available at www.sciencedirect.com
ScienceDirect
journal homepage: www.elsevier.com/locate/funeco

Short Communication

# Polymerase matters: non-proofreading enzymes inflate fungal community richness estimates by up to 15 %

*Alena K. OLIVER*[a], *Shawn P. BROWN*[a], *Mac A. CALLAHAM Jr*[b], *Ari JUMPPONEN*[a,*]

[a] *Division of Biology, 421 Ackert Hall, Kansas State University, Manhattan, KS 66506, USA*
[b] *Center for Forest Disturbance Science, USDA Forest Service, Southern Research Station, 320 Green Street, Athens, GA 30602, USA*

ARTICLE INFO

*Article history:*
Received 17 July 2014
Revision received 4 March 2015
Accepted 5 March 2015
Available online 17 April 2015
*Corresponding editor:*
Björn Lindahl

*Keywords:*
Community composition
DNA polymerase
Fungal communities
Proofreading
Richness

ABSTRACT

Rare taxa overwhelm metabarcoding data generated using next-generation sequencing (NGS). Low frequency Operational Taxonomic Units (OTUs) may be artifacts generated by PCR-amplification errors resulting from polymerase mispairing. We analyzed two Internal Transcribed Spacer 2 (ITS2) MiSeq libraries generated with proofreading (ThermoScientific Phusion®) and non-proofreading (ThermoScientific Phire®) polymerases from the same MiSeq reaction, the same samples, using the same DNA tags, and with two different clustering methods to evaluate the effect of polymerase and clustering tool choices on the estimates of richness, diversity and community composition. Our data show that, while the overall communities are comparable, OTU richness is exaggerated by the use of the non-proofreading polymerase–up to 15 % depending on the clustering method, and on the threshold of low frequency OTU removal. The overestimation of richness also consistently led to underestimation of community evenness, a result of increase in the low frequency OTUs. Stringent thresholds of eliminating the rare reads remedy this issue; exclusion of reads that occurred ≤10 times reduced overestimated OTU numbers to <0.3 %. As a result of these findings, we strongly recommend the use of proofreading polymerases to improve the data integrity as well as the use of stringent culling thresholds for rare sequences to minimize overestimation of community richness.

The adoption of next-generation sequencing (NGS) tools has enabled deep interrogation of hyper-diverse fungal communities (Hibbett et al., 2009). NGS data can be overwhelmed by rare Operational Taxonomic Units (OTUs) that may represent a 'rare biosphere' (Sogin et al., 2006), cryptic taxa, or simply PCR and sequencing artifacts (Tedersoo et al., 2010; Brown et al., 2015). While some rare OTUs may represent true biological variability, the artifact OTUs may lead to a

[*] *Corresponding author.* Division of Biology, 433 Ackert Hall, Kansas State University, Manhattan, KS 66506, USA. Tel.: +1 785 532 6751; fax: +1 785 532 6653.
E-mail address: ari@ksu.edu (A. Jumpponen).

http://dx.doi.org/10.1016/j.funeco.2015.03.003

substantial inflation of richness estimators from NGS data (Huse et al., 2010; Kunin et al., 2010; Quince et al., 2011).

Most metabarcoding data are generated through polymerase chain reaction (PCR) carried out by DNA polymerases that vary in their fidelity. We queried recent publications and observed that less than a third of studies that unambiguously identify the polymerase for amplicon library generation used a proofreading, high-fidelity polymerase (Table S1). As a result, we aimed to quantify the effects of polymerase choice. To do this, proofreading and non-proofreading thermostable hot start polymerases from the same manufacturer were compared, and whether the proofreading enzyme would minimize potentially erroneous sequences resulting from PCR errors in complex environmental templates was examined. 24 experimental units from a long-term experiment, that was designed to evaluate the effects of prescribed fires on ecosystem properties (see Brown et al., 2013; Oliver et al., 2015), were used. Each sample was amplified in triplicate with each of the two polymerases in a two-step PCR reaction (Berry et al., 2011) to generate comparable NGS data. The primary PCR reaction used primers ITS1F (Gardes and Bruns, 1993) and ITS4 (White et al., 1990) (25 cycles) and subsequent secondary PCR reactions (5 cycles) used a nested primer (fITS7) (Ihrmark et al., 2012) and a sample specific 12-bp DNA tag in the reverse primer (ITS4). Each sample was amplified and sequenced with the same DNA-tags; this allowed evaluation of polymerase performance side by side in identical reactions, and testing the difference in the generation of potential PCR artifacts by the two enzymes. Although PCR artifacts generated by the proofreading enzyme cannot be accounted for, we argue that the relative influence of the non-proofreading enzyme can be evaluated by focusing on the differences between polymerases.

Two hot start polymerases that share optimal extension temperatures and are compatible with the green loading dyes incorporated in the PCR buffers from one manufacturer were used: a proofreading Phusion® Green Hot Start II High-Fidelity DNA polymerase and non-proofreading Phire® Green Hot Start II DNA polymerase (Thermo Scientific®, Pittsburgh, PA, USA). The reaction conditions for the 25 μl primary PCR reactions included 25 ng DNA template (5 μl), 200 μM dNTPs, 1 μM of both primers, 5 μl 5× Phusion Green HF Buffer or 5× Phire Green Buffer, 1.5 m M $MgCl_2$, 7.3 μl molecular biology grade water, and 0.5 units polymerase. PCR cycle parameters included an initial denaturing at 98 °C for 30 s, followed by 25 cycles of denaturing at 94 °C for 30 s, annealing at 54 °C for 1 min, extension at 72 °C for 2 min, and a final extension at 72 °C for 8 min. The secondary PCRs were identical except that they included 5 μl primary PCR product as template, nested fITS7 forward primer, tagged reverse primers (ITS4; Table S2), and only five cycles. Three technical replicates per experimental unit were combined after secondary PCRs, and the experimental units pooled into two amplicon libraries (24 experimental units/library; one generated with Phire®, another with Phusion® polymerase) at equal amounts of DNA. Illumina specific adapters and indices were ligated into amplicons using a NEBNext® DNA MasterMix for Illumina (Protocol E6040, New England Biolabs Inc., Ipswich, MA, USA) and sequenced using a MiSeq Reagent Kit v2 (Illumina, San Diego, CA, USA) with 500 cycles at the Integrated Genomics Facility at Kansas State University Manhattan, KS. Paired fastq files for Phusion® (SRR1508275) and Phire® (SRR1508273) libraries are available in the Sequence Read Archive at NCBI (www.ncbi.nlm.nih.gov).

We analyzed the sequence data with the MOTHUR pipeline (v. 1.32.2; Schloss et al., 2009) following suggestions from Schloss et al. (2011) and Kozich et al. (2013). The paired sequences contained in reverse and forward fastq files were aligned into a contig. After contiging the paired-end reads, the Phire® library contained 6 292 965 sequences and the Phusion® library 5 425 946 sequences. The libraries were screened to remove contigs with less than 100 bp overlap, ambiguous bases, any mismatches with primer or DNA-tag sequences (Table S2), sequences shorter than 250 bp, or homopolymers ≥8 bp. Since we did not include the Illumina adapters into our primers, we had no control over the orientation of the ligated amplicons and accounted for this by considering the reverse and forward reads in both orientations. This resulted in datasets with 1 182 870 (Phire®) and 1 113 584 sequences (Phusion®). Remaining sequences were truncated to 250 bp, the two libraries merged, and analyzed together with a total of 48 experimental units – or 24 per library–from this point on. Near identical sequences (>99 % similar) were preclustered to minimize sequencing induced errors (Huse et al., 2010). Unique sequences were screened for chimeras (UCHIME, Edgar et al., 2011) using the abundant sequences as a reference and default parameters (abundance skew = 1.9; minimum divergence ratio = 0.5). The proportion of potential chimeras was recorded for each of the samples, the chimeric sequences were removed, and the experimental units rarefied to 15 000 sequences per experimental unit from each of the Phire® and Phusion® libraries for a total of 720 000 sequences. We calculated a pairwise distance matrix for unique sequences and clustered OTUs at 97 % sequence similarity using the furthest and nearest neighbor algorithms. Furthest neighbor (complete-linkage clustering) assigns all sequences that are at most 3 % distant from all other sequences into an OTU; nearest neighbor (single-linkage clustering) assigns sequences that are at most 3 % distant from the most similar sequence into an OTU. As a result, for the same similarity threshold, the furthest neighbor algorithm yields a greater number of OTUs than the nearest neighbor algorithm. From the subsampled data for each experimental unit in each library and for both clustering methods, we enumerated sequences assigned to OTUs that were represented by 1 sequence, ≤2 sequences, ≤5 sequences, and ≤10 sequences to estimate the numbers of low frequency OTUs that may represent artifacts (Tedersoo et al., 2010; Brown et al., 2015), and to estimate coverage (Good's coverage), richness and diversity (Richness – $S_{Obs}$, complement of Simpson's diversity – 1–D, Evenness – Simpson's $E_D$, extrapolative richness – Chao1; Table S3).

The significance of differences in numbers of putatively chimeric sequences, rare OTUs, richness and diversity estimators generated with Phire® or Phusion® polymerases were tested using both paired t-tests and non-parametric Wilcoxon signed-rank tests in JMP® (version 7.0.2). The conclusions based on these analyses were always congruent and only the more conservative non-parametric tests are presented. To visualize and infer compositional differences in the fungal communities generated from the two polymerases, a

Bray–Curtis distance matrix was derived in MOTHUR for data matrices generated based on both furthest and nearest neighbor clustering. Differences between the two polymerases were tested using Analysis of MOlecular VAriance (AMOVA; PERMANOVA in Anderson, 2001) and visualized using Non-metric Multi-Dimensional Scaling (NMDS) after estimating axis loading scores for the first four ordination axes to obtain a minimum stress $\leq$0.20 (Fig S1A and S1B).

The proportion of potential chimeras did not differ between the ITS2 libraries. The two clustering methods performed as expected: richness estimators and the total number of OTUs were greater with furthest neighbor than with nearest neighbor clustering (Fig 1; Table S3). Although the proportion of low frequency sequences was low ($\leq$2.57 %), the library generated with the non-proofreading Phire® enzyme consistently yielded a greater proportion than that generated with the proofreading Phusion®. This resulted in a consistent inflation of OTU numbers: the Phire®-generated data had greater number of OTUs (10.1 % using furthest neighbor clustering; 15.5 % using nearest neighbor clustering) when singletons were included compared to the Phusion®-generated data (Fig 1). While the inflation was consistent, removal of rare sequences controlled for this: excluding reads that occurred $\leq$2 times resulted in 3.1 % (furthest neighbor clustering) or 5.9 % (nearest neighbor clustering) greater total number of OTUs and excluding reads that occurred $\leq$10 times resulted in <0.3 % greater total number of OTUs, with the non-proofreading enzyme compared to the proofreading enzyme regardless of the clustering method (Fig 1). Overall, the increase in the proportion of rare OTUs in the Phire® dataset resulted in a small (on average 5.0 % using furthest neighbor and 3.5 % using nearest neighbor clustering) but significant inflation of richness and subsequent deflation (on average 6.0 % using furthest neighbor and 8.1 % using nearest neighbor clustering) of evenness (Table S3), after omission of OTUs that were represented by $\leq$10 sequences. However, neither the extrapolative richness (Chao1) nor the diversity (1–D) estimators were strongly affected by the choice of polymerase (Table S3). Similarly, community composition did not differ between the datasets generated with the two polymerases (furthest neighbor clustering AMOVA: $F_{1,\,47} = 0.445$; $P = 0.9890$; nearest neighbor clustering AMOVA: $F_{1,\,47} = 0.487$; $P = 0.9870$): the paired Phire® and Phusion® samples were tightly coupled in our NMDS community visualization regardless of furthest (Fig S1A) or nearest (Fig S1B) neighbor clustering.

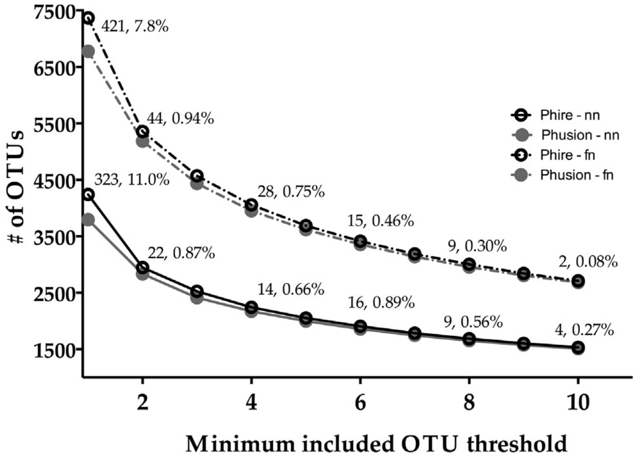

**Fig 1 – Observed total number of OTUs in MiSeq ITS2 amplicon libraries generated from 24 experimental units with the non-proofreading Phire® (open black circles) or proofreading Phusion® (closed gray circles) polymerases and with nearest neighbor (nn – solid lines) furthest neighbor (fn – dashed lines) clustering algorithms. Furthest neighbor OTUs include sequences that are at most 3 % distant from all other sequences; nearest neighbor OTUs include those that are at most 3 % distant from the most similar sequence resulting in OTU counts that are commonly lower than those acquired using furthest neighbor algorithms. Inserted numbers indicate the difference in counts of OTUs represented by rare sequences (singletons, doubletons, etc.) and their proportion relative to the Phusion-generated amplicon library. Total number of rare OTUs with ≤10 sequences was 3 571 (Phire) and 2 982 (Phusion) using furthest neighbor clustering and 2 170 (Phire) and 1 714 (Phusion) using nearest neighbor clustering, suggesting that non-proofreading polymerase generated 589 and 456 additional OTUs prior to the removal of rare OTUs, respectively – 15.5 % and 10.1 % inflation in OTU counts if rare OTUs were included. Note that inserted numbers represent proportion of OTUs in each threshold class, whereas Supplementary Table S3 represents cumulative proportions of sequences.**

Our comparison of proofreading and non-proofreading polymerases strongly indicates that richness, but not composition, is sensitive to polymerase choices in MiSeq amplicon libraries regardless of clustering method. The potential >10 % inflation of OTUs (Fig 1) is alarming and highlights the importance of diligent removal of rare OTUs from NGS data (Tedersoo et al., 2010; Brown et al., 2015). Although the proportion of chimeric reads seemed insensitive to the polymerase choice, controlling for incompatible terminal ends of the sequenced amplicons may provide a powerful tool to screen for potentially erroneous low frequency reads (Carlsen et al., 2012). Further, as deep sequencing of an excess of $10^4$ reads per experimental unit affords for even more stringent and cautious quality control, we suggest that reads occurring in counts $\leq$10 should be removed when possible. While it is impossible to determine the proportion of artifact sequences in the dataset generated with the proofreading enzyme, the combination of higher rare OTU omission thresholds and the selection of polymerases is likely to increase the quality and reliability of the NGS data. Although proofreading enzymes tend to be more costly, more aggressive multiplexing and decreasing costs of sequencing likely counter these costs.

## Acknowledgments

This work was supported by USDA Forest Service Cooperative Agreement 11-CA-11330136-126 to A.J. We would like to thank Evelyn Wenk for coordinating the sampling. Alina Akhunova and Huang Liang at the Integrated Genomics Facility at Kansas State University performed the sequencing. The use of trade names in this contribution does not constitute official endorsement of these products by USDA or any of its subsidiary agencies. Alena Oliver and Ari Jumpponen were partly supported by DOE Contract No. DE–SC0004953. Shawn Brown was supported with a National Science Foundation GK-12 fellowship (NSF DGE-0841414, PI Ferguson). This is contribution no. 15-340-J from the Kansas Agricultural Experiment Station.

## Supplementary data

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.funeco.2015.03.003.

## REFERENCES

Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26, 32–46.

Berry, D., Mahfoudh, K.B., Wagner, M., Loy, A., 2011. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and Environmental Microbiology* 77, 7846–7849.

Brown, S.P., Callaham, M.A., Oliver, A.K., Jumpponen, A., 2013. Deep Ion Torrent sequencing identifies soil fungal community shifts after frequent prescribed fires in a southeastern US forest ecosystem. *FEMS Microbiology Ecology* 86, 557–566.

Brown, S.P., Veach, A.M., Rigdon-Huss, A.R., Grond, K., Lickteig, S.K., Lothamer, K., Oliver, A.K., Jumpponen, A., 2015. Scraping the bottom of the barrel: are rare high throughput sequences artifacts? *Fungal Ecology* 13, 221–225 http://dx.doi.org/10.1016/j.funeco.2014.08.006.

Carlsen, T., Aas, A.B., Lindner, D., Vrålstad, T., Schumacher, T., Kauserud, H., 2012. Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies. *Fungal Ecology* 5, 747–749.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R., 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200.

Gardes, M., Bruns, T.D., 1993. ITS primers with enhanced specificity for basidiomycetes – application to the identification of mycorrhizae and rusts. *Molecular Ecology* 2, 113–118.

Hibbett, D.S., Ohman, A., Kirk, P.M., 2009. Fungal ecology catches fire. *New Phytologist* 184, 279–282.

Huse, S.M., Welch, D.M., Morrison, H.G., Sogin, M.L., 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology* 12, 1889–1898.

Ihrmark, K., Bödeker, I.T.M., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., Stenlid, J., Brandström-Durling, M., Clemmensen, K.E., Lindahl, B.D., 2012. New primers to amplify the fungal ITS2 region – evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology* 82, 666–677.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., Schloss, P.D., 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.

Kunin, V., Engelbrektson, A., Ochman, H., Hugenholtz, P., 2010. Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. *Environmental Microbiology* 12, 118–123.

Oliver, A.K., Callaham, M.A., Jumpponen, A., 2015. Soil fungal communities respond compositionally to recurring frequent prescribed burning in a managed southeastern US forest ecosystem. *Forest Ecology and Management* 345, 1–9. http://dx.doi.org/10.1016/j.foreco.2015.02.020.

Quince, C., Lanzen, A., Davenport, R.J., Turnbaugh, P.J., 2011. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12, 38.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F., 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75, 7537–7541.

Schloss, P.D., Gevers, D., Westcott, S.L., 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* 6, e27310. http://dx.doi.org/10.1371/journal.pone.0027310.

Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, A.M., Neal, P.R., Arrieta, J.M., Herndl, G.J., 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Science* 103, 12115–12120.

Tedersoo, L., Nilsson, R.H., Abarenkov, K., Jairus, T., Sadam, A., Saar, I., Bahram, M., Bechem, E., Chuyong, G., Kõljalg, U., 2010. 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist* 188, 291–301.

White, T.J., Bruns, T., Lee, S., Taylor, J., 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis, M.A.I., Gelfand, D.H., Sninsky, J.J., White, W.J. (Eds.), PCR Protocols: A Guide to Methods and Applications. Academic Press, San Diego, CA, pp. 315–322.