Title: Ultra-high throughput multiplexing and sequencing of >500 bp amplicon regions on the Illumina HiSeq 2500 platform

Authors:

Johanna B. Holm[1], Michael S. Humphrys[1], Courtney K. Robinson[1], Matthew L. Settles[2], Sandra Ott[1], Li Fu[1], Hongqiu Yang[1], Pawel Gajer[1], Xin He[3], Elias McComb[1], Patti E Gravitt[4], Khalil G. Ghanem[5], Rebecca M. Brotman[1], Jacques Ravel[1]*

Affiliations:

[1] Institute for Genome Sciences and Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, Maryland.

[2] University of California, Davis Genome Center, Davis, California.

[3] Department of Epidemiology and Biostatistics, University of Maryland, School of Public Health, College Park, Maryland.

[4] Milken Institute School of Public Health, George Washington University, Washington, DC.

[5] Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD

*Corresponding author: jravel@som.umaryland.edu

Abstract

Amplification, sequencing and analysis of the 16S rRNA gene affords characterization of microbial community composition. As this tool has become more popular and projects have grown in size and scope, greater sample multiplexing is becoming necessary while maintaining high quality sequencing. Here, modifications to the Illumina HiSeq 2500 platform are described that afford greater multiplexing and 300 bp paired-end reads of higher quality than produced by the current Illumina MiSeq platform. To improve the feasibility and flexibility of this method, a 2-Step PCR amplification protocol is also described that allows for targeting of different amplicon regions, thus improving amplification success from low bacterial bioburden samples.

Importance

Amplicon sequencing has become a popular and widespread tool for surveying microbial communities. Lower overall costs associated with higher throughput sequencing have made it a widely-adopted approach, especially for larger projects which necessitate higher sample multiplexing to eliminate batch effect and reduced time to acquire data. The method for amplicon sequencing on the Illumina HiSeq 2500 platform described here provides improved multiplexing capabilities while simultaneously producing greater quality sequence data and lower per sample cost relative to the Illumina MiSeq platform, without sacrificing amplicon length. To make this method more flexible to various amplicon targeted regions as well as improve amplification from low biomass samples, we also present and validate a 2-Step PCR library preparation method.

Introduction

The introduction of the Illumina HiSeq and MiSeq platforms has allowed for the characterization of microbial community composition and structure by enabling in-depth, paired-end sequencing of amplified fragments of the 16S rRNA gene. The Illumina MiSeq instrument produces paired sequence reads up to 300 bp long. However, low amplicon sequence diversity often results in reduced sequence read quality because of the homogenous signals generated across the entire flow cell [1]. The co-sequencing of PhiX DNA can alleviate the problem, but reduces the overall sequence read throughput and multiplexing options. Alternatively, the addition of a "heterogeneity spacer" in the amplification primer offsets the sequence reads by up to 7 bases and simultaneously increases multiplexing capacity by lowering the amount of PhiX control DNA to ~5% [1]. Lower overall costs associated with higher throughput sequencing have made it a widely-adopted approach, especially for larger projects which necessitate higher sample multiplexing to eliminate batch effect and reduced time to acquire data. The high-throughput Illumina HiSeq 2500 platform offers a remedy to this issue but can currently only be used on short amplicons (i.e. the 16S rRNA gene V4 region) due to limitations in read length (maximum of 250 bp PE in Rapid Run Mode on a HiSeq 2500 instrument).

We present here a method that produces high-quality 300 bp paired-end reads from up to 1,568 samples per lane on a HiSeq 2500 instrument set to Rapid Run Mode. To make this method feasible and flexible in sequencing different amplicon regions, libraries are prepared using an improved version of a previously published 1-Step PCR method [1], by using a 2-Step PCR approach. In the 1-Step PCR method, fusion primers that contain both the target amplification primer, the heterogeneity spacer, the barcode, and the sequencing primers have been used to amplify a ready-to-sequence amplicon. However, primers ranging from 90-97 bp in length are expensive, can be subject to degradation leading to poor or no amplification from low biomass samples, and are limited to the targeted amplicon region. The 2-Step PCR library preparation procedure described here is relatively more flexible and improves amplification from low biomass samples because it uses short primers and a small anchor sequence to target the amplicon region of interest in the first amplification step. The barcode, heterogeneity spacer and sequencing primer sequences are introduced via the anchor sequence in a second round of PCR.

To validate this method and its application to low biomass samples, we compared vaginal community state types [2] as defined by metataxonomic profiling of vaginal samples from late and post-reproductive age women [3] targeting the V3-V4 region of the 16S rRNA gene. Samples from each woman were prepared using the 1-Step PCR procedure [1] sequenced on the Illumina MiSeq platform, and the 2-Step PCR procedure sequenced on both the Illumina MiSeq and HiSeq platforms. We sought to evaluate if the within-woman vaginal community state types differ between methods.

Materials & Methods

Late and post-reproductive age vaginal sample collection & genomic DNA extraction

A total of 92 mid-vaginal ESwabs stored in Amies transport medium (Copan) as previously described [3] were utilized in this study. The use of these samples was approved by the University of Maryland Baltimore IRB. Samples were thawed on ice and vortexed briefly. A 0.5 mL aliquot of the cell suspension was transferred to a FastPrep Lysing Matrix B (MP

Biomedicals) tube containing 0.5 mL of PBS (Invitrogen). A cell lysis solution containing 5 μL lysozyme (10 mg/ml; EMD chemicals), 13 μL mutanolysin (11,700 U/ml; Sigma Aldrich), and 3.2 μL lysostaphin (1 mg/ml; Ambi Products, LLC) was added and samples were incubated at 37°C for 30 min. Then, 10 μL Proteinase K (20mg/ml; Invitrogen), 50 μL 10% SDS (Sigma), and 2 μL RNase A (10mg/ml; Invitrogen) were added and samples were incubated at 55°C for an additional 45 min. Cells were lysed by mechanical disruption on a FastPrep homogenizer at 6 m/s for 40 s, and the lysate was centrifuged on a Zymo Spin IV column (Zymo Research). Lysates were further processed on the QIAsymphony platform using the QS DSP Virus/Pathogen Midi Kit (Qiagen) according to the manufacturer's recommendation. DNA quantification was carried out using the Quant-iT PicoGreen dsDNA assay (Invitrogen). Three separate sequencing libraries were constructed from each genomic DNA: one using the 1-Step 16S rRNA gene V3-V4 regions PCR protocol described by Fadrosh *et al.* [1], and two using the 2-Step 16S rRNA gene V3-V4 regions PCR protocol.

Sequencing library construction using 1-Step PCR

Sequencing libraries were constructed by amplifying the 16S rRNA gene V3-V4 regions using the 1-Step PCR amplification protocol previously described [1]. Primer sequences ranged from 90-97 bp depending on the length of the heterogeneity spacer (Table 1). Amplification was performed using Phusion Taq Master Mix (1X, ThermoFisher) with 3% DMSO, 0.4 μM each primer, and 5 μL of genomic DNA. Cycling conditions were as follows: initial denaturation at 98°C for 30 s, 30 cycles of denaturation at 98°C for 15 s, annealing at 58°C for 15 s, and elongation at 72°C for 15 s, followed by a final elongation step at 72°C for 60 s. Amplicons were cleaned and normalized with the SequalPrep kit (Invitrogen) according to the manufacturer's recommendation.

Sequencing library construction using 2-Step PCR

The following library preparation method is a modified version of a method provided by Illumina (https://support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html). The V3-V4 regions of 16S rRNA genes were first amplified from genomic DNA using primers that combine bacterial 338F or 806R sequences previously described [1], a heterogeneity spacer of 0-7 bp, and the Illumina sequencing primers (**Table 2, Step 1**). A single PCR master mix was used for all 16S rRNA gene amplifications as the primers do not contain barcode indices (**Figure 1**). Each PCR reaction contained 1X Phusion Taq Master Mix (ThermoFisher), Step 1 Forward and Reverse primers (0.4 μM each, **Supplementary Table 1a**), 3% DMSO, and 5 μL of genomic DNA. PCR amplification was performed using the following cycling conditions: an initial denaturation at 94°C for 3 min, 20 cycles of denaturation at 94°C for 30 s, annealing at 58°C for 30 s, and elongation at 72°C for 1 min, and a final elongation step at 72°C for 7 min. The resultant amplicons were diluted 1:20, and 1 μL was used in the second step PCR. This second amplification step introduced an 8 bp dual-index barcode to the 16S rRNA gene amplicons (**Supplementary Table 1b**), as well as the flow cell linker adaptors using primers containing a sequence that anneals to the Illumina sequencing primer sequence introduced in step 1 (**Table 2, Step 2** and **Supplementary Tables 1c and 1d** for full oligonucleotide sequences). Each primer was added to a final concentration of 0.4 μM in each sample specific reaction, along with Phusion Taq Master Mix (1X) and 3% DMSO. Phusion Taq Polymerase (ThermoFisher) was used with the following cycling conditions: an initial denaturation at 94°C for 30 s, 10 cycles consisting of denaturation at 94°C for 30 s, annealing at 58°C for 30 s, and elongation at 72°C for 60 s, followed by a final elongation step at 72°C for 5 min (**Figure 1**). Libraries were cleaned using 0.6X SPRI beads (Agencourt) and quantified using a Perkin Elmer LabChip GX Touch HT instrument.

Amplicon success scoring and pooling

Prepared libraries were run on a 2% agarose E-Gel (ThermoFisher, Waltham, MA) and scored for their relative success after amplification (expected ~627 bp, amplicon + linker + spacer + all primer sequences). Based on the score from the gel, a volume of 5 µl, from successful samples, 10 µl from partially success, and 15 µl from low success samples were pooled into an Eppendorf tube. Pooled amplicons were cleaned and normalized using the SequalPrep normalization kit (Life Technologies, Carlsbad, Ca), according to manufacturer's recommendations. The pooled samples were cleaned up with AMPure XP (Agencourt/Beckman Coulter, Brea, CA) beads following manufacturer's instructions and size selected around 600 bp. After size-selection the DNA was eluted in water. To ensure proper size of PCR product the pooled libraries were run on Agilent TapeStation 2200 with a DNA1000 tape for quality assurance.

Sequencing by Illumina MiSeq and sequence data processing

Libraries were sequenced on an Illumina MiSeq instrument using 600 cycles producing 2 x 300 bp paired-end reads. The sequences were de-multiplexed using the dual-barcode strategy, a mapping file linking barcode to samples and split_libraries.py, a QIIME-dependent script [4]. The resulting forward and reverse fastq files were split by sample using the QIIME-dependent script split_sequence_file_on_sample_ids.py, and primer sequences were removed using TagCleaner (version 0.16) [5]. Further processing followed the DADA2 Workflow for Big Data and dada2 (v. 1.5.2) (https://benjjneb.github.io/dada2/bigdata.html, [6], **Supplementary File 1**). Forward and reverse reads were each trimmed using lengths of 255 bp and 225 bp, respectively, filtered to contain no ambiguous bases, trimmed at minimum quality score of two, and the maximum number of expected errors in a read set to 2. Reads were assembled and chimeras for the combined runs removed as per the dada2 protocol.

Sequencing by Illumina HiSeq and sequence data processing

Libraries were sequenced on an Illumina HiSeq 2500 using Rapid Run chemistry and a 515 nm laser barcode reader (a required accessory), and loaded at 8 pmol with 20% diverse library. Paired-end 300 bp reads were obtained using a HiSeq Rapid SBS Kit v2 (2 x 250 bp, 500 cycles kit) combined with a (2 x 50 bp, 100 cycles kit; alternatively, a single 500 bp kit plus 2 x 50 bp kits can be used instead). Within the HiSeq Control Software, under the Run Configuration tab, within the Flow Cell Setup, the Reagent Kit Type was set to “HiSeq Rapid v2”, and the Flow Cell Type to “HiSeq Rapid Flow Cell v2”. Next, within Recipe, the Index Type was set to “Custom”, the Flow Cell Format to Paired End, and the Cycles set to “301”, “8”, “8”, “301”, for Read 1, Index 1, Index 2, and Read 2, respectively (**Supplementary File 2**). Instead of the standard sequencing primers, custom locked nucleic acid primers were used according to the Fluidigm Access Array User Guide Appendices B and C [7]. The sequences were de-multiplexed using the dual-barcode strategy, a mapping file linking barcode to samples (**Supplementary Table 1**), and split_libraries.py, a QIIME-dependent script [4]. The resulting forward and reverse fastq files were split by sample using the QIIME-dependent script split_sequence_file_on_sample_ids.py, and primer sequences were removed using TagCleaner (version 0.16) [5]. Further processing followed the DADA2 Workflow for Big Data and DADA2 (v. 1.5.2) (https://benjjneb.github.io/dada2/bigdata.html, [6]). Forward and reverse reads were each trimmed using lengths of 255 and 225 bp, respectively, filtered to contain no ambiguous bases, a minimum quality score of two was imposed, with the maximum number of expected errors in a read set to 2. Reads were assembled and chimeras for the combined runs were removed as per the DADA2 protocol.

All sequence data are available from NCBI SRA under Accession number SRP159872.

Sequencing Quality Comparisons

To compare the quality of a near-full run of sequences produced by the 2-Step PCR library preparation sequenced on either the Illumina MiSeq or HiSeq 2500 platforms, sample-specific forward and reverse fastq files were analyzed and visualized in R version 3.4.4 (2018-03-15) using the qa function of the ShortRead package v 1.36.1 [8], data.table v 1.11.4 , and ggplot2 v 3.0.0 [9]. Because quality scores were not normally distributed, a Mann-Whitney-Wilcoxon test was applied to test if differences in the quality scores per cycle differed between the two sequencing platforms (R Package: stats, Function: wilcox.test).

Amplification success of low bioburden late and post-reproductive age vaginal samples

The success or failure of amplifying the 16S rRNA gene V3-V4 regions from low biomass vaginal samples of late and post-reproductive age women using the 1-Step or 2-Step protocols was measured by the presence or absence of an amplicon band using agarose gel electrophoresis after the final amplification (in the case of the 2-Step protocol, after the 2nd step). Samples successfully amplified using all three protocols were used for statistical analyses. To test for differences in the quality scores of samples prepared and sequenced by the different methods, a Kruskal-Wallis Rank Sum test was applied.

Distance-based bacterial community comparisons from low bioburden late and post-reproductive vaginal samples

The 1-Step library was sequenced on the Illumina MiSeq Platform and the 2-Step library was sequenced on both the Illumina MiSeq and HiSeq platforms. Sequences were quality-filtered and assembled as described above. For each of the three quality-filtered datasets, amplification sequence variants generated by DADA2 were individually taxonomically classified using the RDP Naïve Bayesian Classifier [10] trained with the SILVA v132 16S rRNA gene database [11]. ASVs of major vaginal taxa were assigned species-level annotations using speciateIT (version 2.0), a novel and rapid per sequence classifier (http://ravel-lab.org/speciateIT), and verified via BLASTn against the NCBI 16S rRNA reference database. Read counts for ASVs assigned to the same taxonomy were summed for each sample. To determine if library preparation methods influenced microbial community β-diversity, samples were assigned a vaginal community state type as defined by Jensen-Shannon distances and clustering via Ward linkage. Agreement of within-subject assigned CSTs between methods was determined using Fleiss' Kappa statistic κ [12] (R package: irr v 0.84). Here $\kappa = 0$ indicates all CST assignments were dissimilar between the libraries, and $\kappa = 1$ indicates identical CST assignments. A $\kappa > 0.75$ is considered excellent agreement.

Results

Comparison of Illumina MiSeq and Illumina HiSeq amplicon sequencing read quality and quantity

To compare the quality of amplicon reads produced via 2-Step PCR and the Illumina MiSeq and HiSeq platforms, each sequencing run was demultiplexed with the same mapping file, and the quality profiles were compared. Significantly greater mean quality scores were observed for 1,536 samples run on the HiSeq platform compared to 444 samples run on the MiSeq platform ($U = 3 \times 10^5$, $p < 2.2\times10^{-16}$, **Figure 2**). The HiSeq 2500 platform produced a greater mean number of quality-filtered sequences per sample than the MiSeq platform, with fewer chimeric sequences detected on average (**Table 3**). Additionally, the HiSeq 2500 sequencing strategy was more cost efficient for large sequencing projects at $3.99 per sample, assuming 2 lanes are run with 1,568 multiplexed samples per lane (**Table 3**).

2-Step PCR amplicon library preparation improves amplification success of low biomass vaginal samples

Of 92 low-biomass vaginal samples collected from late and post-reproductive women, 54% were successfully amplified using the 1-Step PCR protocol, while the 2-Step protocol produced amplifications from 90% of samples (**Table 4**). Of 42 samples that did not amplify by the 1-Step method, 55% were over the age of 51, the average of menopause, and 34 successfully amplified using the 2-Step method, an 80% improvement (**Supplementary Table 2**). Amplicons were not observed from 8 samples regardless of protocol type, and 1 sample was successfully amplified using the 1-Step but not the 2-Step procedure. From all libraries, 1-3% of sequences were detected as chimeras and removed. This yielded on average 11,080 sequences per sample from the 1-Step library, 14,282 sequences per sample from the 2-Step library sequenced on the MiSeq platform, and 50,514 sequences per sample from the 2-Step library sequenced on the HiSeq platform. The 1-Step library consisted of 49 samples, of which 30 had > 500 total sequences and were used for comparative β-diversity analysis (**Table 4**). Consistency of observed CSTs between libraries was tested by using Fleiss' *kappa* for inter-rater reliability, where $\kappa > 0.75$ indicated excellent agreement. Complete agreement between all three methods was observed ($\kappa = 1.0$, **Figure 3**, raw read count taxonomy tables are available in Supplemental Table 3).

Discussion

Large sample sizes and within subject frequent sampling are now becoming the norm for microbiome analyses to increase statistical power. Therefore, higher-throughput capabilities are needed that do not sacrifice sequence quality, afford flexibility to target a diverse set of genes or gene regions, and maintain the ability to sequence longer amplicons for increased taxonomic resolution. Additionally, the less than optimal read quality and per sample read counts generated by the Illumina MiSeq platform necessitated an improved method. The innovative use of the Illumina HiSeq 2500 platform presented here improves on current technologies by producing 300 bp PE reads of high quality and multiplexing of up to 1,568 samples per lane compared to samples multiplexed on an Illumina MiSeq instrument. This new approach affords a greater mean number of significantly higher quality sequences per sample with high multiplexing.

The 2-Step PCR library preparation method described here allows for production of sequencing libraries from various gene targets and low biomass samples. Amplification success of low biomass samples prone to amplification difficulties was improved by 80% when this method was used instead of the traditional 1-Step PCR method. In addition to lower the cost of the shorter primers used in the 2-Step PCR library protocol, which do not require PAGE purification, the 2-Step PCR protocol represents a major improvement. Other investigators have reported the use of 16S rRNA gene fusion amplification primers that contain a universal 16S rRNA sequence, a barcode and sequencer specific adaptors have been previously used to generate large sequence datasets, including those related to the Human Microbiome Project [13, 14]. This 1-Step PCR library construction method suffers from low efficacy of amplification due to the long primer length, which is especially problematic in cases where template targets are in low abundance. A 2-Step PCR library construction wherein a barcode and sequencer specific adaptors sequences are added in a second highly efficient PCR step is preferable. This approach affords flexibility to target any regions of interest with minimal investment as only new primers for the first PCR of the 2-Step library preparation method are needed. Other low-biomass environments that could benefit from the 2-Step PCR procedure include blood and serum [15], respiratory airways [16], skin [17], sub-seafloor sediments [18], and clean rooms [19], among others.

In summary, to demonstrate the comparability of sequence datasets produced via different methods, 16S rRNA gene V3-V4 regions sequence datasets were generated from low-biomass vaginal samples from late and post-reproductive age women using both 1-Step and 2-Step PCR library construction methods and the Illumina HiSeq and MiSeq sequencing platforms. Complete within-subject agreement between the vaginal community state type assignments [2] were observed between all three methods, though a greater number of significantly higher quality sequences were obtained from the 2-Step PCR method sequenced on the Illumina HiSeq 2500 platform. We therefore conclude that while the 2-Step PCR preparation method combined with the Illumina HiSeq 2500 platform is preferred, data generated by 1-Step or 2-Step PCR and sequenced on the Illumina MiSeq or HiSeq 2500 platform can still be combined to successfully obtain meaningful conclusions about the environment and sample types of interest.

Limitations:

The method is extremely high-throughput, and as such might not be suitable for small projects unless these are combined with other samples. Producing a large number of samples ready for pooling requires automation so that time from sample collection to data generation is still reasonable. Overall, automation is required, and this approach might be suitable for microbiome service cores where faster turn-around is needed and running many MiSeq runs is not a viable option.

Acknowledgements

JB Holm was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number F32AI136400. Research reported in this publication was supported in part by the National Institute of Allergy and Infectious Diseases and the National Institute of Nursing Research of the National Institutes of Health under award numbers: U19AI084044, R01AI116799, R21AI107224, R01AI089878 and R01NR015495. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, Ravel J: **An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform.** In: *Microbiome.* vol. 2; 2014: 6.
2. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO *et al*: **Vaginal microbiome of reproductive-age women.** In: *Proc Natl Acad Sci USA.* vol. 108 Suppl 1; 2011: 4680-4687.
3. Brotman RM, Shardell MD, Gajer P, Fadrosh D, Chang K, Silver MI, Viscidi RP, Burke AE, Ravel J, Gravitt PE: **Association between the vaginal microbiota, menopause status, and signs of vulvovaginal atrophy**. In: *Menopause.* vol. 21; 2014: 450-458.
4. Kuczynski J, Stombaugh J, Walters WA, Gonzalez A, Caporaso JG, Knight R: **Using QIIME to analyze 16S rRNA gene sequences from microbial communities**. *Current protocols in microbiology* 2012, **Chapter 1**:Unit 1E 5.

5. Schmieder R, Lim YW, Rohwer F, Edwards R: **TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets**. *Bmc Bioinformatics* 2010, **11**:341.

6. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP: **DADA2: High-resolution sample inference from Illumina amplicon data**. *Nature methods* 2016, **13**(7):581-583.

7. Fluidigm: **Access Array System for Illumina Sequencing Systems**. In.; 2018: 61-68.

8. Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R: **ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data**. *Bioinformatics* 2009, **25**(19):2607-2608.

9. Wickham H: **Ggplot2 : elegant graphics for data analysis**. New York: Springer; 2009.

10. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy**. In: *Applied and Environmental Microbiology.* vol. 73; 2007: 5261-5267.

11. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.** In: *Nucleic Acids Res.* vol. 41; 2013: D590-596.

12. Fleiss JL: **Measuring nominal scale agreement among many raters**. *Psychological bulletin* 1971, **76**(5):378.

13. Consortium THMP: **Structure, function and diversity of the healthy human microbiome**. In., vol. 486; 2012: 207-214.

14. Sinha R, Abnet CC, White O, Knight R, Huttenhower C: **The microbiome quality control project: baseline study design and future directions**. In: *Genome Biol.* Genome Biology; 2015: 1-6.

15. Santiago A, Pozuelo M, Poca M, Gely C, Nieto JC, Torras X, Roman E, Campos D, Sarrabayrouse G, Vidal S *et al*: **Alteration of the serum microbiome composition in cirrhotic patients with ascites**. *Sci Rep* 2016, **6**:25001.

16. Goleva E, Jackson LP, Harris JK, Robertson CE, Sutherland ER, Hall CF, Good JT, Jr., Gelfand EW, Martin RJ, Leung DY: **The effects of airway microbiome on corticosteroid responsiveness in asthma**. *Am J Respir Crit Care Med* 2013, **188**(10):1193-1201.

17. Byrd AL, Belkaid Y, Segre JA: **The human skin microbiome**. *Nat Rev Microbiol* 2018, **16**(3):143-155.

18. Fry JC, Parkes RJ, Cragg BA, Weightman AJ, Webster G: **Prokaryotic biodiversity and activity in the deep subseafloor biosphere**. *FEMS Microbiol Ecol* 2008, **66**(2):181-196.

19. Vaishampayan P, Probst AJ, La Duc MT, Bargoma E, Benardini JN, Andersen GL, Venkateswaran K: **New perspectives on viable microbial communities in low-biomass cleanroom environments**. *ISME J* 2013, **7**(2):312-324.

Table 1. 1-Step PCR Method Primers (5' → 3')

	Illumina MiSeq 3' Flowcell Linker + Illumina 5' Sequencing Primer (CS1/CS2) + Index + Heterogeneity Spacer + 16S rRNA Gene V3-V4 Primer
Forward Primer	AATGATACGGCGACCACCGAGATCTACAC + GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT + Index (8 bp) + Heterogeneity Spacer (0-7 bp) + ACTCCTRCGGGAGGCAGCAG
Reverse Primer	CAAGCAGAAGACGGCATACGAGAT + ACACTCTTTCCCTACACGACGCTCTTCCGATCT + Index (8 bp) + Heterogeneity Spacer (0-7 bp) + GGACTACHVGGGTWTCTAAT

Table 2. 2-Step Protocol PCR Primers (5' → 3')

Step 1*	**Illumina 5' Sequencing Primer (CS1/CS2) + Heterogeneity Spacer + 16S rRNA Gene V3-V4 Primer**
Forward Primer	ACACTGACGACATGGTTCTACA + Heterogeneity Spacer (0-7 bp) + ACTCCTRCGGGAGGCAGCAG
Reverse Primer	TACGGTAGCAGAGACTTGGTCT + Heterogeneity Spacer (0-7 bp) + GGACTACHVGGGTWTCTAAT
Step 2**	**Illumina 3' Flowcell Linker + Index + CS1/CS2 Complement**
Forward Primer	AATGATACGGCGACCACCGAGATCTACAC + INDEX (8 bp) + ACACTGACGACATGGTTCTACA
Reverse Primer	CAAGCAGAAGACGGCATACGAGAT + INDEX (8 bp) + TACGGTAGCAGAGACTTGGTCT

*See Supplementary Table 1b for full oligonucleotide sequences
**See Supplementary Tables 1c & 1d for full forward and reverse oligonucleotides, respectively

Table 3. Sequencing run information for the MiSeq and HiSeq platforms.

Sequencing Platform	**MiSeq**	**HiSeq 2500 RR**
Run Details	2 x 300 bp PE	2 x 250 bp + 2 x 50bp
Mean No. Assembled Sequences per Sample ± SE	14,774 ± 503	52,142 ± 4750
No. Samples in Sequencing Run	444	1,536
Mean Quality Score per Sample ± SE	27.2 ± 0.3*	34.6 ± 0.2*
Mean No. Reads per Sample Pre-QC ± SE	22,880 ± 2006	58,034 ± 1040
Mean No. Reads per Sample Post-QC ± SE	9,938 ± 1042	47,307 ± 848
% Chimeric Sequences Detected	10.8	7.8
Mean No. Non-chimeric, Assembled Sequences per Sample ± SE	8,383 ± 825	42,978 ± 735
Cost of Sequencing per Sample (No. Multiplexed Samples)	$6.38 (384)	$3.99 (1,568)

* Significant. Wilcoxon Rank Sum W = 3×10^5, $p < 2.2 \times 10^{-16}$

Table 4. Summary of sequencing results for low-bioburden, late and post-reproductive age vaginal samples

Library Preparation Method	**1-Step**	**2-Step**	
No. samples attempted to amplify	92	92	
No. samples amplified	49	83	
Sequencing Platform	MiSeq	MiSeq	HiSeq
% Chimeric Sequences Detected	0.70	3.3	3.1
Mean No. Non-chimeric, Assembled Sequences per Sample ± SE	11,080 ± 1506	14,282 ± 483	50,514 ± 4427
Median Quality Score per Sample [Q1-Q3]	36.2 [33.5-37.2]*	34.9 [29.9-36.3]*	37.1 [33.0-38.0]*

*Significant. Kruskal-Wallis H = 187.85, p < 2.2 x 10^{-16}

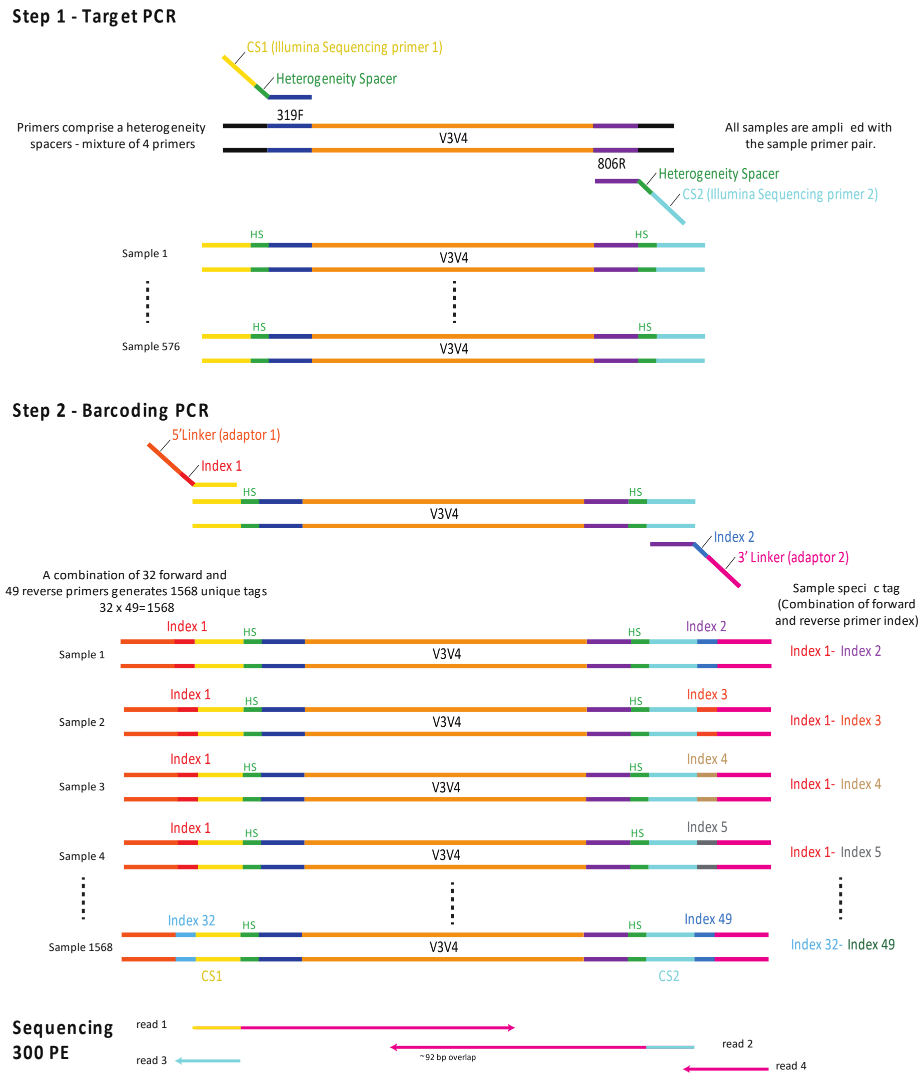
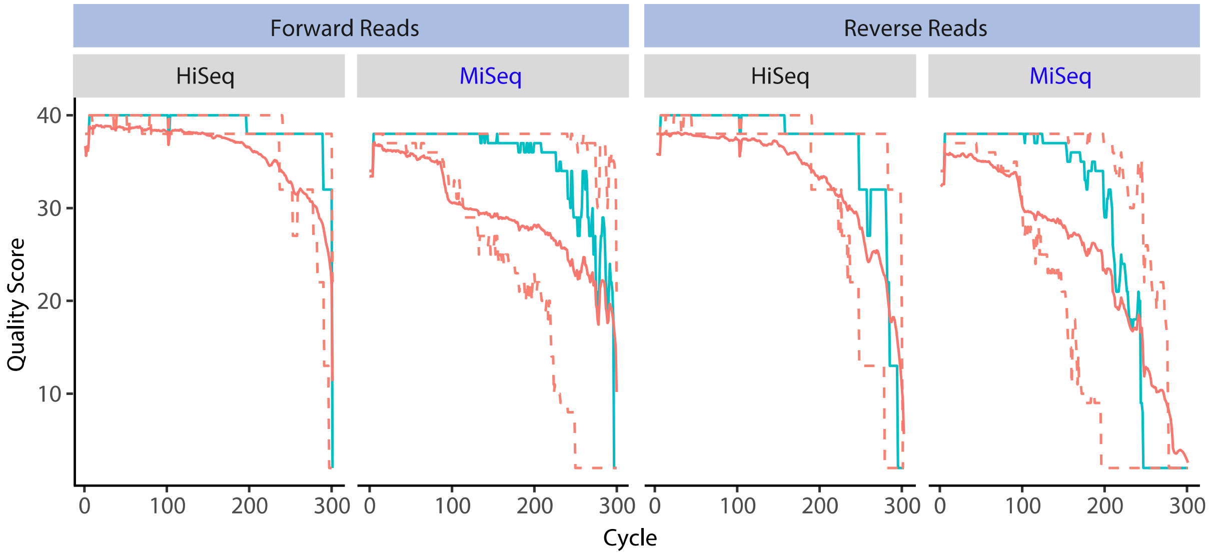


Figure 1. Illumina amplicon library preparation through 2-Step PCR amplification.

360

361



362

Figure 2. Forward and reverse read quality profiles for 300 cycles on the Illumina HiSeq (1,536 samples) and MiSeq (444 samples) platforms. Amplicon libraries were prepared using a 2-Step PCR method. Shown for each cycle are the mean quality score (green line), the median quality score (solid orange line), the quartiles of the quality score distribution (dotted orange lines).

363

364

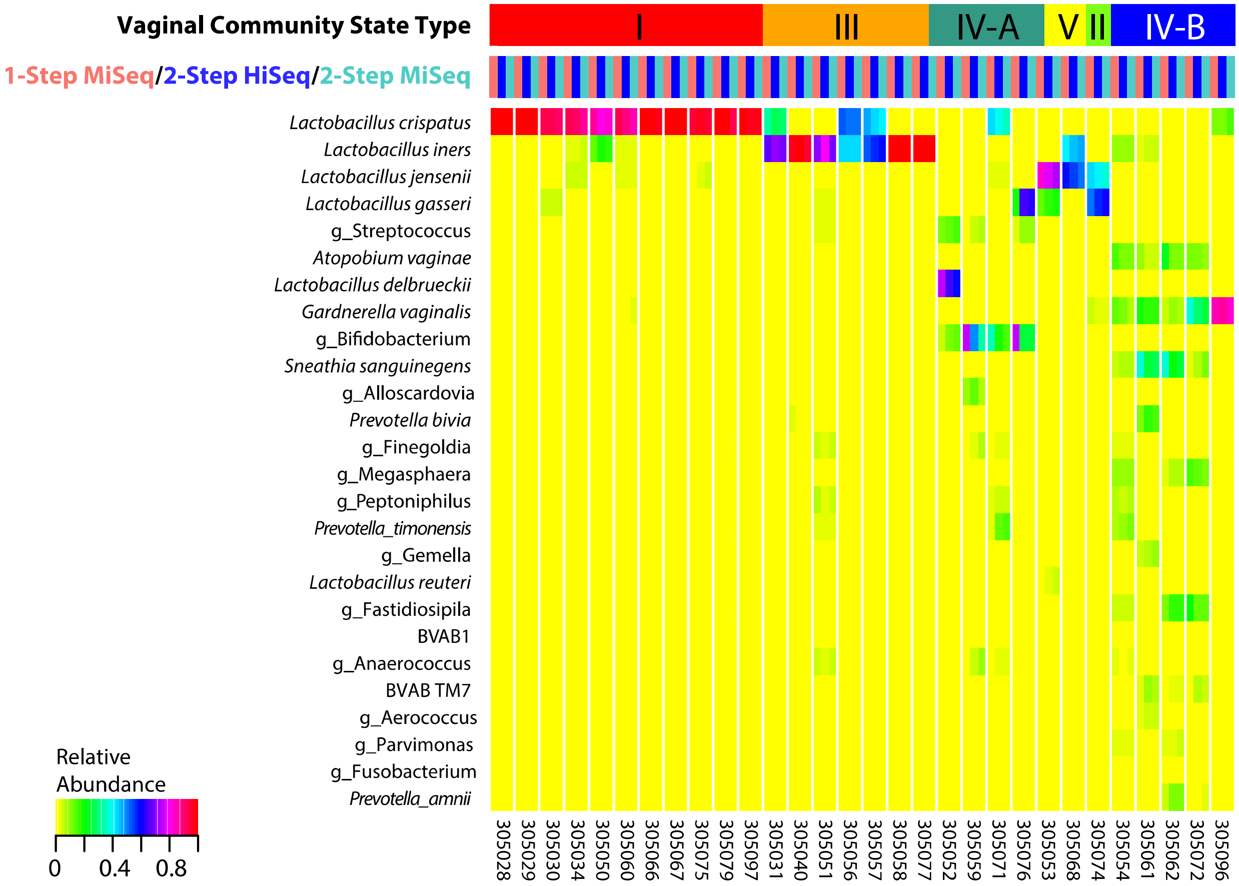


Figure 3. Heatmap of taxon relative abundances (rows) of samples (columns). Subject samples are separated by white lines and samples are ordered by vaginal community state types and as follows: 1-Step MiSeq (pink), 2-Step HiSeq (blue), 2-Step MiSeq (aqua).