Computational and Structural Biotechnology Journal 19 (2021) 5126–5139

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Recovery of microbial community profile information hidden in chimeric sequence reads

Mengfei Ho [a,*], Damee Moon [a,1], Melissa Pires-Alves [a], Patrick D. Thornton [b], Barbara L. McFarlin [b], Brenda A. Wilson [a,*]

[a] *Department of Microbiology, School of Molecular and Cellular Biology, University of Illinois at Urbana-Champaign, United States*
[b] *Department of Human Development Nursing Science, College of Nursing, University of Illinois at Chicago, United States*

ARTICLE INFO

Article history:
Received 25 April 2021
Received in revised form 30 August 2021
Accepted 30 August 2021
Available online 3 September 2021

Keywords:
next-generation sequencing
chimeras
microbial community profiles
Shannon diversity
16S rRNA
metagenomic sequencing
Blastn
QIIME2
Morisita-Horn similarity
vaginal microbiome

ABSTRACT

The next frontier in the field of microbiome studies is identification of all microbes present in the microbiome and accurate determination of their abundance such that microbiome profiles can serve as reliable assessments of health or disease status. PCR-based 16S rRNA gene sequencing and metagenome shotgun sequencing technologies are the prevailing approaches used in microbiome analyses. Each poses a number of technical challenges associated with PCR amplification, sample availability, and cost of processing and analysis. In general, results from these two approaches rarely agree completely with each other. Here, we compare these methods utilizing a set of vaginal swab and lavage specimens from a cohort of 42 pregnant women collected for a pilot study exploring the effect of the vaginal microbiome on preterm birth. We generated the microbial community profiles from the sequencing reads of the V3V4 and V4V5 regions of the 16S rRNA gene in the vaginal swab and lavage samples. For a subset of the vaginal samples from 12 subjects, we also performed metagenomic shotgun sequencing analysis and compared the results obtained from the PCR-based sequencing methods. Our findings suggest that sample composition and complexity, particularly at the species level, are major factors that must be considered when analyzing and interpreting microbiome data. Our approach to sequence analysis includes consideration of chimeric reads, by using our chimera-counting BlastBin program, and enables recovery of microbial content information generated during PCR-based sequencing methods, such that the microbial profiles more closely resemble those obtained from metagenomic read-based approaches.

1. Introduction

It is now well established that microbiomes are strongly connected to general health and microbial-related diseases of the host. The current paradigm in human microbiome studies posits that a shift in the microbial profile is an indication of a change in health status and disease progression [1]. For instance, shifting of the human vaginal microbiota profile from a *Lactobacillus*-dominant community to a diverse non-*Lactobacillus* community is characteristic of bacterial vaginosis and is associated with increased risk of a number of adverse reproductive health outcomes [2–5]. The value of microbiome composition in intestinal health is illustrated by the application of fecal microbiota transplantation, which has shown high efficacy in treating relapsing *Clostridium difficile* infections [6,7] and has become a promising therapy for resetting microbiomes in other intestinal bowel disorders [8–10]. The importance of microbiome composition in immune regulation and tumor progression is also well established [11,12], and oncologists are now exploring microbiome modulation as a means for improving immunotherapy outcomes [13,14]. As demonstrated by the case of PD-1 inhibitor efficacy in cancer therapy in at least three different clinical studies, the microbiome composition can significantly impact the clinical response to chemotherapy treatment [15–17].

The next challenge in microbiome research is to advance beyond simple observation of differences in microbial profile composition to the precise identification of the causal microbial biomarkers [18,19]. A case in point is the finding that the microbiome profiles and signature bacteria identified as responsible for the associated PD-1 response in each of the above-mentioned

[*] Corresponding authors.
E-mail addresses: mho1@illinois.edu (M. Ho), damee.moon@nyulangone.org (D. Moon), melalves@illinois.edu (M. Pires-Alves), pthorn3@uic.edu (P.D. Thornton), bmcfar1@uic.edu (B.L. McFarlin), wilson7@illinois.edu (B.A. Wilson).
[1] Current Address: New York University, Grossman School of Medicine, New York City, United States.

https://doi.org/10.1016/j.csbj.2021.08.050
2001-0370/Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

clinical trial studies [15–17] were different (*Akkermansia muciniphila*, *Faecalibacterium prausnitzii*, and *Bifidobacterium longum*, respectively) [20]. The general perception is that these divergent findings are attributed to variation in experimental parameters, including the study setup, sample collection and handling, genomic DNA preparation, sequencing technology platform and analytical pipeline, and as yet undetermined factors. Although differences between sequencing analysis pipelines was dismissed as a confounding factor in the above PD-1 studies [20], current bioinformatics pipelines for analyzing sequencing data, as a whole, have not reached a point where they can provide a comprehensive and precise picture of the microbiome that can reliably pinpoint a microbial causal agent [18].

Sequence analysis based on the 16S rRNA gene is a powerful tool commonly applied in identifying the composition of the microbiome [1,21]. However, because it is a PCR-based method, it is associated with a number of unavoidable error-generating artifacts, such as those resulting from choice of PCR amplification conditions or 16S rRNA variable region used, as well as the chimeric read phenomenon [21–24]. Direct microbiome metagenome sequence analysis has been posited as a superior method for determining microbial profiles [25,26]. However, this method requires a large amount of sample and suffers from limited availability of microbial genomic DNA compared to contaminating host DNA [27], such that for many large-scale studies, it remains cost-prohibitive and impractical [28]. In addition, long sequence read length and high sequencing coverage is required for meaningful assembly and community representation, and an optimal assembler remains elusive [25,27]. Alternative use of single-cell sequencing has been proposed, but also suffers from poor cost-effectiveness and outcome variability with how the sample is obtained (amount of sample, nature of study) [27]. Consequently, 16S rRNA gene amplicon-based Illumina sequencing methods remain valuable as cost-effective approaches most frequently used for analyzing microbiome profiles [29,30].

Chimeric read formation during PCR amplification is one step where improvement in microbial profiling could be addressed. Most efforts to handle the chimeric read problem have been directed towards improving the PCR amplification and/or DNA sample preparation steps to minimize the number of chimeric reads introduced [22,31–33]. Others have suggested that appropriate selection of the primer sets for amplification of the 16S rRNA gene variable regions could enhance the accuracy of the microbial profile [24,34]. However, the conclusions reached in those reports lack consistency and instead appear to apply only for the samples used in those particular studies, since previous studies noted that discrepancies in abundances for certain phyla were dependent on primer targets [23].

Other efforts have been directed towards improving chimera-filtering algorithms [25,35,36]. Several methods have been developed to identify chimeric sequences, including UCHIME [37], Chimera Slayer [38], Chimera Checker [39], and Bellerophon [40]. When using these programs for sequence analysis, chimeric reads are routinely discarded from the dataset and are not used for subsequent analyses. However, the impact of discounting chimeric reads on sequence data interpretation has not been well addressed [31,41], particularly in cases where the majority of the reads are suspected chimeras. Several datasets in the Mockrobiota data collection [42] showed errors of inflated operational taxonomic unit (OTU) counts. For example, whereas results for the V4 region displayed less than twofold OTU inflation due to chimera formation in one such study [43], the V3V4 and V4V5 regions exhibited as much as 3 and 14 times the number of expected OTUs, respectively, in others [42], suggesting inconsistencies in the chimera-filtering methods in those pipelines. Microbial profile compositions of mock community DNA samples, HM-276D,
HM-277D, and HM-278D, using the best practice procedures based on the V4 region, could deviate by more than fivefold in OTU counts from the expected composition [44].

We asked whether it is possible to recover microbial content information from chimeric reads generated during PCR-based sequencing approaches and assess the impact of recovering the information lost due to chimera formation. Here, we report an approach for microbial profiling using Illumina sequencing analysis of the 16S rRNA gene that includes recovery of information lost to chimeric reads. Our approach assumes that all chimeric reads are generated from PCR amplicons that can be matched to 16S rRNA genes of actual bacteria present in the sample. Toward this end, we chose the vaginal microbiome for our analyses since this microbial community has been extensively characterized and very few, if any, unknown bacterial taxa are expected to be newly discovered. We utilized a set of vaginal swab and lavage specimens from a cohort of 42 pregnant women collected for a pilot study exploring the effect of the vaginal microbiome on preterm birth. We compared the microbial profiles generated from the sequencing reads of the V3V4 and V4V5 regions of the full-length 16S rRNA gene amplicons from the vaginal swab and lavage samples. We examined the effect of PCR cycle number on the microbial profiles of samples from 5 selected subjects with varying compositions with a wide range of chimeric content, and we examined the effect of PCR extension time on the microbial profiles of samples from 18 subjects. For a subset of the vaginal samples from 12 subjects, we also performed metagenomic shotgun sequencing analysis for comparison. Overall, our findings suggest that sample composition and complexity are major determining factors for chimera formation that must be considered when interpreting microbiome data. Further, in order to build a consensus model that more accurately reflects the microbiome, a combination of multiple sequencing technologies and analysis tools should be adopted.

2. Methods

2.1. Sample collection

A total of 42 pregnant women (White = 3, African American = 26, Latina = 8, Asian = 1, unknown = 4) with an average age of 28 years (ranging from 18 to 43 years) were recruited for this study. Of these 42 women, 12 had a history of preterm delivery, and samples from these 12 subjects were subjected to further analysis. Two vaginal samples were collected from each subject at an average gestational age of 19 weeks (ranging from 16 to 23 weeks). A vaginal lavage sample was collected using a 15 mL sterile saline solution. Swab samples were obtained from the overall vaginal canal using vaginal swabs and placed in 1 mL sterile phosphate-buffered saline (PBS). All samples were stored at −80 °C until processed, as described below.

2.2. DNA Isolation/Purification

Each swab sample was extracted three times with 1 mL of PBS, and the combined mixture was centrifuged at 16,100*g* for 5 min. Similarly, each lavage sample was centrifuged at 16,100*g* for 5 min. The resulting pellets were used for further purification. The Genomic DNA Buffer Set (Qiagen, 19060) and Genomic-tip 20/G (Qiagen, 10223) were used for DNA purification. The pellets were resuspended in buffer B1 containing 20 μL of RNase A (10 mg/mL), followed by the addition of 20 μL of lysozyme (100 mg/mL) and 45 μL of proteinase K (20 mg/mL). The samples were incubated at 37 °C for 30 min, followed by the addition of 350 μL of buffer B2, gentle mixing, and incubation at 50 °C for

30 min. Thereafter, the manufacturer's protocol was followed. The genomic DNA was eluted twice with 1 mL of buffer QF (pre-heated to 50 °C to increase yield). The eluted DNA was precipitated by the addition of 700 μL of isopropanol, followed by mixing and centrifugation at 12,000*g* at 4 °C for 15 min. The resulting DNA pellets were washed with 1 mL of cold 70% ethanol and again centrifuged at 12,000*g* at 4 °C for 10 min. The pellets were air-dried for 10 min and subsequently solubilized in 50 μL of TE.

2.3. 16S rRNA gene amplicon preprocessing and sequencing

For 16S rRNA gene PCR amplification, each reaction contained 1 μL of genomic DNA, 25 μL of 2X DreamTaq Green PCR Master Mix (Thermo, K1081), 25 μL of water, and 1 μL of each primer (10 μM, 16S-27F: 5′-AGAGTTTGATYMTGGCTCAG-3′, 16S-1492R: 5′-CGGTTACCTTGTTACGACTT-3′). The thermocycler was set at 95 °C for 5 min, followed by 30 cycles at 95 °C for 15 sec, 47 °C for 30 sec, and 72 °C for 90 sec. The final extension step was set at 72 °C for 5 min. The resulting PCR amplicons were separated by electrophoresis on a 1% agarose gel, followed by extraction of the desired DNA band using the GeneJET gel extraction kit (Thermo, K0691). The 16S rRNA gene amplicons were eluted in 50 μL of water and quantified using a NanoDrop 2000 Spectrophotometer. The samples were diluted to 20 ng/μL and submitted for amplification and sequencing using the V3V4 and V4V5 primer sets for Fluidigm Access Array Amplification sequencing at the UIUC Roy J. Carver Biotechnology Center. The V3V4 primer pairs were 5′-CCTACGGG NGGCWGCAG-3′ and 5′-GGACTACNVGGGTWTCTAAT-3′, and the V4V5 primer pairs were 5′-GTGYCAGCMGCCGCGGTAA-3′ and 5′-CCGTCAATTCMTTTRAGT-3′. The resulting 250-nt paired-end reads obtained from Fluidigm-Illumina MiSeq.v2 sequencing after primer-sorting and demultiplexing were joined by fastp [45] using default parameters.

2.4. Metagenomic sequencing

Freshly purified DNA samples from eleven swab and twelve lavage samples from 12 of the 42 subjects were prepared for metagenomic sequencing, according to the protocol recommended by the W. M. Keck Sequencing Facility of the Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign (UIUC). Sequencing was performed on an Illumina HiSeq2500 with 160-nt paired-end reads and 500-bp fragment sizes. The remainder of the samples was stored at −80 °C until further use.

2.5. 16S rRNA database 16S_RefLib

Our 16S rRNA database, named 16S_RefLib, was initially constructed from the 16S rRNA sequences identified in several early vaginal microbiome studies, including PopSets: 66,878,480 [46], 52,222,145 [47], 63,146,101 [48], 119,352,235 [49]. Using a threshold of 99% identity, similar sequences were de-replicated, keeping the longest sequences and/or replacing with the closest entries in the NCBI dataset of 16S ribosomal RNA sequences (RefSeq). After initial BlastBin analysis of all merged reads, those sequences that did not have any partial match in our 16S_RefLib with > 99% identity were used as queries for BLASTn search using RefSeq dataset. New high-quality hits were added to the 16S_RefLib. The remaining unmatched sequences were used as queries for searching the NCBI nucleotide database. To ensure that the number of chimeric reads obtained using BlastBin for chimera-rich samples was not simply due to the absence of unidentified reference sequences in our 16S_RefLib, we also analyzed the sequencing data from two chimera-rich samples, S27 and S41, comparing the number of OTUs and chimeras obtained using BlastBin versus that obtained using the Silva database (see Supplemental Methods for the case of Samples S27 and S41). High hits of full-length 16S rRNA sequences from whole genome sequencing or from uncultured 16S rRNA sequences were added to the 16S_RefLib. For example, *Clostridium* sp. ND2 was identified from the sequenced genomes, and BVAB1 was identified from uncultured bacterial amplicon sequences frequently found in vaginal samples. The definition line of each sequence includes accession number, genus name, species name, and strain name. For sequences without a known species name, RDP project classifier was used to identify the closest (>99%) genus name or a higher classification name, clone name or other available information, which was used as species name in our database. Examples include: Coriobacteriaceae DNF00809 (KP192306.1), Lachnospiraceae BVAB1 (EF120366.1), and *Prevotella* skin nbw1010e05c1 (GQ047249.1).

2.6. QIIME2 OTU clustering

In the QIIME2 (2020.2.0) workflow, the joined reads were dereplicated and clustered using q2-vsearch (98% identity). Sequences in the 16S_RefLib reference databases were trimmed to generate V3V4 and V4V5 region databases, which were used for open reference clustering. The resulting OTU tables in QZA format were converted into BIOM-formatted tables using QIIME2 export tool. The BIOM tables were then converted into readable TSV-formatted files with "biom convert," included in the QIIME2 package. A shell script was used to assign Genus/Species name for each OTU accession number and identical species names were merged.

2.7. BlastBin OTU clustering

Our BlastBin algorithm allows for the identification and assignment of the source of chimeric reads. The assumption made is that all chimeric reads originate from sequences in the same sample that can be detected and binned by BLASTn. This program bins the reads by BLASTn at adjustable coverage and identity threshold defined by the user. Currently, BlastBin is implemented as a shell script program. For detailed information regarding BlastBin and counting chimeric reads, see Supplemental Materials and Methods.

To generate a mini-base for each dataset from a sample in this study, 1000 joined reads were randomly selected and searched against the 16S_RefLib database. Perfectly matched hits (identity > 99% with coverage > 90%) were used to generate a mini-database, which used to remove any additional perfectly matched sequences from the original dataset. The remaining reads in the reduced dataset were then subjected to the same process again. After 2 iterations of picking 1000 joined reads, all of the remaining reads in the final reduced dataset were used directly for BLASTn search of the 16S_RefLib to identify any remaining matches, which were then also added to the mini-database. The entire original dataset was then searched against this mini-database using BLASTn at 98% identity and word-size 32 to classify fully matched sequences with > 99% coverage. Any remaining sequences that were not matched were assumed to be chimeras derived from sequences in the mini-database. For each of the chimeric reads, the largest possible overlapping fragments of matching sequence hits from BLASTn were identified and used to generate a string of OTUs, each corresponding to a matching fragment. This string of OTUs was then used to generate fractional counts for each source OTU (see Supplemental Materials and Methods). For each sequenced dataset, BlastBin produced a table consisting of an accession number, genus/species name, matched read counts, calculated chimeric read counts, and total matched + chimeric read counts. BlastBin also generated a table according to accession numbers and a table according to genus/species names after merging accession numbers assigned to the same genus/species names.

Multiple datasets were merged using a shell script to produce a TSV OTU table, according to the species names or accession numbers for matched only reads, chimeric reads, or matched + chimeric reads.

2.8. Metagenome assembly pipeline

Our assembly pipeline included the following steps: (1) The Bowtie2 program [50] was used to filter away human sequences. (2) The MEGAHIT v1.2.9 program [51] was used to assemble the remaining paired reads. (3) The resulting contigs from assembly were used as query to BLAST against our curated 16S rRNA gene database (16S_RefLib). (4) Contigs containing 16S rRNA genes were selected using a shell script. (5) The Prokka program [52] was used to annotate all 16S rRNA gene-containing contigs. (6) The multiplicity of the contigs containing full-length gene sequences was used for calculating the relative abundance. For those contigs with truncated 16S rRNA sequences, a weight was applied using the ratio of coverage length over 1500 bp for the presumed full-length. Metagenomic contig-based fractional abundance was calculated from the ratio of each relative abundance over the total abundance (see Supplemental Table 5A). Sequencing reads after removing host DNA were also used as query for direct assignment by BLASTn search with 16S_RefLib as the database, setting identity at 99%, coverage at 90%, and selecting only one top hit. A shell script was used to count the hits and generate an OTU list and the read-based fractional abundance from the BLASTn output (see Supplemental Table 5B).

2.9. Statistical analysis and graphical presentations

Statistical analyses were performed using R version 3.6.3 or 4.1.0 (https://cran.r-project.org). Shannon diversity index (H) was calculated as $H = -\sum p_i \log p_i$ and Gini-Simpson diversity index was calculated as $1 - \sum p_i{}^2$, where p_i is the fraction for the i-th OTU, using the R package vegan version 2.5–7 function diversity (), with index = shannon and simpson, respectively. The Morisita-Horn similarity index (MH) was calculated as $MH = 2\ (\sum x_i\ y_i)/(\sum x_i^2 + \sum y_i^2)$, where x_i , y_i are the abundances for the i-th OTU in sample X and Y, respectively. The MH index was calculated using the R package vegan function vegdist() with method = Horn, and converted into similarity, where $MH = 0$, if no common species were found in the two samples, and $MH = 1$, if the species occur in the same proportions in both samples. ANOVA was performed using the R package stats function aov(). Correlations and t-tests were calculated using Microsoft Excel (version 16.47.1). Graphical presentations of data were generated by using Excel or the R package ggplot2 with compatible packages of gplots, dplyr, reshape, ggpubr, and ggrepel.

3. Results

3.1. Effect of PCR cycle number on chimera formation

PCR amplicons of the full-length 16S rRNA gene from vaginal swab and lavage samples obtained from all 42 subjects in our cohort were generated under standard PCR conditions of 30 cycles and 90-second extension times, using universal bacterial primers, 27F and 1492R. The resulting full-length amplicons were used as templates to generate V3V4 and V4V5 amplicons for Illumina sequencing and analysis. Those reads that were > 98% identical to a known reference sequence in the 16S rRNA gene RefSeq database, to a known genome sequence, or to a known amplicon in published studies, were considered as "matched" to that taxon. Those reads that were not matched were considered to be either chimeric reads or currently unidentified sequences, both denoted as "chimera" reads from here on. The sequencing results for the swab and lavage samples indicated that there was a wide range of chimeric content found among the different samples (see Supplemental Table S1A and S1B).

To determine the effect of PCR amplification conditions on the occurrence of chimeras, a subset of vaginal swab and lavage samples representing samples with chimeric read contents ranging from 5% to 50% were selected: Subject 9 swab (S9), Subject 11 lavage (L11), Subject 13 lavage (L13), Subject 27 swab (S27), and Subject 41 swab (S41). Full-length 16S rRNA gene amplicons of these samples were generated under PCR conditions of 25, 30, 35, or 40 cycles with 90-second extension times, followed by Illumina sequencing of the V3V4 or V4V5 regions. The microbial profiles for these samples are summarized in Supplemental Table S2A. For matched sequences, similar microbial profiles were obtained with each of the different cycle numbers for each of the samples. As shown in Fig. 1A, PCR cycle number had no consistent effect on the fraction of chimeric reads detected in a sample. For those samples having a high match rate with sequences in the 16S_RefLib database (S9 and L11), the number of matched sequences actually increased with cycle number, whereas for those samples having fewer matched reads (and presumably more chimeric reads), there was either little change (L13) or a decrease (S27 and S41) in matched reads.

If PCR cycle number were to have a similar effect on chimera formation for different samples, then one would expect that the correlation of chimera occurrence for the series with different PCR cycle numbers would be similar among different samples as to that observed between the V3V4 and V4V5 reactions within the same sample. The mean correlation of cycle dependency on chimera formation (comparing PCR cycle number 25, 30, 35 or 40) between each reaction pair for V3V4 and V4V5 from the same sample was r = 0.88 ± 0.18 (n = 5); while the mean correlation between any two reaction sets of different samples was r = -0.22 ± 0.61 (n = 40). In contrast, the mean correlation of sample dependency on chimera formation for each pair of cycle numbers was r = 0.97 ± 0.03 (n = 6). ANOVA analysis of chimera content among the 5 samples, the 4 PCR cycle numbers, and the 2 variable regions of the 16S rRNA gene showed p < 2e-16 for samples, p > 0.8 for cycle number, and p > 0.9 for the variable region. Details of these comparisons are summarized in "Correlation and ANOVA Analysis of Five Samples" in the Supplemental Methods. These results suggest that the sample itself has more impact on the extent of chimera formation than the number of PCR cycles used for full-length 16S rRNA gene amplification or the variable region used for sequencing. Chimera formation was expected to occur during the preparation of full-length 16S rRNA gene amplicons, as well as during the preparation of the V3V4 and V4V5 amplicons. We were only able to examine the effect of cycle number on chimera formation during the first amplification step, but we could not quantify the impact of the second amplification step.

The diversity of the samples, calculated as the Shannon diversity using only matched reads (Fig. 1B), showed that those samples containing more matched sequences (S9, L11, and L13) represent less diverse communities. For L11 and L13, the only change in diversity was observed between cycle numbers 25 and 30. Since these samples are dominated primarily by one *Lactobacillus* species, it is reasonable to assume that increasing the cycle number resulted in loss of the minor components and loss of diversity. In contrast, the samples with fewer matched reads and presumably more chimeric reads (S27 and S41) had higher diversity and were refractory to the number of PCR cycles used. Based on the results observed for both chimera occurrence and sample diversity, chimera formation appears to be dependent on the composition and the diversity of the sample, and not on the PCR cycle number used.

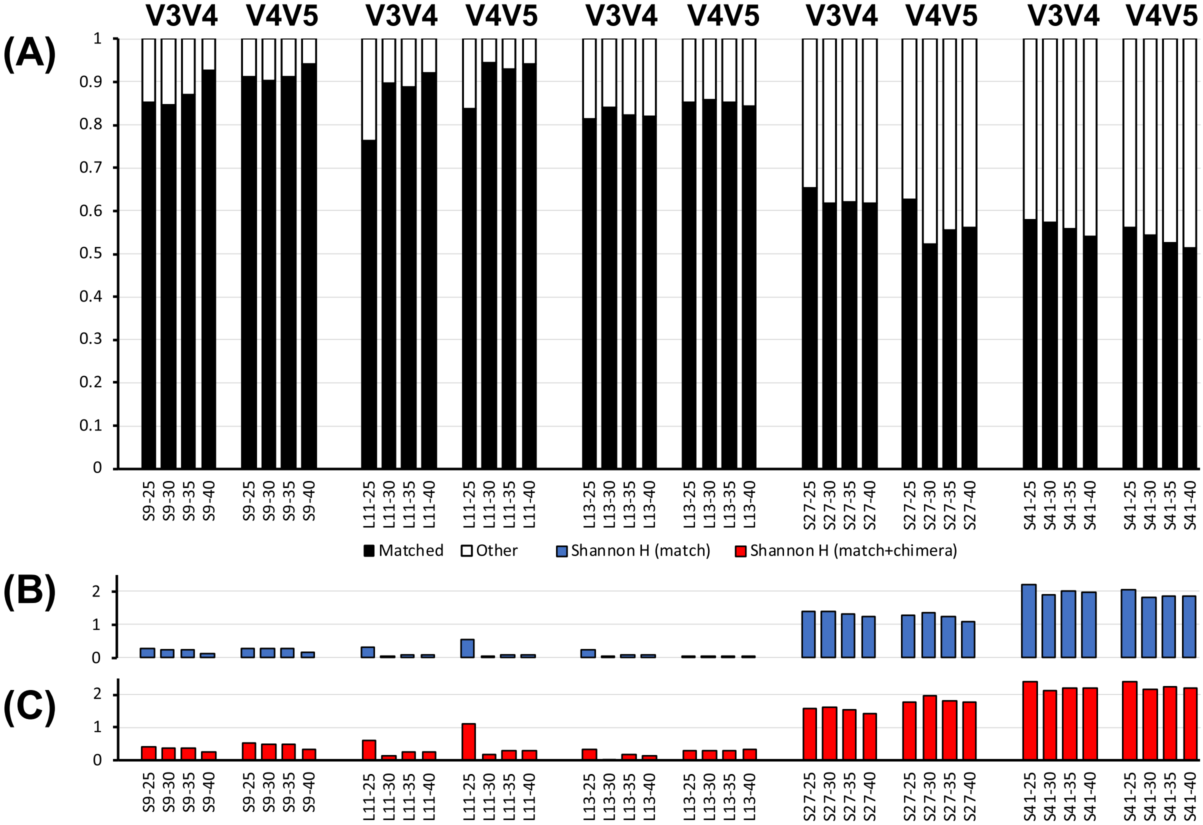


Fig. 1. Dependence of PCR cycle number on chimeric read formation in 5 selected samples. (A) Shown is the fraction of read counts that matched to reference taxa in the 16S_RefLib database (black bars), and those having no matches (white bars), denoted as "Other" and presumed to be chimeric reads. S and L in sample names denote swab and lavage, respectively. Sample numbers (9, 11, 13, 27, or 41) and PCR cycle numbers (25, 30, 35, or 40) are incorporated into the sample name. (B) Shown is the Shannon diversity index H for each of the selected samples (blue bars), calculated only from the matched read counts. (C) Shown is the Shannon diversity index H for each of the selected samples (red bars), calculated from the adjusted read counts after including chimeric reads. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Counting chimeras

In many of the swab and lavage samples from the 42 subjects, we observed that >50% of the total reads could not be assigned at 98% identity to known taxa in the RefSeq database, known genome sequence databases, or our 16S_RefLib database. In some cases, these unmatched sequence reads, presumed to be chimeric reads, reached as much as 80% of the total non-host sequenced reads (Fig. 2A). These chimeric reads were detected and discarded by QIIME2 during the standard sequencing analysis protocol. To determine if it is possible to recover information lost in the discarded chimeric sequence reads, we developed an algorithm, BlastBin, based on NCBI BLASTn [53], that assigns the taxa with matched sequences as the source of the chimeras, as well as assigns a fractional count to each source.

For application of BlastBin, our assumptions were:

(1) Chimera products can only be derived from precursors that can form non-chimera products. That is, if a chimera product is generated, then non-chimera products of the precursors should also be detectable from the same PCR reaction.
(2) BLASTn search can be used to assign the fragments in a chimera product to their corresponding precursor taxa from a database consisting only of the non-chimera precursors that are present in the sample.
(3) For each observed chimera joining site, there are two precursor templates of equal amounts. Therefore, for the detected number (n) of a given bipartite chimera product, if the chimeric process had not occurred, there would be half the number (n/2) of each non-chimera product. For the number (n) of a given tripartite chimera product with two joining sites, if the chimera process had not occurred, there would be a third of the number (n/3) of each of the three non-chimera products. Similarly, the number (n) of a k-partite chimera product with k-1 joining sites can be counted as an equal number of n/k of each non-chimera product. A mathematical rationale for this chimera counting method is provided as "Rationale for Counting Chimera Reads" in Supplemental Methods.

Based on these assumptions, we devised the BlastBin program in shell script, which could be translated into Python (see Supplemental Materials and Methods). BlastBin takes the merged read pairs and generates an OTU table using only the matched sequence reads (Supplemental Table S1B) and a separate OTU table that incorporates the fractional counts recovered from the chimeric reads (Supplemental Table S1C). When chimeric reads are included in the calculation, the diversity of the samples increases (Fig. 2A, compare red dots versus blue dots). From the diversity comparison using scatter plot (Fig. 2B), in all cases the Shannon diversity considering both matched and chimeric reads is greater than that considering only matched reads. This was confirmed when the Gini-Simpson index, which considers the number of species present and the relative abundance of each species, was used for comparison (Fig. 2C). These findings were also observed for the selected samples used for PCR cycle-dependent analysis mentioned above (compare Fig. 1B and 1C, and the corresponding OTU tables Supplemental Table S2A and S2B).

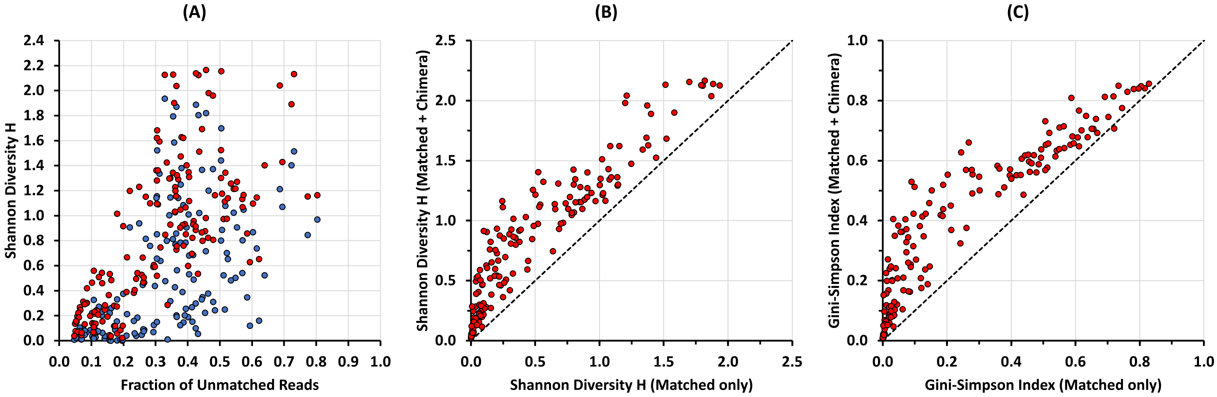


Fig. 2. Interplay of chimera formation and sample diversity. (A) More diverse samples yield more unmatched reads, presumed to be chimeric reads. Shown is the Shannon diversity index H for all swab and lavage samples from the 42 subjects, based on analysis of the V3V4 and V4V5 regions in the full-length 16S rRNA amplicons obtained using standard PCR conditions. The Shannon diversity index H was calculated from the relative abundances using only matched reads (blue dots) or from the adjusted relative abundances considering chimeric reads (red dots) versus the fraction of chimeric reads. (B, C) Counting chimeric reads increases diversity. Shown are the Shannon diversity index H comparison (B) and the Gini-Simpson diversity index comparison (C) calculated from the relative abundance adjusted to include chimeric reads versus that from the relative abundance using only matched reads. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3. Effect of PCR extension time on chimera formation

It is generally believed that longer extension times decrease the amount of incomplete PCR products and minimize formation of chimera products. Thus, to obtain full-length 16S rRNA gene amplicons, extension times of 60-seconds [54], 90-seconds [55], or longer [32] have been recommended as standard conditions by different groups. We tested the effect of using a PCR extension time of 60 or 90 s on a set of samples (18 swabs and the corresponding lavage samples from Subjects 13 through 30, plus 12 swabs from Subjects 31 through 42) under standard PCR amplification conditions. The resulting full-length 16S rRNA gene amplicons were then subjected to Illumina sequencing and analysis of the V3V4 and V4V5 regions. As shown in Fig. 3A, increasing the PCR extension time from 60 to 90 s generally did not alter the number of chimeric reads generated for the majority of the samples. However, in some instances, the 60-second reactions yielded more chimeric reads, whereas for others the 90-second reactions yielded more, as summarized in Fig. 3C.

In most cases, the two PCR extension times yielded similar diversity indices (Fig. 3B), although in some samples, the 60-second reactions appeared to have more diversity than the 90-second reactions (Fig. 3D). Consistent with this finding, the 60-second reactions generally picked up more taxa (Fig. 3E). As was noted above, the diversity increased when chimeric reads were considered in determining the microbial abundance (compare top and bottom panels in Fig. 3B). This held true for both PCR extension times (Fig. 3F). Again, there was no consistency with regard to PCR amplification conditions on the number of matched reads, chimera formation, or diversity of the microbial profiles. Combined, these findings suggest that PCR conditions are not solely responsible for influencing the formation of chimeras or the detection of sample diversity, and instead suggest that sample composition itself contributes to chimera formation and diversity.

3.4. Comparison of microbial profiles derived from PCR-based sequencing versus metagenome shotgun sequencing analyses

PCR-independent metagenomic sequencing is assumed to provide the closest representation of the actual microbial profiles of samples [28,56–59]. However, 16S rRNA genes only account for a small fraction of the entire genome, so only a small fraction of the total reads can be used for ribotyping, and high sequencing depth is required to achieve reliable taxonomic resolution. With that in mind, we performed metagenomic analysis by using Illumina HiSeq metagenome sequencing of the swab and lavage samples from Subjects 1 through 12. In our dataset of 23 samples, 19 ± 13% (range 7% to 53%) of the total reads remained after filtering out host genomic DNA using Bowtie2 [50]. Of the non-host reads, 0.22 ± 0.06% (range 0.11 to 0.32%) could be matched to 16S rRNA genes in our 16S_RefLib database using BLASTn. This equated to only a tiny fraction of the metagenomic shotgun sequencing reads, 0.046 ± 0.036% (range 0.012 to 0.139%), that were usable for bacterial taxon identification (see Table 1).

The non-host DNA reads were assembled using MEGAHIT [51], and the resulting contigs were used for BLASTn search against our 16S_RefLib database to identify contigs containing a 16S rRNA gene, which were then annotated with Prokka to identify the sequence range coding for that 16S rRNA gene. The abundance of bacterial taxa based on the metagenomic assembly was calculated from the 16S rRNA gene coverage as well as the multiplicity of the contigs containing the 16S rRNA gene sequence of that taxon (Supplemental Table S5A). The non-host DNA reads were also used directly for BLASTn search to obtain read-based microbial profiles (Supplemental Table S5B). In comparing the contig-based microbial profiles after assembly with the read-based profiles without assembly, the number of bacterial taxa identified from the assembly-based method appear to be undercounted compared to the numbers by other methods (ranging from 1 to 26), while that from the read-based method appear to be overcounted (ranging from 13 to 79) (see Fig. 4A). In 15 out of 23 cases, the diversities of the samples were comparable between the contig-based and read-based methods (Fig. 4B), suggesting that the majority of the taxa for the read-based counts in these samples were minor components of the microbial community.

The number of bacterial taxa identified from the metagenomic analyses were also compared with that obtained from PCR-based analysis of the V3V4 and V4V5 regions of full-length 16S rRNA gene amplicons generated using 30 or 40 cycles (Fig. 4A, compare also Supplemental Tables S1B and S1C with S5A and S5B). As expected, more taxa were detected in all cases using the read-based metagenomic method than the assembly-based method. In 12 out of 23 cases, the PCR-based sequencing method using 30 cycles yielded similar bacterial taxon counts as those obtained from assembled metagenomic reads, whereas using 40 cycles yielded more bacterial taxa in all cases. Interestingly, the number of taxa observed in PCR amplicons after 30 cycles was consistently less than that observed after 40 cycles. In some cases (S1, L1, S3, L3,

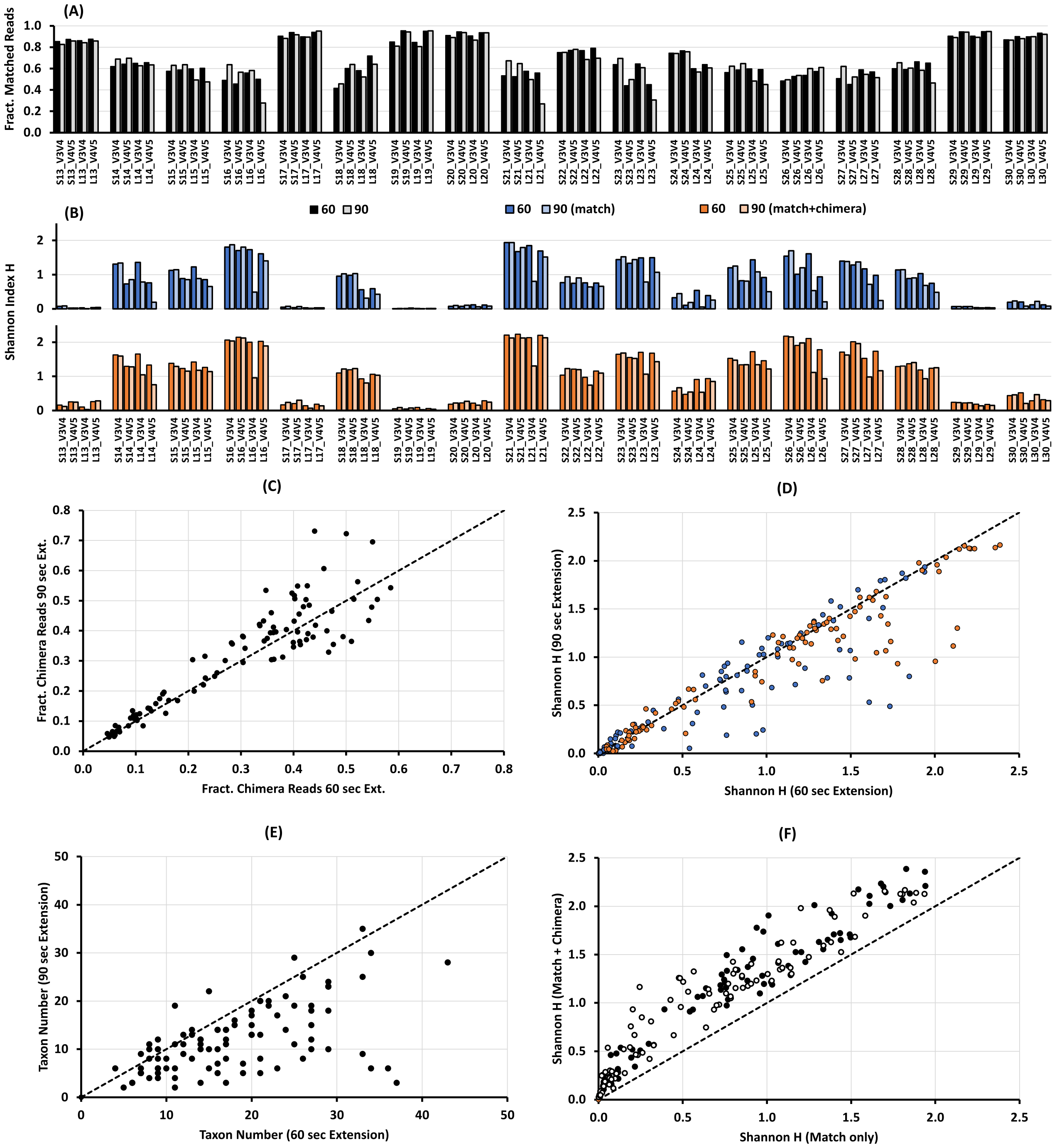


Fig. 3. Dependence of PCR extension time on chimeric read formation. (A) Shown is the fraction of read counts from selected samples that matched to reference taxa in the 16S_RefLib database. S and L in sample names denote swab and lavage, respectively. Subject numbers (13 through 30) and 16S rRNA gene regions (V3V4 or V4V5) are incorporated into the sample name. Full-length 16S rRNA gene amplicons were generated using PCR reactions with extension time of 60 s (black bars) or 90 s (grey bars). (B) Shown are the Shannon diversity index H values for the V3V4 and V4V5 regions of 16S rRNA gene from swab and lavage samples from Subjects 13 through 30 and swabs only from Subjects 31 through 42. H values were calculated from the matched read only counts (top panel, blue bars) or from the adjusted read counts including chimeric reads (bottom panel, orange bars), where PCR extension time of 60 s (dark bars) or 90 s (light bars) was used to generate full-length 16S rRNA gene amplicons. (C) Shown is a scatter plot of the fraction of chimeric reads (solid black circles) obtained from reactions using 60-second extension time versus 90-second extension time, similar to that described for (B). (D) Shown is a scatter plot comparing the effect of PCR extension time (60 versus 90 s) on the Shannon diversity index H values, similar to that described for (B), calculated from the matched read only counts (blue circles) or from the adjusted read counts including chimeric reads (orange circles). (E) Shown is a scatter plot comparing the number of taxa identified using PCR extension times of 60 s versus 90 s for the full-length 16S rRNA gene amplicons, followed by Illumina sequencing of the V3V4 and V4V5 regions in both swab and lavage samples from Subjects 13 through 30, as well as swab samples only from Subjects 31 through 42. (F) Shown is a scatter plot comparing the effect of counting chimeric reads (match only versus match + chimera) on the Shannon diversity index H values, calculated from fractional reads obtained using extension time of 60-seconds (solid black circles) or 90-seconds (open black circles), similar to that described for (B). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Metagenomic shotgun sequencing results for swab and lavage samples from Subjects 1 through 12.

Sample	Total paired reads	Non-host paired reads	Non-host yields	16S rRNA hits	16S rRNA %	16S yields
S1	57,048,685	5,310,534	9%	8088	0.15%	0.014%
L1	51,857,992	4,604,966	9%	6642	0.14%	0.013%
S2	49,442,086	9,719,560	20%	26,826	0.28%	0.054%
L2	46,319,475	15,074,561	33%	35,583	0.24%	0.077%
S3	56,017,859	18,510,934	33%	53,417	0.29%	0.095%
L3	49,973,698	25,917,542	52%	69,688	0.27%	0.139%
S4	41,247,549	8,228,648	20%	26,062	0.32%	0.063%
L4	44,974,048	8,738,985	19%	27,831	0.32%	0.062%
S5	54,453,859	3,869,396	7%	4254	0.11%	0.008%
L5	52,573,652	4,588,747	9%	6187	0.13%	0.012%
L6	54,177,968	11,414,311	21%	28,922	0.25%	0.053%
S7	48,367,288	4,273,106	9%	10,073	0.24%	0.021%
L7	52,099,277	5,402,199	10%	15,270	0.28%	0.029%
S8	53,885,469	5,986,783	11%	10,106	0.17%	0.019%
L8	40,612,001	6,282,090	15%	10,742	0.17%	0.026%
S9	61,537,335	7,658,706	12%	15,703	0.21%	0.026%
L9	53,667,891	8,490,088	16%	17,000	0.20%	0.032%
S10	46,366,307	13,079,450	28%	31,697	0.24%	0.068%
L10	48,902,116	25,712,763	53%	63,467	0.25%	0.130%
S11	49,671,540	8,025,035	16%	18,626	0.23%	0.037%
L11	45,930,502	5,696,707	12%	11,807	0.21%	0.026%
S12	45,147,780	6,069,520	13%	15,547	0.26%	0.034%
L12	41,887,200	5,373,884	13%	10,396	0.19%	0.025%

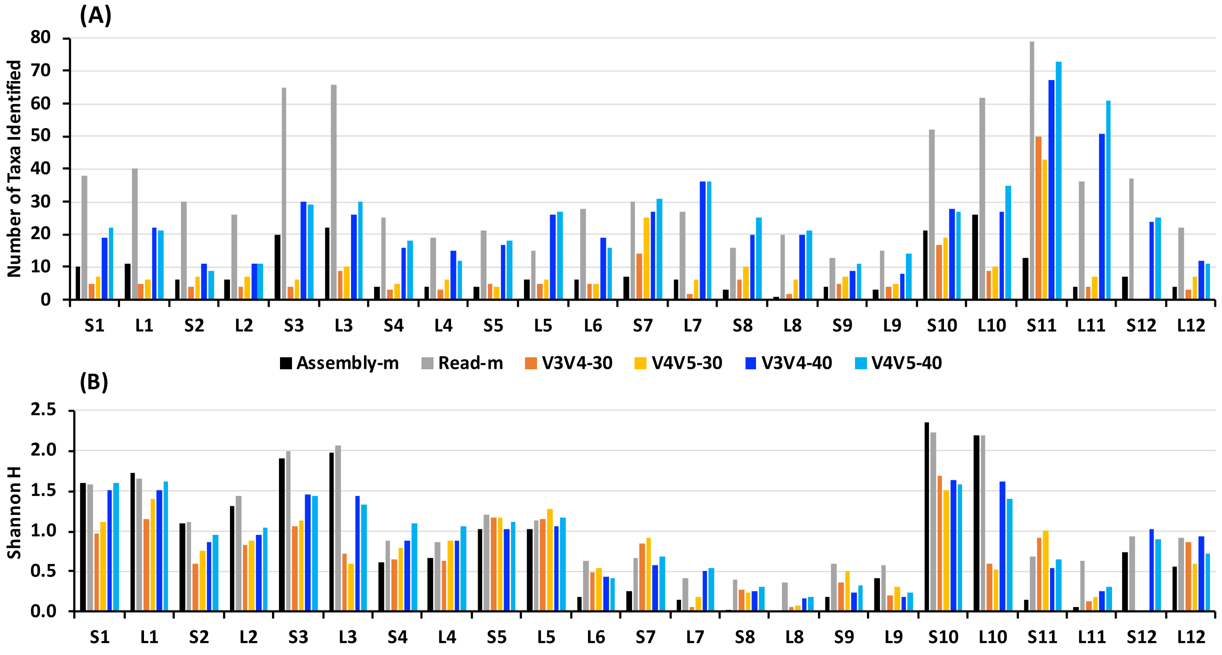


Fig. 4. Comparison of PCR-based and metagenome shotgun sequencing methods for swab and lavage samples from 12 subjects. (A) Number of taxa identified in swab (S) and lavage (L) samples from Subjects 1 through 12, obtained from metagenome shotgun sequencing, using assembled-based reads (Assembly-m) or raw read-based (Read-m), or from PCR-based methods considering chimeras analyzing the V3V4 (dark shade) or V4V5 (light shade) regions of the full-length 16S rRNA gene amplicons generated with 30 (orange) or 40 (blue) cycles. For S6, no genomic DNA was obtained, and for S12, no full-length amplicon was obtained for 30 cycle reaction due to insufficient sample. (B) Shannon diversity index H for each of the samples in (A), calculated from the abundance of taxa identified by each of the respective methods as in (A). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

L7, *L*10), the Shannon diversity was also greater after 40 cycles than 30 cycles, but most of the samples (S2, L2, S4, L4, S5, L5, L6, S7, S8, L8, S9, L9, S11, L11, S12, and L12) displayed similar Shannon diversity across the PCR-based and read-based methods (Fig. 4B). These findings again indicate that at least some of the discrepancies in taxon counts in these samples could be attributed to low-abundance species. The only exceptions to this, where noticeable differences were observed, were the highly diverse profiles of samples from Subjects 3 and 10.

Many of the bacterial taxa present in the metagenomic profiles were not detected in the 16S rRNA gene profiles using 30 cycles for PCR amplification (Fig. 4, also compare Supplemental Tables S4A and S4B with Supplemental Tables S1B and S1C). This is consistent with the lower number of taxa and lower diversity observed for the 16S rRNA gene profiles of these samples. However, after 40 cycles, most of the taxa observed in the corresponding metagenomic profiles could be detected, with the exception of *Gardnerella vaginalis*, *Atopobium vaginae*, and *Saccharibacterium* TM7, which is likely due to bias introduced by the imperfect sequence matching of the universal primers, as previously reported [60]. In this comparison, several of the taxa not found by the metagenomic assembly method, including some with abundances > 1%, could be detected by the metagenomic read-only and the 40-cycle PCR-based methods.

Closer examination of the microbial profiles from Subjects 3 and 10 (Table 2) revealed the intricate variance of the microbial profile composition using different analysis methods. For example, *G. vaginalis* was detected as a major component of the microbial content in the assembly- and read-based metagenomic analyses but was not detected or barely detected by the PCR-based methods, most likely due to poor primer matching, as mentioned above. In this case, the assembly-based method overcounted the abundance compared to the read-based method (3 times for samples from Subject 3 and 2 times for Subject 10), revealing a flaw in using multiplicity of contigs as a basis for counting abundances. In the case of *G. vaginalis*, the reads of other species may have been incorrectly incorporated into the contig containing the 16S rRNA gene for *G. vaginalis*. Using the assembly-based method, the bacterium *Lactobacillus iners* was undercounted, while *Lactobacillus gasseri* and *Lactobacillus crispatus* were not detected. On the other hand, even though *A. vaginae* was present in relatively low abundance in *L*10, the full-length assembly could be detected, indicating that the assembly-based method has the capability of picking up low abundant taxa (even < 1%) so long as no other taxa with similar 16S rRNA sequences are present in the sample. This also appears to be the case for *Prevotella* skin (S3, L3), *Peptoniphilus lacrimalis* (S3, L3, *L*10), *Sneathia amnii* (S3, L3), *Prevotella colorans* (S3, L3), *Mycoplasma hominis* (*L*10), *Bacteroidales* L3 (L3), and a few others.

The similarities between the *Lactobacillus* 16S rRNA gene sequences could result in assembly error of metagenomic sequences. In samples from Subjects 3 and 10, the reads of less abundant *Lactobacillus* species, such as *L. jensenii, L. crispatus, L. gasseri* and *L. coleohominis,* may have been misassembled into the dominant *L. iners* species to the point where they were no longer detected. For samples where there are multiple *Lactobacillus* species of comparable abundances, as was found in samples from Subject 5, misassembly of the highly similar sequence reads could result in high chimera content of the profile. Some 16S rRNA gene sequences identified by metagenomic assembly in samples from Subject 5 could be assigned to a single *Lactobacillus* species based on the highest BLASTn sequence identity match (Table 3). However, upon closer examination, the best hit was<99% for a number of the sequences, and in each case, there were several alternative hits with similar lower identity scores (Table 4), indicating that the reads were likely misassembled. Among the *Lactobacillus* species, *L. coleohominis* is the most distant (<90% identity) from other more abundant *Lactobacillus* species in the samples, and so, in this case, it was still possible to obtain a complete assembly of the 16S rRNA gene, despite its abundance at < 1%. In agreement with the read-based metagenomic analysis for samples from Subject 5, the microbial profiles from the PCR-based method using 40 cycles were dominated by five *Lactobacillus* species, in order of abundance: *L. crispatus*, *L. iners*, *L. jensenii*, *L. gasseri,* and *L. coleohominis.* In total, closer examination of the microbial profiles from Subjects 3, 5 and 10 comparing different analysis methods support the notion that microbial content and their abundances affects the reliability of assembly-based shotgun sequencing analysis.

We compared the microbial profiles generated by the PCR-based methods with the read-based shotgun sequencing method, using the Morisita-Horn (MH) similarity index [61]. As shown in Fig. 5A, the microbial profiles for the samples from Subjects 1 through 12 generated by QIIME2 using 40-cycle reads were more similar to the read-based metagenomic profiles than using the 30-cycle reads. According to the MH similarity indices (Fig. 5B), our BlastBin method using only matched reads yielded similar microbial profiles as the QIIME2 method for both 30- and 40-cycle reads. This is expected since in each case chimeric reads were removed from the analysis. Without or with the consideration of chimeric reads, the profiles generated using 30 or 40 cycle numbers were different from each other. When chimeric reads were considered, the 30-cycle amplicons displayed a greater shift in the MH similarity pattern, while there was less of a shift for the 40-cycle amplicons. The smaller impact of chimera counting on the 40-cycle profiles versus 30-cycle profiles most likely reflects the greater coverage of taxa reached by using 40 PCR cycles, even when only matched reads were used. The outliers found in the MH similarity indices are primarily due to those profiles from the two most diverse samples, those from Subjects 3 and 10 (solid circles in Fig. 5A), again supporting the notion that the nature and complexity of the sample impacts the reliability of the analysis method.

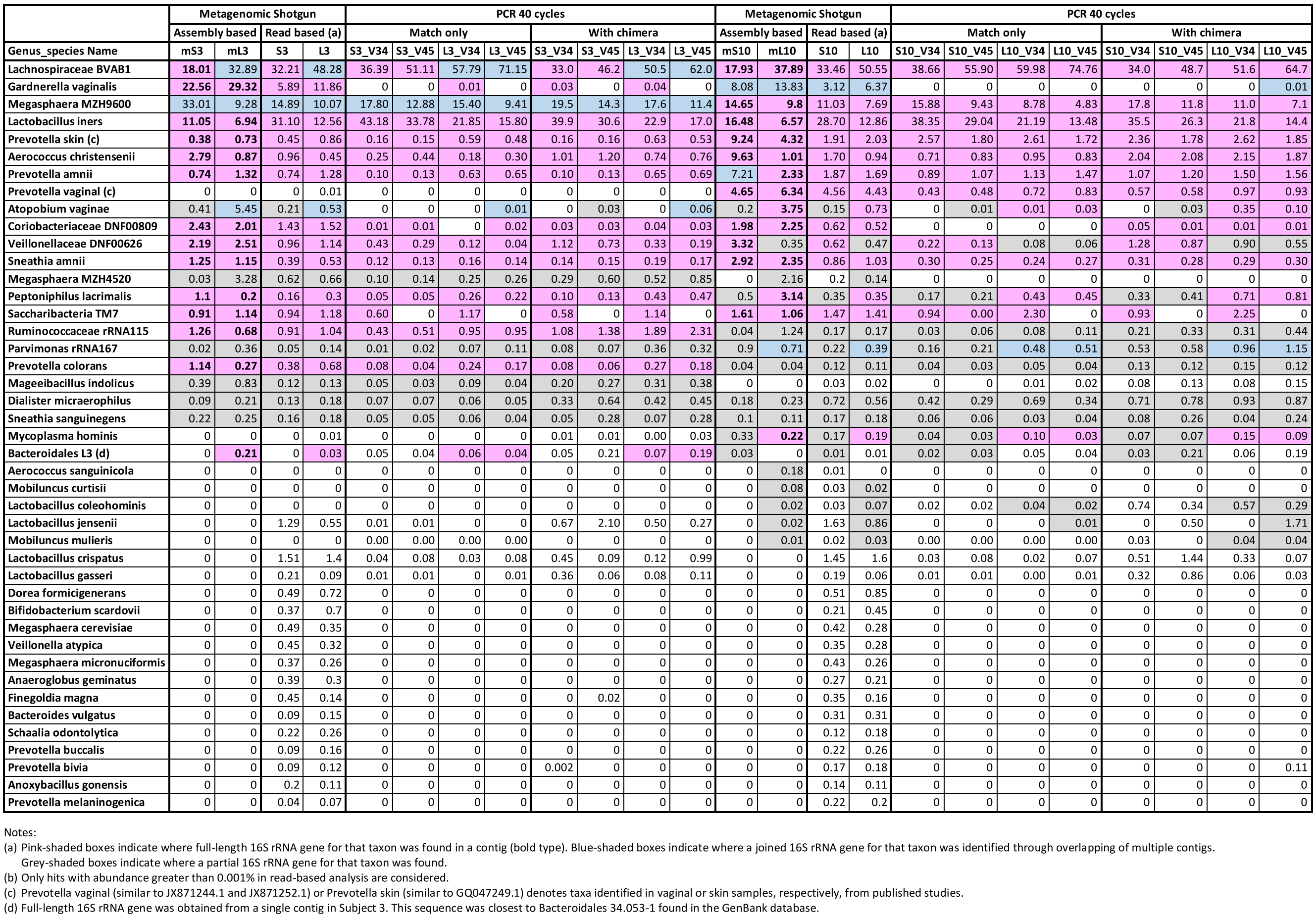
4. Discussion

As has been well established in the field, we found that the formation of chimeric PCR products is an unavoidable outcome of any PCR amplification of complex DNA samples. Our samples yielded a wide range of chimeric content when standard methods were used for PCR amplification and sequencing analysis. However, unlike previous reports that suggested chimera formation could be largely corrected by adjusting the PCR conditions or cycle number used, we found that the differential occurrence of chimeric reads for a given sample depends on the nature of the microbial content and the complexity of the microbial community of the sample, and less so on the PCR conditions such as cycle number or extension times used in obtaining the full-length 16S rRNA amplicons. We also found that the extent of chimera formation was similar whether the V3V4 or V4V5 regions were used to assess the PCR amplification of full-length 16S rRNA gene amplicons. Overall, we found that more diverse samples tended to have more chimeric reads. While there were minor effects of PCR cycle number on chimera formation observed for a number of samples, the effects were inconsistent, with some samples displaying an increase in chimera formation with increased cycle number, while others a decrease. In general, PCR extension time had little effect on chimera formation, even though some samples showed minor differences with 60-sec versus 90-sec cycles. While the 60-sec extension times tended to yield more diverse profiles, the outcomes were inconsistent and varied among samples. Again, these findings point toward sample dependency of chimera formation on the microbial profiles.

Removal of chimeras from sequence analysis is the standard protocol for all current pipelines. However, when the chimeric read content in a microbiome sample can range from 5% to as much as 80%, such as was observed in our vaginal swab and lavage samples from 42 subjects, it is difficult to justify excluding chimeras from microbial community analyses. According to a mathematical rationale for chimera formation (see equation 17 of the mathematical rationale in Supplementary Materials and Methods), after i cycles of amplification, the original ratio between the two species A and B in a sample is $A_0/B_0 = (A_i + 0.5{*}C_i)/(B_i + 0.5{*}C_i)$, where A_i, B_i, and C_i represent the amounts of A, B, and chimera C. When these two species are present at greatly different abundances, $A_0 \gg B_0$, after i cycles of amplification the ratio will become greatly exaggerated compared to the original ($A_i/B_i > A_0/B_0$). The consequence will be that more sequences from the minor component will be lost due to chimera formation.

To determine how the information loss by removing chimeric reads from analysis impacts the interpretation of microbiome data, we first needed to devise a way to count chimeric reads in terms of what they represent with regard to the microbial composition of the original sample. So, we developed the BlastBin algorithm to

Table 2
Comparison of microbial profiles obtained for swab and lavage samples from Subjects 3 and 10.[(a)]



	Metagenomic Shotgun				PCR 40 cycles								Metagenomic Shotgun				PCR 40 cycles							
	Assembly based		Read based (a)		Match only				With chimera				Assembly based		Read based (a)		Match only				With chimera			
Genus_species Name	mS3	mL3	S3	L3	S3_V34	S3_V45	L3_V34	L3_V45	S3_V34	S3_V45	L3_V34	L3_V45	mS10	mL10	S10	L10	S10_V34	S10_V45	L10_V34	L10_V45	S10_V34	S10_V45	L10_V34	L10_V45
Lachnospiraceae BVAB1	**18.01**	32.89	32.21	48.28	36.39	51.11	57.79	71.15	33.0	46.2	50.5	62.0	**17.93**	**37.89**	33.46	50.55	38.66	55.90	59.98	74.76	34.0	48.7	51.6	64.7
Gardnerella vaginalis	**22.56**	**29.32**	5.89	11.86	0	0	0.01	0	0.03	0	0.04	0	8.08	13.83	3.12	6.37	0	0	0	0	0	0	0	0.01
Megasphaera MZH9600	33.01	9.28	14.89	10.07	17.80	12.88	15.40	9.41	19.5	14.3	17.6	11.4	**14.65**	**9.8**	11.03	7.69	15.88	9.43	8.78	4.83	17.8	11.8	11.0	7.1
Lactobacillus iners	**11.05**	**6.94**	31.10	12.56	43.18	33.78	21.85	15.80	39.9	30.6	22.9	17.0	**16.48**	**6.57**	28.70	12.86	38.35	29.04	21.19	13.48	35.5	26.3	21.8	14.4
Prevotella skin (c)	**0.38**	**0.73**	0.45	0.86	0.16	0.15	0.59	0.48	0.16	0.16	0.63	0.53	**9.24**	**4.32**	1.91	2.03	2.57	1.80	2.61	1.72	2.36	1.78	2.62	1.85
Aerococcus christensenii	**2.79**	**0.87**	0.96	0.45	0.25	0.44	0.18	0.30	1.01	1.20	0.74	0.76	**9.63**	**1.01**	1.70	0.94	0.71	0.83	0.95	0.83	2.04	2.08	2.15	1.87
Prevotella amnii	**0.74**	**1.32**	0.74	1.28	0.10	0.13	0.63	0.65	0.10	0.13	0.65	0.69	7.21	**2.33**	1.87	1.69	0.89	1.07	1.13	1.47	1.07	1.20	1.50	1.56
Prevotella vaginal (c)	0	0	0	0.01	0	0	0	0	0	0	0	0	**4.65**	**6.34**	4.56	4.43	0.43	0.48	0.72	0.83	0.57	0.58	0.97	0.93
Atopobium vaginae	0.41	5.45	0.21	0.53	0	0	0	0.01	0	0.03	0	0.06	0.2	**3.75**	0.15	0.73	0	0.01	0.01	0.03	0	0.03	0.35	0.10
Coriobacteriaceae DNF00809	**2.43**	**2.01**	1.43	1.52	0.01	0.01	0	0.02	0.03	0.03	0.04	0.03	**1.98**	**2.25**	0.62	0.52	0	0	0	0	0.05	0.01	0.01	0.01
Veillonellaceae DNF00626	**2.19**	**2.51**	0.96	1.14	0.43	0.29	0.12	0.04	1.12	0.73	0.33	0.19	**3.32**	0.35	0.62	0.47	0.22	0.13	0.08	0.06	1.28	0.87	0.90	0.55
Sneathia amnii	**1.25**	**1.15**	0.39	0.53	0.12	0.13	0.16	0.14	0.14	0.15	0.19	0.17	**2.92**	**2.35**	0.86	1.03	0.30	0.25	0.24	0.27	0.31	0.28	0.29	0.30
Megasphaera MZH4520	0.03	3.28	0.62	0.66	0.10	0.14	0.25	0.26	0.29	0.60	0.52	0.85	0	2.16	0.2	0.14	0	0	0	0	0	0	0	0
Peptoniphilus lacrimalis	**1.1**	**0.2**	0.16	0.3	0.05	0.05	0.26	0.22	0.10	0.13	0.43	0.47	0.5	**3.14**	0.35	0.35	0.17	0.21	0.43	0.45	0.33	0.41	0.71	0.81
Saccharibacteria TM7	**0.91**	**1.14**	0.94	1.18	0.60	0	1.17	0	0.58	0	1.14	0	**1.61**	**1.06**	1.47	1.41	0.94	0.00	2.30	0	0.93	0	2.25	0
Ruminococcaceae rRNA115	**1.26**	**0.68**	0.91	1.04	0.43	0.51	0.95	0.95	1.08	1.38	1.89	2.31	0.04	1.24	0.17	0.17	0.03	0.06	0.08	0.11	0.21	0.33	0.31	0.44
Parvimonas rRNA167	0.02	0.36	0.05	0.14	0.01	0.02	0.07	0.11	0.08	0.07	0.36	0.32	0.9	0.71	0.22	0.39	0.16	0.21	0.48	0.51	0.53	0.58	0.96	1.15
Prevotella colorans	**1.14**	**0.27**	0.38	0.68	0.08	0.04	0.24	0.17	0.08	0.06	0.27	0.18	0.04	0.04	0.12	0.11	0.04	0.03	0.05	0.04	0.13	0.12	0.15	0.12
Mageeibacillus indolicus	0.39	0.83	0.12	0.13	0.05	0.03	0.09	0.04	0.20	0.27	0.31	0.38	0	0	0.03	0.02	0	0	0.01	0.02	0.08	0.13	0.08	0.15
Dialister micraerophilus	0.09	0.21	0.13	0.18	0.07	0.07	0.06	0.05	0.33	0.64	0.42	0.45	0.18	0.23	0.72	0.56	0.42	0.29	0.69	0.34	0.71	0.78	0.93	0.87
Sneathia sanguinegens	0.22	0.25	0.16	0.18	0.05	0.05	0.06	0.04	0.05	0.28	0.07	0.28	0.1	0.11	0.17	0.18	0.06	0.06	0.03	0.04	0.08	0.26	0.04	0.24
Mycoplasma hominis	0	0	0	0.01	0	0	0	0	0.01	0.01	0.00	0.03	0.33	**0.22**	0.17	0.19	0.04	0.03	0.10	0.03	0.07	0.07	0.15	0.09
Bacteroidales L3 (d)	0	**0.21**	0	0.03	0.05	0.04	0.06	0.04	0.05	0.21	0.07	0.19	0.03	0	0.01	0.01	0.02	0.03	0.05	0.04	0.03	0.21	0.06	0.19
Aerococcus sanguinicola	0	0	0	0	0	0	0	0	0	0	0	0	0	0.18	0.01	0	0	0	0	0	0	0	0	0
Mobiluncus curtisii	0	0	0	0	0	0	0	0	0	0	0	0	0	0.08	0.03	0.02	0	0	0	0	0	0	0	0
Lactobacillus coleohominis	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0.03	0.07	0.02	0.02	0.04	0.02	0.74	0.34	0.57	0.29
Lactobacillus jensenii	0	0	1.29	0.55	0.01	0.01	0	0	0.67	2.10	0.50	0.27	0	0.02	1.63	0.86	0	0	0	0.01	0	0.50	0	1.71
Mobiluncus mulieris	0	0	0	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0.01	0.02	0.03	0.00	0.00	0.00	0.00	0.03	0	0.04	0.04
Lactobacillus crispatus	0	0	1.51	1.4	0.04	0.08	0.03	0.08	0.45	0.09	0.12	0.99	0	0	1.45	1.6	0.03	0.08	0.02	0.07	0.51	1.44	0.33	0.07
Lactobacillus gasseri	0	0	0.21	0.09	0.01	0.01	0	0.01	0.36	0.06	0.08	0.11	0	0	0.19	0.06	0.01	0.01	0.00	0.01	0.32	0.86	0.06	0.03
Dorea formicigenerans	0	0	0.49	0.72	0	0	0	0	0	0	0	0	0	0	0.51	0.85	0	0	0	0	0	0	0	0
Bifidobacterium scardovii	0	0	0.37	0.7	0	0	0	0	0	0	0	0	0	0	0.21	0.45	0	0	0	0	0	0	0	0
Megasphaera cerevisiae	0	0	0.49	0.35	0	0	0	0	0	0	0	0	0	0	0.42	0.28	0	0	0	0	0	0	0	0
Veillonella atypica	0	0	0.45	0.32	0	0	0	0	0	0	0	0	0	0	0.35	0.28	0	0	0	0	0	0	0	0
Megasphaera micronuciformis	0	0	0.37	0.26	0	0	0	0	0	0	0	0	0	0	0.43	0.26	0	0	0	0	0	0	0	0
Anaeroglobus geminatus	0	0	0.39	0.3	0	0	0	0	0	0	0	0	0	0	0.27	0.21	0	0	0	0	0	0	0	0
Finegoldia magna	0	0	0.45	0.14	0	0	0	0	0	0.02	0	0	0	0	0.35	0.16	0	0	0	0	0	0	0	0
Bacteroides vulgatus	0	0	0.09	0.15	0	0	0	0	0	0	0	0	0	0	0.31	0.31	0	0	0	0	0	0	0	0
Schaalia odontolytica	0	0	0.22	0.26	0	0	0	0	0	0	0	0	0	0	0.12	0.18	0	0	0	0	0	0	0	0
Prevotella buccalis	0	0	0.09	0.16	0	0	0	0	0	0	0	0	0	0	0.22	0.26	0	0	0	0	0	0	0	0
Prevotella bivia	0	0	0.09	0.12	0	0	0	0	0.002	0	0	0	0	0	0.17	0.18	0	0	0	0	0	0	0	0.11
Anoxybacillus gonensis	0	0	0.2	0.11	0	0	0	0	0	0	0	0	0	0	0.14	0.11	0	0	0	0	0	0	0	0
Prevotella melaninogenica	0	0	0.04	0.07	0	0	0	0	0	0	0	0	0	0	0.22	0.2	0	0	0	0	0	0	0	0

Notes:
(a) Pink-shaded boxes indicate where full-length 16S rRNA gene for that taxon was found in a contig (bold type). Blue-shaded boxes indicate where a joined 16S rRNA gene for that taxon was identified through overlapping of multiple contigs. Grey-shaded boxes indicate where a partial 16S rRNA gene for that taxon was found.
(b) Only hits with abundance greater than 0.001% in read-based analysis are considered.
(c) Prevotella vaginal (similar to JX871244.1 and JX871252.1) or Prevotella skin (similar to GQ047249.1) denotes taxa identified in vaginal or skin samples, respectively, from published studies.
(d) Full-length 16S rRNA gene was obtained from a single contig in Subject 3. This sequence was closest to Bacteroidales 34.053-1 found in the GenBank database.

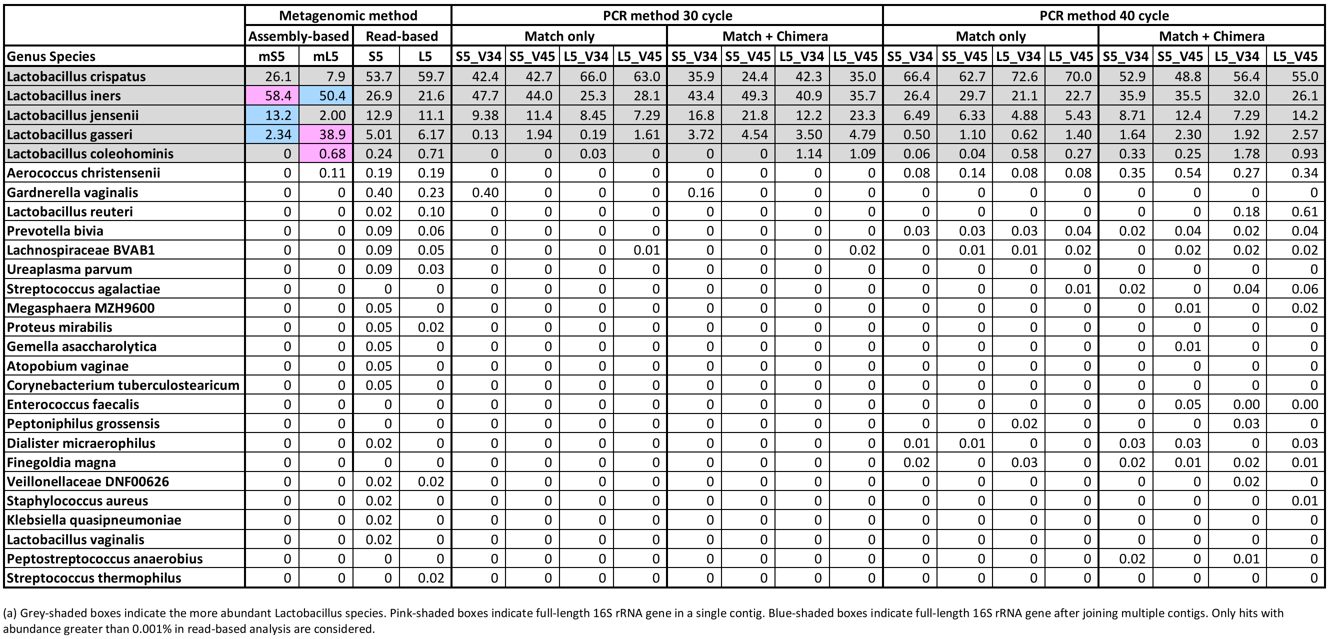
Notes: Pink-shaded boxes indicate where full-length 16S rRNA gene for that taxon was found in a contig (bold type). Blue-shaded boxes indicate where a joined 16S rRNA gene for that taxon was identified through overlapping of multiple contigs. Grey-shaded boxes indicate where a partial 16S rRNA gene for that taxon was found.
(b) Only hits with abundance>0.001% in read-based analysis are considered.
(c) Prevotella vaginal (similar to JX871244.1 and JX871252.1) or Prevotella skin (similar to GQ047249.1) denotes taxa identified in vaginal or skin samples, respectively, from published studies.
(d) Full-length 16S rRNA gene was obtained from a single contig in Subject 3. This sequence was closest to Bacteroidales 34.053–1 found in the GenBank database.

enable the inclusion of chimeric reads in the analysis. An advantage of our BlastBin tool is that it can be integrated into other existing PCR-based sequence analysis workflows, including QIIME2. Further expansion of BlastBin could include the capability of automatically amending the reference library to accommodate new OTUs found in the sample from the NCBI 16S rRNA gene reference library. In all of our cases, we found that counting chimeric reads yielded microbial profiles that were more diverse than when the chimeras were discarded.

Building a mini-database for each dataset is important for the BlastBin process. This step eliminates incorrect attribution of a chimeric read to a taxon that is not present in the sample. Likewise, omission of low-abundance taxa could incorrectly attribute the read as a chimera of others. By sampling 1000 reads in the first iteration, the probability of missing a nonchimeric read is = (1-p)^1000, where p is the abundance of a nonchimeric read. When p is 0.01, the probability of missing this nonchimeric read is 0.00043. When p = 0.005, the probability is 0.0066. To minimize the omission of “real” taxa in the mini-database, we included a final step of searching for possible reference sequences in the remaining unmatched reads. Even with an incomplete mini-database, the counts of dominant taxa generated using BlastBin were consistent (see Supplemental Methods for the case of Samples S27 and S41, analyzed using different mini-databases).

Host sequences in the samples from certain host sites can reach as high as 90%, such as was observed for many of the vaginal samples in this study. The resulting low yields of non-host sequences gave insufficient coverage of bacterial reads that could be used for assembly of all bacterial content in the samples. Another limitation of assembly-based methods is that there are no effective marker genes, other than the rRNA genes, that can be used to count contigs for microbial profiling. Most of the assembled contigs did not contain 16S rRNA genes. Nevertheless, sequence assembly of some contigs from metagenomic sequencing confirmed the presence of bacterial taxa identified through PCR-based and read-based metagenome sequencing using the 16S rRNA genes. If 16S rRNA genes are used to count abundances, misassembled contigs and contigs containing partial 16S rRNA genes also pose a challenge with regard to counting accuracy. We observed significant misassembly of reads in our study, particularly among taxa having high sequence identity, such as those found among several closely related *Lactobacillus* species. Alternatively, we used metagenomic sequencing reads directly for counting taxa to generate microbial profiles. We found for our 23 samples that the metagenomic

Table 3
Comparison of microbial profiles obtained for swab and lavage samples from Subject 5.(a)



	Metagenomic method				PCR method 30 cycle								PCR method 40 cycle							
	Assembly-based		Read-based		Match only				Match + Chimera				Match only				Match + Chimera			
Genus Species	**mS5**	**mL5**	**S5**	**L5**	**S5_V34**	**S5_V45**	**L5_V34**	**L5_V45**	**S5_V34**	**S5_V45**	**L5_V34**	**L5_V45**	**S5_V34**	**S5_V45**	**L5_V34**	**L5_V45**	**S5_V34**	**S5_V45**	**L5_V34**	**L5_V45**
Lactobacillus crispatus	26.1	7.9	53.7	59.7	42.4	42.7	66.0	63.0	35.9	24.4	42.3	35.0	66.4	62.7	72.6	70.0	52.9	48.8	56.4	55.0
Lactobacillus iners	58.4	50.4	26.9	21.6	47.7	44.0	25.3	28.1	43.4	49.3	40.9	35.7	26.4	29.7	21.1	22.7	35.9	35.5	32.0	26.1
Lactobacillus jensenii	13.2	2.00	12.9	11.1	9.38	11.4	8.45	7.29	16.8	21.8	12.2	23.3	6.49	6.33	4.88	5.43	8.71	12.4	7.29	14.2
Lactobacillus gasseri	2.34	38.9	5.01	6.17	0.13	1.94	0.19	1.61	3.72	4.54	3.50	4.79	0.50	1.10	0.62	1.40	1.64	2.30	1.92	2.57
Lactobacillus coleohominis	0	0.68	0.24	0.71	0	0	0.03	0	0	0	1.14	1.09	0.06	0.04	0.58	0.27	0.33	0.25	1.78	0.93
Aerococcus christensenii	0	0.11	0.19	0.19	0	0	0	0	0	0	0	0	0.08	0.14	0.08	0.08	0.35	0.54	0.27	0.34
Gardnerella vaginalis	0	0	0.40	0.23	0.40	0	0	0	0.16	0	0	0	0	0	0	0	0	0	0	0
Lactobacillus reuteri	0	0	0.02	0.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.18	0.61
Prevotella bivia	0	0	0.09	0.06	0	0	0	0	0	0	0	0	0.03	0.03	0.03	0.04	0.02	0.04	0.02	0.04
Lachnospiraceae BVAB1	0	0	0.09	0.05	0	0	0	0.01	0	0	0	0.02	0	0.01	0.01	0.02	0	0.02	0.02	0.02
Ureaplasma parvum	0	0	0.09	0.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Streptococcus agalactiae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.02	0	0.04	0.06
Megasphaera MZH9600	0	0	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0.02
Proteus mirabilis	0	0	0.05	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gemella asaccharolytica	0	0	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0
Atopobium vaginae	0	0	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corynebacterium tuberculostearicum	0	0	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Enterococcus faecalis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0.00	0.00
Peptoniphilus grossensis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0	0	0.03	0
Dialister micraerophilus	0	0	0.02	0	0	0	0	0	0	0	0	0	0.01	0.01	0	0	0.03	0.03	0	0.03
Finegoldia magna	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0.03	0	0.02	0.01	0.02	0.01
Veillonellaceae DNF00626	0	0	0.02	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0
Staphylococcus aureus	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01
Klebsiella quasipneumoniae	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lactobacillus vaginalis	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Peptostreptococcus anaerobius	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0.01	0
Streptococcus thermophilus	0	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(a) Grey-shaded boxes indicate the more abundant Lactobacillus species. Pink-shaded boxes indicate full-length 16S rRNA gene in a single contig. Blue-shaded boxes indicate full-length 16S rRNA gene after joining multiple contigs. Only hits with abundance greater than 0.001% in read-based analysis are considered.

(a) Grey-shaded boxes indicate the more abundant Lactobacillus species. Pink-shaded boxes indicate full-length 16S rRNA gene in a single contig. Blue-shaded boxes indicate full-length 16S rRNA gene after joining multiple contigs. Only hits with abundance>0.001% in read-based analysis are considered.

Table 4
Major *Lactobacillus* species identified in the metagenomic sequence assemblies in samples from Subject 5.

Sequence Name	Acc. Number	Species Name	Query Cov.	% ID
mS5_34910 *	NR_036982.1	*Lactobacillus iners*	98%	96.88
	NR_075051.1	*Lactobacillus gasseri*	99%	94.46
	NR_041800.1	*Lactobacillus crispatus*	97%	94.20
mS5_34776-j	NR_075051.1	*Lactobacillus gasseri*	100%	99.24
	NR_036982.1	*Lactobacillus iners*	98%	94.49
	NR_025087.1	*Lactobacillus jensenii*	99%	93.88
	CP026503.1	*Lactobacillus crispatus* AB70	100%	92.49
	NR_041800.1	*Lactobacillus crispatus* ATCC33820	97%	92.23
mS5_8225-j	NR_025087.1	*Lactobacillus jensenii*	99%	99.61
	NR_075051.1	*Lactobacillus gasseri*	100%	93.64
	CP026503.1	*Lactobacillus crispatus* AB70	100%	93.24
	NR_041800.1	*Lactobacillus crispatus* ATCC33820	97%	92.99
mL5_28841-j *	NR_036982.1	*Lactobacillus iners*	98%	97.27
	NR_075051.1	*Lactobacillus gasseri*	99%	94.71
	NR_025087.1	*Lactobacillus jensenii*	99%	93.57
mL5_32455	NR_042436.1	*Lactobacillus coleohominis*	99%	99.23
	NR_041796.1	*Lactobacillus vaginalis* ATCC49540	99%	94.97
	NR_075036.1	*Lactobacillus reuteri*	99%	94.71
	NR_104927.1	*Lactobacillus fermentum*	100%	93.46
mL5_49026 *	NR_075051.1	*Lactobacillus gasseri*	100%	96.11
	NR_036982.1	*Lactobacillus iners*	98%	94.94
	NR_041800.1	*Lactobacillus crispatus* ATCC33820	97%	94.51

*Asterisks denote possible chimeric sequences. Full-length 16S rRNA gene sequences identified in a single contig or joined from multiple overlapping contigs are denoted as j.

sequence yields for 16S rRNA gene reads ranged from 0.008% to 0.14% with a median value of 0.03%. Thus, the sequencing depth of these reads was relatively low, compared to PCR-based methods. The metagenomic sequencing method generated non-overlapping reads of about 250 nucleotides or shorter, which provided relatively poorer resolution for assignment of taxa compared to the PCR-based method, where the overlapping paired reads provided merged sequences of 400 nucleotides or longer. However, the metagenomic read-based method has fewer of the above-mentioned artifacts that are associated with the assembly-based or PCR-based methods.

We compared microbial profiles generated by metagenomic read-based methods with those generated by the PCR-based methods. We found that use of 40-cycles for generating the full-length 16S rRNA amplicons and inclusion of chimeric reads in V3V4 and V4V5 analysis provided microbial profiles that more closely resembled the corresponding metagenomic read-based profiles. By using 40 PCR cycles, as opposed to 30 cycles, and including chimeric

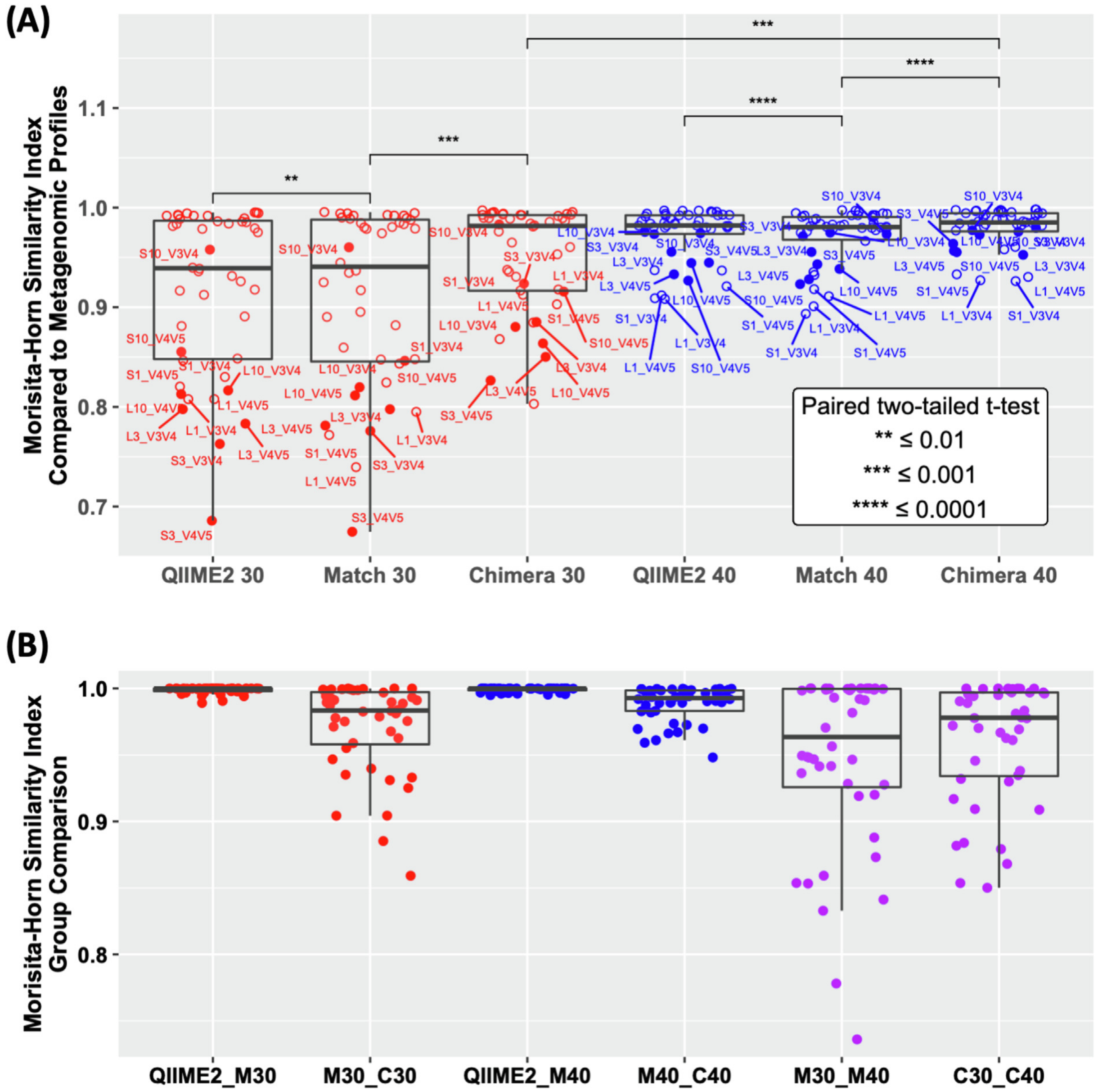


Fig. 5. Comparison of Morisita-Horn similarity indices for microbial profiles. (A) Shown are boxplots of Morisita-Horn similarity indices between the microbial profiles generated using the metagenomic read-based method and each of the indicated methods for 16S amplicons after 30 cycles (red) or 40 cycles (blue): QIIME2 method; BlastBin method considering only matched reads (Match 30, Match 40); and BlastBin method considering matched + chimeric reads (Chimera 30, Chimera 40). Data points for samples from Subjects 3 and 10 are shown as solid circles with labels. Significance was determined by paired two-tailed *t*-test: $^{**}p < 0.01$; $^{***}p < 0.001$; $^{****}p < 0.0001$. (B) Shown are boxplots of Morisita-Horn similarity indices between microbial profiles for each pair of methods, QIIME2 vs BlastBin matched only, BlastBin method matched only vs matched + chimera with 30 cycles (red) or 40 cycles (blue); and BlastBin method with 30 vs 40 cycles (purple), as indicated. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reads, it was possible to detect low-abundance taxa that were present in the metagenomic read-based profile. The 40-cycle microbial profiles with chimeric reads included also displayed higher microbial diversities.

For generating hypothesis-based models for microbiome studies with large sample sets, using PCR-based or metagenomic sequencing approaches can yield valuable insights for further studies. However, each of these approaches have limitations that need to be addressed. Our results have demonstrated that for some microbiome samples the nature and composition of the microbial community can contribute to significant chimera formation. This could skew the data interpretation without considering the amount of chimeric reads present. Overly optimistic expectations of current assembly-based metagenomic sequencing methods need to be reassessed in light of the misassembly problem, where reads could match to multiple hits and lead to inaccurate counting and assignment of taxa, particularly among closely related species. One solution might be to use longer reads, but this comes at the expense of poor quality of reads and increased cost.

In conclusion, the potential power of using microbiome profiles as tools for clinical diagnosis must consider sample-dependent reliability in applying analytical methods. To avoid the biases that may be inherent in clinical samples, it may be necessary to apply both PCR-based and metagenomic shotgun sequencing methods to identify key biomarkers of disease. When PCR-based methods are used, it may be important to include counting of chimeric reads in the microbial profile analysis. Our approach using the chimera-counting tool BlastBin could be incorporated into the analysis workflow for this purpose.

5. Ethics approval and consent to participate

This study was approved by the Institutional Review Boards of the University of Illinois at Urbana-Champaign (IRB#05079 and IRB#16789), the University of Illinois at Chicago (IRB#2014-0527), and the Human Research Protection Office of the US Army Medical Research and Materiel Command (HRPO#A-18934). Informed consent was obtained from all study participants prior to sample collection.

6. Availability of data and materials

All joined sequence reads used in the current study are available in the NCBI Sequence Read Archive SRA: SRP184243 (BioProject: PRJNA521059).

7. Funding

This work was supported in part by the Department of Army DARPA grant #W911NF-14–1-0311 (to B.A.W.), and the Jenner Family Research Fellowship and Dawn M. Carlson MD, MPH Research Fellowship for Women in Science (to D.M.).

8. Authors' contributions

B.A.W. and M.H. designed and conceived the experiments. B.L. M. and P.D.T. recruited human subjects, collected samples, and assisted with clinical assessment and interpretation of data. M.P.-A. and M.H. performed the experiments. M.H. and D.M. performed scripting, bioinformatic analysis and algorithm development. M.H., D.M. and B.A.W. analyzed data and interpreted the results, and generated tables and figures. B.A.W., M.H. and B.L.M. provided reagents, materials, analytical tools, and support. M.H. and B.A.W. wrote the paper. All authors read, edited, and approved the final version of the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank Alvaro Hernandez and Mark Band of the UIUC Roy J. Carver Biotechnology Center for their assistance with sequencing technologies, Samadh Ravangard for clinical assistance with initial subject recruitment and sample collection, and Bill O'Brien for thoughtful discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.08.050.

References

[1] Integrative, H.M.P.R.N.C., The Integrative Human Microbiome Project. Nature, 2019. 569(7758): pp. 641–648.

[2] Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, et al. The vaginal microbiome and preterm birth. Nat Med 2019;25(6):1012–21.

[3] Ma B, Forney LJ, Ravel J. Vaginal microbiome: rethinking health and disease. Annu Rev Microbiol 2012;66(1):371–89.

[4] McClelland RS, Lingappa JR, Srinivasan S, Kinuthia J, John-Stewart GC, Jaoko W, et al. Evaluation of the association between the concentrations of key vaginal bacteria and the increased risk of HIV acquisition in African women from five cohorts: a nested case-control study. Lancet Infect Dis 2018;18(5):554–64.

[5] Smith SB, Ravel J. The vaginal microbiota, host defence and reproductive physiology. J Physiol 2017;595(2):451–63.

[6] McDonald, L.C., et al., Clinical Practice Guidelines for Clostridium difficile Infection in Adults and Children: 2017 Update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA). Clin Infect Dis, 2018. 66(7): p. e1-e48.

[7] Staley C, Kaiser T, Vaughn BP, Graiziger C, Hamilton MJ, Kabage AJ, et al. Durable long-term bacterial engraftment following encapsulated fecal microbiota transplantation to treat clostridium difficile infection. mBio 2019;10(4). https://doi.org/10.1128/mBio.01586-19.

[8] Costello SP, Hughes PA, Waters O, Bryant RV, Vincent AD, Blatchford P, et al. Effect of fecal microbiota transplantation on 8-week remission in patients with ulcerative colitis: a randomized clinical trial. JAMA 2019;321(2):156. https://doi.org/10.1001/jama.2018.20046.

[9] Goll R, Johnsen PH, Hjerde E, Diab J, Valle PC, Hilpusch F, et al. Effects of fecal microbiota transplantation in subjects with irritable bowel syndrome are mirrored by changes in gut microbiome. Gut Microbes 2020;12(1):1794263. https://doi.org/10.1080/19490976.2020.1794263.

[10] Tan P et al. Fecal microbiota transplantation for the treatment of inflammatory bowel disease: an update. Front Pharmacol 2020;11:574533.

[11] Arthur JC, Jobin C. The complex interplay between inflammation, the microbiota and colorectal cancer. Gut Microbes 2013;4(3):253–8.

[12] Ma C, Han M, Heinrich B, Fu Q, Zhang Q, Sandhu M, et al. Gut microbiome-mediated bile acid metabolism regulates liver cancer via NKT cells. Science 2018;360(6391):eaan5931. https://doi.org/10.1126/science:aan5931.

[13] Helmink BA, Khan MAW, Hermann A, Gopalakrishnan V, Wargo JA. The microbiome, cancer, and cancer therapy. Nat Med 2019;25(3):377–88.

[14] Mullard A. Oncologists tap the microbiome in bid to improve immunotherapy outcomes. Nat Rev Drug Discov 2018;17(3):153–5.

[15] Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews MC, Karpinets TV, et al. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. Science 2018;359(6371):97–103.

[16] Matson V, Fessler J, Bao R, Chongsuwat T, Zha Y, Alegre M-L, et al. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. Science 2018;359(6371):104–8.

[17] Routy B, Le Chatelier E, Derosa L, Duong CPM, Alou MT, Daillère R, et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. Science 2018;359(6371):91–7.

[18] Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med 2018;24(4):392–400.

[19] Neville BA, Forster SC, Lawley TD. Commensal Koch's postulates: establishing causation in human microbiota research. Curr Opin Microbiol 2018;42:47–52.

[20] Gharaibeh RZ, Jobin C. Microbiota and cancer immunotherapy: in search of microbial signals. Gut 2019;68(3):385–8.

[21] 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). https://doi.org/10.1186/s40168-017-0267-5.

[22] D'Amore R et al. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics 2016;17:55.

[23] Jumpstart Consortium Human Microbiome Project Data Generation Working, G., Evaluation of 16S rDNA-based community profiling for human microbiome research. PLoS One, 2012. 7(6): p. e39315.

[24] Tremblay J et al. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol 2015;6:771.

[25] D'Argenio V. Human Microbiome Acquisition and Bioinformatic Challenges in Metagenomic Studies. Int J Mol Sci 2018;19(2):383. https://doi.org/10.3390/ijms19020383.

[26] Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human microbiome project. Nature 2017;550(7674):61–6.

[27] Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 2017;35(9):833–44.

[28] Hillmann, B., et al., Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems, 2018. 3(6).

[29] Schriefer AE, Cliften PF, Hibberd MC, Sawyer C, Brown-Kennerly V, Burcea L, et al. A multi-amplicon 16S rRNA sequencing and analysis method for improved taxonomic profiling of bacterial communities. J Microbiol Methods 2018;154:6–13.

[30] Tettamanti Boshier, F.A., et al., Complementing 16S rRNA Gene Amplicon Sequencing with Total Bacterial Load To Infer Absolute Species Concentrations in the Vaginal Microbiome. mSystems, 2020. 5(2).

[31] Bonk F, Popp D, Harms H, Centler F. PCR-based quantification of taxa-specific abundances in microbial communities: Quantifying and avoiding common pitfalls. J Microbiol Methods 2018;153:139–47.

[32] Fujiyoshi S, Muto-Fujita A, Maruyama F. Evaluation of PCR conditions for characterizing bacterial communities with full-length 16S rRNA genes using a portable nanopore sequencer. Sci Rep 2020;10(1):12580.

[33] Haile, S., et al., Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. Nucleic Acids Res, 2019. 47(2): p. e12.

[34] Smyth RP, Schlub TE, Grimm A, Venturi V, Chopra A, Mallal S, et al. Reducing chimera formation during PCR amplification to ensure accurate genotyping. Gene 2010;469(1-2):45–51.

[35] Nearing, J.T., et al., Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ, 2018. 6: p. e5364.

[36] Schloss PD, Gevers D, Westcott SL, Gilbert JA. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE 2011;6(12):e27310.

[37] Edgar, R.C., et al., UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 2011. 27(16): p. 2194-200.

[38] Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 2011;21(3):494–504.

[39] Nilsson RH, Abarenkov Kessy, Veldre Vilmar, Nylinder Stephan, De Wit Pierre, Brosché Sara, et al. An open source chimera checker for the fungal ITS region. Mol Ecol Resour 2010;10(6):1076–81.

[40] Huber T, Faulkner G, Hugenholtz P. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics 2004;20(14):2317–9.

[41] Xue Z, Kable ME, Marco ML, Suen G. Impact of DNA sequencing and analysis methods on 16S rRNA gene bacterial community analysis of dairy products. mSphere 2018;3(5). https://doi.org/10.1128/mSphere.00410-18.

[42] Bokulich, N.A., et al., mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking. mSystems, 2016. 1(5).

[43] Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 2013;79(17):5112–20.

[44] Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol 2016;34(9):942–9.

[45] Chen, S., et al., fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 2018. 34(17): p. i884-i890.

[46] Hyman RW, Fukushima M, Diamond L, Kumm J, Giudice LC, Davis RW. Microbes on the human vaginal epithelium. Proc Natl Acad Sci U S A 2005;102(22):7952–7.

[47] Fredricks DN, Fiedler TL, Marrazzo JM. Molecular identification of bacteria associated with bacterial vaginosis. N Engl J Med 2005;353(18):1899–911.

[48] Zhou X, Brown CJ, Abdo Z, Davis CC, Hansmann MA, Joyce P, et al. Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. ISME J 2007;1(2):121–33.

[49] Zozaya-Hinchliffe M, Martin DH, Ferris MJ. Prevalence and abundance of uncultivated Megasphaera-like bacteria in the human vaginal environment. Appl Environ Microbiol 2008;74(5):1656–9.

[50] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9(4):357–9.

[51] Li, D., et al., MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 2015. 31(10): p. 1674-6.

[52] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30(14):2068–9.

[53] Camacho C et al. BLAST+: architecture and applications. BMC Bioinf 2009;10:421.

[54] Callahan, B.J., et al., High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res, 2019. 47(18): p. e103.

[55] Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA 2011;108(Supplement_1):4516–22.

[56] Jovel J et al. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Front Microbiol 2016;7:459.

[57] Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman JA, et al. Unlocking the potential of metagenomics through replicated experimental design. Nat Biotechnol 2012;30(6):513–20.

[58] Laudadio I, Fulci V, Palone F, Stronati L, Cucchiara S, Carissimi C. Quantitative assessment of shotgun metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome. OMICS 2018;22(4):248–54.

[59] Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun 2016;469(4):967–77.

[60] Frank JA, Reich CI, Sharma S, Weisbaum JS, Wilson BA, Olsen GJ. Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. Appl Environ Microbiol 2008;74(8):2461–70.

[61] Horn HS. Measurement of 'Overlap' in comparative ecological studies. Am Nat 1966;100(914):419–24.