available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/funeco


Short Communication

# Scraping the bottom of the barrel: are rare high throughput sequences artifacts?

*Shawn P. BROWN*[a,*], *Allison M. VEACH*[a], *Anne R. RIGDON-HUSS*[b], *Kirsten GROND*[a], *Spencer K. LICKTEIG*[a], *Kale LOTHAMER*[a], *Alena K. OLIVER*[a], *Ari JUMPPONEN*[a]

[a]*Division of Biology, Kansas State University, 116 Ackert Hall, Manhattan, KS 66506, USA*
[b]*Department of Grain Science and Industry, Kansas State University, 201 Shellenburger Hall, Manhattan, KS 66506, USA*

ARTICLE INFO

*Article history:*
Received 25 March 2014
Revision received 26 June 2014
Accepted 19 July 2014
Available online 5 October 2014
*Corresponding editor:*
Marie Louise Davey

*Keywords:*
Fungi
High-throughput sequencing
Rare biosphere
Singleton

ABSTRACT

Metabarcoding data generated using next-generation sequencing (NGS) technologies are overwhelmed with rare taxa and skewed in Operational Taxonomic Unit (OTU) frequencies comprised of few dominant taxa. Low frequency OTUs comprise a rare biosphere of singleton and doubleton OTUs, which may include many artifacts. We present an in-depth analysis of global singletons across sixteen NGS libraries representing different ribosomal RNA gene regions, NGS technologies and chemistries. Our data indicate that many singletons (average of 38 % across gene regions) are likely artifacts or potential artifacts, but a large fraction can be assigned to lower taxonomic levels with very high bootstrap support (~32 % of sequences to genus with ≥90 % bootstrap cutoff). Further, many singletons clustered into rare OTUs from other datasets highlighting their overlap across datasets or the poor performance of clustering algorithms. These data emphasize a need for caution when discarding rare sequence data *en masse*: such practices may result in throwing the baby out with the bathwater, and underestimating the biodiversity. Yet, the rare sequences are unlikely to greatly affect ecological metrics. As a result, it may be prudent to err on the side of caution and omit rare OTUs prior to downstream analyses.



Next generation sequencing (NGS) permits deep interrogation of hyper-diverse fungal communities (Hibbett et al., 2009). Data generation has become expedient and sequence analysis/annotation more streamlined via available pipelines (e.g. MOTHUR, QIIME). Concurrently sequencing costs have declined, resulting in the democratization of sequencing in ecology (Caporaso et al., 2012). Many new investigators utilize NGS but are often uncertain how to handle rare operational taxonomic units (OTUs). These rarities are common - singletons alone often comprise half of all OTUs.

Rare OTUs may represent the 'rare biosphere' (Sogin et al., 2006) but their validity has been questioned; PCR/sequencing

* *Corresponding author*. Division of Biology, 116 Ackert Hall, Kansas State University, Manhattan, KS 66506, USA. Tel.: +1 785 532 3934; fax: +1 785 532 6653.
E-mail address: spbrown1@ksu.edu (S.P. Brown).
http://dx.doi.org/10.1016/j.funeco.2014.08.006
1754-5048/

artifacts may lead to inflation of the 'rare biosphere' (Huse et al., 2010; Kunin et al., 2010; Quince et al., 2011). However, Zhan et al. (2013) sequenced aquatic communities and spiked the samples with known indicators to test sensitivity. They found that many singletons represented the spiked controls suggesting that not all singletons are artifacts.

To estimate the proportion of artifactual singletons and to test the origin of these singletons (NGS platform or PCR errors), we reanalyzed singletons from sixteen experiments that targeted three nuclear ribosomal RNA gene regions (LSU, ITS1, ITS2) from different sequencing technologies or chemistries (454-FLX, 454-Titanium, and Illumina-MiSeq; Table S1). These datasets included five ITS1 [454-FLX(3) and 454-Titanium(2)], six ITS2 (Illumina-MiSeq), and five Large Subunit variable region D1 (454-Titanium) libraries (see Table S1 for primers and direction of sequencing). The datasets were analyzed using MOTHUR (v.1.32.1; Schloss et al., 2009), denoised (Quince et al., 2011), plus chimera- (UCHIME; Edgar et al., 2011) and sequencing-error screened (pre.cluster; Huse et al., 2010) prior to OTU binning at 97 % similarity. After this quality control, ~50 % of the OTUs were singletons, which we extracted into four fasta files (Supplemental material) containing all comparable singleton sequences (ITS1-FLX, ITS1-Titanium, ITS2 and LSU). LSU libraries were aligned against a modified James et al. (2006) reference (Brown et al., 2014), and gaps removed prior to downstream analyses. Sequences were truncated to equal lengths and subsampled to equal numbers per library (Table S1). Four MiSeq libraries were generated on split-reactions (EcM and Soil Fungi – Australia and EcM of Yellow Pine using two different polymerases) allowing differentiation among sequencing platform-generated artifacts from others.

Each singleton dataset was pairwise-aligned and resultant distance matrices clustered into OTUs at 97 % similarity (using the MOTHUR implemented Average-Neighbor clustering algorithm – UPGMA) to detect overlapping rare OTUs across libraries. It is important to note that the method of OTU binning can dramatically affect the generation of singletons: single-linkage clustering (nearest-neighbor in MOTHUR) produces fewer OTUs with higher average sequence dissimilarity within an OTU, whereas a complete-linkage clustering (furthest-neighbor in MOTHUR) produces more OTUs with higher sequence similarity within an OTU. Average-neighbor clustering (UPGMA) is a "middle ground" algorithm both in terms of OTU numbers and sequence similarity. After clustering, conserved regions (SSU, 5.8S, LSU) were removed from representative sequences for each ITS OTU (including singletons) using the online UNITE Phylogenetic Module ITSx using default online options with the exception that we set the minimal number of domains required to match for extraction to one (unite.ut.ee; Nilsson et al., 2010; Bengtsson-Palme et al., 2013). The extracted OTU sequences were assigned to taxa in MOTHUR using the Naïve Bayesian Classifier (Wang et al., 2007) with the RDP 28s rRNA reference (v.7) or with two ITS databases, Findley (ITS1; Findley et al., 2013) and UNITE plus INSD non-redundant ITS database (ITS1 and ITS2; Kõljalg et al., 2013). The Naïve Bayesian Classifier queries all non-overlapping 8-bp words (k-mers) against a reference dataset and provides bootstrap support estimates to taxonomic levels based on the number of times a queried sequence is placed in the same rank. OTUs were considered artifacts if: (1) OTUs were unclassified at a phylum level (many uncultured sequences may lack phylum level classification thus exaggerating proportion of artifact OTUs); (2) they did not classify

**Table 1 – Percentage of singletons that are artifacts and potential artifacts as well as the percentage of non-artifactual OTUs that are assigned to taxa above 50 %, 75 % and 90 % bootstrap support on all levels of taxonomic levels**

| | LSU-Titanium | ITS1-FLX | ITS1-Titanium | ITS2-MiSeq |
|---|---|---|---|---|
| Percentage of artifacts | 16.87 % | 12.94 % | 13.34 % | 19.10 % |
| Percentage of potential artifacts | 37.93 % | 21.67 % | 13.29 % | 17.20 % |
| Percentage of sequences above bootstrap support thresholds | | | | |
| Phylum (90 %) | 67.80 % | 71.67 % | 74.00 % | 64.27 % |
| Phylum (75 %) | 69.80 % | 80.17 % | 79.86 % | 69.50 % |
| Phylum (50 %) | 73.60 % | 86.33 % | 86.29 % | 79.07 % |
| Class (90 %) | 48.27 % | 62.67 % | 63.71 % | 58.60 % |
| Class (75 %) | 55.60 % | 70.00 % | 71.57 % | 63.77 % |
| Class (50 %) | 63.40 % | 76.83 % | 79.29 % | 70.23 % |
| Order (90 %) | 32.73 % | 52.82 % | 58.86 % | 53.80 % |
| Order (75 %) | 44.07 % | 61.33 % | 67.00 % | 60.33 % |
| Order (50 %) | 56.53 % | 68.17 % | 77.00 % | 66.23 % |
| Family (90 %) | 20.00 % | 48.00 % | 51.71 % | 47.40 % |
| Family (75 %) | 32.40 % | 56.17 % | 60.71 % | 56.07 % |
| Family (50 %) | 47.13 % | 65.50 % | 73.43 % | 64.13 % |
| Genus (90 %) | 10.53 % | 39.17 % | 44.14 % | 37.30 % |
| Genus (75 %) | 18.07 % | 51.17 % | 55.57 % | 48.97 % |
| Genus (50 %) | 36.80 % | 61.33 % | 70.43 % | 61.33 % |