Communication

Diversity of the Japanese Gut Microbiome

Tatsuki Itagaki , Ken-ichiro Sakata , Akira Hasebe * , Yoshimasa Kitagawa

Posted Date: 5 February 2024

doi: 10.20944/preprints202402.0275.v1

Keywords: gut microbiome; compositional data; principal component analysis; machine learning; diversity doi:10.20944/preprints202402.0275.v1

Communication

Diversity of the Japanese Gut Microbiome

Tatsuki Itagaki [1,2], Ken-ichiro Sakata [1], Akira Hasebe [2,*] and Yoshimasa Kitagawa [1]

[1] Oral Diagnosis and Medicine, Faculty of Dental Medicine, Graduate School of Dental Medicine, Hokkaido University, Kita-13 Nishi-7, Kita-ku, Sapporo 060-8586, Japan
[2] Oral Molecular Microbiology, Faculty of Dental Medicine, Graduate School of Dental Medicine, Hokkaido University, Kita-13 Nishi-7, Kita-ku, Sapporo 060-8586, Japan
* Correspondence: akkun@den.hokudai.ac.jp; Tel.: +81-11-706-4240

Simple Summary: A compositional data vector is a special type of multivariate observation in which the elements of the vector are non-negative and sum to a constant, usually taken to be unity. A compositional data only retains relative information. Furthermore, comparisons can only be made between compositional data of the same component. The relevant sample space is the standard simplex. A simplex space is a space that is a generalized form of a triangle. For compositional data, many of the operations defined in Euclidean space are meaningless. Microbiome analyzes have become popular in recent years. Operational Taxonomic Units (OTU) used in microbiome analysis are one type of composition data. The bacterial flora data are compositional, with information only about relative abundance.

Abstract: Firmicutes and Bacteroidetes, the major phyla in the colon, have been observed in humans worldwide. Gut microbiome analyses often use the Firmicutes/Bacteroidetes ratio and principal coordinates analysis (PCoA). However, misinformation is pervasive in the human microbiome literature and analysis. An attempt was made to demonstrate how to analyze using the "National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN JMD) (Public Data)." The results showed that PCoA did not work, and principal component analysis (PCA) was useful for analyzing the gut microbiome alpha diversity.

Keywords: gut microbiome; compositional data; principal component analysis; machine learning; diversity

1. Introduction

Basis is a non-constrained data (counting data) [1–5]. Composition corresponds to the constrained data [1–5]. When considering the ratios of compositions, the coefficient of variation is affected by moments up to the fourth order of basis. The coefficient of variation of ratios is subject to change when the unchanging component is switched between the denominator and numerator [1]. Compositional data cannot be used to calculate sums, differences, products, or quotients [1–5]. When the constituent elements are not the same, the appearance patterns will be different, and the differences in each component cannot be compared using composition data [1–5]. The correlation and rank correlation coefficients do not mean the correlation on compositional data [[2], Supplementary Materials]. The comparisons of the compositional components varied depending on the denominator size [1–5]. In other words, while the proportion is one for the whole, it is common to extract a portion and renormalize the whole to one with the data from that portion, resulting in an apparent change. Changes in percentages represent apparent changes in relative abundance [3–5]. A simple example of calculating percentages is that if one mixes 10% saline and 20% saline, the resulting concentration will be between 10% and 20%, depending on the amount added to each concentration. To determine whether a 10% saline solution or a 20% saline solution contains more salt, we need to know the total

doi:10.20944/preprints202402.0275.v1

amount of saline solution. However, this argument cannot be made if the total quantity is unknown. This indicates that the 20% saline solution had a higher percentage of salt.

Principal component analysis (PCA) and multidimensional scaling (MDS or principal coordinates analysis/PCoA) are often used for gut microbiome analyses [6–23]. Use of these analyzes requires algebraic knowledge. A simplex is a generalization of a triangle or a tetrahedron [3–5]. PCA corresponds to the task of finding a specific Cartesian coordinate axis in a p-dimensional subspace within a n-dimensional vector space and synthesizing principal component vectors on that Cartesian coordinate axis. The ordinary PCA uses n subjects that are independent of each other as the coordinate axes of a n-dimensional vector space, and p types of observed items that are correlated with each other as the coordinate axes of a p-dimensional subspace. In the case of composition data, it is a (p-1)-dimensional subspace [1]. PCA is a commonly used analysis to show relative relationships. In contrast, the multidimensional scaling method uses p observation items that are independent of each other as the coordinate axes of a p-dimensional vector space, and on subjects that are treated as correlated in terms of their similarity or distance to each other as the coordinate axes of a n-dimensional subspace. Considering these points, it is contradictory to apply PCA and multidimensional scaling to the same data at the same time. When constructing a positional relationship in Euclidean space from a distance matrix representing dissimilarity in multidimensional scaling, even the same composition data are often plotted at different positions in Euclidean space. Furthermore, the distances commonly used in multidimensional scaling methods are the Bray-Curtis distance and the Jaccard distance, but these are only used in Euclidean space and cannot be used for compositional data in simplex space. For ordinary medical data involving intestinal flora, it is more reasonable to apply PCA rather than the multidimensional scaling construction method.

Firmicutes and Bacteroidetes, the major phyla in the colon, have been observed in humans worldwide. The five major colonic phyla in Japanese individuals were Firmicutes, Bacteroidetes, Actinobacteria, Proteobacteria, and Fusobacteria. Gut microbiome analysis is a field of information engineering. Moreover, in recent years, bacteriological studies have sometimes been conducted using the results from the gut microbiome analysis [6–23]. However, misinformation is pervasive in the human microbiome literature [6]. Operational taxonomic units (OTU) are the results of clustering the 16S rRNA gene sequences at a certain cutoff value [7]. Many recent microbiome analyses have compared the proportions of compositional data. Many studies apply mathematical incorrect methods (multidimensional scaling, correlation coefficient and arithmetic means) for analyzing the gut microbiome [8–23]. Misinformation on microbiota analysis can cause great harm to bacteriology professionals.

Gut microbiome analyses often use the Firmicutes/Bacteroidetes (F/B) ratio [11–15]. Many studies have explored the relationship between the F/B ratio and the evaluation index [11–15]. The ratio is the same for both basis and composition. The F/B ratio is always the same when the results are given in OTU or in any subcomposition. Using the ratio of Bacteroidetes, the relative abundance of other bacteria was compared to that of Bacteroidetes in previous research [11–15]. However, most studies did not provide reasons for using the ratio of Bacteroidetes [11–15]. Moreover, many microbiota analyses have been performed assuming the same number of bacteria among individuals, with many standardized 10,000 OTU reads [8–23]. However, the assumption that the number of bacteria is the same among individuals is a severe condition that is not necessarily met. The denominator can be chosen arbitrarily for a ratio, as long as it's not zero. If the denominator with the smallest difference between individuals is selected, the ratio only requires consideration of the numerator's effect. A ratio analysis was not done since the same constituent components of the OTU data were little in this study. Therefore, the purpose of this study was to clarify the misconceptions regarding diversity of the Japanese gut microbiome by showing the difference between PCoA and PCA.

2. Materials and Methods

Datasets

The artificial composition data shown in Table 1 was used.

The OTU dataset was extracted from the "NIBIOHN JMD (Public Data)." This study evaluated the fecal samples and metadata of community-dwelling Japanese volunteers from a cohort of the health- and nutrition-based cohort study conducted by the NIBIOHN, hereafter referred to as the NIBIOHN cohort [23]. In the NIBIOHN cohort, 1518 healthy Japanese adult volunteers were analyzed from October 2015 to October 2020 (age range: 19.5–80 years; males: 693; females: 825; 6 regions: Kanagawa, Atsugi; Niigata, Minamiuonuma; Osaka, Osaka; Tokyo, Shinjuku; Yamaguchi, Shunan; Yamanashi, Chuo) [23].

Table 1. The artificial compositional data.

A	B	C	D
0.1	0.1	0.4	0.4
0.1	0.2	0.35	0.35
0.1	0.3	0.3	0.3
0.1	0.4	0.25	0.25
0.1	0.5	0.2	0.2
0.1	0.6	0.15	0.15
0.1	0.7	0.1	0.1
0.1	0.8	0.05	0.05
0.1	0.1	0.4	0.4
0.1	0.2	0.35	0.35
0.1	0.3	0.3	0.3
0.1	0.4	0.25	0.25
0.1	0.5	0.2	0.2
0.1	0.6	0.15	0.15
0.1	0.7	0.1	0.1
0.1	0.8	0.05	0.05
0.1	0.1	0.4	0.4
0.1	0.2	0.35	0.35
0.1	0.3	0.3	0.3
0.1	0.4	0.25	0.25
0.1	0.5	0.2	0.2
0.1	0.6	0.15	0.15
0.1	0.7	0.1	0.1
0.1	0.8	0.05	0.05
0.1	0.1	0.4	0.4
0.1	0.2	0.35	0.35
0.1	0.3	0.3	0.3
0.1	0.4	0.25	0.25
0.1	0.5	0.2	0.2
0.1	0.6	0.15	0.15

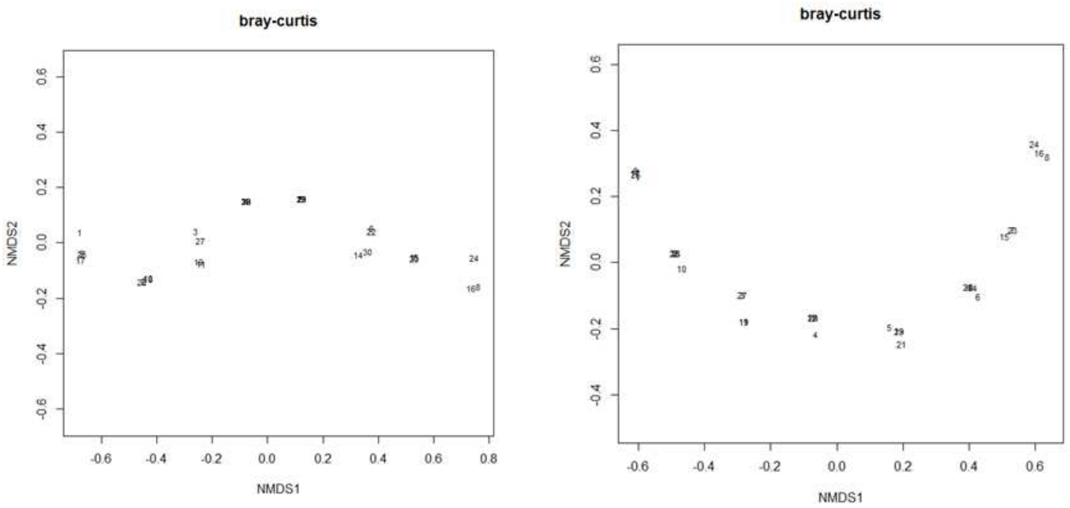
Procedure

All data was analyzed using R version 4.3.1 (2023-06-16 ucrt) and MANTA (Microbiota And pheNoType correlation Analysis platform) [23]. Vegan 2.6-4 package was used. A multidimensional scaling method (principal coordinate analysis/PCoA) was performed using the artificial composition data (Table 1). Subsequently, PCoA and PCA were performed on the OTU Dataset. PCA was performed using sample variance without standardizing the data. Three PCoA charts were shown as nonmeaning figures of an index of the diversity. Scatter plots of principal component scores were shown as an index of α-diversity.

Finally, the relationship between green tea intake and bacterial flora of genus level was investigated using principal component regression analysis. The information on green tea intake was adopted because it had fewer 0s than the others.

3. Results

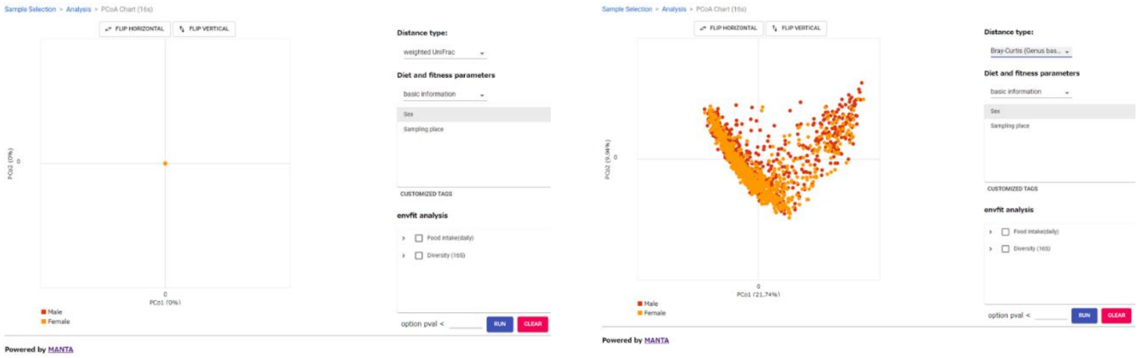
The PCoA results were shown in Figure 1 for the artificial dataset shown in Table 1. Different analysis results were obtained for the same data. Therefore, the results of PCoA do not make any mathematical sense and therefore have no scientific basis.



(**a**) (**b**)

Figure 1. The results of principal coordinates analysis of the same data.

The PCoA results were shown in Figure 2. The contribution rate of principal coordinates 1 and 2 was small. This indicates that dimensionality reduction failed. Moreover, these results had no scientific meaning for the reasons mentioned above [1–5].



Distance type: weighted UniFrac

Diet and fitness parameters: basic information

Distance type: Bray-Curtis (Genus bas...

Diet and fitness parameters: basic information

(**a**) (**b**)

Distance type: Jaccard (Genus based)

Diet and fitness parameters: basic information

envfit analysis

☐ Food intake(daily)

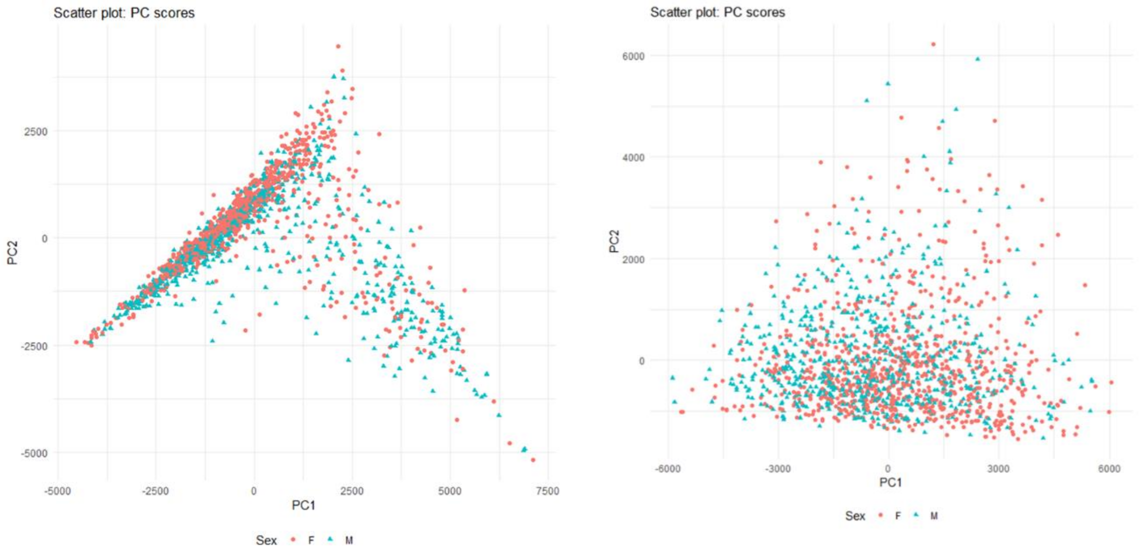
☐ Diversity (16S)

option pval <

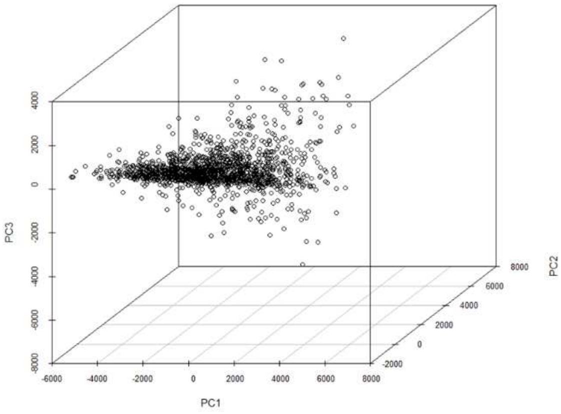
(c)

Figure 2. PCoA results were shown: (**a**); weighted UniFrac distance. (**b**); Bray-Curtis distance. (c); Jaccard distance. All plots were in Euclidean space.

The PCA results were shown in Figure 3. Principal component scores were plotted in simplex space. The contribution rate of principal components 1 and 2 at the genus level was 66.3%. The contribution rate of principal components 1 and 2 at the class level was 81.0%. These plots summarize the results of a 100% stacked bar graph, thus indicating alpha diversity. The adjusted R-squared of principal component regression analysis for green tea intakes was 0.00338 for PC1 and 0.0006333 for PC2, and no relationship was observed (Figure 4). The adjusted R-squared was small, so the fit of the regression model to the data was poor.



(**a**) (**b**)



(c)

Figure 3. PCA results were shown: (**a**); genus. (**b**); class. (c); class. All plots were in simplex space.

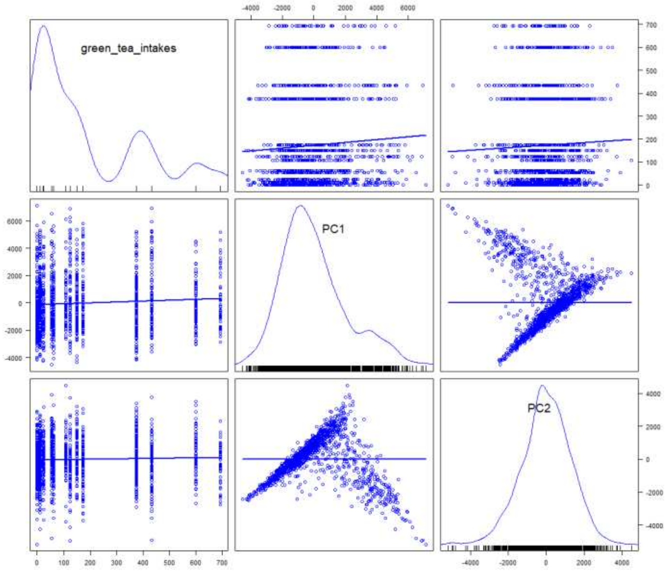


Figure 4. The scatterplots. The regression lines were nearly horizontal.

4. Discussion

An example using artificial data proved that PCoA did not make sense. The analysis of beta and alpha diversity using multidimensional scaling was not good because it assumed no correlation between bacteria [8–23]. ANORSIM, ADONIS, and other analyses may result in small p-values and small values of R and R-squared values. Sometimes, the number of null distributions generated was small, but this only indicated little consistency between the numerical data values, statistical model, and null hypothesis, probably resulting from ignoring correlations between bacteria. On the other, PCA was better because it considered correlations among bacteria. However, the information possessed by compositional data limited PCA because only one composition had only enough information to add up to 100% of the composition [1–5]. A scatterplot of principal component scores was a summary of the 100% stacked bar graph. Thess scatterplots showed alpha diversity (Figure 2).

Analysis using principal component scores is a univariate analysis. It is necessary to select the classification level used for PCA depending on the number of subjects. As shown in Figure 4, there was no relative relationship. However, this was of little scientific significance as it explored the relationship to relative change. It is difficult to show a true relationship between bacterial flora and diet. Absolute changes rather than relative changes should be examined, but there are currently few methods for doing so [[1–5], Supplementary Materials]. Ratio analysis is effective in solving this problem, but the components that take the ratio must have the same absolute amount [[1], Supplementary Materials]. It is difficult to identify components whose absolute amounts is same, and there are not necessarily such kinds of components [[1], Supplementary Materials]. Therefore, the absolute relationship cannot be investigated in this study.

A limitation of this study was the research design. This was a study that collected data on bacterial composition between individuals. Since it is only possible to compare compositions of the same component, it is desirable to conduct research that examines changes over time within individuals.

5. Conclusions

Many studies apply mathematical incorrect methods (multidimensional scaling, correlation coefficient and arithmetic means) for analyzing the gut microbiome. PCA is recommended for compressing these dimensions. For analyzing the gut microbiome, it is desirable to study changes over time within an individual.

Supplementary Materials: The following supporting information can be downloaded at: Table S1: Sample dataset on artificial composition, Table S2: Sample dataset on artificial basis, Table S3: Statistics of the sample data, Figure S1: Absolute variation and apparent variation of F, Figure S2(a): Absolute variation and apparent variation of G, Figure S2(b): The difference between absolute and apparent variation of G, Figure S3(a): Absolute variation and apparent variation of H, Figure S3(b): The difference between absolute and apparent variation of H.

Author Contributions: Conceptualization, T.I. and K.S..; methodology, T.I..; software, T.I..; validation, T.I., A.H. and Y.K.; formal analysis, T.I.; investigation, T.I.; resources, T.I.; data curation, T.I.; writing—original draft preparation, T.I.; writing—review and editing, T.I. and Y.K.; visualization, T.I.; supervision, T.I. and A.H.; project administration, T.I.; funding acquisition, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This retrospective study procedures were performed in accordance with the principles of the Declaration of Helsinki.

Informed Consent Statement: This article does not disclose identifiable information of any of the participants in any form. Hence, consent for publication is not applicable in this case.

Data Availability Statement: The OTU data that support the findings of this study are accessible at https://microbiome.nibiohn.go.jp/jmd-public. (The authors accessed on 2023.07.03.)

Acknowledgments: The authors would like to thank Norio Sugimoto for suggesting the idea of using principal component analysis. He then explained to us the difference between principal component analysis and multidimensional scaling in an easy-to-understand manner.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ohta, T.; Arai, H.; Noda, A. Identification of the unchanging reference component of compositional data from the properties of the coefficient of variation. *Math. Geosci.* **2011**, *43*, 421–434. https://doi.org/10.1007/s11004-011-9332-y
2. Pearson, K. Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc. Lond.* **1897**, *60*, 489–498. https://doi.org/10.1098/rspl.1896.0076
3. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. Series B (Methodol)* **1982**, *44*, 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x
4. Aitchison, J.; Egozcue, J. The statistical analysis of geochemical compositions. *Math. Geol.* **1984**, *16*, 531–564. https://doi.org/10.1007/BF01029316
5. Aitchison, J. Compositional data analysis: Where are we and where should we be heading? *Math. Geol.* **2005**, *37*, 829–850. https://doi.org/10.1007/s11004-005-7383-7

6. Walker, A. W., & Hoyles, L. (2023). Human microbiome myths and misconceptions. Nature Microbiology, 8(8), 1392–1396. https://doi.org/10.1038/s41564-023-01426-7
7. Mysara, M.; Vandamme, P.; Props, R.; Kerckhof, F.-M.; Leys, N.; Boon, N.; Raes, J.; Monsieurs, P. Reconciliation between operational taxonomic units and species boundaries. *FEMS Microbiology Ecology* **2017**, *93*, fix029. https://doi.org/10.1093/femsec/fix029
8. Fan, Y.; Støving, R.K.; Berreira Ibraim, S.; Hyötyläinen, T.; Thirion, F.; Arora, T.; Lyu, L.; Stankevic, E.; Hansen, T.H.; Déchelotte, P.; et al. The gut microbiota contributes to the pathogenesis of anorexia nervosa in humans and mice. *Nature Microbiology* 2023, *8*, 787-802. https://doi.org/10.1038/s41564-023-01355-5
9. Chen, X.; Hashimoto, D.; Ebata, K.; Takahashi, S.; Shimizu, Y.; Shinozaki, R.; Hasegawa, Y.; Kikuchi, R.; Senjo, H.; Yoneda, K.; et al. Reactive granulopoiesis depends on T-cell production of IL-17A and neutropenia-associated alteration of gut microbiota. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2211230119. https://doi.org/10.1073/pnas.2211230119
10. Turnbaugh, P.J.; Ley, R.E.; Mahowald, M.A.; Magrini, V.; Mardis, E.R.; Gordon, J.I. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **2006**, *444*, 1027–1031. https://doi.org/10.1038/nature05414
11. Stojanov, S.; Berlec, A.; Štrukelj, B. The Influence of Probiotics on the Firmicutes/Bacteroidetes Ratio in the Treatment of Obesity and Inflammatory Bowel disease. *Microorganisms* **2020**, *8*. https://doi.org/10.3390/microorganisms8111715
12. Nakayama, J.; Yamamoto, A.; Palermo-Conde, L.A.; Higashi, K.; Sonomoto, K.; Tan, J.; Lee, Y.K. Impact of westernized diet on gut microbiota in children on Leyte island. Front. Microbiol. 2017, 14, 197. https://doi.org/10.3389/fmicb.2017.00197
13. An, J.; Kwon, H.; Kim, Y.J. The Firmicutes/Bacteroidetes Ratio as a Risk Factor of Breast Cancer. *Journal of Clinical Medicine* **2023**, *12*. https:/doi.org/10.3390/jcm12062216
14. Takezawa, K.; Fujita, K.; Matsushita, M.; Motooka, D.; Hatano, K.; Banno, E.; Shimizu, N.; Takao, T.; Takada, S.; Okada, K.; et al. The Firmicutes/Bacteroidetes ratio of the human gut microbiota is associated with prostate enlargement. *The Prostate* **2021**, *81*, 1287-1293. https://doi.org/10.1002/pros.24223
15. Jasirwan, C.O.M.; Muradi, A.; Hasan, I.; Simadibrata, M.; Rinaldi, I. Correlation of gut Firmicutes/Bacteroidetes ratio with fibrosis and steatosis stratified by body mass index in patients with non-alcoholic fatty liver disease. *Biosci Microbiota Food Health* **2021**, *40*, 50-58. https://doi.org/10.12938/bmfh.2020-046
16. Park, S.H.; Kim, K.A.; Ahn, Y.T.; Jeong, J.J.; Huh, C.S., Kim, D.H. Comparative analysis of gut microbiota in elderly people of urbanized towns and longevity villages. *BMC Microbiol.* **2015**, *15*, 1–5. https://doi.org/10.1186/s12866-015-0386-8
17. Park, J.; Kato, K.; Murakami, H.; Hosomi, K.; Tanisawa, K.; Nakagata, T.; Ohno, H.; Konishi, K.; Kawashima, H.; Chen, Y.A.; Mohsen, A.; Xiao, J.Z.; Odamaki, T.; Kunisawa, J.; Mizuguchi, K.; Miyachi, M. Comprehensive analysis of gut microbiota of a healthy population and covariates affecting microbial variation in two large Japanese cohorts. *BMC Microbiol.* **2021**, *2*, 151. https://doi.org/10.1186/s12866-021-02215-0
18. Maruyama, S.; Matsuoka, T.; Hosomi, K.; Park, J.; Nishimura, M.; Murakami, H.; Konishi, K.; Miyachi, M.; Kawashima, H.; Mizuguchi, K.; et al. Characteristic Gut Bacteria in High Barley Consuming Japanese Individuals without Hypertension. *Microorganisms* **2023**, *11*. https://doi.org/10.3390/microorganisms11051246
19. Matsuoka, T.; Hosomi, K.; Park, J.; Goto, Y.; Nishimura, M.; Maruyama, S.; Murakami, H.; Konishi, K.; Miyachi, M.; Kawashima, H.; et al. Relationships between barley consumption and gut microbiome characteristics in a healthy Japanese population: a cross-sectional study. *BMC Nutrition* **2022**, *8*, 23. https://doi.org/10.1186/s40795-022-00500-3
20. Naito, Y.; Takagi, T.; Inoue, R.; Kashiwagi, S.; Mizushima, K.; Tsuchiya, S.; Itoh, Y.; Okuda, K.; Tsujimoto, Y.; Adachi, A.; et al. Gut microbiota differences in elderly subjects between rural city Kyotango and urban city Kyoto: an age-gender-matched study. *Journal of Clinical Biochemistry and Nutrition* **2019**, *65*, 125-131. https://doi.org/10.3164/jcbn.19-26
21. Nishino, K.; Nishida, A.; Inoue, R.; Kawada, Y.; Ohno, M.; Sakai, S.; Inatomi, O.; Bamba, S.; Sugimoto, M.; Kawahara, M.; et al. Analysis of endoscopic brush samples identified mucosa-associated dysbiosis in inflammatory bowel disease. *Journal of Gastroenterology* **2018**, *53*, 95-106. https://doi.org/10.1007/s00535-017-1384-4

22. Takagi, T.; Naito, Y.; Inoue, R.; Kashiwagi, S.; Uchiyama, K.; Mizushima, K.; Tsuchiya, S.; Dohi, O.; Yoshida, N.; Kamada, K.; et al. Differences in gut microbiota associated with age, sex, and stool consistency in healthy Japanese subjects. *Journal of Gastroenterology* **2019**, *54*, 53-63. https://doi.org/10.1007/s00535-018-1488-5
23. Chen, Y.-A.; Park, J.; Natsume-Kitatani, Y.; Kawashima, H.; Mohsen, A.; Hosomi, K.; Tanisawa, K.; Ohno, H.; Konishi, K.; Murakami, H.; et al. MANTA, an integrative database and analysis platform that relates microbiome and phenotypic data. *PLOS ONE* **2020**, *15*, e0243609, https://doi.org/10.1371/journal.pone.0243609