



Construction & assessment of a unified curated reference database for improving the taxonomic classification of bacteria using 16S rRNA sequence data

Shikha Agnihotry^{1,†}, Aditya N. Sarangi^{1,†} & Rakesh Aggarwal^{1,2}

¹Biomedical Informatics Centre & ²Department of Gastroenterology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, Uttar Pradesh, India

Received: January 31, 2018

Background & objectives: For bacterial community analysis, 16S rRNA sequences are subjected to taxonomic classification through comparison with one of the three commonly used databases [Greengenes, SILVA and Ribosomal Database Project (RDP)]. It was hypothesized that a unified database containing fully annotated, non-redundant sequences from all the three databases, might provide better taxonomic classification during analysis of 16S rRNA sequence data. Hence, a unified 16S rRNA database was constructed and its performance was assessed by using it with four different taxonomic assignment methods, and for data from various hypervariable regions (HVRs) of 16S rRNA gene.

Methods: We constructed a unified 16S rRNA database (16S-Udb) by merging non-ambiguous, fully annotated, full-length 16S rRNA sequences from the three databases and compared its performance in taxonomy assignment with that of three original databases. This was done using four different taxonomy assignment methods [mothur Naïve Bayesian Classifier (mothur-nbc), RDP Naïve Bayesian Classifier (rdp-nbc), UCLUST, SortMeRNA] and data from 13 regions of 16S rRNA [seven hypervariable regions (HVR) (V2-V8) and six pairs of adjacent HVRs].

Results: Our unified 16S rRNA database contained 13,078 full-length, fully annotated 16S rRNA sequences. It could assign genus and species to larger proportions (90.05 and 46.82%, respectively, when used with mothur-nbc classifier and the V2+V3 region) of sequences in the test database than the three original 16S rRNA databases (70.88-87.20% and 10.23-24.28%, respectively, with the same classifier and region).

Interpretation & conclusions: Our results indicate that for analysis of bacterial mixtures, sequencing of V2-V3 region of 16S rRNA followed by analysis of the data using the mothur-nbc classifier and our 16S-Udb database may be preferred.

Key words 16S rRNA - bioinformatics - hypervariable regions - metagenomics - microbiota

[†]Equal contribution

Supplementary material available from <http://www.ijmr.org.in/article.asp?issn=0971-5916;year=2020;volume=151;issue=1;page=93;epage=103;aulast=Agnihotry>

Several human body sites contain a variety of organisms, including prokaryotes and eukaryotes, collectively referred to as microbiota. High-throughput genomic sequencing of 16S rRNA is used for profiling of microbiota¹. This bacterial gene has nine hypervariable regions (HVRs) interspersed with conserved nucleotide sequences. Sequences of these HVRs differ between bacterial groups and hence can be used to identify different bacteria². Further, high-throughput platforms allow sequencing of several nucleic acid molecules in parallel. One of the commonly used platforms, namely Illumina, can read DNA sequences beginning at each end of a DNA fragment for up to 300 nucleotides; these paired-end reads can then be merged to obtain sequences of up to 575-nucleotide long, enough to cover one or two adjacent HVRs. Sequencing of one or more HVR regions using this platform followed by taxonomic identification of each sequence by matching with a database of known bacterial 16S rRNA gene sequences is one of the most-frequently used methods for determining the type and abundance of various bacteria present in specimens that contain a mixture of several bacteria³.

Three databases are commonly used to identify bacterial 16S rRNA sequences, namely SILVA⁴, Ribosomal Database Project (RDP)⁵ and Greengenes⁶. These databases overlap partially, with each containing some entries which are absent in the others. The databases also vary in their coverage; for instance, the most abundant genus in the Greengenes is *Prevotella* and that in the SILVA database is *Lachnospira*. Further, several taxonomy assignment methods such as UCLUST⁷, mothur-nbc⁸, rdp-nbc⁹ and SortMeRNA¹⁰ are used for comparing the query sequences to these databases.

Effectiveness of bioinformatic analysis to assign individual reads in a high-throughput 16S rRNA sequence dataset to various bacterial species can be expected to vary with the HVR sequenced, the 16S rRNA reference database used and the taxonomic assignment methods used. A few studies have assessed the effect of varying one factor, *e.g.*, the reference database¹¹, or assignment method¹² on the taxonomic assignment process, at a time; however, no study has comprehensively looked at the effect of varying all the three factors together. Therefore, we analysed the composite effect of varying the three factors. Further, it was hypothesized that the use of a hybrid database which included entries from all the three major 16S rRNA sequence databases, might improve

the phylogenetic assignment. Hence, a unified 16S rRNA database was constructed by merging non-ambiguous, full-length prokaryotic 16S rRNA sequences with complete annotation up to the species level from the three commonly used databases. The relative performance in taxonomic assignment of this new unified database [referred to as 16S unified database (16S-UDb)] and each of its constituents taken individually was compared, using four different taxonomic assignment methods and using different HVR of 16S rRNA.

Material & Methods

Acquisition and preparation of individual 16S rRNA databases: For the Greengenes and SILVA databases, files containing 16S rRNA reference sequences (pre-clustered at 97% threshold) and corresponding taxonomy mapping information in QIIME format were downloaded (ftp://greengenes.microbio.me/greengenes_release/gg_13_5/gg_13_5_otus.tar.gz and https://www.arb-silva.de/fileadmin/silva_databases/qiime/Silva_123_release.zip, respectively). For RDP (release 11.4), unaligned bacterial 16S rRNA sequences were downloaded (https://rdp.cme.msu.edu/download/current_Bacteria_unaligned.fa.gz) and made QIIME-compatible by removing sequences that were shorter than 1200 nucleotides or that contained any ambiguous base, followed by clustering at 97 per cent threshold using VSEARCH¹³. In addition, a taxonomy mapping file in QIIME-compatible format was created by linking RDP sequence identifiers of the representative sequences with 6-level (phylum, class, order, family, genus and species) taxonomic lineage hierarchy.

Construction of a unified 16S rRNA database (16S-UDb): Bacterial 16S rRNA gene sequences were obtained from three databases - Greengenes (v.13.5), RDP (v11.4) and SILVA 'All-Species Living Tree' Project. For each, any sequences shorter than 1200 nucleotide in length, containing any ambiguous base, not classified up to the species level, or labelled as derived from environment or cloned material were purged. Sequences classified up to subspecies or isolate level were grouped to the species level. These curated data from the three databases were merged and clustered at 97 per cent threshold using VSEARCH¹³. Further, a taxonomy mapping file was created by linking sequence identifiers of the representative sequences with 6-level taxonomic lineage hierarchy.

Construction of test dataset: A test dataset was required for comparing the performance of the unified and the individual 16S rRNA databases. For this, a pre-formatted 16S rRNA reference dataset was downloaded from the NCBI ftp site (<ftp://ftp.ncbi.nlm.nih.gov/blast/db/16SMicrobial.tar.gz>), and sequences of 1200 nucleotides or longer in it were extracted using the 'blastdbcmd' utility of NCBI stand-alone BLAST version 2.2.28. The identifier of each sequence was used to query the NCBI taxonomy database (<https://www.ncbi.nlm.nih.gov/taxonomy>) to obtain its taxonomic information. These data were used to create a taxonomy mapping file containing the reference identifiers and their 6-level taxonomic lineage hierarchies up to the species level; any sequences with incomplete species information were removed, and those with taxonomic lineage defined up to subspecies or isolates, were grouped to species level. Any duplicate sequences with 100 per cent identity were removed using CD-HIT¹⁴. This provided a test dataset of non-redundant full-length 16S rRNA sequences classified up to species level.

Construction of test datasets of shorter lengths: Primer-pairs flanking seven individual HVRs (V2 to V8) of bacterial 16S rRNA and six pairs of adjacent HVRs [V23 (*i.e.*, V2+V3), V34, V45, V56, V67, V78] were identified (Supplementary Table I (available from http://www.ijmr.org.in/articles/2020/151/1/images/IndianJMedRes_2020_151_1_93_279138_sm4.pdf)). Their sequences were aligned with the test dataset, using ClustalW program in-built in BioEdit Suite v.7.2.5 (<http://www.mbio.ncsu.edu/BioEdit/BioEdit.zip>), and the nucleotide sequences lying between the primers in primer pairs covering each HVR (V2 to V8; 7 sets) or two adjacent HVRs (*i.e.*, V2+V3, V3+V4, V4+V5, V5+V6, V6+V7 and V7+V8; 6 sets) were extracted. This yielded a total of 13 test datasets. The V1 and V9 regions were ignored since these are rarely used for metagenomic analysis because of their incomplete nature¹⁵.

Comparison of different approaches to identify an optimal pipeline: The performances of various combinations of four taxonomy assignment methods (UCLUST, SortMeRNA, mothur-nbc and rdp-nbc) and four 16S rRNA databases (Greengenes, SILVA, RDP and the new 16S-UDb) in correctly classifying reads included in the 13 HVR/HVR-pair test datasets, up to the genus and species levels were determined. This was

done using 208 runs (=4×4×13) of the assign_taxonomy.py script of QIIME software package¹⁶.

UCLUST-based classifier uses the consensus taxonomy assignment-based approach; it was invoked with the parameters min_consensus_fraction: 0.51, similarity: 0.9 and uclust_max_accepts: 3. SortMeRNA algorithm is based on approximate seeds and accounts for fast and sensitive analyses of rRNA sequences. SortMeRNA was run with parameters: SortMeRNA_coverage: 0.9, SortMeRNA_best_N_alignments: 5 and SortMeRNA_e_value: 1.0. The two naive Bayes-based classifiers (mothur-nbc and rdp-nbc) were retrained using the reference operational taxonomic units (OTUs) clustered at 97 per cent threshold and the corresponding taxonomy mapping files from the above four databases, individually, with minimum confidence score (-c) of 0.80. Performance of each classifier-database combination was calculated as the proportion of sequences in the test dataset that were correctly classified.

Assessment of performance of the 'optimal pipeline' using a real-life dataset: For this, Illumina 16S rRNA gene V3 sequencing data were used from one of our previous studies¹⁷. It contained 20,508,594 high-quality reads obtained from stool specimens of 14 healthy persons [number of median reads per specimen: 272,327 (range 136,563 to 761,961); total reads 4,728,631] and 33 patients with enthesitis-related arthritis [median 397,351 (102,093 to 1,502,380); total reads 15,871,719]. These data were subjected to taxonomy assignment using the optimum approach identified in the previous step. The results obtained were compared with those obtained using the approach used in the original analysis *i.e.*, using UCLUST Consensus Taxonomy Assigner and sub-sampled open-reference OTU picking protocol of QIIME 1.8 against the Greengenes v.13.8 reference OTUs pre-clustered at 97 per cent threshold with the software's default parameters¹⁶. During these analyses, any singleton OTUs (with only one sequence in all specimens taken together), unassigned OTUs and eukaryotic OTUs were removed from the 'biom' files generated by each approach. Further, to reduce noise, OTUs that were observed in fewer than 10 per cent of stool specimens or that accounted for fewer than 0.002 per cent of reads in all the specimens taken together were also purged. The relative performance of the two methods was assessed by comparing the BIOM files generated by each.

Comparison of computational performances of various approaches: The time taken for different combinations of databases and classifiers were assessed using an Intel Corporation Xeon E7 Workstation with 6 processors (Intel Corp., CA, USA) using identical parameters, *i.e.*, pre-filtration of sequences at 60 per cent identity and subsample-based open-reference OTU picking method with the use of 10 per cent subsample.

The study was carried out at the Biomedical Informatics Centre, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, India.

Results & Discussion

Acquisition of 16S rRNA databases: The Greengenes database (v.13.5), when pre-clustered at 97 per cent threshold as inbuilt in QIIME V1 software, contained 99,322 sequences belonging to 1812 unique genera, and the SILVA database (v.123) with similar pre-clustering contained 216,401 sequences belonging to 3541 unique genera. For RDP, the QIIME-compatible data in a similar format contained 145,925 sequences belonging to 2737 unique genera. Figure A shows the number of taxonomic units (by name) at various ranks, *i.e.*, phylum to genus, between the SILVA, RDP and Greengenes taxonomies. This comparison showed that the three databases varied markedly in the taxa represented in each, with only 19.8 per cent of the phyla, 9.7 per cent of classes, 9 per cent of orders, 14.8 per cent of families and 27.6 per cent of genus ranks present in any of the three databases being shared across all the databases.

Construction of integrated 16S database: Of the 4,345,168 bacterial 16S rRNA gene sequences obtained from the three databases (Greengenes v.13.5: 1,262,986 sequences, RDP v.11.4: 3,070,243 sequences and SILVA 'All-Species Living Tree' Project v.123: 11,939 sequences), 2,629,394 (60.5%) were 1200 nucleotide or longer in length and free of ambiguous nucleotides. Of these, 405,538 were fully classified up to the species level. These 405,538 sequences were clustered at 97 per cent threshold using VSEARCH¹³. For the current version of 16S-UDb v1.0, this yielded 13,078 unified sequences from the bacterial kingdom, belonging to 36 phyla, 94 classes, 187 orders, 414 families, 2453 genera and 4881 unique species-like groups. Proteobacteria was the predominant phylum accounting for 38.24 per cent of all the sequences, followed by Firmicutes (28.91%), Actinobacteria (12.72%), Bacteroidetes (10.44%), Cyanobacteria (2.6%), Tenericutes (1.39%),

Spirochaetes (0.83%), Verrucomicrobia (0.47%) and 29 other minor phyla.

The numbers of taxonomic units (by name) shared between a union of the SILVA, RDP, Greengenes taxonomies (labelled as ALL) versus 16S-UDb at phylum to genus ranks are shown in Figure B. The top 10 phyla, class, order and genera in the Greengenes, SILVA, RDP and 16S-UDb databases are shown in Supplementary Table II (available from http://www.ijmr.org.in/articles/2020/151/1/images/IndianJMedRes_2020_151_1_93_279138_sm5.pdf). Of the taxonomic units at various levels contained in 'ALL' but not in 16S-UDb, a large majority were present in only one of the three starting databases, *i.e.*, Greengenes, SILVA or RDP, and very few were represented in two or three of these (Supplementary Table III (available from http://www.ijmr.org.in/articles/2020/151/1/images/IndianJMedRes_2020_151_1_93_279138_sm6.pdf)).

Contribution of each input database to the 16S-UDb database: Greengenes database contributed 3387 sequences belonging to 23 phyla, 45 classes, 79 orders, 150 families, 306 genus and 533 species to the 16S-UDb database. Among these sequences, Firmicutes was the predominant phylum with 43.52 per cent of all the sequences contributed, followed by Proteobacteria (26.28%), Bacteroidetes (13.14%), Actinobacteria (9.86%), Crenarchaeota (1.27%), Spirochaetes (1.18%), Euryarchaeota (1.09%) and Verrucomicrobia (0.83%).

RDP contributed 6254 sequences from 30 phyla, 56 classes, 118 orders, 280 families, 1316 genus and 2880 species. Among these sequences, Proteobacteria (45.68%) was the most dominant phylum, followed by Firmicutes (28.06%), Actinobacteria (11.24%), Bacteroidetes (4.88%), Cyanobacteria (4.05%), Tenericutes (1.92%), Spirochaetes (0.80%) and Planctomycetes (0.56%).

SILVA database contributed 3437 sequences from 29 phyla, 58 classes, 137 orders, 293 families, 1576 genus and 2264 species. Of these, Proteobacteria (36.49%) were the most predominant, followed by Firmicutes (19.84%), Actinobacteria (18.21%), Bacteroidetes (17.89%), Tenericutes (1.45%), Deinococcus-Thermus (0.93%), Verrucomicrobia (0.61%) and Chloroflexi (0.52%).

Construction of gold standard test dataset: The NCBI Bacterial 16S Ribosomal RNA dataset (<ftp://ftp.ncbi>.

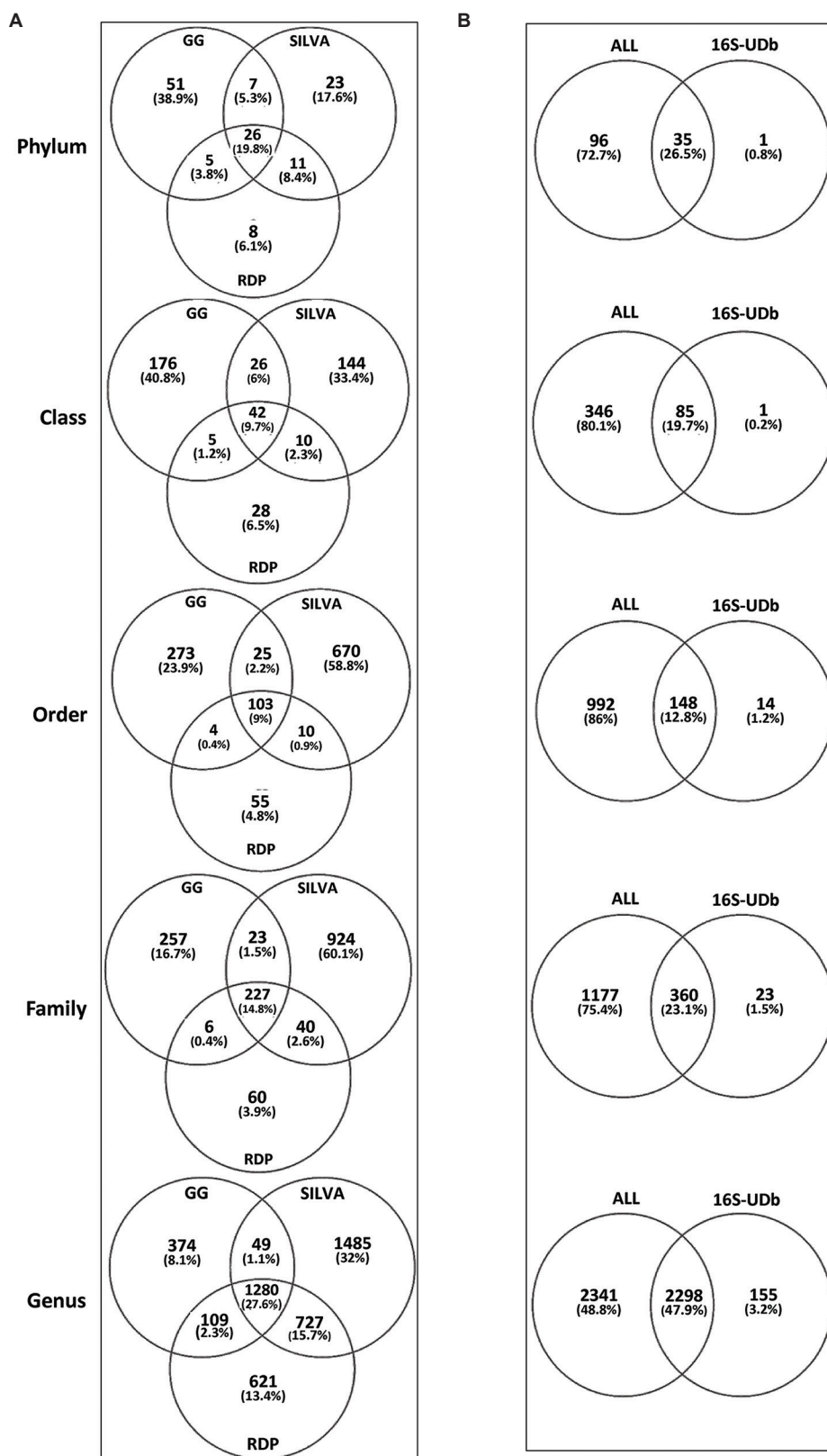


Figure. Comparison of taxonomies based on taxon names found at each rank from phylum to genus. The three taxonomies, SILVA, Ribosomal Database Project (RDP) and Greengenes (GG) commonly used for 16S rRNA based analyses were compared in detail (Panel A) and then a union of these three databases (labelled as ALL) were compared against the unified 16S-UDb database (Panel B).

nlm.nih.gov/blast/db/16SMicrobial.tar.gz) contained a total of 18,775 eligible (1200 nucleotide or longer in length) 16S rRNA sequences, of which 2132 were redundant and 5318 lacked taxonomic classification to species level. From the remaining full-length 16S rRNA sequences, 13 HVR/HVR-pair test datasets (each with 11,325 members) were constructed. The 16S rRNA test dataset comprised sequences from only the bacterial kingdom, related to 36 phyla, 86 classes, 162 orders, 383 families, 2453 genera and 4881 unique species-like groups. Proteobacteria was the dominant phylum with 38.89 per cent of all sequences (n=11325), followed by Actinobacteria (23.69%), Firmicutes (19.51%), Bacteroidetes (10.76%), Tenericutes (1.78%), Spirochaetes (0.91%), Cyanobacteria (0.54%), Verrucomicrobia (0.44%) and 28 other minor phyla. The top 10 phylum, class, order, family and genus groups in this test dataset are shown in Supplementary Table IV (available from http://www.ijmr.org.in/articles/2020/151/1/images/IndianJMedRes_2020_151_1_93_279138_sm7.pdf).

Identification of 'optimal pipeline' using the test dataset: The default reference database for QIIME v1 is a subset of Greengenes rRNA sequences pre-clustered clustered at 97 per cent identity¹⁶. Several meta-analyses and case-control analyses of human microbiome used this 97 per cent identity subset of Greengenes^{17,18} and SILVA¹⁹⁻²¹ as reference databases. Hence, we used the 16S-UDb clustered at 97 per cent threshold for our comparisons.

Table I shows the performance of various combinations of four taxonomy assignment methods and four 16S rRNA databases (including the 16S-UDb clustered at 97 per cent threshold) in correctly classifying the 13 HVR/HVR-pair test datasets at the family, genus and species levels. For UCLUST and SortMeRNA, the SILVA database performed the best, 16S-UDb performed the best; this could indicate that different methods work better with databases of different sizes. However, overall, for each of the 13 datasets, the unified database (16S-UDb) and mothur-nbc classifier combination provided the highest correct classification rate at each taxonomic level.

Using this database-classifier combination (16S-UDb and mothur-nbc), we compared the performance of the 13 test datasets. In this analysis, V2-3 HVR performed the best for classification at the genus and species levels, and the V3-4 HVR did the best for classification at the family level. The HVR

providing the best discrimination depends on the composition of the sample (bacterial mixture)^{22,23}. The strength of our analysis was that we used various tools on both an idealized and a real-life dataset and obtained good performance with this database-classifier pair¹⁵.

Assessment of performance of the 'optimal pipeline' using a real-life dataset: In data from our previous study¹⁷, the original analysis using the Greengenes databases and UCLUST classifier (the GG-UCLUST approach) was able to assign genus and species to 17,439,681 (85.0%) and 8,381,405 (40.9%), respectively, of all reads (numbering 20,508,594). By contrast, the 16S-UDb-mothur-nbc approach was able to assign genus and species to large proportions [17,886,699 (87.2%) and 12,450,839 (60.7%), respectively] of reads in this dataset. The 16S-UDb-mothur-nbc pipeline identified presence of a large number of families, genera and species in the test dataset (50, 94 and 175, respectively; Supplementary Tables V (available from http://www.ijmr.org.in/articles/2020/151/1/images/IndianJMedRes_2020_151_1_93_279138_sm10.pdf) and VI (available from http://www.ijmr.org.in/articles/2020/151/1/images/IndianJMedRes_2020_151_1_93_279138_sm11.pdf)) than the original GG-UCLUST pipeline (47, 69 and 34, respectively). In the results from these two pipelines, 35 families, 59 genera and 32 species were common. In the species-level results from the two approaches, 145 species showed discordance, including two (*Staphylococcus aureus* and *Lactobacillus zaeae*) that were detected only by the GG-UCLUST approach and 143 that were detected only by the 16S-UDb-mothur-nbc approach. These data showed that the 16S-UDb-mothur-nbc approach was more sensitive in identifying the presence of different bacterial species in 16S rRNA sequence datasets, such as those derived from bacterial mixtures.

The relative abundances of the four most frequent phyla (Bacteroidetes, Firmicutes, Proteobacteria and Actinobacteria) and for the 32 species identified by both the approaches (Table II) were generally similar. However, for one species, namely *Faecalibacterium prausnitzii* (Family: *Ruminococcaceae*), the per cent abundances using the 16S-UDb-mothur-nbc approach [healthy: 3.33 (0.51-9.19), ERA: 4.39 (0.07-79.55); $P < 0.01$] were much higher than those using the GG-UCLUST approach [healthy: 0.09 (0.02-0.63), ERA: 0.13 (0.001-1.84); $P < 0.01$]. This indicated that the use of the optimum database and classifier combination identified by us might not only allow detection of a larger number of bacterial

Table I. Accuracy (%) of various combinations of four taxonomy assignment algorithms and four 16S rRNA sequence databases using different hypervariable regions

HVR	16S rRNA database	Accuracy in assigning bacterial family using each classifier				Accuracy in assigning bacterial genus using each classifier				Accuracy in assigning bacterial species using each classifier			
		UCLUST	SortMeRNA	RDP	mothur	UCLUST	SortMeRNA	RDP	mothur	UCLUST	SortMeRNA	RDP	mothur
V2	GG	84.37	82.74	83.71	85.82	57.62	55.34	63.41	69.97	2.85	2.61	9.11	10.08
	SILVA	87.43	85.62	86.30	88.25	77.16	73.03	75.38	82.08	2.10	1.80	12.05	14.86
	RDPdb	79.35	77.51	81.90	83.21	71.67	66.14	77.16	84.27	3.24	3.96	21.23	24.20
	16S-UDb	77.25	81.13	89.06	91.30	60.62	65.64	83.05	88.50	8.71	11.48	42.84	45.61
V3	GG	80.85	79.43	70.35	76.31	53.65	49.55	46.64	57.72	2.11	1.73	6.32	8.00
	SILVA	84.97	83.16	73.43	83.33	70.55	66.08	56.19	67.96	1.46	0.85	7.02	10.35
	RDPdb	76.15	76.11	71.45	77.86	64.18	60.86	60.06	73.87	1.63	1.37	12.94	17.95
	16S-UDb	76.11	77.29	76.31	85.27	54.68	55.76	64.69	76.85	5.63	5.77	30.48	37.49
V4	GG	85.50	85.17	84.47	86.09	59.37	56.60	60.65	66.97	2.24	1.93	6.91	8.39
	SILVA	89.07	87.71	87.74	89.24	78.18	74.17	71.72	79.43	1.47	1.17	7.31	10.38
	RDPdb	82.18	78.75	82.16	83.47	74.27	67.55	71.73	82.09	1.70	1.70	13.40	17.47
	16S-UDb	81.45	80.01	87.67	90.91	59.81	60.49	78.31	85.62	6.30	6.11	34.30	39.83
V5	GG	82.05	81.25	74.17	81.55	51.00	49.18	45.06	57.05	1.34	1.05	3.61	5.84
	SILVA	86.00	83.83	77.36	85.05	65.52	60.72	48.69	65.86	0.84	0.50	3.07	0.55
	RDPdb	77.10	73.97	72.50	79.48	59.87	54.85	51.17	68.88	1.02	0.70	6.70	11.88
	16S-UDb	76.99	79.27	77.62	87.11	52.21	54.27	57.53	73.33	3.53	3.58	18.24	27.57
V6	GG	77.24	76.83	64.85	76.75	47.96	47.34	43.94	57.77	1.72	1.92	6.15	8.40
	SILVA	81.59	82.19	71.28	81.47	64.58	63.81	51.80	67.58	1.33	1.62	7.10	11.00
	RDPdb	73.65	72.68	67.74	76.30	59.74	55.49	52.27	68.06	1.95	2.75	14.05	19.24
	16S-UDb	70.18	78.91	74.76	84.53	50.25	58.63	62.88	75.91	6.63	12.10	27.44	35.06
V7	GG	76.78	77.28	57.45	73.27	47.39	44.68	32.79	48.15	1.17	0.88	2.68	4.78
	SILVA	79.54	79.31	62.75	76.68	59.51	56.47	32.95	52.13	0.68	0.26	2.19	5.23
	RDPdb	70.08	69.53	60.34	72.39	53.33	51.21	39.13	57.91	0.79	0.52	5.74	10.63
	16S-UDb	71.89	74.10	65.15	79.64	46.86	49.10	46.47	63.33	3.41	2.47	14.23	23.45
V8	GG	83.68	80.79	83.22	85.51	56.20	48.11	57.95	65.47	1.97	1.24	6.22	8.07
	SILVA	86.70	83.21	85.33	87.90	73.84	62.23	67.47	75.73	1.23	0.48	6.88	9.76
	RDPdb	79.05	74.81	80.65	82.72	68.01	57.63	70.20	79.93	1.37	1.29	14.08	18.28
	16S-UDb	80.79	80.50	83.99	90.66	56.33	59.08	76.52	83.97	4.53	3.90	34.15	39.50

Contd...

HVR	16S rRNA database	Accuracy in assigning bacterial family using each classifier				Accuracy in assigning bacterial genus using each classifier				Accuracy in assigning bacterial species using each classifier			
		UCLUST	SortMeRNA	RDP	mothur	UCLUST	SortMeRNA	RDP	mothur	UCLUST	SortMeRNA	RDP	mothur
V2+V3	GG	85.86	83.59	85.81	86.55	61.11	55.79	66.78	70.88	2.85	2.52	9.40	10.23
	SILVA	88.73	86.58	88.28	89.25	80.14	73.94	79.43	83.84	2.01	1.60	12.64	15.02
	RDPdb	81.76	78.06	83.04	83.75	75.84	67.73	82.47	87.20	2.79	3.13	21.78	24.28
V3+V4	16S-UDb	80.18	80.72	90.43	91.83	63.13	64.82	86.87	90.05	7.85	9.25	45.08	46.82
	GG	85.77	85.01	86.00	86.81	60.64	55.96	64.73	69.17	2.48	2.19	8.13	9.21
	SILVA	89.29	88.33	88.95	89.71	80.64	75.43	77.47	83.06	1.57	1.21	9.09	11.88
V4+V5	RDPdb	82.94	80.89	83.38	83.88	77.17	68.16	80.91	86.58	1.94	1.66	16.42	20.29
	16S-UDb	83.12	80.11	90.19	91.89	62.75	62.12	85.33	89.47	6.13	5.80	41.61	44.81
	GG	86.15	85.92	86.04	86.81	61.26	62.13	64.04	68.54	2.30	6.23	7.46	8.74
V5+V6	SILVA	89.47	86.31	89.06	89.64	80.09	72.01	75.86	81.41	1.39	1.11	8.23	10.97
	RDPdb	82.97	77.82	83.24	83.81	77.64	65.94	78.08	85.03	1.82	1.92	14.74	18.68
	16S-UDb	82.99	78.14	89.70	91.65	61.85	61.02	83.64	88.54	5.90	6.02	38.58	42.81
V6+V7	GG	74.60	72.52	84.77	86.60	50.67	43.00	60.95	68.06	1.76	1.04	7.25	9.19
	SILVA	88.57	84.28	87.80	89.27	76.97	63.95	73.63	81.20	1.16	0.61	8.64	9.72
	RDPdb	81.26	70.86	82.20	83.47	72.87	56.40	73.69	83.28	1.62	1.49	15.88	20.64
V7+V8	16S-UDb	81.17	67.88	88.52	91.31	59.13	54.15	80.25	87.06	5.36	4.75	37.92	42.77
	GG	83.53	79.67	81.62	85.34	55.22	45.43	56.76	66.35	1.83	1.09	7.09	9.03
	SILVA	87.11	80.49	85.32	88.22	73.45	59.27	68.53	78.41	1.31	0.48	8.89	12.19
V7+V8	RDPdb	78.78	70.75	80.43	82.48	67.99	53.17	69.76	80.59	1.65	1.90	16.43	21.01
	16S-UDb	79.20	74.33	86.35	90.67	56.96	54.06	77.71	85.16	5.24	4.88	37.17	42.09
	GG	84.13	80.01	83.31	85.44	55.85	46.07	58.04	65.57	1.92	0.92	6.21	8.07
V7+V8	SILVA	87.31	81.93	85.33	87.95	74.80	57.41	67.01	75.64	1.11	0.42	6.75	9.62
	RDPdb	78.87	72.33	80.74	82.69	68.16	54.03	70.08	79.86	1.34	1.67	14.05	18.19
	16S-UDb	75.43	80.49	86.75	90.61	56.74	59.14	76.42	83.94	4.49	4.45	34.00	39.37

Values in bold represents the highest values among those obtained using different databases and classifiers for the same HVR and up to the same taxonomic level. GG, greenGenes, RDP, Ribosomal Database Project, 16S-UDb, the new 16S unified database; HVR, hypervariable regions

Table II. Relative abundance (in %) of different taxonomic groups detected in the real-life test dataset using the GG-UCLUST and 16S-Udb-mothur approaches. Data are shown separately for healthy subjects and those with disease

Taxonomic group	Subject group	16S rRNA database and classifier algorithm used					
		Greengenes-UCLUST			16S-UDb-mothur		
		Median	Minimum	Maximum	Median	Minimum	Maximum
p_Bacteroidetes; f_Prevotellaceae; g_Prevotella; s_copri	Healthy	58.17	0.79	77.16	58.2	0.78	77.53
	Patient	8.92	0.02	67.53	8.87	0.02	67.81
p_Bacteroidetes; f_Prevotellaceae; g_Prevotella; s_stercorea	Healthy	3.86	0.03	15.7	3.83	0.03	15.63
	Patient	0.08	1.16E-03	17.93	0.08	1.16E-03	18.81
p_Firmicutes; f_Veillonellaceae; g_Mitsuokella; s_multacida	Healthy	0.19	0	1.04	0.22	6.02E-04	1.05
	Patient	1.75E-03	0	1	1.74E-03	0	1.06
p_Firmicutes; f_Erysipelotrichaceae; g_Eubacterium; s_biforme	Healthy	0.12	3.99E-04	3.85	0.12	3.96E-04	3.91
	Patient	0.08	0	5.62	0.08	0	5.59
p_Proteobacteria; f_Pasteurellaceae; g_Haemophilus; s_parainfluenzae	Healthy	0.1	9.65E-04	2.98	0.1	9.75E-04	2.99
	Patient	0.06	1.86E-03	7.75	0.06	1.85E-03	7.76
p_Firmicutes; f_Lactobacillaceae; g_Lactobacillus; s_ruminis	Healthy	0.1	4.82E-04	5.7	0.1	4.88E-04	5.67
	Patient	0.02	0	2.4	0.02	0	2.39
p_Firmicutes; f_Ruminococcaceae; g_Faecalibacterium; s_prausnitzii	Healthy	0.09	0.02	0.62	3.33	0.51	9.19
	Patient	0.13	1.05E-03	1.84	4.39	0.07	79.55
p_Actinobacteria; f_Coriobacteriaceae; g_Collinsella; s_aerofaciens	Healthy	0.05	7.99E-04	16.22	0.05	7.93E-04	16.18
	Patient	0.17	0	6.72	0.16	0	6.71
p_Actinobacteria; f_Bifidobacteriaceae; g_Bifidobacterium; s_adolescentis	Healthy	0.05	0	8.4	0.05	1.84E-04	8.38
	Patient	0.05	0	8.31	0.05	0	8.32
p_Firmicutes; f_Veillonellaceae; g_Veillonella; s_dispar	Healthy	0.03	2.35E-03	2.85	0.03	9.75E-04	2.5
	Patient	0.11	1.84E-03	3.04	0.08	1.60E-03	3
p_Actinobacteria; f_Bifidobacteriaceae; g_Bifidobacterium; s_longum	Healthy	0.02	1.83E-03	3.14	0.02	1.84E-03	3.12
	Patient	0.37	1.20E-03	34.4	0.37	1.20E-03	34.32
p_Firmicutes; f_Erysipelotrichaceae; g_Bulleidia; s_p-1630-c5	Healthy	0.02	0	0.22	0.02	0	0.22
	Patient	2.54E-04	0	2.61	2.54E-04	0	2.59
p_Bacteroidetes; f_Porphyrimonadaceae; g_Parabacteroides; s_distasonis	Healthy	3.12E-03	0	0.56	3.22E-03	0	0.57
	Patient	0.03	1.67E-04	7.87	0.03	1.66E-04	7.9
p_Bacteroidetes; f_Bacteroidaceae; g_Bacteroides; s_caccae	Healthy	3.08E-03	0	0.45	3.27E-03	0	0.59
	Patient	4.28E-03	0	15.76	0.03	0	15.86
p_Actinobacteria; f_Micrococcaceae; g_Rothia; s_mucilaginososa	Healthy	1.31E-03	0	0.09	1.31E-03	0	0.09
	Patient	2.98E-03	0	0.18	2.98E-03	0	0.17
p_Bacteroidetes; f_Bacteroidaceae; g_Bacteroides; s_fragilis	Healthy	1.25E-03	0	0.07	1.25E-03	0	0.07
	Patient	0.01	0	17.96	0.01	0	17.9
p_Firmicutes; f_Clostridiaceae; g_Clostridium; s_perfringens	Healthy	8.10E-04	0	0.08	0	0	7.05E-04
	Patient	2.84E-04	0	0.03	0	0	0.17
p_Proteobacteria; f_Pasteurellaceae; g_Actinobacillus; s_parahemolyticus	Healthy	6.58E-04	0	0.05	8.89E-04	0	0.05
	Patient	1.73E-04	0	0.1	1.73E-04	0	0.12
p_Bacteroidetes; f_Bacteroidaceae; g_Bacteroides; s_plebeius	Healthy	2.58E-04	0	0.99	2.59E-04	0	0.99
	Patient	0.01	0	52.88	0.01	0	52.68

Contd...

Taxonomic group	Subject group	16S rRNA database and classifier algorithm used					
		Greengenes-UCLUST			16S-UDb-mothur		
		Median	Minimum	Maximum	Median	Minimum	Maximum
p_Firmicutes; f_Lactobacillaceae; g_Lactobacillus; s_mucosae	Healthy	0	0	0.02	0	0	0.02
	Patient	3.51E-04	0	0.45	3.48E-04	0	0.46
p_Actinobacteria; f_Coriobacteriaceae; g_Eggerthella; s_lenta	Healthy	0	0	0.02	0	0	0.12
	Patient	3.29E-04	0	0.75	1.04E-03	0	0.75
p_Firmicutes; f_Veillonellaceae; g_Veillonella; s_parvula	Healthy	0	0	0.07	1.12E-03	0	0.41
	Patient	2.71E-04	0	0.07	2.12E-03	0	0.11
p_Firmicutes; f_Erysipelotrichaceae; g_Eubacterium; s_dolichum	Healthy	0	0	0.01	0	0	0.01
	Patient	2.54E-04	0	0.02	2.54E-04	0	0.02
p_Firmicutes; f_Streptococcaceae; g_Streptococcus; s_anginosus	Healthy	0	0	6.81E-04	0	0	6.81E-04
	Patient	2.47E-04	0	0.05	2.46E-04	0	0.05
p_Bacteroidetes; f_Bacteroidaceae; g_Bacteroides; s_eggerthii	Healthy	0	0	1.01E-03	0	0	1.00E-03
	Patient	2.30E-04	0	2.52	2.29E-04	0	2.49
p_Firmicutes; f_Lachnospiraceae; g_Blautia; s_producta	Healthy	0	0	0.01	0.03	2.12E-03	0.45
	Patient	1.94E-04	0	0.16	0.03	3.90E-04	0.21
p_Firmicutes; f_Clostridiaceae; g_Clostridium; s_neonatale	Healthy	0	0	0	0	0	2.25E-03
	Patient	0	0	0.18	0	0	0.18
p_Proteobacteria; f_Oxalobacteraceae; g_Oxalobacter; s_formigenes	Healthy	0	0	0.01	2.10E-04	0	0.02
	Patient	0	0	0.21	0	0	0.21
p_Proteobacteria; f_Enterobacteriaceae; g_Morganella; s_morganii	Healthy	0	0	1.22E-03	0	0	1.20E-03
	Patient	0	0	0.27	0	0	0.26
p_Proteobacteria; f_Moraxellaceae; g_Acinetobacter; s_johnsonii	Healthy	0	0	2.36E-03	0	0	2.34E-03
	Patient	0	0	0.09	0	0	0.09
p_Proteobacteria; f_Moraxellaceae; g_Acinetobacter; s_lwoffii	Healthy	0	0	2.30E-03	0	0	2.29E-03
	Patient	0	0	0.55	0	0	0.55
p_Verrucomicrobia; f_Verrucomicrobiaceae; g_Akkermansia; s_muciniphila	Healthy	0	0	0.01	0	0	0.01
	Patient	0	0	0.69	0	0	0.69

species in a mixture but also help better assess their abundances.

Computational performances of different approaches:

The 16S-UDb-mothur-nbc approach needed a shorter computational time than the GG-UCLUST approach (59 vs. 84 min, respectively) for completion. For analysis of high-throughput 16S rRNA sequence data for bacterial mixtures, several taxonomy assignment methods, at least three 16S rRNA reference databases and various HVRs of the 16S rRNA gene were used. In this study, a unified database was constructed by merging non-ambiguous, fully annotated, full-length 16S sequences from the three commonly-used databases. The performance of various combinations of 16S sequence databases, HVRs and taxonomy assignment methods

was assessed. Our analysis showed that 16S-UDb (clustered at 97% identity), the unified 16S rRNA database that we created, performed better than the currently-available 16S rRNA databases in that it was able to assign taxonomic lineage up to the family, genus and species levels to a large proportion of sequences in a test database and in a real-life dataset. Further, a combination of this database with mothur-nbc classifier had the best performance among all the database-classifier combinations, as did a region covering the V2 and V3 HVRs compared to the other HVRs.

Our study had a limitation in that we clustered the reference sequences using a 97 per cent threshold. Edgar²⁴ has reported that the use of 100 per cent identity threshold may be better than the usual

97 per cent threshold for accurate prediction. This is because the use of traditional 97 per cent threshold may place sequences from more than one species and/or genera in the same cluster, leading to some erroneous classification at genus and species levels²⁵. Thus, there is a need to repeat this study at a future date using 100 per cent identity threshold instead of the 97 per cent identity threshold.

In conclusion, our analysis shows that sequencing of V2-V3 region of the 16S rRNA, followed by analysis of the sequence data obtained using the mothur-nbc classifier of QIIME v1 and our unified 16S-UDb database (<https://github.com/sarangian/16S-UDb.git>) may be used for analysis of bacterial mixtures, such as those present various body sites and in environmental specimens.

Financial support & sponsorship: None.

Conflicts of Interest: None.

References

- Neufeld JD, Mohn WW. Assessment of microbial phylogenetic diversity based on environmental nucleic acids. In: *Molecular identification, systematics, and population Structure of prokaryotes*. Berlin, Heidelberg: Springer; 2006. p. 219-59.
- Curtis TP, Head IM, Lunn M, Woodcock S, Schloss PD, Sloan WT. What is the extent of prokaryotic diversity? *Philos Trans R Soc Lond B Biol Sci* 2006; 361 : 2023-37.
- Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *J Clin Microbiol* 2007; 45 : 2761-4.
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al*. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* 2013; 41 : D590-6.
- Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, *et al*. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 2014; 42 : D633-42.
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al*. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006; 72 : 5069-72.
- Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010; 26 : 2460-1.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, *et al*. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009; 75 : 7537-41.
- Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007; 73 : 5261-7.
- Kopylova E, Noé L, Touzet H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012; 28 : 3211-7.
- Ritari J, Salojärvi J, Lahti L, de Vos WM. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics* 2015; 16 : 1056.
- Chaudhary N, Sharma AK, Agarwal P, Gupta A, Sharma VK. 16S classifier: A tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. *PLoS One* 2015; 10 : e0116106.
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* 2016; 4 : e2584.
- Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012; 28 : 3150-2.
- Yang B, Wang Y, Qian PY. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* 2016; 17 : 135.
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010; 7 : 335-6.
- Aggarwal A, Sarangi AN, Gaur P, Shukla A, Aggarwal R. Gut microbiome in children with enthesitis-related arthritis in a developing country and the effect of probiotic administration. *Clin Exp Immunol* 2017; 187 : 480-9.
- Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vázquez-Baeza Y, *et al*. Meta-analyses of studies of the human microbiota. *Genome Res* 2013; 23 : 1704-14.
- Wu Y, Chi X, Zhang Q, Chen F, Deng X. Characterization of the salivary microbiome in people with obesity. *PeerJ* 2018; 6 : e4458.
- Singh A, Sarangi AN, Goel A, Srivastava R, Bhargava R, Gaur P, *et al*. Effect of administration of a probiotic preparation on gut microbiota and immune response in healthy women in India: An open-label, single-arm pilot study. *BMC Gastroenterol* 2018; 18 : 85.
- Jangi S, Gandhi R, Cox LM, Li N, von Glehn F, Yan R, *et al*. Alterations of the human gut microbiome in multiple sclerosis. *Nat Commun* 2016; 7 : 12015.
- 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 : 330-9.
- Kim M, Morrison M, Yu Z. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Methods* 2011; 84 : 81-7.
- Edgar RC. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 2018; 34 : 2371-5.
- Edgar RC. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* 2018; 6 : e4652.

For correspondence: Dr Rakesh Aggarwal, Director, Jawaharlal Institute of Postgraduate Medical Education & Research, Dhanvantri Nagar, Puducherry 605 006, India
e-mail: aggarwal.ra@gmail.com

Supplementary Table I. Location and sequences of primers used for extracting data on each hypervariable region (HVR) and combinations of adjacent HVRS

HVR	Location of the forward primer	Sequence of the forward primer	Location of the reverse primer	Sequence of the reverse primer	Reference
V2	119	AGYGGCGNACGGGTGAGTAA	338	TGCTGCCTCCCGTAGGAGT	1
V3	341	CCTACGGGAGGCAGCAG	534	ATTACCGCGGCTGCTGG	2
V4	577	AYTGGGYDTAAAGNG	785	TACNVGGGTATCTAATCC	3,4
V5	784	AGGATTAGATACCCT	907	CCGTCAATTCCTTTGAGTTT	5,6
V6	961	TCGAtGCAACGCGAAGAA	1085	ACATtTCACaACACGAGCTGACGA	7
V7	1099	GYAACGAGCGCAACCC	1238	GTAGRCRCGTGTGTMGCCC	8,9
V8	1050	ATGGCTGTCTCAGCT	1385	ACGGGCGGTGTGTAC	10
V2+V3	119	AGYGGCGNACGGGTGAGTAA	518	ATTACCGCGGCTGCTGG	1,2
V3+V4	341	CCTACGGGRSGCAGCAG	798	GGGGTATCTAATCCC	2,11
V4+V5	577	AYTGGGYDTAAAGNG	907	CCGTCAATTYTTRAGTTT	3,5,6
V5+V6	805	GACTACCAGGGTATCTAATCC	1065	AGGTGCTGCATGGCTGT	9
V6+V7	967	CAACGCGAAGAACCTTACC	1238	GTAGRCRCGTGTGTMGCCC	9
V7+V8	1046	ACAGCCATGCAGCACCT	1406	GACGGGCGGTGWTRCA	9
N, any nucleotide; Y, C or T; D, not C; V, not T; R, A or G; M, A or C; lowercase nucleotide, low confidence base					

References

1. Hsiao WW, Li KL, Liu Z, Jones C, Fraser-Liggett CM, Fouad AF. Microbial transformation from normal oral microbiota to acute endodontic infections. *BMC Genomics* 2012; 13 : 345.
2. He S, Gall DL, McMahon KD. "Candidatus Accumulibacter" population structure in enhanced biological phosphorus removal sludges as revealed by polyphosphate kinase genes. *Appl Environ Microbiol* 2007; 73 : 5865-74.
3. Reddy BV, Kallifidas D, Kim JH, Charlop-Powers Z, Feng Z, Brady SF. Natural product biosynthetic gene diversity in geographically distinct soil microbiomes. *Appl Environ Microbiol* 2012; 78 : 3744-52.
4. Rodrigues JL, Pellizari VH, Mueller R, Baek K, Jesus Eda C, Paula FS, *et al.* Conversion of the Amazon rainforest to agriculture results in biotic homogenization of soil bacterial communities. *Proc Natl Acad Sci U S A* 2013; 110 : 988-93.
5. Sridevi G, Minocha R, Turlapati SA, Goldfarb KC, Brodie EL, Tisa LS, *et al.* Soil bacterial communities of a calcium-supplemented and a reference watershed at the Hubbard Brook Experimental Forest (HBEF), New Hampshire, USA. *FEMS Microbiol Ecol* 2012; 79 : 728-40.
6. Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, Desantis TZ, *et al.* Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World J Gastroenterol* 2010; 16 : 4135-44.
7. 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 : 330-9.

8. Mizrahi-Man O, Davenport ER, Gilad Y. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PLoS One* 2013; 8 : e53608.
9. Youssef N, Sheik CS, Krumholz LR, Najjar FZ, Roe BA, Elshahed MS. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl Environ Microbiol* 2009; 75 : 5227-36.
10. Huws SA, Edwards JE, Kim EJ, Scollan ND. Specificity and sensitivity of eubacterial primers utilized for molecular profiling of bacteria within complex microbial ecosystems. *J Microbiol Methods* 2007; 70 : 565-9.
11. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* 2007; 35 : e120.

Supplementary Table II. Top 10 phyla, class, order and genera of Greengenes, SILVA, RDP and 16S-UDb databases (clustered at 97% threshold)

Rank	Taxon	Greengenes	SILVA	RDP	16S-UDb
Phylum	Proteobacteria	27.69	31.01	29.22	38.24
	Firmicutes	26.84	29.41	27.03	29.91
	Actinobacteria	5.60	8.37	6.60	12.72
	Bacteroidetes	13.05	11.34	12.80	10.44
	Cyanobacteria	2.13	2.09	1.76	1.93
	Tenericutes	0.74	0.57	0.26	1.39
	Spirochaetes	1.45	1.08	1.03	0.83
	Verrucomicrobia	1.76	1.09	1.61	0.47
	Deinococcus-Thermus	N/A	0.19	0.22	0.45
	Planctomycetes	3.42	2.59	2.40	0.38
Class	Gammaproteobacteria	9.11	11.35	10.19	21.20
	Bacilli	4.55	6.48	5.43	17.57
	Actinobacteria	3.73	6.54	6.53	12.59
	Clostridia	21.73	21.48	19.23	11.16
	Alphaproteobacteria	7.69	8.62	7.84	9.26
	Betaproteobacteria	4.04	5.39	4.49	4.56
	Bacteroidia	7.31	5.28	5.55	3.83
	Flavobacteriia	2.23	2.24	2.24	3.56
	Deltaproteobacteria	5.59	4.50	4.78	1.85
	Sphingobacteriia	0.57	1.99	2.27	1.58
Order	Bacillales	2.87	3.65	3.51	12.70
	Actinomycetales	3.59	0.17	4.43	11.90
	Clostridiales	20.98	21.10	18.77	10.32
	Pseudomonadales	1.08	1.94	1.48	5.28

Rank	Taxon	Greengenes	SILVA	RDP	16S-UDb
	Enterobacteriales	0.94	2.32	1.36	4.98
	Lactobacillales	1.60	2.80	1.86	4.79
	Bacteroidales	7.31	5.14	5.55	3.82
	Flavobacteriales	2.22	2.24	2.24	3.56
	Unclassified	8.96	6.53	23.69	3.46
	Burkholderiales	1.99	3.11	2.39	2.97
Family	<i>Enterobacteriaceae</i>	0.94	2.32	1.36	4.97
	<i>Bacillaceae</i>	1.27	1.75	1.23	4.68
	<i>Flavobacteriaceae</i>	1.48	1.86	1.79	3.43
	<i>Pseudomonadaceae</i>	0.62	1.24	0.97	3.36
	<i>Lachnospiraceae</i>	5.78	11.06	7.47	3.30
	<i>Ruminococcaceae</i>	1.27	5.36	5.55	2.58
	<i>Rhodobacteraceae</i>	1.35	1.78	1.49	2.34
	<i>Lactobacillaceae</i>	0.48	0.74	0.50	2.10
	<i>Streptomycetaceae</i>	0.33	0.69	0.41	2.00
	<i>Moraxellaceae</i>	0.46	0.69	0.50	1.90
Genus	<i>Bacillus</i>	0.74	1.28	1.23	5.55
	<i>Pseudomonas</i>	0.49	1.20	0.70	3.27
	<i>Ruminococcus</i>	0.67	0.36	0.30	2.52
	<i>Lactobacillus</i>	0.46	0.72	0.46	1.92
	<i>Prevotella</i>	1.07	0.42	0.98	1.91
	<i>Streptomyces</i>	0.26	0.64	0.46	1.87
	<i>Faecalibacterium</i>	0.28	0.49	0.54	1.83
	<i>Paenibacillus</i>	0.38	0.47	0.32	1.51
	<i>Clostridium</i>	0.80	0.44	0.34	1.43
	<i>Acinetobacter</i>	0.24	0.46	0.23	1.30

Supplementary Table III. Number of taxa contained in a union of the greengenes, silva and RDP databases (*i.e.* ALL) which were not contained in the newly constructed 16S-Udb

Level	Total number of taxa contained in ALL but not in 16S-Udb	Distribution of taxa contained in ALL but not in the newly-constructed 16S-Udb, by the original databases which contained these taxa						
		G	S	R	G∩S	G∩R	S∩R	G∩S∩R
Phylum	97	48	24	6	7	3	9	0
Class	345	159	141	21	22	0	1	1
Order	991	252	667	46	24	0	1	1
Family	1176	226	914	18	15	0	2	1
Genus	2341	334	1341	425	21	42	115	63
G, Greengenes; S, silva; R, RDP; ∩, intersection								

Supplementary Table IV. Top 10 phyla, classes, orders, families and genera contained in the test database

Phylum		Class		Order		Family		Genus	
Name	%	Name	%	Name	%	Name	%	Name	%
Proteobacteria	38.89	Actinobacteria	23.54	Bacillales	8.53	<i>Streptomycetaceae</i>	5.90	<i>Streptomyces</i>	5.61
Actinobacteria	23.69	Gammaproteobacteria	16.63	Streptomycetales	5.89	<i>Flavobacteriaceae</i>	5.06	<i>Bacillus</i>	1.92
Firmicutes	19.51	Alphaproteobacteria	13.25	Micrococcales	5.59	<i>Bacillaceae</i>	4.19	<i>Lactobacillus</i>	1.56
Bacteroidetes	10.76	Bacilli	12.59	Flavobacteriales	5.20	<i>Rhodobacteraceae</i>	3.13	<i>Clostridium</i>	1.52
Tenericutes	1.78	Clostridia	6.00	Clostridiales	5.05	<i>Enterobacteriaceae</i>	2.32	<i>Paenibacillus</i>	1.47
Spirochaetes	0.91	Betaproteobacteria	5.55	Rhizobiales	4.78	<i>Microbacteriaceae</i>	2.23	<i>Pseudomonas</i>	1.46
Deinococcus-Thermus	0.77	Flavobacteriia	5.20	Corynebacteriales	4.14	<i>Paenibacillaceae</i>	2.0	<i>Mycobacterium</i>	1.45
Cyanobacteria	0.54	Deltaproteobacteria	2.48	Lactobacillales	4.06	<i>Pseudonocardiaceae</i>	1.99	<i>Flavobacterium</i>	1.08
Verrucomicrobia	0.44	Sphingobacteriia	2.00	Burkholderiales	3.88	<i>Clostridiaceae</i>	1.98	<i>Mycoplasma</i>	1.06
Thermotogae	0.41	Cytophagia	1.94	Rhodobacterales	3.41	<i>Lactobacillaceae</i>	1.68	<i>Vibrio</i>	0.92

Supplementary Table V. Comparison of family-level abundances (%) between healthy subjects and patients with enthesitis-related arthritis (ERA) using 16S-UDb-MOTHUR approach

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae	69.21	1.00	87.35	9.02	0.03	88.11
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae	5.38	0.96	26.40	4.92	0.11	26.39
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae	5.11	0.53	20.07	10.53	0.09	85.35
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Veillonellaceae	2.54	0.01	8.09	0.77	0.01	12.32
p__Firmicutes; c__Clostridia; o__Clostridiales; f__UC Clostridiales	0.70	0.07	3.61	0.56	2.44E-03	14.75
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae	0.66	0.01	10.85	0.72	3.96E-03	7.54
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Eubacteriaceae	0.59	0.04	7.57	0.74	0.01	7.50
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae	0.57	0.05	1.64	0.23	2.09E-03	16.51
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae	0.33	4.88E-03	2.91	0.15	2.78E-03	3.00
p__Firmicutes; c__UC Firmicutes; o__UC Firmicutes; f__UC Firmicutes	0.26	1.06E-03	1.67	0.26	2.32E-03	5.82
p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae	0.26	0.01	11.97	1.29	0.01	34.53
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Peptostreptococcaceae	0.22	0.01	30.80	0.14	2.09E-03	4.76
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae	0.22	1.98E-03	2.22	0.61	3.48E-03	10.29
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae	0.19	0.01	6.64	1.02	0.01	68.83
p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae	0.10	9.75E-04	3.05	0.06	1.85E-03	7.87
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae	0.10	0.01	5.67	0.10	2.10E-03	20.98
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Sutterellaceae	0.09	5.28E-04	0.69	0.23	3.17E-04	22.46
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae	0.07	0.01	2.77	0.17	0.01	9.86
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae	0.07	3.53E-03	0.37	0.04	7.87E-04	0.92
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae	0.06	0.01	3.08	3.01	0.05	64.04
p__Proteobacteria; c__Gammaproteobacteria; o__Aeromonadales; f__Succinivibrionaceae	0.06	0.00	2.46	0.01	0.00	10.62

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
p__Actinobacteria; c__Coriobacteriia; o__Coriobacteriales; f__Coriobacteriaceae	0.05	7.93E-04	16.20	0.17	3.00E-04	6.71
p__Firmicutes; c__Clostridia; o__Clostridiales; f__UC Clostridiales	0.04	3.17E-03	0.66	0.08	3.48E-04	2.71
p__Bacteroidetes; c__UC Bacteroidetes; o__UC Bacteroidetes; f__UC Bacteroidetes	0.04	0.00	8.04	0.00	0.00	13.74
p__Firmicutes; c__Clostridia; o__UC Clostridia; f__UC Clostridia	0.03	2.64E-04	1.77	0.02	0.00	5.20
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Comamonadaceae	0.03	4.72E-04	0.22	4.52E-03	0.00	7.36
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Acidaminococcaceae	0.02	0.00	0.46	1.58E-03	0.00	0.85
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae	0.01	0.00	0.35	0.10	0.00	6.78
p__Actinobacteria; c__UC Actinobacteria; o__UC Actinobacteria; f__UC Actinobacteria	0.01	0.00	4.21	0.01	0.00	2.35
p__Actinobacteria; c__Actinobacteria; o__Coriobacteriales; f__Coriobacteriaceae	0.01	0.00	0.89	0.01	0.00	0.59
p__Coriobacteriia; c__UC Coriobacteriia; o__Eggerthellales; f__Eggerthellaceae	0.01	0.00	0.47	0.01	0.00	0.49
p__Proteobacteria; c__Deltaproteobacteria; o__Desulfovibrionales; f__Desulfovibrionaceae	4.86E-03	0.00	1.60	2.72E-04	0.00	0.28
p__Firmicutes; c__Clostridia; o__UC Clostridia; f__UC Clostridia	4.23E-03	0.00	0.11	0.00	0.00	0.16
p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Micrococcaceae	1.31E-03	0.00	0.09	2.98E-03	0.00	0.17
p__Proteobacteria; c__UC Proteobacteria; o__UC Proteobacteria; f__UC Proteobacteria	1.28E-03	0.00	0.44	3.51E-03	0.00	8.33
p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Actinomycetaceae	1.07E-03	0.00	0.03	3.90E-03	0.00	1.15
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae 1	9.28E-04	0.00	0.08	3.00E-04	0.00	0.04
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Carnobacteriaceae	8.53E-04	0.00	0.01	2.08E-03	0.00	0.29
p__Tenericutes; c__Mollicutes; o__UC Mollicutes; f__UC Mollicutes	7.49E-04	0.00	8.69	0.01	0.00	12.06
p__Proteobacteria; c__Epsilonproteobacteria; o__Campylobacteriales; f__Campylobacteraceae	6.30E-04	0.00	0.43	9.51E-04	0.00	1.66
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Leuconostocaceae	6.26E-04	0.00	0.09	8.87E-04	0.00	1.63
p__UC Bacteria; c__UC Bacteria; o__UC Bacteria; f__UC Bacteria	3.53E-04	0.00	0.03	2.14E-03	0.00	0.09
p__Firmicutes; c__Bacilli; o__Bacillales; f__Bacillales_Incertae Sedis XI	3.53E-04	0.00	0.01	3.00E-04	0.00	0.17
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Enterococcaceae	2.39E-04	0.00	0.01	0.01	0.00	6.71
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Oxalobacteraceae	2.10E-04	0.00	0.02	0.00	0.00	0.21

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
p__Proteobacteria; c__Deltaproteobacteria; o__UC Deltaproteobacteria; f__UC Deltaproteobacteria	0.00	0.00	0.07	8.16E-04	0.00	1.43
p__Proteobacteria; c__Alphaproteobacteria; o__UC Alphaproteobacteria; f__UC Alphaproteobacteria	0.00	0.00	2.16	5.09E-04	0.00	3.00E-03
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Aerococcaceae	0.00	0.00	2.12E-03	2.83E-04	0.00	0.12
p__Tenericutes; c__Mollicutes; o__Acholeplasmatales; f__Acholeplasmataceae	0.00	0.00	0.06	2.54E-04	0.00	0.17
p__Firmicutes; c__Bacilli; o__UC Bacilli; f__UC Bacilli	0.00	0.00	0.02	1.93E-04	0.00	0.05
p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae	0.00	0.00	0.01	1.48E-04	0.00	0.57
p__Proteobacteria; c__Betaproteobacteria; o__Neisseriales; f__Neisseriaceae	0.00	0.00	2.60E-03	1.48E-04	0.00	0.03
p__Tenericutes; c__Mollicutes; o__Anaeroplasmatales; f__Anaeroplasmataceae	0.00	0.00	0.06	0.00	0.00	0.10
p__Coriobacteriia; c__UC Coriobacteriia; o__Coriobacteriales; f__Atopobacteriaceae	0.00	0.00	0.06	0.00	0.00	1.57
p__Spirochaetes; c__Spirochaetia; o__Spirochaetales; f__Brachyspiraceae	0.00	0.00	0.21	0.00	0.00	3.48E-04
p__Coriobacteriia; c__UC Coriobacteriia; o__Coriobacteriales; f__Coriobacteriaceae	0.00	0.00	0.52	0.00	0.00	0.16
p__Elusimicrobia; c__Elusimicrobia; o__Elusimicrobiales; f__Elusimicrobiaceae	0.00	0.00	0.17	0.00	0.00	0.01
p__Fusobacteria; c__Fusobacteriia; o__Fusobacteriales; f__Fusobacteriaceae	0.00	0.00	6.69E-04	0.00	0.00	1.44
p__Proteobacteria; c__Epsilonproteobacteria; o__Campylobacteriales; f__Helicobacteraceae	0.00	0.00	0.00	0.00	0.00	0.77
p__Tenericutes; c__Mollicutes; o__Mycoplasmatales; f__Mycoplasmataceae	0.00	0.00	0.26	0.00	0.00	0.28
p__Firmicutes; c__Bacilli; o__Bacillales; f__Planococcaceae	0.00	0.00	3.27E-04	0.00	0.00	0.16
p__Firmicutes; c__Bacilli; o__Bacillales; f__Staphylococcaceae	0.00	0.00	0.00	0.00	0.00	0.05
p__Coriobacteriia; c__UC Coriobacteriia; o__UC Coriobacteriia; f__UC Coriobacteriia	0.00	0.00	4.72E-03	0.00	0.00	0.08
p__Verrucomicrobia; c__Verrucomicrobiae; o__Verrucomicrobiales; f__Verrucomicrobiaceae	0.00	0.00	0.01	0.00	0.00	0.69
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__UC Burkholderiales	0.00	0.00	0.33	0.00	0.00	0.11

Supplementary Table VI. Comparison of genus-level abundances (%) between healthy subjects and patients with enthesitis-related arthritis (ERA) using 16S-Udb-MOTHUR approach

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Prevotella copri	58.20	0.78	77.53	8.87	0.02	67.81
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Prevotella stercorea	3.83	0.03	15.63	0.08	1.16E-03	18.81
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Faecalibacterium prausnitzii	3.33	0.51	9.19	4.39	0.07	79.55
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__UC Lachnospiraceae	1.97	0.14	5.37	1.28	0.03	6.25
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Alloprevotella rava	1.43	0.01	5.73	0.01	0.00	8.27
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Prevotella	1.10	4.72E-03	6.27	0.03	6.54E-04	7.13
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Ruminococcus gnavus	1.02	0.21	5.61	1.82	0.03	16.98
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__UC Prevotellaceae	0.95	3.27E-04	3.62	4.53E-03	0.00	2.36
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Veillonellaceae; g__Dialister succinatiphilus	0.91	0.01	7.87	0.22	1.62E-03	7.34
p__Firmicutes; c__Clostridia; o__Clostridiales; f__UC Clostridiales; g__UC Clostridiales	0.70	0.07	3.61	0.56	2.44E-03	14.75
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Roseburia faecis	0.70	0.01	3.23	0.39	0.02	10.65
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae; g__UC Erysipelotrichaceae	0.52	1.89E-03	8.76	0.07	3.60E-04	5.05
p__Firmicutes; c__UC Firmicutes; o__UC Firmicutes; f__UC Firmicutes; g__UC Firmicutes	0.26	1.06E-03	1.67	0.26	2.32E-03	5.82
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Eubacteriaceae; g__Eubacterium eligens	0.25	0.01	1.13	0.28	2.46E-04	4.11
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium cadaveris	0.25	3.30E-03	1.57	0.04	1.09E-03	15.51

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae; g__Mitsuokella multacida	0.22	6.02E-04	1.05	1.74E-03	0.00	1.06
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Peptostreptococcaceae; g__Clostridium bifermentans	0.21	0.01	28.80	0.14	2.09E-03	4.68
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Lachnospira pectinoschiza	0.18	2.44E-03	1.49	0.03	5.18E-04	2.18
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Oscillospira guilliermondii	0.17	3.70E-03	2.00	0.19	0.00	2.03
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__UC Ruminococcaceae	0.17	2.91E-03	5.07	0.06	1.11E-03	4.01
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Subdoligranulum variabile	0.14	0.02	3.51	0.67	3.14E-03	12.37
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Butyrivibrio fibrisolvens	0.13	4.26E-03	0.64	0.09	1.86E-03	2.15
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Intestinimonas butyriciproducens	0.13	0.01	4.54	0.18	1.24E-03	20.38
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Escherichia coli	0.12	0.01	6.63	0.81	0.01	68.35
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Coprococcus catus	0.12	0.03	8.42	0.19	2.66E-03	1.93
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Ruminococcus bromii	0.12	5.28E-04	5.94	0.06	2.21E-04	6.09
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Eubacterium bifforme	0.12	3.96E-04	3.91	0.08	0.00	5.59
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__UC Porphyromonadaceae	0.11	0.00	1.47	1.91E-03	0.00	6.02
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Blautia obeum	0.10	0.01	2.46	0.23	2.51E-03	1.16
p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g__Haemophilus parainfluenzae	0.10	9.75E-04	2.99	0.06	1.85E-03	7.76
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Veillonellaceae; g__Megasphaera indica	0.10	1.66E-03	1.84	2.59E-03	0.00	1.96
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus	0.10	4.88E-04	5.67	0.02	0.00	2.39

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
ruminis						
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Coprococcus eutactus	0.08	1.32E-03	1.30	3.33E-03	0.00	3.90
p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium kashiwanohense	0.06	1.88E-03	2.67	0.32	2.10E-03	10.24
p__Proteobacteria; c__Gammaproteobacteria; o__Aeromonadales; f__Succinivibrionaceae; g__Succinivibrio dextrinosolvens	0.06	0.00	2.46	0.01	0.00	10.62
p__Actinobacteria; c__Coriobacteriia; o__Coriobacteriales; f__Coriobacteriaceae; g__Collinsella aerofaciens	0.05	7.93E-04	16.18	0.16	0.00	6.71
p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium adolescentis	0.05	1.84E-04	8.38	0.05	0.00	8.32
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Eubacteriaceae; g__Eubacterium coprostanoligenes	0.05	5.28E-04	6.56	0.09	4.64E-04	2.65
p__Firmicutes; c__Clostridia; o__Clostridiales; f__UC Clostridiales; g__Flavonifractor plautii	0.04	3.17E-03	0.66	0.08	3.48E-04	2.71
p__Bacteroidetes; c__UC Bacteroidetes; o__UC Bacteroidetes; f__UC Bacteroidetes; g__UC Bacteroidetes	0.04	0.00	8.04	3.00E-03	0.00	13.74
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Dorea formicigenerans	0.04	0.01	2.73	0.06	0.00	1.62
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Eubacteriaceae; g__Eubacterium desmolans	0.03	0.01	1.15	0.08	6.00E-04	2.21
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Turicibacter sanguinis	0.03	9.75E-04	0.37	0.01	0.00	0.88
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Blautia producta	0.03	2.12E-03	0.45	0.03	3.90E-04	0.21
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus	0.03	2.12E-03	2.18	0.08	2.75E-03	4.27
p__Firmicutes; c__Clostridia; o__UC Clostridia; f__UC Clostridia; g__UC Clostridia	0.03	2.64E-04	1.77	0.02	0.00	5.20
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Ruminococcus callidus	0.03	0.00	0.22	1.80E-03	0.00	0.51
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae; g__Veillonella dispar	0.03	9.75E-04	2.50	0.08	1.60E-03	3.00
p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae;	0.02	1.84E-03	3.12	0.37	1.20E-03	34.32

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
g__Bifidobacterium longum						
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Sutterellaceae; g__Sutterella wadsworthensis	0.02	0.00	0.69	0.13	2.84E-04	1.97
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Catenibacterium mitsuokai	0.02	0.00	2.10	2.72E-04	0.00	2.21
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Comamonadaceae; g__UC Comamonadaceae	0.02	4.72E-04	0.22	4.28E-03	0.00	7.36
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Bulleidia p-1630-c5	0.02	0.00	0.22	2.54E-04	0.00	2.59
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides	0.01	2.51E-03	0.19	0.35	0.01	5.95
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium	0.01	3.80E-03	0.10	0.01	0.00	2.82
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium lactatifermentans	0.01	0.00	0.03	0.01	0.00	0.17
p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium angulatum	0.01	0.00	0.12	3.41E-03	0.00	5.22
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides dorei	0.01	1.41E-03	1.46	0.10	1.55E-03	20.71
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Oscillibacter valericigenes	0.01	0.00	0.87	0.01	0.00	2.61
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides uniformis	0.01	0.00	0.52	0.10	7.87E-04	2.75
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Parabacteroides merdae	0.01	0.00	0.20	0.01	0.00	1.70
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__UC Clostridiaceae	0.01	0.00	0.53	5.73E-04	0.00	0.78
p__Actinobacteria; c__UC Actinobacteria; o__UC Actinobacteria; f__UC Actinobacteria; g__UC Actinobacteria	0.01	0.00	4.21	0.01	0.00	2.35
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides ovatus	0.01	0.00	0.19	0.13	0.00	48.14
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Butyrivibrio crossotus	0.01	0.00	1.91	2.73E-03	0.00	8.38
p__Coriobacteriia; c__UC Coriobacteriia; o__Eggerthellales; f__Eggerthellaceae; g__Senegalimassilia anaerobia	0.01	0.00	0.46	4.99E-03	0.00	0.49

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Ruminococcus bicirculans	0.01	0.00	1.93	2.21E-03	0.00	11.24
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Peptostreptococcaceae; g__Peptoclostridium difficile	0.01	0.00	2.00	3.55E-03	0.00	0.11
p__Actinobacteria; c__Actinobacteria; o__Coriobacteriales; f__Coriobacteriaceae; g__Slackia isoflavoniconvertens	0.01	0.00	0.47	3.44E-03	0.00	0.35
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides acidifaciens	0.01	0.00	0.04	0.08	0.00	6.30
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus infantis	4.76E-03	1.20E-03	0.17	0.02	1.37E-03	3.31
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Klebsiella pneumoniae	4.24E-03	1.84E-04	0.04	0.02	7.13E-04	50.63
p__Firmicutes; c__Clostridia; o__UC Clostridia; f__UC Clostridia; g__Howardella ureilytica	4.23E-03	0.00	0.11	0.00	0.00	0.16
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Prevotella ruminicola	4.07E-03	0.00	0.43	1.24E-03	0.00	1.65
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Odoribacter splanchnicus	3.75E-03	0.00	0.15	0.02	0.00	2.39
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus luteciae	3.62E-03	0.00	0.04	0.02	0.00	4.66
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides caccae	3.27E-03	0.00	0.59	0.03	0.00	15.86
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Parabacteroides distasonis	3.22E-03	0.00	0.57	0.03	1.66E-04	7.90
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus salivarius	2.81E-03	0.00	0.01	1.73E-04	0.00	0.10
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Sutterellaceae; g__Sutterella stercoricanis	2.27E-03	0.00	0.35	3.44E-03	0.00	22.14
p__Proteobacteria; c__Deltaproteobacteria; o__Desulfovibrionales; f__Desulfovibrionaceae; g__Desulfovibrio piger	2.25E-03	0.00	1.29	0.00	0.00	0.26
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Roseburia	2.18E-03	0.00	0.01	1.21E-03	0.00	0.01
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__Alistipes	2.13E-03	0.00	0.22	0.03	0.00	3.05

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
finegoldii						
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Comamonadaceae; g__Comamonas	1.91E-03	0.00	0.02	0.00	0.00	0.09
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Eubacteriaceae; g__Eubacterium ventriosum	1.89E-03	0.00	0.03	0.01	0.00	0.17
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Ruminococcus	1.71E-03	0.00	0.09	8.16E-04	0.00	2.59
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium butyricum	1.58E-03	0.00	0.16	2.29E-04	0.00	0.66
p__Actinobacteria; c__Actinobacteria; o__Coriobacteriales; f__Coriobacteriaceae; g__UC Coriobacteriaceae	1.51E-03	0.00	0.31	2.80E-04	0.00	0.44
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Enterobacter hormaechei	1.36E-03	0.00	0.05	0.01	0.00	5.08
p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Micrococcaceae; g__Rothia mucilaginosa	1.31E-03	0.00	0.09	2.98E-03	0.00	0.17
p__Proteobacteria; c__UC Proteobacteria; o__UC Proteobacteria; f__UC Proteobacteria; g__UC Proteobacteria	1.28E-03	0.00	0.44	3.51E-03	0.00	8.33
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides fragilis	1.25E-03	0.00	0.07	0.01	0.00	17.90
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Eubacteriaceae; g__Eubacterium	1.24E-03	0.00	0.07	2.87E-04	0.00	0.24
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae; g__Veillonella parvula	1.12E-03	0.00	0.41	2.12E-03	0.00	0.11
p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Actinomycetaceae; g__Actinomyces odontolyticus	1.07E-03	0.00	0.03	3.90E-03	0.00	1.15
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Veillonellaceae; g__Allisonella histaminiformans	1.04E-03	0.00	0.01	0.00	0.00	0.05
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Cronobacter sakazakii	1.03E-03	0.00	0.05	1.45E-03	0.00	1.53
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus sanguinis	1.00E-03	0.00	0.40	0.01	0.00	0.15
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Prevotella salivae	9.32E-04	0.00	0.98	3.48E-04	0.00	0.36
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae 1; g__Clostridium	9.28E-04	0.00	0.08	3.00E-04	0.00	0.04

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
perfringens						
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Clostridium spiroforme	9.16E-04	0.00	0.17	3.50E-03	0.00	0.53
p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g__Actinobacillus parahaemolyticus	8.89E-04	0.00	0.05	1.73E-04	0.00	0.12
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Carnobacteriaceae; g__Granulicatella adiacens	8.53E-04	0.00	0.01	2.08E-03	0.00	0.29
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__Alistipes putredinis	7.96E-04	0.00	0.08	0.01	0.00	3.50
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Veillonellaceae; g__Megamonas rupellensis	7.90E-04	0.00	4.67	4.18E-03	2.87E-04	4.76
p__Tenericutes; c__Mollicutes; o__UC Mollicutes; f__UC Mollicutes; g__UC Mollicutes	7.49E-04	0.00	8.69	0.01	0.00	12.06
p__Proteobacteria; c__Deltaproteobacteria; o__Desulfovibrionales; f__Desulfovibrionaceae; g__Bilophila wadsworthia	6.88E-04	0.00	0.31	0.00	0.00	0.28
p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g__Haemophilus pittmaniae	6.78E-04	0.00	0.01	1.48E-04	0.00	0.10
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Leuconostocaceae; g__Weissella confusa	6.26E-04	0.00	0.05	6.00E-04	0.00	1.61
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Butyricoccus pullicaecorum	5.88E-04	0.00	0.26	2.80E-04	0.00	0.38
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Barnesiella intestinihominis	5.71E-04	0.00	0.21	5.42E-04	0.00	2.25
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Enterobacter cloacae	5.60E-04	0.00	0.12	2.22E-03	0.00	0.82
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium propionicum	4.37E-04	0.00	0.01	3.90E-04	0.00	0.98
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Acetanaerobacterium elongatum	4.13E-04	0.00	0.02	2.71E-04	0.00	0.07
p__Firmicutes; c__Bacilli; o__Bacillales; f__Bacillales_Incertae Sedis XI; g__Gemella palaticanis	3.53E-04	0.00	0.01	3.00E-04	0.00	0.17
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__UC Enterobacteriaceae	2.96E-04	0.00	0.01	1.68E-03	0.00	1.12

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides plebeius	2.59E-04	0.00	0.99	0.01	0.00	52.68
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Acidaminococcaceae; g__Phascolarctobacterium succinatutens	2.44E-04	0.00	0.38	2.29E-04	0.00	0.85
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Enterococcaceae; g__Enterococcus faecium	2.39E-04	0.00	0.01	3.54E-03	0.00	6.68
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__Alistipes senegalensis	2.35E-04	0.00	0.02	9.34E-05	0.00	0.13
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Ruminococcus champanellensis	2.24E-04	0.00	0.12	0.00	0.00	2.45
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Oxalobacteraceae; g__Oxalobacter formigenes	2.10E-04	0.00	0.02	0.00	0.00	0.21
p__Actinobacteria; c__Coriobacteriia; o__Coriobacteriales; f__Coriobacteriaceae; g__Atopobium rimae	1.64E-04	0.00	0.04	0.00	0.00	1.14
p__Proteobacteria; c__Epsilonproteobacteria; o__Campylobacterales; f__Campylobacteraceae; g__Campylobacter concisus	1.47E-04	0.00	0.01	3.10E-04	0.00	0.11
p__Proteobacteria; c__Epsilonproteobacteria; o__Campylobacterales; f__Campylobacteraceae; g__Campylobacter	1.18E-04	0.00	0.42	2.54E-04	0.00	1.66
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Eubacteriaceae; g__Eubacterium siraeum	9.20E-05	0.00	0.99	1.33E-03	0.00	1.65
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__Alistipes obesi	9.20E-05	0.00	0.15	1.89E-04	0.00	0.20
p__Actinobacteria; c__Actinobacteria; o__Coriobacteriales; f__Coriobacteriaceae; g__Slackia exigua	9.20E-05	0.00	0.10	0.00	0.00	0.08
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Hafnia alvei	7.35E-05	0.00	0.04	4.85E-04	0.00	0.31
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus delbrueckii	0.00	0.00	0.02	4.53E-03	0.00	3.38
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium leptum	0.00	0.00	0.05	3.57E-03	0.00	0.32
p__UC Bacteria; c__UC Bacteria; o__UC Bacteria; f__UC Bacteria; g__Candidatus Saccharibacteria oral taxon	0.00	0.00	0.03	2.14E-03	0.00	0.06
p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae;	0.00	0.00	0.61	2.10E-03	0.00	1.01

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
g__Bifidobacterium bifidum						
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides coprocola	0.00	0.00	0.04	1.78E-03	0.00	9.26
p__Actinobacteria; c__Coriobacteriia; o__Coriobacteriales; f__Coriobacteriaceae; g__Eggerthella lenta	0.00	0.00	0.12	1.04E-03	0.00	0.75
p__Proteobacteria; c__Deltaproteobacteria; o__UC Deltaproteobacteria; f__UC Deltaproteobacteria; g__UC Deltaproteobacteria	0.00	0.00	0.07	8.16E-04	0.00	1.43
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides nordii	0.00	0.00	9.82E-04	6.00E-04	0.00	0.74
p__Proteobacteria; c__Alphaproteobacteria; o__UC Alphaproteobacteria; f__UC Alphaproteobacteria; g__UC Alphaproteobacteria	0.00	0.00	2.16	5.09E-04	0.00	0.00
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Sutterellaceae; g__Parasutterella excrementihominis	0.00	0.00	0.03	3.90E-04	0.00	0.54
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Veillonellaceae; g__Dialister invisus	0.00	0.00	0.04	3.90E-04	0.00	12.13
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus mucosae	0.00	0.00	0.02	3.48E-04	0.00	0.46
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Bulleidia moorei	0.00	0.00	9.75E-04	3.31E-04	0.00	0.10
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__Alistipes indistinctus	0.00	0.00	0.05	2.87E-04	0.00	3.58
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Clostridium ramosum	0.00	0.00	2.93E-03	2.87E-04	0.00	0.15
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Veillonellaceae; g__Megasphaera micronuciformis	0.00	0.00	0.05	2.84E-04	0.00	0.12
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Aerococcaceae; g__Abiotrophia defectiva	0.00	0.00	2.12E-03	2.83E-04	0.00	0.12
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Holdemania massiliensis	0.00	0.00	0.01	2.71E-04	0.00	0.08
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Veillonellaceae; g__Veillonella criceti	0.00	0.00	7.87E-04	2.71E-04	0.00	2.31
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae;	0.00	0.00	0.01	2.54E-04	0.00	0.02

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
g_Eubacterium dolichum						
p_Tenericutes; c_Mollicutes; o_Acholeplasmatales; f_Acholeplasmataceae; g_UC Acholeplasmataceae	0.00	0.00	0.06	2.54E-04	0.00	0.17
p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Streptococcaceae; g_Streptococcus anginosus	0.00	0.00	6.81E-04	2.46E-04	0.00	0.05
p_Bacteroidetes; c_Bacteroidia; o_Bacteroidales; f_Bacteroidaceae; g_Bacteroides eggerthii	0.00	0.00	1.00E-03	2.29E-04	0.00	2.49
p_Bacteroidetes; c_Bacteroidia; o_Bacteroidales; f_Porphyrimonadaceae; g_Butyricimonas virosa	0.00	0.00	0.14	2.19E-04	0.00	0.72
p_Firmicutes; c_Bacilli; o_UC Bacilli; f_UC Bacilli; g_Gemella haemolysans	0.00	0.00	0.02	1.93E-04	0.00	0.05
p_Proteobacteria; c_Gammaproteobacteria; o_Enterobacteriales; f_Enterobacteriaceae; g_Shigella sonnei	0.00	0.00	1.95E-03	1.89E-04	0.00	0.05
p_Bacteroidetes; c_Bacteroidia; o_Bacteroidales; f_Rikenellaceae; g_Alistipes	0.00	0.00	0.02	1.48E-04	0.00	0.05
p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Lactobacillaceae; g_Lactobacillus fermentum	0.00	0.00	2.35E-04	1.48E-04	0.00	0.09
p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Streptococcaceae; g_Lactococcus lactis	0.00	0.00	0.26	1.48E-04	0.00	1.56
p_Firmicutes; c_Negativicutes; o_Selenomonadales; f_Veillonellaceae; g_Megamonas	0.00	0.00	1.20	1.48E-04	0.00	0.94
p_Proteobacteria; c_Betaproteobacteria; o_Neisseriales; f_Neisseriaceae; g_Neisseria meningitidis	0.00	0.00	2.60E-03	1.48E-04	0.00	0.03
p_Firmicutes; c_Negativicutes; o_Selenomonadales; f_Acidaminococcaceae; g_Acidaminococcus fermentans	0.00	0.00	0.46	0.00	0.00	0.00
p_Proteobacteria; c_Gammaproteobacteria; o_Pseudomonadales; f_Moraxellaceae; g_Acinetobacter	0.00	0.00	2.01E-03	0.00	0.00	0.04
p_Proteobacteria; c_Gammaproteobacteria; o_Pseudomonadales; f_Moraxellaceae; g_Acinetobacter baumannii	0.00	0.00	3.34E-04	0.00	0.00	0.03
p_Proteobacteria; c_Gammaproteobacteria; o_Pseudomonadales; f_Moraxellaceae; g_Acinetobacter johnsonii	0.00	0.00	2.34E-03	0.00	0.00	0.09
p_Proteobacteria; c_Gammaproteobacteria; o_Pseudomonadales; f_Moraxellaceae; g_Acinetobacter lwoffii	0.00	0.00	2.29E-03	0.00	0.00	0.55
p_Verrucomicrobia; c_Verrucomicrobiae; o_Verrucomicrobiales;	0.00	0.00	0.01	0.00	0.00	0.69

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
f__Verrucomicrobiaceae; g__Akkermansia muciniphila						
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__Alistipes massiliensis	0.00	0.00	4.47E-04	0.00	0.00	0.13
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__Alistipes onderdonkii	0.00	0.00	0.08	0.00	0.00	0.29
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Anaerobacterium chartisolvans	0.00	0.00	0.13	0.00	0.00	0.51
p__Tenericutes; c__Mollicutes; o__Anaeroplasmatales; f__Anaeroplasmataceae; g__Asteroleplasma anaerobium	0.00	0.00	0.06	0.00	0.00	0.10
p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium breve	0.00	0.00	1.32E-04	0.00	0.00	0.23
p__Spirochaetes; c__Spirochaetia; o__Spirochaetales; f__Brachyspiraceae; g__Brachyspira corvi	0.00	0.00	0.21	0.00	0.00	0.00
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Butyricimonas paravirosa	0.00	0.00	1.34E-03	0.00	0.00	0.21
p__UC Bacteria; c__UC Bacteria; o__UC Bacteria; f__UC Bacteria; g__Candidatus Saccharimonas aalborgensis	0.00	0.00	0.01	0.00	0.00	0.03
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium neonatale	0.00	0.00	2.25E-03	0.00	0.00	0.18
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium perfringens	0.00	0.00	7.05E-04	0.00	0.00	0.17
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium sporosphaeroides	0.00	0.00	0.08	0.00	0.00	0.35
p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium termitidis	0.00	0.00	0.02	0.00	0.00	0.06
p__Coriobacteriia; c__UC Coriobacteriia; o__Coriobacteriales; f__Coriobacteriaceae; g__Collinsella tanakaei	0.00	0.00	0.52	0.00	0.00	0.16
p__Elusimicrobia; c__Elusimicrobia; o__Elusimicrobiales; f__Elusimicrobiaceae; g__Elusimicrobium minutum	0.00	0.00	0.17	0.00	0.00	0.01
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Enterobacter	0.00	0.00	1.00E-03	0.00	0.00	0.09
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales;	0.00	0.00	0.08	0.00	0.00	3.22

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
f__Enterobacteriaceae; g__Enterobacter asburiae						
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Enterococcaceae; g__Enterococcus faecalis	0.00	0.00	0.01	0.00	0.00	2.92
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Enterococcaceae; g__Enterococcus gallinarum	0.00	0.00	0.00	0.00	0.00	0.13
p__Fusobacteria; c__Fusobacteriia; o__Fusobacteriales; f__Fusobacteriaceae; g__Fusobacterium varium	0.00	0.00	6.69E-04	0.00	0.00	1.44
p__Actinobacteria; c__Actinobacteria; o__Coriobacteriales; f__Coriobacteriaceae; g__Gordonibacter pamelaeae	0.00	0.00	0.01	0.00	0.00	0.07
p__Proteobacteria; c__Epsilonproteobacteria; o__Campylobacterales; f__Helicobacteraceae; g__Helicobacter macacae	0.00	0.00	0.00	0.00	0.00	0.77
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Klebsiella oxytoca	0.00	0.00	0.02	0.00	0.00	0.95
p__Firmicutes; c__Bacilli; o__Bacillales; f__Planococcaceae; g__Kurthia gibsonii	0.00	0.00	3.27E-04	0.00	0.00	0.16
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus	0.00	0.00	0.01	0.00	0.00	0.04
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus gasseri	0.00	0.00	1.64E-03	0.00	0.00	0.44
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus oris	0.00	0.00	1.31E-03	0.00	0.00	0.06
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus plantarum	0.00	0.00	0.01	0.00	0.00	20.16
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus pontis	0.00	0.00	4.72E-03	0.00	0.00	0.09
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Morganella morganii	0.00	0.00	1.20E-03	0.00	0.00	0.26
p__Tenericutes; c__Mollicutes; o__Mycoplasmatales; f__Mycoplasmataceae; g__Mycoplasma fermentans	0.00	0.00	0.19	0.00	0.00	0.21
p__Tenericutes; c__Mollicutes; o__Mycoplasmatales; f__Mycoplasmataceae; g__Mycoplasma orale	0.00	0.00	0.07	0.00	0.00	0.07
p__Coriobacteriia; c__UC Coriobacteriia; o__Coriobacteriales; f__Atopobacteriaceae; g__Olsenella scatoligenes	0.00	0.00	0.06	0.00	0.00	1.57
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Parabacteroides goldsteinii	0.00	0.00	0.02	0.00	0.00	0.44

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Paraprevotella clara	0.00	0.00	0.04	0.00	0.00	0.82
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Pediococcus pentosaceus	0.00	0.00	1.34E-03	0.00	0.00	0.18
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Acidaminococcaceae; g__Phascolarctobacterium faecium	0.00	0.00	0.34	0.00	0.00	0.50
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Porphyromonas asaccharolytica	0.00	0.00	0.01	0.00	0.00	0.07
p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Prevotella buccalis	0.00	0.00	6.81E-04	0.00	0.00	0.07
p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Proteus mirabilis	0.00	0.00	0.00	0.00	0.00	0.08
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Veillonellaceae; g__Selenomonas	0.00	0.00	0.24	0.00	0.00	0.00
p__Firmicutes; c__Erysipelotrichia; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Sharpea p-3329-23G2	0.00	0.00	0.17	0.00	0.00	0.00
p__Firmicutes; c__Bacilli; o__Bacillales; f__Staphylococcaceae; g__Staphylococcus gallinarum	0.00	0.00	0.00	0.00	0.00	0.05
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus salivarius	0.00	0.00	0.01	0.00	0.00	0.10
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Sutterellaceae; g__Sutterella	0.00	0.00	1.79E-03	0.00	0.00	0.08
p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__UC Burkholderiales; g__UC Burkholderiales	0.00	0.00	0.33	0.00	0.00	0.11
p__Coriobacteriia; c__UC Coriobacteriia; o__UC Coriobacteriia; f__UC Coriobacteriia; g__UC Coriobacteriia	0.00	0.00	4.72E-03	0.00	0.00	0.08
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Veillonellaceae; g__Veillonella	0.00	0.00	0.00	0.00	0.00	0.37
p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Leuconostocaceae; g__Weissella paramesenteroides	0.00	0.00	0.04	0.00	0.00	0.30

Taxonomic group	Healthy persons			Patients with ERA		
	Median	Minimum	Maximum	Median	Minimum	Maximum
p__Coriobacteriia; c__UC Coriobacteriia; o__Eggerthellales; f__Eggerthellaceae; g__UC Eggerthellaceae	0.00	0.00	2.34E-03	0.00	0.00	0.20
p__Firmicutes; c__Negativicutes; o__Selenomonadales; f__Veillonellaceae; g__UC Veillonellaceae	0.00	0.00	0.43	0.00	0.00	2.21E-03