Bioscience of Microbiota, Food and Health Vol. 32 (2), 69–76, 2013

Up to Species-level Community Analysis of Human Gut Microbiota by 16S rRNA Amplicon Pyrosequencing

Jiro NAKAYAMA[1*], Jiahui JIANG[1], Koichi WATANABE[2], Kangting CHEN[3], Huang NINXIN[3], Kazunori MATSUDA[2], Takashi KURAKAWA[2], Hirokazu TSUJI[2], Kenji SONOMOTO[1] and Yuan-Kun LEE[3]

[1]Laboratory of Microbial Technology, Division of Applied Molecular Microbiology and Biomass Chemistry, Department of Bioscience and Biotechnology, Faculty of Agriculture, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

[2]Yakult Central Institute for Microbiological Research, 1796 Yaho, Kunitachi, Tokyo 186-8659, Japan

[3] Department of Microbiology, Yong Loo Lin School of Medicine, National University of Singapore, 5 Science Drive 2, Singapore 117597

Received September 14, 2012; Accepted October 30, 2012

Pyrosequencing-based 16S rRNA profiling has become a common powerful tool to obtain the community structure of gastrointestinal tract microbiota, but it is still hard to process the massive amount of sequence data into microbial composition data, especially at the species level. Here we propose a new approach in combining the quantitative insights into microbial ecology (QIIME), Mothur and ribosomal database project (RDP) programs to efficiently process 454 pyrosequence data to bacterial composition data up to the species level. It was demonstrated to precisely convert batch sequence data of 16S rRNA V6-V8 amplicons obtained from adult Singaporean fecal samples to taxonomically annotated biota data.

Key words: pyrosequence, 16S rRNA gene, human gut microbiota

Pyrosequence-based 16S rRNA profiling has become a very powerful tool to obtain the structure of complex gut microbiota. However, in most pipelines established already, the 16S rRNA sequence data are processed into operational taxonomic units (OTUs) instead of species units [1–3]. This is reasonable because the partial sequence determined by a pyrosequencer is not enough to identify the species with high confidence, but OTU processing on partial 16S rRNA sequences often loses valuable microbiological information apart from systematic biology. For example, *Bifidobacterium longum* subsp. *longum* and *Bifidobacterium longum* subsp. *infantis*, which are classified into two distinct taxa as well as biotypes, are grouped in the same OTU when their partial 16S rRNA sequences are clustered with an identity of 97%. Therefore, we have proposed our unique approach using RDP seqmatch followed by the SeqmatchQ400 program in which pyrosequenced 16S rRNA sequences are directly subjected to a database search one by one to find the closest species [4]. An *in silico* demonstration of this approach with V6-V8 sequences in the database showed that the species of approximately more than 80% of the strains in the database are correctly identified. Furthermore, the above-mentioned two subspecies of *B. longum* could be identified correctly for more than 80% of the strains in the database. However, this approach is hard to apply for massive data analysis because it requires redundant sequence search. Thus, here we propose a new scheme, as shown in Fig. 1, in which a prepared non-redundant 16S rRNA sequence set is uploaded instead of whole sequences. In this study with this approach, 16S rRNA V6-V8 amplicon sequences obtained by pyrosequencing could be precisely and efficiently converted to bacterial composition data annotated with taxomic information from the phylum to species levels (partially up to the subspecies level) as described below.

Fecal samples were collected from 28 healthy adult Singaporeans aged 20 to 25 years. Feces were collected in individual sterile Faeces Containers (Sarstedt, Nümbrecht, Germany) containing 2 mL RNA*later*® (Ambion, Inc., Austin, TX, USA) and stored at room temperature, taken to the laboratory within 12 hr, and then stored at 4°C until DNA extraction. The Ethics Committee of the National University of Singapore provided ethical clearance for this microbiological research study in accordance with the Helsinki Declaration. Bacterial DNA was extracted from samples by the bead-beating method and purified as described previously [5]. The V6-V8 regions of bacterial 16S rRNA genes were amplified by PCR with a bacterial universal primer set, Q-968F-# (5′-CWSWSWWSHTWACGCGARGAACCTTACC-3′)

*Corresponding author. Mailing address: Jiro Nakayama, Laboratory of Microbial Technology, Division of Applied Molecular Microbiology and Biomass Chemistry, Department of Bioscience and Biotechnology, Faculty of Agriculture, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan. Phone: +81 92-642-3020. E-mail: nakayama@agr.kyushu-u.ac.jp

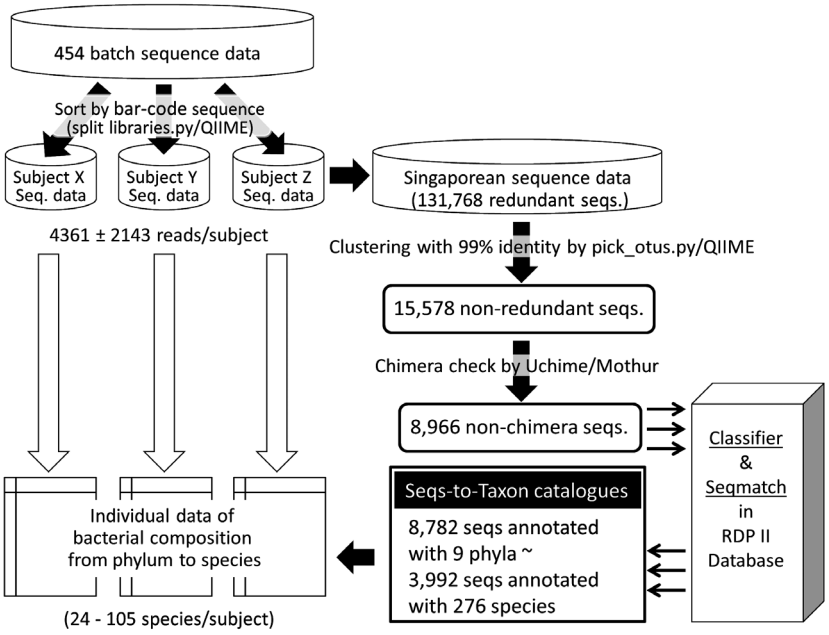

Fig. 1. Flowchart of computational processing of 454 pyrosequence data to individual microbiota data.

and Q-1390R-# (5′-CWSWSWWSHTTGACGGGC GGTGWGTAC-3′) (# indicates a series of 128 barcode sequence tags underlined in the sequence). The PCR was performed in a 50 µL volume containing 10 ng to 100 ng extracted DNA as a template, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 200 µM deoxynucleoside triphosphate (dNTP) mixture, 10 pmol of each primer and 1.25 U TaKaRa Ex *Taq* HS (Takara Bio, Otsu, Shiga, Japan). The PCR condition was as follows: 98°C for 2.5 min; 20 cycles at 98°C for 15 sec, 54°C for 30 sec, and 72°C for 20 sec; and finally 72°C for 5 min. The PCR products were purified using a QIAquick PCR Purification Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's protocol. The purified products were quantified using a NanoDrop ND-1000 microphotometer (NanoDrop Technologies, Wilmington, DE, USA). After that, equal amounts (100 ng) of the amplicons from different samples were pooled and purified prior to pyrosequencing using the ethanol precipitation method. The amplicon mixture DNAs were clonally amplified by emulsion PCR (emPCR) with GS FLX Titanium LV emPCR Kit (Lib-L) v2 according to manufacturer's protocol (454 Life Sciences / Roche Diagnostics). Beads with amplified DNA were loaded onto a GS FLX Titanium PicoTiterPlate with dividers with separate reaction chambers to accommodate two mixture pools. Sequencing was carried out using an FLX Genome Sequencer (454 Life Sciences) with GS FLX Titanium Sequencing Kit XLR70 according to the manufacturer's protocol (454 Life Sciences).

The obtained 454 batch sequence data were sorted into each sample batch by using the QIIME split_library.py script (http://qiime.org/scripts/split_libraries.html) with the barcode sequences. The parameters used in this script were as follows: l (minimum sequence length) = 408, e (maximum number of errors in barcode) = 0, reverse primer mismatches = 3, a (maximum number of ambiguous bases) = 3, L (maximum sequence length) = 500. As a result, 131,768 reads were assigned to the 28 subjects, in which the average ± standard deviation of the number of reads per sample was 4361 ± 2143.

To prepare a nonredundant sequence set, the 131,768 reads were dereplicated within 99% nucleotide sequence identity by using the pick_otus_through_otu_table.py script of QIIME (http://qiime.org/scripts/pick_otus_through_otu_table.html). As a result, a set of 15,578 non-redundant sequences was obtained. Then, these non-redundant sequences were subjected to a PCR chimera check using the Chimera.uchime program in Mothur 1.25.1 (http://www.mothur.org/wiki/Download_mothur), which is one of the fastest chimera check program at present [6–8]. The chimera check was performed using a known 16S rRNA sequence dataset, gg_97_otus_4feb2011.fasta (http://greengenes.lbl.gov/cgi-bin/

nph-index.cgi) as a reference [database (DB) Uchime], following a check in de novo (de novo Uchime). As shown in Fig. 2, a dense cluster was observed in the region where the scores of both database (DB) Uchime and de novo Uchime were higher than approximately 0.5. Frequent sequences read more than 100 times were surrounded by square lines (171 sequencs). Some of these sequences showed scores of more than 0.3, which is a default cut off of this program. In principal, frequently read sequences must have less possibility to be chimeras because PCR chimeras are incidentally generated by misannealing between template and incomplete products during the reaction cycles. Therefore, nine frequent sequences that were read more than 100 times but showed scores of more than 0.3 in both de novo Uchime and DB uchime were subjected to a BLAST search to determine whether these sequences were really chimeras or not. As a result, eight sequences showed 100% identity to 16S rRNA sequences of uncultured organism clones in the NCBI database, and the rest showed 99.5% indentity, suggesting that these were not chimeras. Thus, the default cutoff score of 0.3 was considered to be too strict and was shifted to 0.6 for both de novo and DB Uchime to let most of the frequent sequences be classified as nonchimeras. SG.6_C1696, which had 696 read counts but was still categorized as a chimera was exceptionally removed from the chimera lists; approximately a thousand sequences in the database that came from uncultured fecal bacterial clones showed more than 98% identity to this sequence, suggesting that this sequences represented one of the yet-to-be-cultured members in the human gut. As a result of the chimera check, 6,612 sequences corresponding to 9,671 reads were recognized as PCR chimeras, and the other 8,966 sequences corresponding to 122,097 reads were selected for further analysis as nonchimera sequences.

In order to get taxonomic information, the selected 8,966 sequences were subjected to two web-based searches, RDP Classifier [9] and RDP Seqmatch [10] (http://rdp.cme.msu.edu/). In the RDP Classifier search, the cut off value of the confidence threshold for taxonomic classification was set to be 80% as generally recommended. The number of sequences identified to any known taxa and their total read counts are summarized in Table 1. In the taxonomic level higher than family, almost all sequences were identified to known taxa, and more than 90% of the sequences corresponding to 96.9% of the total nonchimera reads were done at the family level. At the genus level, the identified rate remarkably decreased to 65.5%, while approximately 80% of the nonchimera sequence reads were identified to a known genus.

RDP Seqmatch was employed to gain species

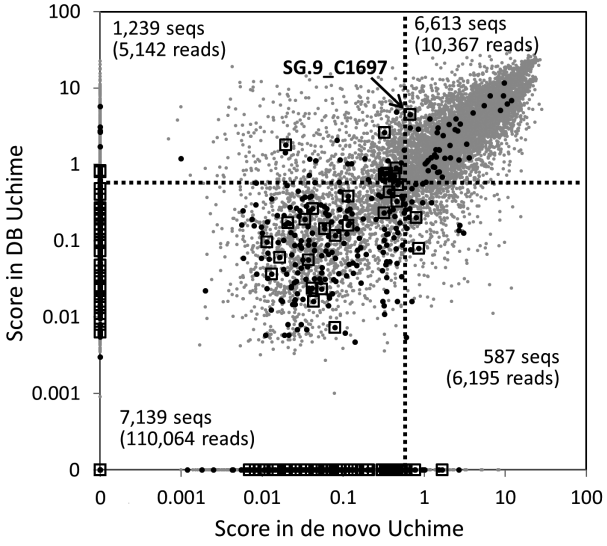

Fig. 2. Result of a chimera check of 15,578 nonredundant sequences. The scores of each sequence in de novo Uchime and DB Uchime were two-dimensionally plotted. Grey dots represent sequences with read counts of less than 10. Black dots represent sequences with the read counts of more than 9, and sequences with more than 99 counts are surrounded by a square. Vertical and horizontal dotted lines show the cut off threshold (score = 0.6) used for the chimera detection except in the case of SG.9_C1697.

level community data. However, less than 50% of the sequences were matched to known species with an S_ab score of more than 0.9, which is the theoretical threshold for species identification. When RDP Seqmatch was performed against the 16S rRNA database including the sequences of uncultured organisms in addition to those of the isolated bacteria, more than 80% of the sequences showed significant identity to 16S rRNA in the database (data not shown). This result suggests that the majority of the unidentified sequences came from yet-to-be cultured bacteria.

Indeed, the RDP Seqmatch search was performed to find the 20 closest 16S rRNA sequences of the cultured strains, and if more than 2 different species showed the same highest score, the one with the highest count in the top 20 list was selected. This algorithm was processed by an excel macro file named SeqmatchQ400 [4]. As a result of the RDP Seqmatch search followed by processing with SeqmatchQ400, a total of 276 species were identified in our Singaporean subjects. The Chao1 value was calculated by using alpha_diversity.py script of QIIME (http://qiime.org/scripts/alpha_diversity.html), which is based on a rarefaction curve [11] of the species frequency. As a result, the total number of described species existing

Table 1. The number and percentage of sequences and reads identified to known taxa*

	Nonchimera[a]	Phylum[b]	Class[b]	Order[b]	Family[b]	Genus[b]	Species[c]
Identified taxa #	-	9	17	25	47	107	276
Sequence #	8,966	8,782 (97.9)	8,689 (96.9)	8,677 (96.8)	8,309 (92.7)	5,869 (65.5)	3,992 (44.5)
Read #	122,097	120,305 (98.5)	119,666 (98.0)	119,651 (98.0)	118,267 (96.9)	95,822 (78.5)	95,112 (77.9)

*The values in parentheses represent percentages of the number of nonchimera OTUs or reads.
[a] Nonchimeras were selected by de novo Uchime (cut-off score = 0.6) and DB Uchime (cut-off score = 0.6) searches except in the cases of SG.6_C1697, which was recognized as a nonchimera by BLAST and RDP Seqmatch searches.
[b] Identification to known taxonomic groups at these hierarchy levels was performed by RDP Classifier with a confidence threshold of 80%.
[c] Identification to known species was performed by RDP Seqmatch with an S_ab score higher than 0.9.

in our Singaporean subjects was estimated to be 315 species. The averaged number of species detected in each individual was 78, and the average Chao1 value of each individual was 117. These numbers are comparative to those previously found by OTU-based estimation using full-length 16S rRNA gene sequences [12]. Among the 276 species, 53 species were commonly detected in more than a half of our subjects (Table 2). This number is also comparative to the data obtained by a metagenomic sequencing, in which 75 species were commonly detected in more than a half of 124 European individuals at the 1% level of genome coverage [13]. As shown in Table 2, 26 species were shared between the common species of our study and the European metagenomic study.

As a final step, individual biota data were profiled based on the catalogue of the 8,782 nonredundant sequences taxonomically annotated. Relative abundance of each taxonomic group was calculated by dividing the read count of identified sequences by the total read count in individuals. Figure 3 shows the population distribution of common genera, families and phyla. Four phyla, *Firmicutes*, *Actinobacteria*, *Bacteroidetes* and *Proteobacteria,* were detected in all subjects. *Firmicutes* mainly consists of two dominant families, *Lachnospiraceae* and *Ruminococcaceae*, and was the most dominant phylum, occupying more than 50% of the total population in most subjects. The next dominant phyla were *Actinobacteria,* which mainly consists of genus *Bifidobacterium,* and *Bacteroidetes,* which mainly consists of genus *Bacteroides.* As shown in Table 2, *Faecalibacterium prausnitzii,* which belongs to the *Ruminococcaceae* family, was the most dominant species and detected in all subjects as in the European metagenomic study [13]. *Bifidobacterium adolescentis*, *B. longum* subsp. *longum* and *B. pseudocatenulatum* were found to be dominant and ranked within the top 10 species in the genus *Bifidobacterium*. This is different from the European metagenomic study in which no *Bifidobacterium* species were ranked within the top 10, suggesting the difference in gut bacterial community between Asian and European individuals.

To examine the quantitative accuracy, the 454 population data of two dominant genera, *Bacteroides* and *Bifidobacterium*, were compared with those measured by quantitative real-time PCR. As shown in Fig. 4, these two groups of data were consistent with each other even though *Bacteroides* was more sensitive and *Bifidobacterium* was less sensitive in the pyrosequencing-based analysis than in quantitative real-time PCR. This is probably due to the difference in the copy numbers of 16S rRNA genes on the chromosome; the average among 16 *Bifidobacterium* species in the rrNDB database (http://rrndb.mmg.msu.edu/search.php) is 3.56, while that among 11 *Bacteroides* species is 6.0.

In conclusion, the newly designed approach combining the QIIME, Mothur and RDP programs precisely and efficiently converted the 454 pyrosequence data to bacterial composition data up to the species level. Details of the whole process are summarized in Table 3. The total processing time was less than 12 hr, and the largest amount of time was spent on the DB Uchime and RDP Seqmatch searches. Ongoing rapid fulfillment of 16S rRNA sequence database should be going to compensate the blank of unidentified bacterial groups appeared in our process. The output data annotated with known bacterial taxa from phylum to species (subspecies in part) should supply missing information more relevant to microbiology in addition to systematic understanding of the microbial community structure realized by the well-established OTU-based profiling technology.

ACKNOWLEDGEMENTS

This research was supported by Grants-in-aid from the Yakult Bio-Science Foundation (to J.N.), for Kyushu University Interdisciplinary Programs in Education and Projects in Research Development, for Exploratory Research from the Japan Society for the Promotion of Science (to J.N.), and for a Research for Promoting Technological Seeds from the Japan Science and Technology Agency (to J.N.) and by a Department of Microbiology Education Grant from the

Table 2. Species commonly detected in more than a half of the 28 Singaporean subjects

Rank[a]	Species	Abundance[b] (%)	Carriers[c]	Metagenome rank[d]
1	*Faecalibacterium prausnitzii*	9.27	28	1
2	*Bacteroides vulgatus*	5.17	23	4
3	*Collinsella aerofaciens*	4.90	22	20
4	*Bifidobacterium adolescentis*	4.90	25	31
5	*Clostridium clostridioforme*	4.01	28	NL
6	*Bifidobacterium longum* subsp *longum*	3.66	27	NL
7	*Bifidobacterium pseudocatenulatum*	2.72	22	61
8	*Eubacterium hadrum*	2.32	27	NL
9	*Eubacterium rectale*	2.04	23	11
10	*Bacteroides dorei*	1.95	22	27
11	*Gemmiger formicilis*	1.72	20	NL
12	*Bifidobacterium stercoris*	1.70	20	NL
13	*Roseburia faecis*	1.67	22	NL
14	*Blautia wexlerae*	1.57	27	NL
15	*Bacteroides uniformis*	1.56	28	6
16	*Ruminococcus bromii*	1.35	14	33
17	*Dorea longicatena*	1.27	21	13
18	*Ruminococcus torques*	1.09	27	17
19	*Ruminococcus obeum*	0.91	26	30
20	*Ruminococcus gnavus*	0.84	21	64
21	*Escherichia coli*	0.76	19	55
22	*Clostridium mayombei*	0.73	26	NL
23	*Bacteroides ovatus*	0.70	25	23
24	*Megasphaera elsdenii*	0.67	14	NL
25	*Blautia luti*	0.64	23	NL
26	*Roseburia inulinivorans*	0.62	23	NL
27	*Parabacteroides distasonis*	0.61	27	21
28	*Streptococcus salivarius*	0.58	27	NL
29	*Eubacterium eligens*	0.56	20	NL
30	*Lachnospira pectinoschiza*	0.55	22	NL
31	*Bacteroides fragilis*	0.44	16	40
32	*Dorea formicigenerans*	0.37	23	3
33	*Clostridium nexile*	0.36	17	69
34	*Parabacteroides merdae*	0.31	17	28
35	*Phascolarctobacterium faecium*	0.30	22	NL
36	*Alistipes putredinis*	0.29	15	19
37	*Bacteroides caccae*	0.28	17	32
38	*Bacteroides thetaiotaomicron*	0.25	28	26
39	*Clostridium bartlettii*	0.21	17	53
40	*Klebsiella pneumoniae*	0.19	16	NL
41	*Sutterella wadsworthensis*	0.18	15	NL
42	*Flavonifractor plautii*	0.17	25	NL
43	*Coprococcus catus*	0.17	17	NL
44	*Eubacterium ventriosum*	0.14	20	35
45	*Streptococcus parasanguinis*	0.14	23	NL
46	*Eggerthella lenta*	0.11	20	NL
47	*Bilophila wadsworthia*	0.10	24	NL
48	*Clostridium disporicum*	0.09	14	NL
49	*Odoribacter splanchnicus*	0.08	17	NL
50	*Alistipes massiliensis*	0.05	14	NL
51	*Clostridium leptum*	0.05	15	46
52	*Clostridium innocuum*	0.05	17	NL
53	*Actinomyces odontolyticus*	0.03	17	NL

[a] Ranked by the relative abundance in the third column
[b] Average of relative abundance among the 28 Singaporean subjects
[c] The number of subjects in whom the species was detected
[d] Rank in the metagenomic catalogue of the study of 124 European individuals (13). NL means not listed in the catalogue.

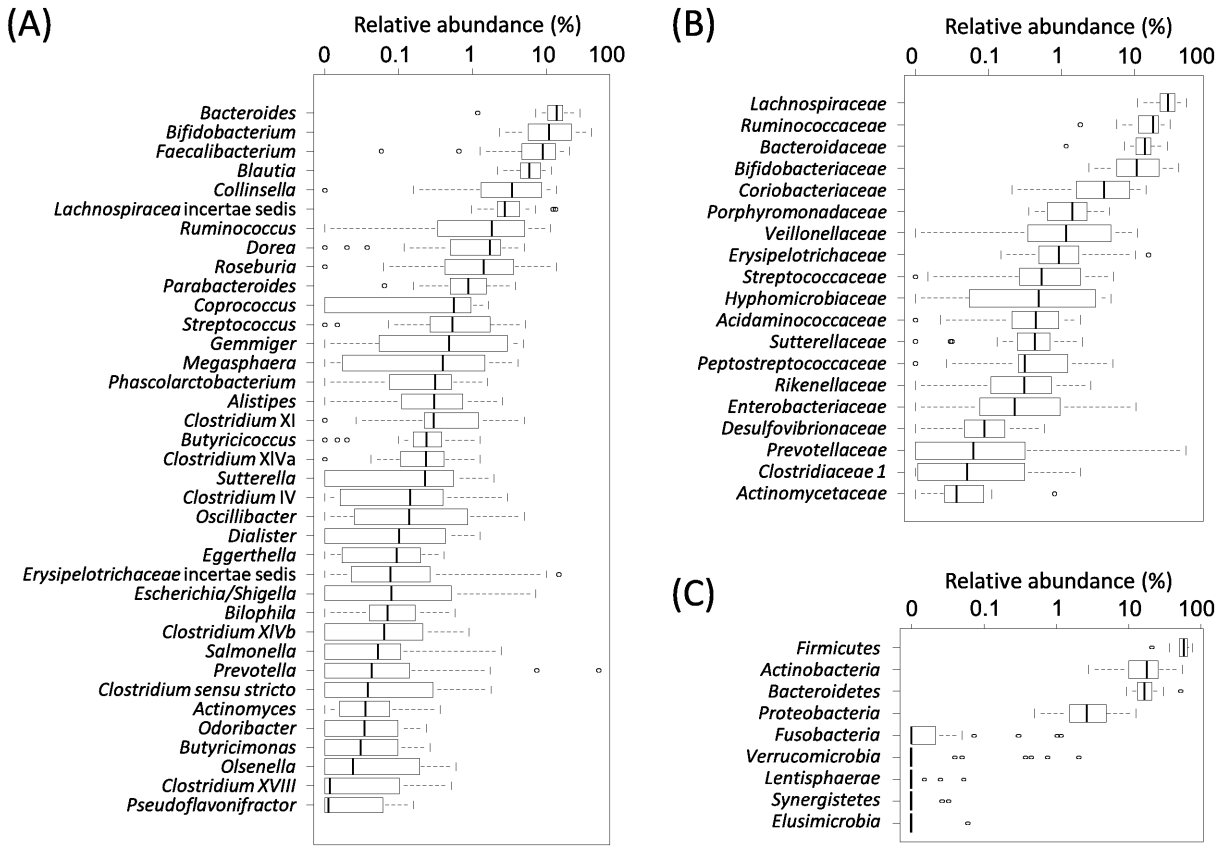

Fig. 3. Population distribution of 37 (A), 19 (B) and 4 (C) common genera, families and phyla, respectively, among the 28 Singaporean subjects. The relative abundance of each taxonomic group was calculated by dividing the read counts of identified sequences by the individual`s total read number. The 37 genera, 19 families, and 4 phyla were selected as they were detected in more than a half of our Singaporean subjects.

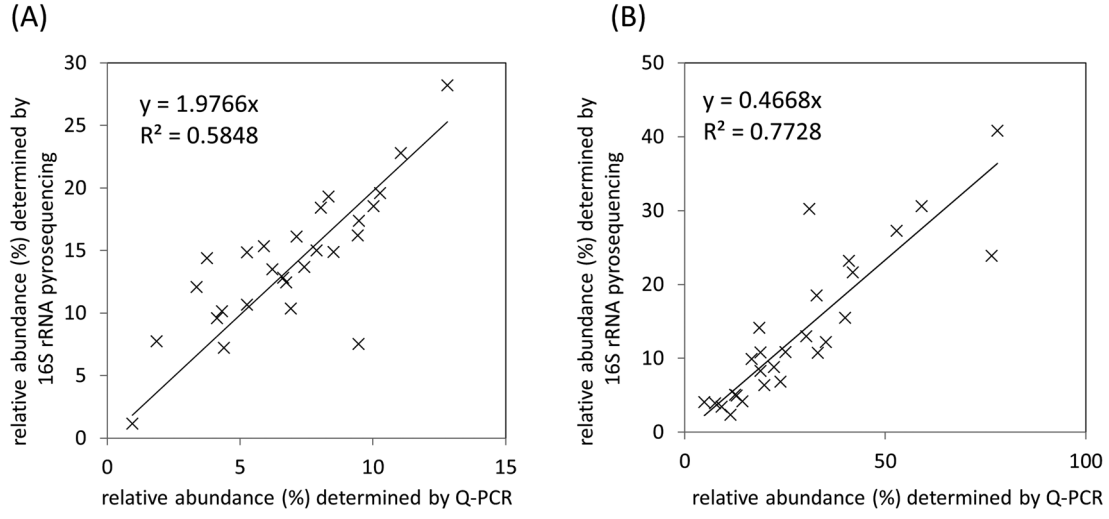

Fig. 4. Comparison of the relative abundances of *Bacteroides* (A) and *Bifidobacterium* (B) determined by 16S rRNA amplicon pyrosequencing with those determined by quantitative real-time PCR. The relative abundance in the pyrosequencing data was calculated by dividing the number of reads identified to genus *Bacteroides* or *Bifidobacterium* by the total read counts in each subject. In the quantitative PCR, group-specific primers targeting the *Bacteroides fragilis* group and genus *Bifidobacterium* were used, respectively [5].

Table 3. Summary of the computational processing of 454 pyrosequence data

Step	Process	Program	Program source (URL or e-mail)	Calculator machine	# of query seqs.	Output	Time
1	Barcode sorting	split_libraries.py	QIIME (http://qiime.org/scripts/split_libraries.html)	Windows 64 bit PC*	1,583,218**	106 to 8618 reads per subject	10 min
2	OTU clustering	pick_otus_through_otu_table.py	QIIME (http://qiime.org/scripts/pick_otus_through_otu_table.html)	Windows 64 bit PC	131,768	15,578 redundant seqs.	17 min
3	de novo chimera check	de novo Uchime	Mothur 1.25.1 (http://www.mothur.org/wiki/Download_mothur)	Windows 64 bit PC	15,578	7,200 chimeric seqs.	6 min
4	DB chimera check	DB Uchime	Mothur 1.25.1 (http://www.mothur.org/wiki/Download_mothur)	Windows 64 bit PC	15,578	7,852 chimeric seqs. (6,612 chimeric seqs.)***	7 hrs
5	Sequence search up to genus level	RDP Classifier	RDP II (http://rdp.cme.msu.edu/)	RDP host computer	8,966	9 phyla–109 genera	10 min
6	Sequence search in species level	RDP Seqmatch	RDP II (http://rdp.cme.msu.edu/)	RDP host computer	8,966	3,992 seqs identified with know species	3 hrs
7	Data processing of RDP Seqmatch	SeqmatchQ400	Kyushu Univ (nakayama @ agr.kyushu-u.ac.jp)	Windows 64 bit PC	8,945	276 species	30 min

*Intel Core i7-3930 K CPU (3.20 GHz).
**Batch sequence data included all sequences from 2 x half PicoTiterPlate regions, which contained 256 samples including non-Singaporean samples.
***The number of chimeric sequences determined by taking into account both de novo and DB chimera checks.

National University of Singapore (to Y.K.L). The authors thank the Asian Microbiome Project (AMP) team [Dr. Y. C. Tsai (National Yang-Ming University, Taipei, Taiwan) Dr. F. Z. Ren (China Agricultural University, Beijing, China), Dr. E. S. Rahayu (Gadjah Mada University, Yogyakarta, Indonesia), Dr. C. C. Liao (Food Industry R&D Institute, Hsinchu, Taiwan), and Dr. S. Nitisinprasert (Kasetsart University, Bangkok, Thailand)] for their support of this study.

REFERENCES

1. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335–336. [Medline] [CrossRef]
2. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537–7541. [Medline] [CrossRef]
3. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. Nature 486: 222–227. [Medline]
4. Nakayama J. 2010. Pyrosequence-based 16S rRNA profiling of gastro-intestinal microbiota. Biosci Microflora 29: 83–96.
5. Matsuki T, Watanabe K, Fujimoto J, Takada T, Tanaka R. 2004. Use of 16S rRNA gene-targeted group-specific primers for real-time PCR analysis of predominant bacteria in human feces. Appl Environ Microbiol 70: 7220–7228. [Medline] [CrossRef]
6. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27: 2194–2200. [Medline] [CrossRef]
7. Schloss PD, Gevers D, Westcott SL. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE 6: e27310. [Medline] [CrossRef]
8. Wright ES, Yilmaz LS, Noguera DR. 2012. DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. Appl Environ Microbiol 78: 717–725. [Medline] [CrossRef]
9. 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. [Medline] [CrossRef]
10. 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. [Medline] [CrossRef]

11. Chao A. 1984. Non-parametric estimation of the number of classes in a population. Scand J Stat 11: 783–791.
12. 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. [Medline] [CrossRef]
13. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Jian M, Zhou Y, Li Y, Zhang X, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59–65. [Medline] [CrossRef]