High Fidelity DNA Polymerase Introduces Bias into 16S rRNA Gene Sequencing Results

Marc A Sze 1 and Patrick D Schloss 1†

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

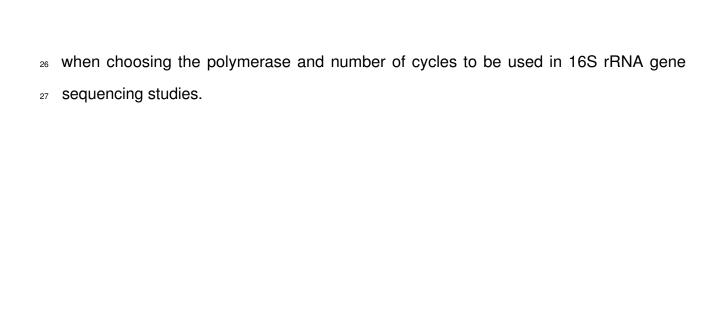
1 Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI

Co-author e-mails:

• marcsze@med.umich.edu

Abstract

- Background. Bias is introduced at different stages of the 16S rRNA gene sequencing
- workflow. Commonly studied sources of bias include preservation and DNA extraction
- 4 methods. Although the number of cycles and high fidelity (HiFi) DNA polymerase are
- studied less often, they are still important sources of bias in this workflow. Here, we
- 6 examine how both cycle number and polymerase can change the bacterial community and
- 7 introduce bias to the final obtained results.
- 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
- ¹⁰ 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 both fecal samples and mock communities.
- 14 It also assessed polymerase dependent differences in the Bray-Curtis index, error rate,
- sequence error prevalence, and chimera prevalence. Our analysis also examined chimera
- prevalence correlation with the number of OTUs.
- 17 **Results.** When analyzing fecal samples we observed a different number of OTUs between
- HiFi DNA polymerases at 35 cycles (P-value < 0.0001). Our analysis identified these
- polymerase dependent differences in the number of OTUs as early as 20 cycles in the
- mock communities (P-value = 0.002). Chimera prevalence varied by polymerase and this
- variation persisted after chimera removal using VSEARCH. We also observed positive
- 22 correlations between chimera prevalence and the number of OTUs which was not affected
- 23 by chimera removal with VSEARCH.
- ²⁴ Conclusions. HiFi DNA polymerase dependent differences in the number of OTUs and
- 25 chimera prevalence makes comparison across studies difficult. Care should be exercised



28 Introduction

Accounting for bias is critical in order to reach an understanding of bacterial community changes using 16S rRNA gene sequencing results. Differentiating between bias, reproducibility, and standardization is important since often times these three can be 31 confused and used interchangeably with each other. Bias can change the observed results 32 in a way that is reproducible and standardized. For example, if one group uses one brand of DNA extraction kit for their 16S rRNA gene sequencing, their results may be biased versus another group not using the same brand kit but within their group they can still have 35 reproducible results (Salter et al., 2014). Our 16S rRNA gene seguencing methods are biased even when these workflows are standardized to increase reproducibility. In order to 37 interpret specific studies within the broader context of the overall field, assessing bias at different parts of the 16S rRNA gene sequencing workflow is critical.

Many parts of the 16S rRNA gene sequencing workflow contribute bias to the results and are studied extensively. A typical 16S rRNA gene sequencing workflow can be divided into preservation, extraction, PCR, and sequencing steps. Generally, not using a preservation media and leaving samples at room temperature supports overgrowth of low abundance members of the fecal bacterial community (Amir et al., 2017). Similarly, this overgrowth can still occur if the preservation media does not adequately inhibit growth (Song et al., 2016; Luo et al., 2016). Reports have also shown that changes in specific community members might occur due to differing susceptibility to freeze thaw cycles amongst microbes (Gorzelak et al., 2015). Additionally, reagent contamination can add community members and the contribution of these contaminant members grows larger with lower biomass samples (Salter et al., 2014). Overall, biases due to either preservation or extraction tend to be smaller than the overall biological signal being measured (Song et al., 2016; Bassis et al., 2017). However, the contribution of PCR bias to this overall workflow is not well characterized since these studies use the same PCR approach while varying preservation

or extraction method.

Identifying the biases in the PCR stage of 16S rRNA gene sequencing is important because 55 a large body of literature shows that there are a variety of steps during PCR that can change the observed results (Eckert & Kunkel, 1991; Burkardt, 2000). Many of these 57 sources of biases are made worse as the number of cycles increases (Wang & Wang, 1996; Haas et al., 2011; Kebschull & Zador, 2015). For example, the selective amplification of AT-rich over GC-rich sequences can exaggerate the difference between 16S rRNA genes higher in AT versus those higher in GC (Polz & Cavanaugh, 1998). Both amplification error and non-specific amplification (e.g. incorrect amplicon size products) can also increase as the number of cycles increases which can drastically change commonly used diversity measures (Acinas et al., 2005; Santos et al., 2016). Additionally, chimeras can form from an aborted extension step followed by a subsequent priming error and secondary extension and will artificially increase community diversity (Haas et al., 2011). Although these differences are not necessarily dependent on primer and DNA polymerase used, there are also biases that are.

The intrinsic properties to primers and DNA polymerases chosen can also introduce bias. Primers have variable region dependent binding affinities for different bacteria and depending on the primer pair do not 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). Additionally, there 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 previously mentioned sources of PCR-based bias.

A recent study found clear differences between normal and high fidelity DNA polymerase and that optimization of the PCR protocol could reduce error and chimera generation (Gohl

et al., 2016). This study also found that regardless of DNA polymerase, the number of
Operational Taxonomic Units (OTUs) or taxa generated were not easily reduced using the
authors chosen bioinformatic pipeline (Gohl et al., 2016). It is natural to extend this line of
inquiry and ask if biases in the number of OTUs and chimeras are also dependent on the
type of high fidelity DNA polymerase. There is some reason to think that this may be the
case since many of these HiFi DNA polymerases 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).

Although bias introduced due to differences between DNA polymerase and high fidelity DNA polymerase has been investigated for 16S rRNA gene sequencing, the bias caused due to differences between specific high fidelity DNA polymerases has not been. This study will specifically address how high fidelity DNA polymerases can bias observed bacterial community results derived from 16S rRNA gene sequencing. We will accomplish this by examining if any of five different types of high fidelity 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 were previously published (Seekatz et al., 2016). Other than confirmation that none of these individuals had a Clostridium difficle infection, clinical data 100 and other types of meta data were not utilized or accessed for this study. All samples 101 were extracted using the MOBIOTM PowerMag Microbiome RNA/DNA extraction kit (now 102 Qiagen, MD, USA). The ZymoBIOMICSTM Microbial Community DNA Standard (Zymo, CA, 103 USA) was used for mock communities and was made up of *Pseudomonas aeruginosa*, 104 Escherichia coli, Salmonella enterica, Lactobacillus fermentum, Enterococcus faecalis, 105 Staphylococcus aureus, Listeria monocytogenes, and Bacillus subtilis at equal genomic 106 DNA abundance (http://www.zymoresearch.com/microbiomics/microbial-standards/ 107 zymobiomics-microbial-community-standards). 108

PCR Protocol: The five different high fidelity DNA polymerases (hereto referred to as 109 polymerases) that were tested included AccuPrimeTM (ThermoFisher, MA, USA), KAPA 110 HIFI (Roche, IN, USA), Phusion (ThermoFisher, MA, USA), Platinum (ThermoFisher, MA, 111 USA), and Q5 (New England Biolabs, MA, USA). The PCR cycle conditions for Platinum 112 and Accuprime followed a previously published protocol (Kozich et al., 2013) (https://github. 113 com/SchlossLab/MiSeq WetLab SOP/blob/master/MiSeq WetLab SOP v4.md). The 114 polymerases activation time was 2 minutes, unless a different activation was specified. For 115 Kappa and Q5, a previously published protocol was used (Gohl et al., 2016). For Phusion, 116 the company defined conditions were used except for extension time, where the Accuprime 117 and Platinum settings were used. 118

The cycle conditions for both 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 (Schloss et al., 2009; Cole et al., 2013; Rognes et al., 2016).

Analysis Workflow: All samples were rarefied to 1000 sequence. 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.

58 Results

177

The Number of OTUs are Dependent on Polymerase: A consistent difference in the range of number of OTUs, that was dependent on the polymerase used was observed 160 for fecal samples [Figure 1]. Additionally, there was a trend for lower number of cycles 161 (15-20) to result in a reduced range in the number of OTUs versus higher number of 162 cycles (25, 30, and 35) between polymerases [Figure 1]. For fecal samples, there was 163 a significant difference between polymerases at 35 cycles (F-stat > 16.35, P-value = 164 9.7e-05) [Table S1]. No differences between the specific groups was found based on a 165 Tukey post-hoc test (P-value > 0.05) [Table S2]. This polymerase dependent difference 166 in the range of number of OTUs also was observed in the mock community samples 167 [Figure 2]. Regardless if fecal or mock communities were used, the same polymerases 168 had high (Platinum) and low (Accuprime) number of OTUs and this was consistent across 169 the number of cycles used [Figure 1-2 & Table S1-S4]. In contrast to the results obtained 170 with fecal samples, differences between polymerases were observed as early as 20 cycles 171 in the mock community samples (F-stat = 15.82, P-value = 0.002) [Table S3]. For the mock 172 community, the majority of differences in the number of OTUs were between Kappa and 173 Platinum and the other polymerases across different number of cycles [Table S4]. Based 174 on these observations in fecal and mock communities, it is clear that 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 were large differences in the community composition between 20 and 25 cycles [Figure 3A]. Additionally, when data was available for the mock communities, there were 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, we found that there was 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 186 (P-value > 0.05). Using PERMANOVA to test for community differences based on the 187 number of cycles within polymerases, only Phusion had cycle dependent differences 188 (P-value = 0.03. For fecal samples, Phusion was one of two polymerases that had enough 189 sequences to be rarefied to 1000 at 15 cycles. Overall, these data suggest that there are 190 small polymerase differences in Bray-Curtis index that are dependent on the number of 191 cycles. 192

Sequence Error is Dependent on both Polymerase and Cycle Number: Differences in the median error rate varied by polymerase across the number of cycles used [Table 194 S5]. The highest error rates were for the Kappa HiFi DNA polymerase [Figure 4]. The differences in the median error rate between the different polymerases was dependent on the number of cycles with Platinum and Kappa having the most differences versus 197 other polymerases [Figure 4 and Table S6]. The total sequences with at least one error 198 was also cycle number dependent [Table S7 & S8. These differences in sequences with 199 at least one error were mostly due to differences in AccuprimeTM and Kappa versus the 200 other polymerases [Table S7 & S8]. Finally, we did not observe a polymerase dependent 201 difference on base substitution rate [Figure S1]. These results suggest that sequence 202 error is dependent on polymerase. We next wanted to assess whether chimeras also were 203 dependent on polymerase and whether this affected OTU number. 204

Prevalence of Chimeric Sequences are Polymerase Dependent and Correlate with
the Number of OTUs: There were significant differences in the chimera prevalence based
on polymerase used at all numbers of cycles used (P-value < 0.05) [Table S9]. Differences
in chimera prevalence between Platinum and all other polymerases accounted for the
majority of these differences [Table S10]. Accuprime[™] had the lowest chimera prevalence

of all polymerases regardless of whether chimera removal with VSEARCH was used [Figure 5A & 5B]. 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 AccuprimeTM, 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 [Figure 1-2 & 4-5]. 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 specific diversity differences between studies can be attributed to differences in polyermase used. Based on our observations metrics that measure within sample diversity 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. Using this metric our observations generally found no differences in the overall bacterial community composition for the number of cycles used. One possible reason for this outcome was that our study did not have enough power to detect differences due to low sample number in each group. Another reason was 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 246 polymerase family, this is unlikely to be the only contributor. For example, the highest and 247 lowest chimera rates both belonged to a family A polymerase (Platinum and AccuprimeTM 248 respectively) (Ishino & Ishino, 2014). Additionally, based on the material safety data 249 sheet (MSDS) the differences between the two polymerases are not immediately apparent. 250 Both polymerases contain a recombinant *Tag* DNA polymerase, a *Pyrococcus* spp GB-D 251 polymerase and a platinum *Taq* antibody. With everything else being equal, it is possible that differences in how the recombinant Tag 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 257 chimera prevalence that can be masked by biological differences. The sequence error 258 introduced by the polymerase is small and likely to be smaller than the biological variation 259 within a specific study, which would be consistent with previous findings for preservation 260 and DNA extraction methods (Salter et al., 2014; Song et al., 2016; Luo et al., 2016). 261 However, the chimarea prevalence for some polymerases (e.g. Platinum) are relatively large and might be greater than the oberseved biological variation within a specific study. The choice of polymerase can be as important a consideration as either preservation or DNA extraction method used because similar to using different preservation methods or 265 different DNA extraction kits, the type of polymerase can add bias to the observed bacterial 266 community. Although avoiding polymerases that yield a high number of OTUs and chimera 267 prevalence might provide better standardization of results it does come at a cost. 268

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 272 metrics used to measure the microbial community consistently vary based on polymerase. 273 Standardizing multiple workflows to one specific polymerase could be detrimental since 274 some polymerases may work better in certain situations over others. Arguably, the degree 275 of workflow standardization across studies and research group needs to be approached 276 on a study by study basis and not every project needs to use the exact same approach. All 277 aspects of the 16S rRNA gene sequencing workflow need to be customized for the specific 278 microbial community that will be sampled. Although a diversity of approaches may make 279 reproducibility and replicability more difficult it will help to avoid systematic biases from 280 occuring due to widespread standardization of approaches. 281

282 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.

91 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.
- 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.
- 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 analysis. *Nucleic Acids Research* 42:D633–D642. DOI: 10.1093/nar/gkt1244.
- Eckert KA., Kunkel TA. 1991. DNA polymerase fidelity and the polymerase chain reaction.

 Genome Research 1:17–24. DOI: 10.1101/gr.1.1.17.
- 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.
- Gorzelak MA., Gill SK., Tasnim N., Ahmadi-Vand Z., Jay M., Gibson DL. 2015. Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. *PLOS ONE* 10:e0134802. DOI: 10.1371/journal.pone.0134802.
- Haas BJ., Gevers D., Earl AM., Feldgarden M., Ward DV., Giannoukos G., Ciulla D.,
 Tabbaa D., Highlander SK., Sodergren E., Methe B., DeSantis TZ., Petrosino JF.,
 Knight R., and BWB. 2011. Chimeric 16S rRNA sequence formation and detection in
 sanger and 454-pyrosequenced PCR amplicons. *Genome Research* 21:494–504. DOI:
 10.1101/gr.112730.110.
- 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.
- Kebschull JM., Zador AM. 2015. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*:gkv717. DOI: 10.1093/nar/gkv717.
- 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*

- ³⁴⁴ 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:
- 349 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.
- 352 *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:
- ³⁵⁷ Community ecology package.
- Polz MF., Cavanaugh CM. 1998. Bias in template-to-product ratios in multitemplate PCR.
- 359 Applied and Environmental Microbiology 64:3724–3730.
- R Core Team. 2017. R: A language and environment for statistical computing. Vienna,
- 361 Austria: R Foundation for Statistical Computing.
- Real R., Vargas JM. 1996. The probabilistic basis of jaccards index of similarity. *Systematic*
- ³⁶³ *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.
- 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.
- ³⁹⁵ DOI: 10.1128/mspheredirect.00073-17.
- ³⁹⁶ Wickham H. 2017. *Tidyverse: Easily install and load 'tidyverse' packages*.

- 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
 legend represents the color code for the 5 different HiFi DNA polymerases used. The
 range in the number of OTUs detected in the different fecal samples increased as cycle
 number increased. This range was also 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 legend represents the color code for the 5 different HiFi DNA polymerases used.
 The dotted black line represents the number of OTUs detected when only the references
 sequences 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.

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