# Comparative evaluation of analytical pipelines for illumina short- and nanopore long-read 16S rRNA gene amplicon sequencing with mock microbial communities

# Comparative evaluation of analytical pipelines for Illumina short- and Nanopore long-read 16S rRNA gene amplicon sequencing with mock microbial communities

Yusuke Ota[a], Kei Yasunaga[a], Samiratu Mahazu[a,b], Isaac Prah[a], Satoshi Nagai[c], Takaya Hayashi[d], Masato Suzuki[e], Mitsunori Yoshida[f], Yoshihiko Hoshino[f], Yukihiro Akeda[g], Toshihiko Suzuki[h], Yoshiaki Gu[i], Ryoichi Saito[a*]

[a] Department of Molecular Microbiology and Immunology, Tokyo Medical and Dental University, Tokyo, Japan

[b] Department of Parasitology and Tropical Medicine, Tokyo Medical and Dental University, Tokyo, Japan

[c] Fisheries Technology Institute, Japan Fisheries Research and Education Agency, Kanagawa, Japan

[d] Department of Molecular Virology, Tokyo Medical and Dental University, Tokyo, Japan

[e] Antimicrobial Resistance Research Center, National Institute of Infectious Diseases, Tokyo, Japan

[f] Department of Mycobacteriology, Leprosy Research Center, National Institute of Infectious Diseases, Tokyo, Japan

[g] Department of Bacteriology I, National Institute of Infectious Diseases, Tokyo, Japan

[h] Department of Bacterial Pathogenesis, Tokyo Medical and Dental University, Tokyo, Japan

[i] Department of Infectious Diseases, Tokyo Medical and Dental University, Tokyo, Japan

[*]**Address correspondence to**:

Ryoichi Saito

Department of Molecular Microbiology and Immunology, Tokyo Medical and Dental University,
1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan
Tel/Fax: +81(03)5803-5368
E-mail: r-saito.mi@tmd.ac.jp

**Running Title:** Comparison of 16S amplicon sequencing pipelines

**Abstract**

Utility of a recently developed long-read pipeline, Emu, was assessed using an expectation-maximization algorithm for accurate read classification. We compared it to conventional short- and long-read pipelines, using well-characterized mock bacterial samples. Our findings highlight the necessity of appropriate data-processing for taxonomic descriptions, expanding our understanding of the precise microbiome.

**Keywords:** Emu; full-length 16S rRNA metabarcoding analysis; microbiome

**Text**

The 16S rRNA gene-based metabarcoding strategy, commonly used for understanding bacterial taxonomy, offers species-level classification necessary for accurate microbiome interpretation (Castellarin et al., 2012; Pantha et al., 2021; Scher et al., 2013). Illumina short-read sequencing platforms and QIIME2 (bioinformatics package) are widely used for processing and analyzing microbiome data (Bolyen et al., 2019); however, short reads (≤600 bp) provide limited species-level information (Sadowsky et al., 2017; Winand et al., 2019). Nanopore long-read sequencers address this drawback with the cloud-based analysis platform EPI2ME applicable for long-read data analysis (Ciuffreda et al., 2021). However, EPI2ME shows high sequencing errors, 39.50% misclassified and 25.46% unclassified reads at the species level, when using microbial-community DNA standard (Winand et al., 2019). In contrast, Emu, a novel Nanopore long-read 16S rRNA metabarcoding pipeline developed for species-level microbial community profiling using the expectation-maximization algorithm, enables correct generation of taxonomic outlines with reduced sequencing errors (Curry et al., 2022). Nevertheless, the usefulness of this workflow compared with

that of other bioinformatic approaches using well-characterized mock samples has not been extensively examined.

Here, we aimed to evaluate the utility of Emu in comparison with taxonomic results of current short- and long-read analysis pipelines using ZymoBIOMICS Microbial Community DNA Standard (Zymo Research Corp., Irvine, CA, USA) with known 16S rRNA gene compositions. DNA mixture of each plasmid-cloned single copy of the 16S rRNA genes derived from *Eggerthella lenta* ATCC 43055, *Staphylococcus aureus* ATCC 29213, *Limosilactobacillus fermentum* ATCC 9338, *Bacteroides fragilis* ATCC 25285, *Clostridioides difficile* R20291, *Pseudomonas aeruginosa* ATCC 27853, *Escherichia coli* ATCC 25922, and a *Campylobacter jejuni* clinical strain was prepared to reduce the effects of PCR bias (Nagai et al., 2022).

The V3-V4 region of the 16S rRNA gene was sequenced using the Illumina MiSeq platform, per the manufacturer's instructions (Illumina, 2015). Taxonomic assignment was performed using amplicon sequence variants with the QIIME2 Naive Bayes classifier pre-trained on the SILVA reference database (release 138) (Bokulich et al., 2018; Quast et al., 2013). Full-length 16S rRNA gene sequencing was performed using the 16S Barcoding Kit containing primer set, MinION sequencer, R9.4.1 flow cell, and MinKNOW v21.11.7 (Oxford Nanopore Technologies). The number of sequences from each sample was adjusted using SeqKit v2.2.0 (Shen et al., 2016). Generated long-read data were processed using Emu v3.4.4 (Curry et al., 2022) and EPI2ME v.3.5.7 (Ciuffreda et al., 2021). We used FastQC (Andrew, 2010) to confirm the expected sequence length distribution of short- and long-read data (Supplementary Fig. 1).

In the commercial sample analysis, all bacterial taxa at the genus level were identified using the three pipelines, barring *Salmonella* when using QIIME2 (Fig. 1A and 1C). At the species level, Emu identified all taxa in both commercial and in-house samples, whereas EPI2ME failed to detect

74 *E. coli* in the commercial sample and QIIME2 failed to classify species other than *L. fermentum* and
75 *B. fragilis* (Fig. 1B and 1D). Thus, QIIME2 use is considered challenging for species-level
76 discrimination owing to incomplete coverage of the 16S rRNA gene (Sadowsky et al., 2017).
77 EPI2ME identified *Listeria monocytogenes* in commercial and *E. coli* and *C. difficile* in in-house
78 samples at low abundance (<1%). Furthermore, 16.8% of EPI2ME reads were misidentified as those
79 of *Listeria welshimeri* in the commercial sample, consistent with previous study findings (Nanopore,
80 2016); these reads could have been derived from *L. monocytogenes.* These *Listeria* spp. show 98.8%
81 similarity in their 16S rRNA sequences (Collins et al., 1991), demanding a stricter classification
82 approach for bacteria with highly homologous 16S rRNA sequences. EPI2ME does not offer
83 removal or correction of erroneous sequences leading to increased misclassified reads (Winand et
84 al., 2019). Conventional analytical pipelines lack discriminability resulting in limited taxa
85 identification and increased misclassified and unclassified reads. Contrastingly, Emu employs a
86 homology-aware alignment likelihood algorithm capable of highly accurate taxonomic classification
87 based on read alignments to multiple reference sequences (Curry et al., 2022). This approach enables
88 better classification by contributing to reduced false positives and improved discrimination between
89 genetically similar bacterial species (Curry et al., 2022). Indeed, the F-scores (Almeida et al., 2018)
90 calculated from the precision and recall were best in the Emu workflow compared with those of
91 QIIME2 and EPI2ME at the genus and species levels (Fig. 2).
92 PCR primer selection for gene amplification contributes to varying results during 16S rRNA
93 metabarcoding analyses (Park et al., 2021). Additionally, the primers used here (Oxford Nanopore
94 Technologies) mismatched with particular bacterial 16S rRNA genes (Nanopore, 2016; Winand et
95 al., 2019). Thus, PCR efficiency in the library preparation step before using each pipeline may affect
96 correlation results, including abundance rank evaluation of the bacterial components. The existence

of multiple heterogeneous 16S rRNA copies within a genome can lead to experimental bias (Ibal et al., 2019). Therefore, we prepared samples with the cloned 16S rRNA of each bacterium to reduce PCR amplification bias and determine variations in taxonomic results among the three pipelines tested as another measure for comparative evaluation of closeness to the true value (Nagai et al., 2022). Unification of the number of 16S rRNA gene copies enables evaluation of variations in the theoretical value. The coefficients of variation, the ratio of the standard deviation to the mean, for the existence ratio using QIIME2, Emu, and EPI2ME were 44.6%, 37.1%, and 71.6%, respectively, at the genus level, and 178.5%, 37.1%, and 79.5%, respectively, at the species level. Emu proved superior in terms of reflecting relative bacterial abundance. These observations also suggest that a mock sample with the same copy number of 16S rRNA gene, alongside DNA concentration, might provide precise quality control of the metabarcoding analysis workflow without bias stemming from multiple gene copies.

A study limitation was the use of only mock samples with pure bacterial DNA. Clinical and environmental samples typically contain an assortment of bacterial species (Castellarin et al., 2012; Pantha et al., 2021; Scher et al., 2013). Some primers used for metabarcoding analysis reportedly amplify off-target sequences derived from human DNA (Walker et al., 2020). In future microbial diversity studies, performance evaluation of pipelines and effects of artifacts should include clinical and environmental samples.

In conclusion, the Nanopore long-read pipeline Emu enabled accurate species-level allocation and abundance representation during 16S rRNA metabarcoding with the lowest variation in mock microbial communities compared to short-read-based QIIME2 and long-read-based EPI2ME workflows. For quality management of metabarcoding analytical workflows, this study suggests the use of plasmid DNA mock sample with equal 16S rRNA gene copy numbers. Our findings emphasize

the importance of appropriate data processing and evaluation for taxonomic investigations in representing actual microbiome profiles.

**Acknowledgments**

We would like to thank Editage (www.editage.jp) for English language editing.

**Funding**

This work was partially supported by the Japan Agency for Medical Research and Development (AMED) [grant numbers JP20wm0125007 (TH, TS, RS); JP20wm0225004 (MS, YH, TS, RS); JP20wm0225013 (YA, TS, RS); JP23wm0225022 (MS, MY, YH, TS, RS); JP23gm1610003, JP23fk0108642, JP23fk0108665, JP23fk0108683, and JP23wm0325037 (MS); JP23fk0108608; JP23wm0125007; JP23wm0325054; JP23gm1610003; JP23gm1610007; and JP23fk0108673 (YH)]; the Japan Society for the Promotion of Science (JSPS) KAKENHI grant [grant numbers JP20K08818 (RS) and JP23H03551 (YO, RS)]; and the Morinomiyako Medical Research Foundation (YO). The funders had no role in the study design, data collection and interpretation, or the decision to submit the work for publication.

**Author contributions**

RS: Conceptualization, Methodology. YO: Investigation, Formal analysis. KY, SM, IP, SN, and YG: Formal analysis. YO, TH, MS, MY, YH, YA, TS, and RS: Funding acquisition. YO and RS: Writing – original draft, Writing – review & editing. All authors revised the drafts of the manuscript and approved the final version.

## Data availability statement

The 16S rRNA amplicon sequencing data included in this study have been deposited in the NCBI's Sequence Read Archive (SRA) under accession numbers SRR23636350, SRR23636351, SRR23636352, and SRR23636353.

## Conflicts of interest

None to declare.

## References


Almeida, A., Mitchell, A.L., Tarkowska, A., Finn, R.D., 2018. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. GigaScience 7, giy054. https://doi.org/10.1093/gigascience/giy054.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed 26 March 2024).

Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G.A., Gregory Caporaso, J., 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6, 90. https://doi.org/10.1186/s40168-018-0470-z.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodriguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson,

D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., 2nd, Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vazquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9.

Castellarin, M., Warren, R.L., Freeman, J.D., Dreolini, L., Krzywinski, M., Strauss, J., Barnes, R., Watson, P., Allen-Vercoe, E., Moore, R.A., Holt, R.A., 2012. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. Genome Res. 22, 299–306. https://doi.org/10.1101/gr.126516.111.

Ciuffreda, L., Rodriguez-Perez, H., Flores, C., 2021. Nanopore sequencing and its application to the study of microbial communities. Comput. Struct. Biotechnol. J. 19, 1497–1511. https://doi.org/10.1016/j.csbj.2021.02.020.

Collins, M.D., Wallbanks, S., Lane, D.J., Shah, J., Nietupski, R., Smida, J., Dorsch, M., Stackebrandt, E., 1991. Phylogenetic analysis of the genus *Listeria* based on reverse

transcriptase sequencing of 16S rRNA. Int. J. Syst. Bacteriol. 41, 240–246. https://doi.org/10.1099/00207713-41-2-240.

Curry, K.D., Wang, Q., Nute, M.G., Tyshaieva, A., Reeves, E., Soriano, S., Wu, Q., Graeber, E., Finzer, P., Mendling, W., Savidge, T., Villapol, S., Dilthey, A., Treangen, T.J., 2022. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. Nat. Methods 19, 845–853. https://doi.org/10.1038/s41592-022-01520-4.

Ibal, J.C., Pham, H.Q., Park, C.E., Shin, J.H., 2019. Information about variations in multiple copies of bacterial 16S rRNA genes may aid in species identification. PLoS One 14, e0212090. https://doi.org/10.1371/journal.pone.0212090.

Illumina, 2015. 16S metagenomic sequencing library preparation. Part# 15044223 Rev. B.

Nagai, S., Sildever, S., Nishi, N., Tazawa, S., Basti, L., Kobayashi, T., Ishino, Y., 2022. Comparing PCR-generated artifacts of different polymerases for improved accuracy of DNA metabarcoding. Metabarcoding Metagenom. 6, e77704.

Oxford Nanopore, 2016. Barcode of life: simple laboratory and analysis workflows for 16S and CO1 analysis. https://nanoporetech.com/resource-centre/barcode-life-simple-laboratory-and-analysis-workflows-16s-and-co1-analysis (accessed 3 March 2023).

Pantha, K., Acharya, K., Mohapatra, S., Khanal, S., Amatya, N., Ospina-Betancourth, C., Butte, G., Shrestha, S.D., Rajbhandari, P., Werner, D., 2021. Faecal pollution source tracking in the holy Bagmati River by portable 16S rRNA gene sequencing. NPJ Clean Water. 4, 12.

Park, C., Kim, S.B., Choi, S.H., Kim, S., 2021. Comparison of 16S rRNA gene based microbial profiling using five next-generation sequencers and various primers. Front. Microbiol. 12, 715500. https://doi.org/10.3389/fmicb.2021.715500.

211 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glockner, F.O.,
212 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-
213 based tools. Nucleic Acids Res. 41, D590–596. https://doi.org/10.1093/nar/gks1219.
214 Sadowsky, M.J., Staley, C., Heiner, C., Hall, R., Kelly, C.R., Brandt, L., Khoruts, A., 2017. Analysis
215 of gut microbiota - An ever changing landscape. Gut Microbes 8, 268–275.
216 https://doi.org/10.1080/19490976.2016.1277313.
217 Scher, J.U., Sczesnak, A., Longman, R.S., Segata, N., Ubeda, C., Bielski, C., Rostron, T.,
218 Cerundolo, V., Pamer, E.G., Abramson, S.B., Huttenhower, C., Littman, D.R., 2013.
219 Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis.
220 Elife 2, e01202. https://doi.org/10.7554/eLife.01202.
221 Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q
222 File Manipulation. PLoS One 11, e0163962. https://doi.org/10.1371/journal.pone.0163962.
223 Walker, S.P., Barrett, M., Hogan, G., Flores Bueso, Y., Claesson, M.J., Tangney, M., 2020. Non-
224 specific amplification of human DNA is a major challenge for 16S rRNA gene sequence
225 analysis. Sci. Rep. 10, 16356. https://doi.org/10.1038/s41598-020-73403-7.
226 Winand, R., Bogaerts, B., Hoffman, S., Lefevre, L., Delvoye, M., Braekel, J.V., Fu, Q., Roosens,
227 N.H., Keersmaecker, S.C., Vanneste, K., 2019. Targeting the 16S rRNA gene for bacterial
228 identification in complex mixed samples: comparative evaluation of second (Illumina) and
229 third (Oxford Nanopore Technologies) generation sequencing technologies. Int. J. Mol. Sci.
230 21. https://doi.org/10.3390/ijms21010298.

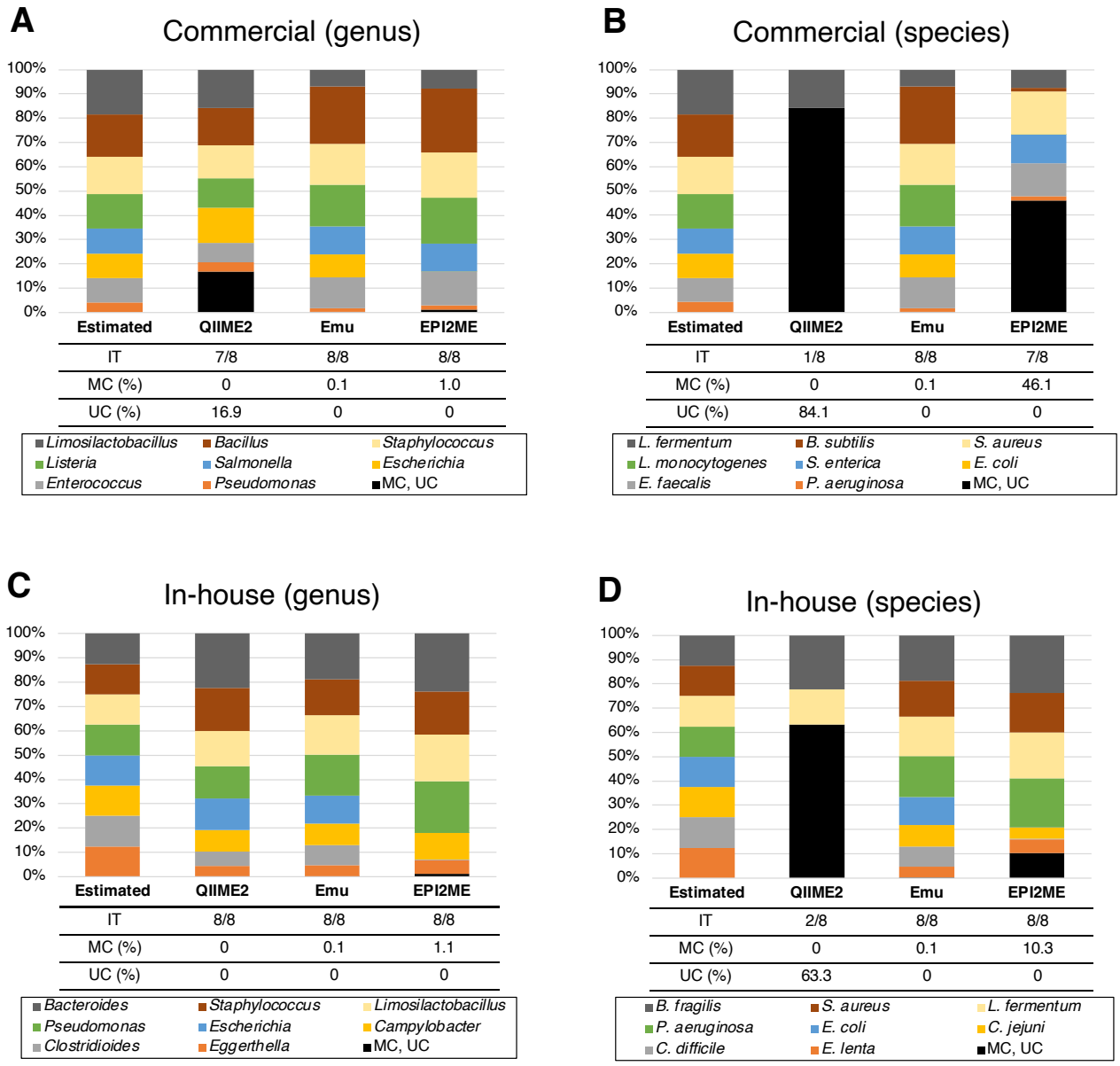
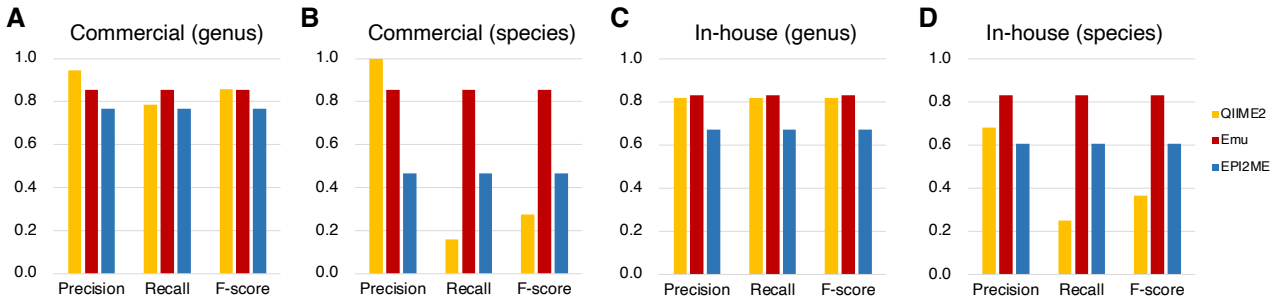


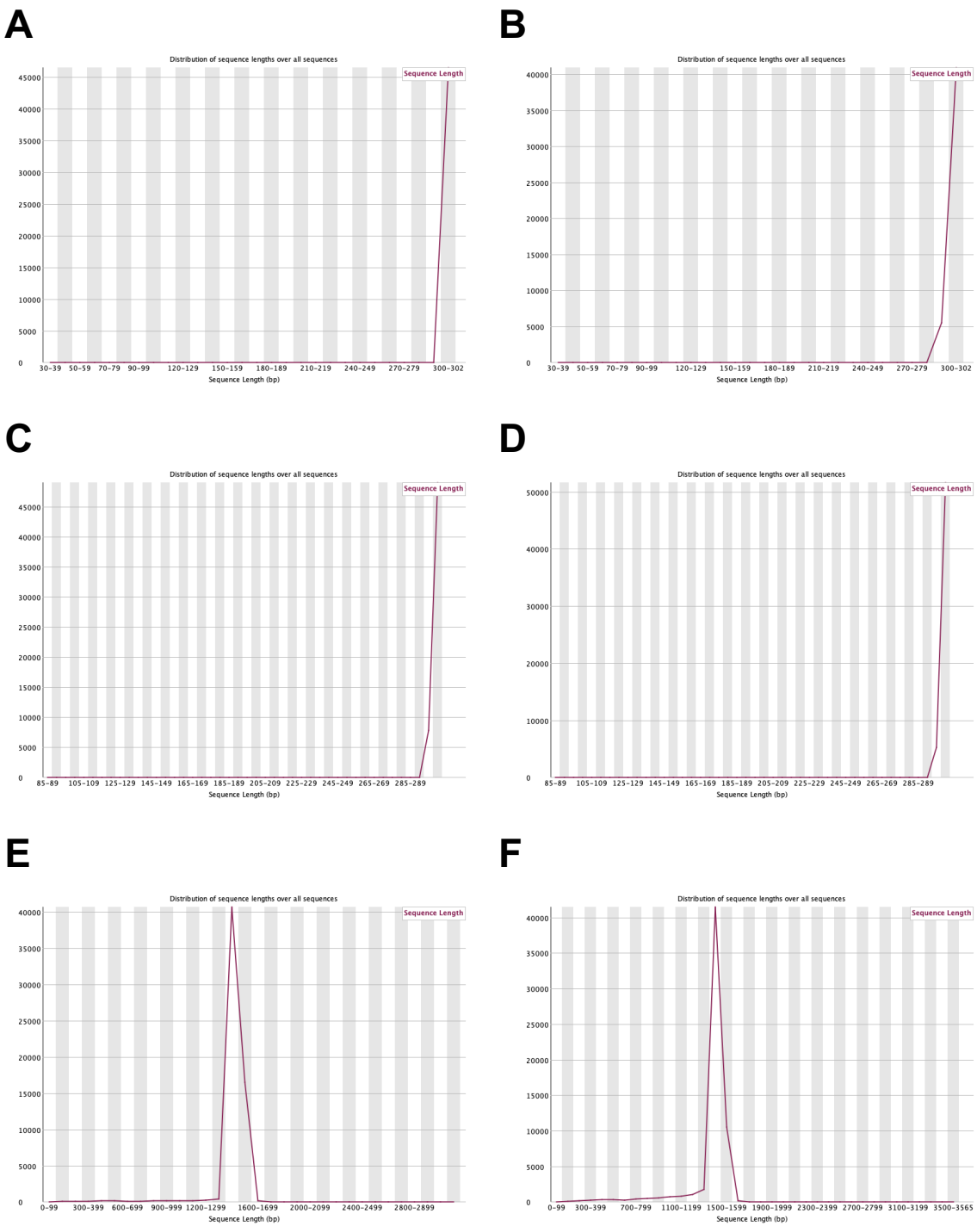

**Fig. 1.** Relative abundance of each bacterial taxon measured using each tested pipeline in the commercial (A: genus level, B: species level) and in-house (C: genus level, D: species level) samples, with the distribution of the identified taxa (IT) and unclassified (UC) and misclassified (MC) reads obtained from each pipeline.




236

237 **Fig. 2.** Comparison of precision, recall, and F-score of each tested pipeline in the commercial (A:

238 genus level, B: species level) and in-house (C: genus level, D: species level) samples.




239

240 **Supplementary Fig. 1. Evaluation of sequence length distribution using FastQC.** Short paired-
241 end reads of the commercial sample (A, B), short paired-end reads of the in-house sample (C, D),
242 long reads of the commercial sample (E), and long reads of the in-house sample (F).