Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution

Yoshiyuki Matsuo[1*], Shinnosuke Komiya[2,3], Yoshiaki Yasumizu[4,5], Yuki Yasuoka[5], Katsura Mizushima[6], Tomohisa Takagi[6], Kirill Kryukov[7], Aisaku Fukuda[8], Yoshiharu Morimoto[2], Yuji Naito[6], Hidetaka Okada[3], Hidemasa Bono[9,10], So Nakagawa[7] and Kiichi Hirota[1]

[1] Department of Human Stress Response Science, Institute of Biomedical Science, Kansai Medical University, Hirakata, Japan.

[2] HORAC Grand Front Osaka Clinic, Osaka, Japan.

[3] Obstetrics and Gynecology, Kansai Medical University Graduate School of Medicine, Hirakata, Japan.

[4] Department of Experimental Immunology, Immunology Frontier Research Center, Osaka University, Osaka, Japan.

[5] Faculty of Medicine, Osaka University, Osaka, Japan.

[6] Molecular Gastroenterology and Hepatology, Kyoto Prefectural University of Medicine, Kyoto, Japan.

[7]Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Japan.

[8] IVF Osaka Clinic, Osaka, Japan.

[9] Database Center for Life Science (DBCLS), Research Organization of Information and Systems, Mishima, Japan.

[10] Program of Biomedical Science, Graduate School of Integrated Sciences for Life, Hiroshima University, Higashi-Hiroshima, Japan.

* Correspondence:

Yoshiyuki Matsuo

ysmatsuo-kyt@umin.ac.jp

[1]Department of Human Stress Response Science, Institute of Biomedical Science, Kansai Medical University, 2-5-1 Shin-machi, Hirakata, Osaka 573-1010, Japan.

Abstract

Species-level genetic characterization of complex bacterial communities has important clinical applications. In the present study, we assessed the performance of full-length 16S rRNA gene analysis of human gut microbiota using the nanopore long-read sequencer MinION™. A new strategy for library construction with an optimized primer set overcame PCR-associated bias and produced accurate taxonomic classifications of a broad range of bacterial species. Our present microbiome study, comparing the discriminatory power of full-length and short-read sequencing, clearly illustrated the analytical advantage of sequencing the full-length 16S rRNA gene, which provided higher species-level resolution and accuracy.

Keywords:

16S rRNA, gut microbiota, metagenome, MinION™, nanopore sequencing

Background

Recent advances in DNA sequencing technology have had a revolutionary impact on clinical microbiology [1]. Next-generation sequencing (NGS) technology enables parallel sequencing of DNA on a massive scale to generate vast quantities of accurate data. NGS platforms are now increasingly used in the field of clinical research [2]. Metagenomic sequencing offers numerous advantages over traditional culture-based techniques that have long been the standard test for detecting pathogenic bacteria. This method is particularly useful for characterizing uncultivable bacteria and novel pathogens [3].

Among the metagenomic sequencing strategies, amplicon sequencing of the 16S ribosomal RNA (rRNA) gene has proven to be a reliable and efficient option for taxonomic classification [4, 5]. The bacterial 16S rRNA gene contains nine variable regions (V1 to V9) that are separated by highly conserved sequences across different taxa. For bacterial identification, the 16S rRNA gene is first amplified by polymerase chain reaction (PCR) with primers annealing to conserved regions and then sequenced. The sequencing data are subjected to bioinformatic analysis in which the variable regions are used to discriminate between bacterial taxa [6].

Since the conventional parallel-type short-read sequencer cannot yield reads covering the full length of the 16S rRNA gene [7], several regions of it have been targeted for sequencing, which often causes ambiguity in taxonomic classification [8]. New sequencing platforms have overcome these technical restrictions, particularly those affecting read length. A prime example is the MinION™ sequencer from Oxford Nanopore Technologies, which is capable of producing long sequences with no theoretical read length limit [9-11]. MinION™ sequencing targets the entire 16S rRNA gene, allowing the identification of bacteria with more accuracy and sensitivity [12, 13]. Furthermore, MinION™ produces sequencing data in real time, which reduces turnaround time for data processing [14, 15].

Given these features of MinION™ sequencing, we had previously conducted full-length 16S amplicon sequencing analyses using the MinION™ platform coupled to a bioinformatics pipeline, which allowed us to identify bacterial pathogens with a total analysis time of under two hours [16]. However, we also found that the approach of using the commercial 16S Barcoding Kit (SQK-RAB204) available from Oxford Nanopore Technologies has a limited ability to detect particular taxa such as *Bifidobacterium* [16]. This is probably due to sequence mismatches in the primer used for 16S gene amplification [17]. *Bifidobacterium* play an essential role in maintaining gut homeostasis, and deviations or aberrancies in the *Bifidobacterium* composition have been closely associated with human diseases including obesity, allergy, and inflammatory disorders [18]. Based on their health-promoting effects, several strains of *Bifidobacterium* have been utilized as probiotics [19]. Within these contexts, the species-level characterization of *Bifidobacterium* diversity in human gut microbiota is very important in clinical practice.

Our metagenomic sequencing approach using MinION™ has been tested only with pre-characterized mock bacterial DNA and a limited number of pathogenic bacteria from a patient-derived sample [20]. Its applicability to highly complex bacterial communities has not yet been thoroughly investigated. Therefore, in this study we modified our existing protocol for 16S amplicon sequencing by MinION™ and applied it to human gut microbiota with a complex bacterial composition [21], including *Bifidobacterium*, to determine whether full-length 16S rRNA gene sequencing with MinION™ is an effective characterization tool.

Methods

Mock bacterial community DNA

A mixture of bacterial DNA (10 Strain Even Mix Genomic Material, MSA-1000) was obtained from the American Type Culture Collection (ATCC, Manassas, VA, USA), comprising genomic DNA prepared from the following ten bacterial strains: *Bacillus cereus* (ATCC 10987), *Bifidobacterium adolescentis* (ATCC 15703), *Clostridium beijerinckii* (ATCC 35702), *Deinococcus radiodurans* (ATCC BAA-816), *Enterococcus faecalis* (ATCC 47077), *Escherichia coli* (ATCC 700926), *Lactobacillus gasseri* (ATCC 33323), *Rhodobacter sphaeroides* (ATCC 17029), *Staphylococcus epidermidis* (ATCC 12228), and *Streptococcus mutans* (ATCC 700610).

Fecal DNA

DNA was extracted from six human fecal samples using the NucleoSpin® Microbial DNA Kit (Macherey-Nagel, Düren, Germany), as described previously [22]. Briefly, human feces stored using the Feces Collection Kit (Techno Suruga Lab, Shizuoka, Japan) were subjected to mechanical disruption by bead-beating, and DNA was isolated using silica membrane spin columns. Extracted DNA was purified with the Agencourt AMPure® XP (Beckman Coulter, Brea, CA, USA).

Metagenomic sequencing on the MinION™ platform

Four-primer PCR with rapid adapter attachment chemistry generated 16S gene amplicons with modified 5′ ends for simplified post-PCR adapter attachment following the manufacturer's instructions with slight modifications. For amplification of the V1-V9 region of the 16S rRNA gene, the following inner primers were used, with 16S rRNA gene-specific sequences underlined: forward primer (S-D-Bact-0008-c-S-20 [23]) with anchor sequence 5′-TTTCTGTTGGTGCTGATATTGC<u>AGRGTTYGATYMTGGCTCAG</u>-3′ and reverse

124 primer (1492R) with anchor sequence 5′-
125 ACTTGCCTGTCGCTCTATCTTCCGGYTACCTTGTTACGACTT-3′. For
126 amplification of the V3-V4 region, the following inner primers were used, with 16S
127 rRNA gene-specific sequences underlined: 341F with anchor sequence 5′-
128 TTTCTGTTGGTGCTGATATTGCCCTACGGGNGGCWGCAG-3′ and 806R with
129 anchor sequence 5′-
130 ACTTGCCTGTCGCTCTATCTTCGGACTACHVGGGTWTCTAAT-3′. PCR
131 amplification of 16S rRNA genes was conducted using the KAPA2G™ Robust HotStart
132 ReadyMix PCR Kit (Kapa Biosystems, Wilmington, MA, USA) in a total volume of 25
133 µl containing inner primer pairs (50 nM each) and the barcoded outer primer mixture
134 (3%) from the PCR Barcoding Kit (SQK-PBK004; Oxford Nanopore Technologies,
135 Oxford, UK). Amplification was performed with the following PCR conditions: initial
136 denaturation at 95 °C for 3 min, 5 cycles of 95 °C for 15 sec, 55 °C for 15 sec, and
137 72 °C for 30 sec, 30 cycles of 95 °C for 15 sec, 62 °C for 15 sec, and 72 °C for 30 sec,
138 followed by a final extension at 72 °C for 1 min. Amplified DNA was purified using
139 AMPure® XP (Beckman Coulter) and quantified by a NanoDrop® 1000 (Thermo
140 Fischer Scientific, Waltham, MA, USA). A total of 100 ng of DNA was incubated with
141 1 µl of Rapid Adapter at room temperature for 5 min. The prepared DNA library (11 µl)
142 was mixed with 34 µl of Sequencing Buffer, 25.5 µl of Loading Beads, and 4.5 µl of
143 water, loaded onto the R9.4 flow cell (FLO-MIN106; Oxford Nanopore Technologies),
144 and sequenced on the MinION™ Mk1B. MINKNOW software ver. 1.11.5 (Oxford
145 Nanopore Technologies) was used for data acquisition.
146
147 **Metagenomic sequencing on the MiSeq™ platform**
148 Sequencing libraries were constructed as described previously [22]. Briefly, the V3-V4
149 regions of the 16S rRNA gene were amplified using a 16S (V3–V4) Metagenomic
150 Library Construction Kit for NGS (Takara Bio Inc, Kusatsu, Japan). The following

151 primers were used (16S rRNA gene-specific sequences are underlined): 341F with
152 overhang adapter 5′-
153 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-
154 3′ and 806R with overhang adapter 5′-
155 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGACTACHVGGGTWTCT
156 AAT-3′. The second PCR was performed using the Nextera® XT Index Kit (Illumina,
157 San Diego, CA, USA) for sample multiplexing with index adapters. The libraries were
158 sequenced on the MiSeq™ platform using the MiSeq™ Reagent Kit v3 (2 × 250 bp;
159 Illumina).
160
161 **Bioinformatics analysis**
162 Albacore software ver. 2.3.4 (Oxford Nanopore Technologies) was used for basecalling
163 the MinION™ sequencing data (FAST5 files) to generate pass reads (FASTQ format)
164 with a mean quality score > 7. The adapter and barcode sequences were trimmed using
165 the EPI2ME Fastq Barcoding workflow ver. 3.10.4 (Oxford Nanopore Technologies).
166 The reads were filtered by size using SeqKit software ver. 0.10.0 [24], retaining 1300-
167 1950 bp sequences for the V1-V9 region and 350-600 bp sequences for the V3-V4
168 region, based on the size distribution of 16S rRNA gene sequences in the SILVA
169 database ver. 132 [25, 26]. The processed reads from each set were analyzed using our
170 bioinformatics pipeline [27], as described previously [14, 15]. Briefly, FASTQ files
171 were converted to FASTA files. Simple repetitive sequences were masked using the
172 TANTAN program ver. 18 with default parameters [28]. To remove reads derived from
173 human DNA, we searched each read against the human genome (GRCh38) using
174 minimap2 ver. 2.14 with a map-ont-option [29]. Then, unmatched reads were regarded
175 as reads derived from bacteria. For each read, a minimap2 search with 5850
176 representative bacterial genome sequences stored in the GenomeSync database
177 (Additional File 1) [30] was performed. For each read, we kept the alignment with the

178 highest minimap2 score, and taxa were determined based on the NCBI taxonomy
179 database [31]. Low-abundance taxa with less than 0.01% of total reads were discarded
180 from the analysis.
181
182 **Statistical analyses**
183 Differences between groups were evaluated by one-way analysis of variance (ANOVA)
184 followed by Dunnett's test for multiple comparisons. The Pearson correlation
185 coefficient was computed to compare the bacterial compositions analyzed by different
186 sequencing methods. Statistical significance was defined by a P-value < 0.05. Statistical
187 analyses were performed with Prism8 (GraphPad Software, Inc. La Jolla, CA, USA).
188
189 **Statement of ethics**
190 The Sunkaky Institutional Review Board approved this study (No. 2017-27). All
191 participants provided written informed consent.
192

Results

Classification of the mock bacterial community

The 16S rRNA gene sequence of *Bifidobacterium* has three base mismatches with the 27F forward primer provided in the commercial sequencing kit (16S Barcoding Kit, SQK-RAB204, Oxford Nanopore Technologies; Additional File 2: Supplementary Fig. S1a), which biases amplification toward underrepresentation of *Bifidobacterium* species (Additional File 2: Supplementary Fig S2, Additional File 3: Supplementary Table S1-S3). To overcome this drawback, we introduced three degenerate bases to the 16S rRNA gene-specific sequences of the primer (Additional File 2: Supplementary Fig. S1b). The competence of the modified primer set was then evaluated by metagenomic analysis of a ten-species mock community. The V1-V9 region of the 16S rRNA gene was amplified by the four-primer PCR method with the rapid adapter attachment chemistry and sequenced (Fig. 1a). MinION™ sequencing generated 8651 pass reads (Table 1). Following adapter trimming and size selection, 6972 reads (80.6% of pass reads with an average lead length of 1473 bp) were retained for bacterial identification. Full-length 16S amplicon sequencing with the modified primer set led to the successful identification of all the ten bacterial species, including *Bifidobacterium adolescentis* (Fig. 1b, 1c, Additional File 3: Supplementary Table S4). At the species level, 97.2% of analyzed reads were correctly assigned to each bacterial taxon included in the mock community, demonstrating the excellent discriminatory power of this full-length sequencing method for species identification (Fig. 1c).

We compared the resolution of full-length and short-read 16S amplicon sequencing for the taxonomic classification of bacteria. The V3-V4 region was amplified by four-primer PCR from the ten-species mock community DNA, and the samples were sequenced on MinION™. After removing the adapter/barcode sequences and filtering reads by length, 96189 reads with an average length of 454.9 bp for downstream analysis were yielded (Table 1). In contrast to full-length sequencing with the highest

220 resolution, a significant number of V3-V4 reads could not be classified down to the
221 genus level, but could only be assigned to a higher taxonomic rank (Fig. 1b, 1c). Most
222 reads derived from *Enterococcus faecalis* and *Escherichia coli* were not assigned down
223 to each taxon, as more than two species produced the same similarity score for the V3-
224 V4 sequence read queries. Such reads were classified to the parent taxon, as more
225 specific classification was impossible (Additional File 3: Supplementary Table S5). The
226 classifications were not affected by increasing the number of analyzed reads to 10000
227 (Additional File 2: Supplementary Fig. S3, Additional File 3: Supplementary Table S6).
228 For nine of the ten bacterial strains constituting the mock community, each subset of
229 V1-V9 sequencing reads classified to the specific genus were assigned with almost
230 complete accuracy (98.2-100%) to the corresponding species (Fig. 2). V3-V4 short-read
231 sequencing showed a discriminatory power comparable to that of V1-V9 full-length
232 sequencing in the classification of four genera (*Bacillus*, *Deinococcus*, *Rhodobacter*,
233 and *Streptococcus*) with more than 98% of reads correctly assigned. However, the V3-
234 V4 region was not suitable for species-level identification of other taxa, such as
235 *Clostridium* and *Staphylococcus*. Only 0.2% of the V3-V4 reads belonging to the genus
236 *Clostridium* were assigned to *Clostridium beijerinckii*, a component of the mock
237 community. In contrast, 92.3% of *Clostridium* reads obtained from V1-V9 full-length
238 sequencing were correctly classified at the species level. These results suggest a lower
239 resolution of the V3-V4 region for species-level classification, emphasizing the
240 advantage of long-read sequencing for obtaining an accurate representation of the
241 sample bacterial composition.
242
243 **Classification of human fecal bacteria**
244 We assessed the performance of our full-length 16S amplicon sequencing approach in
245 the context of a highly complex bacterial community. The V1-V9 region was amplified
246 by four-primer PCR from six human fecal samples (F1-F6) and analyzed by MinION™

247 sequencing. (Table 2). In Fig. 3, the numbers of species detected are plotted against the
248 numbers of reads analyzed. The curve started to plateau at around 20000 reads. There
249 was a highly significant correlation between the read numbers 20000 and 30000
250 (Pearson's correlation coefficient $r > 0.999$, Additional File 4: Supplementary Table S7).
251 Based on these observations, randomly sampled 20000 reads were used in further
252 analysis to determine the bacterial composition of the human gut.
253 For comparison, amplicon sequencing of the V3-V4 region was also conducted using
254 the MinION™ (Table 2) and the Illumina MiSeq™ platform (Table 3). The processed
255 reads from each data set were allocated to the reference bacterial genome using our
256 bioinformatics pipeline to determine the bacterial compositions (Additional File 5 for
257 V1-V9 MinION™ sequencing, Additional File 6 for V3-V4 MinION™ sequencing, and
258 Additional File 7 for V3-V4 MiSeq™ sequencing). From MiSeq™ sequencing data, the
259 bacterial composition was also analyzed by the operational taxonomic unit (OTU)-
260 based approach using the QIIME 2 (ver. 2019.7) pipeline (Additional File 2:
261 Supplementary Fig. S4, Additional File 8) [32, 33]. Although *Bacteroides* was
262 underrepresented in the OTU-based analysis, the two analytical methods (our
263 bioinformatics pipeline and OTU-based method) produced similar taxonomic profiles in
264 the dominant phylotypes for the MiSeq™ data. This result confirmed the validity of our
265 method for the taxonomic classification of the bacterial community.
266 The three sequencing methods (V1-V9 MinION™ sequencing, V3-V4 MinION™
267 sequencing, and V3-V4 MiSeq™ sequencing) revealed similar profiles for the six fecal
268 samples at the genus level (Fig. 4). Statistically significant similarities have been found
269 in the relative genus abundances across these sequencing methods. Thus, at the genus
270 level, V1-V9 full-length MinION™ sequencing exhibited a discriminatory power
271 comparable to that of high-quality short-read sequencing with MiSeq™ technology.
272
273 **The species-level taxonomic resolution achieved by full-length sequencing of the**

16S rRNA gene using MinION™

While genus classification using long versus short reads was relatively comparable, we observed considerable differences across amplified regions in the species-level profiling of human gut microbiota. As shown in Fig. 5, the number of ambiguous reads that were not assigned to species but could be classified at a higher level was significantly greater in the V3-V4 data set in comparison than in the V1-V9 data set.

When species compositions of the dominant taxa (*Bifidobacterium*, *Blautia*, and *Bacteroides*) were analyzed, the V1-V9 and V3-V4 sequencing produced comparable results for *Blautia* (Additional File 2: Supplementary Fig. S5, Additional File 4: Supplementary Table S8) and *Bacteroides* genus (Additional File 2: Supplementary Fig. S6, Additional File 4: Supplementary Table S9) in most of the fecal samples. For *Bifidobacterium*, there appeared to be considerable deviations in the relative abundances of some species depending on the sequencing method used (Fig. 6, Additional File 4: Supplementary Table S10). Notably, most of the *Bifidobacterium* reads generated by V1-V9 MinION™ sequencing were classified into the *Bifidobacterium* species that were isolated from human sources [18, 34]. A significant number of the V3-V4 reads, however, were assigned to *Bifidobacterium* species of non-human origin (Additional File 2: Supplementary Fig. S7). From V3-V4 MiSeq™ sequencing data, the OTU-based classification analysis using the QIIME 2 pipeline also revealed a lower resolution of short-read sequencing for taxonomic separation of *Bifidobacterium* genus. Except for *Bifidobacterium longum*, *Bifidobacterium* species could not be reliably identified by the V3-V4 sequencing strategy (Additional File 2: Supplementary Fig. S8). These results suggest that species classification of *Bifidobacterium* based on V3-V4 sequencing can potentially lead to misidentification and biased community profiling that lacks biological significance.

Discussion

16S rRNA amplicon sequencing is a powerful strategy for taxonomic classification of bacteria and has been extensively employed for analyzing metagenomic samples from environmental and clinical sources [5, 35, 36]. We assessed the performance of MinION™ sequencing by comparing the resolution of the V1-V9 and V3-V4 reads for the taxonomic classification of bacteria. Long-read MinION™ sequencing with the optimized primer set successfully identified *Bifidobacterium* species leading to a better representation of the species composition of the mock community. For improving the classification results, the reads were filtered by length to eliminate those outside the expected size range. Typically, extremely short reads possess only one primer-binding site, suggesting that they are derived from incomplete sequencing. There also exist unexpectedly longer reads with a continuous sequence structure in which two 16S amplicons are linked end-to-end. Because these reads can potentially result in unclassified reads or misclassification, they were eliminated before alignment to the reference sequences of the bacterial genome.

We also modified library construction for MinION™ sequencing with a four-primer PCR strategy, which enabled ligase-free adapter attachment to occur in a single-step reaction. The four-primer PCR generates amplicons with particular chemical modifications at the 5' ends to which adapter molecules can be attached non-enzymatically. Unlike the ligation-based approach, the PCR products amplified by the four-primer method are subjected directly to the adapter attachment reaction without repairing their 5' ends, substantially reducing the time required for sample preparation. Furthermore, because the protocol is free of Nanopore's transposase-based technology (e.g. Rapid Sequencing Kit, SQK-RAD004) that cleaves DNA molecules to produce chemically modified ends for library construction, the PCR products are kept intact, enabling sequencing of the entire amplified region. Thus, the four-primer PCR-based method allowed us to perform amplicon sequencing on the MinION™ platform with

327 user-defined arbitrary primer pairs, taking advantage of the rapid adapter attachment
328 chemistry. This method can be applied to a wide range of sequence-based analyses,
329 including detection of functional genetic markers like antimicrobial resistance genes
330 and identification of genetic variations in targeted loci [11, 37, 38].
331 Our present microbiome study, comparing the discriminatory power of the V1-V9 and
332 V3-V4 reads sequenced on the MinION™ platform, clearly illustrated the advantage of
333 sequencing the entire 16S rRNA gene. The full-length 16S gene sequencing provided
334 better resolution than short-read sequencing for discriminating between members of
335 certain bacterial taxa, including *Clostridium*, *Enterococcus*, *Escherichia*, and
336 *Staphylococcus*. Consistently, comprehensive *in silico* experiments using sequencing a
337 sequencing data set consisting of different regions of the 16S rRNA gene have shown
338 that the choice of the regions to be sequenced substantially affects the classification
339 results [6, 39]. As shown here and in previous publications, short-read sequencing of the
340 16S rRNA gene may be a reasonable option for providing a rough estimation of
341 bacterial diversity. However, it was not suited for analysis requiring species-level
342 resolution and accuracy, which was afforded by sequencing the entire 16S rRNA gene.
343 In metagenomic analysis of the human fecal samples, we used the taxonomic resolution
344 of the V3-V4 region sequenced with MiSeq™, which generates highly accurate reads,
345 as a benchmark for the taxonomic resolution of the full-length 16S gene sequenced with
346 MinION™. The relative abundance of dominant bacterial taxa was highly similar at the
347 genus level between full-length MinION™ and short-read MiSeq™ sequencing.
348 Despite the lower read quality, the full-length sequencing by MinION™ enabled
349 reliable identification of bacterial genera with an accuracy comparable to MiSeq™
350 technology.
351 At the species level, MinION™ long-read sequencing had better resolution for accurate
352 identification of the composition of human gut microbiota. Composition profiles of
353 *Bifidobacterium*, one of the dominant genera present in the human gut [40], appeared to

differ considerably between the two sequencing platforms. While most MinION™ V1-V9 reads were assigned to *Bifidobacterium* species of human origin, a significant number of the MiSeq™ V3-V4 reads were assigned to non-human *Bifidobacterium* species [34]. Such improbable errors in species classification may be attributed to the lower resolution provided by the V3-V4 region, but in fact, metagenomic analysis of mouse gut microbiota has revealed that V3-V4 reads sequenced on MiSeq™ are not well-suited for classifying *Bifidobacterium* species, consistent with our findings [41].

Conclusions

Our modified protocol for 16S amplicon sequencing overcame known limitations, such as the primer-associated bias toward the underrepresentation of *Bifidobacterium*, and enabled taxonomic classification across a broad range of bacterial species. Benchmarking with MiSeq™ sequencing technology demonstrated the analytical advantage of sequencing the full-length 16S gene with MinION™, which provided the requisite species-level resolution and accuracy. With the recent progress in nanopore sequencing chemistry and base-calling algorithms, sequencing accuracy is continuously improving [42, 43]. This will soon enable us to exploit the full potential of MinION™ long-read sequencing technology. High-quality long sequences will allow better discrimination between closely related species, and even bacterial strains, in metagenomic analyses.

375 **Abbreviations**

376 NGS: next-generation sequencing; OTU: operational taxonomic unit; PCR: polymerase
377 chain reaction; rRNA: ribosomal RNA

378

379 **Additional files**

380 **Additional File 1:** Representative bacterial genomes stored in the GenomeSync
381 database.

382 **Additional File 2: Fig. S1.** Sequence heterogeneities of the 27F primer-annealing site
383 in 16S rRNA genes. **Fig. S2.** Evaluation of 16S rRNA PCR primers for identification of
384 bacterial species. **Fig. S3.** Effect of read number on taxonomic classification. **Fig. S4.**
385 Rarefaction curves of observed OTUs in 16S V3-V4 amplicon sequencing of human
386 fecal samples using the MiSeq™ platform. **Fig. S5.** Species composition of *Blautia* in
387 human fecal samples. **Fig. S6.** Species composition of *Bacteroides* in human fecal
388 samples. **Fig. S7.** Deviations in the relative abundances of *Bifidobacterium* species in
389 human fecal samples. **Fig. S8.** Comparison of species composition of fecal
390 *Bifidobacterium* between classification methods.

391 **Additional File 3: Tables S1-S6.** Taxonomic assignment of the mock community
392 analyzed by MinION™ sequencing.

393 **Additional File 4: Table S7.** Correlations between numbers of reads and numbers of
394 detected species in metagenomic sequencing of human fecal samples. **Table S8.**
395 Comparison of species composition of fecal *Blautia* between sequencing methods.
396 **Table S9.** Comparison of species composition of fecal *Bacteroides* between sequencing
397 methods. **Table S10.** Comparison of species composition of fecal *Bifidobacterium*
398 between sequencing methods.

399 **Additional File 5:** Taxonomic profile of human fecal samples from MinION™
400 sequencing (amplicons: V1-V9).

401 **Additional File 6:** Taxonomic profile of human fecal samples from MinION™

sequencing (amplicons: V3-V4).

Additional File 7: Taxonomic profile of human fecal samples from MiSeq™ sequencing (amplicons: V3-V4).

Additional File 8: Taxonomic profiles of human fecal samples from MiSeq™ sequencing (amplicons: V3-V4, taxonomic classification by OTU-based analysis using the QIIME 2 pipeline).

Declarations

Ethics approval and consent to participate

This study was approved by the Sunkaky Institutional Review Board (No. 2017-27).

Consent for publication

All participants provided written informed consent.

Availability of data and materials

The sequence datasets supporting the conclusions of this article are available in the DDBJ DRA database (https://www.ddbj.nig.ac.jp/dra/index-e.html) under accession numbers DRR225043 to DRR225065.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Japan Society for the Promotion of Science KAKENHI Grant Number JP19K09339 to YMa and the branding program as a world-leading research university on intractable immune and allergic diseases supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

429

430 **Authors' contributions**

431 YMa, KK, AF, YMo, YN, SN and KH designed and supervised the study. SK, AF, YMo,
432 and HO contributed to sample collection. YMa, KM, and TT conducted the
433 experiments. YMa, YYasumizu, YYasuoka, and SN analyzed the data. YMa wrote the
434 manuscript. YYasumizu, HB, SN, and KH contributed to editing the manuscript. All
435 authors read and approved the final manuscript.

436

437 **Acknowledgements**

438 We are grateful to Tadashi Imanishi (Tokai University School of Medicine) and Shino
439 Matsukawa (Kyoto University Hospital) for helpful comments and discussion. We
440 would like to thank Editage (www.editage.com) for English language editing.

441

442 **References**

443 1. Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet. 2019;20(6):341-
444 55.
445 2. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al.
446 Performance comparison of benchtop high-throughput sequencing platforms.
447 Nat Biotechnol. 2012;30(5):434-9.
448 3. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical
449 microbiology with bacterial genome sequencing. Nat Rev Genet.
450 2012;13(9):601-12.
451 4. Clarridge JE, 3rd. Impact of 16S rRNA gene sequence analysis for identification
452 of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev.
453 2004;17(4):840-62, table of contents.
454 5. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et
455 al. Predictive functional profiling of microbial communities using 16S rRNA
456 marker gene sequences. Nat Biotechnol. 2013;31(9):814-21.
457 6. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et
458 al. Evaluation of 16S rRNA gene sequencing for species and strain-level
459 microbiome analysis. Nat Commun. 2019;10(1):5029.
460 7. Ravi RK, Walton K, Khosroheidari M. MiSeq: A Next Generation Sequencing
461 Platform for Genomic Analysis. Methods Mol Biol. 2018;1706:223-32.
462 8. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, et
463 al. Experimental and analytical tools for studying the human microbiome. Nat
464 Rev Genet. 2011;13(1):47-58.
465 9. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. J
466 Exp Bot. 2017;68(20):5419-29.
467 10. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, et al. Rapid draft
468 sequencing and real-time nanopore sequencing in a hospital outbreak of

469 Salmonella. Genome Biol. 2015;16:114.
470 11. Leggett RM, Alcon-Giner C, Heavens D, Caim S, Brook TC, Kujawska M, et al.
471 Rapid MinION profiling of preterm microbiota and antimicrobial-resistant
472 pathogens. Nat Microbiol. 2020;5(3):430-42.
473 12. Benitez-Paez A, Sanz Y. Multi-locus and long amplicon sequencing approach to
474 study microbial diversity at species level using the MinION portable nanopore
475 sequencer. Gigascience. 2017;6(7):1-12.
476 13. Shin H, Lee E, Shin J, Ko SR, Oh HS, Ahn CY, et al. Elucidation of the bacterial
477 communities associated with the harmful microalgae Alexandrium tamarense
478 and Cochlodinium polykrikoides using nanopore sequencing. Sci Rep.
479 2018;8(1):5323.
480 14. Mitsuhashi S, Kryukov K, Nakagawa S, Takeuchi JS, Shiraishi Y, Asano K, et
481 al. A portable system for rapid bacterial composition analysis using a nanopore-
482 based sequencer and laptop computer. Sci Rep. 2017;7(1):5657.
483 15. Nakagawa S, Inoue S, Kryukov K, Yamagishi J, Ohno A, Hayashida K, et al.
484 Rapid sequencing-based diagnosis of infectious bacterial species from
485 meningitis patients in Zambia. Clin Transl Immunology. 2019;8(11):e01087.
486 16. Kai S, Matsuo Y, Nakagawa S, Kryukov K, Matsukawa S, Tanaka H, et al.
487 Rapid bacterial identification by direct PCR amplification of 16S rRNA genes
488 using the MinION nanopore sequencer. FEBS Open Bio. 2019;9(3):548-57.
489 17. Kim SW, Suda W, Kim S, Oshima K, Fukuda S, Ohno H, et al. Robustness of
490 gut microbiota of healthy adults in response to probiotic intervention revealed by
491 high-throughput pyrosequencing. DNA Res. 2013;20(3):241-53.
492 18. Arboleya S, Watkins C, Stanton C, Ross RP. Gut Bifidobacteria Populations in
493 Human Health and Aging. Front Microbiol. 2016;7:1204.
494 19. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. Host-bacterial
495 mutualism in the human intestine. Science. 2005;307(5717):1915-20.

496 20. Tanaka H, Matsuo Y, Nakagawa S, Nishi K, Okamoto A, Kai S, et al. Real-time
497 diagnostic analysis of MinION-based metagenomic sequencing in clinical
498 microbiology evaluation: a case report. JA Clin Rep. 2019;5(1):24.
499 21. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al.
500 Enterotypes of the human gut microbiome. Nature. 2011;473(7346):174-80.
501 22. Takagi T, Naito Y, Inoue R, Kashiwagi S, Uchiyama K, Mizushima K, et al.
502 Differences in gut microbiota associated with age, sex, and stool consistency in
503 healthy Japanese subjects. J Gastroenterol. 2019;54(1):53-63.
504 23. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al.
505 Evaluation of general 16S ribosomal RNA gene PCR primers for classical and
506 next-generation sequencing-based diversity studies. Nucleic Acids Res.
507 2013;41(1):e1.
508 24. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for
509 FASTA/Q File Manipulation. PLoS One. 2016;11(10):e0163962.
510 25. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA
511 ribosomal RNA gene database project: improved data processing and web-based
512 tools. Nucleic Acids Res. 2013;41(Database issue):D590-6.
513 26. Silva reference files. https://mothur.org/wiki/silva_reference_files/
514 27. Genome Search Toolkit. http://kirill-kryukov.com/study/tools/gstk/
515 28. Frith MC. A new repeat-masking method enables specific detection of
516 homologous sequences. Nucleic Acids Res. 2011;39(4):e23.
517 29. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
518 2018;34(18):3094-100.
519 30. GenomeSync. http://genomesync.org
520 31. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res.
521 2012;40(Database issue):D136-43.
522 32. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello

523 EK, et al. QIIME allows analysis of high-throughput community sequencing
524 data. Nat Methods. 2010;7(5):335-6.
525 33. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et
526 al. Reproducible, interactive, scalable and extensible microbiome data science
527 using QIIME 2. Nat Biotechnol. 2019;37(8):852-7.
528 34. Milani C, Lugli GA, Duranti S, Turroni F, Bottacini F, Mangifesta M, et al.
529 Genomic encyclopedia of type strains of the genus Bifidobacterium. Appl
530 Environ Microbiol. 2014;80(20):6290-302.
531 35. Human Microbiome Project C. A framework for human microbiome research.
532 Nature. 2012;486(7402):215-21.
533 36. Srinivasan R, Karaoz U, Volegova M, MacKichan J, Kato-Maeda M, Miller S, et
534 al. Use of 16S rRNA gene for identification of a broad range of clinically
535 relevant bacterial pathogens. PLoS One. 2015;10(2):e0117617.
536 37. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al.
537 Multiplex PCR method for MinION and Illumina sequencing of Zika and other
538 virus genomes directly from clinical samples. Nat Protoc. 2017;12(6):1261-76.
539 38. Cornelis S, Gansemans Y, Deleye L, Deforce D, Van Nieuwerburgh F. Forensic
540 SNP Genotyping using Nanopore MinION Sequencing. Sci Rep. 2017;7:41759.
541 39. Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya
542 TI. The effect of 16S rRNA region choice on bacterial community
543 metabarcoding results. Sci Data. 2019;6:190007.
544 40. Ventura M, Turroni F, Motherway MO, MacSharry J, van Sinderen D. Host-
545 microbe interactions that facilitate gut colonization by commensal
546 bifidobacteria. Trends Microbiol. 2012;20(10):467-76.
547 41. Shin J, Lee S, Go MJ, Lee SY, Kim SC, Lee CH, et al. Analysis of the mouse gut
548 microbiome using full-length 16S rRNA amplicon sequencing. Sci Rep.
549 2016;6:29681.

42. Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis: state of the art, applications and challenges. Brief Bioinform. 2018;19(6):1256-72.

43. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19(1):90.

558 **Figure legends**

559 **Fig. 1** Metagenomic analysis of the 16S rRNA gene amplicons using MinION™
560 nanopore sequencing. **a** Workflow of 16S rRNA amplicon sequencing on the MinION™
561 platform. Sequencing libraries are generated by the four-primer PCR-based strategy,
562 enabling simplified post-PCR adapter attachment. At the initial stage of PCR, the 16S
563 rRNA gene is amplified with the inner primer pairs. The resulting PCR products are
564 targeted for amplification with the outer primers to introduce the barcode and tag
565 sequences at both ends, to which adapter molecules can be attached in a single-step
566 reaction. **b, c** Taxonomic assignments of a mock community analyzed by MinION™
567 sequencing. The V1-V9 or V3-V4 region of the 16S rRNA gene was amplified from a
568 pre-characterized mock community sample comprising ten bacterial species and
569 sequenced on the MinION™ platform. Three thousand reads were randomly selected
570 from the processed data set and aligned directly to the reference genome database of
571 5850 representative bacterial species. The pie charts represent taxonomic profiles at the
572 (b) genus and (c) species levels. Slices corresponding to misclassified (assigned to
573 bacteria not present in the mock community) or unclassified (not classified at the
574 species level but placed in a higher taxonomic rank) reads are exploded. The relative
575 abundance (%) of each taxon is shown.

576

577 **Fig. 2** Accurate taxonomic assignment afforded by full-length MinION™ sequencing of
578 the 16S rRNA gene. Classification accuracy compared between full-length (V1-V9) and
579 partial (V3-V4) 16S sequencing data obtained from composition profiling of the ten-
580 species mock community. The donut charts show the proportions of reads correctly
581 assigned to the species constituting the mock community. The percentage of correctly
582 classified reads is shown in the center hole. NA: not assigned (no reads were classified
583 in *Escherichia* genus).

584

585 **Fig. 3** Metagenomic analysis of human gut microbiota. Six human fecal samples (F1-
586 F6) were subjected to full-length 16S rRNA amplicon sequencing via MinION™.
587 Numbers of detected species are plotted against numbers of reads used for taxonomic
588 classification.
589
590 **Fig. 4** Comparison of taxonomic profiles of human gut microbiota between sequencing
591 methodologies. Six fecal samples (F1-F6) were analyzed by sequencing the entire 16S
592 rRNA gene using MinION™ (N_V1-V9). For comparison, the V3-V4 region was
593 sequenced on MinION™ (N_V3-V4) or MiSeq™ platforms (I_V1-V9). Randomly
594 sampled 20000 reads from each data set were allocated to the reference genome
595 database of 5850 representative bacterial species. A heat map shows the relative genus
596 abundance (%) of classified reads. The 15 most abundant taxa are shown. The Pearson
597 correlation coefficient (*r*) between sequencing methods was computed. Asterisks
598 indicate significant correlations at $P < 0.05$.
599
600 **Fig. 5** Comparison of taxonomic resolution. The percentages of ambiguous reads not
601 assigned to the species level are plotted for six fecal samples analyzed by MinION™
602 (N_V1-V9 and N_V3-V4) or MiSeq™ (I_V3-V4). Horizontal bars represent mean
603 values. * $P < 0.05$ (statistically significant).
604
605 **Fig. 6** Species composition of *Bifidobacterium* in six fecal samples. MinION™ V1-V9
606 sequencing confers species-level resolution for bacterial composition profiling. Results
607 obtained by the three sequencing methods are shown. The legends show the 14 most
608 abundant *Bifidobacterium* species.

Table 1 MinION™ sequencing statistics for the mock community sample

Primers	Pass reads				Trimmed reads		Filtered reads	
	No. of reads	Min (bp)	Avg (bp)	Max (bp)	No. of reads	Avg (bp)	No. of reads	Avg (bp)
V1-V9	8651	237	1497	3292	8455	1367.1	6972 (80.6%)	1473
V3-V4	101372	180	585.7	1977	99937	451.8	96189 (94.9%)	454.9

Min: minimum read length, Avg: average read length, Max: maximum read length

Table 2 Statistics of MinION™ sequencing data for human fecal samples

Sample	Pass reads				Trimmed reads		Filtered reads	
	No. of reads	Min (bp)	Avg (bp)	Max (bp)	No. of reads	Avg (bp)	No. of reads	Avg (bp)
F1/N_V1-V9	104895	186	1521.1	4549	103100	1386.4	89752 (85.6%)	1463.7
F2/N_V1-V9	84065	169	1393.8	4253	82458	1259.8	60326 (71.8%)	1461.4
F3/N_V1-V9	76968	168	1474.3	4829	74479	1343.3	60,713 (78.9%)	1465.5
F4/N_V1-V9	114060	168	1541.7	4836	111436	1410.6	100569 (88.2%)	1469.9
F5/N_V1-V9	85912	177	1536.0	4877	83038	1405.4	74168 (86.3%)	1474.2
F6/N_V1-V9	108938	213	1525.1	4866	106857	1393.5	93146 (85.5%)	1467.4
F1/N_V3-V4	52864	160	568.8	2759	52283	435.2	48494 (91.7%)	442.5
F2/N_V3-V4	92816	174	583.4	2886	91989	442.8	89016 (95.9%)	444.7
F3/N_V3-V4	60200	163	568.5	2062	59435	434.6	55706 (92.5%)	441.1
F4/N_V3-V4	83021	202	578.0	2050	81734	446.1	77995 (93.9%)	450.0
F5/N_V3-V4	78409	167	578.4	1796	76135	447.8	72526 (92.5%)	453.1
F6/N_V3-V4	74931	114	580.3	2246	73946	446.1	71330 (95.2%)	449.1

N: Oxford Nanopore MinION™, Min: minimum read length, Avg: average read length, Max: maximum read length

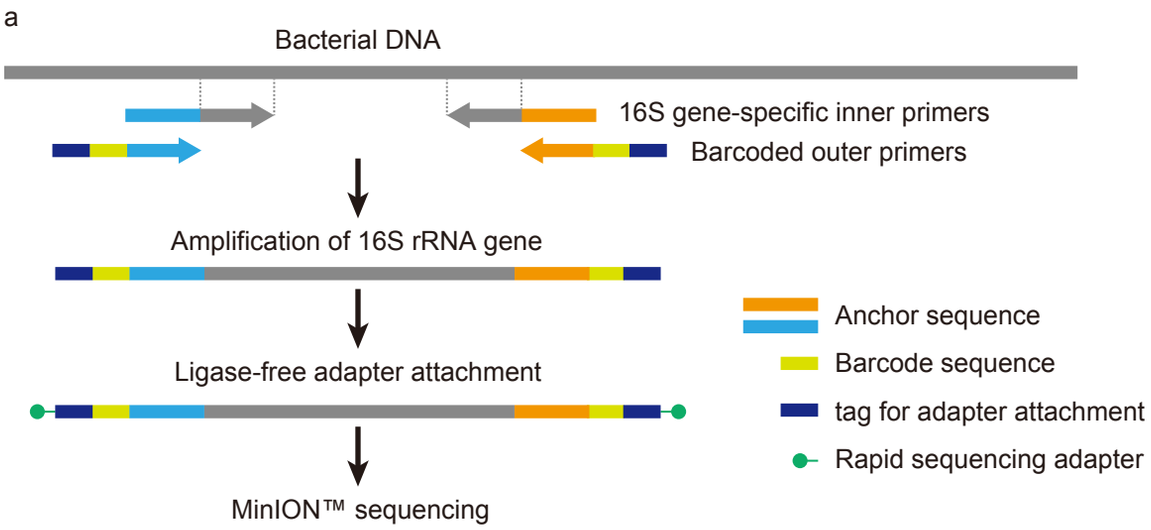
Table 3 Statistics of MiSeq™ sequencing data for human fecal samples

Sample	Paired reads	Merged reads		Filtered reads	
	No. of reads	No. of reads	Avg (bp)	No. of reads	Avg (bp)
F1/I_V3-V4	66242	63821	449.3	63778 (96.3%)	449.5
F2/I_V3-V4	68824	66640	447.6	66490 (96.6%)	448.3
F3/I_V3-V4	132057	128095	446.9	127999 (96.9%)	447.1
F4/I_V3-V4	103532	100945	451.4	100853 (97.4%)	451.7
F5/I_V3-V4	72136	70521	451	70459 (97.7%)	451.3
F6/I_V3-V4	52182	50907	449	50841 (97.4%)	449.5

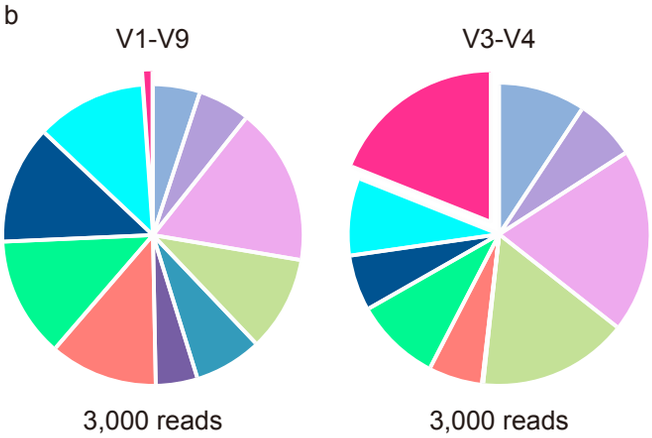
I: Illumina MiSeq™, Avg: average read length

Fig. 1

a

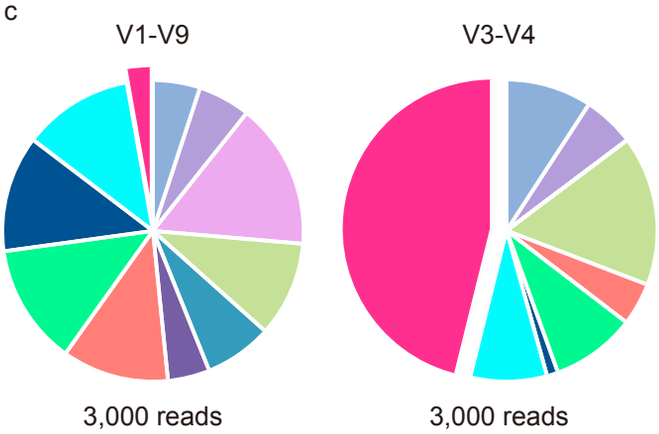


b



Genus	V1-V9	V3-V4
Bacillus	5.07%	9.30%
Bifidobacterium	5.63%	6.60%
Clostridium	16.97%	19.70%
Deinococcus	10.23%	16.10%
Enterococcus	7.30%	0.13%
Escherichia	4.50%	0.00%
Lactobacillus	11.63%	5.77%
Rhodobacter	12.97%	9.17%
Staphylococcus	12.73%	5.97%
Streptococcus	11.87%	8.30%
mis- or unclassified	1.10%	18.97%

c



Species	V1-V9	V3-V4
B. cereus	5.07%	9.20%
B. adolescentis	5.63%	5.53%
C. beijerinckii	15.67%	0.03%
D. radiodurans	10.23%	16.07%
E. faecalis	7.30%	0.00%
E. coli	4.50%	0.00%
L. gasseri	11.47%	4.57%
R. sphaeroides	12.97%	9.07%
S. epidermidis	12.50%	1.20%
S. mutans	11.87%	8.20%
mis- or unclassified	2.80%	46.13%

Fig. 2

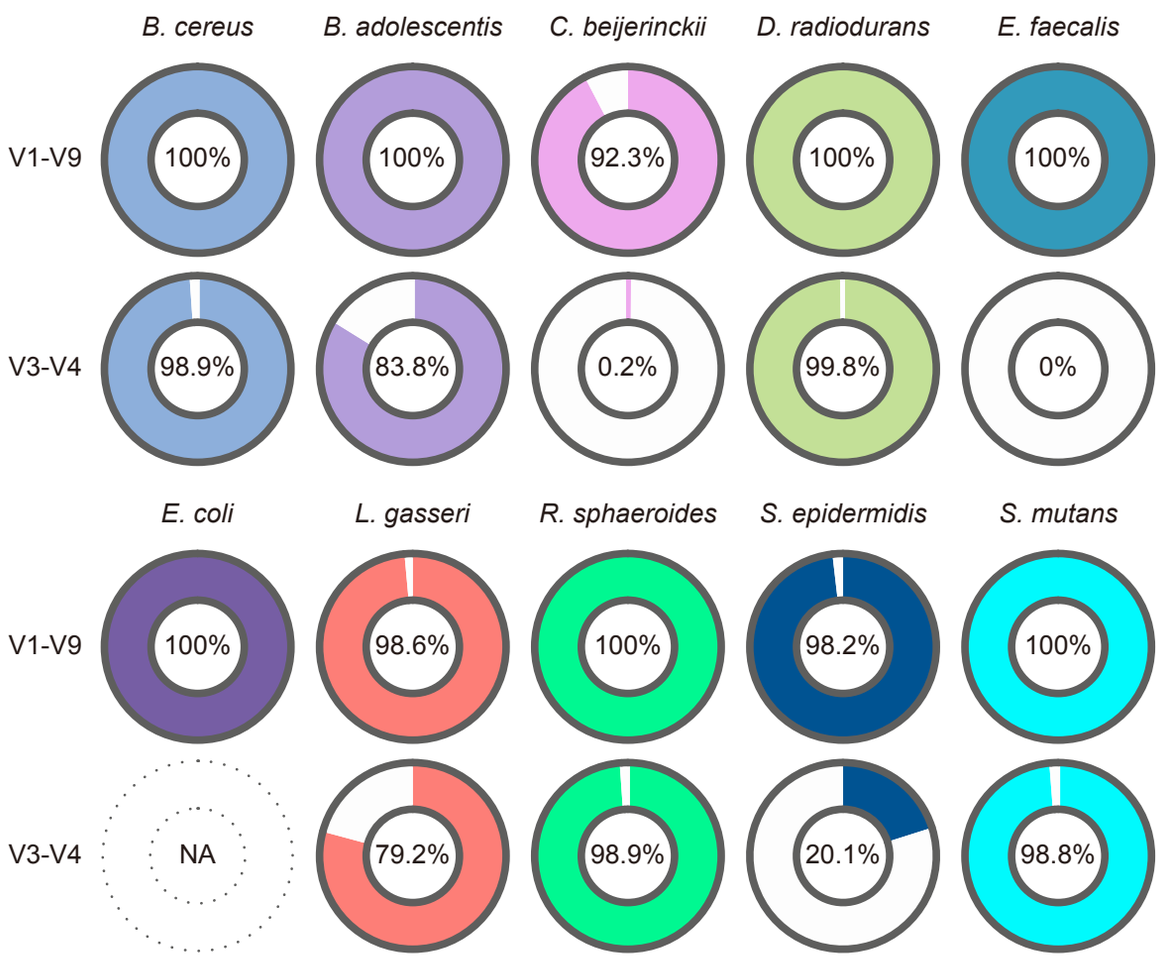

B. cereus
B. adolescentis
C. beijerinckii
D. radiodurans
E. faecalis
V1-V9
100%
100%
92.3%
100%
100%
V3-V4
98.9%
83.8%
0.2%
99.8%
0%
E. coli
L. gasseri
R. sphaeroides
S. epidermidis
S. mutans
V1-V9
100%
98.6%
100%
98.2%
100%
V3-V4
NA
79.2%
98.9%
20.1%
98.8%

Fig. 3

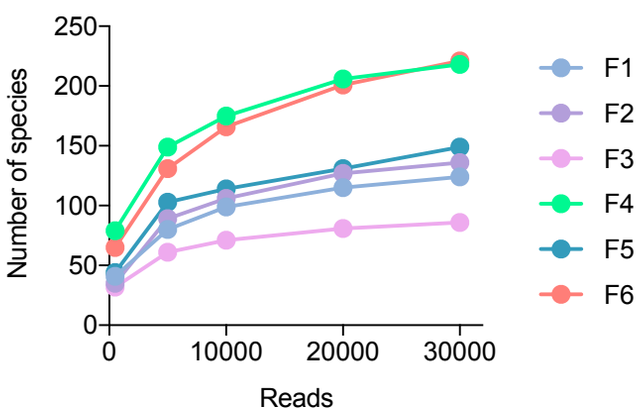

250
200
150
100
50
0
Number of species
0
10000
20000
30000
Reads
F1
F2
F3
F4
F5
F6

Fig. 4

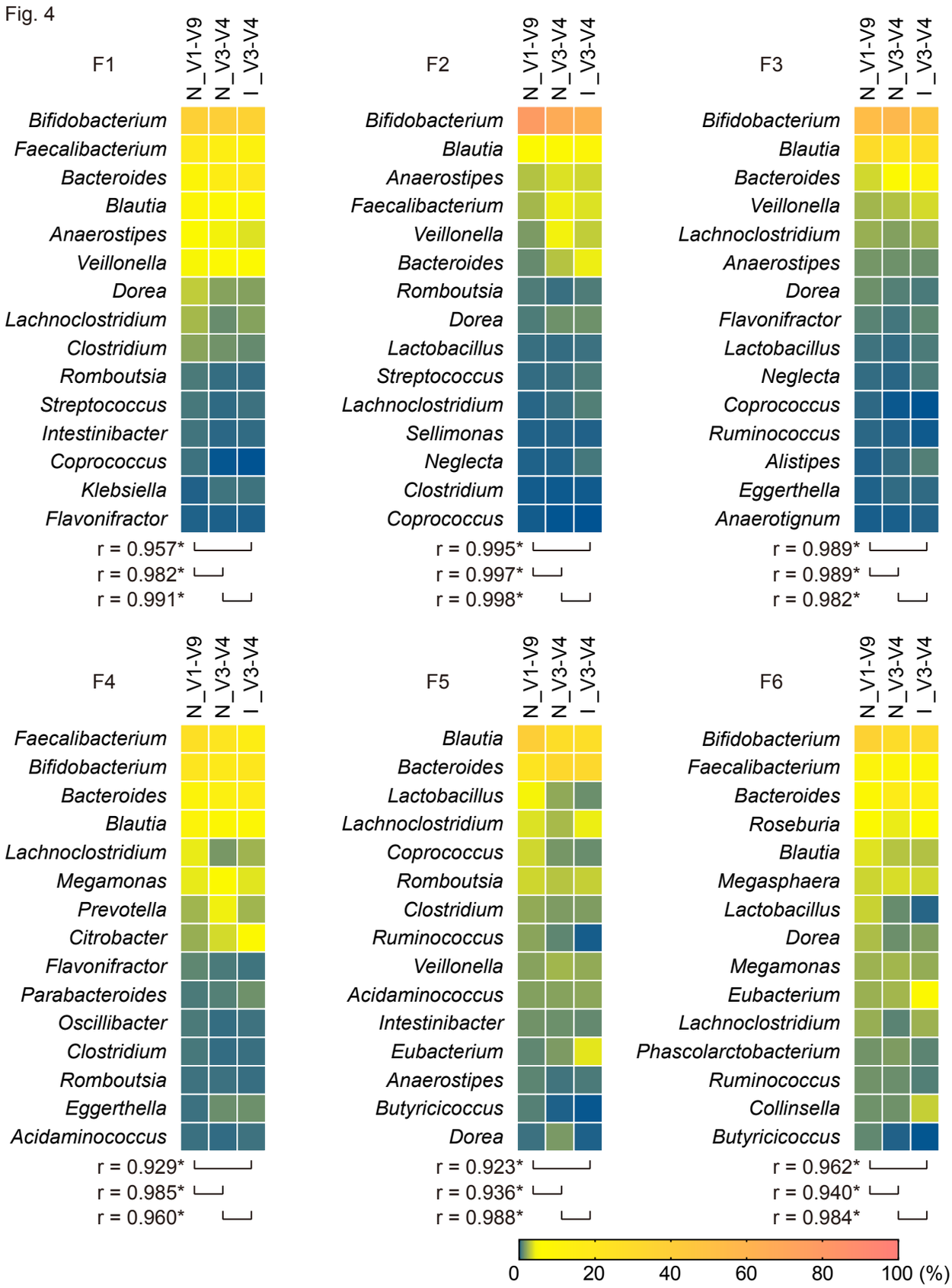

F1
N_V1-V9
N_V3-V4
I_V3-V4
Bifidobacterium
Faecalibacterium
Bacteroides
Blautia
Anaerostipes
Veillonella
Dorea
Lachnoclostridium
Clostridium
Romboutsia
Streptococcus
Intestinibacter
Coprococcus
Klebsiella
Flavonifractor
r = 0.957*
r = 0.982*
r = 0.991*
F2
N_V1-V9
N_V3-V4
I_V3-V4
Bifidobacterium
Blautia
Anaerostipes
Faecalibacterium
Veillonella
Bacteroides
Romboutsia
Dorea
Lactobacillus
Streptococcus
Lachnoclostridium
Sellimonas
Neglecta
Clostridium
Coprococcus
r = 0.995*
r = 0.997*
r = 0.998*
F3
N_V1-V9
N_V3-V4
I_V3-V4
Bifidobacterium
Blautia
Bacteroides
Veillonella
Lachnoclostridium
Anaerostipes
Dorea
Flavonifractor
Lactobacillus
Neglecta
Coprococcus
Ruminococcus
Alistipes
Eggerthella
Anaerotignum
r = 0.989*
r = 0.989*
r = 0.982*
F4
N_V1-V9
N_V3-V4
I_V3-V4
Faecalibacterium
Bifidobacterium
Bacteroides
Blautia
Lachnoclostridium
Megamonas
Prevotella
Citrobacter
Flavonifractor
Parabacteroides
Oscillibacter
Clostridium
Romboutsia
Eggerthella
Acidaminococcus
r = 0.929*
r = 0.985*
r = 0.960*
F5
N_V1-V9
N_V3-V4
I_V3-V4
Blautia
Bacteroides
Lactobacillus
Lachnoclostridium
Coprococcus
Romboutsia
Clostridium
Ruminococcus
Veillonella
Acidaminococcus
Intestinibacter
Eubacterium
Anaerostipes
Butyricicoccus
Dorea
r = 0.923*
r = 0.936*
r = 0.988*
F6
N_V1-V9
N_V3-V4
I_V3-V4
Bifidobacterium
Faecalibacterium
Bacteroides
Roseburia
Blautia
Megasphaera
Lactobacillus
Dorea
Megamonas
Eubacterium
Lachnoclostridium
Phascolarctobacterium
Ruminococcus
Collinsella
Butyricicoccus
r = 0.962*
r = 0.940*
r = 0.984*
0
20
40
60
80
100 (%)

Fig. 5

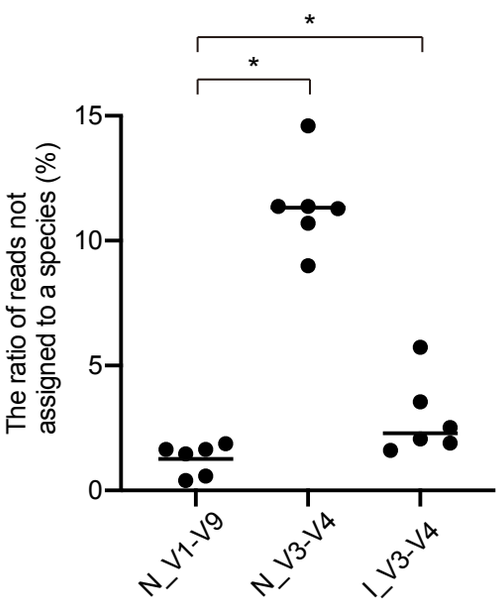
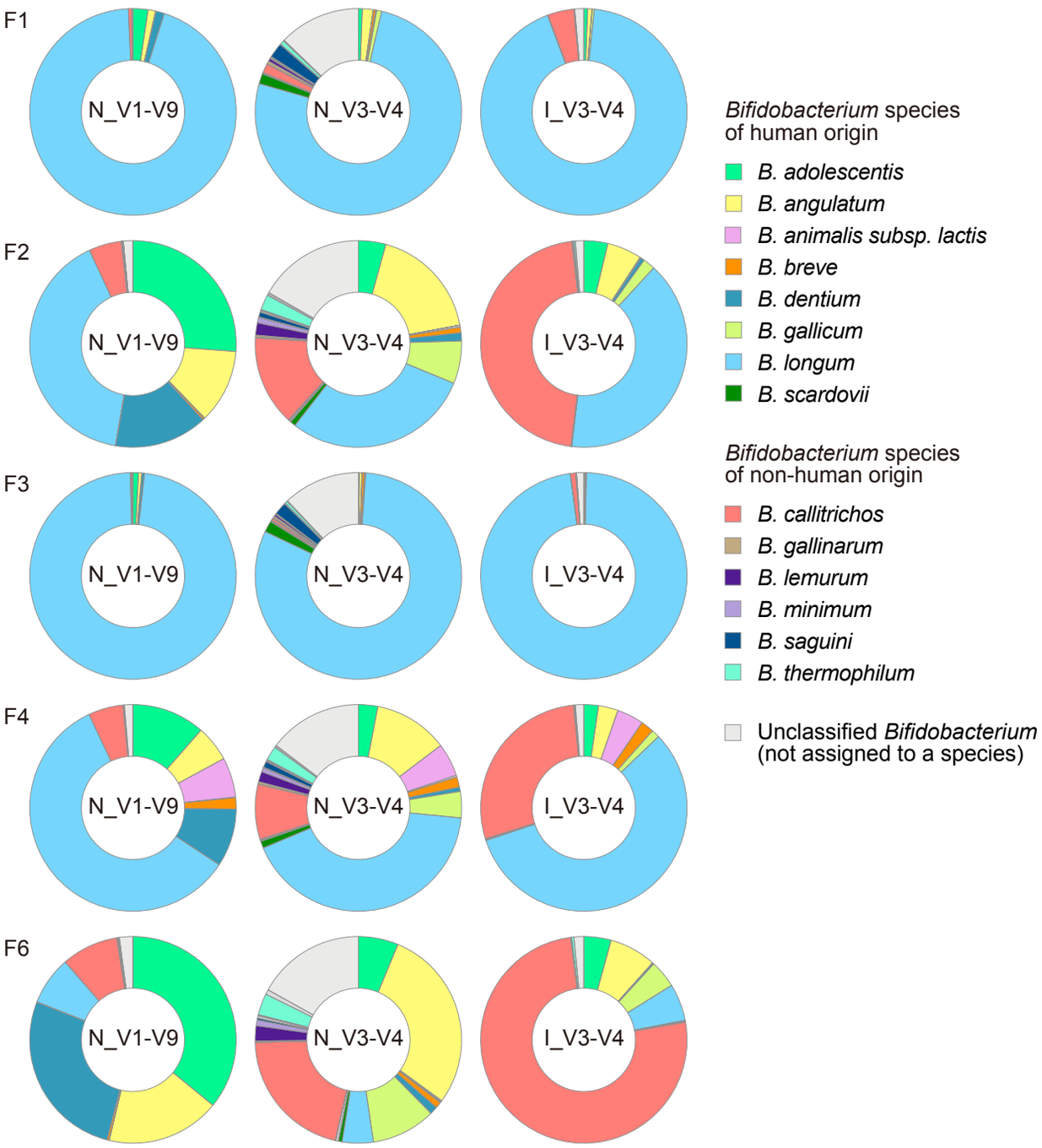

*
*
15
10
5
0
The ratio of reads not
assigned to a species (%)
N_V1-V9
N_V3-V4
I_V3-V4

Fig. 6

a


F1
F2
F3
F4
F6
N_V1-V9
N_V3-V4
I_V3-V4
Bifidobacterium species of human origin
B. adolescentis
B. angulatum
B. animalis subsp. lactis
B. breve
B. dentium
B. gallicum
B. longum
B. scardovii
Bifidobacterium species of non-human origin
B. callitrichos
B. gallinarum
B. lemurum
B. minimum
B. saguini
B. thermophilum
Unclassified Bifidobacterium (not assigned to a species)