The Influence of High Fidelity DNA Polymerase on 16S rRNA Gene Sequencing

Marc A Sze[1] and Patrick D Schloss[1†]

† To whom correspondence should be addressed: pschloss@umich.edu

1 Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI

Co-author e-mails:

- marcsze@med.umich.edu

Abstract

Background. Research has identified various methodological steps along the 16S rRNA gene survey pipeline that can change microbial community results. Although cycle number and high fidelity (HiFi) DNA polymerase are studied less often, they are still important sources of bias. Here, we critically examine both cycle number and HiFi DNA polymerase for biases that may influence downstream bacterial community results.

Methods. DNA from fecal samples (n = 4) were extracted using a PowerMag DNA extraction kit with a 10 minute bead beating step and amplified at 15, 20, 25, 30, and 35 cycles using Accuprime, Kappa, Phusion, Platinum, or Q5 HiFi DNA polymerase. Mock communities (technical replicates n = 4) consisting of previously isolated whole genomes of 8 different bacteria were also amplified using the same approach. First, the number of OTUs (Operational Taxonomic Units) were examined for both fecal samples and mock communities. Next, Bray-Curtis index, the error rate, sequence error prevalence, and chimera prevalence were investigated. Finally, the chimera prevalence correlation with number of OTUs was assessed.

Results. At 35 cycles there were significant differences between HiFi DNA polymerase for fecal samples (P-value < 0.0001). These HiFi dependent differences in the number of OTUs could be identified as early as 20 cycles in the mock communities (P-value = 0.002). Chimera prevalence varied by HiFi DNA polymerase and these differences were still observed after chimera removal using VSEARCH. Additionally, the chimera prevalence had a strong positive correlation with the number of OTUs and this association was not changed by chimera removal with VSEARCH.

Conclusions. Due to HiFi DNA polymerase dependent differences in the number of OTUs and chimera prevalence, common diversity metrics could have values that are not comparable across studies.

Introduction

Recently, there has been an increasing focus on standardizing methodological approaches in microbiota research (Kim et al., 2017; Hugerth & Andersson, 2017). Investigating ways that 16S rRNA gene surveys can be made more reproducible has been a predominant focal point of this push towards standardization (Lauber et al., 2010; Salter et al., 2014; Song et al., 2016; Gohl et al., 2016). Due to this focus, 16S rRNA gene sequencing has been much maligned for introduced biases. Yet, many of these same considerations also affect metagenomic sequencing (Nayfach & Pollard, 2016; Costea et al., 2017). Between the two approaches similar bias considerations include preservation media, storage conditions, DNA extraction kit, PCR, and sequence library preparation. For these overlapping considerations, biases identified for 16S rRNA gene sequencing will also likely influence metagenomic sequencing results.

The most commonly studied biases are preservation media, storage conditions, and DNA extraction kits. Although DNA extraction kits have consistently been shown to add bias to downstream analysis the current literature on preservation media and storage conditions has been mixed, with some studies showing biases while others do not (Lauber et al., 2010; Dominianni et al., 2014; Salter et al., 2014; Sinha et al., 2015; Song et al., 2016; Luo et al., 2016; Bassis et al., 2017; Costea et al., 2017). The study of these specific biases has become so large, aggregating them all together has become a difficult task with some researchers providing resources to actively track new findings (e.g. Microbiome Digest - https://microbiomedigest.com/microbiome-papers-collection/microbiome-techniques/sample-storage/). It is evident that these three sources of bias have been extensively covered within the literature but they are not the only sources that should be critically examined. The type of DNA polymerase chosen could also have a wide ranging affect on downstream results due to error rates and chimeras that may not be easily resolved using bioinformatic approaches.

A recent study in *Nature Biotechnology* found clear differences between normal and high fidelity (HiFi) DNA polymerase and that you could reduce error and chimera generation by optimizing the PCR protocol (Gohl et al., 2016). This study also found that, regardless of DNA polymerase, the number of Operational Taxonomic Units (OTUs) generated were not easily removed using the authors chosen bioinformatic pipeline (Gohl et al., 2016). It is natural to extend this line of inquiry and ask whether different HiFi DNA polymerase contribute different biases to downstream sequencing results. There is some reason to think that this may be the case since many of these HiFi DNA polymerase come from different families (e.g. *Taq* belongs to the family A polymerases) and may intrinsically have different error rates that cannot be completely removed with modifications (Ishino & Ishino, 2014). In this study we critically examine if any of five different HiFi DNA polymerases introduce significant biases into 16S rRNA gene surveys, if this is a cycle dependent phenomenon, and whether they can be removed using a standard bioinformatic pipeline.

Materials & Methods

Human and Mock Samples: A single fecal sample was obtained from 4 individuals who were part of the Enterics Research Investigational Network (ERIN). The processing and storage of these samples have been previously published (Seekatz et al., 2016). Other than confirmation that none of these individuals had a *Clostridium difficle* infection, clinical data and other types of meta data were not utilized or accessed for this study. All samples were extracted using the MOBIO^{TM} PowerMag Microbiome RNA/DNA extraction kit (now Qiagen, MD, USA). The ZymoBIOMICS^{TM} Microbial Community DNA Standard (Zymo, CA, USA) was used in this study and is made up of *Pseudomonas aeruginosa*, *Escherichia coli*, *Salmonella enterica*, *Lactobacillus fermentum*, *Enterococcus faecalis*, *Staphylococcus aureus*, *Listeria monocytogenes*, and *Bacillus subtilis* at equal genomic DNA abundance (http://www.zymoresearch.com/microbiomics/microbial-standards/zymobiomics-microbial-community-standards).

PCR Protocol: The five different HiFi DNA polymerases that were tested included AccuPrime^{TM} (ThermoFisher, MA, USA), KAPA HIFI (Roche, IN, USA), Phusion (ThermoFisher, MA, USA), Platinum (ThermoFisher, MA, USA), and Q5 (New England Biolabs, MA, USA). The PCR cycle conditions for Platinum and Accuprime followed a previously published protocol (Kozich et al., 2013) (https://github.com/SchlossLab/MiSeq_WetLab_SOP/blob/master/MiSeq_WetLab_SOP_v4.md). The HiFi DNA polymerase activation time was 2 minutes, unless a different activation was specified. For Kappa and Q5, the protocol previously published by Gohl and colleagues was used (Gohl et al., 2016). For Phusion, the company defined conditions were used except for extension time, where the Accuprime and Platinum settings were used.

Both fecal and mock samples cycle conditions started at 15 and increased by 5 up to 35 cycles with amplicons used at each 5-step increase for sequencing. The fecal PCR consisted of all 4 samples at 15, 20, 25, 30, and 35 cycles for each HiFi DNA polymerase (total samples = 100). The mock communities had 4 replicates at 15, 20, 25, and 35 cycles and 10 replicates for 30 cycles for all HiFi DNA polymerase (total samples = 130). No mock community sample had enough PCR product at 15 cycles for adequate 16S rRNA gene sequencing.

Sequence Processing: The mothur software program was utilized for all sequence processing steps (Schloss et al., 2009). Generally, the protocol followed what has been previously published (Kozich et al., 2013) (https://www.mothur.org/wiki/MiSeq_SOP). Two major differences from the stated protocol were the use of VSEARCH instead of UCHIME for chimera detection and the use of the OptiClust algorithm instead of average neighbor for OTU generation at 97% similarity (Edgar et al., 2011; Rognes et al., 2016; Westcott & Schloss, 2017). Sequence error was determined using the ‘seq.error’ command on mock samples after chimera removal and classification to the RDP to remove non-bacterial sequences (Schloss et al., 2009; Cole et al., 2013; Rognes et al., 2016).

Analysis Workflow: The total number of OTUs after sub-sampling was analyzed for both the fecal and mock community samples. Cycle dependent affects on Bray-Curtis indices were next assessed for the fecal samples looking at both overall cycle differences and within individual differences for the previous cycle (e.g. 20 versus 25, 25 versus 30, etc.). For these community based measures, the fecal samples were analyzed at 4 different sub-sampling sequence depths (1000, 5000, 10000, and 15000) while the mock community samples were analysed at 3 levels (1000, 5000, 10000). Next, based on these observations we wanted to analyze potential reasons for these differences. First, analysis of general sequence error rate, number of sequences with an error, and base substitution were assessed in the mock community for each HiFi DNA polymerase. After assessing these errors, the total number of chimeras was determined after sequence processing and correlations to the number of OTUs also assessed.

Statistical Analysis: All analysis was done with the R (v 3.4.3) software package (R Core Team, 2017). Data transformation and graphing was completed using the tidyverse package (v 1.2.1) and colors selected using the viridis package (v 0.4.1) (Garnier, 2017; Wickham, 2017). Differences in the total number of OTUs were analyzed using an ANOVA with a tukey post-hoc test. For the fecal samples the data was normalized to each individual by cycle number to account for the biological variation between people. Bray-Curtis distance matrices were generated using mothur after 100 sub-samplings at 1000, 5000, 10000, and 15000 sequence depth. The distance matrix data was analyzed using PERMANOVA with the vegan package (v 2.4.5) (Oksanen et al., 2017) and Kruskal-Wallis tests within R. For both error and chimera analysis, samples were tested using Kruskal-Wallis with a Dunns post-hoc test. Where applicable correction for multiple comparison utilized the Benjamini-Hochberg method (Benjamini & Hochberg, 1995).

Reproducible Methods: The code and analysis can be found here https://github.com/SchlossLab/Sze_PCRSeqEffects_XXXX_2017. The raw sequences can be found in the SRA at the following accession number SRP132931.

Results

The Number of OTUs is Dependent on HiFi DNA Polymerase: There was a HiFi DNA polymerase dependent difference that was consistent across sub-sampling [Figure 1]. Lower cycle numbers (15-20) tended to result in less differences between HiFI DNA polymerase versus higher cycle numbers (25, 30, and 35) [Figure 1]. At 35 cycles, all sub-sampling levels had significant differences between HiFi DNA polymerases (P-value < 0.0001) [Table S1]. Some differences between HiFi DNA polymerase we identifed at 25 and 30 cycles but the sub-sampling depth had to be 5000 or higher (P-value < 0.05) [Table S1]. Most of the differences observed at 35 cycles were between Platinum and other HiFi DNA polymerases, based on a Tukey post-hoc test (P-value < 0.05) [Table S2]

This HiFi DNA polymerase dependent difference in the number of OTUs was also observed in the mock community samples with the same DNA polymerases having high (Platinum) and low (Accuprime) number of OTUs [Figure 2 & Table S3]. In contrast to fecal samples, differences between HiFi DNA polymerase were observed as early as 20 cycles and as low as a sub-sampling depth of 1000 sequences (P-value = 0.002) [Table S3]. For both different cycle numbers and sub-sampling depths, the majority of differences were between Platinum and the other HiFi DNA polymerases [Table S4]. The lowest number of OTUs identified was from Accuprime^{TM} while the highest was from Platinum and this was consistent for both fecal and mock samples across sub-sampling depth and cycle number [Figure 1 & 2]. Based on these observations it is clear that different HiFi DNA polymerase result in different total number of OTUs observed within a sample.

Minimal Bray-Curtis Differences are Detected by Cycle Number: Overall, there were a few sub-sampling dependent differences in bacterial communities between cycle number . First, independent of HiFi DNA polymerase there were differences for the same fecal sample between 20 versus 25 cycles that was dependent on sub-sampling depth (sub-sampled to 1000 = 0.51 (0.4 - 0.79) (median (25% - 75% quantile)), sub-sampled to 5000 = 0.43 (0.33 - 0.63), sub-sampled to 10000 = 0.4 (0.24 - 0.43)) [Figure 3A]. Second, where data is available for the mock communities, there were larger difference between 20 and 25 cycles (sub-sampled to 1000 = 0.88 (0.42 - 0.91)) [Figure 3B]. Third, using PERMANOVA to test for differences within HiFi DNA polymerase groups based on cycle number, only Phusion had cycle dependent differences at 1000 and 5000 sub-sampling depth (P-value = 0.03 and 0.01, respectively). Phusion was one of only two HiFi DNA polymerases that that had fecal samples for the 1000 sub-sampling depth at 15 cycles. These stated differences between the next 5-cycle increment do not persist once 25 cycles are reached [Figure 3].

Next, we assessed whether there were any differences between 5-cycle increments within each individual or sample replicate. For both fecal and mock samples, we found that there was no detectable difference in Bray-Curtis index when comparing to the previous 5-cycle increment (P-value > 0.05). However, Phusion at 1000 sub-sampling depth had a P-value = 0.02 before multiple comparison correction. Similar to the PERMANOVA analysis, at higher sub-sampling depths these differences in Bray-Curtis indices disappear. Overall, HiFi DNA polymerase differences in Bray-Curtis index are dependent on both sub-sampling depth and cycle number.

Sequence Error is Dependent on both HiFi DNA Polymerase and Cycle Number: Differences by HiFi DNA polymerase in the median average per base error varied without a clear pattern across sub-sampling depth [Table S5]. The highest per base median average error rates were for the Kappa HiFi DNA polymerase [Figure 4]. This error rate was minimally affected by both the ‘pre.cluster’ step and chimera removal by VSEARCH [Figure 4]. The differences in the per base error rate between the various HiFi DNA polymerase was cycle dependent with Platinum having the largest differences versus other HiFi DNA polymerases [Figure 4B-C and Table S6]. The total sequences with at least one error had was also cycle number dependent and mostly alleviated by the use of the ‘pre.cluster’ step [Figure S1]. Differences in sequences with an error before this ‘pre.cluster’ step were due to large differences in Accuprime^{TM} and Platinum versus the other HiFi DNA polymerases [Figure S1 & Table S7 & S8]. Investigation of whether there were HiFi DNA polymerase dependent effects on base substitution found that there were no biases in the types of substitution made [Figure S2]. Although though sequence error is dependent on HiFi DNA polymerase it can be corrected using exsiting bioinformatic approaches.

Chimeric Sequences are HiFi DNA Polymerase Dependent and Correlate with Number of OTUs: At all levels of sub-sampling and cycle numbers there were significant differences in the chimera prevalence based on HiFi DNA polymerase used (P-value < 0.05) [Table S9]. Differences between Platinum and all other HiFi DNA polymerases accounted for the majority of these differences and was independent of cycle number and sub-sampling depth [Table S10]. Across sub-sampling depth and cycle number Accuprime^{TM} had the lowest chimera prevalence of all the HiFi DNA Polymerases regardless of whether ‘pre.cluster’ or chimera removal with VSEARCH was used [Figure 5].

For all HiFi DNA polymerases, a positive correlation was observed between chimeric sequences and number of OTUs, with this correlation being strongest for Accuprime, Platinum and Phusion HiFi DNA Polymerase [Figure 6]. The R^2 value between the number of OTUs and chimeric sequences did not change with the use of ‘pre.cluster’ or with the removal of chimeras using VSEARCH [Figure 6]. Taken together, this data suggests that a strong correlation exists between the number of OTUs and the prevalence of chimeric sequences.

Discussion

These observations build upon previous studies (Gohl et al., 2016) by showing that different HiFi DNA polymerases have significant differences in the number of OTUs and that changes in total OTUs correlate with chimeras not removed after sequence processing [Figure 1-2 & 5]. HiFi DNA polymerase dependent differences in total number of OTUs are important to consider since many diversity metrics rely on this measure as part of their calculations. Our observations show that HiFi DNA polymerase can have a noticeable affect on the OTUs generated and these differences are consistent across sub-sampling depth and PCR cycle number [Figure 2-4]. Importantly, high biomass samples may underestimate the biases present within low biomass samples. An indication that this may be the case is the earlier detection of differences in total number of OTUs between HiFi DNA polymerases in the mock versus fecal samples.

Although we did not observe strong differences in the Bray-Curtis index the data suggests that there may be differences between 15 and 20 cycles versus higher cycle numbers, such as 30 cycles, that are commonly used. There was few differences within individuals between corresponding 5-cycle increments (e.g. 15 to 20, 20 to 25, etc.). However, there is a clear trend that suggests that the community at 20 cycles is different then the community at 25 cycles [Figure 3]. These findings, in conjunction with the PERMANOVA results, suggest that cycle number can change bacterial community calculations but that these differences are minimal once 25 cycles are reached. Further, for some HiFI DNA polymerases, increasing the sub-sampling depth may reduce some of these observed community differences at lower cycle numbers.

Increasing the cycle number also exacerbated chimera prevalence differences between the different HiFi DNA polymerases [Figure 5]. The chimera prevalence was strongly correlated with the number of OTUs which is relied upon heavily for many community metric calculations. However, Bray-Curtis analysis with PERMANOVA showed few differences based on HiFi DNA polymerase. It is possible that many of the increased number of OTUs are not highly abundant allowing the Bray-Curtis index to be able to successfully down-weight these respective OTUs (Minchin, 1987). The choice of downstream diversity metric could be an important consideration in helping to mitigate these observed changes due to high chimera prevalence in specific HiFi DNA polymerases.

Our observations suggest that there are clear HiFi DNA polymerase dependent differences in both per base error rate and chimera prevalence that cannot be removed using bioinformatic approaches [Figure 4 & 5]. Although it may be a natural assumption that the variation may be due to the DNA polymerase family, the highest chimera rate, from Platinum, was a family A polymerase while the lowest, from Accuprime, was also an A family polymerase (Ishino & Ishino, 2014). In fact, from the material safety data sheet (MSDS), the differences between the two is not immediately apparent. Both Accuprime and Platinum contain a recombinant *Taq* DNA polymerase, a *Pyrococcus* spp GB-D polymerase and a platinum *Taq* antibody. It is possible that differences in how the recombinant *Taq* was generated could be the main reason for the differences in chimera rate.

Conclusion

Our findings show that the number of OTUs will be dependent on both HiFi DNA polymerase and cycle number chosen. Care should be taken when choosing a HiFi DNA polymerase for 16S rRNA gene surveys since intrinsic differences can change the number of OTUs observed and influence diversity based metrics that do not down-weight rare observations.

Acknowledgements

The authors would like to thank all the study participants in ERIN whose samples were utilized. We would also like to thank Judy Opp and April Cockburn for their effort in sequencing the samples as part of the Microbiome Core Facility at the University of Michigan. Salary support for Marc Sze came from the Canadian Institute of Health Research and the Michigan Institute for Clinical and Health Research Postdoctoral Translational Scholar Program.

References

Bassis CM., Nicholas M. Moore., Lolans K., Seekatz AM., Weinstein RA., Young VB., Hayden MK. 2017. Comparison of stool versus rectal swab samples and storage conditions on bacterial community profiles. *BMC Microbiology* 17. DOI: 10.1186/s12866-017-0983-9.

Benjamini Y., Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:289–300.

Cole JR., Wang Q., Fish JA., Chai B., McGarrell DM., Sun Y., Brown CT., Porras-Alfaro A., Kuske CR., Tiedje JM. 2013. Ribosomal database project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research* 42:D633–D642. DOI: 10.1093/nar/gkt1244.

Costea PI., Zeller G., Sunagawa S., Pelletier E., Alberti A., Levenez F., Tramontano M., Driessen M., Hercog R., Jung F-E., Kultima JR., Hayward MR., Coelho LP., Allen-Vercoe E., Bertrand L., Blaut M., Brown JRM., Carton T., Cools-Portier S., Daigneault M., Derrien M., Druesne A., Vos WM de., Finlay BB., Flint HJ., Guarner F., Hattori M., Heilig H., Luna RA., Hylckama Vlieg J van., Junick J., Klymiuk I., Langella P., Chatelier EL., Mai V., Manichanh C., Martin JC., Mery C., Morita H., O'Toole PW., Orvain C., Patil KR., Penders J., Persson S., Pons N., Popova M., Salonen A., Saulnier D., Scott KP., Singh B., Slezak K., Veiga P., Versalovic J., Zhao L., Zoetendal EG., Ehrlich SD., Dore J., Bork P. 2017. Towards standards for human fecal sample processing in metagenomic studies. *Nature Biotechnology*. DOI: 10.1038/nbt.3960.

Dominianni C., Wu J., Hayes RB., Ahn J. 2014. Comparison of methods for fecal microbiome biospecimen collection. *BMC Microbiology* 14:103. DOI:

10.1186/1471-2180-14-103.

Edgar RC., Haas BJ., Clemente JC., Quince C., Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27:2194–2200. DOI: 10.1093/bioinformatics/btr381.

Garnier S. 2017. *Viridis: Default color maps from 'matplotlib'*.

Gohl DM., Vangay P., Garbe J., MacLean A., Hauge A., Becker A., Gould TJ., Clayton JB., Johnson TJ., Hunter R., Knights D., Beckman KB. 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature Biotechnology* 34:942–949. DOI: 10.1038/nbt.3601.

Hugerth LW., Andersson AF. 2017. Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Frontiers in Microbiology* 8. DOI: 10.3389/fmicb.2017.01561.

Ishino S., Ishino Y. 2014. DNA polymerases as useful reagents for biotechnology â the history of developmental research in the field. *Frontiers in Microbiology* 5. DOI: 10.3389/fmicb.2014.00465.

Kim D., Hofstaedter CE., Zhao C., Mattei L., Tanes C., Clarke E., Lauder A., Sherrill-Mix S., Chehoud C., Kelsen J., Conrad M., Collman RG., Baldassano R., Bushman FD., Bittinger K. 2017. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* 5. DOI: 10.1186/s40168-017-0267-5.

Kozich JJ., Westcott SL., Baxter NT., Highlander SK., Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq illumina sequencing platform. *Applied and Environmental Microbiology*

79:5112–5120. DOI: 10.1128/aem.01043-13.

Lauber CL., Zhou N., Gordon JI., Knight R., Fierer N. 2010. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiology Letters* 307:80–86. DOI: 10.1111/j.1574-6968.2010.01965.x.

Luo T., Srinivasan U., Ramadugu K., Shedden KA., Neiswanger K., Trumble E., Li JJ., McNeil DW., Crout RJ., Weyant RJ., Marazita ML., Foxman B. 2016. Effects of specimen collection methodologies and storage conditions on the short-term stability of oral microbiome taxonomy. *Applied and Environmental Microbiology* 82:5519–5529. DOI: 10.1128/aem.01132-16.

Minchin PR. 1987. An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio* 69:89–107. DOI: 10.1007/bf00038690.

Nayfach S., Pollard KS. 2016. Toward accurate and quantitative comparative metagenomics. *Cell* 166:1103–1116. DOI: 10.1016/j.cell.2016.08.007.

Oksanen J., Blanchet FG., Friendly M., Kindt R., Legendre P., McGlinn D., Minchin PR., O'Hara RB., Simpson GL., Solymos P., Stevens MHH., Szoecs E., Wagner H. 2017. *Vegan: Community ecology package*.

R Core Team. 2017. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rognes T., Flouri T., Nichols B., Quince C., Mahé F. 2016. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* 4:e2584. DOI: 10.7717/peerj.2584.

Salter SJ., Cox MJ., Turek EM., Calus ST., Cookson WO., Moffatt MF., Turner P., Parkhill J., Loman NJ., Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* 12. DOI:

10.1186/s12915-014-0087-z.

Schloss PD., Westcott SL., Ryabin T., Hall JR., Hartmann M., Hollister EB., Lesniewski RA., Oakley BB., Parks DH., Robinson CJ., Sahl JW., Stres B., Thallinger GG., Horn DJV., Weber CF. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75:7537–7541. DOI: 10.1128/aem.01541-09.

Seekatz AM., Rao K., Santhosh K., Young VB. 2016. Dynamics of the fecal microbiome in patients with recurrent and nonrecurrent clostridium difficile infection. *Genome Medicine* 8. DOI: 10.1186/s13073-016-0298-8.

Sinha R., Chen J., Amir A., Vogtmann E., Shi J., Inman KS., Flores R., Sampson J., Knight R., Chia N. 2015. Collecting fecal samples for microbiome analyses in epidemiology studies. *Cancer Epidemiology Biomarkers & Prevention* 25:407–416. DOI: 10.1158/1055-9965.epi-15-0951.

Song SJ., Amir A., Metcalf JL., Amato KR., Xu ZZ., Humphrey G., Knight R. 2016. Preservation methods differ in fecal microbiome stability, affecting suitability for field studies. *mSystems* 1:e00021–16. DOI: 10.1128/msystems.00021-16.

Westcott SL., Schloss PD. 2017. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. *mSphere* 2:e00073–17. DOI: 10.1128/mspheredirect.00073-17.

Wickham H. 2017. *Tidyverse: Easily install and load 'tidyverse' packages*.

Figure 1: Normalized Fecal Number of OTUs. The x-axis represents the different sub-sampling depths used and the y-axis is the normalized within individual number of OTUs. The red line represents the overall mean Z-score normalized number of OTUs for each respective HiFi DNA polymerase. The dashed black line represents the overall Z-score normalized mean number of OTUs.

Figure 2: Mock Sample Variability in Number of OTUs based on HiFi DNA Polymerase. A) Sub-sampled to 1000 reads. B) Sub-sampled to 5000 reads. C) Sub-sampled to 10000 reads. The dotted line represents the number of OTUs generated when the mock reference sequences are run through the pipeline.

Figure 3: Community Differences by Five-Cycle Intervals and Sub-sampling Depth. A) Fecal samples within person difference based on the next 5-cycle PCR interval. B) Mock samples within replicate difference based on the next 5-cycle PCR interval.

Figure 4: HiFi DNA Polymerase Per Base Error Rate in Mock Samples. A) Error rate before the merger of sequences with pre.cluster and the removal of chimeras with VSEARCH. B) Error rate before the removal of chimeras with VSEARCH. C) Full pipeline. The error bars represent the 75% interquartile range of the median.

Figure 5: HiFi DNA Polymerase Chimera Prevalence in Mock Samples. A) Chimera sequence percentage before the merger of sequences with pre.cluster and the removal of chimeras with VSEARCH. B) Chimera sequence percentage before the removal of chimeras with VSEARCH. C) Full pipeline. The error bars represent the 75% interquartile range of the median.

Figure 6: The Correlation between Number of OTUs and Chimeras. A) Correlation before the merger of sequences with pre.cluster and the removal of chimeras with VSEARCH. B) Correlation before the removal of chimeras with VSEARCH. C) Correlation with full pipeline.

Figure S1: HiFi DNA Polymerase Sequence Error Prevalence in Mock Samples. A) Sequence error prevalence before the merger of sequences with pre.cluster and the removal of chimeras with VSEARCH. B) Sequence error prevalence before the removal of chimeras with VSEARCH. C) Full pipeline. The error bars represent the 75% interquartile range of the median.

Figure S2: HiFi DNA Polymerase Nucleotide Subsitutions in Mock Samples.