An Improved Dual-Indexing Approach for Multiplexed 16S rRNA Gene Sequencing on the Illumina MiSeq Platform

Douglas W. Fadrosh[1], Bing Ma[1], Pawel Gajer[1], Sandra Ott[1], Naomi Sengamalay[1], Rebecca M. Brotman[2] and Jacques Ravel[1,*]

[1] Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore MD
[2] Institute for Genome Sciences, Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore MD

* To whom correspondence should be addressed: jravel@som.umaryland.edu

Abstract

Background: Taking advantage of affordable high-throughput next-generation sequencing technologies to characterize microbial community composition often requires the development of novel methods to overcome technical limitations inherent to these platforms. Sequencing low sequence diversity libraries such as 16S rRNA amplicons has been problematic on the Illumina Miseq platform and often generates low quality sequences.

Results: Here we present an improved dual-indexing amplification and sequencing approach to assess the composition of microbial communities from clinical samples using the V3-V4 region of the 16S rRNA gene on the Illumina MiSeq platform.

Conclusion: This approach addresses the known “low sequence diversity” issue associated with sequencing 16S rRNA gene amplicons on the MiSeq platform and provides a flexible and cost-effective option.

Background

The development of methods to detect fastidious or non-cultivable organisms through amplification and determination of the sequence of conserved genes, or culture-independent profiling, has precipitated a revolution in biology. It was recognized decades ago that the number of microbes seen on direct staining of environmental or human samples often exceeded by many orders of magnitude the number that could be cultured (termed "the great plate-count anomaly") [1]. Culture-independent profiling of bacterial communities generally relies on amplification and sequencing of the 16S ribosomal RNA (rRNA) gene and has greatly increased appreciation for the complexity hidden in even seemingly simple microbial consortia. Advancement in next-generation sequencing technologies, both in terms of throughput and sequence read length and accuracy, has had a major impact in the field by enabling large numbers of samples to be examined at greater depth.

The Illumina Miseq platform provides researchers with a scalable, high-throughput and streamlined sequencing platform to survey community composition from clinical and environmental samples. However, known limitations with the MiSeq platform associated with the sequencing of low sequence diversity samples has hampered its potential to sequence marker gene amplicon libraries such as 16S rRNA gene amplicons. The “low sequence diversity” issue arises in the first twelve cycles of a MiSeq 16S rRNA gene amplicons sequencing run during which successful cluster generation and phasing/pre-phasing steps are dependent on heterogeneous base composition of targeted amplicons. Because of the nature of the 16S rRNA gene, amplicons libraries are highly homogenous and are required to be co-sequenced with a heterogeneous control library comprised of phage PhiX (random library), normally combined 1:1 with the amplicon sample. This increases the quality of the sequencing reads but at the expense of having half of the sequence reads lost to a non-targeted template. Despite these limitations, the MiSeq sequencing platform’s data yield and

250 bp paired-end read sequencing technologies continue to be attractive to researchers and enables the high resolution characterization of microbial communities with effective read lengths comparable to that obtained on the Roche/454 pyrosequencing platform, for a fraction of the cost.

The most widely used 16S rRNA-based MiSeq sequencing strategies include single- [2, 3] or dual-indexing (http://www.mothur.org/wiki/MiSeq_SOP) approaches targeting the V4 hypervariable region of the 16S rRNA gene. These strategies leverage custom 16S rRNA gene PCR primers that enable multiplexing of samples and direct sequencing on the MiSeq instrument, but do not fully maximize the potential, or directly address the known limitations of the sequencing technology. The single-indexing strategy requires large numbers of barcoded primers (one per sample) and custom sequencing primers, increasing costs and limiting flexibility. While the current dual-indexing approach reduces the number of primers needed, it still requires the addition of a control library to modulate the sequence heterogeneity of the library being sequenced.

Here, we address the technical limitations of the MiSeq platform for 16S rRNA gene sequencing and present a more flexible and cost-effective approach to sequence barcoded 16S rRNA amplicons by leveraging dual-indexed primers with built-in heterogeneity spacers that reduce the required amount of control library to less than 20% of the total sequencing library.

Methods

V3-V4 Amplification and Sequencing Strategy

The 16S rRNA gene is well characterized in most known bacteria and consists of nine hypervariable regions flanked by regions of more conserved sequence. To maximize the

effective length of the MiSeq's 250bp paired-end sequencing reads, a region of approximately 469 bp encompassing the V3 and V4 hypervariable regions of the 16S rRNA gene was targeted for sequencing. This region provides ample information for taxonomic classification of microbial communities from specimens associated with human microbiome studies and was used by the Human Microbiome Project [4].

To amplify and sequence the V3-V4 hypervariable regions of the 16S rRNA gene, primers were designed that contained 1) linker sequence allowing amplicons to bind to the flow cell and be sequenced using the standard Illumina HP10 or HP11 sequencing primers; 2) a 12 bp autocorrecting index sequence from Caporaso et al. [2]; 3) a 0-7 bp "heterogeneity spacer"; and 4) 16S rRNA gene-specific sequence (**Figure 1A**, **Additional File 1**). Genomic DNA extracted from clinical vaginal and stool samples were amplified, normalized using the SequalPrep Normalization Kit (Life Technologies, Carlsbad, CA) and pooled prior to sequencing on the MiSeq platform (see **Additional File 2** for detailed description of the methods). Samples were collected under several protocols approved by the Institutional Review Board at the University of Maryland School of Medicine. All samples were de-identified. The amplicon pool was prepared for sequencing with AMPure XT beads (Beckman Coulter Genomics, Danvers, MA) while the size and quantity of the amplicon library were assessed on the LabChip GX (Perkin Elmer, Waltham, MA) and with the Library Quantification Kit for Illumina (Kapa Biosciences, Woburn, MA), respectively. PhiX Control library (v3) (Illumina, San Diego, CA) was combined with the amplicon library in a 1:4 (20%) ratio and clustered to a density of ~570. The libraries were sequenced using a hard-coded matrix file (**Additional File 3**) on a 250 bp paired-end run on the MiSeq instrument (Illumina, San Diego, CA) using the standard Illumina sequencing primers (**Figure 1A**). The use of an in-line indexing strategy eliminates the need for a third index read. Sequencing data was available within approximately 48 h.

Sequence data analysis

Image analysis, base calling and data quality assessment was performed on the MiSeq instrument. Approximately 9.5 million sequence reads from clinical samples were pre-processed to remove low quality base calls and manipulated to generate files that could be further processed using QIIME (version 1.6.0) [5] as described in **Figure 1C**. Briefly, the index sequence contained in the first 12 bp of each paired-end read was extracted. The two indexes were concatenated and subsequently appended to the beginning of each read of the mate pair to form a new 24 bp barcode unique to each sample. Additional read pre-processing included 1) removal of primer sequence, 2) truncation of reads not having an average quality of 20 over a 30 bp sliding window based on the phred algorithm [6, 7] implemented previously [8, 9], 3) removal of trimmed reads having less than 75% of their original length, and 4) removal of the mate of reads that were discarded for having less than 75% original length resulted in nearly 94% of reads being retained. Samples processed consisted of 10 vaginal and 10 rectal swabs. QIIME (version 1.6.0) [5] was used for all sequence processing steps, including quality trimming, demultiplexing, and taxonomic assignments. Additional quality trimming in QIIME were performed using the following criteria: 1) truncate sequence before 3 consecutive low quality bases and re-evaluate for length, 2) no ambiguous base calls, and 3) minimum sequence length of 150 bp after trimming, 4) remove sequences with less than 60% identity to a pre-built Greengenes database of 16S rRNA gene sequences (Oct, 2012 version) [10]. Further data processing included denoising by clustering similar sequences with less than 3% dissimilarity using USEARCH [11] and *de novo* chimera detection was conducted in UCHIME v5.1 [12]. The taxonomic ranks were assigned to each sequence using Ribosomal Database Project (RDP) Naïve Bayesian Classifier v.2.2 [13], using 0.8 confidence values as the cutoff to the Greengenes database.

Results and Discussion

The MiSeq system provides a powerful sequencing platform for the rapid, high throughput and in-depth characterization of microbial community composition using 16S rRNA gene amplicon sequencing. Its adoption has been slow mainly because of limitations inherent to the technology. Two critical software processing steps for the generation of high quality data on the Miseq, cluster identification and phasing/prephasing rate determination, require a balanced base composition through at least cycle 12 of the run. With this requirement, low sequence diversity 16S rRNA gene amplicon libraries do not sequence well on the MiSeq and the resulting data is of significantly lower overall quality than a more random library (i.e. a metagenomic sample). This problem is even further compounded in clinical and environmental samples that are dominated by one or very few types of bacteria. Current methods for sequencing 16S rRNA gene amplicons on the MiSeq require a 1:1 mixture of PhiX Control library v3 and amplicon library to modulate the overall sample base composition to help facilitate a successful sequencing run.

The dual-indexing amplification strategy, combined with the heterogeneity spacer design presented here addresses this limitation by providing a much more balanced base composition not only through the first 12 cycles, but the entire duration of the run, increasing the overall quality of the sequence data (**Figure S1**). In designing the primer system where the first 12 bases sequenced in each read are the in-line index, it is possible to select index combinations that ensure an equal proportion of each base throughout the first 12 cycles of the run. This approach also allows for multiplexing large numbers of samples at a reduced initial investment. Here, we could process 576 samples on a single sequencing run with the combination of 24 forward and 24 reverse primers. The addition of the 0-7 bp “heterogeneity spacer” between the index sequence and the 16S rRNA gene-specific sequence allows the

16S rRNA gene portion of the amplicons from an equal proportion of samples to be sequenced out of phase, further dampening the effect of the “low sequence diversity” issue associated with the MiSeq platform (**Figure 1B**). This feature dramatically reduces the ratio of PhiX Control library (v3) to amplicon library as overall sample base composition is more heterogenous. Currently, a 4:1 amplicon:PhiX mixture has been tested, however lower amounts of PhiX control library can be used and should produce reads of similar overall quality. This improved strategy increases read quality as well as the number of usable reads by up to 30% or more. Taxonomic assignments of 10 vaginal and 10 rectal swabs sequenced in this run are described in **Figure 2**, while the entire run quality is show in **Figure S1**. The run generated 8,332,670 reads that passed the stringent quality criteria described above (566,407 [6.02%] were filtered out in the preprocessing steps and 514,615 (5.47%) were filtered out during QIIME quality control steps). After trimming the adaptor and primer sequences, read 1 was 219 bp (SD 6 bp) and read 2 was 207 bp (SD 18 bp). A total of 352 samples were processed on the run providing an average of 15,676 reads per samples.

To process the dual-indexed sequencing reads, we specifically designed a workflow to process the two read files individually allowing for the downstream assembly or paired-read stitching, depending on the degree of mate-pair overlap. The advantage here lies in the ability to 1) maximize the information from paired-end reads that improves taxonomic assignment; 2) avoid the erroneous taxonomic assignments that assembling mate paired reads prior to demultiplexing would yield, because of the internal region of the stitched reads is often of lower quality score; 3) allow for a strict yet more flexible quality control should the sequence quality of one read is substantially different from its mate.

Current methods for sequencing 16S rRNA gene amplicons on the MiSeq instrument use custom sequencing primers complementary to the universal 16S rRNA gene universal primer [2]. This custom setup creates potential problems when trying to multiplex samples on

the same run that target different regions of the 16S rRNA gene. Our approach uses the standard Illumina HP10 and HP11 sequencing primers allowing this system to be easily adapted to target any hypervariable region as well as the ability to multiplex amplicons from different regions, or any other samples using the standard sequencing primers, on a single sequencing run.

Supporting data

The sequence data set generated for this project and supporting the results of this article are available from the sequence read archive (SRA) under BioProject PRJNA203369 (SRA082708 and SRP023530).

Additional Files

Additional File 1. Table listing the primer sequences

Additional File 2. Detailed methods and protocols

Additional File 3. Matrix file for the MiSeq instrument (.xml format)

Figure S1. Quality and base composition assessment of amplification approaches. (A) Average quality plot of a dual-indexed 16S rRNA gene amplicon library sequenced on a paired-end 250 bp MiSeq run using a hard coded matrix file, a cluster density of ~570, and a PhiX Control Library (v3) spike-in of ~20%. (B) Base composition plot of the MiSeq run from library shown in panel A. (C) Base composition plot from a 250bp paired-end MiSeq run prepared from a 16S rRNA gene amplicons pool that employed the strategy described by Caporaso et al. [2].

Competing interests

The authors declare that they have no competing interests

Author's contributions

DF and JR designed the study. DF, SO and NS performed the amplification and sequencing. DF, BM, PG, JR performed analyses. DF, BM, RMB and JR contributed to manuscript preparation.

Acknowledgements

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases grant of the National Institutes of Health under Awards Number UH2AI083264, U19AI084044, R33AI078798 and K01AI080974. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

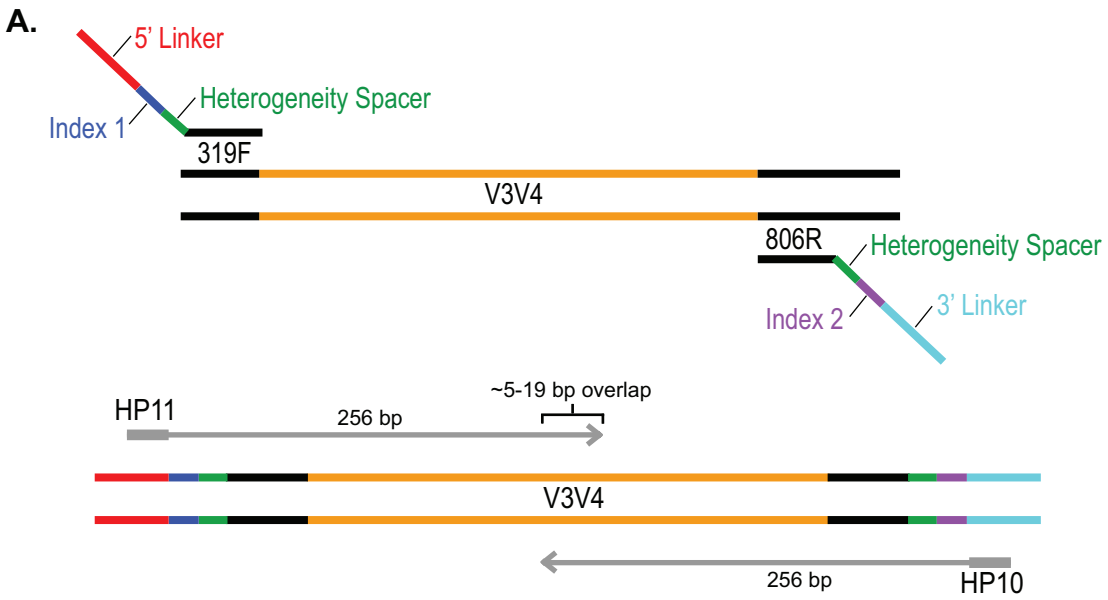
Literature cited

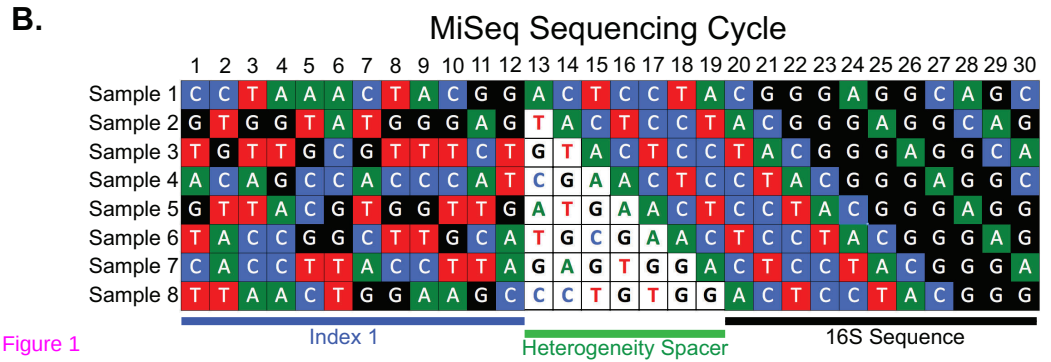
1. Staley JT, Konopka A: **Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats**. *Annual review of microbiology* 1985, **39**:321-346.
2. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M *et al*: **Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms**. *ISME J* 2012, **6**(8):1621-1624.
3. Illumina: **High-speed multiplexed 16S microbial sequencing on the MiSeq System**. In., vol. Application Note: Sequencing; 2012.
4. Consortium HMP: **Structure, function and diversity of the healthy human microbiome**. *Nature* 2012, **486**(7402):207-214.
5. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI *et al*: **QIIME allows analysis of high-throughput community sequencing data**. *Nature methods* 2010, **7**(5):335-336.
6. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment**. *Genome Res* 1998, **8**(3):175-185.
7. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 1998, **8**(3):186-194.
8. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
9. [https://github.com/lh3/seqtk]

10. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P: **An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea**. *Isme J* 2012, **6**(3):610-618.
11. Edgar RC: **Search and clustering orders of magnitude faster than BLAST**. *Bioinformatics* 2010, **26**(19):2460-2461.
12. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R: **UCHIME improves sensitivity and speed of chimera detection**. *Bioinformatics* 2011, **27**(16):2194-2200.
13. 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):5261-5267.

Figure 1. Dual-indexed 16S rRNA gene amplification, sequencing and data analysis strategy. **(A)** Dual-indexed PCR amplification primers targeting the V3-V4 hypervariable region of the 16S rRNA gene contain a heterogeneity spacer region and linker sequence and are optimized for sequencing on the Illumina Miseq. Using this approach enables sequencing using the standard Illumina HP10 and HP11 sequencing primers allowing for additional sequencing flexibility. **(B)** Schematic showing the first thirty sequencing cycles of eight mock amplicons prepared using the dual-indexed approach. This diagram illustrates how the index sequence and heterogeneity spacer (colored letters, white background) helps to alleviate the "low sequence diversity" issue associated with the MiSeq platform by creating a more even base composition at each cycle of the run. **(C)** Flow diagram outlining the sequence data analysis process.

Figure 2. Taxonomic assignment of clinical samples. Quality filtered sequence reads from ten representative vaginal **(A)** and stool samples **(B)** were taxonomically classified using QIIME. The distribution of taxa obtained for the vaginal samples at the genus level, and stool samples at the order level are consistent with known community taxonomic classifications for these sample types.


A.
5' Linker
Heterogeneity Spacer
Index 1
319F
V3V4
806R
Heterogeneity Spacer
Index 2
3' Linker
~5-19 bp overlap
HP11
256 bp
V3V4
256 bp
HP10


B.
MiSeq Sequencing Cycle
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Sample 1 C C T A A A C T A C G G A C T C C T A C G G G A G G C A G C
Sample 2 G T G G T A T G G G A G T A C T C C T A C G G G A G G C A G
Sample 3 T G T T G C G T T T C T G T A C T C C T A C G G G A G G C A
Sample 4 A C A G C C A C C C A T C G A A C T C C T A C G G G A G G C
Sample 5 G T T A C G T G G T T G A T G A A C T C C T A C G G G A G G
Sample 6 T A C C G G C T T G C A T G C G A A C T C C T A C G G G A G
Sample 7 C A C C T T A C C T T A G A G T G G A C T C C T A C G G G A
Sample 8 T T A A C T G G A A G C C C C T G T G G A C T C C T A C G G G
Index 1
Heterogeneity Spacer
16S Sequence

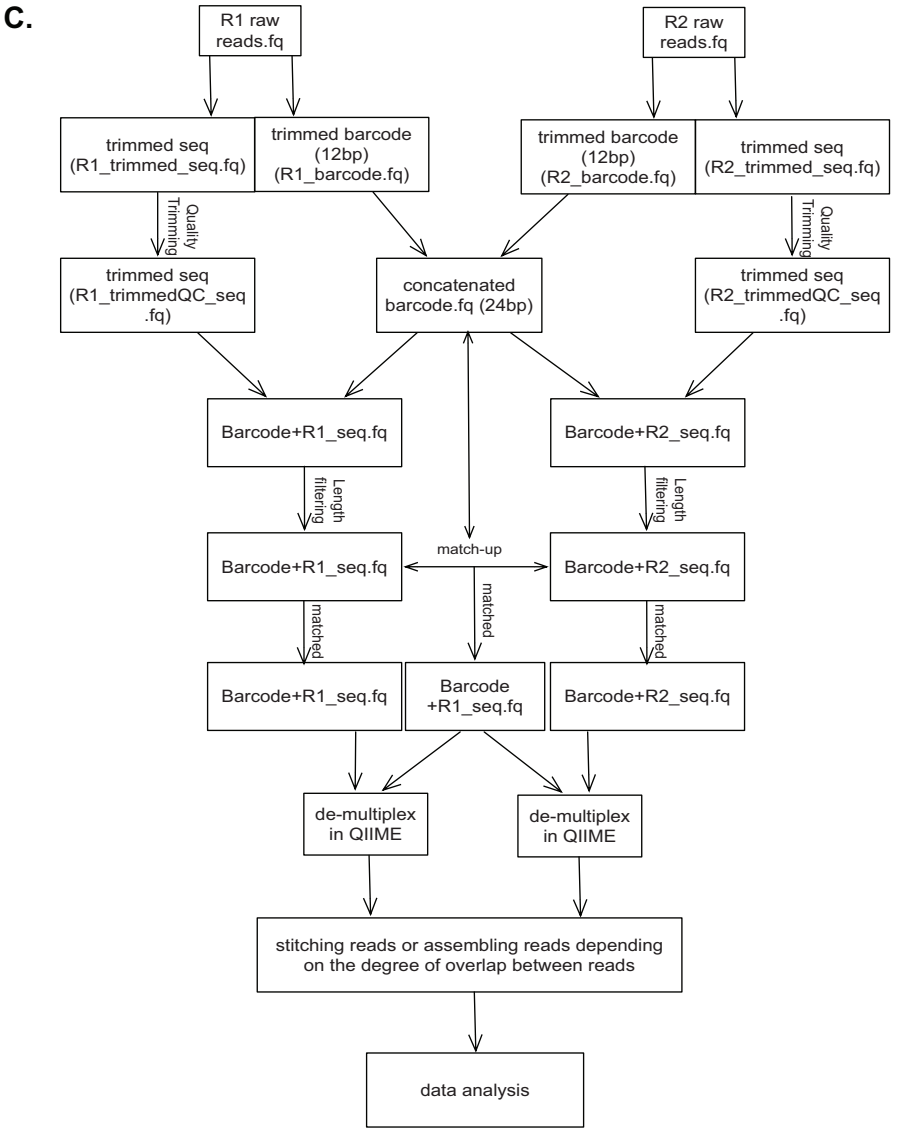

C.
R1 raw reads.fq
R2 raw reads.fq
trimmed seq (R1_trimmed_seq.fq)
trimmed barcode (12bp) (R1_barcode.fq)
trimmed barcode (12bp) (R2_barcode.fq)
trimmed seq (R2_trimmed_seq.fq)
Quality Trimming
Quality Trimming
trimmed seq (R1_trimmedQC_seq .fq)
concatenated barcode.fq (24bp)
trimmed seq (R2_trimmedQC_seq .fq)
Barcode+R1_seq.fq
Barcode+R2_seq.fq
Length filtering
Length filtering
Barcode+R1_seq.fq
match-up
Barcode+R2_seq.fq
matched
matched
matched
Barcode+R1_seq.fq
Barcode +R1_seq.fq
Barcode+R2_seq.fq
de-multiplex in QIIME
de-multiplex in QIIME
stitching reads or assembling reads depending on the degree of overlap between reads
data analysis

Figure 1

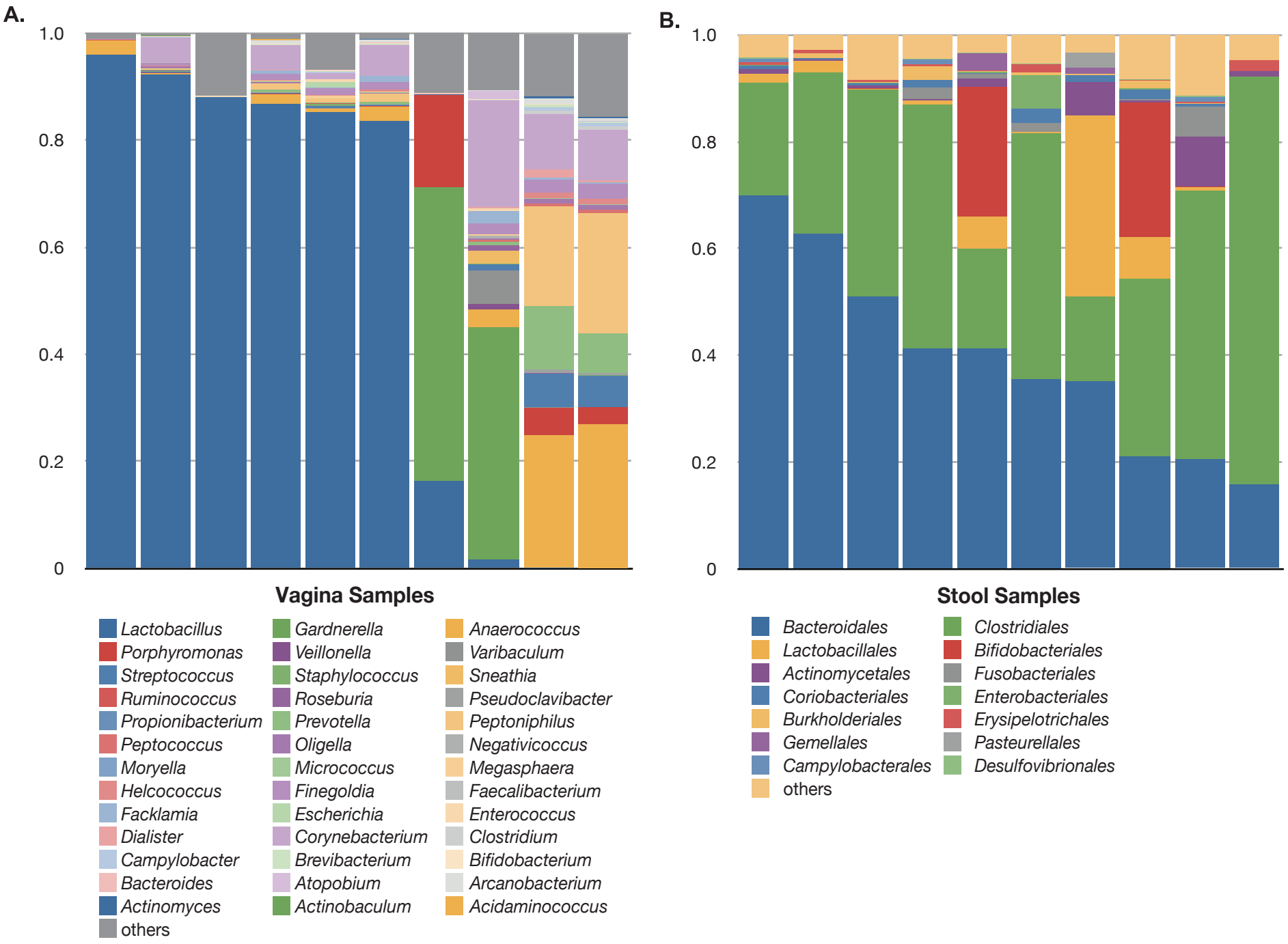

A.
1.0
0.8
0.6
0.4
0.2
0
Vagina Samples
Lactobacillus
Porphyromonas
Streptococcus
Ruminococcus
Propionibacterium
Peptococcus
Moryella
Helcococcus
Facklamia
Dialister
Campylobacter
Bacteroides
Actinomyces
others
Gardnerella
Veillonella
Staphylococcus
Roseburia
Prevotella
Oligella
Micrococcus
Finegoldia
Escherichia
Corynebacterium
Brevibacterium
Atopobium
Actinobaculum
Anaerococcus
Varibaculum
Sneathia
Pseudoclavibacter
Peptoniphilus
Negativicoccus
Megasphaera
Faecalibacterium
Enterococcus
Clostridium
Bifidobacterium
Arcanobacterium
Acidaminococcus
B.
1.0
0.8
0.6
0.4
0.2
0
Stool Samples
Bacteroidales
Lactobacillales
Actinomycetales
Coriobacteriales
Burkholderiales
Gemellales
Campylobacterales
others
Clostridiales
Bifidobacteriales
Fusobacteriales
Enterobacteriales
Erysipelotrichales
Pasteurellales
Desulfovibrionales

Figure 2

Additional files provided with this submission:

Additional file 1: Additional File 1.txt, 5K
http://www.microbiomejournal.com/imedia/1685499519101071/supp1.txt
Additional file 2: Additional File 2.docx, 159K
http://www.microbiomejournal.com/imedia/1265255424101071/supp2.docx
Additional file 3: Additional File 3.xml, 3K
http://www.microbiomejournal.com/imedia/1120502180101071/supp3.xml
Additional file 4: Figure S1.pdf, 245K
http://www.microbiomejournal.com/imedia/1025014752101071/supp4.pdf