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 amplicon-sequencing applications have grown in the total number of samples, growth in sample multiplexing is becoming necessary while maintaining high sequence quality and sequencing depth. Here, modifications to the Illumina HiSeq 2500 platform are described which produce greater multiplexing capabilities 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 high throughput sequencing have made it a widely-adopted approach, especially for projects which necessitate 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 ITS region, and other marker genes. 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 high throughput sequencing have made it a widely-adopted approach, especially for projects which

necessitate sample multiplexing to eliminate batch effect and reduced time to acquire data, which is often the case in sequencing cores. The Illumina HiSeq 2500 platform with its high throughput offers a remedy to limitations in multiplexing 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) [2].

We present a method that produces high-quality 300 bp paired-end reads (median Q-score 37.1) 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 a modified version of previously published 1-Step PCR [1] and 2-Step PCR (https://support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html) methods. 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, and are associated with 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 more flexible and improves amplification from low biomass samples because the 1st step primers are short, target the amplicon region of interest, and contain the heterogeneity spacer and Illumina Sequencing Primer. The barcodes and flow-cell linker sequences are introduced in a second round of PCR by using the Illumina Sequencing primer as a target.

A previously published 2-Step PCR method [2] used triple barcode-indexing, produced 2 x 250 bp paired-end reads on the Illumina HiSeq 2500 platform, and reported a taxon-specific sequencing bias of the first step primers which differed in both barcode sequence and heterogeneity spacer length. The method we present here uses 8 bp dual-indexing as described by Fadrosh *et al.* [1] wherein the forward index is never used as a reverse index, produces 2 x 300 bp paired end reads by modifying the Illumina HiSeq 2500 sequencing method, and attempts to control for amplification biases by implementing 1) an equimolar ratio of all PCR step 1 primers (which differ only in the length of the heterogeneity spacers) provided to each sample to reduce biases imposed by the heterogeneity spacer, and 2) introduction of barcode sequences are in the second PCR step for library preparation.

69 In addition to the benefit of flexibility in choice of gene target, we show that the 2-Step PCR
70 method improves amplification success of low biomass samples relative to the 1-Step PCR
71 method. Additionally, we show that the 2-Step PCR method does not significantly bias the
72 measured microbial community by comparing vaginal community state types [3] as defined by
73 taxonomic profiling of vaginal samples of pre- and post-menopausal women [4] targeting the V3-
74 V4 region of the 16S rRNA gene. Post-menopausal vaginal samples tend to be lower in
75 absolute bacterial load relative to pre-menopausal samples [5, 6], making amplification
76 challenging. Samples from each woman were prepared using the 1-Step PCR procedure [1]
77 sequenced on the Illumina MiSeq platform, and the 2-Step PCR procedure sequenced on both
78 the Illumina MiSeq and HiSeq platforms. In addition to comparing the quality of libraries
79 sequenced on the Illumina HiSeq and MiSeq platforms, we also sought to measure 1) improved
80 amplification efficiency of samples prepared by the 2-Step PCR method compared to the 1-Step
81 method and 2) the differences in intra-individual vaginal community state types between
82 methods. Finally, we demonstrate the precision of this method using a comparative mock
83 community analysis.
84
85 **Materials & Methods**
86 *Overall Study Design*
87 First, to determine if the choice of library preparation method improved amplification of low-
88 biomass samples, we specifically processed 92 vaginal samples using the dual-indexing 1-Step
89 [1] and 2-Step (described below) library preparation methods. The success of amplifying the
90 16S rRNA V3V4 region from genomic DNA was evaluated for each method.
91
92 To then determine if the choice of library preparation method or sequencing platform impacted
93 the observed sample microbial composition of these samples, we sequenced the libraries of
94 samples successfully produced using both 1-Step and 2-Step methods on the Illumina MiSeq
95 (1-Step and 2-Step) and HiSeq (2-Step only) platforms. The same 2-Step library was
96 sequenced on the MiSeq and HiSeq platforms. The compositions of samples for which high-
97 quality data were obtained from all three methods were statistically compared.
98
99 To further validate if sequencing platform impacted observed microbial compositions, we also
100 produced ten separate V3-V4 16S rRNA gene amplicon libraries from the ZymoBIOMICS
101 Microbial Community DNA Standard (Zymo Research, Irvine, CA) using the 2-Step library
102 preparation method, and sequenced each library on separate runs of the Illumina HiSeq

platform. We compared the microbial compositions of these samples to theoretical values reported by Zymo as well as to V3-V4 amplicon libraries of the same standard prepared and sequenced by Zymo Research on the Illumina MiSeq platform (see **Supplementary File 7** for library preparation and sequencing methods).

Finally, to compare the sequencing quality and per sample read statistics (per sample number and quality of reads) produced by the Illumina MiSeq and HiSeq 2500 platforms, amplicon libraries from vaginal samples were produced using the 2-Step PCR method and sequenced on both the Illumina MiSeq (276 out of possible 576 samples) and HiSeq 2500 (1,194 out of possible 1,568 samples) platforms. All amplicon libraries targeted the 16S rRNA gene V3-V4 regions from human vaginal samples.

Genomic DNA extraction

Clinician-collected mid-vaginal ESwabs were stored in Amies transport medium (Copan, Murrieta, CA) as previously described [4]. The study was approved by the University of Maryland Baltimore and the Johns Hopkins School of Public Health Institutional Review Board. 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, Santa Ana, CA) tube containing 0.5 mL of PBS (Invitrogen, Carlsbad, CA). A cell lysis solution containing 5 μL lysozyme (10 mg/ml; EMD Chemicals, Gibbstown, NJ), 13 μL mutanolysin (11,700 U/ml; Sigma Aldrich, St. Louis, MO), and 3.2 μL lysostaphin (1 mg/ml; Ambi Products, LLC, Lawrence, NY) 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 Aldrich, St. Louis, MO), and 2 μL RNase A (10mg/ml; Invitrogen, Carlsbad, CA) 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 at 7000 x g for 1 min. (Zymo Research, Irvine, CA). Lysates were further processed on the QIAsymphony platform using the QS DSP Virus/Pathogen Midi Kit (Qiagen, Hilden, GER) according to the manufacturer's recommendation. DNA quantification was carried out using the Quant-iT PicoGreen dsDNA assay (Invitrogen).

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

137 90-97 bp depending on the length of the heterogeneity spacer (**Table 1**). Amplification was
138 performed using Phusion Taq Master Mix (1X, ThermoFisher, Waltham, MA) with 3% DMSO,
139 0.4 μM each primer, and 5 μL of genomic DNA. A standard volume of genomic DNA was used
140 for each library because genomic DNA concentration was not indicative of the number of 16S
141 rRNA gene targets (**Supplementary File 1**). Cycling conditions were as follows: initial
142 denaturation at 98°C for 30 s, 30 cycles of denaturation at 98°C for 15 s, annealing at 58°C for
143 15 s, and elongation at 72°C for 15 s, followed by a final elongation step at 72°C for 60 s.
144
145 *Sequencing library construction using 2-Step PCR*
146 The following library preparation method is a modified version of a method provided by Illumina
147 (https://support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.htm
148 l). The V3-V4 regions of 16S rRNA gene were targeted from genomic DNA using primers
149 bacterial 338F and 806R combined with a heterogeneity spacer of 0-7 bp, and the Illumina
150 Sequencing Primers (**Table 2, Step 1**). A single PCR master mix containing an equal ratio of all
151 primers, which vary by the length of the heterogeneity spacer, was used for all samples. This
152 strategy reduces any amplification biases that may be introduced by the differing lengths of the
153 heterogeneity spacers, and is efficient because the primers do not contain barcode indices
154 (**Figure 1**). Each PCR reaction contained 1X Phusion Taq Master Mix (ThermoFisher), Step 1
155 Forward and Reverse primers (0.4 μM each, **Supplementary Table 1a**), 3% DMSO, and 5 μL
156 of genomic DNA. This standard volume of genomic DNA was used for each library because
157 genomic DNA concentration was not indicative of the number of 16S rRNA gene targets
158 (**Supplementary File 1**). PCR amplification was performed using the following cycling
159 conditions: an initial denaturation at 94°C for 3 min, 20 cycles of denaturation at 94°C for 30 s,
160 annealing at 58°C for 30 s, and elongation at 72°C for 1 min, and a final elongation step at 72°C
161 for 7 min. We used only 20 PCR cycles because biases in microbial community profiles have
162 been reported with higher number of cycles [2]. The resultant amplicons were diluted 1:20, and
163 1 μL was used in the second step PCR. This second amplification step introduces an 8 bp dual-
164 index barcode to the 16S rRNA gene V3-V4 regions amplicons (**Supplementary Table 1b**), as
165 well as the flow cell linker adaptors using primers containing a sequence that anneals to the
166 Illumina sequencing primer sequence introduced in step 1 (**Table 2, Step 2** and for full
167 oligonucleotide sequences see **Supplementary Tables 1c and 1d**). Each primer was added to
168 a final concentration of 0.4 μM in each sample specific reaction, along with Phusion Taq Master
169 Mix (1X) and 3% DMSO. Phusion Taq Polymerase (ThermoFisher) was used with the following

170 cycling conditions: an initial denaturation at 94°C for 30 s, 10 cycles consisting of denaturation
171 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
172 elongation step at 72°C for 5 min (**Figure 1**).
173
174 *Amplicon library pooling for sequencing*
175 For a large number of samples, library purification and quantification for each sample would be
176 time and labor-intensive. To streamline the process, we visualize libraries on 2% agarose E-Gel
177 (ThermoFisher) and determine the relative amplification success at the expected ~627 bp band
178 size (amplicon + spacer + all primer sequences + linker). Strong, clear bands indicate
179 successful amplification, a weak or fuzzy band indicates intermediate and no band indicates low
180 amplification success. We then standardize the volume of each sample to pool to either 5, 10, or
181 15 µL of each sample depending on the high, intermediate, or low amplification success of that
182 sample, respectively (see **Supplementary File 2** for labeled gel example). The pooled samples
183 were cleaned up with AMPure XP (Agencourt/Beckman Coulter, Brea, CA) beads following
184 manufacturer's instructions and size selected around 600 bp. After size-selection the DNA was
185 eluted in water. To ensure proper size of PCR product the pooled libraries were run on Agilent
186 TapeStation 2200 with a DNA1000 tape for quality assurance.
187
188 Pooled libraries prepared by 1-Step PCR were sequenced on the Illumina MiSeq platform and
189 those prepared by 2-Step PCR were sequenced on both the Illumina HiSeq and MiSeq
190 platforms following the procedures outlined above.
191
192 *Amplification success of vaginal samples using the 1-Step and 2-Step PCR library preparation*
193 *methods*
194 To determine the success or failure of amplifying the 16S rRNA gene V3-V4 regions from
195 vaginal samples which include samples with low absolute bacterial load using the 1-Step or 2-
196 Step protocols, we evaluated the presence or absence of an amplicon band using agarose gel
197 electrophoresis after the final amplification (in the case of the 2-Step protocol, after the 2nd step).
198 Estimates of absolute bacterial abundance in vaginal samples were determined using real-time
199 quantitative PCR as previously described [7]. **Supplementary File 2** contains an example
200 electrophoresis gel labeled with the volume of the library used for pooling. Samples labeled with
201 "20" show no bands, and in this analysis, represent a failure of amplification. All other samples
202 represent successful amplifications.
203

204 *Sequencing by Illumina MiSeq and sequence data processing*
205 Libraries were sequenced on an Illumina MiSeq instrument using 600 cycles producing 2 x 300
206 bp paired-end reads. The sequences were de-multiplexed using the dual-barcode strategy, a
207 mapping file linking barcode to samples and split_libraries.py, a QIIME-dependent script [8]. The
208 resulting forward and reverse fastq files were split by sample using the QIIME-dependent script
209 split_sequence_file_on_sample_ids.py, and primer sequences were removed using TagCleaner
210 (version 0.16) [9]. Further processing followed the DADA2 Workflow for Big Data and DADA2 (v.
211 1.5.2) (https://benjjneb.github.io/dada2/bigdata.html, [10], **Supplementary File 3**).
212
213 *Sequencing by Illumina HiSeq and sequence data processing*
214 Libraries were sequenced on an Illumina HiSeq 2500 using Rapid Run chemistry and a 515 nm
215 laser barcode reader (a required accessory), and loaded at 8 pmol with 20% PhiX library.
216 Paired-end 300 bp reads were obtained using a HiSeq Rapid SBS Kit v2 (2 x 250 bp, 500
217 cycles kit) combined with a (2 x 50 bp, 100 cycles kit; alternatively, a single 500 bp kit plus 2 x
218 50 bp kits can be used). In the HiSeq Control Software, under the Run Configuration tab, within
219 the Flow Cell Setup, the Reagent Kit Type was set to “HiSeq Rapid v2”, and the Flow Cell Type
220 to “HiSeq Rapid Flow Cell v2”. Next, within Recipe, the Index Type was set to “Custom”, the
221 Flow Cell Format to Paired End, and the Cycles set to “301”, “8”, “8”, “301”, for Read 1, Index 1,
222 Index 2, and Read 2, respectively (**Supplementary File 4**). Instead of the standard sequencing
223 primers, custom locked nucleic acid primers were used according to the Fluidigm Access Array
224 User Guide Appendices B and C [11] (the Fluidigm system itself not required). These primers
225 are required for sequencing under the modified conditions, so that the CS1 and CS2 regions
226 can be used as primer binding regions (to produce reads 1 and 2, see **Figure 1**). The
227 sequences were de-multiplexed using the dual-barcode strategy, a mapping file linking barcode
228 to samples (**Supplementary Table 1**), and split_libraries.py, a QIIME-dependent script [8]. The
229 resulting forward and reverse fastq files were split by sample using the QIIME-dependent script
230 split_sequence_file_on_sample_ids.py, and primer sequences were removed using TagCleaner
231 (version 0.16) [9]. Further processing followed the DADA2 Workflow for Big Data and DADA2 (v.
232 1.5.2) [10].
233
234 *Intra-individual distance-based bacterial community comparisons of vaginal samples*
235 Samples successfully amplified using both library preparation methods were used for
236 comparative analyses. The 1-Step libraries were sequenced on the Illumina MiSeq Platform and
237 the 2-Step libraries were sequenced on both the Illumina MiSeq and HiSeq platforms.

238 Sequences were quality-filtered and assembled as described above. To fairly compare the
239 Illumina HiSeq and MiSeq platforms, lengths of 255 bp and 225 bp were chosen for hard
240 trimming of forward and reverse reads, respectively, because these were the lengths at which
241 median quality scores decreased below 20 for the worst library (see **Supplementary File 5**,
242 specifically the 2-Step MiSeq F and R reads quality decreases dramatically at approximately
243 these lengths, and so the same lengths were applied to all three methods). Individual reads
244 were further truncated at the base where a quality score of 2 was observed and filtered to
245 contain no ambiguous bases. Additionally, the maximum number of expected errors in a read
246 was set to 2. Reads were assembled only if the overlap between forward and reverse reads,
247 which occurs in the conserved region between V3 and V4, was 100% identical. Chimeras for
248 combined runs removed as per the dada2 protocol. A Kruskal-Wallis test was applied to test if
249 differences in the per sample quality scores differed between the three methods (R Package:
250 stats, Function: kruskal.test). For each of the three quality-filtered datasets, amplification
251 sequence variants (ASVs) generated by DADA2 were individually taxonomically classified using
252 the RDP Naïve Bayesian Classifier [12] trained with the SILVA v128 16S rRNA gene sequence
253 database [13]. ASVs of major vaginal taxa were assigned species-level annotations using
254 speciateIT (version 2.0), a novel and rapid per sequence classifier (http://ravel-
255 lab.org/speciateIT), and verified via BLASTn against the NCBI 16S rRNA gene sequence
256 reference database. Read counts for ASVs assigned to the same taxonomy were summed for
257 each sample. To test for differences in the quality scores of samples prepared and sequenced
258 by the different methods, a Kruskal-Wallis Rank Sum test was applied. To determine if library
259 preparation methods influenced microbial community β-diversity, samples were assigned a
260 vaginal community state type as defined by Jensen-Shannon distances and clustering via Ward
261 linkage [3]. Clusters of Jensen-Shannon distances were visualized using t-Stochastic Neighbor
262 Embedding [14] using 5,000 iterations and perplexity set to 30. Agreement of within-subject
263 assigned CSTs between methods was determined using Fleiss' Kappa statistic κ [15] (R
264 package: irr v 0.84). Here $\kappa = 0$ indicates all CST assignments were dissimilar between the
265 libraries, and $\kappa = 1$ indicates identical CST assignments. A $\kappa > 0.75$ is considered excellent
266 agreement.
267
268 *Comparison of mock community microbial compositions between Illumina HiSeq runs*
269 We used the ZymoBIOMICS Microbial Community DNA Standard (Zymo Research) as a mock
270 community for this analysis. To maintain consistency in taxonomic annotations, we used BLAST
271 and the NCBI Reference database to classify each sequence variant in these analyses. Specific

272 single nucleotide variants produced different taxonomic classifications in the following taxa due
273 to truncation to the V3-V4 amplicon regions: *Bacillus subtilis* to *Bacillus mojavensis*, *Listeria*
274 *monocytogenes* to *Listeria welshimeri*, *Escherichia coli* to *Escherichia fergusonii*, however, we
275 manually verified the identity of the sequence variants. To determine if the mock community
276 amplicon library compositions produced by the our 2-Step library preparation and HiSeq
277 sequencing methods were within the same variation observed by Zymo Research, we
278 statistically compared the distributions of Jensen-Shannon distances between the Zymo-MiSeq
279 samples and the reported theoretical values and our HiSeq samples and the theoretical values
280 using a Mann-Whitney-Wilcoxon test (R Package: stats, Functions: wilcox.test).
281
282 *Sequencing Quality Comparisons of Illumina HiSeq and Illumina MiSeq Sequencing of 2-Step*
283 *PCR Amplicon Libraries*
284 To compare the sequence quality produced on the near-full Illumina MiSeq and HiSeq runs, the
285 per cycle mean, median, and 1st and 3rd quartiles were calculated from quality scores of sample-
286 specific forward and reverse fastq files in R version 3.4.4 (2018-03-15) using the qa function of
287 the ShortRead package v 1.36.1 [16], data.table v 1.11.4 , and ggplot2 v 3.0.0 [17] (R notebook
288 html available upon request). Because quality scores were not normally distributed, a Mann-
289 Whitney-Wilcoxon test was applied to test if differences in the quality scores per cycle differed
290 between the two sequencing platforms (R Package: stats, Functions: shapiro.test and
291 wilcox.test).
292
293 All sequence data are available from NCBI SRA under Accession number SRP159872.
294
295
296 **Results**
297 *2-Step PCR amplicon library preparation improves amplification success of low biomass vaginal*
298 *samples*
299 Amplification failure was more common in the 1-Step PCR amplification protocol due to the long
300 primers which degrade over time thus reducing amplification efficiency, especially for low
301 biomass samples (i.e. low absolute bacterial load). **Supplementary File 2** contains an example
302 electrophoresis gel labeled with the volume of the library used for pooling. Samples labeled with
303 "20" show no bands, and in this analysis, represent a failure of amplification. All other samples
304 represent successful amplifications. Of 92 low-biomass vaginal samples (mean subject age
305 48.9), 54% were successfully amplified using the 1-Step PCR protocol, while the 2-Step
306 protocol produced amplifications from 90% of samples (**Table 3**). Of 42 vaginal samples that did

307 not amplify by the 1-Step method, 55% were from women over the age of 51, the average age
308 of menopause. Thirty-four of these samples successfully amplified using the 2-Step method, an
309 80% improvement (**Supplementary Table 2**). A pan-bacterial qPCR analysis confirmed the
310 significantly lower number of 16S rRNA gene targets in the samples which were amplified by 2-
311 Step PCR but not 1-Step PCR (U= 790.5, p = 0.03, **Supplementary File 6)**. Subjects in this
312 group were also significantly older (U=484, p = 0.001). Amplicons were not observed from 8
313 samples regardless of protocol type, and 1 sample was successfully amplified using the 1-Step
314 but not the 2-Step procedure.
315
316 *Samples successfully amplified by both 1-Step and 2-Step library preparation methods yield*
317 *similar sequencing metrics on both Illumina MiSeq and HiSeq platforms*
318 Samples successfully amplified using both library preparation methods (n=49) were used to
319 compare the 1-Step and 2-Step library preparation methods sequenced on the Illumina HiSeq
320 (2-Step only) and MiSeq platforms (1-Step and 2-Step). From each combination of methods,
321 0.7-3% of sequences were detected as chimeras and removed. This yielded on average 11,080
322 sequences per sample from the 1-Step library sequenced on the MiSeq platform, 14,282
323 sequences per sample from the 2-Step library sequenced on the MiSeq platform, and 50,514
324 sequences per sample from the 2-Step library sequenced on the HiSeq platform (**Table 3**). Due
325 to low total read counts from some samples, only 30 samples containing > 500 total sequences
326 in each method were used for comparative β-diversity analysis between the three methods.
327 Consistency of observed vaginal community state types (CSTs) between libraries was tested
328 using Fleiss' *kappa* for inter-rater reliability, where $\kappa > 0.75$ indicated excellent agreement.
329 Complete agreement between all three methods was observed and samples clustered primarily
330 by vaginal community state type and subject as opposed to library preparation method or
331 sequencing platform ($\kappa = 1.0$, **Figure 2**, raw read count taxonomy tables are available in
332 **Supplementary Table 3**).
333
334 *Mock community libraries prepared via 2-Step PCR and sequenced on Illumina HiSeq are not*
335 *different than those sequenced on Illumina MiSeq*
336 In order to verify that consistency of results was not simply due to sample type, we also
337 compared the microbial compositions of the Zymobiomics Microbial DNA Standard obtained on
338 the Illumina HiSeq and MiSeq platforms. We used theoretical values reported by Zymo, as well
339 as compositional data produced from 16S rRNA gene V3-V4 regions amplicon libraries
340 sequenced on the Illumina MiSeq (prepared, sequenced, and provided by Zymo). The raw read

count taxonomy table is available in **Supplementary Table 4**. The distribution of Jensen-Shannon distances between Zymo-prepared, MiSeq-sequenced microbiota composition and theoretical composition did not significantly differ from the distribution of distances between the 2-Step-prepared, Illumina HiSeq-sequenced and theoretical microbiota compositions ($U = 29$, $p = 0.9578$, **Supplementary File 8**).

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

To compare the quality of amplicon sequence 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 sequence reads quality profiles were compared. The two runs had 183 samples in common. Significantly greater mean quality scores of both forward and reverse reads were observed for 1,194 samples run on the HiSeq platform compared to 276 samples run on the MiSeq platform ($p < 2.2 \times 10^{-16}$, **Figure 3**). 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 4**). Additionally, the HiSeq 2500 sequencing strategy was more cost efficient (nearly 40% cheaper per sample), assuming 2 lanes are run with 1,568 multiplexed samples per lane (**Table 4**). These results were also consistent across multiple sequencing runs (**Supplementary File 9**).

Discussion

Microbiome analyses large enough to achieve adequate statistical power are becoming more desirable, and reduced sequencing costs make these analyses feasible. Therefore, ultra-high-throughput sequencing capabilities are needed that do not sacrifice sequence quality. Ideally, such methods would allow for flexibility to target a diverse set of genes or gene regions (for example, ITS regions, the 16S and 23S rRNA genes, the *cpn60* gene [18, 19] among others) while also maintaining the ability to sequence longer amplicons (i.e., the 16S rRNA gene V3-V4 region). The method presented here improves on current technologies by producing consistent high-quality, 300 bp paired-end reads. Relative to the Illumina MiSeq platform, sequencing on the Illumina HiSeq platform produced a greater number of reads per sample, of significantly higher quality, with the capability to multiplex up to 2 x 1,568 samples. The innovative use of the Illumina HiSeq 2500 platform as presented here and by Muinck *et al.* [2] allow for ultra-high-throughput sequencing of amplicon libraries.

In addition, the 2-Step PCR library preparation method described here makes production of

375 sequencing libraries from various gene targets and samples containing low bacterial loads easy
376 through the use of unindexed, target specific primers in the first round of PCR. Amplification
377 success of samples with low bacterial loads are prone to amplification difficulties, and
378 amplification using the longer primers required in the traditional 1-Step protocol [1] exacerbate
379 the problem because of primer degradation and poor annealing due to the long overhang
380 unprimed sequence. Using the 2-Step PCR approach, we showed an 80% improvement of
381 samples containing low bacterial loads over the 1-Step PCR method. In addition, the shorter
382 primers used in the 2-Step PCR library protocol do not require PAGE purification, lowering the
383 overall cost of the method relative to the 1-Step PCR protocol. Other low-biomass environments
384 that could benefit from this 2-Step PCR procedure include blood and serum [20], respiratory
385 airways [21], skin [22], sub-seafloor sediments [23], and clean rooms [24].
386
387 In summary, to demonstrate the comparability of sequence datasets produced via different
388 methods, 16S rRNA gene V3-V4 regions sequence datasets were generated from low-biomass
389 vaginal samples women using both 1-Step and 2-Step PCR library construction methods and
390 the Illumina HiSeq and MiSeq sequencing platforms. Complete within-subject agreement
391 between the vaginal community state type assignments [3] were observed between all three
392 methods, though a greater number of significantly higher quality sequences were obtained from
393 the 2-Step PCR method sequenced on the Illumina HiSeq 2500 platform. We also show that
394 resulting microbial compositions of mock community samples are not significantly altered when
395 amplicon libraries are prepared using the 2-Step library preparation method and sequenced on
396 the Illumina HiSeq platform. We therefore conclude that while the 2-Step PCR preparation
397 method combined with the Illumina HiSeq 2500 platform is preferred, data generated by 1-Step
398 or 2-Step PCR and sequenced on the Illumina MiSeq or HiSeq 2500 platform can be combined
399 to successfully obtain meaningful conclusions about the environment and sample types of
400 interest (given that the same region is targeted).
401
402 Limitations:
403 The method is extremely high-throughput, and as such might not be suitable for small projects
404 unless these are combined with other samples. Producing a large number of samples ready for
405 pooling requires automation so that time from sample collection to data generation is still
406 reasonable. Overall, automation is required, and this approach might be suitable for microbiome
407 service cores where faster turn-around is needed and running many MiSeq runs is not a viable
408 option because of potential batch effect.

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. de Muinck EJ, Trosvik P, Gilfillan GD, Hov JR, Sundaram AYM: **A novel ultra high-throughput 16S rRNA gene amplicon sequencing library preparation method for the Illumina HiSeq platform**. *Microbiome* 2017, **5**(1):68.
3. 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.
4. 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.
5. Zhang R, Daroczy K, Xiao B, Yu L, Chen R, Liao Q: **Qualitative and semiquantitative analysis of Lactobacillus species in the vaginas of healthy fertile and postmenopausal Chinese women**. *J Med Microbiol* 2012, **61**(Pt 5):729-739.
6. Hillier SL, Lau RJ: **Vaginal microflora in postmenopausal women who have not received estrogen replacement therapy**. *Clin Infect Dis* 1997, **25 Suppl 2**:S123-126.
7. Liu CM, Aziz M, Kachur S, Hsueh PR, Huang YT, Keim P, Price LB: **BactQuant: an enhanced broad-coverage bacterial quantitative real-time PCR assay**. *Bmc Microbiol* 2012, **12**:56.
8. 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.
9. 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.
10. 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.
11. Fluidigm: **Access Array System for Illumina Sequencing Systems**. In.; 2018: 61-68.
12. 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.

13. 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.
14. Maaten Lvd, Hinton G: **Visualizing data using t-SNE**. *Journal of machine learning research* 2008, **9**(Nov):2579-2605.
15. Fleiss JL: **Measuring nominal scale agreement among many raters**. *Psychological bulletin* 1971, **76**(5):378.
16. 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.
17. Wickham H: **Ggplot2 : elegant graphics for data analysis**. New York: Springer; 2009.
18. Johnson LA, Chaban B, Harding JC, Hill JE: **Optimizing a PCR protocol for cpn60-based microbiome profiling of samples variously contaminated with host genomic DNA**. *BMC Res Notes* 2015, **8**:253.
19. Schellenberg J, Links MG, Hill JE, Hemmingsen SM, Peters GA, Dumonceaux TJ: **Pyrosequencing of chaperonin-60 (cpn60) amplicons as a means of determining microbial community composition**. *Methods Mol Biol* 2011, **733**:143-158.
20. 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.
21. 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.
22. Byrd AL, Belkaid Y, Segre JA: **The human skin microbiome**. *Nat Rev Microbiol* 2018, **16**(3):143-155.
23. 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.
24. 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

496

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**
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. Summary of sequencing results for vaginal samples. **Supplementary Figure 6** summarizes the pre-quality filtering per-cycle quality scores.

Library Preparation Method	**1-Step**	**2-Step**	
No. samples attempted to amplify	92	92	
No. samples successfully 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, $p < 2.2 \times 10^{-16}$

497

498

499

500

501

502

503

504

Table 4. 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	13,116 ± 479	49,851± 895
No. Samples	276	1194
% Chimeric Sequences Detected	2.8	7.7
Mean No. Non-chimeric, Assembled Sequences per Sample ± SE	12,737 ± 463	45,988 ± 787
Median Quality Score [Q25-Q75] – Forward Reads	35.7 [33-37]*	36.1 [35-38]*
Median Quality Score [Q25-Q75] – Reverse Reads	33.9 [25-36]†	33.2 [31-37]†
Cost of Sequencing per Sample (No. Multiplexed Samples)	$6.38 (384)	$3.99 (1,568)

* Significant. Wilcoxon Rank Sum W = 70352, $p < 2.2 \times 10^{-16}$
† Significant. Wilcoxon Rank Sum W = 76453, $p < 2.2 \times 10^{-16}$

505

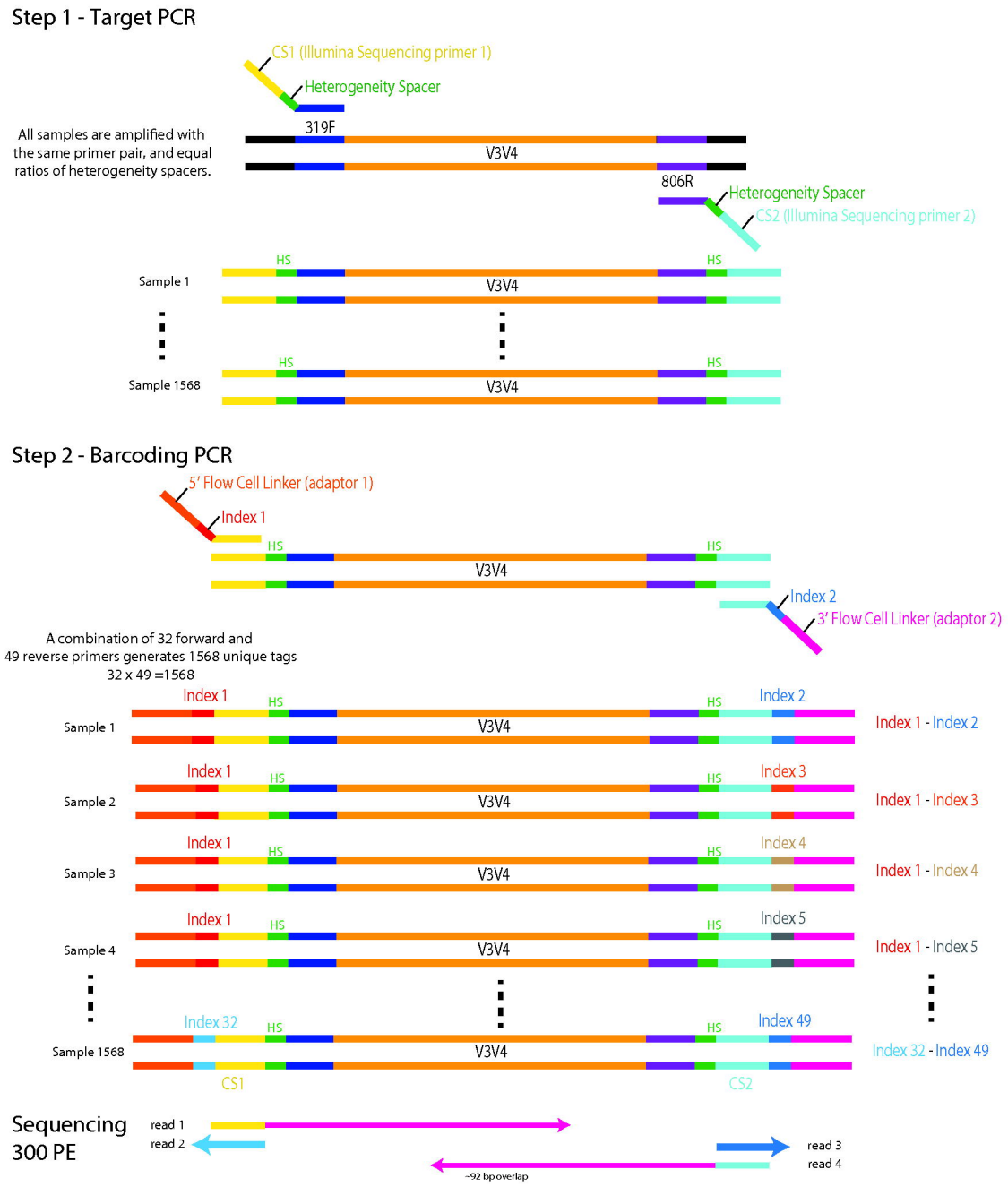

Figure 1. Illumina amplicon library preparation through 2-Step PCR amplification. In the Step 1 PCR, the target gene is amplified using primers that contain the heterogeneity space, and the CS1 Illumina Sequencing primers. The 2nd PCR targets the CS1 Illumina Sequencing primer to add the indices and the Illumina flow-cell linker sequence. Sequencing proceeds wherein Reads 1 and 4 contain the forward and reverse target gene sequence, respectively, and Reads 2 and 3 contain the first and second barcode indices, respectively.

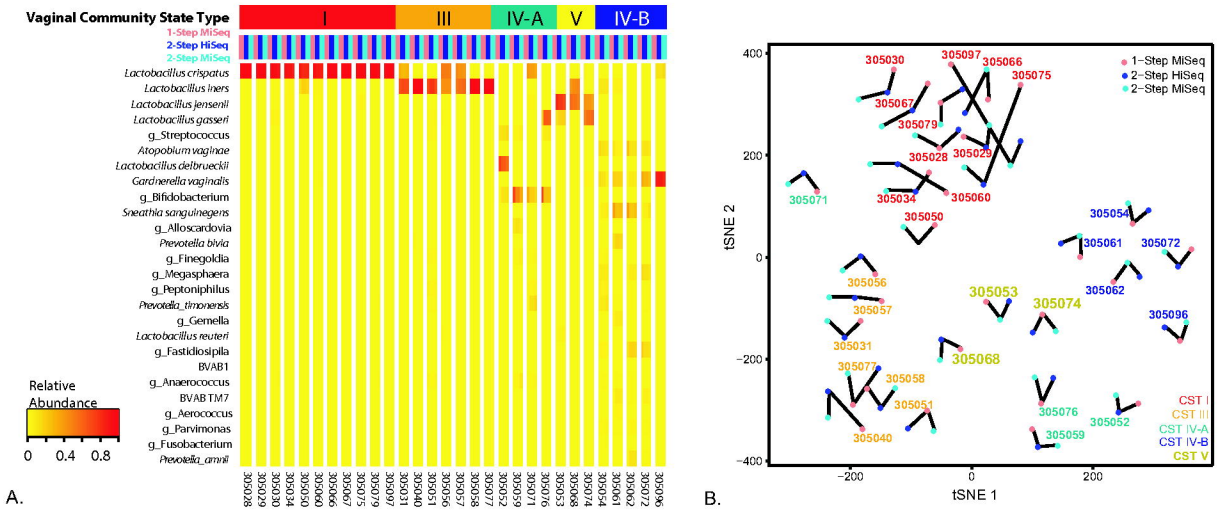

Figure 2. A Heatmap of taxon relative abundances (rows) of samples (columns). Subject samples are separated by white lines and samples are ordered by vaginal community state types and as follows: 1-Step MiSeq (pink), 2-Step HiSeq (blue), 2-Step MiSeq (aqua). B) tSNE representation of Jensen-Shannon distances between samples from the same subject. Samples primarily cluster by vaginal CST.

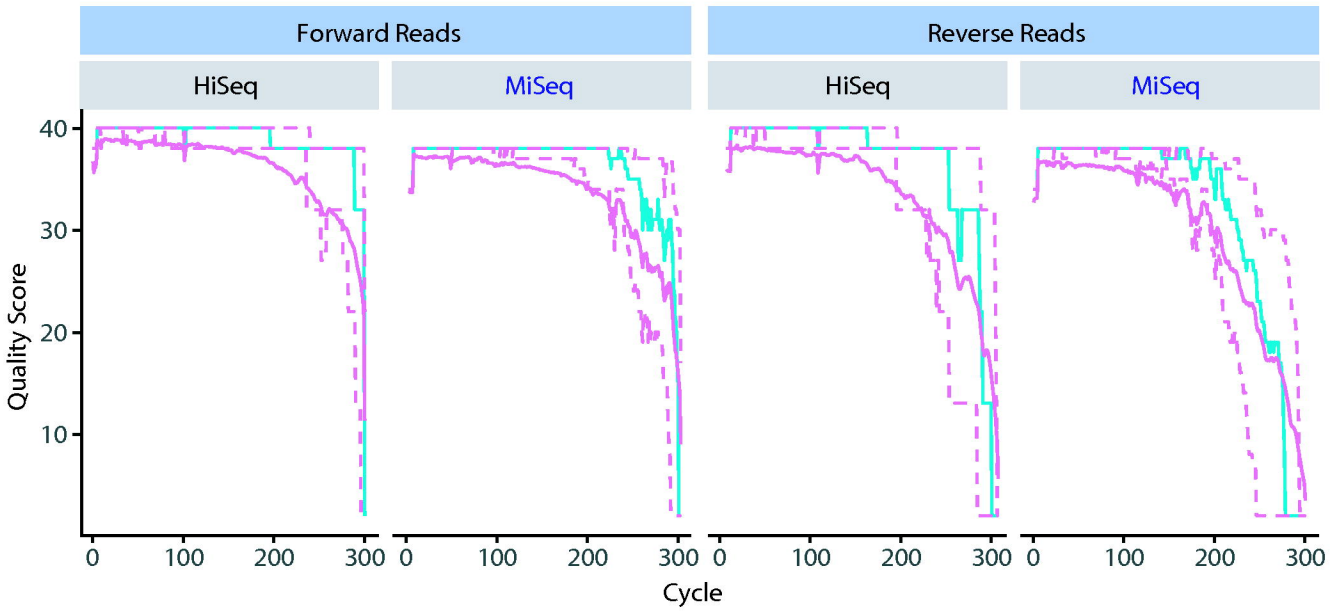

Figure 3. Forward and reverse read quality profiles for 300 cycles on the Illumina HiSeq (1,536 samples) and MiSeq (444 samples) platforms. Amplicon libraries were prepared using a 2-Step PCR method. Shown for each cycle are the mean quality score (green line), the median quality score (solid purple line), the quartiles of the quality score distribution (dotted purple lines).