Error Introduced into 16S rRNA Gene Sequencing Results Varies by High Fidelity DNA Polymerase Used

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. It is challenging to compare 16S rRNA gene sequencing data across studies and one of the reasons for this is due to error. There are many different places throughout the workflow where error can be introduced into the pipeline. Here, we focus on studying how the number of cycles and high fidelity (HiFi) DNA polymerase introduce error by varying cycle number and polymerase used to amplify 16S rRNA genes in human fecal and mock community samples.

Methods. We extracted DNA from fecal samples (n=4) 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. Amplification of mock communities (technical replicates n=4) consisting of previously isolated whole genomes of 8 different bacteria used the same approach. The analysis initially examined the number of Operational Taxonomic Units (OTUs) for fecal samples and mock communities. It also assessed polymerase dependent differences in the Bray-Curtis index, error rate, sequence error prevalence, chimera prevalence, and the correlation between chimera prevalence and number of OTUs.

Results. When analyzing fecal samples we observed that the range in the number of OTUs detected was not consistent between HiFi DNA polymerases at 35 cycles (Accuprime = 84 - 106 (min - max) versus Phusion = 84 - 136). Additionally, the median number of OTUs vaired by HiFi DNA polymerase used (P-value < 0.0001). When analyzing mock community samples the variation in the number of OTUs detected by the polymerases was observable as early as 20 cycles (P-value = 0.002). There also was a large range in the number of OTUs amplified by the polymerases at 35 cycles (Accuprime = 15 - 20 versus Phusion = 14 - 73). Chimera prevalence in mock communities varied by polymerase with differences being most notable at 35 cycles (Kappa = 5.71% (median) versus Platinum = 26.62%) and this variation persisted after chimera removal using VSEARCH. We also observed positive correlations between chimera prevalence and the number of OTUs with Platinum having the highest (Rˆ2 = 0.974) and Kappa having the worst (Rˆ2 = 0.478).

Conclusions. Although the variation in the number of OTUs in fecal samples could be due to certain polymerases capturing the biological variability better than others, this is unlikely to be the

main reason for our observed differences. In mock community samples, the strong correlation between chimera prevalence and the number of OTUs suggests that this is the main reason for differences between the polymerases. Ultimately, this variation makes comparison across studies difficult and care should be exercised when choosing the polymerase and number of cycles in 16S rRNA gene sequencing studies.

Introduction

The bacterial community is reported to vary between case and control for a number of diseases (Turnbaugh et al., 2008; Sze et al., 2015; Baxter et al., 2016; Bonfili et al., 2017). However, for diseases like obesity, the taxa identified have varied widely depending on the study (Turnbaugh et al., 2008; Zupancic et al., 2012). Some of this variation could be due to error introduced during the 16S rRNA gene sequencing workflow. Yet, standardizing a 16S rRNA gene sequencing workflow will ultimately result in a standardized and reproducible bias due to choices made on the methods used for preservation, extraction, PCR, and sequencing. Within this context, all 16S rRNA gene sequencing methods are biased even when these workflows are standardized to increase reproducibility. In order to interpret specific studies within the broader context of the overall field, assessing error at different parts of the 16S rRNA gene sequencing workflow is critical.

A typical 16S rRNA gene sequencing workflow can be divided into preservation, extraction, PCR, and sequencing steps. The preservation and extraction stages of the 16S rRNA gene sequencing workflow have been the most extensively studied (Salter et al., 2014; Song et al., 2016; Bassis et al., 2017; Kim et al., 2017). For preservation and extraction stages of the workflow, it has been consistently found that there are biases based on the kits used, but that these differences are smaller than the overall biological difference measured between samples with different kits (Song et al., 2016; Bassis et al., 2017). Since these studies use the same PCR approach while varying preservation or extraction method, the contribution of PCR bias to this overall workflow is not well characterized.

There is a large body of literature that shows there are biases due to primer and number of cycles chosen for the PCR stage of 16S rRNA gene sequencing (Eckert & Kunkel, 1991; Burkardt, 2000). Primers have variable region dependent binding affinities which causes an inability to detect specific bacteria (e.g. V1-V3 does not detect *Haemophilus influenzae* and V3-V5 does not detect *Propionibacterium acnes*) (Sze et al., 2015 (Table S4); Meisel et al., 2016). Another source of error is the selective amplification of AT-rich over GC-rich sequences which exaggerate the difference between 16S rRNA genes higher in AT versus those higher in GC content (Polz & Cavanaugh, 1998). Many of these sources of biases are made worse as the number of cycles increases (Wang &

Wang, 1996; Haas et al., 2011; Kebschull & Zador, 2015). For example, both amplification error and non-specific amplification (e.g. incorrect amplicon size products) also can increase as the number of cycles increases. This will increase the number of Operational Taxonomic Units (OTUs) observed and drastically change the values obtained from commonly used diversity measures (Acinas et al., 2005; Santos et al., 2016). Additionally, as the number of cycles increases more chimeras can form from an aborted extension step that causes a priming error and subsequent secondary extension (Haas et al., 2011). These chimeras will artificially increase community diversity by increasing the number of OTUs that are observed (Haas et al., 2011). In addition to these sources of errors, there also are multiple families of DNA polymerases that have their own error rate and proof reading capacity (Ishino & Ishino, 2014). Interestingly, the influence that these different DNA polymerases can have on the observed 16S rRNA gene sequencing results have not been well studied like some of the other sources of PCR-based bias.

A recent study found differences in the number of OTUS and chimeras between normal and high fidelity DNA polymerases (Gohl et al., 2016). The authors could reduce the difference between the two polymerases by optimizing the annealing and extension steps within the PCR protocol (Gohl et al., 2016). Yet, within this study there was no comparison made between different high fidelity DNA polymerases. Due to this gap, it is natural to extend this line of inquiry and test if biases in the number of OTUs and chimeras also are dependent on the type of high fidelity DNA polymerase. This study will investigate how high fidelity DNA polymerases can bias observed bacterial community results derived from 16S rRNA gene sequencing. We will accomplish this by examining the number of OTUs, error rate, number of sequences with an error, and chimera prevalence at varying number of cycles in five different high fidelity DNA polymerases

Materials & Methods

Human and Mock Samples: Fecal samples were obtained from 4 individuals who were part of the Enterics Research Investigational Network (ERIN). The processing and storage of these samples were previously published (Seekatz et al., 2016). Other than confirmation that none of these individuals had a *Clostridium difficile* infection, clinical data and other types of meta data were not utilized or accessed for this study. All samples were extracted using the MOBIO™ PowerMag Microbiome RNA/DNA extraction kit (now Qiagen, MD, USA). The ZymoBIOMICS™ Microbial Community DNA Standard (Zymo, CA, USA) was used for mock communities and was 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 high fidelity DNA polymerases (hereto referred to as polymerases) that were tested included AccuPrime™ (ThermoFisher, MA, USA), KAPA HIFI (Roche, IN, USA), Phusion (ThermoFisher, MA, USA), Platinum (ThermoFisher, MA, USA), and Q5 (New England Biolabs, MA, USA). The polymerases activation time was 2 minutes, unless a different activation was specified by the manufacturer. The annealing and extension time 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). For Kappa and Q5, the annealing and extension time were from a previously published protocol (Gohl et al., 2016). For Phusion, the company defined activation and annealing times were used while the extension time followed the Accuprime and Platinum settings.

The number of cycles in the PCR for fecal and mock samples started at 15 and increased by 5 up to 35 cycles, with amplicons used at each 5-step increase for sequencing. The PCR of fecal DNA samples consisted of all 4 samples at 15, 20, 25, 30, and 35 cycles for each polymerase (total sample n=100). The mock communities had 4 replicates at 15, 20, 25, and 35 cycles and 10 replicates for 30 cycles for all polymerases (total samples n=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 used for all sequence processing steps (Schloss et al., 2009). The protocol has been previously published (Kozich et al., 2013) (https://www.mothur.org/wiki/MiSeq_SOP). Two major differences from the published 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 to compare back to the reference 16S sequences (Schloss et al., 2009; Cole et al., 2013; Rognes et al., 2016).

Analysis Workflow: To adjust for unequal sequencing, all samples were rarefied to 1000 sequences for downstream analysis. The total number of OTUs was analyzed for both the fecal and mock community samples. For fecal samples, cycle dependent affects on Bray-Curtis indices were assessed for cycle group and within individual differences from the previous cycle (e.g. 20 versus 25, 25 versus 30). Based on these observations we analyzed potential reasons for these differences. Analysis of the mock community of each polymerase for sequence error rate, number of sequences with an error, base substitution, and numbers of chimeras before and after chimera removal with VSEARCH was assessed. Additionally, the correlation between the number of chimeras and the number of OTUs was also assessed.

Statistical Analysis: All analysis was done with the R (v 3.4.4) 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 are Dependent on Polymerase: Differences in the range of the number of OTUs detected for fecal samples is dependent on the polymerase used (e.g. Accuprime at 35 cycles = 84 - 106 versus Phusion at 35 cycles = 84 - 136) [Figure 1]. Additionally, there is a trend for lower number of cycles (15-20) to result in a reduced range in the number of OTUs versus higher number of cycles (25, 30, and 35) for all polymerases (e.g. Phusion at 15 cycles = 10 - 19 versus Phusion at 35 cycles = 84 - 136) [Figure 1]. There is an overall difference in the number of OTUs detected within fecal samples between polymerases at 35 cycles (F-stat > 16.35, P-value = 9.7e-05) [Table S1]. Using a Tukey post-hoc test to identify which polymerase groups were different from each other at 35 cycles, no difference was found (P-value > 0.05) [Table S2]. The polymerase dependent difference in the range of the number of OTUs also was observed in the mock community samples [Figure 2]. The closest the polymerases came to the total of 8 OTUs created by the mock reference 16S sequences was at 25 and 30 cycles [Figure 2]. Regardless of if fecal or mock communities were used, the same polymerases generated high and low number of OTUs and this was consistent across the number of cycles used [Figure 1-2 & Table S1-S2]. In contrast to the results obtained with fecal samples, differences between polymerases for the number of OTUs created were observed as early as 20 cycles in the mock community (F-stat = 15.82, P-value = 0.002) [Table S1]. Using a Tukey post-hoc test, the majority of differences for the number of OTUs detected in the mock community is largely due to Kappa and Platinum differences versus the other polymerases across the different number of cycles [Table S2]. Based on these observations in fecal and mock communities, it is clear that using different polymerases result in a different total number of OTUs within a sample.

Minimal Bray-Curtis Differences are Detected and are Dependent on Cycle Number: A few small differences based on the number of cycles were detected in overall bacterial community composition. Within the same fecal sample and independent of polymerases, there are large differences in the community composition between 20 and 25 cycles [Figure 3A]. Additionally, when data was available for the mock communities, there are also large observed differences between 20 and 25 cycles [Figure 3B]. These community differences within a sample disappear

when comparing 25 to 30 cycles and do not persist past 25 cycles [Figure 3]. Although these trends are clearly noticeable, there is no detectable difference in Bray-Curtis index when comparing to the previous 5-cycle increment for both fecal and mock communities after multiple comparison correction (P-value > 0.05). Using PERMANOVA to test for community differences based on the number of cycles within polymerases, only Phusion had cycle dependent differences (P-value = 0.03. For fecal samples, Phusion was one of two polymerases that had enough sequences to be rarefied to 1000 at 15 cycles. Overall, these data suggest that there are small polymerase differences in Bray-Curtis index that are dependent on the number of cycles.

Sequence Error is Dependent on both Polymerase and Cycle Number: Differences in the median error rate varied by polymerase, with Kappa being the highest [Figure 4 & Table S3]. The differences in the median error rate between the polymerases also is dependent on the number of cycles, with Platinum and Kappa having the most differences versus other polymerases [Figure 4 and Table S4]. The total sequences with at least one error also is dependent on the number of cycles [Table S3 & S4]. These differences in sequences with at least one error is mostly due to differences in Accuprime^{TM} and Kappa versus the other polymerases [Table S3 & S4]. The differences in error rates were not due to polymerase dependent differences in base substitution rate [Figure S1]. These results suggest that sequence error is dependent on polymerase and the number of cycles, but is not attributable to a specific type. Based on these results, we examined whether chimeras also were dependent on polymerase and whether this also could affect the number of OTUs.

Prevalence of Chimeric Sequences are Polymerase Dependent and Correlate with the Number of OTUs: There is significant differences in the chimera prevalence based on polymerase at all the number of cycles used (P-value < 0.05) [Table S3]. Differences in chimera prevalence between Platinum and all other polymerases accounted for the majority of these differences [Table S4]. Accuprime^{TM} had the lowest chimera prevalence of all polymerases regardless of whether chimera removal with VSEARCH was used [Figure 5A & 5B]. Additionally, there was a plateau in the total percent of chimeras that were removed that was similar for all polymerases [Figure 5C]. A positive correlation was observed between chimeric sequences and the number of OTUs for all polymerases [Figure 6]. This positive correlation was strongest for Accuprime^{TM}, Platinum, and

Phusion [Figure 6]. This data suggests that chimera prevalence depends on polymerase used and confirms that the number of OTUs is dependent on the prevalence of these chimeric sequences.

Discussion

In this study we show that the number of OTUs, error rate, and chimera prevalence depends on polymerase used [Figure 1-5]. These differences are important because many diversity metrics rely on the number of OTUs or other measures dependent on error rate and chimera prevalence as part of their metric calculations (e.g. richness). Additionally, the earlier detection of differences in total number of OTUs between polymerases in the mock versus fecal samples might indicate that high biomass samples may underestimate the biases present within low biomass samples. We observed that undetected chimeras that were not identified and removed using standard bioinformatic approaches cause many of these differences. This suggests that some of the specific diversity differences between studies can be attributed to differences in polyermase used. Based on these observations, metrics that measure within sample diversity like richness depend on polymerase but this may not be the case for metrics that assess between sample diversity.

There were few differences that depend on polymerase for between sample diversity, as measured by the Bray-Curtis index. Our observations generally found no differences in the overall bacterial community composition for the number of cycles used. One possible reason for this outcome is that our study did not have enough power to detect differences due to low sample number in each group. Another reason is that many of the OTUs are likely not highly abundant, allowing the Bray-Curtis index to be able to successfully down-weight chimeric OTUs (Minchin, 1987). The choice of downstream diversity metric could be an important consideration in helping to mitigate these observed polymerase dependent differences in chimera prevalence. Metrics that solely use presence/absence of OTUs (e.g. Jaccard (Real & Vargas, 1996)) may be less robust to chimera prevalence and by extension total number of OTU differences in polymerases. When choosing a distance metric, careful consideration of the biases introduced from the PCR step of the 16S rRNA gene sequencing workflow need to be taken into account. With differences in the number of OTUs and chimera prevalence depending on polymerase used, it might be easier to avoid specific DNA polymerase families altogether.

Although the variation in error rate and chimera prevalence may be due to the DNA polymerase family, this is unlikely to be the only contributor. For example, the highest and lowest chimera

rates both belonged to a family A polymerase (Platinum and Accuprime^{TM} respectively) (Ishino & Ishino, 2014). Additionally, based on the material safety data sheet the differences between the two polymerases are not immediately apparent. Both polymerases contain a recombinant *Taq* DNA polymerase, a *Pyrococcus* spp GB-D polymerase and a platinum *Taq* antibody. With everything else being equal, it is possible that differences in how the recombinant *Taq* was generated could be a contributing factor for the differences observed between the polymerases. We are unlikely to avoid adding polymerase dependent bias to 16S rRNA gene sequencing results, however, these differences may not be large enough to mask the actual biological signal.

The majority of polymerases we studied add small increases in the number of OTUs and chimera prevalence and may be masked by biological differences. The sequence error introduced by the polymerase is also small and likely to be smaller than the biological variation within a specific study, which would be consistent with previous findings for preservation and DNA extraction methods (Salter et al., 2014; Song et al., 2016; Luo et al., 2016). The choice of polymerase should be an important consideration in the creation of a 16S rRNA gene sequencing workflow because using different polymerases can add error and bias to the downstream observations. Although standardization of the workflow may partially solve this problem by introducing a consistent bias to all samples, it does come at a cost.

Methods can be standardized but they commonly contain bias that is reproducible and may miss important associations. Bias can be easily reproduced and can be found in every step of the 16S rRNA gene sequencing workflow (Salter et al., 2014; Gohl et al., 2016; Luo et al., 2016; Amir et al., 2017). This study shows that specific diversity metrics used to measure the microbial community consistently vary based on polymerase. Standardizing multiple workflows to one specific polymerase could be detrimental since some polymerases may work better in certain situations over others. Arguably, the degree of workflow standardization across studies and research group needs to be approached on a study by study basis and not every project needs to use the exact same approach. All aspects of the 16S rRNA gene sequencing workflow need to be customized for the specific microbial community that will be sampled. Although a diversity of approaches may make reproducibility and replicability more difficult it will help to avoid systematic biases from occuring due to widespread standardization of approaches.

Conclusion

Our observations fill a gap in knowledge on the bias introduced to 16S rRNA gene sequencing results due to differences in polymerases. We found that the number of OTUs and the chimera prevalence is dependent on both polymerase and cycle number chosen. Care should be taken when choosing a polymerase for 16S rRNA gene surveys because their intrinsic differences can change the number of OTUs observed and influence diversity based metrics that do not down-weight rare observations. Knowing the inherent bias associated with different polymerases allows for better interpretation of the relationsip between an individual study to their respective field of research.

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 A. Sze came from the Canadian Institute of Health Research and NIH grant UL1TR002240. Salary support for Patrick D. Schloss came from NIH grants P30DK034933 and 1R01CA215574.

References

Acinas SG., Sarma-Rupavtarm R., Klepac-Ceraj V., Polz MF. 2005. PCR-induced sequence artifacts and bias: Insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology* 71:8966–8969. DOI: 10.1128/aem.71.12.8966-8969.2005.

Amir A., McDonald D., Navas-Molina JA., Debelius J., Morton JT., Hyde E., Robbins-Pianka A., Knight R. 2017. Correcting for microbial blooms in fecal samples during room-temperature shipping. *mSystems* 2:e00199–16. DOI: 10.1128/msystems.00199-16.

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.

Baxter NT., Ruffin MT., Rogers MAM., Schloss PD. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine* 8. DOI: 10.1186/s13073-016-0290-3.

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.

Bonfili L., Cecarini V., Berardi S., Scarpona S., Suchodolski JS., Nasuti C., Fiorini D., Boarelli MC., Rossi G., Eleuteri AM. 2017. Microbiota modulation counteracts alzheimer's disease progression influencing neuronal proteolysis and gut hormones plasma levels. *Scientific Reports* 7. DOI: 10.1038/s41598-017-02587-2.

Burkardt H-J. 2000. Standardization and quality control of PCR analyses. *Clinical Chemistry and Laboratory Medicine* 38. DOI: 10.1515/cclm.2000.014.

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

297 analysis. *Nucleic Acids Research* 42:D633–D642. DOI: 10.1093/nar/gkt1244.

298 Eckert KA., Kunkel TA. 1991. DNA polymerase fidelity and the polymerase chain reaction. *Genome*
299 *Research* 1:17–24. DOI: 10.1101/gr.1.1.17.

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

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

303 Gohl DM., Vangay P., Garbe J., MacLean A., Hauge A., Becker A., Gould TJ., Clayton JB., Johnson
304 TJ., Hunter R., Knights D., Beckman KB. 2016. Systematic improvement of amplicon marker gene
305 methods for increased accuracy in microbiome studies. *Nature Biotechnology* 34:942–949. DOI:
306 10.1038/nbt.3601.

307 Haas BJ., Gevers D., Earl AM., Feldgarden M., Ward DV., Giannoukos G., Ciulla D., Tabbaa D.,
308 Highlander SK., Sodergren E., Methe B., DeSantis TZ., Petrosino JF., Knight R., and BWB. 2011.
309 Chimeric 16S rRNA sequence formation and detection in sanger and 454-pyrosequenced PCR
310 amplicons. *Genome Research* 21:494–504. DOI: 10.1101/gr.112730.110.

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

313 Kebschull JM., Zador AM. 2015. Sources of PCR-induced distortions in high-throughput sequencing
314 data sets. *Nucleic Acids Research*:gkv717. DOI: 10.1093/nar/gkv717.

315 Kim D., Hofstaedter CE., Zhao C., Mattei L., Tanes C., Clarke E., Lauder A., Sherrill-Mix S.,
316 Chehoud C., Kelsen J., Conrad M., Collman RG., Baldassano R., Bushman FD., Bittinger K.
317 2017. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* 5. DOI:
318 10.1186/s40168-017-0267-5.

319 Kozich JJ., Westcott SL., Baxter NT., Highlander SK., Schloss PD. 2013. Development of a
320 dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the
321 MiSeq illumina sequencing platform. *Applied and Environmental Microbiology* 79:5112–5120. DOI:

10.1128/aem.01043-13.

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.

Meisel JS., Hannigan GD., Tyldsley AS., SanMiguel AJ., Hodkinson BP., Zheng Q., Grice EA. 2016. Skin microbiome surveys are strongly influenced by experimental design. *Journal of Investigative Dermatology* 136:947–956. DOI: 10.1016/j.jid.2016.01.016.

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

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*.

Polz MF., Cavanaugh CM. 1998. Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology* 64:3724–3730.

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

Real R., Vargas JM. 1996. The probabilistic basis of jaccards index of similarity. *Systematic Biology* 45:380–385. DOI: 10.1093/sysbio/45.3.380.

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.

Santos QMB-d los., Schroeder JL., Blakemore O., Moses J., Haffey M., Sloan W., Pinto AJ.

2016. The impact of sampling, PCR, and sequencing replication on discerning changes in drinking water bacterial community over diurnal time-scales. *Water Research* 90:216–224. DOI: 10.1016/j.watres.2015.12.010.

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.

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.

Sze MA., Dimitriu PA., Suzuki M., McDonough JE., Campbell JD., Brothers JF., Erb-Downward JR., Huffnagle GB., Hayashi S., Elliott WM., Cooper J., Sin DD., Lenburg ME., Spira A., Mohn WW., Hogg JC. 2015. Host response to the lung microbiome in chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine* 192:438–445. DOI: 10.1164/rccm.201502-0223oc.

Turnbaugh PJ., Hamady M., Yatsunenko T., Cantarel BL., Duncan A., Ley RE., Sogin ML., Jones WJ., Roe BA., Affourtit JP., Egholm M., Henrissat B., Heath AC., Knight R., Gordon JI. 2008. A core gut microbiome in obese and lean twins. *Nature* 457:480–484. DOI: 10.1038/nature07540.

Wang GCY., Wang Y. 1996. The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology* 142:1107–1114. DOI: 10.1099/13500872-142-5-1107.

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*.

Zupancic ML., Cantarel BL., Liu Z., Drabek EF., Ryan KA., Cirimotich S., Jones C., Knight R., Walters WA., Knights D., Mongodin EF., Horenstein RB., Mitchell BD., Steinle N., Snitker S., Shuldiner AR., Fraser CM. 2012. Analysis of the gut microbiota in the old order amish and its relation to the metabolic syndrome. *PLoS ONE* 7:e43052. DOI: 10.1371/journal.pone.0043052.

Figure 1: Total Number of OTUs in Fecal Samples by Number of Cycles. The x-axis represents the different number of cycles used and the y-axis is the number of OTUs. The points represent the median value of all four fecal samples. The lines represent the range of the minimum and maximum number of OTUs detected within the four fecal samples. The range in the number of OTUs detected in the different fecal samples increased as cycle number increased. This increased range also was larger for specific HiFi DNA polymerases.

Figure 2: Total Number of OTUs in Mock Samples by Number of Cycles. The x-axis represents the different number of cycles used and the y-axis is the number of OTUs. The points represent the median value of all four fecal samples. The lines represent the range of the minimum and maximum number of OTUs detected within the four fecal samples. The dotted black line represents the number of OTUs detected when only the references sequences for the mock community are clustered. The range in the number of OTUs detected in the mock samples increased as cycle number increased. This range was also larger for specific HiFi DNA polymerases.

Figure 3: Bray-Curtis Community Differences by Five-Cycle Intervals. A) within person differences based on the next 5-cycle PCR interval in fecal samples. B) Within replicate difference based on the next 5-cycle PCR interval in Mock samples.

Figure 4: HiFi DNA Polymerase Error Rate in Mock Samples. The error bars represent the 75% interquartile range of the median error rate.

Figure 5: HiFi DNA Polymerase Chimera Prevalence in Mock Samples. A) Precentage of chimeric sequences without the removal of chimeras with VSEARCH. C) Percentage of chimeric sequences with the removal of chimeras with VSEARCH. C) The total percent of chimeric sequences removed with VSEARCH by cycle number. The error bars represent the 75% interquartile range of the median.

Figure 6: The Correlation between Number of OTUs and Chimeras in Mock Samples.

Figure S1: HiFi DNA Polymerase Nucleotide Subsitutions in Mock Samples.