BMC Microbiology

RESEARCH ARTICLE Open Access

Biological observations in microbiota analysis are robust to the choice of 16S rRNA gene sequencing processing algorithm: case study on human milk microbiota

Shirin Moossavi[1,2,3,4,5*], Faisal Atakora[3,6], Kelsey Fehr[3,6] and Ehsan Khafipour[7,8]

Abstract

Background: In recent years, the microbiome field has undergone a shift from clustering-based methods of operational taxonomic unit (OTU) designation based on sequence similarity to denoising algorithms that identify exact amplicon sequence variants (ASVs), and methods to identify contaminating bacterial DNA sequences from low biomass samples have been developed. Although these methods improve accuracy when analyzing mock communities, their impact on real samples and downstream analysis of biological associations is less clear.

Results: Here, we re-processed our recently published milk microbiota data using Qiime1 to identify OTUs, and Qiime2 to identify ASVs, with or without contaminant removal using *decontam*. Qiime2 resolved the mock community more accurately, primarily because Qiime1 failed to detect *Lactobacillus*. Qiime2 also considerably reduced the average number of ASVs detected in human milk samples (364 ± 145 OTUs vs. 170 ± 73 ASVs, $p <$ 0.001). Compared to the richness, the estimated diversity measures had a similar range using both methods albeit statistically different (inverse Simpson index: 14.3 ± 8.5 vs. 15.6 ± 8.7, $p = 0.031$) and there was strong consistency and agreement for the relative abundances of the most abundant bacterial taxa, including *Staphylococcaceae* and *Streptococcaceae*. One notable exception was *Oxalobacteriaceae*, which was overrepresented using Qiime1 regardless of contaminant removal. Downstream statistical analyses were not impacted by the choice of algorithm in terms of the direction, strength, and significance of associations of host factors with bacterial diversity and overall community composition.

Conclusion: Overall, the biological observations and conclusions were robust to the choice of the sequencing processing methods and contaminant removal.

Keywords: Qiime1, Qiime2, Decontam, Reproducibility, Microbiome, Milk microbiota, Human milk, CHILD cohort

* Correspondence: moossavs@myumanitoba.ca; shirin.moossavi@gmail.com
[1]Digestive Oncology Research Center, Digestive Disease Research Institute, Tehran University of Medical Sciences, Tehran, Iran
[2]Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, MB, Canada
Full list of author information is available at the end of the article

Background

Amplicon sequencing targeting bacterial 16S rRNA gene has so far been the most commonly used sequencing method for microbiome studies. In recent years there have been novel developments in several aspects of sequencing processing including a shift from clustering-based methods of operational taxonomic unit (OTU) designation based on sequence similarity (commonly > 97%) [1, 2] to denoising algorithms which identify exact amplicon sequence variants (ASVs) [3–9]; thereby increasing ecological precision. Performance of the denoising methods has been assessed mostly on mock communities [7–12]. However, the impact of these different methods on characterizing real biospecimens and downstream analysis of biological associations is less clear. The detection of true biological and ecological variations appears to be robust to the choice of sequencing processing method (OTU vs. ASV) in a few studies on soil, mouse feces, and human intestinal biopsies [12–14], but head-to-head comparisons are lacking for most human microbiota communities.

Another issue receiving increasing attention in sequencing-based microbiome analysis is contamination introduced during DNA extraction and library preparation. This is especially of concern for low biomass samples where the signal-to-noise ratio is very low [15, 16]. In this case, reproducible downstream analyses plausibly depend on the identification and removal of potential contaminants. Milk is a low biomass sample and thus highly susceptible to reagent contaminants in culture-independent sequencing-based microbiota profiling [17]. To our knowledge, the majority of previously published milk microbiota studies are based on OTU-picking methods and have not assessed the potential reagent contaminants [18]. Therefore, the comparability and generalizability of different studies in terms of the milk microbiota composition and association with maternal and infant characteristics are not known. To address these knowledge gaps, we re-processed our recently published 16S rRNA gene sequencing milk microbiota dataset [19] using Qiime1 closed-reference OTU picking and Qiime2 denoising method with or without contaminant removal using *decontam* [20]. We adhered to the quality control process and taxonomy assignment threshold commonly used by these methods (97% for Qiime1 and 99% for Qiime2) to examine the real world impact of these different approaches on downstream analysis. The datasets resulting from these four approaches (Fig. 1a) were used to assess the comparability of results in terms of microbiota features (taxonomy, alpha, and beta diversity) and test the hypothesis that biological associations are robust to the choice of upstream data processing.

Results

We analyzed 18 replicates (3 per PCR plate in 2 sequencing runs) of a mock community consisting of 8 different bacterial species with a known composition (ZymoBIOMICS™ Microbial Community Standard, Zymo Research, USA). While the Qiime1 method detected an average (SD) of 223 (50) OTUs in the mock community samples, Qiime2 performance was closer to the expected composition, detecting 12 (3) ASVs (Table S1). Although contaminant removal did not considerably reduce the number of OTUs, it decreased the average (SD) ASVs to 9 (3) effectively eliminating the potential contaminants (Table S1). Overall, there was a good agreement between the observed and expected taxonomic composition with both methods (Fig. 1b). However, two notable differences were observed. Most prominently, Qiime2 performed better at identifying *Lactobacillus:* the actual contribution of this genus to the mock community was 19%; the estimated relative abundance was < 1% using Qiime1 compared to 16% using Qiime2. Moreover, the proportion of identified taxa not belonging to the mock community (likely contaminants) was higher with Qiime1 (~ 12% vs. 0.1% in Qiime2). Neither method could accurately identify *Escherichia coli* or *Salmonella enterica* present in the mock community; however, for both methods, the relative abundance of taxa identified as *Enterobacteriaceae* was within the range of the combined relative abundances of these two enteric bacteria. Thus, overall, Qiime2 provided a more accurate representation of the mock community (Table S1 and Fig. 1b) in agreement with previous studies [7–12].

Overall when comparing the four approaches, the mean depth of sequencing per sample was slightly lower in Qiime2 compared to Qiime1 both before and after contaminant removal. The differences in library size within each method before and after *decontam* were negligible (Figure S1). Despite differences in the initial number of OTUs/ASVs in total and on average, Qiime1 and Qiime2 resulted in a similar number of remaining OTUs/ASVs (298 and 299 respectively before contaminant removal) after filtering taxa with less than 0.01% mean relative abundance (Table S1). This suggests the majority of "noisy true reads" (true reads containing sequencing errors [10]) initially retained by Qiime1 were eliminated by applying abundance-based filtering. In agreement with the literature [10, 12], there was a considerable difference in the number of observed OTUs/ASVs prior to filtering very low abundant taxa (364 ± 145 average OTUs vs. 170 ± 73 average ASVs per sample, Table S1). The bacterial richness at OTU/ASV level remained higher in Qiime1 vs. Qiime2 even after data filtering, regardless of contaminant removal (394 ± 91 vs. 148 ± 44, $p < 0.001$) (Fig. 1c). In contrast, milk microbiota

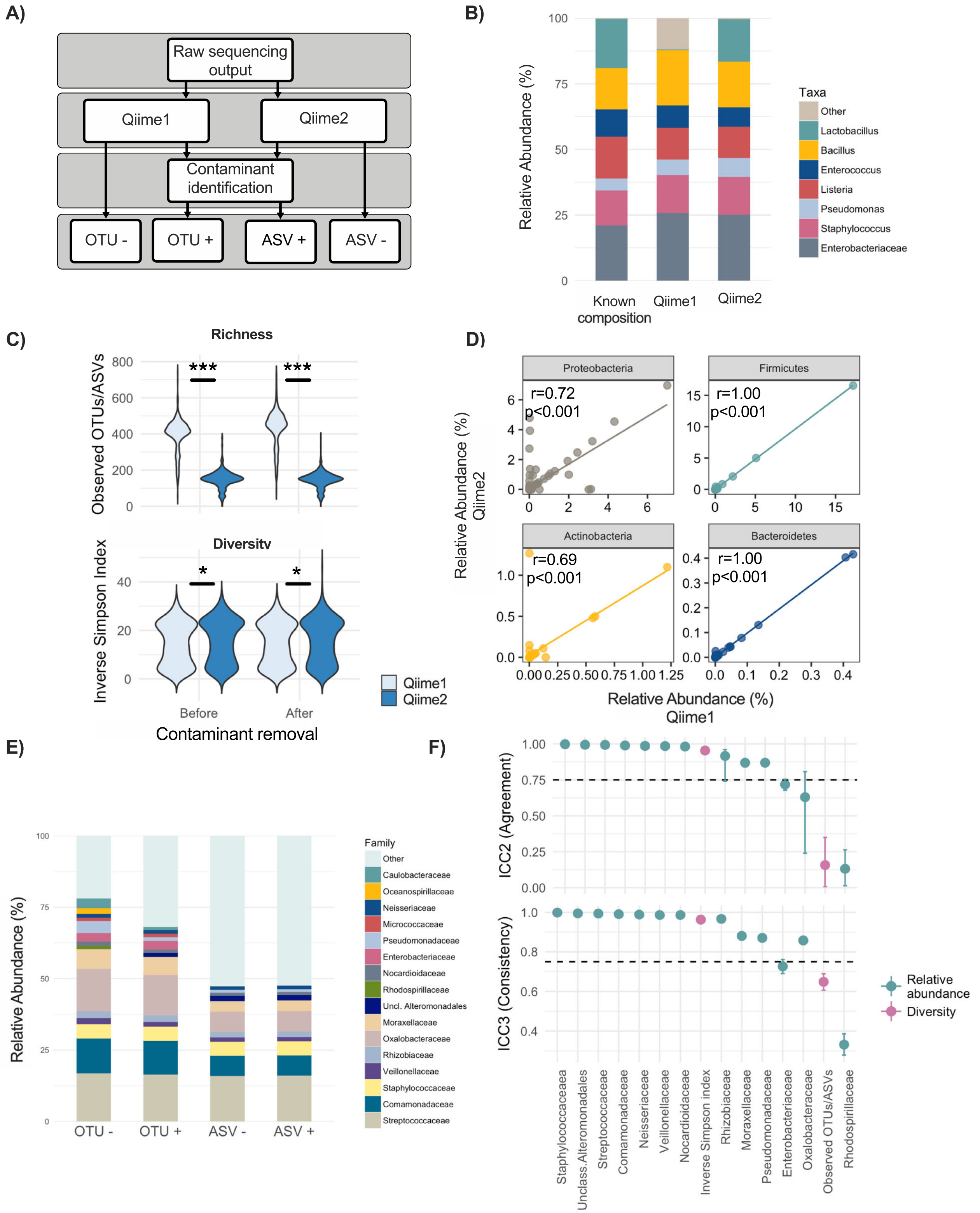


Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Microbiota composition in a mock community and human milk samples using a clustering-based method (Qiime1) and a denoising algorithm (Qiime2) with and without contaminant removal. **a** Schematic of the study design. **b** Composition of the mock community by Qiime1 and Qiime2 prior to contaminant removal (each dataset = combined data from 8 replicates). **c** Comparison of milk microbiota richness (observed OTUs/ASVs) and diversity (inverse Simpson index) between Qiime1 and Qiime2 with and without contaminant removal. **d** Correlation of the relative abundances of milk genera between Qiime1 and Qiime2 prior to contaminant removal (See also Figures. S2 and S3 and Tables S2 and S3) (each dataset = combined data from 393 milk samples). Each dot represents a classified genus. Contaminant removal doesn't impact the associations (not shown). **e** Comparison of the composition of abundant families (> 1% mean relative abundance) between Qiime1 and Qiime2 with or without contaminant removal. Contaminant removal reduced the relative abundance of certain low-abundance taxa (e.g. *Caulobacteraceae* and *Rhodospirillaceae*) and proportion of Other taxa (OTUs with less than 1% mean relative abundance) estimated by Qiime1, but generally did not affect the microbiota profile estimated by Qiime2. **f** Agreement and consistency between methods by intraclass correlation for alpha diversity and 13 most abundant families. * $p < 0.05$, *** $p < 0.001$

diversity was slightly but significantly higher with Qiime2 regardless of contaminant removal (14.3 ± 8.5 vs. 15.6 ± 8.7 $p < 0.05$) (Fig. 1c) suggesting that the number of the abundant taxa remained consistent in both Qiime1 and Qiime2 methods.

Next, we compared the relative abundance (Fig. 1d) and prevalence (Figure S2) of genera belonging to the major milk phyla (Figure S3 and Table S2) between methods and observed high degrees of correlation, especially for Firmicutes, Actinobacteria, and Bacteroidetes (Table S3). In comparing abundant genera (> 0.01% mean relative abundance) (Figure S3 and Table S4), relative abundances remained highly correlated (Figure S4). Overall, the relative abundances of the most abundant families (> 1% mean relative abundance) including *Streptococcaceae* and *Staphylococcaceae* were not considerably impacted by the choice of the sequencing processing method, with and without contaminant removal (Fig. 1e and Figure S5). However, there were notable differences between methods for some other taxa. For example, at the family level, *Oxalobacteriaceae* was detected at higher relative abundances by Qiime1 compared to Qiime2 (14.1% ± 8.4% vs. 7.2% ± 4.1%, Figure S5), while *Enterobacteriaceae* and *Caulobacteraceae* were only observed as top abundant families using Qiime1 (Fig. 1e). In addition, *Oxalobacteriaceae* and *Comamonadaceae* were not assigned taxonomy at genus level using Qiime1, whereas some members of these families including *Acidovorax* (family *Comamonadaceae*), *Ralstonia*, and *Massilia* (*Oxalobacteriaceae*) were resolved by Qiime2. In contrast, *Methylibium* (family *Comamonadaceae*) and *Erwinia* (family *Enterobacteriaceae*) were only identified by Qiime1 as abundant taxa. Overall, the total proportion of less abundant OTUs/ASVs (< 1% mean relative abundance) was higher in Qiime2 (Fig. 1e) while the number of true abundant taxa was less biased, possibly due to the lack of binning of multiple similar sequence variants into an OTU. In agreement with this interpretation, contaminant removal considerably increased the proportion of less abundant taxa only in Qiime1 suggesting that some of the abundant taxa were consistent of contaminants (Table S1).

Overall, agreement and consistency between different methods (Fig. 1f) were quite low for milk microbiota richness, highlighting the sensitivity of this measure to the choice of bioinformatics method. Nevertheless, a very high inter-class correlation for inverse Simpson index (0.95) and relative abundances for the majority of the abundant families (above 0.75 for 10/13 families) suggests an acceptable degree of agreement and consistency, which would be required for downstream analyses to generate comparable results.

Given the differences in microbiota alpha diversity and taxonomic structure resulting from the choice of processing approaches (Fig. 1), we explored whether the processing approach also influenced the association of microbiota and metadata variables. To do this, we 1) assessed the association of mode of breastfeeding with milk microbiota beta diversity and 2) compared the association of maternal, infant, and early life factors, breastfeeding practices, and other milk components with milk microbiota richness, diversity, and overall composition as previously described [19]. We have previously identified mode of breastfeeding to be significantly associated with milk microbiota beta diversity [19]. Here, we observed similar beta diversity association patterns with mode of breastfeeding regardless of the sequencing processing method or contaminant removal (Fig. 2). Overall, there were high degrees of concordance in the direction, strength, and significance of association between milk microbiota diversity and overall composition with the independent variables assessed using both methods, regardless of contaminant removal (Fig. 3). However, some method-related differences were observed for associations with microbiota richness. While the direction and strength of associations using Qiime2 without contaminant removal were comparable to the Qiime1, contaminant removal generally resulted in weaker associations and lower effect size estimates when using Qiime2-processed data (e.g. for maternal atopy, infant sex, and mode of breastfeeding). Occasionally, the direction of association was also different in Qiime2/decontaminated compared to the other processing methods (e.g. for prenatal smoking and fatty acids profile) (Fig. 3).

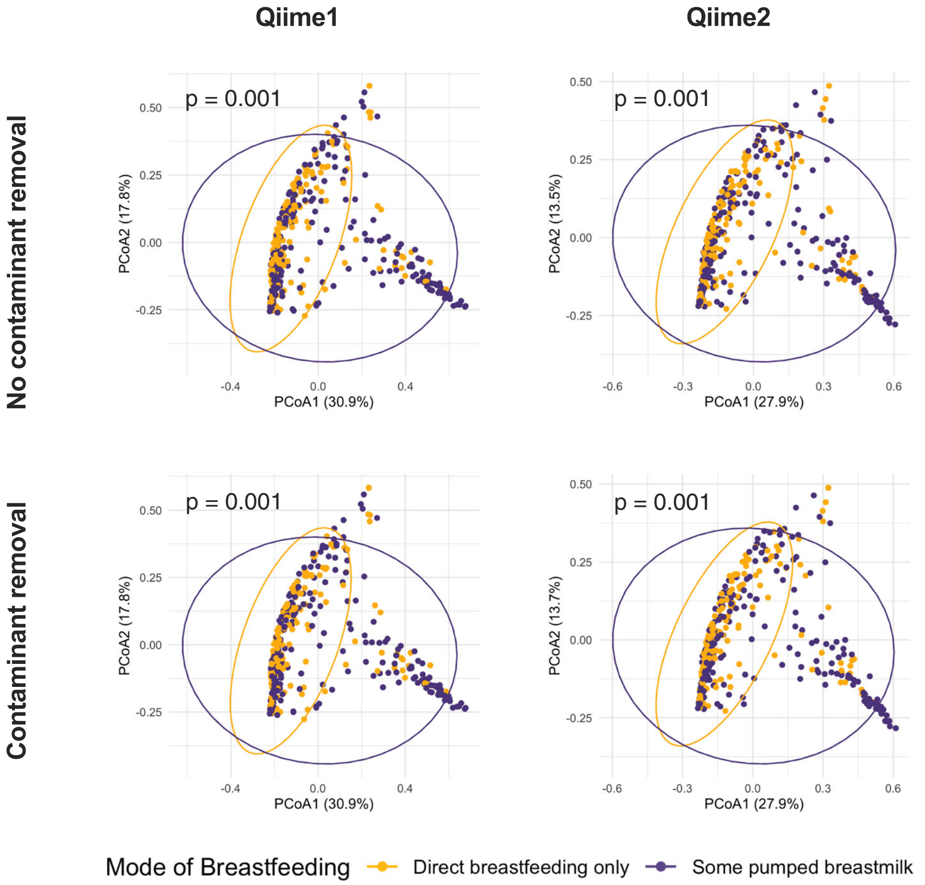
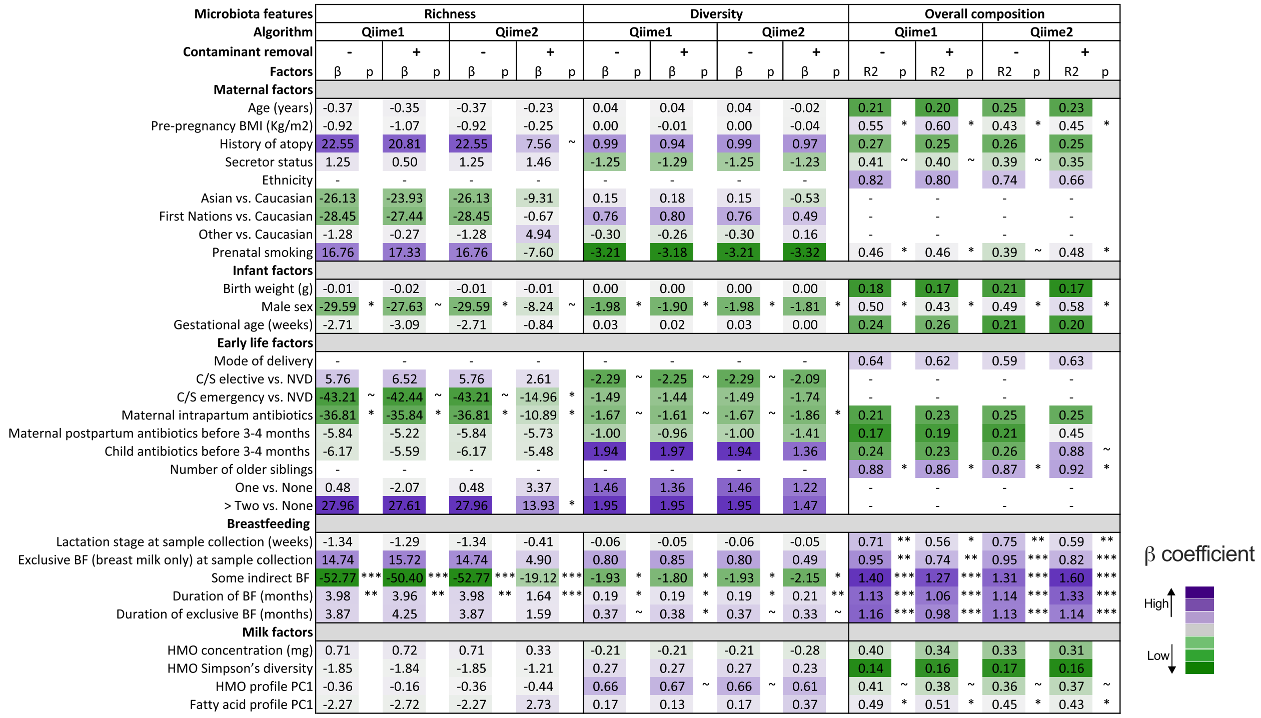


Fig. 2 Impact of four sequence processing approaches on the association of mode of breastfeeding with milk microbiota beta diversity. We reprocessed our published 16S rRNA gene sequencing milk microbiota dataset [19] using Qiime1 and Qiime2 with or without contaminant removal resulting in four datasets (see also Fig. 1a). The Association of mode of breastfeeding with milk microbiota beta diversity was assessed on Bray-Curtis dissimilarity matrix and was tested by permutational ANOVA (PERMANOVA)

Discussion

We compared the milk microbiota composition and associations with maternal, infant, early life, and milk factors on data processed by a clustering-based method into OTUs vs. a denoising method resulting in ASVs. Additionally, we compared the impact of contaminant removal on the statistical conclusions. Overall, richness was strongly impacted by the choice of the bioinformatics approach while statistical contaminant removal had minimal additional impact. There was an acceptable agreement and consistency in the relative abundances of the dominant milk bacteria and milk diversity. Additionally, the main conclusions remained robust to the choice of data processing.

Sequencing-based microbiome studies are highly influenced by the various bioinformatics and data processing choices [21, 22]; specifically, with the recent shift from clustering-based OTU picking methods to denoising algorithms identifying ASVs with higher ecological accuracy. Generally, it is not clear how the results of previously published milk microbiota studies using OTU-picking methods would compare to the more recent results including ours using ASV methods. In a head-to-head comparison of an OTU-picking method with a denoising algorithm, we observed high degrees of agreement and consistency in milk microbial features regardless of the methods. Our results suggest that although milk microbiota richness and some members of the microbial community are strongly influenced by the choice of sequencing processing method, there is high agreement and consistency between methods for estimating microbiota diversity and quantifying the majority of the most abundant taxa. Addition of a contaminant removal step resulted in minor shifts in the composition of the most abundant taxa as well as association of richness with a few of the variables assessed when using Qiime2-processed data. The overall conclusions of the study and the main determinants of milk microbiota composition (e.g. consistent association of feeding mode with milk microbiota composition [19]) remained robust to the choice of OTU vs. ASV methods with or without contaminant removal.



Microbiota features	Richness								Diversity								Overall composition							
Algorithm	**Qiime1**				**Qiime2**				**Qiime1**				**Qiime2**				**Qiime1**				**Qiime2**			
Contaminant removal	**-**		**+**		**-**		**+**		**-**		**+**		**-**		**+**		**-**		**+**		**-**		**+**	
Factors	β	p	β	p	β	p	β	p	β	p	β	p	β	p	β	p	R2	p	R2	p	R2	p	R2	p
Maternal factors																								
Age (years)	-0.37		-0.35		-0.37		-0.23		0.04		0.04		0.04		-0.02		0.21		0.20		0.25		0.23	
Pre-pregnancy BMI (Kg/m2)	-0.92		-1.07		-0.92		-0.25		0.00		-0.01		0.00		-0.04		0.55	*	0.60	*	0.43	*	0.45	*
History of atopy	22.55		20.81		22.55		7.56	~	0.99		0.94		0.99		0.97		0.27		0.25		0.26		0.25	
Secretor status	1.25		0.50		1.25		1.46		-1.25		-1.29		-1.25		-1.23		0.41	~	0.40	~	0.39	~	0.35	
Ethnicity	-		-		-		-		-		-		-		-		0.82		0.80		0.74		0.66	
Asian vs. Caucasian	-26.13		-23.93		-26.13		-9.31		0.15		0.18		0.15		-0.53		-		-		-		-	
First Nations vs. Caucasian	-28.45		-27.44		-28.45		-0.67		0.76		0.80		0.76		0.49		-		-		-		-	
Other vs. Caucasian	-1.28		-0.27		-1.28		4.94		-0.30		-0.26		-0.30		0.16		-		-		-		-	
Prenatal smoking	16.76		17.33		16.76		-7.60		-3.21		-3.18		-3.21		-3.32		0.46	*	0.46	*	0.39	~	0.48	*
Infant factors																								
Birth weight (g)	-0.01		-0.02		-0.01		-0.01		0.00		0.00		0.00		0.00		0.18		0.17		0.21		0.17	
Male sex	-29.59	*	-27.63	~	-29.59	*	-8.24	~	-1.98	*	-1.90	*	-1.98	*	-1.81	*	0.50	*	0.43	*	0.49	*	0.58	*
Gestational age (weeks)	-2.71		-3.09		-2.71		-0.84		0.03		0.02		0.03		0.00		0.24		0.26		0.21		0.20	
Early life factors																								
Mode of delivery	-		-		-		-		-		-		-		-		0.64		0.62		0.59		0.63	
C/S elective vs. NVD	5.76		6.52		5.76		2.61		-2.29	~	-2.25	~	-2.29	~	-2.09		-		-		-		-	
C/S emergency vs. NVD	-43.21	~	-42.44	~	-43.21	~	-14.96	*	-1.49		-1.44		-1.49		-1.74		-		-		-		-	
Maternal intrapartum antibiotics	-36.81	*	-35.84	*	-36.81	*	-10.89	*	-1.67	~	-1.61	~	-1.67	~	-1.86	*	0.21		0.23		0.25		0.25	
Maternal postpartum antibiotics before 3-4 months	-5.84		-5.22		-5.84		-5.73		-1.00		-0.96		-1.00		-1.41		0.17		0.19		0.21		0.45	
Child antibiotics before 3-4 months	-6.17		-5.59		-6.17		-5.48		1.94		1.97		1.94		1.36		0.24		0.23		0.26		0.88	~
Number of older siblings	-		-		-		-		-		-		-		-		0.88	*	0.86	*	0.87	*	0.92	*
One vs. None	0.48		-2.07		0.48		3.37		1.46		1.36		1.46		1.22		-		-		-		-	
> Two vs. None	27.96		27.61		27.96		13.93	*	1.95		1.95		1.95		1.47		-		-		-		-	
Breastfeeding																								
Lactation stage at sample collection (weeks)	-1.34		-1.29		-1.34		-0.41		-0.06		-0.05		-0.06		-0.05		0.71	**	0.56	*	0.75	**	0.59	**
Exclusive BF (breast milk only) at sample collection	14.74		15.72		14.74		4.90		0.80		0.85		0.80		0.49		0.95	**	0.74	**	0.95	***	0.82	***
Some indirect BF	-52.77	***	-50.40	***	-52.77	***	-19.12	***	-1.93	*	-1.80	*	-1.93	*	-2.15	*	1.40	***	1.27	***	1.31	***	1.60	***
Duration of BF (months)	3.98	**	3.96	**	3.98	**	1.64	***	0.19	*	0.19	*	0.19	*	0.21	**	1.13	***	1.06	***	1.14	***	1.33	***
Duration of exclusive BF (months)	3.87		4.25		3.87		1.59		0.37	~	0.38	*	0.37	~	0.33	~	1.16	***	0.98	***	1.13	***	1.14	***
Milk factors																								
HMO concentration (mg)	0.71		0.72		0.71		0.33		-0.21		-0.21		-0.21		-0.28		0.40		0.34		0.33		0.31	
HMO Simpson's diversity	-1.85		-1.84		-1.85		-1.21		0.27		0.27		0.27		0.23		0.14		0.16		0.17		0.16	
HMO profile PC1	-0.36		-0.16		-0.36		-0.44		0.66		0.67	~	0.66	~	0.61		0.41	~	0.38	~	0.36	~	0.37	~
Fatty acid profile PC1	-2.27		-2.72		-2.27		2.73		0.17		0.13		0.17		0.37		0.49	*	0.51	*	0.45	*	0.43	*

Fig. 3 Impact of four sequence processing approaches on observed associations of milk microbiota richness (observed OTUs/ASVs), diversity (inverse Simpson index), and overall composition with maternal, infant, early life, breastfeeding, and milk factors. We re-processed our published 16S rRNA gene sequencing milk microbiota dataset [19] using Qiime1 and Qiime2 with or without contaminant removal resulting in four datasets (see also Fig. 1a). Beta coefficients from univariate linear regression (richness and diversity) or R^2 from redundancy analysis (overall composition) are presented and colour coded within each microbiota feature. Results of Qiime2 with contaminant removal are originally reported in Moossavi et al. [19]. BF, breastfeeding; BMI, body mass index; C/S, Cesarean section; HMO, human milk oligosaccharide; NVD, normal vaginal delivery; PC1, Principal Component 1 * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

While the overall agreement between different data processing approaches was high, some differences stood out. For example, the relative abundance of *Oxalobacteriaceae*, an environmental bacteria and a common reagent contaminant, was lower using denoising methods, potentially in line with higher classification accuracy of the denoising approaches. Notably, there was not a difference in the relative abundance of *Oxalobacteriaceae* before and after identification and removal of potential reagent contaminants using *decontam*, potentially highlighting the limitation of these methods for low biomass samples [23]. Computational methods, while helpful, do not replace the need for careful study design, sample handling, and reagent controls. Additionally, culture-based methods such as culture-enriched molecular profiling may inform the sequencing results of low biomass samples [24].

While new bioinformatic methods are typically evaluated using mock communities, we have provided a real-word comparison of two commonly used sequencing processing approaches (Qiime1 vs. Qiime2). A limitation of our study is that both approaches are sequencing-dependent and prone to reagent contaminants, and therefore neither can verifiably identify "true" milk taxa. Thus, while we have provided evidence that the statistical results obtained by the two approaches are comparable, it is important that the microbiological and ecological implications be studied using controlled experimental designs.

Conclusion

In summary, we have shown that Qiime2 resolved the mock community more accurately and there were high degrees of agreement and consistency in milk microbiota features regardless of the choice of the sequencing processing approach (OTU vs. ASV). In light of our observation that the associations with metadata and the main conclusions were robust to the choice of sequence processing approaches and contamination removal, previous studies of milk microbiota and potentially other low biomass samples using OTU picking approaches are likely valid both in terms of the composition of the abundant taxa and associations, especially for metrics that put less emphasis on richness.

Methods

Study design

We used our published data on milk microbiota [19] (SRA accession number: PRJNA481046) in the CHILD cohort [25]. Each mother provided one sample of milk at 3–4 months postpartum [mean (SD) 17 (5) weeks postpartum] in a sterile milk container provided by the CHILD study. Milk microbiota was profiled by sequencing the V4 hypervariable region of 16S rRNA gene on a MiSeq platform (Illumina, San Diego, CA, USA) as previously described [19].

Microbiome sequencing processing

Overlapping paired-end reads were separately processed with a clustering-based (Qiime1) and a denoising algorithm (Qiime2). In the clustering-based approach, paired-end reads were merged using the PANDAseq assembler [26]. Sequences with low quality base calling scores (< 20) as well as those containing ambiguous bases in the overlapping region were discarded. The subsequent fastq file was processed using the open-source software Qiime v1.9.1 [27]. Assembled reads were demultiplexed according to the barcode sequences and chimeric reads were filtered using UCHIME [28]. Reads were clustered into OTUs using closed-reference OTU picking based on 97% similarity using UCLUST [29]. Representative sequences from each OTU were assigned a taxonomy using RDP Classifier [30] and aligned to the 2013 release of the Greengenes reference database at 97% sequence similarity [31] using PyNAST [32]. In the denoising approach, overlapping paired-end reads were processed with DADA2 pipeline [7] using the open-source software Qiime 2 v.2018.6 (https://Qiime2.org) [27]. Unique ASVs were assigned taxonomy and aligned to the 2013 release of the Greengenes reference database at 99% sequence similarity [31].

OTU/ASV table pre-processing and filtering

Initial pre-processing of the OTU/ASV table was conducted using the Phyloseq package [33]. As previously reported [19], the mean (SD) sequencing depth was 47, 710 (18,643). Samples with less than 25,000 sequencing reads were excluded ($n = 35$) and the remaining samples ($n = 393$) were rarefied to the minimum 25,000 sequencing reads per sample. OTUs/ASVs only present in the mock community or negative controls and OTUs/ASVs belonging to phylum Cyanobacteria, family of mitochondria, and class of chloroplast were removed. OTUs/ASVs with less than 20 reads across the entire dataset ($n = 393$ samples) were also removed. The numbers of sequencing reads of taxa were then relativized to the total sum of 25,000. This dataset was used for analysis unless otherwise specified.

Reagent contaminant removal

Potential reagent contaminants were identified using *decontam* package based on either the frequency of the OTUs/ASVs in the negative control or the negative correlation with DNA concentration using default parameters [20].

Performance assessment on mock community

The baseline performance of each method was assessed on DNA extracted from a high biomass mock community consisting of 8 bacterial species with known relative abundances (ZymoBIOMICS™ Microbial Community Standard, Zymo Research, USA).

Statistical analysis

Depth of sequencing and alpha diversity (observed OTUs/ASVs and inverse Simpson index) were compared between methods ($n = 4$ datasets) using Student's t test. Within the 4 most abundant phyla, the prevalence (percentage of samples containing the taxa) and average relative abundance of classified genera were compared between Qiime1 and Qiime2 prior to contaminant removal using Pearson correlation. Agreement and consistency of community alpha diversity and relative abundances of the most abundant families were assessed by interclass correlation by 2-way random and fixed single measurement models using Psych package [34, 35]. Association of mode of breastfeeding with milk microbiota beta diversity was assessed on Bray-Curtis dissimilarity matrix and was tested by permutational ANOVA (PERMANOVA) using the vegan package [36]. Separately for each method (within the $n = 393$ milk samples), the association of maternal, infant, early life, breastfeeding, and milk factors was assessed by linear regression (for microbiota alpha diversity) and redundancy analysis (RDA, for microbiota composition). RDA was performed with 1000 permutations using the vegan package [36] following zero-replacement and centre log-ratio transformation [37, 38].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12866-020-01949-7.

Additional file 1: Table S1. Comparison of the number of OTUs/ASVs in the mock community, negative controls, and milk microbiota datasets processed by clustering-based OTU method (Qiime1) and a denoising algorithm (Qiime2) with or without contaminant removal. **Table S3.** Comparison of prevalence and relative abundance of shared genera in milk microbiota processed by clustering-based OTU method (Qiime1) and a denoising algorithm (Qiime2) without contaminant removal. **Figure S1.** Comparison of the library size on datasets processed by clustering-based OTU method (Qiime1) and a denoising algorithm (Qiime2) with or without contaminant removal. **Figure S2.** Comparison of the prevalence of bacterial genera on datasets processed by clustering-based OTU method (Qiime1) and a denoising algorithm (Qiime2) without contaminant removal. **Figure S3.** Distribution of classified genera in Qiime1 and Qiime2 processed datasets without contaminant removal. **Figure S4.**

Comparison of the relative abundance and prevalence of abundant bacterial genera (> 0.01% mean relative abundance) on datasets processed by clustering-based OTU method (Qiime1) and a denoising algorithm (Qiime2) without contaminant removal. **Figure S5.** Comparison of the composition of abundant taxa (> 1% mean relative abundance) on datasets processed by clustering-based OTU method (Qiime1) and a denoising algorithm (Qiime2) with or without contaminant removal.

Additional file 2: Table S2. Prevalence of classified genera in datasets processed by Qiime1 and Qiime2 after excluding OTUs/ASVs with less than 20 reads across each dataset. **Table S4.** Prevalence of classified genera in datasets processed by Qiime1 and Qiime2 after excluding OTUs/ASVs with less than 20 reads across each dataset and mean relative abundance of less than 0.01%.

Abbreviations

ASV: Amplicon sequence variants; OTU: Operational taxonomic unit; PCR: Polymerase chain reaction; RDA: Redundancy analysis; SD: Standard deviation

Acknowledgements

We are grateful to all the families who took part in this study, and the whole CHILD team, which includes interviewers, nurses, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, and receptionists. We thank Padmaja Subbarao (University of Toronto) Director of the CHILD study; CHILD site directors: Meghan B. Azad (University of Manitoba), Stuart E. Turvey (University of British Columbia), Theo J. Moraes (University of Toronto), Allan B. Becker (University of Manitoba), Piushkumar J. Mandhane (University of Alberta), Malcolm R. Sears (McMaster University); Diana Lefebvre (McMaster University) for managing the CHILD database; Leah Stiemsma for creating the antibiotic variables and Lorena Vehling for creating the breastfeeding variables.

Authors' contributions

SM conceptualized the study. SM, FA, KF performed statistical analysis. SM wrote the original draft. FA, KF, EK revised and edited the manuscript. All authors read and approved the final manuscript.

Funding

No funding was obtained specifically for this method comparison project. The Canadian Institutes of Health Research (CIHR) and the Allergy, Genes and Environment Network of Centres of Excellence (AllerGen NCE) provided core support for the CHILD Study. Milk microbiota sequencing was funded by the Heart and Stroke Foundation and Canadian Lung Association, in partnership with the Canadian Respiratory Research Network and AllerGen NCE (Meghan B. Azad). Infrastructure at the Gut Microbiome Laboratory was supported by grants from the Canadian Foundation for Innovation. SM holds a Research Manitoba Doctoral Studentship. These entities had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

Availability of data and materials

The datasets generated and analysed during the current study are available in the Sequence Read Archive of NCBI repository (accession number PRJNA481046).

Ethics approval and consent to participate

Written informed consent was obtained from the participants. The original study was approved by the Human Research Ethics Boards at McMaster University, the Hospital for Sick Children, and the Universities of Manitoba, Alberta, and British Columbia.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Digestive Oncology Research Center, Digestive Disease Research Institute, Tehran University of Medical Sciences, Tehran, Iran. [2]Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, MB, Canada. [3]Children's Hospital Research Institute of Manitoba, Winnipeg, MB, Canada. [4]Developmental Origins of Chronic Diseases in Children Network (DEVOTION), Winnipeg, MB, Canada. [5]Present Address Department of Physiology and Pharmacology & Mechanical and Manufacturing Engineering, University of Calgary, Calgary, AB, Canada. [6]Department of Pediatrics and Child Health, University of Manitoba, Winnipeg, MB, Canada. [7]Department of Animal Science, University of Manitoba, Winnipeg, MB, Canada. [8]Present Address Microbiome Research and Technical Support, Cargill Animal Nutrition, Diamond V brand, Cedar Rapids, USA.

Received: 23 January 2020 Accepted: 18 August 2020
Published online: 18 September 2020

References

1. Stackebrandt E, Goebel BM. A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol. 1994;44(4):846–9.
2. Kopylova E, Navas-Molina JA, Mercier C, Xu ZZ, Mahe F, He Y, et al. Open-source sequence clustering methods improve the state of the art. mSystems. 2016;1(1):e00003–e00015. https://doi.org/10.1128/mSystems.00003-15.
3. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11(12):2639–43. https://doi.org/10.1038/ismej.2017.119.
4. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, et al. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol. 2013;4(12). https://doi.org/10.1111/2041-210X.12114.
5. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J. 2015;9(4):968–79. https://doi.org/10.1038/ismej.2014.195.
6. Tikhonov M, Leach RW, Wingreen NS. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. ISME J. 2015;9(1):68–80. https://doi.org/10.1038/ismej.2014.117.
7. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3. https://doi.org/10.1038/nmeth.3869.
8. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems. 2017;2(2):e00191–e00116. https://doi.org/10.1128/mSystems.00191-16 .
9. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. 2016:081257. https://doi.org/10.1101/081257.
10. Caruso V, Song X, Asquith M, Karstens L, Gibbons SM. Performance of microbiome sequence inference methods in environments with varying biomass. mSystems. 2019;4(1):e00163–e00118. https://doi.org/10.1128/mSystems.00163-18 .
11. Chen W, Zhang CK, Cheng Y, Zhang S, Zhao H. A comparison of methods for clustering 16S rRNA sequences into OTUs. PLoS One. 2013;8(8):e70837. https://doi.org/10.1371/journal.pone.0070837.
12. Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ. 2018;6:e5364. https://doi.org/10.7717/peerj.5364.
13. Allali I, Arnold JW, Roach J, Cadenas MB, Butz N, Hassan HM, et al. A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. BMC Microbiol. 2017;17(1):194. https://doi.org/10.1186/s12866-017-1101-8.
14. Glassman SI, Martiny JBH. Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. mSphere. 2018;3(4):e00148–e00118. https://doi.org/10.1128/mSphere.
15. Karstens L, Asquith M, Caruso V, Rosenbaum JT, Fair DA, Braun J, et al. Community profiling of the urinary microbiota: considerations for low-biomass samples. Nat Rev Urol. 2018;15(12):735–49. https://doi.org/10.1038/s41585-018-0104-z.
16. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in low microbial biomass microbiome studies: issues and recommendations. Trends Microbiol. 2019;27(2):105–17. https://doi.org/10.1016/j.tim.2018.11.003.

17. Dahlberg J, Sun L, Persson Waller K, Ostensson K, McGuire M, Agenas S, et al. Microbiota data from low biomass milk samples is markedly affected by laboratory and reagent contamination. PLoS One. 2019;14(6):e0218257. https://doi.org/10.1371/journal.pone.0218257.
18. Sakwinska O, Bosco N. Host microbe interactions in the lactating mammary gland. Front Microbiol. 2019;10:1863. https://doi.org/10.3389/fmicb.2019.01863.
19. Moossavi S, Sepehri S, Robertson B, Bode L, Goruk S, Field CJ, et al. Composition and variation of the human milk microbiome is influenced by maternal and early life factors. Cell Host Microbe. 2019;25:324–35.
20. Davis NM, Proctor D, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. 2018;6(1):226. https://doi.org/10.1186/s40168-018-0605-2.
21. Walker AW, Martin JC, Scott P, Parkhill J, Flint HJ, Scott KP. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. Microbiome. 2015;3:26. https://doi.org/10.1186/s40168-015-0087-4.
22. Siegwald L, Caboche S, Even G, Viscogliosi E, Audebert C, Chabe M. The impact of bioinformatics pipelines on microbiota studies: Does the analytical "microscope" affect the biological interpretation? Microorganisms. 2019;7(10). https://doi.org/10.3390/microorganisms7100393.
23. Moossavi S, Fehr K, Moraes TJ, Khafipour E, Azad MB. Repeatability and reproducibility assessment in a large-scale population-based microbiota study: case study on human milk microbiota. bioRxiv. 2020. p. 052035. https://doi.org/10.1101/2020.04.20.
24. de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, et al. Recognizing the reagent microbiome. Nat Microbiol. 2018;3(8): 851–3. https://doi.org/10.1038/s41564-018-0202-y.
25. Subbarao P, Anand SS, Becker AB, Befus AD, Brauer M, Brook JR, et al. The Canadian healthy infant longitudinal development (CHILD) study: examining developmental origins of allergy and asthma. Thorax. 2015;70(10):998–1000. https://doi.org/10.1136/thoraxjnl-2015-207246.
26. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics. 2012;13: 31. https://doi.org/10.1186/1471-2105-13-31.
27. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6. https://doi.org/10.1038/nmeth.f.303.
28. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011;27(16): 2194–200. https://doi.org/10.1093/bioinformatics/btr381.
29. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1. https://doi.org/10.1093/bioinformatics/btq461.
30. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7. https://doi.org/10.1128/AEM.00062-07.
31. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069–72. https://doi.org/10.1128/AEM.03006-05.
32. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics. 2010;26(2):266–7. https://doi.org/10.1093/bioinformatics/btp636.
33. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 2013;8(4): e61217. https://doi.org/10.1371/journal.pone.0061217.
34. Koo TK, Li MY. A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012.
35. Revelle W. psych: Procedures for Personality and Psychological Research. Evanston: Northwestern University; 2018. https://CRAN.R-project.org/package=psych Version =1.8.12.
36. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. Vegan: Community Ecology Package. R package version 2.4–3. 2017.
37. Palarea-Albaladejo J, Martin-Fernandez JA. zCompositions -- R package for multivariate imputation of left-censored data under a compositional approach. Chemom Intell Lab Syst. 2015;143:85–96.
38. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can J Microbiol. 2016;62(8): 692–703. https://doi.org/10.1139/cjm-2015-0821.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.