What the Taq? The Influence of Different Hi-Fidelity Taq Polymerase on 16S rRNA Gene Sequencing

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

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

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

Co-author e-mails:

- marcsze@med.umich.edu

Abstract

Background.

Methods.

Results.

Conclusions.

Introduction

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) and the processing and storage of these samples have been published previously (Seekatz et al., 2016). 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 in this study and is made up of *Pseudomonas aeruginosa*, *Escherichia coli*, *Salmonella enterica*, *Lactobacillus fermentum*, *Enterococcus faecalis*, *Staphylococcus aureus*, *Listeria monocytogenes*, and *Bacillus subtilis* at equal genomic DNA abundance (http://www.zymoresearch.com/microbiomics/microbial-standards/zymobiomics-microbial-community-standards).

PCR Protocol: The five different high fidelity (HiFi) Taq DNA polymerase that were tested were 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 PCR cycle conditions for Platinum and Accuprime followed a previously published protocol (Kozich et al., 2013) (https://github.com/SchlossLab/MiSeq_WetLab_SOP/blob/master/MiSeq_WetLab_SOP_v4.md). If the HiFi Taq had a specific activation time that was different then 2 minutes that was used instead. For Kappa and Q5 the protocol previously published by Gohl and colleagues was used (Gohl et al., 2016). For Phusion the company defined conditions were used but the same extension time as that used for Accuprime and Platinum was used.

The 30 cycle default was used but the cycle conditions started at 15 and increased by 5 up to 35 cycles and was used for both fecal and mock samples. The fecal PCR consisted of all 4 samples at 15, 20, 25, 30, and 35 cycles for each Taq (total samples = 100). Although, the

mock communities also had 4 replicates used for 15, 20, 25, and 35 cycles, 10 replicates were used for 30 cycles for all Taq (total samples = 130). For all the mock community samples there was not enough PCR product at 15 cycles for adequate sequencing.

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

Statistical Analysis: All analysis was done with the R (v 3.4.2) software package (R Core Team, 2017). Data tranformation and graphing was completed using the tidyverse package (v 1.1.1) and colors selected using the viridis package (v 0.4.0) (Garnier, 2017; Wickham, 2017). 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 different people. 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).

Analysis Workflow: The total number of OTUs after sub-sampling was analyzed for both the fecal and mock community samples. From these observations we wanted to next analyze potential reasons as to why some of these differences may have occured. First, analysis of general sequence error rate, number of sequences with an error, and base subsitution were assessed in the mock community for each Taq. After assessing these

errors, the total number of chimeras was determined after sequence processing. The fecal samples were analyzed at 4 different sub-sampling levels, 1000, 5000, 10000, and 15000 while the mock community samples were analysed at 3 levels, 1000, 5000, 10000.

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 accesssion number **need to upload still**.

Results

The Number of OTUs is Dependent on HiFi Taq Used: After normalization by individual, for each cycle number, we observed that for fecal samples the number of OTUs identified was dependent upon the HiFi Taq used and this difference increased as the depth of sub-sampling increased [Figure 1]. Lower cycle numbers (15-20) resulted in less differences between Taq while cycle numbers of 25, 30, and 35 had larger clearer defined differences [Figure 1]. Only 35 cycles had HiFi Taq differences that were significantly different at all sub-sampling levels (P-value < 0.0001) [Table S1]. At sub-sampling depth of 5000 or higher 25 and 30 cycles had HiFi Taq differences (P-value < 0.05) [Table S1]. Using a Tukey post-hoc test only at 35 cycles were significant differences found to be mainly driven by Platinum being different than all other Taq across sub-sampling depth (P-value < 0.05) [Table S2].

This Taq dependent difference in the number of OTUs was also observed in the mock community samples with the same Taq polymerases being high (Platinum) and low (Accuprime) respectively [Figure 2 & Table S3]. Conversely, differences between HiFi Taq were observed as early as 20 cycles and a sub-sampling depth of 1000 sequences (P-value = 0.002) [Table S3]. Using a Tukey post-hoc test differences between Platinum and the other HiFi Taqs was the major driver of the differences seen at different cycle numbers and sub-sampling depths [Table S4]. Both fecal and mock samples consistently showed that across sub-sampling depth and cycle number the lowest number of OTUs identified was from Accuprime™ while the highest was from Platinum for both fecal and mock samples [Figure 1 & 2].

Sequence Error is Dependent on both Taq and Cycle Number Used: Differences by HiFi Taq in the median average per base error varied without a clear pattern across sub-sampling depth [Table S5]. Generally, the highest values were for the Kappa HiFI Taq

[Figure 3A]. Sub-sampling depth seems to have little effect on this rate with both 5000 and 10000 sub-sampled sequences showing similar results [Figure 3B-C]. There were small differences between the various HiFi Taq at lower cycle number but larger differences at higher cycle number with Platinum having large differences between all other HiFi Taq [Figure 3B-C and Table S6].

The total seqeunces with at least one error had multiple differences at different cycle numbers and sub-sampling depth driven by large differences in Accuprime^{TM} and Platinum versus the other HiFi Taq tested [Figure S1 & Table S7 & S8]. Although Accuprime^{TM} had the lowest per base error rate it had either the largest or second largest number of sequences with at least one error regardless of cycle number or sub-sampling depth [Figure S1]. Investigation of whether there were Taq dependent effects on base subsitution found that there was no clear bias and this was indpendent of sub-sampling depth [Figure S2-S4]. Further, the variation in subsitution error seems to reduce as the sub-sampling depth increases [Figure S2-S4].

Chimeric Sequences Corelate with OTUs and are HiFi Taq Dependent: After chimera removal using VSEARCH and removal of sequences that did not classify as bacteria we assessed the percentage of sequences that were still chimeric within our mock community. At all levels of sub-sampling and cycle number there were significant differences between the HiFi Taq used (P-value < 0.05) [Table S9]. Using a Dunn's post-hoc test the vast majority of these differences were driven by Platinum being different then all other HiFi Taq across cycle number and sub-sampling depth [Table S10]. Generally, across sub-sampling depth and cycle number Accuprime^{TM} had the lowest chimera prevalence of all the HiFi Taq [Figure 4].

For all Taqs a positive correlation was observed between chimeric sequences and number of OTUs, with this correlation being strongest for Platinum and Phusion HiFi Taq [Figure 5]. In general, the R^2 value between the number of OTUs and chimeric sequences became

stronger as sub-sampling depth increased [Figure 5]. Taken together this data suggests that a strong correlation exists between the number of OTUs and chimera sequence prevalence. Interestingly, Accuprime^{TM} not only had one of the lowest prevalence of chimeric sequences but also, consistently, had the lowest correlations between the number of OTUs and chimeric sequences across sub-sampling depth [Figure 5].

Discussion

Conclusion

Acknowledgements

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

References

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

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

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.

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.

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

Rognes T., Flouri T., Nichols B., Quince C., Mahé F. 2016. VSEARCH: A versatile open

source tool for metagenomics. *PeerJ* 4:e2584. DOI: 10.7717/peerj.2584.

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.

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*.

Table 1:

Table 2:

Figure 1: .

Figure 2: .

Figure 3: .

Figure 4: .

Figure 5: .

Figure 6: .

Figure 7: .

Figure S1: .

Figure S2: .

Figure S3: .

Figure S4: .

Figure S5: .

Figure S6: .

Figure S7: .