Assessing the Differences in 16S rRNA Gene Sequencing Due to High Fidelity DNA Polymerase

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. A typical 16S rRNA gene sequencing workflow can be divided into preservation, extraction, amplification, and sequencing steps. At each of these stages error can be introduced that will change the underlying bacterial community composition results. In this study, we focus on the amplification step's contribution to this overall error. To accomplish this we assessed 16S rRNA gene sequencing results in human fecal and mock community samples after using different high fidelity (HiFi) DNA polymerases and number of amplification cycles.

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. We first assessed GC dependent differences, error rate, sequence error prevalence, chimera prevalence, and correlation between chimera prevalence and number of Operational Taxonomic Units (OTUs) by polymerase and number of cycles. Next, differences in the number of OTUs was examined based on the polymerase and number of cycles used. Additionally, differences in the bacterial community composition by the Bray-Curtis index also was assessed based on polymerase and number of cycles. Based on these findings individual taxa differences based on polymerase and number of cycles was investigated. Finally, Random Forest models were created to test whether the bacterial community was better at classifying polymerases, number of cycles, or individual donor. We also assessed whether the most important taxa in the polymerase and number of cycle Random Forest models were also important in the model for individual donors.

Results. Predictably, we found noticeable differences in relative abundance based on high and low GC content (P-value $\leqslant 0.04$). 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 (R^2 = 0.974) and Kappa having the worst (R^2 = 0.478). 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). When analyzing fecal samples we observed smaller differences in the number of OTUs. Additionally, the median number of OTUs only varied by HiFi DNA polymerase used at 35 cycles and not at 20 cycles like the mock community (P-value < 0.0001). Random Forest models were most successful at classifying individual donor samples rather than polymerase or number of cycles used (P-value $\leqslant$ 5.49e-07). Additionally, the most important OTUs in the polymerase and number of cycle models were not the most important in the individual donor sample model.

Conclusions. Although there were 16S rRNA gene sequencing differences based on polymerase and number of cycles used, bacterial community composition differences are mostly only detectable in mock communities. Collectively, these results provide evidence that smaller biological differences between groups, based on 16S rRNA gene sequencing of fecal samples, can be consistently detected regardless of polymerase and number of cycles used.

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. 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 errors in the observed bacterial community composition 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 error to this overall workflow is not as well characterized.

There is a large body of literature that shows there are errors 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 error 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). For simplicity, the use of the term 'error' within the context of this study will be used as a broad term for changes in the downstream sequencing results due to the amplification step (e.g. higher GC content genomes, base substitutions, chimeras, changes from the expected relative abundance, etc.).

Interestingly, the influence 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 error. 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 of this study could reduce the difference between 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 on the same samples. The authors instead focused their efforts on testing specific types of amplification methods. Additionally, the fecal samples that were analyzed, as a model of a more complex community, may overestimate the biological variation in human samples due to the comparison of captive versus semi-captive red-shanked doucs. Due to these gaps, it makes sense to extend the previous line of inquiry to include a more detailed examination of two specific areas. First, whether differences in PCR error are detected based on number of cycles used and high fidelity polymerases. Second, whether these differences in error are enough to obscure smaller biological signals than what has been previously reported (Gohl et al., 2016).

This study will investigate whether there are error differences that are dependent on high fidelity DNA polymerases and whether these differences obscure the biological variation in the fecal bacterial communities between individuals. We accomplish this by first examining GC-based amplification error, sequence error (e.g. substitutions), chimera prevalence, the number of OTUs, Bray-Curtis index, and OTU-based differences at varying number of cycles in five different high fidelity DNA polymerases in mock communities. In addition to these metrics we also created Random Forest models of the mock community to assess how successful classification based on 16S rRNA gene sequencing of number of cycles and type of polymerase could be. We then assessed if similar

differences in the number of OTUs, Bray-Curtis index, and OTUs could be detected in human fecal samples. Additionally, we also created Random Forest models to assess how successful they could be at classifying fecal samples by number of cycles, polymerase used, or individual. Collectively, our observations suggest that although detectable differences occur based on number of cycles or high fidelity DNA polymerase used, they are smaller than the biological differences between individuals.

Results

Error differences due to number of cycles or polymerase used are detectable in mock communities. There was a significant difference in relative abundance between high/low GC content based on either the V4 16S rRNA gene region or the whole genome [Figure 1 & Table S1]. However, using only the GC content of the V4 region resulted in less differences than when using the whole genome [Table S1]. The differences between high and low GC content groups were significant as early as 20 cycles in all polymerases except Q5 [Table S1]. The highest (*Staphylococcus*) and lowest (*Pseudomonas*) GC content bacteria were the most divergent from the expected relative abundance of 0.12 [Figure 1]. These observations suggest that the high fidelity DNA polymerases tested do not prevent GC-based amplification error.

The median error rate varied by polymerase, with Kappa having the highest error rate of all the polymerases across the number of cycles [Figure 2A & Table S2]. Thus, the majority of the differences observed across the number of cycles was between Kappa and other polymerases [Figure 2A & Table S3]. The total sequences that contained at least one error was also polymerase dependent, with the majority of differences being between Kappa or Accuprime and the other polymerases [Table S2 & S3]. These differences in error rates were not due to polymerase dependent differences in base substitution rate [Figure S1]. Collectively, the results suggest that sequence error is dependent on polymerase, persists across the number of cycles used, and are not due to any bias towards a specific type of base substitution.

We next examined whether chimeras also were dependent on polymerase and whether this could affect the total number of OTUs. We observed significant differences in the chimera prevalence based on polymerase across the number of cycles used (P-value < 0.05) [Table S2]. Differences in chimera prevalence between Platinum and all other polymerases accounted for the majority of these differences [Table S3]. Accuprime^{TM} had the lowest chimera prevalence of all polymerases regardless of whether chimera removal with VSEARCH was used [Figure 2B & 2C]. The number of cycles used clearly increased chimera prevalence for all polymerases but the rate of increase differed [Figure 2B & 2C]. There was also a plateau in the total percent of chimeras that were removed that was similar for all polymerases [Figure 2D]. Additionally, a positive correlation was

observed between chimeric sequences and the number of OTUs for all polymerases [Figure 2E]. This positive correlation was strongest for Accuprime^{TM}, Platinum, and Phusion [Figure 2E]. This data suggests that chimera prevalence depends on polymerase, is made worse by increasing the number of cycles, and confirms previous reports that the number of OTUs is dependent on the prevalence of these chimeric sequences.

Bacterial community composition differences based on number of cycles or polymerase used are less obvious in mock community samples. Due to the close relationship between chimera prevalence and the number of OTUs, we also observed a polymerase dependent difference in the range of the number of OTUs in the mock community samples based on number of cycles and polymerase [Figure 3A]. The closest any of the polymerases came to a total of 8 OTUs, created by running the mock reference 16S sequences through mothur processing, was at 25 and 30 cycles [Figure 3A]. 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 S4]. Using a Tukey post-hoc test, the majority of differences for the number of OTUs detected in the mock community was largely due to Kappa and Platinum versus the other polymerases across the number of cycles used [Table S5].

We next investigated the effect polymerases and number of cycles had on the bacterial community composition in more detail by using the Bray-Curtis index. We observed clear differences in the mock community based on polymerase used (P-value = 1e-04) as well as differences based on polymerase and number of cycles (P-value = 1e-04). However, when investigating specific cycles within each polymerase we observed no difference in the bacterial community composition (P-value > 0.05) [Figure 3B]. Finding specific differences in overall bacterial community composition based on number of cycles and polymerase, we next examined if specific taxa vary depending on polymerase and number of cycles. We found that OTUs that taxonomically classified to *Salmonella*, *Escherichia*, *Enterococcus*, and *Staphylococcus* differed based on polymerase used [Table S6 & S7]. These specific OTUs were all members of the mock community and the differences were most pronounced at 25 and 30 cycles [Table S7]. The majority of the differences between taxa were due to differences between a single polymerase (Kappa) and the other polymerases tested [Table S7].

Based on these observations, we created Random Forest models to classify samples based on number of cycles or polymerase used. When assessing the Random Forest models there was a difference between models built to classify number of cycles versus those built to classify polymerase used [P-value = 5.06e-07]. Although there was a difference, neither model was successful at correctly classifying samples based on 16S rRNA gene sequencing data (cycles model, probability of correct classification = 0.42 (0.4 - 0.43)(min - max) versus polymerase model, probability of correct classification = 0.39 (0.37 - 0.4)). Our observations in mock communities suggest that although differences are detected based on number of cycles or polymerase used, the overall community-wide differences are subtle.

Few differences based on number of cycles and polymerase used were identified in fecal samples. Unlike mock community samples, an overall difference in the number of OTUs was only detected between polymerases at 35 cycles in fecal samples (F-stat > 16.35, P-value = 9.7e-05) [Table S4], and a Tukey post-hoc test failed to identify any differences between specific polymerases used (P-value > 0.05) [Table S5]. However, differences in the range of the number of OTUs detected for fecal samples was dependent on the polymerase used (e.g. Accuprime at 35 cycles = 84 - 106 versus Phusion at 35 cycles = 84 - 136) [Figure 4A]. There was also 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 4A]. Based on these observations, there are only small differences in the total number of OTUs detected in fecal samples that are based on number of cycles or polymerases used.

When using the Bray-Curtis index, we observed that polymerases and number of cycles had no affect on the bacterial community composition in fecal samples (P-value = 1). There was also no interaction between individuals, number of cycles, and polymerase used (P-value = 1). When using PERMANOVA to test for community differences based on any of the number of cycles within polymerases, only Phusion had cycle dependent differences (P-value = 0.03). Additionally, Phusion was one of two polymerases that had enough sequences to be rarefied to 1000 sequences at 15 cycles. Within each respective 5-cycle increment comparison there was no difference in Bray-Curtis index between the polymerases (P-value > 0.05) [Figure 4B]. However, there was an

overall decrease in Bray-Curtis index by cycle comparison group (e.g. the 15 cycle versus 20 cycle group compared to the 30 cycle versus 35 cycle group) (P-value < 0.01) [Figure 4B]. Using a Dunn's post-hoc test the 15 cycle versus 20 cycle and 20 cycle versus 25 cycle comparison groups had a higher Bray-Curtis index than the 30 cycle versus 35 cycle comparison group (P-value < 0.05). In addition to assessing community composition by Bray-Curtis index, we also investigated differences in OTUs based on polymerase at each number of cycle used and no significant differences could be found (P-value $\geqslant$ 0.01, corrected P-value = 1). Overall, these results suggest that in fecal samples the number of cycles and polymerase used only minimally change the bacterial community composition.

Finally, we built Random Forest models to classify samples for the number of cycles, polymerase used, or individual it was obtained from. Using 16S rRNA gene sequencing data there was a large disparity in the model's ability to correctly classify samples based on polymerase used, number of cycles, and individual (P-value $\leqslant$ 5.49e-07). There was a clear difference in the success of the model used to classify individuals (probability of correct classification = 0.87 (0.86 - 0.9)) versus the models created to classify number of cycles or polymerase (cycles model, probability of correct classification = 0.26 (0.25 - 0.27)(min - max) and polymerase model, probability of correct classification = 0.18 (0.17 - 0.2)). Additionally, the top 10 most important OTUs to the polymerase and number of cycle models were not the most important OTUs to the model that classified individuals [Figure 5]. This suggests that OTUs that change the most due to the number of cycles and polymerase in fecal samples are not the same as OTUs that are part of the biological variation between individuals [Figure 5]. Collectively, our observations show that despite a variety of errors associated with the amplification process, it is still possible to detect the bacterial community variation between individuals.

Discussion

Previous studies have reported GC dependent amplification differences (Polz & Cavanaugh, 1998) and we confirm that these differences still persist in high fidelity DNA polymerase when analyzing mock community samples [Figure 1]. Additionally, these specific polymerases have different sequence error rates and chimera prevalence [Figure 2]. Chimera prevalence in particular was strongly associated with the total number of OTUs detected in our mock community and vaired by polymerase [Figure 2E & 3A]. Additionally, the bacterial community composition varied by number of cycles and polymerase in these mock community samples. Clearly, it is possible to identify a large number of differences based on number of cycles and polymerase used when analyzing mock community samples. However, these differences may not be a major concern when analyzing more complex community samples.

Fecal samples showed a similar increase in median and range for the number of OTUs, based on number of cycles and polymerase used, as that found when analyzing mock community samples [Figure 4A]. However, there were few differences between the polymerases within each of the amplification cycles [Figure 4A]. Additionally, these differences did not translate to large bacterial community composition changes based on number of cycles or polymerase. The differences that were identifed were small and occured based on the number of cycles, independent of the polymerase used [Figure 4B]. There was also no OTU-based differences observed within the number of cycles between any of the polymerases for fecal samples. Interestingly, Random Forest models built on this OTU data had the highest classification success for individuals versus number of cycles or polymerase. In addition to the better model performance of Random Forest for individuals, the most important OTUs to this model were not the most important OTUs for the other models tested [Figure 5]. Collectively, these observations show that differences based on the number of cycles and between high fidelity DNA polymerases can be found. However, when using 16S rRNA gene sequencing these technical differences can be smaller than the biological differences between fecal bacterial communities within individuals.

The small differences observed due to error rate and chimera prevalence may be due to the actual DNA polymerase family being used within the PCR mix. DNA polymerases from distinct families

have different binding affinities and error correction capacity (Ishino & Ishino, 2014). However, based on the observations within our study this is unlikely to be the only contributor. Within our study the highest and lowest chimera prevalence both belonged to a family A polymerase (Platinum and Accuprime^{TM} respectively) (Ishino & Ishino, 2014). Additionally, based on the information supplied by the respective manufacturers, the differences between the two PCR mixtures are not immediately apparent. Both PCR mixtures contain a recombinant *Taq* DNA polymerase, a *Pyrococcus* spp GB-D polymerase and a platinum *Taq* antibody. Since it is not possible to know everything about the mixture beyond what was willingly provided by the manufacturer, it is possible that differences in how the recombinant *Taq* was generated or other compounds within the PCR mixture could be a contributing factor for the differences in error rate and chimera prevalence. Beyond the choice of the type of polymerase, there may be other ways to reduce the affect of polymerase dependent error rates and chimera prevalence on the downstream results.

Based on polymerase used, our observations generally found no difference when using the Bray-Curtis index to measure the bacterial community. Specifically, there was no difference in distance between successive 5-cycle increments within samples between the polymerases when using the Bray-Curtis index [Figure 4]. One possible reason for this outcome is that many of the OTUs generated by polymerase dependent error rates and chimera prevalence are likely not highly abundant, allowing the Bray-Curtis index to be able to successfully down-weight these OTUs (Minchin, 1987). The choice of downstream diversity metric could be an important consideration in helping to mitigate some of the observed polymerase dependent differences in error rate and chimera prevalence. Metrics that solely use presence/absence of OTUs (e.g. Jaccard (Real & Vargas, 1996), richness) may be less robust to polymerase dependent error rates and chimera prevalence. 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. One possible way to better choose polymerases may be based on the DNA polymerase family used in the PCR mixture.

Collectively, our data shows that using distinct high fidelity DNA polymerases will result in a different number of OTUs being detected. The differences in the number of OTUs are primarily due to chimera prevalence and sequence error rates that are distinct to the specific polymerase. Despite

this variation in the number of OTUs, our Random Forest models were able to correctly classify individuals regardless of the number of cycles and polymerase used. These results suggest that even though there is a large variation in the high fidelity DNA polymerases used across studies, biological variation similar in effect to fecal bacterial community composition differences between individuals will still be consistent.

Conclusion

Although care should always be taken when choosing a polymerase for 16S rRNA gene sequencing, our observations show that the differences between a variety of polymerases are dwarfed by the actual biological variation in fecal communities between individuals. Therefore, if the biological signal of interest is similar to differences in fecal bacterial communities found between individuals, then the type of high fidelity DNA polymerase used will only minimally change the results.

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^{TM} PowerMag Microbiome RNA/DNA extraction kit (now Qiagen, MD, USA). The ZymoBIOMICS^{TM} 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^{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 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 also 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 of *P. aeruginosa*, *E. coli*, *S. enterica*, *L. fermentum*, *E. faecalis*, *S. aureus*, *L. monocytogenes*, and *B. subtilis* (Schloss et al., 2009; Cole et al., 2013; Rognes et al., 2016). For simplicity, the use of the term 'error' within the context of this study moving forward encompasses changes in the downstream sequencing results due to the amplification step (e.g. higher GC content genomes, base substitutions, deletions, insertions, non-specific amplification, chimeras, etc.).

Analysis workflow. To adjust for unequal sequencing, all samples were rarefied to 1000 sequences for downstream analysis. The analysis of the mock community of each polymerase for GC-based amplification differences, 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. The total number of OTUs, taxa differences, and Bray-Curtis indices were analyzed for both the fecal and mock community samples. Finally, Random Forest models were created to assess whether classification of polymerase, number of cycles, or individual could be performed best using 16S rRNA gene sequencing data. For the model used to classify individuals, all samples were included from all number of cycles and polymerases but only the individual labels were used. Both number of cycles and polymerase models included all samples but only the number of cycles or polymerase label was used for each respective model. Additionally, overlap between the most important OTUs to the three models was assessed using mean decrease in accuracy (MDA).

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). High and low GC content was determined based on the median GC percentage of either the V4 region or

the whole genome of the bacterial species used in the mock community. Differences in the total number of OTUs were analyzed using an ANOVA with a tukey post-hoc test. For the comparison of the number of OTUS in fecal samples the data was normalized to each individual by cycle number to account for the biological variation. Bray-Curtis distance matrices were generated using mothur. 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. The Random Forest models were run using the caret package (v 6.0.78) (Jed Wing et al., 2017). A total of 100 10-fold CV runs on different 80/20 splits of the data was run to generate a range of the Logloss value. The probability of a correct call was obtained from this Logloss value by taking the negative natural logarithm. 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 on the SRA at the following accession number SRP132931.

Acknowledgements

The authors would like to thank all the study participants in ERIN whose samples were utilized. We also would 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. Additional thanks to members of the Schloss lab and Dr. Marcy Balunas for reading earlier drafts of the manuscript and providing helpful critiques. 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.

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 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.

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.

Jed Wing MKC from., Weston S., Williams A., Keefer C., Engelhardt A., Cooper T., Mayer Z., Kenkel B., R Core Team., Benesty M., Lescarbeau R., Ziem A., Scrucca L., Tang Y., Candan C., Hunt. T. 2017. *Caret: Classification and regression training*.

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.

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.

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: Relative abundance differences due to GC is consistent across polymerases. At each number of cycle the points and lines represent the median and with the minimum and maximum relative abundance respectively. Regardless of the number of cycle used a consistent difference between high and low GC bacteria was observed. The dotted line represents the actual relative abundance that each bacterium should be at within the mock community sample.

Figure 2: Error rate and chimera prevalence vary by polymerase and affect the number of observed OTUs. A) The error bars represent the 75% interquartile range of the median error rate. B) Percentage of chimeric sequences without the removal of chimeras with VSEARCH. C) Percentage of chimeric sequences with the removal of chimeras with VSEARCH. D) The total percent of chimeric sequences removed with VSEARCH by cycle number. The error bars represent the 75% interquartile range of the median. E) Chimera prevalence and the the observed number of OTUs are strongly correlated in mock community samples.

Figure 3: Subtle differences based on number of cycles and polymerase used are detected in bacterial community composition of mock samples. A) The range in the number of OTUs detected in the mock samples increased as cycle number increased. This range was larger for specific HiFi DNA polymerases. The points represent the median number of OTUs and the lines represent the range from the minimum to maximum number of OTUs detected within the four technical replicates. The dotted black line represents the number of OTUs detected when only the references sequences for the mock community are clustered. A) Within replicate difference based on the next 5-cycle PCR interval in mock samples. The lines represent the range of the minimum and maximum Bray-Curtis index value for each PCR 5-cycle increment comparison. The closer a sample is to a Bray-Curtis index of 1.00 the more dissimilar the bacterial community is of the two compared number of cycles.

Figure 4: Subtle differences based on number of cycles and polymerase used are detected in bacterial community composition of fecal samples. A) The range in the number of OTUs detected in the different fecal samples increased as cycle number increased. This range was larger for specific HiFi DNA polymerases. The points represent the median number of OTUs and the lines represent the range from the minimum to maximum number of OTUs detected within the four fecal

samples. B) Within person differences based on the next 5-cycle PCR interval in fecal samples. The points represent the median Bray-Curtis index for the samples. The lines represent the range of the minimum and maximum Bray-Curtis index value for each PCR 5-cycle increment comparison. The closer a sample is to a Bray-Curtis index of 1.00 the more dissimilar the bacterial community is of the two compared number of cycles.

Figure 5: Important OTUs for the number of cycles and polymerase models are not the most important OTUs in the model classifying individuals. Color highlights the top 10 OTUs in number of cycles and polymerase used and where they fall in the model used to classify individuals. The majority of these top 10 OTUs have 0 importance to the model used to classify individuals.

Figure S1: No preference for specific base subsitutions was observed across the polymerases in mock community samples.