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
1754-5048/

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