**Identification of household bacterial community and analysis of species shared with human microbiome**

**Current Microbiology**

**Yoon-Seong Jeon[1,2], Jongsik Chun[1,2,3] and Bong-Soo Kim[1]**

(1) *ChunLab, Inc., Seoul National University, Seoul, Republic of Korea*

(2) *Interdisciplinary Graduate Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea*

(3) *School of Biological Sciences, Seoul National University, Seoul, Republic of Korea*

e-mail: bongsoo.kim@chunlab.com

## Supplementary Methods

### Pyrosequence Data processing

### Trimming primer sequences

The profiles of the V1−V3 regions of the 16S rRNA gene were used to trim primer sequences by hmm-search within HMMER (Eddy, 2011). Trimming primer sequences is essential for removing the Roche-454 adaptor sequence and primer sequences used for amplification in primer regions that could generate inaccurate results. Sequence reads without target primer sequences were eliminated in subsequent steps because these reads could be generated by sequencing errors.

### Assembly of reads into representative sequences

Pyrosequencing can generate homopolymers in sequences, which may bias the results of microbial community analysis. To correct this problem, we generated representative sequences from clusters by the following process. (1) Each read was converted to an artificial sequence that had homopolymeric regions condensed to a single base. (2) Identical sequences and subsequences of longer sequences were clustered. (3) A consensus sequence was generated for each cluster with the original sequences

of the cluster by using multiple alignments. (4) Generated consensus sequences were arranged by sequence length, and they were clustered again, allowing less than two base mismatches based on the error rate of 454 sequencing technology, which is reported at 0.5% (Huse, *et al.*, 2007). (5) The longest consensus sequence was selected from the clustered sequences as the representative sequence. Representative sequences were used to assign taxonomic positions.

**Taxonomic assignments of individual reads**

Each read was identified using hierarchical taxonomic information in the EzTaxon-e database (Kim *et al.*, 2012) and robust pairwise global sequence alignment. The 5 sequences most similar to each pyrosequencing read were identified by a BLASTN search against the EzTaxon-e database, and the pairwise similarities between the query and the 5 most similar sequences were calculated by global pairwise alignment (Myers & Miller, 1988). Taxonomic classifications were carried out using the criteria of ≥97% similarity for species, ≥94% for genus, ≥90% for family, ≥85% for order, ≥80% for class, and ≥75% for phylum. If the sequence similarity was below the criteria value, the sequence was assigned to the "unclassified" group at the corresponding taxonomic ranks.

**Filtering chimera sequences**

Artificial products (chimera sequences) are generated during PCR amplification. They can affect the analysis of microbial communities. To remove chimera sequences, sequences that did not match to the EzTaxon-e database at 97% similarity were subjected to a chimera check process. The EzTaxon-e database was used as first screening tool to choose chimera sequences, because it consists of manually curated high quality and non-chimeric sequences (Kim, *et al.*, 2012). The detection of chimeras was conducted using the UCHIME program (Edgar, *et al.*, 2011).

**Calculation of diversity indices**

The cutoff value for determining operational taxonomic units (OTUs) is generally 97% sequence similarity. These OTUs can be considered as species in pyrosequencing data analysis. Therefore, we calculated the diversity indices of samples using the following three different methods.

(1) CD-HIT method: The CD-HIT program (Li & Godzik, 2006) was used to define OTUs

(2) TBC method: The TBC (Taxonomy-based clustering) program (Lee, *et al.*, 2012) was used to calculate OTUs.

(3) TDC-TBC method: Both CD-HIT and TBC methods are de novo clustering methods, which ignore the real taxonomic identification of each read. Taxonomy-dependent clustering (TDC) can overcome this problem by using information on taxonomic identification. Each read is identified against the EzTaxon-e database, and unclassified reads at the species level (<97% similarity) are subjected to clustering as OTUs using the TBC method. Therefore, the TDC-TBC method is a combination of database-dependent and de novo clustering, and the reads determined by this method are considered real species and artificial OTUs. This hybrid approach can maximize the information on real species diversity in samples in which the 16S rRNA similarity values between species are often higher than 97 % (e.g. species belonging to the family *Enterobacteriaceae*).

The results of the three different calculations are presented in Supplementary Table 1.

## References


Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**: e1002195.

Edgar RC, Haas BJ, Clemente JC, Quince C & Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194-2200.

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

Kim OS, Cho YJ, Lee K*, et al.* (2012) Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int J Syst Evol Microbiol* **62**: 716-721.

Lee JH, Yi H, Jeon YS, Won S & Chun J (2012) TBC: a clustering algorithm based on prokaryotic taxonomy. *J Microbiol* **50**: 181-185.

Li W & Godzik A (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658-1659.
Myers EW & Miller W (1988) Optimal Alignments in Linear-Space. *Computer Applications in the Biosciences* **4**: 11-17.