Bias of the Mock Community Composed of Common Strains Isolated from Clinical Specimens

Jeong Hwan Shin (jhsmile@paik.ac.kr)
Inje University College of Medicine

Eun-Young Kim
Inje University College of Medicine

Si Hyun Kim
Semyung University

Jaewook Yang
Inje University College of Medicine

Research Article

Keywords: Metagenomics, Bioinformatics, Detection, Diagnosis, Diversity

Posted Date: January 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1304839/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

**Background:** The mock communities (MCs) bias of common species from clinical specimens should be considered when we use a microbiome in the clinical microbiology laboratory. The aims of this study were to MCs using clinically important common species and to investigate the bias of MC use in the clinical laboratory.

**Results:** Five MCs incorporating 32 bacterial strains isolated from clinical specimens were included. We analyzed the diversity of operational taxonomic units (OTU), the relative abundance, and the taxonomic assignment using paired-end sequencing of the 16S rRNA V3−V4 region. The Shannon index revealed the best correlation with the actual number of MC species for diversity. We determined that the OTU with relative abundance (log) of 0.001 is most appropriate to explain the community. The relative abundance was higher in *Bacteroidetes* and *Fusobacteria*; however, low relative abundance was shown in *Aeromonas caviae*, *Burkholderia cepacia*, *Enterobacter cloacae*, *Pseudomonas aeruginosa*, and *Clostridioides difficile*. It is suggested that the relative abundance was low in gram-positive bacteria and those with 60%−70% GC content but was not related to genome size or 16S rRNA copy number.

**Conclusions:** We investigated the characteristics of MC composed as common clinical isolates and confirmed the bias of MC. These data could be used for microbiome research on clinical specimens.

Background

Attempts to use microbiome analysis in a clinical laboratory are increasing with the enhancement of next-generation sequencing (NGS) technology [1-3]. However, the method has a few serious limitations for such use, such as a lack of standardization of the research process, distortion of bioinformatics, and inconsistent results [4-6].

Recently, it was recommended to use a mock community (MC) to overcome these limitations and increase the reliability of the results [7, 8]. An MC is composed of several bacterial strains, and we can establish the standardization of experimental process including DNA extraction, PCR, NGS, and bioinformatics pipelines using MC [7-11]. We could compare the data among researchers and reduce unnecessary duplication by others [12-14].

The bias limits can be reduced by using MC [12, 14], although it cannot replace the whole microbial community characteristics of specimens [15]. A few commercial MCs are provided by the American Type Culture Collection, BEI resource, and Zymo Research, although these are limited to use for basic research in specific sites such as oral tissue, skin, gut, and vagina. It is important to consider the GC contents, genome size, results of gram strain, and 16S rRNA copy number when we design a new MC.

The aims of this study were to design MCs using clinically important common species and to investigate the bias of MCs for use in the clinical laboratory.

Methods

Mock community

Five MCs using 32 bacteria were constructed to include clinically common and important bacteria (Table 1). All bacterial species were confirmed by biochemical testing, MALDI-TOF MS, and 16S rRNA sequencing.

The criteria for each MC were as follows: the genome size (range 1.8Mb to 5.8Mb; mean: 3.9Mb) in MC1, s GC content (range 27.0%−66.7%; mean: 44.18%) in MC2, 16S rRNA copy number (range 4−12; mean: 6) in MC 3, and 4, a total of 32 bacteria and an even distribution of gram staining results in MC 5.

All five MCs contained a total of $8.0 \times 10^8$ cells and consisted of 9−32 bacterial species. MC5 contained all 32 bacteria (each $1.6 \times 10^7$ CFU) with 12 gram-positive and 20 gram-negative bacteria. The phylum is composed of 1 Actinobacteria, 11 *Firmicutes*, 3 *Bacteroidetes*, 1 Fusobacteria, and 16 Proteobacteria. MC1 contained 9 bacteria (each $8.9 \times 10^7$ CFU) with 4 *Firmicutes*, 1 *Bacteroidetes*, and 4 Proteobacteria. MC2 contained 12 bacteria (each $6.7 \times 10^7$ CFU) with 4 *Firmicutes*, 1 Fusobacteria, and 8
Proteobacteria. MC3 contained 16 bacteria (each $5.0 \times 10^7$ CFU) with 6 *Firmicutes*, 2 *Bacteroidetes*, 1 *Fusobacteria*, and 7 *Proteobacteria*. MC4 contained 18 bacteria (each $4.4 \times 10^7$ CFU) of species not included in MC3 except *Streptococcus oralis* and *Aeromonas caviae*. For GC contents and genome size, NCBI registration information for the type strain of each species was used, and the 16s rRNA copy number was based on the average of rrnDB (Ribosomal RNA Database; University of Michigan). The GC contents, genome size, and 16S rRNA copy number of all MCs consisted of an average of 44.7% (STDEV 2.4), 3.9 Mb (STDEV 0.2), and 6.5 (STDEV 0.3), respectively.

We used phosphate-buffered saline as a negative control for DNA library construction.

### NGS and bioinformatic pipelines

DNA extraction from MCs was performed using the Maxwell 16 LEV Blood DNA Kit (Promega, USA). The DNA library construction and NGS analysis were performed by paired-end sequencing of V3−V4 regions according to the manual of the Illumina MiSeq System. The EzBioCloud pipeline (https://www.ezbiocloud.net/contents/16smtp) was used for NGS data analysis. In pre-processing, merging of paired-end reads was performed using the VSEARCH program, and low-quality reads ($<Q25$) were filtered [16]. As many as 100,000 reads were used for data analysis, and QC processing was performed to exclude low-quality, non-target, and chimeric amplicons. The UCHIME with chimera-free reference DB was used to detect any chimera [17]. Operational taxonomic unit (OTU) picking was performed by UCLUST using an open-reference method. Taxonomic assignment was performed through VSEARCH using the EzBioCloud 16S database and determined on the basis of 97% of 16S similarity [16, 18].

### Diversity of OTU

The species richness of α-diversity was estimated according to the ACE [19], Chao1 [20], and Jackknife [21] methods. The species evenness was calculated by the Shannon and Simpson index [22].

### Relative abundance and interpretation of taxonomic assignment

We used “fold error” to calculate the bias of relative abundance. Fold error is defined as the relative abundance of the ideal value, and the ideal value is calculated the same way for each MC (100 divided by the number of species). The Ez-BioCloud 16S database uses a concept of “group,” It classifies a group of several bacterial species that cannot be distinguished by 16S rRNA sequences. Of the 32 species included in this study, 23 were identified as a “group”: *Klebsiella aerogenes*, *Klebsiella pneumoniae*, and *Salmonella enterica* were included in the *Enterobacteriaceae* group (Table 1). The *Staphylococcus aureus* group included *S. aureus* and *S. epidermidis*, and the *Streptococcus pneumoniae* group included *S. pneumoniae* and *S. pyogenes*. For these species, we divided the total relative abundance of the group by the number of species for calculation of the fold error. Others were defined as the results of misidentified species other than those of the expected MC species.

### Statistics

Correlation analysis was calculated by the Pearson correlation coefficient, and polynomial logistics regression analysis was performed for bacterial variance. The p value was calculated by a two-side test, and it was considered significant when it equaled $< 0.05$.

### Results

#### Change of NGS raw data during bioinformatics pipeline

Total reads for all MCs were at least 100,000 each, ranging from 123,255 (MC2) to 184,711 (MC5) (Table 2). After pre-filtering, total reads were between 84,090 (MC4) and 100,000 (MC1, 3, and 5). All 5 MCs showed sufficient valid reads of more than
70,000. The percent of identified reads for MC2 and MC4 were 98.9% and 98.0%, respectively. The identification rates of MC1 (67.4%), MC3 (87.1%), and MC5 (89.6%) were relatively low. However, this was caused by the inclusion of a strain of *Sphingobacterium*, a suspected new species. When we treat the result of this strain (*Sphingobacterium*_uc) as correct, the final identification rates of all MCs were more than 98%. There are differences in the identified reads among 5 MCs although each MC contains the same CFUs of bacteria. The rates of misidentified reads ranged from 0.9% in MC3 to 3.6% in MC4.

**Taxonomic assignment**

The number of OTUs among the 5 MCs was between 65 (MC4) and 126 (MC5), and the number of species was 48 (MC1)−89 (MC5) (Table 2, Figure 1A). On average, 1.27 (MC2) to 2.10 (MC1 OTUs were matched to one species.

The ratio of the number of detected species to that of expected species was between 2.8 (MC4 and MC5) and 5.3 (MC1). The ratio of those at the genus level was between 1.8 (MC5) and 4.0 (MC1). These ratios were slightly decreased when an MC contained a large number of species.

**Calculation of α-diversity by OTU**

The OTU, ACE, Chao1, and Jackknife indices reflecting species richness showed a moderate correlation (r = 0.56−0.62) with the number of expected MC species (Table 2). For species evenness, the correlation between the Shannon index (range: 1.85−2.65) and the number of expected MC species was high (r = 0.82).

There were few changes in the number of OTUs if the valid reads exceeded 60,000 in all MCs (Figure 1A). The OTU ranks with relative abundance (log) > 0.001 were 16 (MC1), 18 (MC2 and MC3), 26 (MC4), and 38 (MC5) (Table 2, Figure 1B).

**Relative abundance at the species level**

There were significant differences in the relative abundance, although all expected MC species were identified to the species level (Figures 2 and 3). At the phylum, the relative abundance was highest in *Bacteroidetes* and also high in *Fusobacteria* within each MC. The relative abundance of *Fusobacterium* was highest (26.42%) in MC2 because MC2 does not contain *Bacteroidetes*. In this study, *Bacteroidetes* compromise *Bacteroides fragilis* (6.8−7.8), *Chryseobacterium gleum* group (4.4−4.8), and *Sphingobacterium* uc (2.0−2.8). The *Morganella morgani* group (1.6−1.7) and *Acinetobacter baumannii* (1.0−2.49) showed high fold error except in *Bacteroidetes* and *Fusobacterium*.

The proportion of species with fold error between 0.5 and 1.5 was 33% (3 of 9), 50% (6 of 12), 13% (2 of 16), 28% (5 of 18), and 13% (4 of 32) among MC1, MC2, MC3, MC4, and MC5, respectively. *Staphylococcus aureus* (MC1), *B. cepacia* (MC2), *C. difficile* (MC3), *E. cloacae* (MC4), and *P. aeruginosa* (MC4 and MC5) had the lowest fold error in each MC (Figure 2). The cut-off value of the relative abundance that can identify all species among MCs was the highest in MC1 (1.94%) and lowest in MC5 (0.01%) (Table 2).

A total of 422 species with 30 included in MCs were identified in the negative control (Figure 2). The relative abundance of the 30 species of MCs ranges from a minimum of 0.001% (*Enterococcus faecalis*) to a maximum of 0.58% (*Enterobacteriaceae* group).

**Misidentified results at the species level of 5 MCs**

We defined the misidentified results in the MCs when the final identification results showed a different species in the same genus or a different genus from a specific species. A different species in the same genus was confirmed in most species, and most of them were of the *Enterobacteriaceae* group, *Enterococcus*, and *Streptococcus* among MCs (Figure 4). However, the relative abundance of these was low: between 0.46% (MC3) and 1.11% (MC5). *Corynebacterium striatum*, *Clostridiodides difficile*, *Clostridium perfringens*, *Aeromonas caviae*, *Hemophilus influenzae*, and *Stenotrophomonas maltophilia* were identified.
at the species level. *Pantoea, Erwinia, Cronobacter, Cosenzaea,* and *Raoultella* were identified although they were not included in the MCs (Table 3).

**Difference of relative abundance by bacterial characteristics**

The fold error was significantly lower in gram-positive bacteria (Figure 3). There was no difference of fold error by 16S rRNA copy number or genome size. Bacteria with GC contents of 60% to 70% showed significantly lower fold error.

**Discussion**

For the diagnostic use of the microbiome in the clinical microbiology laboratory, it is necessary to verify the experimental procedures and data analysis using bacteria isolated from clinical specimens [23, 24]. In this study, we tried to confirm the bias of microbiomes using 5 MCs incorporating 32 clinical isolates.

It is well known that at least 60,000 raw reads are necessary to analyze a microbiome using biological samples. It has been reported that more than 100,000 raw reads are essential for reliable community analysis [25]. In this study, we could obtain more than 100,000 raw reads with 70,000 valid reads. We believe that 60,000 valid reads are enough to analyze the microbial community because there are no changes of OTU, and the rarefaction threshold was reached at 60,000 reads in all MCs.

The OTU to species ratio was 1.3–2.1. This is attributable to the misperception of sequencing error as OTU [26]. We could identify all species included in the 5 MCs when analyzing OTUs with a relative abundance (log) of 0.001 or higher. We believe that the use of the relative abundance (log) 0.001 can eliminate sequencing errors, and we could perform the community analysis accurately.

The species diversity can be expressed as the species richness and evenness. The richness index was different from the actual number of species in MC when we use ACE, CHAO, and Jackknife. When we calculated the Shannon and Simpson indices for species evenness, the former was highly correlated with the actual number of species in the MC. So, we concluded that the Shannon index is the best indicator of the species diversity for an MC as in a previous report [27].

For many bacteria, a partial sequence (V3–V4 region) of the 16S rRNA gene is identical or very similar in many bacteria, so it is common that one strain is expressed as two species, although it may have the opposite effect. However, the Ez-Biocloud 16S database has the advantage of reducing this erroneous result with the use of a “group.” This composes as a group for several species that cannot be distinguished by 16S rRNA amplicons. In this study, 23 of 32 species were reported as a group in this database. It is possible to analyze the microbiome data to the species level if we use the Ez-Biocloud 16S database including the concept of “group,” although many previous reports allow the identification only to the genus level.

The relative abundance was different even though the same contents of bacteria were included in each MC. We confirmed that the relative abundance was extremely low in *P. aeruginosa, E. cloacae, A. cavie, C. difficile, B. cepacia,* and *S. aureus* (Figure 2). So, it should be noted that it is likely to be calculated markedly less than the original for these species when we use clinical specimens for the microbiome. In addition, these should be considered to determine the cut-off value of the community.

It has been reported that relative abundance differs according to phylum, bacterial cell wall, GC content, genome size, and 16S rRNA copy number [28, 29]. We also confirmed the phylum of *Bacteroidetes* (*B. fragilis, Sphingobacterium* sp., *C. indologenes*), and *Fusobacterium* (*F. nucleatum*) showed high relative abundance. In addition, the relative abundance of gram-positive bacteria and GC contents of 60% to 70% was low. This bias can occur during the process of DNA extraction and PCR amplification for microbiome analysis using NGS [8, 29].

**Conclusion**

We investigated the characteristics of MC composed as common clinical isolates and confirmed the bias of MC. These data could be used for microbiome research on clinical specimens.
Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and material
All the assemblies were deposited into the NCBI SRA database with the following accession number: PRJNA801011

Competing interests
The authors declare that they have no competing interests

Funding
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2019R1I1A3A01061803) and by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HR21C1003).

Authors' contributions
EYK created a mock community and analyzed metagenome data. SHK and JY checked the analyzed data and reviewed the draft. JHS presented the conceptualization and methodology of the study. EYK and JHS were the major contributors to the drafting of the manuscript. All authors read and approved the final manuscript.

Acknowledgements
Not applicable

References
1. Schlaberg R. Microbiome Diagnostics. Clin Chem. 2020;66(1):68-76; doi: 10.1373/clinchem.2019.303248.
2. Gu W, Miller S, Chiu CY. Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection. Annu Rev Pathol. 2019;14:319-38; doi: 10.1146/annurev-pathmechdis-012418-012751.
3. Young VB. The role of the microbiome in human health and disease: an introduction for clinicians. BMJ. 2017;356:j831; doi: 10.1136/bmj.j831.
4. Golebiewski M, Tretyn A. Generating amplicon reads for microbial community assessment with next-generation sequencing. J Appl Microbiol. 2020;128(2):330-54; doi: 10.1111/jam.14380.
5. Han D, Gao P, Li R, Tan P, Xie J, Zhang R, et al. Multicenter assessment of microbial community profiling using 16S rRNA gene sequencing and shotgun metagenomic sequencing. J Adv Res. 2020;26:111-21; doi: 10.1016/j.jare.2020.07.010.
6. Brooks JP, Edwards DJ, Harwich MD, Jr., Rivera MC, Fettweis JM, Serrano MG, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol. 2015;15:66; doi: 10.1186/s12866-015-0351-6.

7. Park SC, Won S. Evaluation of 16S rRNA Databases for Taxonomic Assignments Using Mock Community. Genomics Inform. 2018;16(4):e24; doi: 10.5808/GI.2018.16.4.e24.

8. Ducarmon QR, Hormung BVH, Geelen AR, Kuijper EJ, Zwittink RD. Toward Standards in Clinical Microbiota Studies: Comparison of Three DNA Extraction Methods and Two Bioinformatic Pipelines. mSystems. 2020;5(1); doi: 10.1128/mSystems.00547-19.

9. Boukilius NA, Rideout JR, Mercurio WG, Shiffer A, Wolfe B, Maurice CF, et al. mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking. mSystems. 2016;1(5); doi: 10.1128/mSystems.00062-16.

10. Kim SH, Jeong HS, Kim YH, Song SA, Lee JY, Oh SH, et al. Evaluation of DNA extraction methods and their clinical application for direct detection of causative bacteria in continuous ambulatory peritoneal dialysis culture fluids from patients with peritonitis by using broad-range PCR. Ann Lab Med. 2012;32(2):119-25; doi: 10.3343/alm.2012.32.2.119.

11. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiol. 2016;16(1):123; doi: 10.1186/s12866-016-0738-z.

12. Bouwers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, et al. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics. 2015;16:856; doi: 10.1186/s12864-015-2063-6.

13. Yeh YC, Needham DM, Sieradzki ET, Fuhrman JA. Taxon Disappearance from Microbiome Analysis Reinforces the Value of Mock Communities as a Standard in Every Sequencing Run. mSystems. 2018;3(3); doi: 10.1128/mSystems.0023-18.

14. McGovern E, Waters SM, Blackshields G, McCabe MS. Evaluating Established Methods for Rumen 16S rRNA Amplicon Sequencing With Mock Microbial Populations. Front Microbiol. 2018;9:1365; doi: 10.3389/fmicb.2018.01365.

15. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584; doi: 10.7717/peerj.2584.

16. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011;27(16):2194-200; doi: 10.1093/bioinformatics/btr381.

17. Chao A. Estimating the population size for capture-recapture data with unequal catchability. Biometrics. 1979;45(5):52; doi: 10.2307/1936861.

18. Maguran AE, Queiroz HL, Hércos AP. Relationship between evenness and body size in species rich assemblages. Biol Lett. 2013;9(6):20130856; doi: 10.1098/rsbl.2013.0856.

19. Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, et al. Optimizing methods and dodging pitfalls in microbiome research. Microbiome. 2017;5(1):52; doi: 10.1186/s40168-017-0267-5.

20. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol. 2007;45(9):2761-4; doi: 10.1128/JCM.01228-07.

21. Singer GAC, Fahner NA, Barnes JG, McCarthy A, Hajibabaei M. Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. Sci Rep. 2019;9(1):5991; doi: 10.1038/s41598-019-42455-9.
26. Nguyen NP, Warnow T, Pop M, White B. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. NPJ Biofilms Microbiomes. 2016;2:16004; doi: 10.1038/npjbiofilms.2016.4.

27. Feranchuk S, Belkova N, Potapova U, Kuzmin D, Belikov S. Evaluating the use of diversity indices to distinguish between microbial communities with different traits. Res Microbiol. 2018;169(4-5):254-61; doi: 10.1016/j.resmic.2018.03.004.

28. Laursen MF, Dalgaard MD, Bahl Ml. Genomic GC-Content Affects the Accuracy of 16S rRNA Gene Sequencing Based Microbial Profiling due to PCR Bias. Front Microbiol. 2017;8:1934; doi: 10.3389/fmicb.2017.01934.

29. Nearing JT, Comeau AM, Langille MGI. Identifying biases and their potential solutions in human microbiome studies. Microbiome. 2021;9(1):113; doi: 10.1186/s40168-021-01059-0.

Tables

Table 1

Composition and characteristics of mock communities
| Phylum          | Species                        | Gram stain | 16S rRNA GCN\(^*\) | Genome size (Mb) | GC content (%) | MC  |
|-----------------|--------------------------------|------------|---------------------|------------------|----------------|-----|
| Actinobacteria  | Corynebacterium striatum       | +          | 4.0                 | 2.86             | 59.3           | MC4 MC5 |
| Firmicutes      | Bacillus cereus                | +          | 13.6                | 5.76             | 35.0           | MC1 MC4 MC5 |
|                 | Clostridioides difficile       | +          | 11.9                | 4.17             | 28.6           | MC3 MC5 |
|                 | Clostridium perfringens        | +          | 9.9                 | 3.47             | 28.1           | MC2 MC4 MC5 |
|                 | Enterococcus faecalis          | +          | 4.0                 | 2.97             | 37.4           | MC1 MC2 MC3 MC5 |
|                 | Enterococcus faecium           | +          | 6.0                 | 2.91             | 37.9           | MC3 MC5 |
|                 | Staphylococcus aureus\(^*\)    | +          | 5.6                 | 2.84             | 32.7           | MC2 MC4 MC5 |
|                 | Staphylococcus epidermidis\(^*\) | +      | 5.8                 | 2.52             | 32.0           | MC1 MC4 MC5 |
|                 | Streptococcus agalactiae       | +          | 6.8                 | 2.08             | 35.4           | MC4 MC5 |
|                 | Streptococcus oralis\(^**\)    | +          | 4.0                 | 1.97             | 41.1           | MC1 MC3 MC4 MC5 |
|                 | Streptococcus pneumoniae\(^**\) | +      | 4.0                 | 2.09             | 39.6           | MC2 MC3 MC5 |
|                 | Streptococcus pyogenes         | +          | 5.8                 | 1.79             | 38.4           | MC3 MC5 |
| Bacteroidetes   | Bacteroides fragilis           | -          | 6.0                 | 5.27             | 43.4           | MC4 MC5 |
|                 | Sphingobacterium sp.           | -          | 5.3                 | 5.27             | 39.8           | MC1 MC3 MC5 |
|                 | Chryseobacterium indologenes   | -          | 6.0                 | 4.89             | 37.2           | MC3 MC5 |
| Fusobacteria    | Fusobacterium nucleatum        | -          | 4.5                 | 2.41             | 27.0           | MC2 MC3 MC5 |
| Proteobacteria  | Acinetobacter baumannii        | -          | 5.9                 | 3.97             | 39.0           | MC2 MC3 MC5 |
|                 | Aeromonas caviae               | -          | 10.0                | 4.54             | 61.4           | MC1 MC2 MC3 MC4 MC5 |
|                 | Burkholderia cepacia           | -          | 5.5                 | 8.49             | 66.7           | MC2 MC3 MC4 MC5 |
|                 | Enterobacter cloacae           | -          | 8.0                 | 4.96             | 55.1           | MC4 MC5 |
|                 | Escherichia coli               | -          | 7.0                 | 5.13             | 50.6           | MC2 MC3 MC4 MC5 |
|                 | Haemophilus influenzae         | -          | 6.0                 | 1.85             | 38.0           | MC1 MC2 MC3 MC4 MC5 |
|                 | Klebsiella aerogenes\(^***\)  | -          | 8.0                 | 5.26             | 55.0           | MC1 MC3 MC5 |
|                 | Klebsiella pneumoniae\(^***\) | -          | 8.0                 | 5.59             | 57.2           | MC2 MC4 MC5 |
|                 | Moraxella catarrhalis          | -          | 4.0                 | 1.92             | 41.6           | MC3 MC5 |
|                                |     |     |     |          |      |      |
|--------------------------------|-----|-----|-----|----------|------|------|
| **Morganella morganii**        | -   | 4.0 | 3.96| 51.0     | MC3  | MC5  |
| **Neisseria gonorrhoeae**      | -   | 4.0 | 2.14| 52.4     | MC2  | MC4  | MC5  |
| **Proteus mirabilis**          | -   | 6.9 | 3.96| 38.8     | MC3  | MC5  |
| **Salmonella enterica**        | -   | 6.9 | 4.8 | 52.1     | MC1  | MC4  | MC5  |
| **Serratia marcescens**        | -   | 7.0 | 5.2 | 59.8     | MC3  | MC5  |
| **Stenotrophomonas maltophilia**| -  | 3.8 | 4.57| 66.4     | MC4  | MC5  |

*, **, ***: 16s rRNA V3-V4 region sequence not differentiated.
†: gene copy number.

Table 2

Metagenomic results of 5 MCs
|                      | PBS (0) | MC1 (9) | MC2 (12) | MC3 (16) | MC4 (18) | MC5 (32) |
|----------------------|---------|---------|----------|----------|----------|----------|
| **Reads by NGS data processing** |         |         |          |          |          |          |
| Raw data (total reads) | 23,760  | 170,464 | 123,255  | 165,036  | 124,951  | 184,711  |
| Pre-filter* (Q25)      | 17,846  | 100,000 | 90,529   | 100,000  | 84,090   | 100,000  |
| (QC) Low quality amplicons | 3,464   | 161     | 192      | 131      | 113      | 137      |
| (QC) Non-target amplicons | 2      | 0       | 1        | 0        | 0        | 0        |
| (QC) Chimeric amplicons | 587    | 11,143  | 20,051   | 13,476   | 9,807    | 13,697   |
| Valid reads            | 13,793  | 88,696  | 70,285   | 86,393   | 74,170   | 86,166   |
| Identified reads† (%)‡ | 12,884  | 59,796  | 69,531   | 75,259   | 72,655   | 77,186   |
| Valid reads            | 13,793  | 88,696  | 70,285   | 86,393   | 74,170   | 86,166   |
| Misidentified reads§ (%)‡ | 13,793  | 2349   | 1583     | 735      | 2645     | 2369     |

| **Taxonomic assignment (detected/expected, fold)** |         |         |          |          |          |          |
| OTU (r=0.56)* | 441 | 101 | 76 | 101 | 65 | 126 |
| Phylum | 12 | 5/3 (1.7) | 4/3 (1.3) | 5/4 (1.3) | 4/4 (1.0) | 5/5 (1.0) |
| Genus | 200 | 36/9 (4.0) | 34/12 (2.8) | 28/13 (3.1) | 34/16 (2.1) | 46/25 (1.8) |
| Species | 422 | 48/9 (5.3) | 60/12 (5.0) | 50/16 (3.1) | 50/18 (2.8) | 89/32 (2.8) |

| **Species richness** |         |         |          |          |          |          |
| ACE (r=0.62)* | 449.32 | 107.89 | 85.45 | 104.69 | 71.64 | 139.74 |
| CHAO (r=0.61)* | 499.57 | 102.69 | 80.80 | 101.85 | 66.75 | 133.31 |
| Jackknife (r=0.60)* | 466.00 | 113.00 | 92.00 | 109.00 | 73.00 | 146.00 |

| **Species evenness** |         |         |          |          |          |          |
| Expected Shannon (r=0.98)* | - | 2.2 | 2.49 | 2.77 | 2.89 | 3.47 |
| Shannon (r=0.82)* | 4.67 | 1.85 | 2.33 | 2.21 | 2.07 | 2.65 |
| Expected Simpson (r=-0.92)* | - | 0.11 | 0.08 | 0.06 | 0.06 | 0.03 |
| Simpson (r=-0.51)* | 0.03 | 0.21 | 0.13 | 0.15 | 0.23 | 0.11 |

| **OTU rank** | Relative abundance (log) > 0.001 | 286 | 16 | 18 | 18 | 26 | 38 |

| **Cut off value for MC species** |         |         |          |          |          |          |
| The lowest relative abundance (%) | - | 1.94 | 1.85 | 0.03 | 0.06 | 0.01 |

| **Relative abundance > 0.01** |         |         |          |          |          |          |
| Identification of species | 422 | 24 | 32 | 35 | 35 | 59 |
| Identification of MC species | - | 8** | 12 | 15** | 16** | 28** |
| Misidentification of species | 422 | 16 (12) | 20 (13) | 20 (15) | 19 (8) | 31 (20) |
(Similar species) ‡‡

*, After pre-filtering, no more than 100,000 reads were used for analysis.

†, Number of reads identified at the species level.

‡, Ratio of each read divided by valid reads.

§, Number of reads misidentified at the species level.

¶, Correlation coefficient with number of MC species.

**, Contains species indistinguishable from the 16s rRNA V3-V4 region.

††, Identified bacteria as species similar to MC species.

Table 3

| Genus       | Misidentified species          | MC1 | MC2 | MC3 | MC4 | MC5 |
|-------------|--------------------------------|-----|-----|-----|-----|-----|
| Lactococcus | *Lactococcus lactis* group     | 0.01|     |     |     |     |
| Citrobacter | *Citrobacter koseri*          |     |     |     |     |     |
| Cosenzaea   | *Cosenzaea myxofaciens*       |     |     |     |     |     |
| Cronobacter | *Cronobacter dublinensis* group |   |     |     |     |     |
|             | *Cronobacter sakazakii*       | 0.13| 0.14|     |     |     |
| Erwinia     | *Erwinia aphidicola*          |     |     |     |     |     |
|             | *Erwinia injecta*             | 0.06|     | 0.02|     |     |
|             | *Erwinia tasmaniensis*        |     |     |     | 0.01|     |
| Kosakonia   | *Kosakonia cowanii* group      |     |     | 0.09| 0.02|     |
| Pantoea     | *AKIU_s*                      | 0.01|     |     |     |     |
|             | *OCMY_s*                      |     | 0.04| 0.02|     |     |
|             | *Pantoea agglomerans* group   | 0.08|     | 0.11| 0.08|     |
|             | *Pantoea coffeiphila*         | 0.25| 0.03| 0.11| 0.04|     |
|             | *Pantoea theicola*            | 0.02|     | 0.01|     |     |
|             | *Pantoea UC*                  | 0.06| 0.06| 0.25| 0.11|     |
| Pectobacterium | *Pectobacterium carotovorum* group | 0.05|     | 0.01|     |     |
| Pseudescherichia | *Pseudescherichia vulneris* |     |     |     | 0.01|     |
| Raoultella  | *Raoultella planticola* group | 0.12| 0.03| 0.02|     |     |

Figures
Figure 1. α-Diversity of 5 MCs. (A) Rarefaction curve. (B) Rank abundance.

The following symbols are recommended:

- MC1
- MC2
- MC3
- MC4
- MC5

Figure 1

Please See image above for figure legend.
### Relative abundance

|                      | PBS | MC1 | MC2 | MC3 | MC4 | MC5 |
|----------------------|-----|-----|-----|-----|-----|-----|
| *Corynebacterium minitissum* group | 0.09 | 2.09 | 2.15 |
| *Bacillus cereus* group | 0.04 | 5.97 | 2.43 | 0.18 |
| *Clostridioides difficile* group | 0.12 | 13.92 | 6.68 | 0.70 |
| *Clostridium perfringens* group | 0.01 | 7.59 | 2.93 | 1.74 | 1.24 |
| *Enterococcus faecalis* group | 0.25 | 2.55 | 2.12 |
| *Enterococcus faecium* group | 0.04 | 1.94 | 9.38 | 2.46 | 0.55 |
| *Staphylococcus aureus* group | 0.37 | 20.39 | 6.00 | 0.65 | 1.72 | 4.74 |
| *Streptococcus pyogenes* group | 0.06 | 5.12 | 3.70 |
| *Bacteroides fragilis* group | 0.12 | 43.50 | 21.13 |
| *Chryseobacterium glucum* group | 0.23 | 27.23 | 14.93 |
| *Sphingobacterium_usage* group | 0.09 | 30.58 | 12.51 | 9.07 |
| *Fusobacterium nucleatum* group | 0.13 | 26.42 | 14.88 | 8.75 |
| *Actinobacter baumannii* group | 0.06 | 8.33 | 9.56 | 7.79 |
| *Aeromonas caviae* group | 0.14 | 2.27 | 2.17 | 0.05 | 1.04 | 0.02 |
| *Burkholderia cepacia* group | 0.34 | 1.85 | 0.38 | 0.10 |
| *Enterobacter cloacae* group | 0.06 | 0.02 |
| *Enterobacteriaceae* group | 0.58 | 24.85 | 3.85 | 0.93 | 15.57 | 9.47 |
| *Escherichia coli* group | 0.32 | 5.69 | 3.97 | 2.23 |
| *Haemophilus influenzae* group | 0.11 | 3.76 | 6.16 | 1.66 | 0.28 |
| *Moraxella nonliquefaciens* group | 0.12 | 2.97 | 0.73 |
| *Morganella morganii* group | 0.15 | 11.18 | 5.09 |
| *Neisseria gonorrhoeae* group | 0.14 | 11.06 | 1.99 | 0.18 |
| *Proteus mirabilis* group | 0.07 | 8.49 | 1.30 |
| *Pseudomonas aeruginosa* group | 0.08 | 0.06 | 0.01 |
| *Serratia marcescens* group | 0.07 | 1.25 | 0.94 |
| *Stenotrophomonas maltophilia* group | 0.07 | 3.00 | 0.67 |
| Others | 0.16 | 2.65 | 2.25 | 0.85 | 3.57 | 2.75 |

### Fold error

|                      | MC1 | MC2 | MC3 | MC4 | MC5 |
|----------------------|-----|-----|-----|-----|-----|
| *Corynebacterium minitissum* group | 0.54 | 0.44 | 0.06 | 0.22 |
| *Bacillus cereus* group | 0.00 | 0.02 |
| *Clostridioides difficile* group | 1.67 | 1.20 | 0.40 |
| *Clostridium perfringens* group | 0.68 | 0.35 | 0.28 | 0.40 |
| *Enterococcus faecalis* group | 0.17 | 1.13 | 0.22 | 0.09 |
| *Enterococcus faecium* group | 1.84 | 0.72 | 0.05 | 1.57 | 0.76 |
| *Streptococcus pyogenes* group | 0.82 | 0.18 |
| *Bacteroides fragilis* group | 7.82 | 6.75 |
| *Chryseobacterium glucum* group | 2.75 | 2.00 | 2.90 |
| *Sphingobacterium_usage* group | 2.17 | 2.38 | 2.80 |
| *Fusobacterium nucleatum* group | 1.00 | 1.53 | 2.49 |
| *Actinobacter baumannii* group | 0.20 | 0.26 | 0.01 | 0.19 | 0.01 |
| *Aeromonas caviae* group | 0.22 | 0.07 | 0.03 |
| *Burkholderia cepacia* group | 0.01 | 0.01 |
| *Enterobacter cloacae* group | 1.12 | 0.46 | 0.15 | 1.40 | 1.01 |
| *Escherichia coli* group | 0.68 | 0.71 | 0.71 |
| *Haemophilus influenzae* group | 0.34 | 0.74 | 0.30 | 0.09 |
| *Moraxella nonliquefaciens* group | 0.48 | 0.23 |
| *Morganella morganii* group | 1.79 | 1.63 |
| *Neisseria gonorrhoeae* group | 1.33 | 0.36 | 0.06 |
| *Proteus mirabilis* group | 1.36 | 0.41 |
| *Pseudomonas aeruginosa* group | 0.01 | 0.00 |
| *Serratia marcescens* group | 0.20 | 0.30 |
| *Stenotrophomonas maltophilia* group | 0.54 | 0.22 |

**Figure 2.** Relative abundance and fold errors of 5 MCs.

**Figure 2**

Please See image above for figure legend.
Figure 3. Fold error by bacterial variances.

**Figure 3**

Please See image above for figure legend.
Figure 4. Relative abundance of misidentified bacteria showing a different species in the same genus.

*Figure 4*

Please See image above for figure legend.