the inter-subject variations of gut microbiota, supporting high intra-specificity and stability of gut microbiota [7, 150]. This robustness of gut microbiota of adults is in contrast with the profound effect of antibiotic administration on adult gut microbiota [151] and the observed response of gut microbiota of infants fed with probiotics, in which the infant gut microbiota composition was considerably affected by probiotics [136]. A short-term dietary intervention study showed that in controlled feeding of the same diet to subjects over 10 days a marked change was observed within 1 day after the intervention initiation [152]. In this study, no significant difference was observed between samples before (S00) and first samples (S01) after the intervention initiation (data not shown). It would be valuable to analyze fecal samples collected within a few days after administration of probiotics for evaluation of the short-term effect of probiotics.

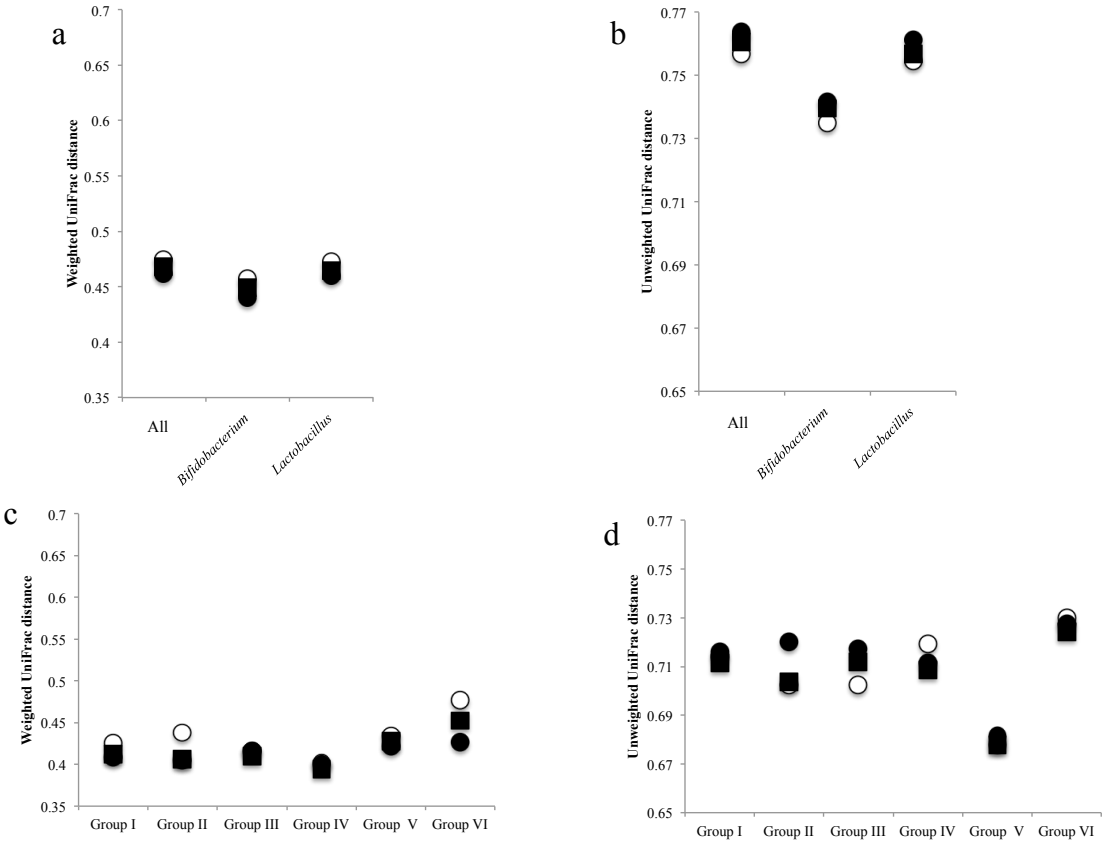

Figure 5.4 Average UniFrac distance within Pro (-) and Pro(+) and between Pro(-) and Pro(+) for each group, type of probiotics and all subject.

Average UniFrac distance between any pair of the three distances for type of probiotics and all subject (a and b), and each group (c and d). Open circle, closed circle, and closed square indicate average UniFrac distance within Pro (-), within Pro (+) and between Pro(-) and Pro(+) samples, respectively.

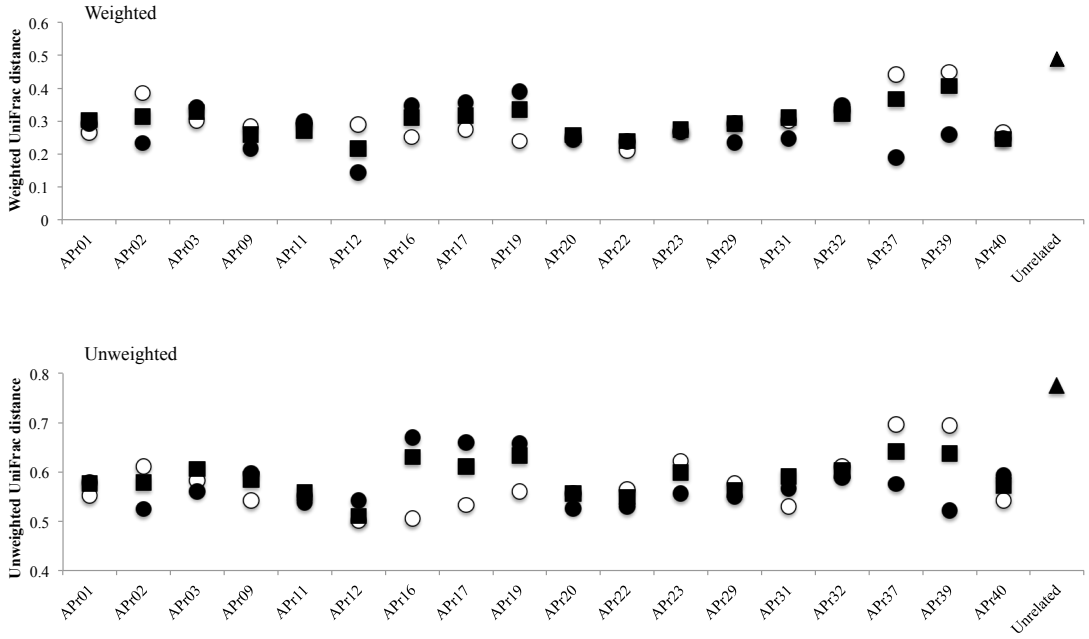

Figure 5.5 Average UniFrac distance within Pro(-) and Pro(+) and between Pro(-) and Pro(+) for each subject.

Open circles, closed circles, and closed squares indicate average UniFrac distance within Pro(-), within Pro(+) and between Pro(-) and Pro(+) samples, respectively. Closed triangles indicate average UniFrac distance between samples (S00) of 18 unrelated individuals.

Table 5.4 Statistical evaluation of differences in UniFrac distances between each intra-subject and 18 unrelated subjects

Weighted UniFrac		Unweighted UniFrac	
Subject	*P*-value	Subject	*P*-value
APr01	1.1273E-09	APr01	1.9660E-12
APr02	1.4501E-06	APr02	5.8962E-12
APr03	3.3515E-08	APr03	5.0581E-16
APr09	6.9384E-11	APr09	1.8923E-10
APr11	3.8418E-11	APr11	2.2021E-25
APr12	4.3680E-12	APr12	5.5241E-14
APr20	2.7362E-21	APr20	8.2730E-18
APr22	1.2724E-18	APr22	1.6304E-17
APr23	9.4860E-12	APr23	8.8633E-10
APr29	2.8113E-13	APr29	1.4073E-16
APr31	1.8091E-12	APr31	1.3290E-11
APr32	3.0355E-06	APr32	5.6107E-10
APr16	3.3705E-14	APr16	3.6037E-08
APr17	1.5936E-05	APr17	6.9927E-06
APr19	2.7300E-04	APr19	8.5021E-06
APr37	4.0409E-04	APr37	3.8670E-05
APr39	8.0596E-04	APr39	1.7585E-06
APr40	1.2277E-19	APr40	8.4807E-14

The Welch's *t* test was used to calculate the *P*-values

5.3.2 Identification of bacterial species having significant increase and decrease by probiotics administration

Although these results suggested that administration of probiotics had almost no effect on the overall structure of gut microbiota, it would be possible to identify bacterial species largely responding to the administrated probiotics at OTU/species level. Therefore, OTUs having an increase and a decrease between Pro(+) and Pro(-) samples by comparing the number of 16S reads in each OTU were surveyed. I first enumerated the OTUs showing ≥2-fold change between the Pro(-) and Pro(+) samples for each subject, and the quantity difference was also obtained by subtracting the 16S read number of the Pro(+) samples by that of the Pro(-) samples. This is because OTUs showing a high quantity difference but less fold change may also have substantial influence on gut microbiota composition. It was found that several OTUs significantly changed by probiotics administration, including OTUs assigned to both indigenous and the administrated strains (Figure 5.6). I listed 88 OTUs (7.5% of a total of 1,175 OTUs) having significant change of ≥3-fold, among which 30 OTUs changed with ≥10-fold (Figure 5.7). I excluded six OTUs assigned to the administrated strains from the 30 OTUs and obtained 25 OTUs assigned to the indigenous species including OTU00072 assigned to *Streptococcus salivarius* that showed significant change in two subjects (Table 5.5). I also found seven OTUs having significant difference in quantity between the both samples (Table 5.6). Of the combined 32 OTUs (2.7%), 18 were increased and 14 were decreased by probiotics administration. Many of the OTUs showing a significant increase were assigned to minority species in the Pro(-) samples, but some increased up nearly 7% in abundance (*e.g.*, OTU00372 assigned to *Eubacterium rectale*). On the other hand, the OTUs showing a significant decrease were almost undetected in the Pro(+) samples. Phylum-level species assignment showed that species belonging to the phylum *Firmicutes* were most largely affected by both probiotics, and all species belonging to the phylum *Bacteroidetes* were affected only by *Lactobacillus* probiotics (Table 5.7). The 32 OTUs were assigned to 27

indigenous species, among which four species (*Clostridium clostridioforme*, *Eubacterium eligens*, *Eubacterium rectale*, and *Faecalibacterium prausnitzii*) were assigned by eight different OTUs and one species (*S. salivarius*) was assigned by the two same OTUs as described above. These data suggested that two closely related species each other are contained in the four bacterial species. All of these species except for *S. salivarius* were found to show significant change only in one subject, indicating that response of the indigenous species to probiotics is highly individual-specific (Figure 5.8). Two different OTUs (OTU02677 and OTU02748) assigned to *F. prausnitzii*, of which the reduction is known to be correlated with inflammatory bowl disease [153], were found to both decrease and increase in the same subject (APr40) by probiotic administration, suggesting that these two phylogenetically close species may have the diversity of response to probiotic action. The distribution of the 32 OTUs in the subjects was also examined. The results revealed that four subjects (APr11, 20, 22, and 37) did not have such OTUs and eight subjects had only one OTU, while four subjects (APr01, 02, 31 and 39) had more than four OTUs showing significant change (Figure 5.9), suggesting their uneven distribution in the 18 subjects. These data imply existence of the sensitive and less sensitive responders to probiotic action and if so, it would be interesting to investigate the relation between gut microbiota type and its response to probiotics.

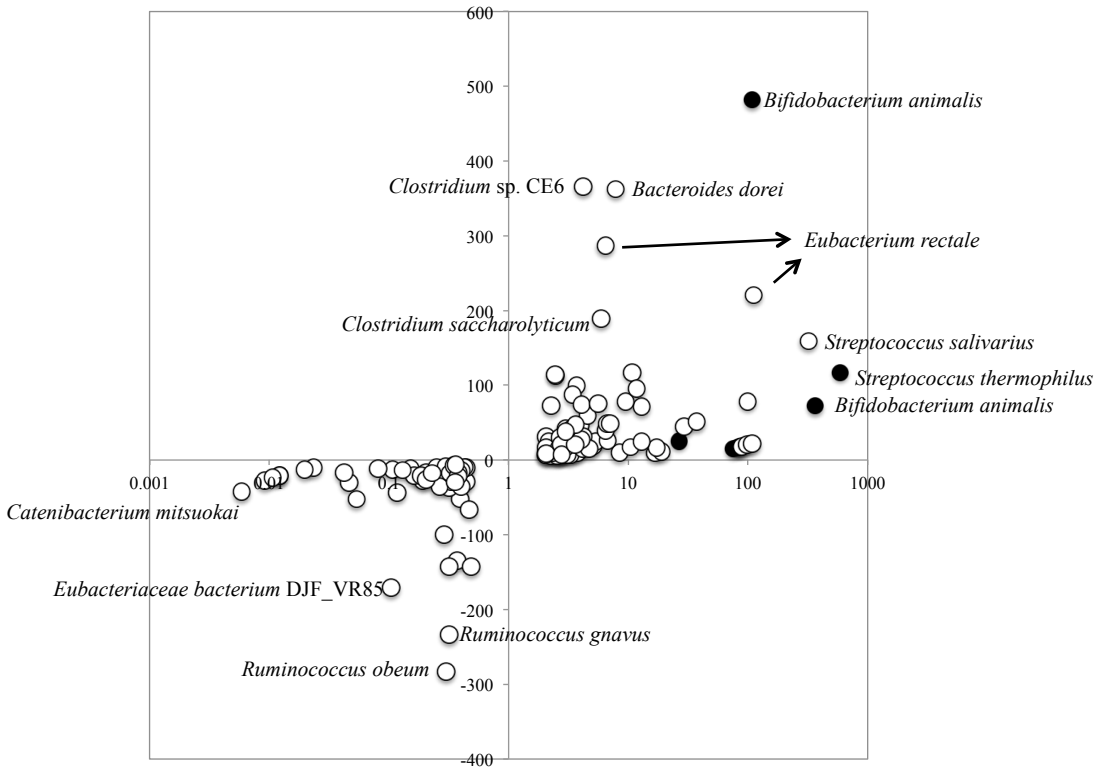

Figure 5.6 OTUs having change of ≥2-fold and their difference in quantity between Pro(-) and Pro(+) samples.

The x-axis represents the difference (read number) in quantity between Pro(+) and Pro(-) samples. The y-axis represents the scale of fold change between Pro(+) and Pro(-) samples. Closed and open circles indicate each administrated probiotic and indigenous species.

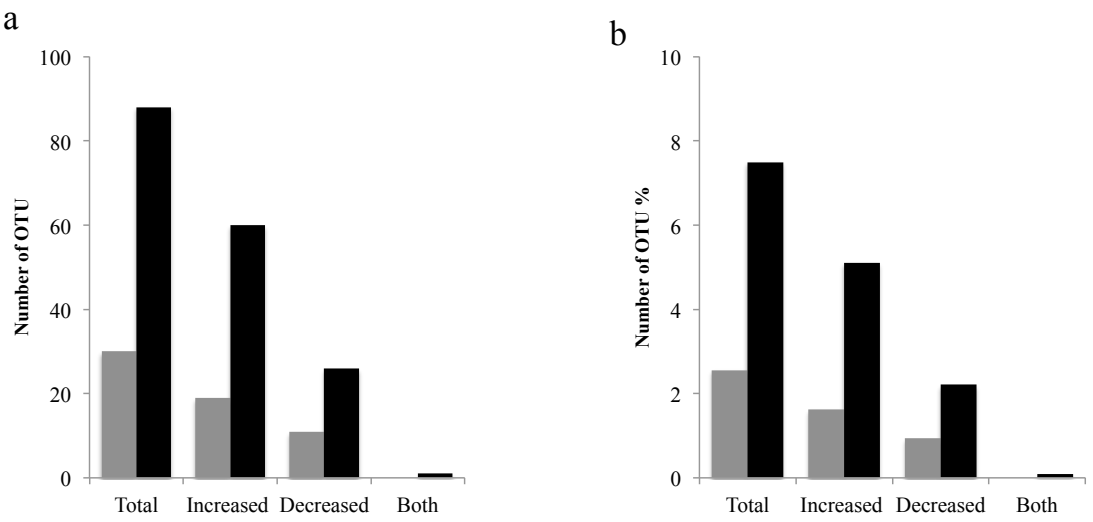

Figure 5.7 Number and ratio of OTUs having significant change by probiotics administration.

Numbers of OTUs having change of ≥3-fold (Black bars) and ≥10-fold (Gray bars) by probiotics administration were shown. Statistical significance was determined using the Welch's *t* test with $P<0.05$. Numbers indicate the number of OTUs (a) and the ratio % for total OTU number (b). Both indicate OTUs showing both increase and decrease.

Table 5.5 OTUs and assigned indigenous species having significant change of ≥10-fold by probiotics administration

OTU ID	Subject	Group	Average read # Pro(+)	SD Pro(+)	Average read # Pro(-)	SD Pro(-)	Fold	Increase/ Decrease	*P*-value	Best-hit species	Identity in %
OTU01567	APr31	V	21.75	13.02	0.00	0.00	-	Increase	0.022	*rumen bacterium* NK3B98	88.24
OTU00072*	APr32	V	159.25	93.32	0.50	1.00	318.50	Increase	0.021	*Streptococcus salivarius*	100
OTU02549	APr39	VI	10.00	4.69	0.60	1.34	16.67	Increase	0.012	*Erysipelotrichaceae bacterium* 5_2_54FAA	93.48
OTU00374	APr02	I	11.50	7.85	0.60	1.34	19.17	Increase	0.034	*Eubacterium eligens*	100
OTU00304	APr02	I	79.50	49.98	0.80	1.30	99.38	Increase	0.026	*Lachnospiraceae bacterium* 5_1_63FAA	98.78
OTU02677	APr40	VI	17.25	8.34	1.00	1.73	17.25	Increase	0.014	*Faecalibacterium prausnitzii*	96.20
OTU00378	APr02	I	52.00	20.35	1.40	3.13	37.14	Increase	0.007	*Eubacterium eligens*	98.78
OTU00679	APr23	IV	46.75	37.16	1.60	3.58	29.22	Increase	0.046	*Bacteroides uniformis*	97.77
OTU00380	APr31	V	18.75	12.79	1.80	2.49	10.42	Increase	0.037	*Clostridia bacterium* S130(2)-2	92.35
OTU01386	APr39	VI	26.00	16.57	2.00	1.22	13.00	Increase	0.031	*Firmicutes bacterium* DJF_VR50	96.93
OTU00372	APr02	I	222.50	92.41	2.00	2.45	111.25	Increase	0.009	*Eubacterium rectale*	96.32
OTU00384	APr39	VI	77.25	27.87	6.00	8.40	12.88	Increase	0.006	*Clostridium clostridioforme*	96.01
OTU00072*	APr31	V	105.00	68.17	9.00	5.61	11.67	Increase	0.033	*Streptococcus salivarius*	100
OTU00854	APr39	VI	129.50	67.61	12.00	10.61	10.79	Increase	0.019	*Clostridium clostridioforme*	96.09
OTU00288	APr01	I	0	-	10.60	11.04	-	Decrease	0.049	*Parabacteroides* sp. D13	99.19
OTU00176	APr01	I	1.00	1.41	12.40	8.85	0.08	Decrease	0.022	*Coprobacillus cateniformis*	99.67
OTU01283	APr16	III	0.25	0.50	12.60	9.66	0.02	Decrease	0.023	*Ruminococcus* sp. DJF_VR52	96.60
OTU01990	APr29	V	0.75	1.50	17.80	14.13	0.04	Decrease	0.027	*Mitsuokella multacida*	94.89
OTU01404	APr17	III	0	-	20.40	21.05	-	Decrease	0.048	*butyrate-producing bacterium* SM6/1	95.40
OTU00257	APr01	I	0	-	20.60	15.19	-	Decrease	0.019	*Bacteroides* sp. 20_3	94.64
OTU02748	APr40	VI	0	-	23.20	21.84	-	Decrease	0.038	*Faecalibacterium prausnitzii*	99.35
OTU00331	APr03	I	0	-	27.40	26.66	-	Decrease	0.042	*Bifidobacterium adolescentis*	97.42
OTU00605	APr19	III	1.50	1.29	32.20	26.96	0.05	Decrease	0.032	*Lachnospiraceae bacterium* 4_1_37FAA	99.08
OTU02241	APr31	V	0	-	42.20	39.26	-	Decrease	0.037	*Catenibacterium mitsuokai*	94.44
OTU01169	APr12	II	3.00	5.20	55.60	46.22	0.05	Decrease	0.032	*Parabacteroides merdae*	95.89

*OTU00072 assigned to *Streptococcus salivarius* showed a significant increase of ≥10-fold in two subjects.

Table 5.6 OTUs and assigned indigenous species having significant quantity difference in quantity by probiotics administration

OTU ID	Subject	Group	Average read # Pro(+)	SD Pro(+)	Average read # Pro(-)	SD Pro(-)	Fold change	Difference between Pro(+) and Pro(-)	*P*-value	Best-hit species	Identity in %
OTU00144	APr09	II	414.75	119.33	52.75	104.17	7.9	362	0.002	*Bacteroides dorei*	99.36
OTU00241	APr02	I	339.75	218.4	52.8	50.08	6.4	286.95	0.038	*Eubacterium rectale*	99.39
OTU00015	APr39	VI	227	87.2	38.4	36.27	5.9	188.6	0.008	*Clostridium saccharolyticum*	91.72
OTU00049	APr01	I	480	245.51	114.6	67.59	4.2	365.4	0.027	*Clostridium* sp. CE6	99.04
OTU00022	APr39	VI	108.75	52.56	342.2	243.71	0.3	-233.45	0.049	*Ruminococcus gnavus*	99.08
OTU00393	APr19	III	122	98.04	405.6	209.96	0.3	-283.6	0.019	*Ruminococcus obeum*	95.98
OTU02208	APr31	V	20	40	190.2	173.64	0.1	-170.2	0.047	*Eubacteriaceae bacterium* DJF_VR85	96.17

OTUs having quantity difference of ≥150 reads and fold change of <10 between Pro(-) and Pro(+) samples were listed.

Table 5.7 Phylum-level species assignment of OTUs having significant fold change or quantity difference by administration of probiotics

Type of Probiotics	Change	*Number of varied OTUs	Fold change (≥10-fold)				Difference (≥150 reads)		
			Firmicutes	*Actinobacteria*	*Bacteroidetes*	Unclassified bacterium	Number of varied OTUs	*Firmicutes*	*Bacteroidetes*
Lactobacillus	**Increase**	9	7	0	1	1	3	2	1
	Decrease	7	3	1	3	0	1	1	0
	Total	16	10	1	4	1	4	3	1
Bifidobacterium	**Increase**	5	5	0	0	0	1	1	0
	Decrease	4	4	0	0	0	2	2	0
	Total	9	9	0	0	0	3	3	0
All	**Increase**	14	12	0	1	1	4	3	1
	Decrease	11	7	1	3	0	3	3	0
	Total	25	19	1	4	1	7	6	1

*Administrated probiotic strains were excluded, and only OTUs of *P*-value <0.05 were shown.

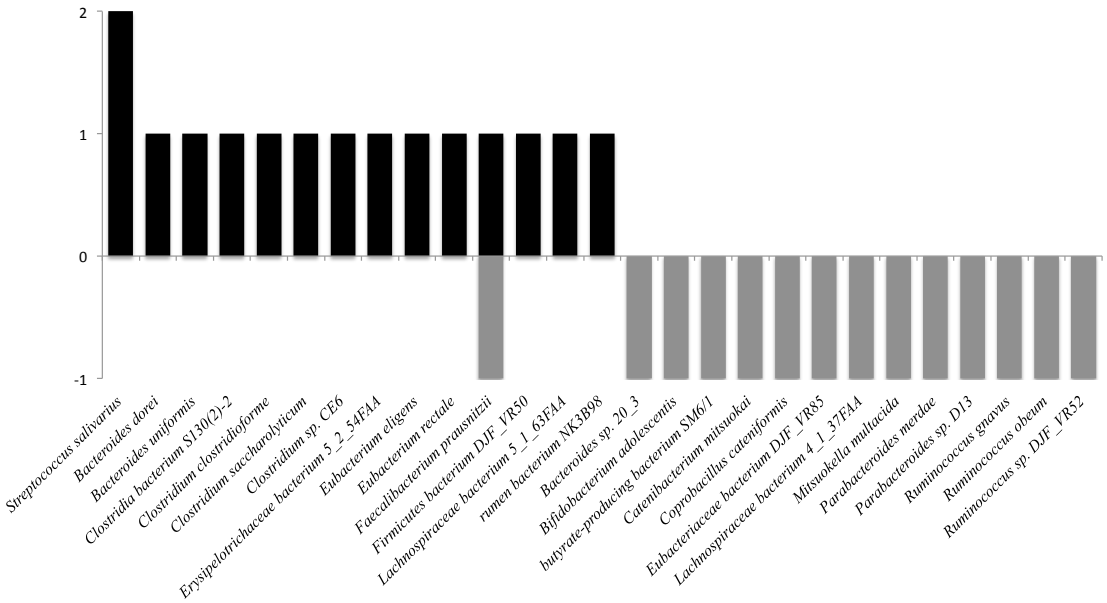

Figure 5.8 Distribution of bacterial species having significant change in subjects.
The y-axis indicates the number of subjects to which 27 bacterial species having significant change by probiotics administration were assigned.

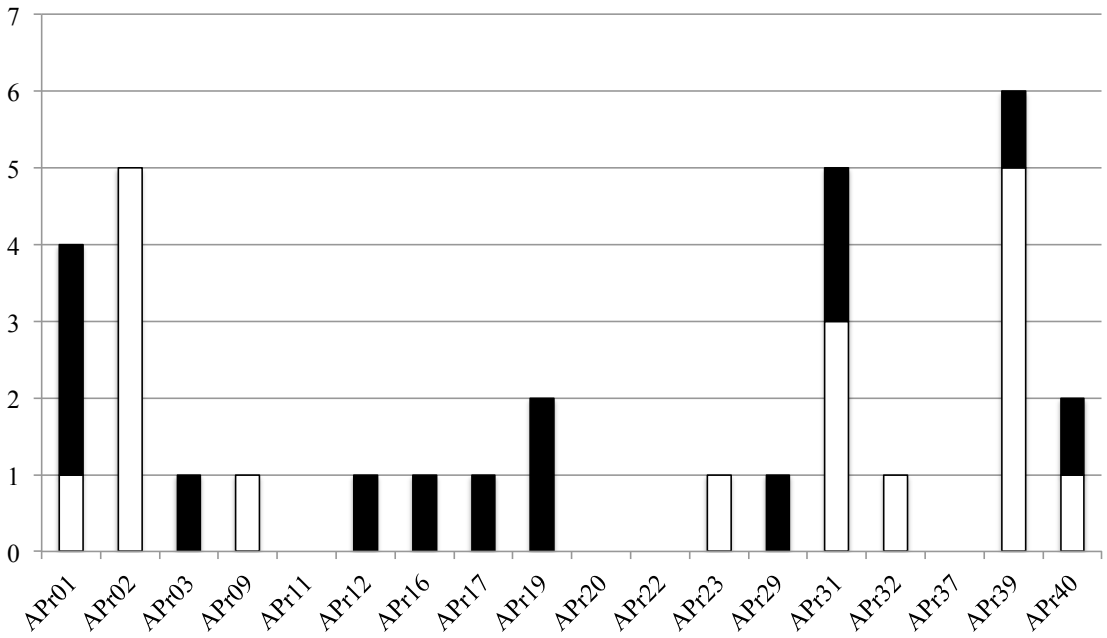

Figure 5.9 Distribution of 32 OTUs having significant change in every subject.
The y-axis indicates the number of OTUs having significant change between Pro(-) and Pro(+) samples in each subject. Open and closed bars indicate increased and decreased OTUs, respectively.

6 Conclusion

I conducted researches for this dissertation along the three important keywords, which is 16S rRNA gene, next-generation sequencing technologies (NGS), and human gut microbiota. Bacterial 16S rRNA gene is important for the analysis of bacterial community. The analysis of human gut microbiota using NGS provided us the high-throughput DNA-based approach to comprehensively evaluate the overall structure and its change.

In chapter 3, I constructed genomic-based 16S rRNA gene database called GRD, which is useful for 16S-based taxonomic analysis of bacterial community. GRD was made by the unique method that was extraction of 16S rRNA genes from genome sequences directly. I revealed that the quality of annotation by GRD was higher than that by GenBank, and GRD was composed of high-quality 16S rRNA genes as compared with other three publically available databases. Furthermore, I found boundary identity of 16S rRNA genes between each taxonomic level, which is useful and reliable for 16S-based taxonomical assignment of species.

In chapter 4, I developed improved universal primer 27Fmod, which provided higher quantitative accuracy than the conventional primer 27F for the analysis of the bacterial composition in human gut microbiota.

In chapter 5, I analyzed change of gut microbiota of healthy adults fed with probiotics by using the improved 454 pyrosequencing platform with the improved quantitative accuracy for evaluation of the overall bacterial composition. This study using large datasets enabled us to evaluate the effect of probiotics on gut microbiota of healthy adults more comprehensively and precisely than the previous probiotic intervention researches in which the analysis exclusively focused on only several limited bacterial species by using conventional methods. This data further supports the high inter-subject variability and the high intra-subject stability that is the current common view for the feature of adult gut microbiota.

References

1. Savage DC: **Microbial ecology of the gastrointestinal tract.** *Annu Rev Microbiol* 1977, **31:**107-133.
2. Human Microbiome Project C: **Structure, function and diversity of the healthy human microbiome.** *Nature* 2012, **486:**207-214.
3. Human Microbiome Project C: **A framework for human microbiome research.** *Nature* 2012, **486:**215-221.
4. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al: **A human gut microbial gene catalogue established by metagenomic sequencing.** *Nature* 2010, **464:**59-65.
5. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome.** *Science* 2006, **312:**1355-1359.
6. O'Hara AM, Shanahan F: **The gut flora as a forgotten organ.** *EMBO Rep* 2006, **7:**688-693.
7. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, et al: **A core gut microbiome in obese and lean twins.** *Nature* 2009, **457:**480-484.
8. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI: **Obesity alters gut microbial ecology.** *Proc Natl Acad Sci U S A* 2005, **102:**11070-11075.
9. Sartor RB: **Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis.** *Nat Clin Pract Gastroenterol Hepatol* 2006, **3:**390-407.
10. Spor A, Koren O, Ley R: **Unravelling the effects of the environment and host genotype on the gut microbiome.** *Nat Rev Microbiol* 2011, **9:**279-290.
11. Dicksved J, Halfvarson J, Rosenquist M, Jarnerot G, Tysk C, Apajalahti J, Engstrand L, Jansson JK: **Molecular analysis of the gut microbiota of identical twins with Crohn's disease.** *ISME J* 2008, **2:**716-727.
12. Perry S, de Jong BC, Solnick JV, de la Luz Sanchez M, Yang S, Lin PL, Hansen LM, Talat N, Hill PC, Hussain R, et al: **Infection with *Helicobacter pylori* is associated with protection against tuberculosis.** *PLoS One* 2010, **5:**e8804.
13. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.** *J Mol Biol* 1975, **94:**441-448.
14. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors.** *Proc Natl Acad Sci U S A* 1977, **74:**5463-5467.

15. Maxam AM, Gilbert W: **A new method for sequencing DNA.** *Proc Natl Acad Sci U S A* 1977, **74:**560-564.
16. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ, et al: **The DNA sequence of human chromosome 22.** *Nature* 1999, **402:**489-495.
17. Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, et al: **The DNA sequence of human chromosome 21.** *Nature* 2000, **405:**311-319.
18. Deloukas P, Matthews LH, Ashurst J, Burton J, Gilbert JG, Jones M, Stavrides G, Almeida JP, Babbage AK, Bagguley CL, et al: **The DNA sequence and comparative analysis of human chromosome 20.** *Nature* 2001, **414:**865-871.
19. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
20. Heilig R, Eckenberg R, Petit JL, Fonknechten N, Da Silva C, Cattolico L, Levy M, Barbe V, de Berardinis V, Ureta-Vidal A, et al: **The DNA sequence and analysis of human chromosome 14.** *Nature* 2003, **421:**601-607.
21. Hillier LW, Fulton RS, Fulton LA, Graves TA, Pepin KH, Wagner-McPherson C, Layman D, Maas J, Jaeger S, Walker R, et al: **The DNA sequence of human chromosome 7.** *Nature* 2003, **424:**157-164.
22. Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, Wilming L, Jones MC, Horton R, Hunt SE, Scott CE, et al: **The DNA sequence and analysis of human chromosome 6.** *Nature* 2003, **425:**805-811.
23. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456:**53-59.
24. Thompson JF, Steinmann KE: **Single molecule sequencing with a HeliScope genetic analysis system.** *Curr Protoc Mol Biol* 2010, **Chapter 7:**Unit7 10.
25. Rusk N: **Torrents of sequence.** *Nat Meth* 2011, **8:**44-44.
26. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, et al: **A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning.** *Genome Res* 2008, **18:**1051-1063.
27. Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H: **Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore.** *Proc Natl Acad Sci U S A* 2009, **106:**7702-7707.

28. Shokralla S, Spall JL, Gibson JF, Hajibabaei M: **Next-generation sequencing technologies for environmental DNA research.** *Mol Ecol* 2012, **21:**1794-1805.
29. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, et al: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
30. Nyren P: **The history of pyrosequencing.** *Methods Mol Biol* 2007, **373:**1-14.
31. Ahmadian A, Ehn M, Hober S: **Pyrosequencing: history, biochemistry and future.** *Clin Chim Acta* 2006, **363:**83-94.
32. Schuwirth BS, Borovinskaya MA, Hau CW, Zhang W, Vila-Sanjurjo A, Holton JM, Cate JH: **Structures of the bacterial ribosome at 3.5 A resolution.** *Science* 2005, **310:**827-834.
33. Klein DJ, Moore PB, Steitz TA: **The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit.** *J Mol Biol* 2004, **340:**141-177.
34. Alberts B, Wilson JH, Hunt T: *Molecular biology of the cell.* 5th edn. New York: Garland Science; 2008.
35. Kurland CG: **Molecular characterization of ribonucleic acid from *Escherichia coli* ribosomes: I. Isolation and molecular weights.** *Journal of Molecular Biology* 1960, **2:**83-91.
36. Brownlee GG, Sanger F, Barrell BG: **Nucleotide sequence of 5S-ribosomal RNA from *Escherichia coli.*** *Nature* 1967, **215:**735-736.
37. Stackebrandt E, Goebel BM: **A Place for DNA-DNA Reassociation and 16s Ribosomal-Rna Sequence-Analysis in the Present Species Definition in Bacteriology.** *International Journal of Systematic Bacteriology* 1994, **44:**846-849.
38. Ludwig W, Schleifer KH: **Bacterial phylogeny based on 16S and 23S rRNA sequence analysis.** *FEMS Microbiol Rev* 1994, **15:**155-173.
39. Giovannoni SJ, Britschgi TB, Moyer CL, Field KG: **Genetic diversity in Sargasso Sea bacterioplankton.** *Nature* 1990, **345:**60-63.
40. Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, et al: **The phylogeny of prokaryotes.** *Science* 1980, **209:**457-463.
41. Kaczanowska M, Ryden-Aulin M: **Ribosome biogenesis and the translation process in *Escherichia coli.*** *Microbiol Mol Biol Rev* 2007, **71:**477-494.
42. Woese CR, Fox GE, Zablen L, Uchida T, Bonen L, Pechman K, Lewis BJ, Stahl D: **Conservation of primary structure in 16S ribosomal RNA.** *Nature* 1975, **254:**83-86.

43. Van de Peer Y, Chapelle S, De Wachter R: **A quantitative map of nucleotide substitution rates in bacterial rRNA.** *Nucleic Acids Res* 1996, **24:**3381-3391.

44. Stern S, Powers T, Changchien LM, Noller HF: **RNA-protein interactions in 30S ribosomal subunits: folding and function of 16S rRNA.** *Science* 1989, **244:**783-790.

45. Smit S, Widmann J, Knight R: **Evolutionary rates vary among rRNA structural elements.** *Nucleic Acids Res* 2007, **35:**3339-3354.

46. Chakravorty S, Helb D, Burday M, Connell N, Alland D: **A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria.** *J Microbiol Methods* 2007, **69:**330-339.

47. Schloss PD: **The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies.** *PLoS Comput Biol* 2010, **6:**e1000844.

48. Tringe SG, Hugenholtz P: **A renaissance for the pioneering 16S rRNA gene.** *Curr Opin Microbiol* 2008, **11:**442-446.

49. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N: **Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia.** *Science* 1985, **230:**1350-1354.

50. Mullis KB, Faloona FA: **Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction.** *Methods Enzymol* 1987, **155:**335-350.

51. Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM, Jansson JK: **Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw.** *Nature* 2011, **480:**368-371.

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

53. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, et al: **QIIME allows analysis of high-throughput community sequencing data.** *Nat Meth* 2010, **7:**335-336.

54. Patin NV, Kunin V, Lidstrom U, Ashby MN: **Effects of OTU clustering and PCR artifacts on microbial diversity estimates.** *Microb Ecol* 2013, **65:**709-719.

55. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ: **Microbial diversity in the deep sea and the underexplored "rare biosphere".** *Proc Natl Acad Sci U S A* 2006, **103:**12115-12120.

56. Newton RJ, Kent AD, Triplett EW, McMahon KD: **Microbial community dynamics in a humic lake: differential persistence of common freshwater phylotypes.** *Environ Microbiol* 2006, **8:**956-970.

57. Yokoyama H, Moriya N, Ohmori H, Waki M, Ogino A, Tanaka Y: **Community analysis of hydrogen-producing extreme thermophilic anaerobic microflora enriched from cow manure with five substrates.** *Appl Microbiol Biotechnol* 2007, **77:**213-222.

58. Grice EA, Kong HH, Renaud G, Young AC, Program NCS, Bouffard GG, Blakesley RW, Wolfsberg TG, Turner ML, Segre JA: **A diversity profile of the human skin microbiota.** *Genome Res* 2008, **18:**1043-1050.

59. Zhang H, DiBaise JK, Zuccolo A, Kudrna D, Braidotti M, Yu Y, Parameswaran P, Crowell MD, Wing R, Rittmann BE, Krajmalnik-Brown R: **Human gut microbiota in obesity and after gastric bypass.** *Proc Natl Acad Sci U S A* 2009, **106:**2365-2370.

60. Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, Yatsunenko T, Niazi F, Affourtit J, Egholm M, Henrissat B, Knight R, Gordon JI: **Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins.** *Proc Natl Acad Sci U S A* 2010, **107:**7503-7508.

61. Nam YD, Jung MJ, Roh SW, Kim MS, Bae JW: **Comparative analysis of Korean human gut microbiota by barcoded pyrosequencing.** *PLoS One* 2011, **6:**e22109.

62. Hong PY, Wheeler E, Cann IK, Mackie RI: **Phylogenetic analysis of the fecal microbial community in herbivorous land and marine iguanas of the Galapagos Islands using 16S rRNA-based pyrosequencing.** *ISME J* 2011, **5:**1461-1470.

63. Lozupone C, Knight R: **UniFrac: a new phylogenetic method for comparing microbial communities.** *Appl Environ Microbiol* 2005, **71:**8228-8235.

64. Shine J, Dalgarno L: **The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites.** *Proc Natl Acad Sci U S A* 1974, **71:**1342-1346.

65. Lin YH, Chang BC, Chiang PW, Tang SL: **Questionable 16S ribosomal RNA gene annotations are frequent in completed microbial genomes.** *Gene* 2008, **416:**44-47.

66. Lim K, Furuta Y, Kobayashi I: **Large variations in bacterial ribosomal RNA genes.** *Mol Biol Evol* 2012, **29:**2937-2948.

67. Salman V, Amann R, Shub DA, Schulz-Vogt HN: **Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria.** *Proc Natl Acad Sci U S A* 2012, **109:**4203-4208.

68. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM: **The Ribosomal Database Project: improved alignments and new tools for rRNA analysis.** *Nucleic Acids Res* 2009, **37:**D141-145.

69. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.** *Nucleic Acids Res* 2013, **41:**D590-596.

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

71. Pearson WR: **Flexible sequence similarity searching with the FASTA3 program package.** *Methods Mol Biol* 2000, **132:**185-219.

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

73. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.

74. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol* 2013, **30:**772-780.

75. Thompson JD, Gibson TJ, Higgins DG: **Multiple sequence alignment using ClustalW and ClustalX.** *Curr Protoc Bioinformatics* 2002, **Chapter 2:**Unit 2 3.

76. Goto N, Prins P, Nakao M, Bonnal R, Aerts J, Katayama T: **BioRuby: bioinformatics software for the Ruby programming language.** *Bioinformatics* 2010, **26:**2617-2619.

77. de la Bastide M, McCombie WR: **Assembling genomic DNA sequences with PHRAP.** *Curr Protoc Bioinformatics* 2007, **Chapter 11:**Unit11 14.

78. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, et al: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.** *Mol Syst Biol* 2011, **7:**539.

79. Federhen S: **The NCBI Taxonomy database.** *Nucleic Acids Res* 2012, **40:**D136-143.

80. Yarza P, Ludwig W, Euzeby J, Amann R, Schleifer KH, Glockner FO, Rossello-Mora R: **Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses.** *Syst Appl Microbiol* 2010, **33:**291-299.

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

82. Droege M, Hill B: **The Genome Sequencer FLX System--longer reads, more applications, straight forward bioinformatics and more complete data sets.** *J Biotechnol* 2008, **136:**3-10.

83. Andersson AF, Lindberg M, Jakobsson H, Backhed F, Nyren P, Engstrand L: **Comparative analysis of human gut microbiota by barcoded pyrosequencing.** *PLoS One* 2008, **3:**e2836.

84. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R: **Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex.** *Nat Methods* 2008, **5:**235-237.

85. Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'Toole PW: **Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions.** *Nucleic Acids Res* 2010, **38:**e200.

86. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R: **Forensic identification using skin bacterial communities.** *Proc Natl Acad Sci U S A* 2010, **107:**6477-6481.

87. Zhou HW, Li DF, Tam NF, Jiang XT, Zhang H, Sheng HF, Qin J, Liu X, Zou F: **BIPES, a cost-effective high-throughput method for assessing microbial diversity.** *ISME J* 2011, **5:**741-749.

88. Hattori M, Taylor TD: **The human intestinal microbiome: a new frontier of human biology.** *DNA Res* 2009, **16:**1-12.

89. Frank JA, Reich CI, Sharma S, Weisbaum JS, Wilson BA, Olsen GJ: **Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes.** *Appl Environ Microbiol* 2008, **74:**2461-2470.

90. Hill JE, Fernando WM, Zello GA, Tyler RT, Dahl WJ, Van Kessel AG: **Improvement of the representation of bifidobacteria in fecal microbiota metagenomic libraries by application of the cpn60 universal primer cocktail.** *Appl Environ Microbiol* 2010, **76:**4550-4552.

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

92. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132:**365-386.

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

94. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23:**2947-2948.

95. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, et al: **Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons.** *Genome Res* 2011, **21:**494-504.

96. Schloss PD, Gevers D, Westcott SL: **Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies.** *PLoS One* 2011, **6:**e27310.

97. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: **Accuracy and quality of massively parallel DNA pyrosequencing.** *Genome Biol* 2007, **8:**R143.

98. Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, Martin JF: **Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing.** *BMC Genomics* 2011, **12:**245.

99. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT: **Accurate determination of microbial diversity from 454 pyrosequencing data.** *Nat Methods* 2009, **6:**639-641.

100. Diaz PI, Dupuy AK, Abusleme L, Reese B, Obergfell C, Choquette L, Dongari-Bagtzoglou A, Peterson DE, Terzi E, Strausbaugh LD: **Using high throughput sequencing to explore the biodiversity in oral bacterial communities.** *Mol Oral Microbiol* 2012, **27:**182-201.

101. Preidis GA, Versalovic J: **Targeting the human microbiome with antibiotics, probiotics, and prebiotics: gastroenterology enters the metagenomics era.** *Gastroenterology* 2009, **136:**2015-2031.

102. Patel RM, Lin PW: **Developmental biology of gut-probiotic interaction.** *Gut Microbes* 2010, **1:**186-195.

103. Gerritsen J, Smidt H, Rijkers GT, de Vos WM: **Intestinal microbiota in human health and disease: the impact of probiotics.** *Genes Nutr* 2011, **6:**209-240.

104. Sanders ME, Heimbach JT, Pot B, Tancredi DJ, Lenoir-Wijnkoop I, Lahteenmaki-Uutela A, Gueimonde M, Banares S: **Health claims substantiation for probiotic and prebiotic products.** *Gut Microbes* 2011, **2:**127-133.

105. Aureli P, Capurso L, Castellazzi AM, Clerici M, Giovannini M, Morelli L, Poli A, Pregliasco F, Salvini F, Zuccotti GV: **Probiotics and health: an evidence-based review.** *Pharmacol Res* 2011, **63:**366-376.

106. Rauch M, Lynch SV: **The potential for probiotic manipulation of the gastrointestinal microbiome.** *Curr Opin Biotechnol* 2012, **23:**192-201.

107. Fujimura KE, Slusher NA, Cabana MD, Lynch SV: **Role of the gut microbiota in defining human health.** *Expert Rev Anti Infect Ther* 2010, **8:**435-454.

108. Deshpande GC, Rao SC, Keil AD, Patole SK: **Evidence-based guidelines for use of probiotics in preterm neonates.** *BMC Med* 2011, **9:**92.

109. Bron PA, van Baarlen P, Kleerebezem M: **Emerging molecular insights into the interaction between probiotics and the host intestinal mucosa.** *Nat Rev Microbiol* 2012, **10:**66-78.

110. Thomas DW, Greer FR, American Academy of Pediatrics Committee on N, American Academy of Pediatrics Section on Gastroenterology H, Nutrition: **Probiotics and prebiotics in pediatrics.** *Pediatrics* 2010, **126:**1217-1231.

111. Indrio F, Neu J: **The intestinal microbiome of infants and the use of probiotics.** *Curr Opin Pediatr* 2011, **23:**145-150.

112. Saxelin M, Tynkkynen S, Mattila-Sandholm T, de Vos WM: **Probiotic and other functional microbes: from markets to mechanisms.** *Curr Opin Biotechnol* 2005, **16:**204-211.

113. Nagpal R, Kumar A, Kumar M, Behare PV, Jain S, Yadav H: **Probiotics, their health benefits and applications for developing healthier foods: a review.** *FEMS Microbiol Lett* 2012.

114. Bisanz JE, Reid G: **Unraveling how probiotic yogurt works.** *Sci Transl Med* 2011, **3:**106ps141.

115. Bron PA, Kleerebezem M: **Engineering lactic acid bacteria for increased industrial functionality.** *Bioeng Bugs* 2011, **2:**80-87.

116. Kleerebezem M, Vaughan EE: **Probiotic and gut lactobacilli and bifidobacteria: molecular approaches to study diversity and activity.** *Annu Rev Microbiol* 2009, **63:**269-290.

117. Ventura M, O'Flaherty S, Claesson MJ, Turroni F, Klaenhammer TR, van Sinderen D, O'Toole PW: **Genome-scale analyses of health-promoting bacteria: probiogenomics.** *Nat Rev Microbiol* 2009, **7:**61-71.

118. Snydman DR: **The safety of probiotics.** *Clin Infect Dis* 2008, **46 Suppl 2:**S104-111; discussion S144-151.

119. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R: **Diversity, stability and resilience of the human gut microbiota.** *Nature* 2012, **489:**220-230.

120. Clemente JC, Ursell LK, Parfrey LW, Knight R: **The impact of the gut microbiota on human health: an integrative view.** *Cell* 2012, **148:**1258-1270.

121. Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, Pettersson S: **Host-gut microbiota metabolic interactions.** *Science* 2012, **336:**1262-1267.

122. Walter J, Ley R: **The human gut microbiome: ecology and recent evolutionary changes.** *Annu Rev Microbiol* 2011, **65:**411-429.

123. Hooper LV, Littman DR, Macpherson AJ: **Interactions between the microbiota and the immune system.** *Science* 2012, **336:**1268-1273.

124. Tannock GW, Munro K, Harmsen HJ, Welling GW, Smart J, Gopal PK: **Analysis of the fecal microflora of human subjects consuming a probiotic product containing *Lactobacillus rhamnosus* DR20.** *Appl Environ Microbiol* 2000, **66:**2578-2588.

125. Garcia-Albiach R, Pozuelo de Felipe MJ, Angulo S, Morosini MI, Bravo D, Baquero F, del Campo R: **Molecular analysis of yogurt containing *Lactobacillus delbrueckii* subsp. *bulgaricus* and *Streptococcus thermophilus* in human intestinal microbiota.** *Am J Clin Nutr* 2008, **87:**91-96.

126. Alvaro E, Andrieux C, Rochet V, Rigottier-Gois L, Lepercq P, Sutren M, Galan P, Duval Y, Juste C, Dore J: **Composition and metabolism of the intestinal microbiota in consumers and non-consumers of yogurt.** *Br J Nutr* 2007, **97:**126-133.

127. Rochet V, Rigottier-Gois L, Levenez F, Cadiou J, Marteau P, Bresson JL, Goupil-Feillerat N, Dore J: **Modulation of *Lactobacillus casei* in ileal and fecal samples from healthy volunteers after consumption of a fermented milk containing *Lactobacillus casei* DN-114 001Rif.** *Can J Microbiol* 2008, **54:**660-667.

128. Rochet V, Rigottier-Gois L, Ledaire A, Andrieux C, Sutren M, Rabot S, Mogenet A, Bresson JL, Cools S, Picard C, et al: **Survival of *Bifidobacterium animalis* DN-173 010 in the faecal microbiota after administration in lyophilised form or in fermented product - a randomised study in healthy adults.** *J Mol Microbiol Biotechnol* 2008, **14:**128-136.

129. Ouwehand AC, Bergsma N, Parhiala R, Lahtinen S, Gueimonde M, Finne-Soveri H, Strandberg T, Pitkala K, Salminen S: ***Bifidobacterium* microbiota and parameters of immune function in elderly subjects.** *FEMS Immunol Med Microbiol* 2008, **53:**18-25.

130. Firmesse O, Mogenet A, Bresson JL, Corthier G, Furet JP: ***Lactobacillus rhamnosus* R11 consumed in a food supplement survived human digestive transit without modifying microbiota equilibrium as assessed by real-time polymerase chain reaction.** *J Mol Microbiol Biotechnol* 2008, **14:**90-99.

131. Lahtinen SJ, Tammela L, Korpela J, Parhiala R, Ahokoski H, Mykkanen H, Salminen SJ: **Probiotics modulate the *Bifidobacterium* microbiota of elderly nursing home residents.** *Age (Dordr)* 2009, **31:**59-66.

132. Savard P, Lamarche B, Paradis ME, Thiboutot H, Laurin E, Roy D: **Impact of *Bifidobacterium animalis* subsp. *lactis* BB-12 and, *Lactobacillus acidophilus* LA-5-containing yoghurt, on fecal bacterial counts of healthy adults.** *Int J Food Microbiol* 2011, **149:**50-57.

133. Yamano T, Iino H, Takada M, Blum S, Rochat F, Fukushima Y: **Improvement of the human intestinal flora by ingestion of the probiotic strain *Lactobacillus johnsonii* La1.** *Br J Nutr* 2006, **95:**303-312.

134. Engelbrektson AL, Korzenik JR, Sanders ME, Clement BG, Leyer G, Klaenhammer TR, Kitts CL: **Analysis of treatment effects on the microbial ecology of the human intestine.** *FEMS Microbiol Ecol* 2006, **57:**239-250.

135. Marzotto M, Maffeis C, Paternoster T, Ferrario R, Rizzotti L, Pellegrino M, Dellaglio F, Torriani S: ***Lactobacillus paracasei* A survives gastrointestinal passage and affects the fecal microbiota of healthy infants.** *Res Microbiol* 2006, **157:**857-866.

136. Cox MJ, Huang YJ, Fujimura KE, Liu JT, McKean M, Boushey HA, Segal MR, Brodie EL, Cabana MD, Lynch SV: ***Lactobacillus casei* abundance is associated with profound shifts in the infant gut microbiome.** *PLoS One* 2010, **5:**e8745.

137. Culligan EP, Hill C, Sleator RD: **Probiotics and gastrointestinal disease: successes, problems and future prospects.** *Gut Pathog* 2009, **1:**19.

138. Gareau MG, Sherman PM, Walker WA: **Probiotics and the gut microbiota in intestinal health and disease.** *Nat Rev Gastroenterol Hepatol* 2010, **7:**503-514.

139. McNulty NP, Yatsunenko T, Hsiao A, Faith JJ, Muegge BD, Goodman AL, Henrissat B, Oozeer R, Cools-Portier S, Gobert G, et al: **The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins.** *Sci Transl Med* 2011, **3:**106ra106.

140. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11:**31-46.

141. Hamady M, Lozupone C, Knight R: **Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data.** *ISME J* 2010, **4:**17-27.

142. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R: **Experimental and analytical tools for studying the human microbiome.** *Nat Rev Genet* 2012, **13:**47-58.

143. Morita H, Kuwahara T, Ohshima K, Sasamoto H, Itoh K, Hattori M, Hayashi T, Takami H: **An Improved DNA Isolation Method for Metagenomic Analysis of the Microbial Flora of the Human Intestine.** *Microbes and Environments* 2007, **22:**214-222.

144. Maidak BL, Cole JR, Lilburn TG, Parker CT, Jr., Saxman PR, Farris RJ, Garrity GM, Olsen GJ, Schmidt TM, Tiedje JM: **The RDP-II (Ribosomal Database Project).** *Nucleic Acids Res* 2001, **29:**173-174.

145. Edgar RC: **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics* 2010, **26:**2460-2461.

146. del Campo R, Bravo D, Canton R, Ruiz-Garbajosa P, Garcia-Albiach R, Montesi-Libois A, Yuste FJ, Abraira V, Baquero F: **Scarce evidence of yogurt lactic acid bacteria in human feces after daily yogurt consumption by healthy volunteers.** *Appl Environ Microbiol* 2005, **71:**547-549.

147. Oozeer R, Leplingard A, Mater DD, Mogenet A, Michelin R, Seksek I, Marteau P, Dore J, Bresson JL, Corthier G: **Survival of *Lactobacillus casei* in the human digestive tract after consumption of fermented milk.** *Appl Environ Microbiol* 2006, **72:**5615-5617.

148. Marco ML, de Vries MC, Wels M, Molenaar D, Mangell P, Ahrne S, de Vos WM, Vaughan EE, Kleerebezem M: **Convergence in probiotic *Lactobacillus* gut-adaptive responses in humans and mice.** *ISME J* 2010, **4:**1481-1484.

149. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, Nalin R, Jarrin C, Chardon P, Marteau P, et al: **Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach.** *Gut* 2006, **55:**205-211.

150. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP, et al: **Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes.** *DNA Res* 2007, **14:**169-181.

151. Jernberg C, Lofmark S, Edlund C, Jansson JK: **Long-term ecological impacts of antibiotic administration on the human intestinal microbiota.** *ISME J* 2007, **1:**56-66.

152. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, et al: **Linking long-term dietary patterns with gut microbial enterotypes.** *Science* 2011, **334:**105-108.

153. Sokol H, Seksik P, Furet JP, Firmesse O, Nion-Larmurier I, Beaugerie L, Cosnes J, Corthier G, Marteau P, Dore J: **Low counts of *Faecalibacterium prausnitzii* in colitis microbiota.** *Inflamm Bowel Dis* 2009, **15:**1183-1189.

Appendix

Table 1. Strains having difference in 16S copy numbers in the individual genome between GRD and GenBank

Name	GRD	GenBank	Difference*
*Clostridium_difficile*_M120_uid158361	11	0	11
*Clostridium_difficile*_2007855_uid158365	9	0	9
*Clostridium_difficile*_CF5_uid158359	9	0	9
*Vibrio_fischeri*_MJ11_uid58907	10	1	9
*Bacillus_coagulans*_2_6_uid68053	9	1	8
*Azospirillum_brasilense*_Sp245_uid162161	11	5	6
*Staphylococcus_aureus*_ST398_uid159247	6	0	6
*Paenibacillus_polymyxa*_M1_uid162159	14	10	4
*Pseudomonas_putida*_DOT_T1E_uid171260	8	4	4
*Actinobacillus_pleuropneumoniae*_serovar_5b_L20_uid58789	6	3	3
*Actinobacillus_pleuropneumoniae*_serovar_7_AP76_uid59231	6	3	3
*Alteromonas_macleodii*__Deep_ecotype__uid58251	6	4	2
*Bacillus_subtilis*_BSn5_uid62463	11	9	2
*Heliobacterium_modesticaldum*_Ice1_uid58279	10	8	2
*Methanosarcina_mazei*_Go1_uid57893	3	1	2
*Staphylococcus_aureus*_HO_5096_0412_uid162163	6	4	2
*Staphylococcus_pseudintermedius*_HKU10_03_uid62125	6	4	2
*Acinetobacter_baumannii*_TYTH_1_uid176498	6	5	1
*Bacillus_amyloliquefaciens_plantarum*_AS43_3_uid183682	10	9	1
*Bacillus_amyloliquefaciens*_TA208_uid158701	7	6	1
*Bacillus_amyloliquefaciens*_XH7_uid158881	8	7	1
*Bacillus_amyloliquefaciens*_Y2_uid165195	10	9	1
*Clostridium_difficile*_R20291_uid40921	10	9	1
*Desulfotomaculum_reducens*_MI_1_uid58277	9	8	1
*Haemophilus_influenzae*_10810_uid86647	6	5	1
*Listeria_monocytogenes*_serotype_4a_L99_uid161953	6	5	1
*Pantoea_ananatis*_LMG_20103_uid46807	7	6	1
*Thermomicrobium_roseum*_DSM_5159_uid59341	2	1	1
*Vibrio_harveyi*_ATCC_BAA_1116_uid58957	11	10	1
*Pseudomonas_fluorescens*_SBW25_uid158693	5	6	-1
*Lactobacillus_casei*_BD_II_uid162119	5	10	-5
*Lactobacillus_casei*_LC2W_uid162121	5	10	-5

*(Difference) = (GRD) – (GenBank)

Table 2. Strains having difference in the average length ≥ 100 bp between GRD and GenBank

Name	GRD	GenBank	Difference*
*Pyrobaculum*_1860_uid82379	3,606	1,499	2,107
*Fervidobacterium_pennivorans*_DSM_9078_uid78143	2,041	1,273	768
*Pyrobaculum_aerophilum*_IM2_uid57727	2,215	1,498	717
*Pyrobaculum_arsenaticum*_DSM_13514_uid58409	2,214	1,499	715
*Pyrobaculum_neutrophilum*_V24Sta_uid58421	2,204	1,495	709
*Aeropyrum_pernix*_K1_uid57757	2,204	1,501	703
*Pyrobaculum_oguniense*_TE7_uid84411	2,202	1,506	696

Name	GRD	GenBank	Difference*
*Thermus_scotoductus*_SA_01_uid62273	2,177	1,539	638
*Micrococcus_luteus*_NCTC_2665_uid59033	2,231	1,616	615
*Neisseria_lactamica*_020_06_uid60851	1,541	986	555
*Neisseria_meningitidis*_053442_uid58587	1,541	986	555
*Corynebacterium_resistens*_DSM_45100_uid50555	1,967	1,544	423
*Corynebacterium_aurimucosum*_ATCC_700975_uid59409	1,849	1,543	306
*Bifidobacterium_animalis_lactis*_BB_12_uid158871	1,538	1,232	306
*Burkholderia_glumae*_BGR1_uid59397	1,533	1,236	297
*Thiomonas*_3As_uid178369	1,529	1,264	265
*Candidatus_Cloacamonas_acidaminovorans*_uid62959	1,602	1,355	247
*Burkholderia_gladioli*_BSR3_uid66301	1,533	1,296	237
*Leptospira_biflexa*_serovar_Patoc__Patoc_1__Paris__uid58993	1,500	1,276	224
*Desulfitobacterium_dehalogenans*_ATCC_51507_uid82553	1,647	1,427	220
*Streptococcus_intermedius*_JTH08_uid168614	1,554	1,335	219
*Streptococcus_equi_zooepidemicus*_MGCS10565_uid59263	1,552	1,335	217
*Spirochaeta_thermophila*_DSM_6192_uid53037	1,547	1,338	209
*Desulfitobacterium_dichloroeliminans*_LMG_P_21439_uid82555	1,687	1,479	208
*Clostridium_sticklandii*_DSM_519_uid59585	1,561	1,359	202
*Mycoplasma_hyorhinis*_HUB_1_uid51695	1,525	1,349	176
*Borrelia_garinii*_BgVir_uid162165	1,536	1,367	169
*Clostridium*_BNL1100_uid84307	1,554	1,394	160
*Thermobacillus_composti*_KWC4_uid74021	1,605	1,449	156
*Synechococcus*_JA_2_3B_a_2_13__uid58537	1,478	1,323	155
*Melissococcus_plutonius*_ATCC_35311_uid66803	1,555	1,402	153
*Synechococcus*_JA_3_3Ab_uid58535	1,474	1,324	150
*Clostridium_clariflavum*_DSM_19732_uid82345	1,516	1,367	149
*Actinobacillus_pleuropneumoniae*_serovar_5b_L20_uid58789	1,686	1,541	145
*Campylobacter_curvus*_525_92_uid58669	1,713	1,575	138
*Streptococcus_agalactiae*_NEM316_uid61585	1,547	1,409	138
*Chlorobium_luteolum*_DSM_273_uid58175	1,500	1,363	137
*Clostridium_cellulolyticum*_H10_uid58709	1,654	1,518	136
*Caldisphaera_lagunensis*_DSM_15908_uid183486	1,524	1,388	136
*Natrinema_pellirubrum*_DSM_15624_uid74437	1,496	1,361	135
*Caldivirga_maquilingensis*_IC_167_uid58711	1,687	1,554	133
*Natronobacterium_gregoryi*_SP2_uid74439	1,494	1,361	133
*Streptococcus_pyogenes*_NZ131_uid59035	1,551	1,419	132
*Halovivax_ruber*_XH_70_uid184819	1,494	1,362	132
*Lactococcus_lactis_cremoris*_SK11_uid57983	1,547	1,416	131
*Streptococcus_pneumoniae*_ATCC_700669_uid59287	1,544	1,413	131
*Streptococcus_pneumoniae*_CGSP14_uid59181	1,543	1,412	131
*Corynebacterium_diphtheriae*_241_uid83607	1,520	1,392	128

Name	GRD	GenBank	Difference*
*Natronococcus_occultus*_SP4_uid184863	1,491	1,363	128
*Corynebacterium_diphtheriae*_NCTC_13129_uid57691	1,519	1,392	127
*Aciduliprofundum*_MAR08_339_uid184407	1,495	1,370	125
*Candidatus_Portiera_aleyrodidarum*_BT_B_uid173859	1,541	1,418	123
*Anaerobaculum_mobile*_DSM_13181_uid168323	1,531	1,408	123
*Mycobacterium_smegmatis*_JS623_uid184820	1,524	1,402	122
*Streptococcus_parasanguinis*_FW213_uid163997	1,550	1,429	121
*Sphaerochaeta_pleomorpha*_Grapes_uid82365	1,541	1,420	121
*Mesotoga_prima*_MesG1_Ag_4_2_uid52599	1,529	1,408	121
*Gramella_forsetii*_KT0803_uid58881	1,528	1,407	121
*Mycobacterium_chubuense*_NBB4_uid168322	1,520	1,399	121
*Mycobacterium_rhodesiae*_NBB3_uid75107	1,520	1,399	121
*Microcoleus*_PCC_7113_uid183114	1,491	1,370	121
*Nostoc*_PCC_7524_uid182933	1,488	1,367	121
*Rivularia*_PCC_7116_uid182929	1,487	1,366	121
*Desulfosporosinus_acidiphilus*_SJ4_uid156759	1,578	1,458	120
*Desulfosporosinus_meridiei*_DSM_13257_uid75097	1,567	1,447	120
*Desulfosporosinus_orientis*_DSM_765_uid82939	1,566	1,446	120
*Geobacillus*_HH01_uid188479	1,558	1,438	120
*Desulfomonile_tiedjei*_DSM_6799_uid168320	1,557	1,437	120
*secondary_endosymbiont_of_Heteropsylla_cubana*_Thao2000_uid172738	1,547	1,427	120
*Marinitoga_piezophila*_KA3_uid81629	1,541	1,421	120
*Spirochaeta_africana*_DSM_8902_uid81779	1,537	1,417	120
*Pleurocapsa*_PCC_7327_uid183006	1,490	1,370	120
*Dactylococcopsis_salina*_PCC_8305_uid183341	1,489	1,369	120
*Cyanobium_gracile*_PCC_6307_uid182931	1,485	1,365	120
*Oscillatoria_acuminata*_PCC_6304_uid183003	1,490	1,371	119
*Cylindrospermum_stagnale*_PCC_7417_uid183111	1,488	1,369	119
*Chamaesiphon*_PCC_6605_uid183005	1,487	1,368	119
*Synechococcus*_PCC_7502_uid183008	1,484	1,365	119
*Streptococcus_suis*_GZ1_uid161937	1,554	1,436	118
*Frateuria_aurantia*_DSM_6220_uid81775	1,545	1,427	118
*Rahnella_aquatilis*_CIP_78_65___ATCC_33071_uid86855	1,543	1,425	118
*Enterobacteriaceae_bacterium*_FGI_57_uid185181	1,542	1,424	118
*secondary_endosymbiont_of_Ctenarytaina_eucalypti*_uid172737	1,542	1,424	118
*Serratia_marcescens*_FGI94_uid185180	1,541	1,423	118
*Dechlorosoma_suillum*_PS_uid81439	1,540	1,422	118
*Acinetobacter_baumannii*_MDR_TJ_uid162739	1,537	1,419	118
*Halobacteroides_halobius*_DSM_5150_uid184862	1,536	1,418	118
*Pseudomonas_stutzeri*_RCH2_uid184342	1,536	1,418	118
*Thioflavicoccus_mobilis*_8321_uid184343	1,536	1,418	118

Name	GRD	GenBank	Difference*
*Thiocystis_violascens*_DSM_198_uid74025	1,529	1,411	118
*Terriglobus_roseus*_DSM_18391_uid168183	1,500	1,382	118
*alpha_proteobacterium*_HIMB59_uid175778	1,494	1,376	118
*Sinorhizobium_meliloti*_GR4_uid184823	1,484	1,366	118
*Sulfurospirillum_barnesii*_SES_3_uid168117	1,512	1,396	116
*Edwardsiella_tarda*_FL6_60_uid159657	1,544	1,429	115
*Thermus_thermophilus*_JL_18_uid162129	1,519	1,405	114
*Methanomethylovorans_hollandica*_DSM_15978_uid184864	1,479	1,365	114
*Corynebacterium_efficiens*_YS_314_uid62905	1,521	1,408	113
*Singulisphaera_acidiphila*_DSM_18658_uid81777	1,514	1,401	113
*Thermus_oshimai*_JL_2_uid178948	1,507	1,394	113
*Methanoregula_formicicum*_SMSP_uid184406	1,471	1,358	113
*Streptococcus_pneumoniae*_Hungary19A_6_uid59117	1,544	1,434	110
*Streptococcus_pneumoniae*_P1031_uid59123	1,544	1,434	110
*Flexibacter_litoralis*_DSM_6794_uid168257	1,528	1,418	110
*Alistipes_finegoldii*_DSM_17242_uid168180	1,524	1,414	110
*Echinicola_vietnamensis*_DSM_17526_uid184076	1,522	1,412	110
*Deinococcus_peraridilitoris*_DSM_19664_uid183485	1,506	1,396	110
*Prevotella_dentalis*_DSM_3688_uid184818	1,534	1,425	109
*Aequorivita_sublithincola*_DSM_14238_uid168181	1,524	1,415	109
*Belliella_baltica*_DSM_15883_uid168182	1,522	1,413	109
*Owenweeksia_hongkongensis*_DSM_17368_uid82951	1,520	1,411	109
*Solitalea_canadensis*_DSM_3403_uid81783	1,520	1,411	109
*Ornithobacterium_rhinotracheale*_DSM_15997_uid168256	1,517	1,408	109
*Nanoarchaeum_equitans*_Kin4_M_uid58009	1,504	1,395	109
*alpha_proteobacterium*_HIMB5_uid175779	1,475	1,366	109
*Candidatus_Carsonella_ruddii*_HC_isolate_Thao2000_uid172734	1,549	1,442	107
*Candidatus_Carsonella_ruddii*_CE_isolate_Thao2000_uid172732	1,544	1,441	103
*Candidatus_Carsonella_ruddii*_CS_isolate_Thao2000_uid172733	1,538	1,435	103
*Candidatus_Carsonella_ruddii*_PC_isolate_NHV_uid172736	1,536	1,433	103
*Bacillus_cereus*_NC7401_uid82815	1,551	1,450	101
*Candidatus_Carsonella_ruddii*_HT_isolate_Thao2000_uid172735	1,540	1,439	101
*Buchnera_aphidicola*__Cinara_tujafilina__uid68101	1,552	1,452	100
*Bacillus_thuringiensis*_Al_Hakam_uid58795	1,551	1,451	100
*Bacillus_amyloliquefaciens*_TA208_uid158701	1,457	1,557	-100
*Bacillus_amyloliquefaciens*_Y2_uid165195	1,454	1,555	-101
*Acinetobacter_baumannii*_TYTH_1_uid176498	1,355	1,491	-136
*Clostridium_difficile*_CD196_uid41017	1,493	1,632	-139
*Bacillus_subtilis*_BSn5_uid62463	1,417	1,558	-141
*Sorangium_cellulosum*__So_ce_56__uid61629	1,557	1,718	-161

Name	GRD	GenBank	Difference*
*Pseudomonas_fluorescens*_SBW25_uid158693	1,537	1,763	-226
*Clostridium_difficile*_R20291_uid40921	1,397	1,643	-246
*Alteromonas_macleodii*__Deep_ecotype__uid58251	1,276	1,533	-257
*Corynebacterium_diphtheriae*_HC03_uid84299	1,519	1,877	-358
*Corynebacterium_diphtheriae*_BH8_uid84311	1,519	1,878	-359
*Azospirillum_brasilense*_Sp245_uid162161	942	1,411	-469
*Methanococcus_maripaludis*_X1_uid70729	1,488	2,482	-994
*Brachyspira_pilosicoli*_B2904_uid175255	1,513	3,256	-1,743
*Brachyspira_pilosicoli*_95_1000_uid50609	1,513	3,281	-1,768
*Brachyspira_pilosicoli*_WesB_uid175256	1,513	3,281	-1,768

*(Difference) = (GRD) – (GenBank)

Table 3. Strains having more than 10 copies of 16S rRNA genes in the genome

Name	16S copy number
*Bacillus_thuringiensis*_MC28_uid176369	15
*Brevibacillus_brevis*_NBRC_100599_uid59175	15
*Photobacterium_profundum*_SS9_uid62923	15
*Bacillus_cereus*_03BB102_uid59299	14
*Bacillus_cereus*_AH187_uid58753	14
*Bacillus_cereus*_B4264_uid58757	14
*Bacillus_cereus*_NC7401_uid82815	14
*Bacillus_thuringiensis*_Al_Hakam_uid58795	14
*Bacillus_thuringiensis*_BMB171_uid49135	14
*Bacillus_thuringiensis*_Bt407_uid177931	14
*Bacillus_thuringiensis*_HD_789_uid173860	14
*Bacillus_thuringiensis*_serovar_finitimus_YBT_020_uid158875	14
*Bacillus_thuringiensis*_serovar_konkukian_97_27_uid58089	14
*Bacillus_weihenstephanensis*_KBAB4_uid58315	14
*Clostridium_beijerinckii*_NCIMB_8052_uid58137	14
*Paenibacillus_mucilaginosus*_3016_uid89377	14
*Paenibacillus_polymyxa*_M1_uid162159	14
*Paenibacillus_polymyxa*_SC2_uid59583	14
*Shewanella_violacea*_DSS12_uid47085	14
*Bacillus_cereus*_ATCC_14579_uid57975	13
*Bacillus_cereus*_E33L_uid58103	13
*Bacillus_cereus*_FRI_35_uid173403	13
*Bacillus_cereus*_G9842_uid58759	13
*Bacillus_cereus*_Q1_uid58529	13
*Bacillus_cytotoxicus*_NVH_391_98_uid58317	13
*Bacillus_thuringiensis*_serovar_chinensis_CT_43_uid158151	13
*Paenibacillus_mucilaginosus*_K02_uid162117	13
*Paenibacillus_mucilaginosus*_KNP414_uid68311	13
*Aliivibrio_salmonicida*_LFI1238_uid59251	12
*Bacillus_cereus*_AH820_uid58751	12
*Bacillus_cereus*_ATCC_10987_uid57673	12
*Bacillus_cereus*_F837_76_uid83611	12
*Bacillus_megaterium*_QM_B1551_uid15862	12
*Bacillus_thuringiensis*_HD_771_uid173374	12
*Clostridium_difficile*_BI1_uid158363	12
*Paenibacillus*_JDR_2_uid59021	12
*Paenibacillus_polymyxa*_E681_uid53477	12

Name	16S copy number
*Shewanella_sediminis*_HAW_EB3_uid58835	12
*Solibacillus_silvestris*_StLB046_uid168516	12
*Vibrio_fischeri*_ES114_uid58163	12
*Azospirillum_brasilense*_Sp245_uid162161	11
*Bacillus_anthracis*__Ames_Ancestor__uid58083	11
*Bacillus_anthracis*_A0248_uid59385	11
*Bacillus_anthracis*_Ames_uid57909	11
*Bacillus_anthracis*_CDC_684_uid59303	11
*Bacillus_anthracis*_Sterne_uid58091	11
*Bacillus_cereus*_biovar_anthracis_CI_uid50615	11
*Bacillus_megaterium*_DSM319_uid48371	11
*Bacillus_megaterium*_WSH_002_uid159841	11
*Bacillus_subtilis*_BSn5_uid62463	11
*Clostridium_acetobutylicum*_ATCC_824_uid57677	11
*Clostridium_acetobutylicum*_DSM_1731_uid68293	11
*Clostridium_acetobutylicum*_EA_2018_uid159515	11
*Clostridium_botulinum*_B_Eklund_17B_uid59159	11
*Clostridium_botulinum*_E3_Alaska_E43_uid59157	11
*Clostridium_difficile*_630_uid57679	11
*Clostridium_difficile*_M120_uid158361	11
*Clostridium_lentocellum*_DSM_5427_uid49117	11
*Desulfosporosinus_meridiei*_DSM_13257_uid75097	11
*Shewanella_pealeana*_ATCC_700345_uid58705	11
*Vibrio*_Ex25_uid41601	11
*Vibrio_harveyi*_ATCC_BAA_1116_uid58957	11
*Vibrio_parahaemolyticus*_BB22OP_uid184822	11
*Vibrio_parahaemolyticus*_RIMD_2210633_uid57969	11
*Aeromonas_hydrophila*_ATCC_7966_uid58617	10
*Aeromonas_veronii*_B565_uid66323	10
*Alkaliphilus_metalliredigens*_QYMF_uid58171	10
*Bacillus_amyloliquefaciens*_DSM_7_uid53535	10
*Bacillus_amyloliquefaciens*_FZB42_uid58271	10
*Bacillus_amyloliquefaciens_plantarum*_AS43_3_uid183682	10
*Bacillus_amyloliquefaciens_plantarum*_CAU_B946_uid84215	10
*Bacillus_amyloliquefaciens_plantarum*_YAU_B9601_Y2_uid159001	10
*Bacillus_amyloliquefaciens*_Y2_uid165195	10
*Bacillus_anthracis*_H9401_uid162021	10
*Bacillus_cellulosilyticus*_DSM_2522_uid43329	10
*Bacillus_coagulans*_36D1_uid54335	10
*Bacillus*_JS_uid162189	10
*Bacillus_subtilis*_168_uid57675	10
*Bacillus_subtilis*_BSP1_uid184010	10
*Bacillus_subtilis*_natto_BEST195_uid183001	10
*Bacillus_subtilis*_QB928_uid173926	10
*Bacillus_subtilis*_RO_NN_1_uid158879	10
*Bacillus_subtilis_spizizenii*_TU_B_10_uid73967	10
*Clostridium_botulinum*_BKT015925_uid66203	10
*Clostridium_difficile*_CD196_uid41017	10
*Clostridium_difficile*_R20291_uid40921	10
*Clostridium_novyi*_NT_uid58643	10
*Clostridium_perfringens*_13_uid57681	10
*Clostridium_perfringens*_SM101_uid58117	10
*Desulfotomaculum_acetoxidans*_DSM_771_uid59109	10
*Geobacillus_HH01*_uid188479	10
*Geobacillus_thermodenitrificans*_NG80_2_uid58829	10

Name	16S copy number
*Geobacillus_thermoglucosidasius*_C56_YS93_uid48129	10
*Geobacillus*_WCH70_uid59045	10
*Geobacillus*_Y4_1MC1_uid55779	10
*Heliobacterium_modesticaldum*_Ice1_uid58279	10
*Lysinibacillus_sphaericus*_C3_41_uid58945	10
*Psychromonas_ingrahamii*_37_uid58521	10
*Shewanella_baltica*_BA175_uid52601	10
*Shewanella_baltica*_OS117_uid162025	10
*Shewanella_baltica*_OS155_uid58259	10
*Shewanella_baltica*_OS185_uid58743	10
*Shewanella_baltica*_OS195_uid58261	10
*Shewanella_baltica*_OS223_uid58775	10
*Shewanella_baltica*_OS678_uid50553	10
*Shewanella_halifaxensis*_HAW_EB4_uid59007	10
*Shewanella_woodyi*_ATCC_51908_uid58721	10
*Vibrio_fischeri*_MJ11_uid58907	10

Table 4. Strains having one copy of 16S rRNA gene in the genome

Name	Domain
*Acidianus_hospitalis*_W1_uid66875	*Archaea*
*Acidilobus_saccharovorans*_345_15_uid51395	*Archaea*
*Aciduliprofundum_boonei*_T469_uid43333	*Archaea*
*Aciduliprofundum*_MAR08_339_uid184407	*Archaea*
*Aeropyrum_pernix*_K1_uid57757	*Archaea*
*Archaeoglobus_fulgidus*_DSM_4304_uid57717	*Archaea*
*Archaeoglobus_profundus*_DSM_5631_uid43493	*Archaea*
*Archaeoglobus_veneficus*_SNP6_uid65269	*Archaea*
*Caldisphaera_lagunensis*_DSM_15908_uid183486	*Archaea*
*Caldivirga_maquilingensis*_IC_167_uid58711	*Archaea*
*Candidatus_Korarchaeum_cryptofilum*_OPF8_uid58601	*Archaea*
*Candidatus_Nitrosopumilus*_AR2_uid176130	*Archaea*
*Candidatus_Nitrosopumilus_koreensis*_AR1_uid176129	*Archaea*
*Candidatus_Nitrososphaera_gargensis*_Ga9_2_uid176707	*Archaea*
*Cenarchaeum_symbiosum*_A_uid61411	*Archaea*
*Desulfurococcus_fermentans*_DSM_16532_uid75119	*Archaea*
*Desulfurococcus_kamchatkensis*_1221n_uid59133	*Archaea*
*Desulfurococcus_mucosus*_DSM_2162_uid62227	*Archaea*
*Ferroglobus_placidus*_DSM_10642_uid40863	*Archaea*
*Fervidicoccus_fontis*_Kam940_uid162201	*Archaea*
*Halalkalicoccus_jeotgali*_B3_uid50305	*Archaea*
*Halobacterium*_NRC_1_uid57769	*Archaea*
*Halobacterium_salinarum*_R1_uid61571	*Archaea*
*Halorhabdus_utahensis*_DSM_12940_uid59189	*Archaea*
*Hyperthermus_butylicus*_DSM_5456_uid57755	*Archaea*
*Ignicoccus_hospitalis*_KIN4_I_uid58365	*Archaea*
*Ignisphaera_aggregans*_DSM_17230_uid51875	*Archaea*
*Metallosphaera_cuprina*_Ar_4_uid66329	*Archaea*

Name	Domain
*Metallosphaera_sedula*_DSM_5348_uid58717	*Archaea*
*Methanoculleus_bourgensis*_MS2_uid171377	*Archaea*
*Methanoculleus_marisnigri*_JR1_uid58561	*Archaea*
*Methanopyrus_kandleri*_AV19_uid57883	*Archaea*
*Methanoregula_boonei*_6A8_uid58815	*Archaea*
*Methanoregula_formicicum*_SMSP_uid184406	*Archaea*
*Methanosaeta_harundinacea*_6Ac_uid81199	*Archaea*
*Nanoarchaeum_equitans*_Kin4_M_uid58009	*Archaea*
*Natronomonas_pharaonis*_DSM_2160_uid58435	*Archaea*
*Nitrosopumilus_maritimus*_SCM1_uid58903	*Archaea*
*Picrophilus_torridus*_DSM_9790_uid58041	*Archaea*
*Pyrobaculum_aerophilum*_IM2_uid57727	*Archaea*
*Pyrobaculum_arsenaticum*_DSM_13514_uid58409	*Archaea*
*Pyrobaculum_calidifontis*_JCM_11548_uid58787	*Archaea*
*Pyrobaculum_islandicum*_DSM_4184_uid58635	*Archaea*
*Pyrobaculum_neutrophilum*_V24Sta_uid58421	*Archaea*
*Pyrobaculum_oguniense*_TE7_uid84411	*Archaea*
*Pyrococcus_abyssi*_GE5_uid62903	*Archaea*
*Pyrococcus_furiosus*_COM1_uid169620	*Archaea*
*Pyrococcus_furiosus*_DSM_3638_uid57873	*Archaea*
*Pyrococcus_horikoshii*_OT3_uid57753	*Archaea*
*Pyrococcus*_NA2_uid66551	*Archaea*
*Pyrococcus*_ST04_uid167261	*Archaea*
*Pyrococcus_yayanosii*_CH1_uid68281	*Archaea*
*Pyrolobus_fumarii*_1A_uid73415	*Archaea*
*Staphylothermus_hellenicus*_DSM_12710_uid45893	*Archaea*
*Staphylothermus_marinus*_F1_uid58719	*Archaea*
*Sulfolobus_acidocaldarius*_DSM_639_uid58379	*Archaea*
*Sulfolobus_acidocaldarius*_N8_uid189027	*Archaea*
*Sulfolobus_acidocaldarius*_Ron12_I_uid189028	*Archaea*
*Sulfolobus_islandicus*_HVE10_4_uid162067	*Archaea*
*Sulfolobus_islandicus*_L_D_8_5_uid43679	*Archaea*
*Sulfolobus_islandicus*_L_S_2_15_uid58871	*Archaea*
*Sulfolobus_islandicus*_M_14_25_uid58849	*Archaea*
*Sulfolobus_islandicus*_M_16_27_uid58851	*Archaea*
*Sulfolobus_islandicus*_M_16_4_uid58841	*Archaea*
*Sulfolobus_islandicus*_REY15A_uid162071	*Archaea*
*Sulfolobus_islandicus*_Y_G_57_14_uid58923	*Archaea*
*Sulfolobus_islandicus*_Y_N_15_51_uid58825	*Archaea*
*Sulfolobus_solfataricus*_98_2_uid167998	*Archaea*
*Sulfolobus_solfataricus*_P2_uid57721	*Archaea*
*Sulfolobus_tokodaii*_7_uid57807	*Archaea*
*Thermococcus*_4557_uid70841	*Archaea*
*Thermococcus*_AM4_uid54735	*Archaea*

Name	Domain
*Thermococcus_barophilus*_MP_uid54733	*Archaea*
*Thermococcus*_CL1_uid168259	*Archaea*
*Thermococcus_gammatolerans*_EJ3_uid59389	*Archaea*
*Thermococcus_kodakarensis*_KOD1_uid58225	*Archaea*
*Thermococcus_onnurineus*_NA1_uid59043	*Archaea*
*Thermococcus_sibiricus*_MM_739_uid59399	*Archaea*
*Thermofilum_pendens*_Hrk_5_uid58563	*Archaea*
*Thermogladius*_1633_uid167488	*Archaea*
*Thermoplasma_acidophilum*_DSM_1728_uid61573	*Archaea*
*Thermoplasma_volcanium*_GSS1_uid57751	*Archaea*
*Thermoproteus_tenax*_Kra_1_uid74443	*Archaea*
*Thermoproteus_uzoniensis*_768_20_uid65089	*Archaea*
*Thermosphaera_aggregans*_DSM_11486_uid48993	*Archaea*
*Vulcanisaeta_distributa*_DSM_14429_uid52827	*Archaea*
*Vulcanisaeta_moutnovskia*_768_28_uid63631	*Archaea*
*Acidobacterium_capsulatum*_ATCC_51196_uid59127	*Bacteria*
*Acidobacterium*_MP5ACTX9_uid50551	*Bacteria*
*Acidothermus_cellulolyticus*_11B_uid58501	*Bacteria*
*alpha_proteobacterium*_HIMB5_uid175779	*Bacteria*
*alpha_proteobacterium*_HIMB59_uid175778	*Bacteria*
*Anaplasma_centrale*_Israel_uid42155	*Bacteria*
*Anaplasma_marginale*_Florida_uid58577	*Bacteria*
*Anaplasma_marginale*_Maries_uid57629	*Bacteria*
*Anaplasma_phagocytophilum*_HZ_uid57951	*Bacteria*
*Atopobium_parvulum*_DSM_20469_uid59195	*Bacteria*
*Blattabacterium__Blaberus_giganteus__*uid165873	*Bacteria*
*Blattabacterium__Blatta_orientalis__*Tarazona_uid188115	*Bacteria*
*Blattabacterium__Blattella_germanica__*Bge_uid41533	*Bacteria*
*Blattabacterium__Cryptocercus_punctulatus__*Cpu_uid81083	*Bacteria*
*Blattabacterium__Mastotermes_darwiniensis__*MADAR_uid77127	*Bacteria*
*Blattabacterium__Periplaneta_americana__*BPLAN_uid41287	*Bacteria*
*Borrelia_afzelii*_HLJ01_uid177930	*Bacteria*
*Borrelia_afzelii*_PKo_uid159867	*Bacteria*
*Borrelia_bissettii*_DN127_uid71231	*Bacteria*
*Borrelia_burgdorferi*_B31_uid57581	*Bacteria*
*Borrelia_burgdorferi*_JD1_uid161197	*Bacteria*
*Borrelia_burgdorferi*_N40_uid161241	*Bacteria*
*Borrelia_burgdorferi*_ZS7_uid59429	*Bacteria*
*Borrelia_crocidurae*_Achema_uid162335	*Bacteria*
*Borrelia_duttonii*_Ly_uid58791	*Bacteria*
*Borrelia_garinii*_BgVir_uid162165	*Bacteria*
*Borrelia_garinii*_NMJW1_uid177081	*Bacteria*
*Borrelia_garinii*_PBi_uid58125	*Bacteria*
*Borrelia_hermsii*_DAH_uid59225	*Bacteria*

Name	Domain
*Borrelia_recurrentis*_A1_uid58793	*Bacteria*
*Borrelia_turicatae*_91E135_uid58311	*Bacteria*
*Brachyspira_hyodysenteriae*_WA1_uid59291	*Bacteria*
*Brachyspira_intermedia*_PWS_A_uid158369	*Bacteria*
*Brachyspira_murdochii*_DSM_12563_uid48819	*Bacteria*
*Brachyspira_pilosicoli*_95_1000_uid50609	*Bacteria*
*Brachyspira_pilosicoli*_B2904_uid175255	*Bacteria*
*Brachyspira_pilosicoli*_P43_6_78_uid184077	*Bacteria*
*Brachyspira_pilosicoli*_WesB_uid175256	*Bacteria*
*Bradyrhizobium_japonicum*_USDA_110_uid57599	*Bacteria*
*Bradyrhizobium*_S23321_uid158167	*Bacteria*
*Buchnera_aphidicola*__Cinara_tujafilina__uid68101	*Bacteria*
*Buchnera_aphidicola*_5A__Acyrthosiphon_pisum__uid59285	*Bacteria*
*Buchnera_aphidicola*_Ak__Acyrthosiphon_kondoi__uid158533	*Bacteria*
*Buchnera_aphidicola*_APS__Acyrthosiphon_pisum__uid57805	*Bacteria*
*Buchnera_aphidicola*_Bp__Baizongia_pistaciae__uid57827	*Bacteria*
*Buchnera_aphidicola*_Cc__Cinara_cedri__uid58579	*Bacteria*
*Buchnera_aphidicola*_JF98__Acyrthosiphon_pisum__uid158845	*Bacteria*
*Buchnera_aphidicola*_JF99__Acyrthosiphon_pisum__uid158847	*Bacteria*
*Buchnera_aphidicola*_LL01__Acyrthosiphon_pisum__uid158843	*Bacteria*
*Buchnera_aphidicola*_Sg__Schizaphis_graminum__uid57913	*Bacteria*
*Buchnera_aphidicola*_TLW03__Acyrthosiphon_pisum__uid158849	*Bacteria*
*Buchnera_aphidicola*_Tuc7__Acyrthosiphon_pisum__uid59283	*Bacteria*
*Buchnera_aphidicola*_Ua__Uroleucon_ambrosiae__uid158535	*Bacteria*
*Caldisericum_exile*_AZM16c01_uid158173	*Bacteria*
*Candidatus_Amoebophilus_asiaticus*_5a2_uid58963	*Bacteria*
*Candidatus_Blochmannia_chromaiodes*_640_uid185308	*Bacteria*
*Candidatus_Blochmannia_floridanus*_uid57999	*Bacteria*
*Candidatus_Blochmannia_pennsylvanicus*_BPEN_uid58329	*Bacteria*
*Candidatus_Blochmannia_vafer*_BVAF_uid62083	*Bacteria*
*Candidatus_Carsonella_ruddii*_CE_isolate_Thao2000_uid172732	*Bacteria*
*Candidatus_Carsonella_ruddii*_CS_isolate_Thao2000_uid172733	*Bacteria*
*Candidatus_Carsonella_ruddii*_HC_isolate_Thao2000_uid172734	*Bacteria*
*Candidatus_Carsonella_ruddii*_HT_isolate_Thao2000_uid172735	*Bacteria*
*Candidatus_Carsonella_ruddii*_PC_isolate_NHV_uid172736	*Bacteria*
*Candidatus_Carsonella_ruddii*_uid58773	*Bacteria*
*Candidatus_Chloracidobacterium_thermophilum*_B_uid73587	*Bacteria*
*Candidatus_Hodgkinia_cicadicola*_Dsem_uid59311	*Bacteria*
*Candidatus_Koribacter_versatilis*_Ellin345_uid58479	*Bacteria*
*Candidatus_Methylomirabilis_oxyfera*_uid161981	*Bacteria*
*Candidatus_Midichloria_mitochondrii*_IricVA_uid68687	*Bacteria*
*Candidatus_Moranella_endobia*_PCIT_uid68739	*Bacteria*
*Candidatus_Mycoplasma_haemolamae*_Purdue_uid171259	*Bacteria*
*Candidatus_Nitrospira_defluvii*_uid51175	*Bacteria*

Name	Domain
*Candidatus_Pelagibacter*_IMCC9063_uid66305	*Bacteria*
*Candidatus_Pelagibacter_ubique*_HTCC1062_uid58401	*Bacteria*
*Candidatus_Portiera_aleyrodidarum*_BT_B_uid173859	*Bacteria*
*Candidatus_Portiera_aleyrodidarum*_BT_B_uid176373	*Bacteria*
*Candidatus_Portiera_aleyrodidarum*_BT_QVLC_uid175570	*Bacteria*
*Candidatus_Portiera_aleyrodidarum*_BT_QVLC_uid176374	*Bacteria*
*Candidatus_Puniceispirillum_marinum*_IMCC1322_uid47081	*Bacteria*
*Candidatus_Rickettsia_amblyommii*_GAT_30V_uid156845	*Bacteria*
*Candidatus_Ruthia_magnifica*_Cm__Calyptogena_magnifica__uid58645	*Bacteria*
*Candidatus_Sulcia_muelleri*_CARI_uid52535	*Bacteria*
*Candidatus_Sulcia_muelleri*_DMIN_uid47075	*Bacteria*
*Candidatus_Sulcia_muelleri*_GWSS_uid58943	*Bacteria*
*Candidatus_Sulcia_muelleri*_SMDSEM_uid59393	*Bacteria*
*Candidatus_Uzinura_diaspidicola*_ASNER_uid186740	*Bacteria*
*Candidatus_Vesicomyosocius_okutanii*_HA_uid59427	*Bacteria*
*Candidatus_Zinderia_insecticola*_CARI_uid52459	*Bacteria*
*Cardinium_endosymbiont_cEper1_of_Encarsia_pergandiella*_uid175524	*Bacteria*
*Chlamydia_psittaci*_01DC12_uid179070	*Bacteria*
*Chlamydia_psittaci*_84_55_uid175571	*Bacteria*
*Chlamydia_psittaci*_GR9_uid175572	*Bacteria*
*Chlamydia_psittaci*_M56_uid175576	*Bacteria*
*Chlamydia_psittaci*_MN_uid175573	*Bacteria*
*Chlamydia_psittaci*_VS225_uid175574	*Bacteria*
*Chlamydia_psittaci*_WC_uid175577	*Bacteria*
*Chlamydia_psittaci*_WS_RT_E30_uid175575	*Bacteria*
*Chlamydophila_abortus*_S26_3_uid57963	*Bacteria*
*Chlamydophila_caviae*_GPIC_uid57783	*Bacteria*
*Chlamydophila_felis*_Fe_C_56_uid57971	*Bacteria*
*Chlamydophila_pecorum*_E58_uid66295	*Bacteria*
*Chlamydophila_pneumoniae*_AR39_uid57809	*Bacteria*
*Chlamydophila_pneumoniae*_CWL029_uid57811	*Bacteria*
*Chlamydophila_pneumoniae*_J138_uid57829	*Bacteria*
*Chlamydophila_pneumoniae*_LPCoLN_uid159529	*Bacteria*
*Chlamydophila_pneumoniae*_TW_183_uid57997	*Bacteria*
*Chlamydophila_psittaci*_01DC11_uid159527	*Bacteria*
*Chlamydophila_psittaci*_02DC15_uid159521	*Bacteria*
*Chlamydophila_psittaci*_08DC60_uid159525	*Bacteria*
*Chlamydophila_psittaci*_6BC_uid159845	*Bacteria*
*Chlamydophila_psittaci*_6BC_uid63621	*Bacteria*
*Chlamydophila_psittaci*_C19_98_uid159523	*Bacteria*
*Chlamydophila_psittaci*_CP3_uid175578	*Bacteria*
*Chlamydophila_psittaci*_Mat116_uid189026	*Bacteria*
*Chlamydophila_psittaci*_NJ1_uid175579	*Bacteria*
*Chlorobium_chlorochromatii*_CaD3_uid58375	*Bacteria*

Name	Domain
*Chlorobium_phaeovibrioides*_DSM_265_uid58129	*Bacteria*
*Chloroherpeton_thalassium*_ATCC_35110_uid59187	*Bacteria*
*Conexibacter_woesei*_DSM_14684_uid43467	*Bacteria*
*Coxiella_burnetii*_CbuG_Q212_uid58893	*Bacteria*
*Coxiella_burnetii*_CbuK_Q154_uid58895	*Bacteria*
*Coxiella_burnetii*_Dugway_5J108_111_uid58629	*Bacteria*
*Coxiella_burnetii*_RSA_331_uid58637	*Bacteria*
*Coxiella_burnetii*_RSA_493_uid57631	*Bacteria*
*Cycloclasticus*_P1_uid176368	*Bacteria*
*Dehalococcoides*_BAV1_uid58477	*Bacteria*
*Dehalococcoides*_CBDB1_uid58413	*Bacteria*
*Dehalococcoides_ethenogenes*_195_uid57763	*Bacteria*
*Dehalococcoides*_GT_uid42115	*Bacteria*
*Dehalococcoides*_VS_uid42393	*Bacteria*
*Dehalogenimonas_lykanthroporepellens*_BL_DC_9_uid48131	*Bacteria*
*Desulfarculus_baarsii*_DSM_2075_uid51371	*Bacteria*
*Desulfobacca_acetoxidans*_DSM_11109_uid65785	*Bacteria*
*Desulfococcus_oleovorans*_Hxd3_uid58777	*Bacteria*
*Desulfomonile_tiedjei*_DSM_6799_uid168320	*Bacteria*
*Ehrlichia_canis*_Jake_uid58071	*Bacteria*
*Ehrlichia_chaffeensis*_Arkansas_uid57933	*Bacteria*
*Ehrlichia_ruminantium*_Gardel_uid58245	*Bacteria*
*Ehrlichia_ruminantium*_Welgevonden_uid58013	*Bacteria*
*Ehrlichia_ruminantium*_Welgevonden_uid58243	*Bacteria*
*Elusimicrobium_minutum*_Pei191_uid58949	*Bacteria*
*Erythrobacter_litoralis*_HTCC2594_uid58299	*Bacteria*
*Francisella_noatunensis_orientalis_Toba*_04_uid164779	*Bacteria*
*Gemmatimonas_aurantiaca*_T_27_uid58813	*Bacteria*
*Gloeobacter_violaceus*_PCC_7421_uid58011	*Bacteria*
*Granulicella_mallensis*_MP5ACTX8_uid49957	*Bacteria*
*Helicobacter_hepaticus*_ATCC_51449_uid57737	*Bacteria*
*Helicobacter_pylori*_2017_uid161151	*Bacteria*
*Helicobacter_pylori*_2018_uid161159	*Bacteria*
*Helicobacter_pylori*_908_uid159985	*Bacteria*
*Hydrogenobacter_thermophilus*_TK_6_uid159875	*Bacteria*
*Hydrogenobacter_thermophilus*_TK_6_uid45927	*Bacteria*
*Hyphomicrobium_denitrificans*_ATCC_51888_uid50325	*Bacteria*
*Hyphomicrobium*_MC1_uid68453	*Bacteria*
*Hyphomonas_neptunium*_ATCC_15444_uid58433	*Bacteria*
*Ignavibacterium_album*_JCM_16511_uid162097	*Bacteria*
*Jannaschia*_CCS1_uid58147	*Bacteria*
*Leifsonia_xyli*_CTCB07_uid57759	*Bacteria*
*Melioribacter_roseus*_P3M_uid170941	*Bacteria*
*Methylacidiphilum_infernorum*_V4_uid59161	*Bacteria*

Name	Domain
*Methylibium_petroleiphilum*_PM1_uid58085	*Bacteria*
*Methylocystis*_SC2_uid174072	*Bacteria*
*Micavibrio_aeruginosavorus*_ARL_13_uid73585	*Bacteria*
*Microlunatus_phosphovorus*_NM_1_uid68055	*Bacteria*
*Moorella_thermoacetica*_ATCC_39073_uid58051	*Bacteria*
*Mycobacterium_abscessus*_ATCC_19977_uid61613	*Bacteria*
*Mycobacterium_africanum*_GM041182_uid68839	*Bacteria*
*Mycobacterium_avium*_104_uid57693	*Bacteria*
*Mycobacterium_avium_paratuberculosis*_K_10_uid57699	*Bacteria*
*Mycobacterium_bovis*_AF2122_97_uid57695	*Bacteria*
*Mycobacterium_bovis*_BCG_Korea_1168P_uid189029	*Bacteria*
*Mycobacterium_bovis*_BCG_Mexico_uid86889	*Bacteria*
*Mycobacterium_bovis*_BCG_Pasteur_1173P2_uid58781	*Bacteria*
*Mycobacterium_bovis*_BCG_Tokyo_172_uid59281	*Bacteria*
*Mycobacterium_canettii*_CIPT_140010059_uid70731	*Bacteria*
*Mycobacterium_canettii*_CIPT_140060008_uid184829	*Bacteria*
*Mycobacterium_canettii*_CIPT_140070008_uid184832	*Bacteria*
*Mycobacterium_canettii*_CIPT_140070010_uid184828	*Bacteria*
*Mycobacterium_canettii*_CIPT_140070017_uid184830	*Bacteria*
*Mycobacterium_indicus_pranii*_MTCC_9506_uid175523	*Bacteria*
*Mycobacterium_intracellulare*_ATCC_13950_uid167994	*Bacteria*
*Mycobacterium_intracellulare*_MOTT_02_uid89387	*Bacteria*
*Mycobacterium_intracellulare*_MOTT_64_uid89385	*Bacteria*
*Mycobacterium_leprae*_Br4923_uid59293	*Bacteria*
*Mycobacterium_leprae*_TN_uid57697	*Bacteria*
*Mycobacterium_liflandii*_128FXT_uid59005	*Bacteria*
*Mycobacterium_marinum*_M_uid59423	*Bacteria*
*Mycobacterium_massiliense*_GO_06_uid170732	*Bacteria*
*Mycobacterium*_MOTT36Y_uid164001	*Bacteria*
*Mycobacterium_tuberculosis*_CCDC5079_uid161943	*Bacteria*
*Mycobacterium_tuberculosis*_CCDC5180_uid161941	*Bacteria*
*Mycobacterium_tuberculosis*_CDC1551_uid57775	*Bacteria*
*Mycobacterium_tuberculosis*_CTRI_2_uid161997	*Bacteria*
*Mycobacterium_tuberculosis*_F11_uid58417	*Bacteria*
*Mycobacterium_tuberculosis*_H37Ra_uid58853	*Bacteria*
*Mycobacterium_tuberculosis*_H37Rv_uid170532	*Bacteria*
*Mycobacterium_tuberculosis*_H37Rv_uid57777	*Bacteria*
*Mycobacterium_tuberculosis*_KZN_1435_uid59069	*Bacteria*
*Mycobacterium_tuberculosis*_KZN_4207_uid83619	*Bacteria*
*Mycobacterium_tuberculosis*_KZN_605_uid54947	*Bacteria*
*Mycobacterium_tuberculosis*_RGTB327_uid157907	*Bacteria*
*Mycobacterium_tuberculosis*_RGTB423_uid162179	*Bacteria*
*Mycobacterium_tuberculosis*_UT205_uid162183	*Bacteria*
*Mycobacterium_ulcerans*_Agy99_uid62939	*Bacteria*

Name	Domain
*Mycoplasma_arthritidis*_158L3_1_uid58005	*Bacteria*
*Mycoplasma_bovis*_Hubei_1_uid68691	*Bacteria*
*Mycoplasma_conjunctivae*_HRC_581_uid59325	*Bacteria*
*Mycoplasma_genitalium*_G37_uid57707	*Bacteria*
*Mycoplasma_genitalium*_M2288_uid173372	*Bacteria*
*Mycoplasma_genitalium*_M2321_uid173373	*Bacteria*
*Mycoplasma_genitalium*_M6282_uid173371	*Bacteria*
*Mycoplasma_genitalium*_M6320_uid173370	*Bacteria*
*Mycoplasma_haemocanis*_Illinois_uid82367	*Bacteria*
*Mycoplasma_haemofelis*_Langford_1_uid62461	*Bacteria*
*Mycoplasma_haemofelis*_Ohio2_uid162029	*Bacteria*
*Mycoplasma_hyopneumoniae*_168_uid162053	*Bacteria*
*Mycoplasma_hyopneumoniae*_232_uid58205	*Bacteria*
*Mycoplasma_hyopneumoniae*_7448_uid58039	*Bacteria*
*Mycoplasma_hyopneumoniae*_J_uid58059	*Bacteria*
*Mycoplasma_hyorhinis*_GDL_1_uid87003	*Bacteria*
*Mycoplasma_hyorhinis*_HUB_1_uid51695	*Bacteria*
*Mycoplasma_hyorhinis*_MCLD_uid162087	*Bacteria*
*Mycoplasma_hyorhinis*_SK76_uid181997	*Bacteria*
*Mycoplasma_mobile*_163K_uid58077	*Bacteria*
*Mycoplasma_penetrans*_HF_2_uid57729	*Bacteria*
*Mycoplasma_pneumoniae*_309_uid85495	*Bacteria*
*Mycoplasma_pneumoniae*_FH_uid162027	*Bacteria*
*Mycoplasma_pneumoniae*_M129_uid57709	*Bacteria*
*Mycoplasma_pulmonis*_UAB_CTIP_uid61569	*Bacteria*
*Mycoplasma_suis*_Illinois_uid61897	*Bacteria*
*Mycoplasma_suis*_KI3806_uid63665	*Bacteria*
*Mycoplasma_wenyonii*_Massachusetts_uid170731	*Bacteria*
*Neorickettsia_risticii*_Illinois_uid58889	*Bacteria*
*Neorickettsia_sennetsu*_Miyayama_uid57965	*Bacteria*
*Nitrobacter_hamburgensis*_X14_uid58293	*Bacteria*
*Nitrobacter_winogradskyi*_Nb_255_uid58295	*Bacteria*
*Nitrosomonas*_AL212_uid55727	*Bacteria*
*Nitrosomonas_europaea*_ATCC_19718_uid57647	*Bacteria*
*Nitrosomonas_eutropha*_C91_uid58363	*Bacteria*
*Nitrosomonas*_Is79A3_uid68745	*Bacteria*
*Nitrosospira_multiformis*_ATCC_25196_uid58361	*Bacteria*
*Oligotropha_carboxidovorans*_OM4_uid162135	*Bacteria*
*Oligotropha_carboxidovorans*_OM5_uid59155	*Bacteria*
*Oligotropha_carboxidovorans*_OM5_uid72795	*Bacteria*
*Olsenella_uli*_DSM_7084_uid51367	*Bacteria*
*Opitutus_terrae*_PB90_1_uid58965	*Bacteria*
*Orientia_tsutsugamushi*_Boryong_uid61621	*Bacteria*
*Orientia_tsutsugamushi*_Ikeda_uid58869	*Bacteria*

Name	Domain
*Parvibaculum_lavamentivorans*_DS_1_uid58739	*Bacteria*
*Parvularcula_bermudensis*_HTCC2503_uid51641	*Bacteria*
*Phenylobacterium_zucineum*_HLK1_uid58959	*Bacteria*
*Phycisphaera_mikurensis*_NBRC_102666_uid157331	*Bacteria*
*Pirellula_staleyi*_DSM_6068_uid43209	*Bacteria*
*Polaromonas*_JS666_uid58207	*Bacteria*
*Polynucleobacter_necessarius*_asymbioticus_QLW_P1DMWA_1_uid58611	*Bacteria*
*Polynucleobacter_necessarius*_STIR1_uid58967	*Bacteria*
*Prochlorococcus_marinus*_AS9601_uid58307	*Bacteria*
*Prochlorococcus_marinus*_CCMP1375_uid57995	*Bacteria*
*Prochlorococcus_marinus*_MIT_9211_uid58309	*Bacteria*
*Prochlorococcus_marinus*_MIT_9215_uid58819	*Bacteria*
*Prochlorococcus_marinus*_MIT_9301_uid58437	*Bacteria*
*Prochlorococcus_marinus*_MIT_9312_uid58357	*Bacteria*
*Prochlorococcus_marinus*_MIT_9515_uid58313	*Bacteria*
*Prochlorococcus_marinus*_NATL1A_uid58423	*Bacteria*
*Prochlorococcus_marinus*_NATL2A_uid58359	*Bacteria*
*Prochlorococcus_marinus*_pastoris_CCMP1986_uid57761	*Bacteria*
*Prosthecochloris_aestuarii*_DSM_271_uid58151	*Bacteria*
*Pseudoxanthomonas_spadix*_BD_a59_uid75113	*Bacteria*
*Ramlibacter_tataouinensis*_TTB310_uid68279	*Bacteria*
*Rhodopirellula_baltica*_SH_1_uid61589	*Bacteria*
*Rhodopseudomonas_palustris*_HaA2_uid58439	*Bacteria*
*Rhodothermus_marinus*_DSM_4252_uid41729	*Bacteria*
*Rhodothermus_marinus*_SG0_5JP17_172_uid72767	*Bacteria*
*Rickettsia_africae*_ESF_5_uid58799	*Bacteria*
*Rickettsia_akari*_Hartford_uid58161	*Bacteria*
*Rickettsia_australis*_Cutlack_uid158039	*Bacteria*
*Rickettsia_bellii*_OSU_85_389_uid58681	*Bacteria*
*Rickettsia_bellii*_RML369_C_uid58405	*Bacteria*
*Rickettsia_canadensis*_CA410_uid88063	*Bacteria*
*Rickettsia_canadensis*_McKiel_uid58159	*Bacteria*
*Rickettsia_conorii*_Malish_7_uid57633	*Bacteria*
*Rickettsia_felis*_URRWXCal2_uid58331	*Bacteria*
*Rickettsia_heilongjiangensis*_054_uid70839	*Bacteria*
*Rickettsia_japonica*_YH_uid73963	*Bacteria*
*Rickettsia_massiliae*_AZT80_uid86751	*Bacteria*
*Rickettsia_massiliae*_MTU5_uid58801	*Bacteria*
*Rickettsia_montanensis*_OSU_85_930_uid158043	*Bacteria*
*Rickettsia_parkeri*_Portsmouth_uid158045	*Bacteria*
*Rickettsia_peacockii*_Rustic_uid59301	*Bacteria*
*Rickettsia_philipii*_364D_uid89383	*Bacteria*
*Rickettsia_prowazekii*_BuV67_CWPP_uid158063	*Bacteria*
*Rickettsia_prowazekii*_Chernikova_uid158053	*Bacteria*

Name	Domain
*Rickettsia_prowazekii*_Dachau_uid158057	*Bacteria*
*Rickettsia_prowazekii*_GvV257_uid158051	*Bacteria*
*Rickettsia_prowazekii*_Katsinyian_uid158055	*Bacteria*
*Rickettsia_prowazekii*_Madrid_E_uid61565	*Bacteria*
*Rickettsia_prowazekii*_Rp22_uid161945	*Bacteria*
*Rickettsia_prowazekii*_RpGvF24_uid158065	*Bacteria*
*Rickettsia_rhipicephali*_3_7_female6_CWPP_uid156977	*Bacteria*
*Rickettsia_rickettsii*__Sheila_Smith__uid58027	*Bacteria*
*Rickettsia_rickettsii*_Arizona_uid86655	*Bacteria*
*Rickettsia_rickettsii*_Brazil_uid88069	*Bacteria*
*Rickettsia_rickettsii*_Colombia_uid86653	*Bacteria*
*Rickettsia_rickettsii*_Hauke_uid86659	*Bacteria*
*Rickettsia_rickettsii*_Hino_uid86657	*Bacteria*
*Rickettsia_rickettsii*_Hlp_2_uid88067	*Bacteria*
*Rickettsia_rickettsii*_Iowa_uid58961	*Bacteria*
*Rickettsia_slovaca*_13_B_uid82369	*Bacteria*
*Rickettsia_slovaca*_D_CWPP_uid158159	*Bacteria*
*Rickettsia_typhi*_B9991CWPP_uid158357	*Bacteria*
*Rickettsia_typhi*_TH1527_uid158161	*Bacteria*
*Rickettsia_typhi*_Wilmington_uid58063	*Bacteria*
*Roseobacter_denitrificans*_OCh_114_uid58597	*Bacteria*
*Roseobacter_litoralis*_Och_149_uid54719	*Bacteria*
*Rubrobacter_xylanophilus*_DSM_9941_uid58057	*Bacteria*
*Salinibacter_ruber*_DSM_13855_uid58513	*Bacteria*
*Salinibacter_ruber*_M8_uid47323	*Bacteria*
*Salmonella_enterica*_serovar_Weltevreden_2007_60_3289_1_uid178014	*Bacteria*
*secondary_endosymbiont_of_Ctenarytaina_eucalypti*_uid172737	*Bacteria*
*secondary_endosymbiont_of_Heteropsylla_cubana*_Thao2000_uid172738	*Bacteria*
*Segniliparus_rotundus*_DSM_44985_uid49049	*Bacteria*
*Serratia_symbiotica*__Cinara_cedri__uid82363	*Bacteria*
*Simkania_negevensis*_Z_uid68451	*Bacteria*
*Sphingopyxis_alaskensis*_RB2256_uid58351	*Bacteria*
*Starkeya_novella*_DSM_506_uid48815	*Bacteria*
*Streptococcus_pseudopneumoniae*_IS7493_uid71153	*Bacteria*
*Synechococcus*_PCC_6312_uid182934	*Bacteria*
*Synechococcus*_RCC307_uid61609	*Bacteria*
*Syntrophus_aciditrophicus*_SB_uid58539	*Bacteria*
*Terriglobus_saanensis*_SP1PR4_uid53251	*Bacteria*
*Thalassobaculum*_L2_uid182483	*Bacteria*
*Thermocrinis_albus*_DSM_14484_uid46231	*Bacteria*
*Thermodesulfobacterium*_OPB45_uid68283	*Bacteria*
*Thermodesulfovibrio_yellowstonii*_DSM_11347_uid59257	*Bacteria*
*Thermosynechococcus_elongatus*_BP_1_uid57907	*Bacteria*
*Thermotoga_lettingae*_TMO_uid58419	*Bacteria*

Name	Domain
*Thermotoga_maritima*_MSB8_uid57723	*Bacteria*
*Thermotoga_naphthophila*_RKU_10_uid42777	*Bacteria*
*Thermotoga_neapolitana*_DSM_4359_uid59065	*Bacteria*
*Thermotoga_petrophila*_RKU_1_uid58655	*Bacteria*
*Thermotoga*_RQ2_uid58935	*Bacteria*
*Thermotoga_thermarum*_DSM_5069_uid68449	*Bacteria*
*Thioalkalivibrio*_K90mix_uid46181	*Bacteria*
*Thioalkalivibrio_nitratireducens*_DSM_14787_uid184011	*Bacteria*
*Thioalkalivibrio_sulfidophilus*_HL_EbGr7_uid59179	*Bacteria*
*Thiomonas*_3As_uid178369	*Bacteria*
*Thiomonas_intermedia*_K12_uid48825	*Bacteria*
*Tropheryma_whipplei*_TW08_27_uid57961	*Bacteria*
*Tropheryma_whipplei*_Twist_uid57705	*Bacteria*
*uncultured_Termite_group_1_bacterium*_phylotype_Rs_D17_uid59059	*Bacteria*
*Wolbachia_endosymbiont_of_Culex_quinquefasciatus*_Pel_uid61645	*Bacteria*
*Wolbachia_endosymbiont_of_Drosophila_melanogaster*_uid57851	*Bacteria*
*Wolbachia_endosymbiont_of_Onchocerca_ochengi*_uid171829	*Bacteria*
*Wolbachia_endosymbiont_TRS_of_Brugia_malayi*_uid58107	*Bacteria*
*Wolbachia_wRi*_uid59371	*Bacteria*