Pyrosequence-based 16S rRNA profiling has become a very powerful tool for visualizing the community structure of gastro-intestinal (GI) tract microbiota. The system was established with newly designed universal primers with barcode sequences and newly modified algorithms to convert batch sequence data to bacterial population data. *In silico* primer match simulation indicates that the primers, Q-968F, Q-1046R, and Q-1390R, match to almost all 16S rRNAs in the database within one base mismatch, with especially high coverage ratios for the four biggest common phyla in human GI tract. Also, the new SeqmatchQ100 algorithm correctly assigns almost all of the target 60-base sequences of the 16S rRNA V6 region to the corresponding genus except for those from the *Enterobacteriaceae* family and the *Enterococcus* genus. Furthermore, the SeqmatchQ400 algorithm efficiently provides species-level population data from a 400-base sequence of the 16S rRNA V6–V8 region with the exceptions of the *Enterobacteriaceae* and *Enterococcaceae* families. A barcode-sequence tag strategy was used to analyze up to 128 samples at a time. With these newly prepared tools, pyrosequence-based 16S rRNA profiling displays community structures of GI-tract microbiota. For instance, establishment of bifidus flora in newborn infants and dynamics in the microbial community structure after weaning were effectively demonstrated by 16S rRNA profiling. In future, this analytical system should be of use for monitoring changes in GI-tract bacterial composition which may be influenced by diets, drugs, or sickness.

Key words: intestinal microbiota; pyrosequence; 16S rRNA; infant; *Bifidobacterium*

**INTRODUCTION**

In the mid 1990s, several molecular techniques for profiling bacterial community structure, which were mostly based on 16S rRNA sequence diversity and phylogeny, were introduced to intestinal microbiology. Notably, fingerprinting methods such as denaturing gradient gel electrophoresis (DGGE) and terminal-restriction fragment length polymorphism (T-RFLP) have become very popular due to their ease of use (14). At the same time, the database of 16S rRNA sequences of the gastro-intestinal (GI) tract bacteria, including uncultured members, has been increasing day by day and it is now supposed to cover the majority of common GI tract members (2, 4). As well as DGGE and T-RFLP, sequence-based composition analysis, that is, random sequencing of 16S rRNA gene (16S rDNA) clone libraries, has often been used to analyze bacterial community structure, which provides us more accurate and reliable taxonomic information based on 16S rRNA sequence and phylogeny (5). However, this procedure is constrained by its cost and time, since more than several hundred clones must be sequenced just to gain the information of percent-level population groups. Recent developments in sequencing technology have broken through these hurdles. Pyrosequencing has become one of the most promising techniques applicable for 16S rRNA-based bacterial composition analysis (9). Pyrosequencing monitors the luminescence of sequencing reactions in a pico-titer plate, each well of which includes a bead capturing millions of copies of clonal DNA fragments previously amplified by an oil emulsion PCR (10). This single molecule-based and cloning-independent approach enables ultra-high throughput sequencing analysis, providing up to 350,000 reads of 100 bases per run in the initial version and 1,000,000 reads of 400 bases per run in the latest version. In this review, pyrosequencing-based structure analysis of gastro-intestinal microbiota, which has been recently developed by our group at Kyushu University, is introduced with its methodology, strategy, detailed protocol and application examples.
GENERAL SCHEME OF PYROSEQUENCE-BASED 16S rRNA PROFILING OF GI-TRACT MICROBIOTA WITH A BARCODE-SEQUENCE TAG STRATEGY

Figure 1 shows a flowchart of the pyrosequence-based 16S rRNA profiling of gastrointestinal microbiota. As in DGGE, T-RFLP and some other molecular-based microbial composition analyses, it begins with isolation of total bacterial DNA from a sample, e.g., feces or some other GI-tract-related sample, then amplifies a certain region of 16S rDNA from the DNA mixture, eventually analyzing the composition of amplicons obtained. A unique point in the pyrosequencing-based approach is to mix amplicons from different samples into one batch. After the batch sequencing, collected sequences are sorted back into a sequence pool for each original sample by the barcode-sequence tags attached to both ends of the PCR fragment. Thus far, we have designed two sets of 48-barcode sequence tags and 128-barcode sequence tags, as shown in Fig. 2, which allows bacterial composition analysis of 48 and 128 samples at a time, respectively. After obtaining the sample sequence data, each sequence is searched against a database to acquire taxonomic information for the 16S rDNA fragment and the number of read sequences belonging to each bacterial group is counted. Finally, bacterial composition is profiled based on the count of each bacterial group. The details of the process determining population data from batch sequence data will be explained below.

Fig. 1. Flowchart of pyrosequence-based 16S rRNA profiling of gastrointestinal microbiota.

Fig. 2. Design of universal primers for pyrosequence-based 16S rRNA profiling of GI tract microbiota.
DESIGN OF UNIVERSAL PRIMERS

In the process of the pyrosequence-based 16S rRNA profiling, PCR amplification of 16S rDNA is one of the most critical steps. In order to gain bacterial composition data with high accuracy, it is crucial to avoid any bias in the PCR amplification, even though it is considerably difficult to amplify all 16S rDNAs with the same efficiency from a sample containing a large variety of bacteria. In addition, in order to acquire valuable taxonomic information from the limited length of readable sequences (up to approximately 400 bases in the latest version of pyrosequencer), it is necessary to amplify a segment with highly variable sequences, such as the V4 or V6 region. Taken together, a hypervariable region surrounded by highly conserved regions should be targeted for the PCR-pyrosequence approach. In our studies, V6 and V6–V8 regions were chosen as the first and second targets, respectively, and a series of bacterial universal primers were designed to match as many bacterial 16S rRNAs in the database as possible (Q-968F, Q-1046R and Q-1390R were modified from the prototype primers A-967F and B-1046R, and 1392r, respectively) (6, 11). Figure 2 shows the list of primers with the result of in silico primer matching to 16S rRNAs of four common phyla of GI tract bacteria. The primer matching was performed by using the Probe Match tool on the RDP-II website (http://rdp.cme.msu.edu/probematch/search.jsp). As shown in Fig. 2, the designed primers are highly matched to most of the 16S rRNAs in the database, especially the four big phyla in the GI tract. Taken together, we consider Q-968F/Q-1046R and Q-968F/Q-1390R quite valid as universal primer sets for the 16S rRNA profiling of GI-tract microbiota.

In our system, the barcode sequence was added to the 5'-end of both primers, which enables sorting from both directions. As in the case of normal PCR primer design, all tagged primers were designed not to have successive G/C sequences which cause primer misannealing and the non-formation of the primer dimer at the annealing temperature. In the 48-barcode set, polybases, e.g., TTTT or TTTTT, were not avoided, which results in misreads of the number of polybases at low but significantly higher frequency than normal sequencing errors. Thus, polybases were avoided in the 128-barcodes of the second version. Furthermore, in the second version, an extra base, H (A, T, or C), indicated by an asterisk in Fig. 2, was added to the 3'-end of the barcode region to further confirm the barcode assignment.

PATHS FROM BATCH SEQUENCE DATA TO POPULATION DATA

Figure 3 is a flow diagram of the conversion of batch sequence data obtained by pyrosequencing to bacterial population data. There are several paths in the process. In the initial step, trimming of primer sequences is commonly performed. In the following step, there are
three paths, Path 1 to 3, as shown in Fig. 3.

**Path 1**

The batch-trimmed sequences are sorted into sample sequence data. Then, each sequence in each sample sequence data set is subjected to taxon classification (RDP classifier) or a similar sequence finder algorithm (Blast or Seqmatch). RDP classifier is available at the RDP-II website, and can conventionally convert batch sequence data to bacterial population data at any hierarchal level from genus to phylum. For Blast searches, the local blast with the downloaded dataset of 16S rRNA sequences would be of use, in which each sample sequence is subjected to the similarity search to the local 16S rRNA database. Seqmatch searches can be used to find the closest 16S rRNA in the database as well. The taxonomic information of the closest 16S rRNA is downloaded and used to make bacterial composition profiles. Since Seqmatch has been demonstrated to be more accurate than Blast searches (3) and is available online through the RDP-II website (http://pyro.cme.msu.edu/), RDP Seqmatch is being used more than Blast at our laboratory. How to create population data from the results of Seqmatch searches will be explained in the following section.

**Path 2**

In Path 2, the sample sequence data is dereplicated by clustering highly similar sequences. The clustering is generally carried out by the complete linkage clustering method based on the calculation of cluster distance following calculation of individual sequence pair distances. In the pyrosequencing pipeline at RDP-II website (http://pyro.cme.msu.edu/), up to 500,000 sequences can be clustered at a time through the process of batch sequence alignment. At the end of clustering, one representative sequence is generated for each cluster. The representative sequences are subjected to a similarity search as in Path 1 and a bacterial composition profile is obtained. The DNA distance level for clustering is variable and depends on the objective of the clustering, which can be to make clusters of exactly identical sequences (DNA distance=0), ones within reads which may contain a small number of sequencing errors (DNA distance=0.01–0.02), ones within putative species (DNA distance=0.03 to 0.04), and ones within higher-order taxonomic groups, e.g., genus to phylum (DNA distance >0.06). Since the sequence data obtained by pyrosequencing is not accurate enough to distinguish one or two sequence errors from microheterogeniety in 16S rRNA sequences, DNA distances ≤0.02 are not very efficient to make valuable cluster. If this kind of fine clustering is necessary, it is necessary to filter out the low quality sequences before clustering.

**Path 3**

Sequence clustering is performed with batch sequence data before sorting and the representative sequence of each cluster is subjected to a sequence similarity search as in the other tracks. As a result of these procedures, taxonomic information for each cluster is catalogued. Sample sequence data is prepared as in Path 1. The sequences in the sample sequence data are classified according to the catalogue and the bacterial composition of each sample is profiled.

**Pros and Cons**

Each path has both advantages and disadvantages. In terms of the preciseness of the bacterial composition data obtained, Path 1 is the best because other paths pass through the process of dereplication in the clustering, which might diminish important sequence information critical for taxonomy. Especially when distinguishing closely-related species, e.g., *Lactobacillus gasseri* and *Lactobacillus johnsonii*, Path 1 should be followed. However, many more sequences are subjected to the similarity search in Path 1 than on the other paths. For example, it may take more than ten minutes to get a result for 1,000 sequences of 400 bases in the RDP Seqmatch. Path 3 is useful for cataloging the inhabiting bacterial members, in addition to gaining bacterial population data of each sample. A drawback of Path 3 is that it is necessary to restart from the initial clustering with new sequences and it is tiresomeness to combine new data to existing data. On the other hand in Paths 1 and 2, it is easy because population analysis is carried out individually for each sample. To choose whether to follow Path 1 or 2, it is important to consider the objective of the population profiling. For information regarding bacterial composition with the names of bacteria, Path 1 would be appropriate. On the other hand, when bacterial community structure based on 16S rRNA sequence phylogeny is more important than taxonomy, Path 2 would be appropriate. However, in our experience, the three paths show only small differences in the resulting data.

**ALGORITHMS TO CONVERT SAMPLE SEQUENCE DATA TO POPULATION DATA**

**SeqmatchQ100**

As mentioned above, the V6 region is hypervariable and surrounded by highly conserved regions, which
should be of great use for the PCR-based 16S rRNA profiling for visualizing the bacterial community structure of an ecological sample. Indeed, several papers have already reported valuable findings for bacterial community structures obtained by V6 sequence-based 16S rRNA profiling (11, 13). However, it was reported that the V6 short sequence is not enough to gain accurate and confident classification via the common classifier algorithms, such as RDP classifier or RDP Seqmatch (12). Therefore, we added some extra steps to the original Seqmatch algorithm and the new algorithm was named SeqmatchQ100. The flow diagram of the SeqmatchQ100 algorithm is shown in Fig. 4. To evaluate this algorithm, simulations using the RDP database were performed. When all sequences of type strains (4,856 sequences) deposited in the RDP database were subjected to SeqmatchQ100, 4,853 sequences corresponding to 99.96% of total queries were classified to their correct phylum. The coverage percentage for the correct classification was 96.18%, 91.55%, and 85.78% for order, family, and genus levels, respectively. Furthermore, the same simulation was performed with the sequences of all isolates of GI-tract-related bacterial groups including type and non-type strains and the results were compared to those obtained with RDP classifier (Fig. 4B). SeqmatchQ100 showed higher accuracy than RDP classifier in the genus classification of all bacterial groups except for genus Enterococcus and family Enterobacteriaceae which includes Escherichia coli. Since Enterobacteriaceae and Enterococcus are important groups of GI-tract bacteria, it is necessary to use these two algorithms to gain a correct overview of the community structure.

**SeqmatchQ400**

In the latest version of pyrosequencer, FLX454 titanium, around 400 bases can be read at a single pass, which can cover the V6–V8 region surrounded by highly conserved regions available for the annealing positions of universal primers. The sequences of the V6–V8 region are much more informative than that of the V6 region alone and enable more accurate and precise taxonomic classification. Even though one base difference can distinguish phylotypes in principle, the fact that the 16S rRNA sequences are not always completely conserved within the species sometimes makes the species-level identification difficult. In order to overcome this problem, we have added extra steps to the original RDP Seqmatch algorithm, in which the species showing the highest matching score is assigned to the query sequence, and if more than two different species show the same highest score, the one showing the highest counts in the

![Fig. 4. SeqmatchQ100 to convert sample sequence data of 16S rRNA V6 region to population data at the genus level. (A) Flow diagram of SeqmatchQ100. (B) Comparison of genus classification accuracy between SeqmatchQ100 (black bar) and RDP classifier (grey bar). 16S rRNA sequences of isolated strains belonging to the indicated bacterial group were downloaded from the RDP database and their V6 sequences corresponding to the E. coli 16S rRNA position no. 986 to no. 1045 were subjected to both algorithms. Percentages of correctly identified strains per query strain are graphed. The numbers in parenthesis indicate the number of correctly identified strains per query strain in the SeqmatchQ100 simulation.](image-url)
Fig. 5. SeqmatchQ400 to convert sample sequence data of 16S rRNA of the V6–V8 region to population data of bacterial species. (A) Flow diagram of SeqmatchQ400. (B) Simulation of genus and species identification by SeqmatchQ400. 16S rRNA sequences of isolates belonging to the indicated bacterial group were downloaded from the RDP database and their V6–V8 sequences corresponding to the E. coli 16S rRNA position no. 986 to no. 1391 were subjected to the SeqmatchQ400 algorithm to identify their genus or species. Self-matching between the query and database which must give the highest score of $S_{ab} = 1.0$ was eliminated from this simulation. Percentages of correctly identified strains per query are graphed.

Table 1. Simulation of species identification of *Bifidobacterium* strains by SeqmatchQ400 algorithm with 16S rDNA V6–V8 sequences

| Genus               | Species                          | no. of input strains | no. of correctly identified strains | hit (%) | comment                              |
|---------------------|----------------------------------|----------------------|-------------------------------------|---------|--------------------------------------|
| *Bifidobacterium*   | *animalis*                       | 22                   | 22                                  | 100     | includes subsp. *lactis* and subsp. *animalis* |
| *Bifidobacterium*   | *bifidum*                        | 9                    | 9                                   | 100     |                                     |
| *Bifidobacterium*   | *thermophilum*                   | 6                    | 6                                   | 100     |                                     |
| *Bifidobacterium*   | *catenulatum*                    | 3                    | 3                                   | 100     |                                     |
| *Bifidobacterium*   | *minimum*                        | 3                    | 3                                   | 100     |                                     |
| *Bifidobacterium*   | *asteroides*                     | 3                    | 3                                   | 100     |                                     |
| *Bifidobacterium*   | *coryneforme*                    | 3                    | 3                                   | 100     |                                     |
| *Bifidobacterium*   | *scardovii*                      | 2                    | 2                                   | 100     |                                     |
| *Bifidobacterium*   | *mongoliense*                    | 2                    | 2                                   | 100     |                                     |
| *Bifidobacterium*   | *breve*                          | 18                   | 17                                  | 94      |                                     |
| *Bifidobacterium*   | *adolescentis*                   | 11                   | 10                                  | 91      |                                     |
| *Bifidobacterium*   | *pseudolongum*                   | 8                    | 7                                   | 88      | includes subsp. *pseudolongum* and globosum |
| *Bifidobacterium*   | *longum* subsp *longum*          | 21                   | 18                                  | 86      |                                     |
| *Bifidobacterium*   | *longum* subsp *infantis*        | 10                   | 8                                   | 80      |                                     |
| *Bifidobacterium*   | *subtile*                        | 3                    | 2                                   | 67      |                                     |
| *Bifidobacterium*   | *pseudocatenulatum*              | 2                    | 1                                   | 50      |                                     |
| *Bifidobacterium*   | *dentium*                        | 2                    | 1                                   | 50      |                                     |
| *Bifidobacterium*   | *magnus*                         | 2                    | 1                                   | 50      |                                     |
| *Bifidobacterium*   | *saeculare*                      | 2                    | 1                                   | 50      |                                     |
| *Bifidobacterium*   | *cuniculi*                       | 2                    | 1                                   | 50      |                                     |
| *Bifidobacterium*   | *longum* subsp *suis*            | 2                    | 1                                   | 50      |                                     |
| *Bifidobacterium*   | *crudilactis*                    | 3                    | 0                                   | 0       |                                     |
| *Bifidobacterium*   | *psychraerophilum*               | 2                    | 0                                   | 0       |                                     |
| *Bifidobacterium*   | *indicum*                        | 2                    | 0                                   | 0       |                                     |
| **total**           |                                  | 143                  | 121                                 | 85      |                                     |
Table 2. Species identification of *Lactobacillus* strains by seqmatchQ algorithm with 16S rDNA V6–V8 sequences

| Genus               | Species                  | no. of input strains | no. of correctly identified strains | hit (%) | comment                        |
|---------------------|--------------------------|----------------------|-------------------------------------|---------|---------------------------------|
| *Lactobacillus*     | reuteri                  | 44                   | 44                                  | 100     |                                 |
|                     | salivarius               | 40                   | 40                                  | 100     |                                 |
|                     | crispatus                | 24                   | 24                                  | 100     |                                 |
|                     | kefiranofaciens          | 24                   | 24                                  | 100     |                                 |
|                     | johnsonii                | 20                   | 20                                  | 100     |                                 |
|                     | gassei                   | 20                   | 20                                  | 100     |                                 |
|                     | vaccinostercus           | 15                   | 15                                  | 100     |                                 |
|                     | kefiri                   | 12                   | 12                                  | 100     |                                 |
|                     | mucosae                  | 12                   | 12                                  | 100     |                                 |
| *Lactobacillus*     | heligardii               | 11                   | 11                                  | 100     |                                 |
|                     | aviarus                  | 10                   | 10                                  | 100     |                                 |
| *Lactobacillus*     | parabuchneri             | 10                   | 10                                  | 100     |                                 |
|                     | amylovorus               | 10                   | 10                                  | 100     |                                 |
|                     | murinus                  | 9                    | 9                                   | 100     |                                 |
| *Lactobacillus*     | paracollinoides          | 9                    | 9                                   | 100     |                                 |
| *Lactobacillus*     | cruyrau                  | 9                    | 9                                   | 100     |                                 |
|                     | vaginalis                | 8                    | 8                                   | 100     |                                 |
| *Lactobacillus*     | rossiae                  | 7                    | 7                                   | 100     |                                 |
| *Lactobacillus*     | sanfranciscensis         | 6                    | 6                                   | 100     |                                 |
| *Lactobacillus*     | pontis                   | 4                    | 4                                   | 100     |                                 |
| *Lactobacillus*     | gallinarum               | 4                    | 4                                   | 100     |                                 |
| *Lactobacillus*     | kitasatonis              | 4                    | 4                                   | 100     |                                 |
| *Lactobacillus*     | lindneri                 | 4                    | 4                                   | 100     |                                 |
| *Lactobacillus*     | intestinalis             | 3                    | 3                                   | 100     |                                 |
| *Lactobacillus*     | coryniformis             | 3                    | 3                                   | 100     |                                 |
| *Lactobacillus*     | panis                    | 2                    | 2                                   | 100     |                                 |
| *Lactobacillus*     | manihotivorans           | 2                    | 2                                   | 100     |                                 |
| *Lactobacillus*     | acetotolerans            | 2                    | 2                                   | 100     |                                 |
| *Lactobacillus*     | ruminis                  | 2                    | 2                                   | 100     |                                 |
| *Lactobacillus*     | zymae                    | 2                    | 2                                   | 100     |                                 |
| *Lactobacillus*     | homohiochii              | 2                    | 2                                   | 100     |                                 |
| *Lactobacillus*     | namurensis               | 2                    | 2                                   | 100     |                                 |
| *Lactobacillus*     | plantarum                | 283                  | 282                                 | 100     |                                 |
| *Lactobacillus*     | helveticus               | 308                  | 305                                 | 99      |                                 |
| *Lactobacillus*     | casei                    | 180                  | 178                                 | 99      |                                 |
| *Lactobacillus*     | fermentum                | 263                  | 260                                 | 99      |                                 |
| *Lactobacillus*     | brevis                   | 73                   | 72                                  | 99      |                                 |
| *Lactobacillus*     | delbrueckii              | 69                   | 68                                  | 99      |                                 |
| *Lactobacillus*     | sakei                    | 60                   | 59                                  | 98      |                                 |
| *Lactobacillus*     | farcininis               | 4                    | 3                                   | 75      |                                 |
| *Lactobacillus*     | mali                     | 4                    | 3                                   | 75      |                                 |
| *Lactobacillus*     | paralimentarius          | 7                    | 5                                   | 71      |                                 |
| *Lactobacillus*     | acidophilus              | 17                   | 11                                  | 65      |                                 |
| *Lactobacillus*     | paracasei                | 60                   | 5                                   | 8       | mostly identified as *L. casei* |
| *Lactobacillus*     | rhamnosus                | 25                   | 1                                   | 4       | mostly identified as *L. casei* |
| *Lactobacillus*     | pentosus                 | 61                   | 2                                   | 3       | mostly identified as *L. plantarum* |
| *Lactobacillus*     | buchneri                 | 7                    | 0                                   | 0       | identified as *L. parabuchneri* |
| *Lactobacillus*     | curvatus                 | 7                    | 0                                   | 0       | identified as *L. sake*         |
| *Lactobacillus*     | sunkii                   | 6                    | 0                                   | 0       | identified as *L. parabuchneri* |
| *Lactobacillus*     | paraplantarum            | 5                    | 0                                   | 0       | identified as *L. plantarum*    |
| *Lactobacillus*     | animalis                 | 5                    | 0                                   | 0       | identified as *L. murinus*      |
| *Lactobacillus*     | zeae                     | 4                    | 0                                   | 0       | mostly identified as *L. casei* |
| *Lactobacillus*     | parakefiri               | 3                    | 0                                   | 0       | identified as *L. parabuchneri* |
| *Lactobacillus*     | fractivorans             | 2                    | 0                                   | 0       | identified as *L. homohiochii*  |
| *Lactobacillus*     | kimchii                  | 2                    | 0                                   | 0       | identified as *L. paralimentarius* |
| *Lactobacillus*     | alimentarius             | 2                    | 0                                   | 0       | identified as *L. paralimentarius* |
| others              |                          | 106                  | 78                                  | 74      |                                 |
| total               |                          | 1899                 | 1668                                | 88      |                                 |
top 20 list is selected (Fig. 5A). Figure 5B and Tables 1 and 2 show the results of simulation of species and genus identification by SeqmatchQ400. In this simulation, 16S rRNA sequences of all isolates of the GI-tract-related bacterial group downloaded from the RDP database were subjected to this newly developed algorithm. Except for the families Enterococcaceae and Enterobacteriaceae, more than 90% of strains were correctly identified at genus level and 80% at species level. This result suggests that the modified algorithm enhanced the accuracy of species identification, especially for those species located in the taxonomic clade which share highly similar 16S rRNA sequences, e.g., the pair of Bifidobacterium longum subsp. longum and B. longum subsp. infantis, or the pair of Lactobacillus gasseri and L. johnsonii. However, we note that there are some bacterial groups which cannot be differentiated at the species level only by using the V6–V8 sequences, e.g., the group of Lactobacillus casei or Lactobacillus plantarum group, as shown in Table 2. Simulations, as reported in Tables 1 and 2, are important to check the accuracy of identification of each species.

**EXPERIMENTAL PROCEDURE**

Figure 6 shows an experimental protocol for the pyrosequencing-based bacterial composition analysis of GI tract microbiota. It begins with DNA extraction, which is most critical step that would be the key to successfully gain accurate population data. Bacterial cells are lysed by beating with zirconium beads and bacterial DNA is purified by PCR inhibitor absorbent and spin column in a Qiagen Stool DNA purification kit. The purified DNA is used as the template in the following two-step PCR. In the 1st PCR, the V6 or V6–V8 fragment of 16S rDNA is amplified with primers without the barcode tag and in the 2nd PCR, the sample tag is attached to the PCR using primers with the barcode sequences. In these PCR steps, it is important to use as few PCR cycles as possible in order to avoid amplification saturation causing a distortion of the original population of the 16S rDNAs. In order to avoid primer misannealing before starting PCR, hot-start Taq polymerase is used. The amplicon is purified, and then its concentration is measured precisely. In our laboratory, a microphotometer, NanoDrop 2000, enabling measurement with one microliter of solution, is used. In this step, it is also important to ensure the pool contains exactly equal amounts of amplicons from 48 or 128 different samples. The amplicon pool is subjected to an oil emulsion PCR to make millions of copies of clonal DNA fragment which are captured by a fine bead. The beads are spread onto pico-titer plate, one bead per well, and the pyrosequence reaction is performed in each well to generate sequential luminescence which is monitored.
by a CCD camera. In our laboratory, 1/4 region of one-million-well titer plate is used for one experiment analyzing 128 samples, in which at least more than 1,000 reads per sample are expected to be obtained.

The batch sequence data from the pyrosequencer is subjected to 'pipeline initial process' available at the RDP website (http://pyro.cme.msu.edu/) to generate sample sequence data. The sample sequence data is then subjected to RDP classifier and RDP Seqmatch, both also available at the RDP website (http://rdp.cme.msu.edu/). For RDP classifier, the confidence threshold is set to 50% for the V6 short sequences and 70% for the V6–V8 sequences. For RDP Seqmatch, the search is performed against the 16S rRNA sequence database of isolates. For each query sequence, the 20 highest matches are downloaded. The result is downloaded as text data, imported to an excel file and then subjected to the remaining processes of SeqmatchQ100 or SeqmatchQ400. Finally, population data is obtained for the genus and species levels, respectively.

APPLICATION EXAMPLES

Observation of changing microbiota with aging

The pyrosequencing-based bacterial composition analyses of GI-tract microbiota have been performed with fecal samples from infants and adults. The data obtained by Path 1 with SeqmatchQ400 based on the V6–V8 sequences are shown in Figs. 7 to 10 and Table 3. The result of RDP classifier indicates the bacterial composition in all the hierarchy levels of taxonomy from phylum to genus. Figure 7(A) shows the changes in phylum composition of feces from different ages. There was a peak of the abundance of Actinobacteria at the age of three months, at which the microbiota was mostly occupied by this phylum. Before and after this period, phylum Firmicutes constituted the major microbiota as well as Actinobacteria. The bacterial composition data of the lower taxonomy levels indicate that the major bacterial group of phylum Firmicutes before three months of age was class Bacilli, e.g., Streptococcus and after that it was class Clostridia, which is a non-monophyletic anaerobic group of Firmicutes. Figure 7(B) and (C) shows the population changes of the genus Bifidobacterium and the class Clostridia, respectively. These two plots clearly show that bifidus flora was established in the first three months but drastically declined due to the increase of clostridia after weaning. It is also noted that the abundance of Actinobacteria in adult feces was significantly higher than in infant feces after weaning. Species level analysis, described in section 3 below, revealed this was mostly due to the colonization of

**Fig. 7.** Changes in fecal microbiota with aging. The data were obtained from bacterial composition analysis of RDP classifier with 16S rRNA V6–V8 sequences obtained by pyrosequencing. (A) Changes in phylum composition. The data are the averages of six infants (w1 to y3) and ten adults. (B) Changes in the relative abundance of the genus Bifidobacterium. The data of six subjects were individually graphed. (C) Changes in the relative abundance of the class Clostridia. The data of six subjects were individually graphed.

**Fig. 8.** Bacterial species diversity and richness in infant and adult feces. Species diversity was determined as Shannon’s index (H') calculated using the species distribution in each sample and species richness was determined as the number of species found. The species composition data were obtained by pyrosequence-based 16S rRNA profiling using the V6–V8 sequence.

*Bifidobacterium adolescentis* that is an adult-type *Bifidobacterium* species.

Figure 8 shows the changes in bacterial species diversity and richness with the aging. The species
Richness is represented by the number of species detected in the pyrosequence-based 16S rRNA profiling, and the species diversity was calculated as the Shannon-Wiener’s index (http://pyro.cme.msu.edu/chaol/form.spr). Both indexes displayed the same changing pattern, in which the lowest value occurred at three months of age and drastically increased after weaning. The diversity and richness at three years old were similar to those of adults, suggesting that the microbiota is stabilized within a couple of years after the weaning and becomes more like adult flora than infants’. The lowest values at three months of age reflect the establishment of bifidus flora and their increase after weaning indicates the decay of bifidus flora. Both diversity and richness at three years old were similar to those of adults.

Distance between the V6–V8 sequences from the GI tract and database

Figure 9 shows the distribution of matching scores (S_ab) of the V6–V8 sequences obtained by pyrosequencing to the closest isolates in the RDP database. Although the threshold of the S_ab value for the species identification cannot be clearly determined, a value lower than 0.90 suggests less chance to belong to the known species. As shown in Fig. 9(A), in the case of newborn babies, approximately 80% of total queries showed S_ab higher than 0.95, suggesting that bacteria colonizing with newborns mostly belong to known cultured species. However, with aging, the population with S_ab lower than 0.90 increased, suggesting that a significant portion of bacteria inhabiting aged humans are as yet uncultured species. Figure 9(B) shows profiling of the rat cecum. It is interesting that there is a big fraction in the range between 0.70 and 0.75. This fraction mostly consists of one operational taxonomic unit showing close but low similarity to the known species, Allobaculum stercoricanis. This kind of S_ab profiling is very suggestive and helpful for checking the distribution of uncultured fractions. However, it should be noted that the fractions with low S_ab scores more or less come from PCR chimera or low quality sequences. It is also important to check chimera and the sequence quality of reads showing low S_ab scores.

Species identified in infant and adult feces

Table 3 and Fig. 10 show a list of identified species and their distribution, respectively, in feces of infants and adults determined by pyrosequencing-based microbiota analysis. As shown in Fig. 10(A), in all infants at the aged three months, more than 90% of microbiota consisted of Bifidobacterium species, indicating that the bifidus flora is established in these subjects. As shown in Table 3(A), the Bifidobacterium species showing the highest counts in our subjects before weaning was B. bifidum. This disagrees with the data in some previous reports showing B. breve as the most dominant Bifidobacterium, and B. bifidum as a subdominant species, in infants in Japan (1, 7, 8). This feature might be specific to our subjects who were born in a rural area in Japan. The following dominant Bifidobacterium species, B. breve, B. longum subsp. longum, and B. pseudocatenulatum, were commonly present in all six tested subjects. The fifth
dominant *Bifidobacterium* species, *B. longum* subsp. *infantis*, was detected in five out of six subjects tested but constituted less than 1% of the total population, except for one subject for whom approximately 25% of fecal biota corresponded to this species. After weaning, *B. bifidum* and *B. breve* dramatically decreased in most subjects, suggesting that these species are highly oriented to the colonic environment of infants taking only milk nutrients. On the other hand, *B. pseudocatenulatum* and *B. longum* subsp. *longum* were still present to some extent in our infant subjects after the weaning and also in adult subjects. In our adult subjects, *Bifidobacterium adolescentis* was detected in nine out ten subjects as the most dominant species, whereas it was rarely present in our infant subjects even after weaning. This suggests that *B. adolescentis* has preference for the environment of the adult colon.

As shown in Fig. 7(B) and (C), the majority of microbiota was changed from *Bifidobacterium* to class *Clostridia* after weaning. There seems to be an antagonism between these two bacterial groups, which may be a causal relation. The most and second most dominant species in our infant subjects after the weaning at the aged one to three years were *Faecalibacterium prausnitzii* and *Blautia wexlerae*, which belong to class *Clostridia*. In addition to these two dominant species, there were a number of species belonging to class *Clostridia*, which included genus *Clostridium*, *Ruminococcus* and *Eubacterium*, appearing in this period. There is a big question about the function of these *Clostridia* in the intestine, as to whether they are of benefit or burden to the host. In addition, it is important to know the causal relationships among food change, *Bifidobacterium* decrease, *Clostridia* increase, and host maturation.

**CONCLUSION**

With the development of DNA sequencing technology, large scale sequencing of 16S rRNA from the GI-tract bacterial community has enabled accurate bacterial composition analysis. The barcode-sequence tag strategy will upgrade easily to higher multiple versions, e.g., 256 tags or 1024 tags from the present 128-barcode system. One of the advantages of the pyrosequencing-based microbiota analysis is that it allows digital-base analysis using nucleotide sequence data which provide much more reliable taxonomic data than other fingerprinting methods such as DGGE and T-RFLP. For instance, only one base of difference in the 16S rRNA fragment allows differentiation of closely-related species. With the increase of the 16S rRNA sequence database, taxonomical identification of GI-tract-inhabiting
bacteria will become more reliable. However, it will be important to maintain a high quality database free of low quality sequence data or wrong taxonomic information, which may lead to miss-assignment of 16S rRNA reads.

It is also notable that there are still large portion of GI tract microbiota which belong to uncultured groups. It is important to carefully consider the matching score in the results of sequence match analyses, in which $S_{ab} < 0.90$

### Table 3. Top 20 bacterial species detected in feces by pyrosequence-based 16S rDNA profiling

#### (A) Infants before weaning

| Rank | Genus             | Species                     | $S_{ab}$ | Count*   | Prevalence in 6 subjects |
|------|-------------------|-----------------------------|---------|----------|--------------------------|
|      |                   |                             |         |          | w1  m1  m3               |
| 1    | *Bifidobacterium* | *bifidum*                   | 1.000   | 4602     | 4  4  6                  |
| 2    | *Bifidobacterium* | *breve*                     | 1.000   | 2750     | 2  5  6                  |
| 3    | *Bifidobacterium* | *longum subsp. longum*      | 1.000   | 1849     | 3  6  6                  |
| 4    | *Bifidobacterium* | *pseudocatenulatum*         | 1.000   | 1380     | 3  6  4                  |
| 5    | *Clostridium*     | *perfringens*               | 1.000   | 506      | 2  2  0                  |
| 6    | *Streptococcus*   | *salivarius*                | 1.000   | 442      | 6  4  4                  |
| 7    | *Streptococcus*   | *thermophilus*              | 0.980   | 380      | 4  4  2                  |
| 8    | *Enterococcus*    | *aerogenes*                 | 1.000   | 327      | 1  1  1                  |
| 9    | *Bifidobacterium* | *longum subsp. infantis*    | 1.000   | 264      | 2  5  5                  |
| 10   | *Staphylococcus*  | *epidermidis*               | 1.000   | 211      | 6  2  2                  |
| 11   | *Raoultella*      | *ornithinolytica*           | 0.985   | 204      | 1  3  1                  |
| 12   | *Leuconostoc*     | *citreum*                   | 1.000   | 162      | 2  1  0                  |
| 13   | *Staphylococcus*  | *hominis*                   | 0.990   | 134      | 2  3  1                  |
| 14   | *Staphylococcus*  | *aureus*                    | 1.000   | 96       | 3  4  4                  |
| 15   | *Salmonella*      | *enterica*                  | 1.000   | 86       | 1  2  4                  |
| 16   | *Enterococcus*    | *faecalis*                  | 1.000   | 79       | 3  3  5                  |
| 17   | *Collinsella*     | *aerofaciens*               | 0.977   | 62       | 2  1  1                  |
| 18   | *Veillonella*     | *ratti*                     | 0.985   | 46       | 1  1  1                  |
| 19   | *Lactobacillus*   | *gasseri*                   | 1.000   | 44       | 0  2  0                  |
| 20   | *Veillonella*     | *parvula*                   | 0.972   | 42       | 2  2  1                  |

* The number of reads showing highest similarity to 16S rRNA of the indicated species. In total, 14,400 reads (800 reads per sample) were used for this profiling and the species with the twenty highest counts are listed.

#### (B) Infants after weaning

| Rank | Genus             | Species                     | $S_{ab}$ | Count*   | Prevalence in 6 subjects |
|------|-------------------|-----------------------------|---------|----------|--------------------------|
|      |                   |                             |         |          | y1  y1  y3               |
| 1    | *Faecalibacterium*| *prausnitzii*               | 1.000   | 3475     | 6  6  6                  |
| 2    | *Blautia*         | *wexlerae*                  | 1.000   | 1519     | 4  6  6                  |
| 3    | *Bifidobacterium* | *pseudocatenulatum*         | 1.000   | 1012     | 3  5  5                  |
| 4    | *Bifidobacterium* | *longum subsp. longum*      | 1.000   | 633      | 6  6  6                  |
| 5    | *Bifidobacterium* | *breve*                     | 0.987   | 615      | 6  3  1                  |
| 6    | *Clostridium*     | *brettii*                   | 1.000   | 442      | 6  5  6                  |
| 7    | *Ruminococcus*    | *gnarus*                    | 0.970   | 341      | 3  6  5                  |
| 8    | *Bifidobacterium* | *bifidum*                   | 0.982   | 329      | 4  2  2                  |
| 9    | *Ruminococcus*    | *obeum*                     | 1.000   | 327      | 4  5  6                  |
| 10   | *Ruminococcus*    | *bromii*                    | 1.000   | 316      | 3  5  4                  |
| 11   | *Eubacterium*     | *elignis*                   | 1.000   | 265      | 2  3  4                  |
| 12   | *Clostridium*     | *clostridioforme*           | 0.987   | 221      | 4  4  6                  |
| 13   | *Clostridium*     | *indolis*                   | 0.835   | 207      | 3  4  6                  |
| 14   | *Eubacterium*     | *rectale*                   | 1.000   | 206      | 2  3  4                  |
| 15   | *Clostridium*     | *nixile*                    | 0.944   | 185      | 3  6  4                  |
| 16   | *Bacteroides*     | *ovatus*                    | 1.000   | 174      | 2  3  3                  |
| 17   | *Clostridium*     | *glycolicium*               | 1.000   | 162      | 3  2  4                  |
| 18   | *Dysgonomonas*    | *capnocytophagoides*        | 0.774   | 153      | 1  0  1                  |
| 19   | *Bacteroides*     | *vulgatus*                  | 1.000   | 149      | 3  4  2                  |
| 20   | *Clostridium*     | *ramosum*                   | 1.000   | 146      | 6  6  6                  |

* The number of reads showing highest similarity to 16S rRNA of the indicated species. In total, 14,400 reads (800 reads per sample) were used for this profiling and the species with the twenty highest counts are listed.
indicates that the 16S rRNA sequence may be different from any species in the database. Indeed, our fecal data indicate that a significant portion corresponds to a putative uncultured fraction. In terms of the above points, 16S rRNA sequence-based microbiota analysis is advantageous for obtaining phylogenetic information of 16S rRNA even for the uncultured bacterial group.

Furthermore, the digital population data obtained by the sequence-based analysis are useful for deriving a series of mathematical simulation data for the microbial community structure, e.g., species richness or species diversity. These kinds of simulated data are of use in expressing the changes in community structure. With these features, the pyrosequencing-based microbiota analysis should be of use to monitor the impact on GI tract microbiota of diets, drugs, or host health condition, that would be important themes in GI tract bacteriology.

However, it should be noted that the pyrosequencing-based method is applicable only to dominant groups whose relative abundance is higher than 0.1 to 1.0%. Also, since the data obtained by the sequence-based method is a relative ratio not the absolute population, combination with target specific quantitative real-time PCR method should yield more quantitative data with a much lower detection limit.

**Acknowledgements.** I thank Mr. Naoshige Sakamoto, Miss Mikiko Teshima, and Miss Takako Kobayashi for their contribution to this research. I also thank all subjects who provided the faecal samples. This research was supported by a grant-in-aid for Exploratory Research from the Japan Society for the Promotion of Science, a grant for Research for Promoting Technological Seeds from the Japan Science and Technology Agency, and grants from the Morinaga Hoshikai, the Mishima Kaiun Memorial Foundation, and the Nisshin Seifun Foundation.

**REFERENCES**

(1) Benno Y, Sawada K, Mitsuoka T. 1984. The intestinal microflora of infants: composition of fecal flora in breast-fed and bottle-fed infants. Microbiol Immunol 28: 975–986.

(2) Blaut M, Collins MD, Welling GW, Dore J, van Loo J, de Vos W. 2002. Molecular biological methods for studying the gut microbiota: the EU human gut flora project. Br J Nutr 87 Suppl 2: S203–S211.

(3) Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM. 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res 33: D294–D296.

(4) Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. 2005. Diversity of the human intestinal microbial flora. Science 308: 1635–1638.

(5) Hayashi H, Sakamoto M, Kitahara M, Benno Y. 2003.

| Rank | Genus          | Species                | S_ab | Count | Prevalence in 10 subjects |
|------|----------------|------------------------|------|-------|--------------------------|
| 1    | Bifidobacterium | adolescentis           | 1.00 | 1471  | 9                        |
| 2    | Blautia        | wexlerae               | 1.000| 900   | 9                        |
| 3    | Bifidobacterium | pseudocatenulatum      | 1.000| 616   | 9                        |
| 4    | Faecalibacterium | prausnitzii           | 1.000| 605   | 10                       |
| 5    | Collinsella    | aerofaciens            | 1.000| 471   | 9                        |
| 6    | Bifidobacterium | longum subsp. longum   | 1.000| 456   | 10                       |
| 7    | Blautia        | luti                   | 0.914| 334   | 10                       |
| 8    | Ruminococcus   | obeum                  | 0.990| 322   | 10                       |
| 9    | Dorea          | longicatena            | 1.000| 292   | 9                        |
| 10   | Clostridium    | indolis                | 0.820| 163   | 7                        |
| 11   | Eubacterium    | rectale                | 1.000| 117   | 9                        |
| 12   | Clostridium    | fimetarium             | 0.837| 116   | 8                        |
| 13   | Ruminococcus   | gnavus                 | 0.970| 110   | 7                        |
| 14   | Megamonas      | funiformis             | 0.983| 103   | 2                        |
| 15   | Eubacterium    | contortum              | 0.827| 82    | 7                        |
| 16   | Ruminococcus   | torques                | 1.000| 77    | 9                        |
| 17   | Bifidobacterium | longum-subsp-suis      | 0.930| 76    | 7                        |
| 18   | Robisoniella   | peoriensis             | 0.845| 75    | 9                        |
| 19   | Clostridium    | clostridiforme         | 1.000| 71    | 9                        |
| 20   | Bifidobacterium | bifidum               | 1.000| 66    | 5                        |

* The number of reads showing highest similarity to 16S rRNA of the indicated species. In total, 8,000 reads (800 reads per sample) were used for this profiling and the species with the twenty highest counts are listed.
Molecular analysis of fecal microbiota in elderly individuals using 16S rDNA library and T-RFLP. Microbiol Immunol 47: 557–570.

(6) Lane DJ. 1991. 16S/23S rRNA sequencing, In Nucleic acid techniques in bacterial systematics, Stackebrandt E, Goodfellow M (eds), John Wiley & Sons, Chichester, United Kingdom.

(7) Matsuki T, Watanabe K, Tanaka R, Fukuda M, Oyaizu H. 1999. Distribution of bifidobacterial species in human intestinal microflora examined with 16S rRNA-gene-targeted species-specific primers. Appl Environ Microbiol 65: 4506–4512.

(8) Mikami K, Takahashi H, Kimura M, Itozaki M, Izuchi K, Shibata R, Sudo N, Matsumoto H, Koga Y. 2009. Influence of maternal bifidobacteria on the establishment of bifidobacteria colonizing the gut in infants. Pediatr Res 65: 669–674.

(9) Petrosino J, Highlander S, Luna RA, Gibbs, RA, Versalovic J. 2009. Metagenomic pyrosequencing and microbial identification. Clin Chem 55: 856–866.

(10) Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat Biotechnol 26: 1135–1145.

(11) Sogin ML, Morrison HG, Huber JA, Welch MD, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proc Natl Acad Sci USA 103: 12115–12120.

(12) Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73: 5261–5267.

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

(14) Zoetendal EG, Collier CT, Koike S, Mackie RI, Gaskins HR. 2004. Molecular ecological analysis of the gastrointestinal microbiota: a review. J Nutr 134: 465–472.