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

Shawn P. Brown, Allison M. Veach, Anne R. Rigdon-Huss, Kirsten Grond, Spencer K. Lickteig, Kale Lothamer, Alena K. Oliver, Ari Jumpponen

Published Version Information

Citation: Brown, S. P., Veach, A. M., Rigdon-Huss, A. R., Grond, K., Lickteig, S. K., Lothamer, K., et al. (2015). Scraping the bottom of the barrel: Are rare high throughput sequences artifacts? Fungal Ecology, 13, 221-225.

Digital Object Identifier (DOI): 10.1016/j.funeco.2014.08.006

Publisher's Link: http://www.sciencedirect.com/science/article/pii/S1754504814001068

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, Manhattan KS 66506
b Department of Grain Science and Industry, Kansas State University, Manhattan, KS 66506
* Corresponding author

Running Head: Are high-throughput singletons artifacts?

* Corresponding Author

Shawn P. Brown
Division of Biology
116 Ackert Hall
Kansas State University
Manhattan, KS 66506
Phone: 785-532-3934
Fax: 785-532-6653

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.

Keywords: fungi, singleton, high-throughput sequencing, rare biosphere

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 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 to a phylum at 50% bootstrap support or higher; or, 3) the ITS sequences could not be mapped to ITS1 or ITS2 region (ITSx). Furthermore, sequences from the ITS1 libraries were considered artifacts if these conditions were met for taxonomy labels from both reference databases. Additionally, singletons were considered *potential artifacts* if they received < 50% bootstrap support at the family level. We report statistics on the proportion of singletons classified to all taxonomic levels at > 50%, 75%, and 90% bootstrap support (Table 1).

Many singletons from the sixteen libraries clustered at 97% with at least one other sequence at rates seemingly driven by gene region [LSU(Titanium) – 11.5%; ITS1 (FLX) – 0.83%; ITS1 (Titanium) – 0.43%; ITS2 (MiSeq) – 2.27%] reflecting variability of clustering efficiencies across gene regions. Singletons that clustered together often originated from within the same original library suggesting that they are a result of algorithm performance that provides non-exact clustering solutions. The more conserved LSU likely performs better with these algorithms.

We queried our sequences against databases to estimate assignment robustness through bootstrapping. Overall, the proportion of artifact sequences (<50% support for phylum level classification) was much lower (12.94% - 19.10%; Table 1) than expected based on previous estimates suggesting that ~80% of singletons may be artifacts (Tedersoo *et al*. 2010). This is unexpected: our liberal inclusion of unclassified phyla as artifacts likely inflated the number of artifact singletons. The combined proportion of artifacts and *potential artifacts* was largely affected by region: LSU (54.80%) had a greater proportion of questionable sequences than ITS regions (Table 1). Interestingly, many sequences that were not considered artifacts or *potential artifacts* were assigned to lower taxonomic levels with high bootstrap support. The proportion of sequences with a genus-level affinity with ≥90% bootstrap support ranged from 10.53%-44.14%, a level of support unlikely for true artifacts.

Our analyses, similarly to Tedersoo *et al.* (2010), indicate that many singletons are likely artifacts. However, our estimates are less than half of the ~80% estimate of Tedersoo and coworkers. There are many underlying reasons for this discrepancy. The early 454-datasets explored how to analyze NGS data (*e.g.,* Buee *et al.* 2009; Jumpponen & Jones 2009; Tedersoo *et al.* 2010). Lessons from those analyses have led to recommendations on tools to utilize NGS data in fungal ecology (Nilsson *et al*. 2011; Lindahl *et al,* 2013), including adoption of denoising (Quince *et al.* 2011), standard chimera removal (Edgar *et al*. 2011) and preclustering (Huse *et al.* 2010). Noteworthy is that Tedersoo *et al.* included a BLAST-based post-hoc chimera check. However, this method is less sensitive as it relies on database accession quality, whereas pre-OTU binning methods (UCHIME; Edgar *et al.* 2011) rely on NGS-acquired data itself. Additionally, our study differs in other important ways; we neither had anchor taxa from the same samples nor performed the detailed phylogenetic analyses. Instead, we relied on the Naïve Bayesian Classifier, an approach that parallels the phylogenetic approach. Nonetheless, our results highlight the importance of appropriate quality controls to minimize artifacts.

Many ‘global singleton’ sequences clustered into new non-singleton OTUs. Whilst the underlying reasons remain unclear, we suggest two primary explanations. First, fungal communities are hyper-diverse (Jumpponen & Jones 2009), include large numbers of low frequency taxa, and are locally or regionally distinct (Meiser *et al.* 2014). Second, clustering relies on imperfect heuristic algorithms that permit non-exact solutions for OTU membership, especially in large and complex datasets. This allows stochastic OTU memberships and sequences may be placed into different OTUs each time a dataset is clustered.

Our results suggest that half of the singletons may represent true target taxa. However, we cannot determine if artifact singletons result from sequencing platform errors. Singletons may also represent off-target amplification as evidenced by the common occurrence of sequences, from which ITS regions could not be extracted with ITSx. A surprisingly high proportion of queried sequences had no extractable ITS regions (5.33% for ITS1-FLX; 2.86% for ITS1-Titanium; 4.53% for ITS2-MiSeq; Table S2). Similar proportions of non-target LSU sequences are likely but tools to evaluate this were not explored here. Interestingly, absence of extractable ITS regions were not solely due to non-target amplification: many discarded sequences were fungal, although no ITS regions could be excised using ITSx. More than 90% of our ITS1-FLX and all of our ITS1-Titanium sequences that failed to extract were fungal based on BLASTn analyses (see Table S3 for complete list of sequences that failed to be extracted using ITSx and the best BLASTn taxonomic strings). ITS2 had the highest non-target amplification: 61.03% of the sequences that failed to be extracted were not fungal (Table S3). Additional sequences failed to extract that were actually ITS2 fungal sequences. Peculiarly, all but two of the fungal sequences discarded because of failed ITS2 extraction belonged to Agaricomycetes (primarily Russulaceae and Thelephoraceae) suggesting that the Hidden Markov Models (HMM) in ITSx may fail to recognize this class fully. Alternatively, this could be explained by insufficient 5’ LSU length upstream of the priming site causing the HMMs to fail for some Agaricomycetes. The remaining artifacts are likely PCR errors - polymerase mis-pairs, deletions, or insertions (Eckert & Kunkel 1991) and chimeras that evaded detection.

To investigate if these singletons represent true biological or artificial variability (platform specific variability, indels due to polymerase slippage, or homopolymeric reads), we aligned singletons against representative sequences of the 100 most abundant OTUs from the original datasets. The mismatches among singletons and the representative sequences of the common OTUs generated on 454 and Illumina platforms appeared stochastically distributed across the alignments suggesting that they were unlikely a result of poor read quality in the read termini. Singletons generated using 454 technologies differed from abundant OTUs frequently because of inconsistent homopolymer lengths and/or single nucleotide differences. In contrast to 454-sequencing, differences in the Illumina-generated singletons were most often nucleotide differences with no evidence of inconsistent homopolymer lengths. Based on these findings it is impossible to determine the source of the variability as polymerase slippage, suboptimal platform performance or true biological variability could result in similar outcomes.

Removal of rare sequences may underestimate observed and extrapolated richness (Unterseher *et al.* 2011). Rare taxa also affect community pairwise distances commonly visualized by ordination tools. Conversely, singleton exclusion may minimally affect community composition (Shade *et al.*

2013) or multivariate analyses (Gobet *et al.* 2010; Lindahl *et al.* 2013). Although removal of singletons may not substantially affect the visualization of community composition, rare sequences may be necessary for more accurate biodiversity estimates, if they represent real taxa but biodiversity estimates from sequence data are capricious (Haegeman *et al.* 2013).

We conclude that for most hypothesis-driven experiments that compare experimental conditions, rare taxa present a minor issue: excluding them unlikely sways strong community responses. However, if estimation of biodiversity is crucial, careful manual examination and annotation of the infrequent sequences is required. One must strike a balance: is it better to err on the side of caution and throw the baby out with the bathwater (exclude rare sequences) or to analyze the rare sequences and scrape the bottom of large pools of sequence data to account for every last unculturable fungus that occurs in the data if even only once? Due to the minimal effect these rare sequences have in community analyses, we concur with previous suggestions to remove all singletons and expand this recommendation to remove other highly rare (n=10) sequences in datasets as modern sequencing depth allows for such stringent practices.

Acknowledgements

The authors' wish to thank all who were involved in field sampling and data generation for these projects. SPB is supported by a NSF GK-12 Fellowship (NSF DGE-0841414), ARHR by KABB grant (GAGS92152), and AJ by DOE-BER grant (DE-SC0004953). This is Kansas Agricultural Experimental Station Publication number XXXXX.

Table 1. Percentage of singletons that are artifacts and potential artifacts as well as the percentage of non-artifactual OTUs than 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%

References

Bengtsson-Palme J., Ryberg M., Hartmann M., Branco S., Wang Z., Godhe A., De Wit P., Sánchez-García M., Edersberger I., de Sousa F., Amend A.S., Jumpponen A., Unterseher M., Kristiansson E., Abarenkov K., Bertrand Y.J.K., Sanli K., Eriksson K.M., Vik U., Veldre V., Nilsson R.H., 2013. Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution* **4**: 914-919.

Brown S.P., Rigdon-Huss A.R., Jumpponen A., 2014. Analyses of ITS and LSU gene regions provide congruent results on fungal community responses. *Fungal Ecology* **9**: 65-68**.**

Buée M., Reich M., Murat C., Morin E., Nilsson R.H., Uroz S., Martin F., 2009. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist* **184**: 449-456.

Caporaso J.G., Lauber C.L., Walters W.A., Berg-Lyons D., Huntley J., Fierer N., Owens S.M., Betley J., Fraser L., Bauer M., Gormley N., Gilbert J.A., Smith G., Knight R., 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal* **6**: 1621-1624.

Eckert K.A., Kunkel T.A., 1991. DNA polymerase fidelity and the polymerase chain reaction. *Genome Research* **1**: 17-24.

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.

Findley K., Oh J., Yang J., Conlan S., Deming C., Meyer J.A., Schoenfeld D., Nomicos E., Park M, NIH Intramural Sequencing Center Comparative Sequencing Program, Kong H.H., Segre J.A., 2013. Topographic diversity of fungal and bacterial communities in human skin. *Nature* **498**: 367-370.

Gobet A., Quince C., Ramette A., 2010. Multivariate cutoff levels analysis (MultoCoLA) of large community data sets. *Nucleic Acid Research* **38**(15): e155.

Haegeman B., Hamelin J., Moriarty J., Neal P., Dushoff J., Weitz J.S., 2013. Robust estimation of microbial diversity in theory and practice. *The ISME Journal* **7**: 1092-1101.

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.

Jumpponen A., Jones K.L., 2009. Massively parallel 454 sequencing indicated hyperdiverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. *New Phytologist* **184**: 438-448.

Kõljalg U., Nilsson R.H., Abarenkov K., Tedersoo L., Taylor A.F.S., Bahram M., Bates S.T., Bruns T.D., Bengtsson-Palme J., Callaghan T.M., Douglas B., Drenkhan T., Eberhardt U., Dueñas M., Grebenc T., Griffith G.W., Hartmann M., Kirk P.M., Kohout P., Larsson E., Lindahl B.D., Lücking R., Nguyen N.H., Niskanen T., Oja J., Peay K.G., Peintner U., Peterson M., Põldmaa K., Saag L., Saar I., Schüßler A., Scott J.A., Senés C., Smith M.E., Suija A., Taylor D.L., Telleria M.T., Weiss M., Larsson K., 2013. Toward a unified paradigm for sequence-based identification of fungi. *Molecular Ecology* **22**: 5271-5277.

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.

Lindahl B.D., Nilsson R.H., Tedersoo L., Abarenkov K., Carlsen T., Kjøller R., Kõljalg U., Pennanen T., Rosendahl S., Stenlid J., Kauserud H., 2013. Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *New Phytologist* **199**: 288-299.

Meiser A., Bálint M., Schmitt I., 2014. Meta-analysis of deep-sequenced fungal communities indicates limited taxon sharing between studies and the presence of biogeographic patterns. *New Phytologist* **201**: 623-635.

Nilsson R.H., Veldre V., Hartmann M., Unterseher M., Amend A., Bergsten J., Kristiansson E., Ruberg M., Jumpponen A., Abarenkov K., 2010. An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. *Fungal Ecology* **3**: 284-287.

Nilsson R.H., Tedersoo L., Lindahl B.D., Kjøller R., Carlsen T., Quince C., Abarenkov K., Pennanen T., Stenlid J., Bruns T., Larsson K., Kõljalg U., Kauserud H., 2011. Towards standardization of the description and publication of next-generation sequencing datasets of fungal communities. *New Phytologist* **191**: 314-318.

Porras-Alfaro A., Liu K., Kuske C.R., Xie G., 2014. From genus to phylum: large-subunit and internal transcribed spacer rRNA operon regions show similar classification accuracies influenced by database composition. *Applied and Environmental Microbiology* **80**: 829-840.

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 E.B., 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.

Shade A., Hogan C.S., Klimowicz A.K., Linske M., McManus P.S., Handelsman J., 2012. Culturing captures members of the soil rare biosphere. *Environmental Microbiology* **14**: 2247-2252.

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.

Unterseher M., Jumpponen A., Öpik M., Tedersoo L., Moora M., Dormann C.F., Schnittler M., 2011. Species abundance distributions and richness estimations in fungal metagenomics – lessons learned from community ecology. *Molecular Ecology* 20: 275-285.

Wang Q., Garrity G.M., Tiedje J.M. Cole J.R., 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into new bacterial taxonomy. *Applied and Environmental Microbiology* 73: 5267-5267.

Zhan A., Hulák M., Sylvester F., Huang X., Adebayo A.A., Abbot, C.L., Adamowicz S.J., Heath D.D., Cristescu M.E., Maclsaac H.J., 2013. High-sensitivity pyrosequencing of detection of rare species in aquatic communities. *Methods in Ecology and Evolution* 4: 58-565.