Commentary

Probing promise versus performance in longer read fungal metabarcoding

In the rapidly evolving world of methodologies to study fungi and other microorganisms, there has been growing interest in the adoption of so-called 'third-generation' technologies for high throughput amplicon sequencing (also known as metabarcoding). This interest is largely based on the capacity for longer sequence read lengths (> 500 bp), which have the potential to provide more accurate phylogenetic inference than current 'second generation' technologies (James *et al.*, 2016; Schloss *et al.*, 2016; Singer *et al.*, 2016). Despite this attraction, higher error rates and the high cost per base pair have inhibited the widespread adoption of this 'next' in microbial metabarcoding. Given the challenges in adopting any new technology, careful benchmarking tests are needed to determine whether significantly greater insights can be gleaned, or if currently established methods remain sufficient. In this issue of *New Phytologist*, Tedersoo *et al.* (2018; pp. 1370–1385) conduct the first of these benchmarking tests for fungi and other eukaryotes, directly comparing Illumina MiSeq (i.e. second generation) and PacBio datasets (i.e. third generation) generated from the same soil samples and analyzed for taxonomic richness and composition. While their collective analyses indicate that longer amplicons can be successfully generated with relatively low error rates, they also demonstrate many technical issues with PacBio-based data, which require careful attention.

'... continued efforts to improve the metabarcoding capacity of third generation technologies are needed, as longer reads will ultimately improve both our taxonomic and ecological understanding of fungal communities.'

The Tedersoo *et al.* (2018) study had a set of specific objectives that largely mirrored previous benchmarking studies by this research group (e.g. Tedersoo *et al.*, 2010, 2015). For their comparisons, they utilized a 24-species mock community sample, and 13 forest soil samples taken from Papua New Guinea, which had previously been analyzed via MiSeq sequencing in Tedersoo *et al.* (2015). Coupled with those samples, which were analyzed for fungi, the authors also assessed 20 Estonian forest nursery samples for oomycete diversity. In the fungal analyses they targeted a wide range of rRNA gene regions (ITS1, ITS2, full ITS, SSU + ITS + LSU) using 14 different primer pairs, while they focused on a single ITS + LSU dataset in the oomycete analysis. Along with those variables, which emphasized differences in gene region target and amplicon length, they also compared different aspects of the PacBio sequencing process (diffusion loading vs Magbead loading) and platform (RSII vs Sequel) that can influence sequence read length abundances as well as total sequence read counts.

Many of their results matched those expected. For example, the new Tedersoo *et al.* (2018) study clearly demonstrates that the taxonomic richness and composition of fungi in samples sequenced using PacBio technology is highly dependent on specific primer combination. This has been well documented with other sequencing technologies (Tedersoo *et al.*, 2015), reiterating the need for researchers to apply caution when comparing fungal richness estimates and compositional patterns across studies. They also showed that despite advances in error rate correction, some biases in PacBio sequences remain, particularly in insertion–deletion (i.e. indel) rates, compared to those generated from MiSeq. Importantly, however, they presented two results which indicated that longer reads provide notable improvements on current options for fungal identification. Specifically, the precision of the genus-level identifications of fungi and other eukaryotic taxa was 33% higher when using full-length ITS sequences compared to using either ITS1 or ITS2 reads only. Similarly, the addition of either SSU or LSU data also facilitated taxonomic assignment at higher taxonomic levels, with > 50% of 'unknown' samples (based on ITS data) receiving assignments to a phylum when any of these rRNA coding regions were considered. Methodologically, the authors observed that the newer PacBio sequencing platform (Sequel) provided five times more sequence reads per library than the RSII platform. Furthermore, using the Magbead instead of diffusion loading notably reduced the short fragment bias in PacBio libraries. Despite these improvements, the authors were unable to get the longest read targets (SSU + ITS + LSU) to amplify well, precluding use of longer sequence reads as a way to syntonize the currently region-specific approach of arbuscular mycorrhizal fungal metabarcoding (Schlaeppi *et al.*, 2016).

While the analyses presented by Tedersoo *et al.* (2018) provide important details about the current utility of PacBio technology for improving fungal taxonomy in metabarcoding, they do not address two important properties for its use by fungal ecologists. First, the ability to multiplex many ecologically independent samples into a single metabarcode run has been arguably the key breakthrough in dramatically up-scaling patterns of α- and β-diversity for fungi (Kennedy *et al.*, 2012; Tedersoo *et al.*, 2014). In Illumina-based datasets, for example, it is routine to multiplex hundreds of ecologically independent samples into a single MiSeq library and obtain millions of sequence reads, which, following quality

This article is a Commentary on Tedersoo *et al.*, **217**: 1370–1385.

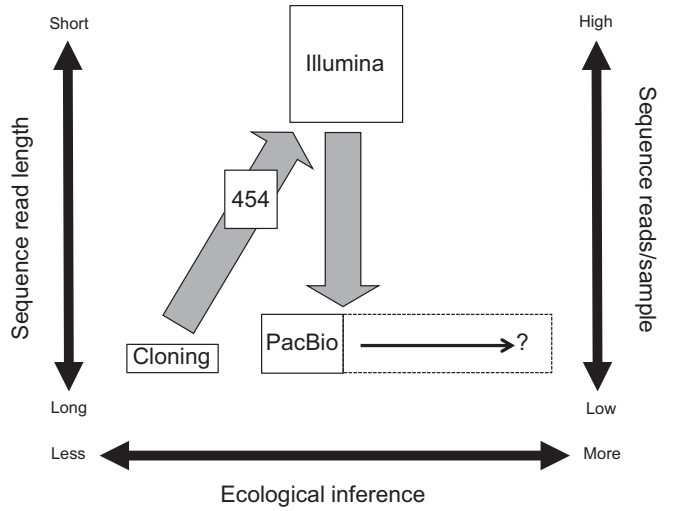

Fig. 1 Progression of metabarcoding sequencing methods for fungi. Box height is proportional to number of ecologically independent samples that can be analyzed in a given study. Other technologies such as Ion Torrent have similar capacities to Illumina, while MinION has similar long-read capacities to PacBio.

filtering, results in thousands of high quality reads per sample. In the current Tedersoo *et al.* (2018) study, the authors multiplexed between 30 and 47 independent samples per library and, following quality filtering, libraries containing 440 and 103 mean and median sequence reads per sample, respectively. Given the high diversity that is present in many fungal communities (Peay *et al.*, 2016), this raises a question of whether long read PacBio sequencing (i.e. targeting full ITS + flanking rRNA subunit regions) can provide sufficient sequence read depth per sample to accurately capture ecological differences in species richness and community composition (Fig. 1). A second question is how phylogenetically-based analyses of fungal community composition could also influence the ecological interpretation of data obtained from longer sequences. Having reads that include both the ITS region, which captures species-level variation for many taxa (Schoch *et al.*, 2012), along with the flanking LSU or SSU regions, which can be aligned for kingdom-level phylogenetic analyses, would give fungal ecologists the ability to couple accurate taxonomic assignments with the phylogenetically-based analysis techniques widely adopted in prokaryotic metabarcoding (Lozupone & Knight, 2005).

To address these issues, we compared the results of 47 samples that we collected from four distinct habitats, which were sequenced in parallel using both MiSeq and PacBio technologies. For the PacBio sequencing, we targeted the full ITS + LSU region using the ITS1F-TW13 primer pair, whereas for the MiSeq data we targeted the ITS1 region using the ITS1F-ITS2 primer pair. Here, we focused the most common metrics used in ecological analyses, namely species richness and community composition. For the former, we analyzed a 25 species mock community sample that was previously assessed in Nguyen *et al.* (2015). For the latter, we compared fungal community composition at two ecological scales. We first examined differences among samples from forest soil, tree litter, tree roots and tree wood. Based on comparable previous analyses, we expected community composition to differ significantly between all these habitats (Cline *et al.*, 2017). We then assessed differences in fungal community composition within each habitat: soil collected in *Quercus* vs *Pinus* forests, decaying *Quercus* vs *Pinus* litter from the forest floor, *Quercus* vs *Pinus* roots, and live vs decaying *Betula* wood. See the Supporting Information Methods S1 for additional details on sampling, sequencing, and bioinformatics processing.

Our results in terms of sequence read depth per sample matched our expectations for the two sequencing technologies. For example, following quality filtering, there was a total of 97 593 mock community sequence reads in the Illumina dataset vs only 220 reads in the PacBio dataset (Table 1). For the Illumina dataset, we recovered 24 of the 25 members of the mock community when the data was not rarefied. However, when the mock community data was rarefied to 1000 and 100 reads, species recovery dropped to 23 and 17 species, respectively. For the PacBio dataset, we were able to recover only 18 species in the mock community. Notably, this

Table 1 Comparison of sequence read counts and species richness of the same mock community sample sequenced on either Illumina MiSeq or PacBio Sequel platforms

Sequencing technology Species	Illumina MiSeq (not rarefied)	Illumina MiSeq (rarefaction = 1000)	Illumina MiSeq (rarefaction = 100)	PacBio Sequel (not rarefied)
Amanita muscaria	18 114	195	20	73
Thelephora terrestris	10 904	57	10	10
Cortinarius sp.	9251	89	10	27
Helvella vespertina	8463	85	9	29
Pholiota spumosa	7063	81	6	10
Lactarius sp.	6417	63	4	4
Tricholoma sp.	6321	63	8	16
Suillus americanus	6176	53	3	11
Xerocomus subtomentosus	5849	55	0	4
Suillus granulatus	3414	41	3	1
Paxillus cuprinus	3312	36	4	17
Helvella dryophila	3225	31	0	6
Boletus edulis	2186	13	7	0
Suillus luteus	1937	12	2	3
Entoloma abortivum	1330	78	5	5
Suillus grevillei	957	7	1	1
Suillus laricinus	769	17	2	0
Suillus grisellus	586	5	1	1
Wilcoxina mikolae	514	2	0	0
Suillus spectabilis	314	5	0	0
Laccaria laccata	275	3	0	1
Phaeoclavulina curta	186	3	1	0
Leucopaxillus gentianeus	17	1	0	1
Hygrophorus russula	13	0	0	0
Sequence sum	97 593	995	96	220
Total species	24	23	17	18

The composition is based on Nguyen *et al.* (2015), except that *Cantherellus* sp. and *Leccinum* sp. were removed before amplification. For the Illumina MiSeq data, rarefaction was applied at 1000 and 100 sequences (sequence sums in those columns do not exactly match the level of rarefaction because sequence reads matching to non-mock taxa were excluded). *Leucopaxillus albissimus* was not successfully amplified on either platform.

included five species represented by a single sequence read, which, based on the recommendations of Tedersoo *et al.* (2018), should be removed. Collectively, these differences in species richness patterns between the Illumina and PacBio datasets suggest that, at present, metabarcoding with third generation technologies is not yet sufficient to accurately characterize fungal community richness, most likely due to the much lower sequence throughput compared to second generation technologies.

To understand how sequence depth and length influenced fungal community composition, we first assessed whether rarefaction of the Illumina dataset to the level of sequence reads per sample present in the PacBio dataset changed any ecological interpretations about community dissimilarity. We found that when we compared the rarefied Illumina dataset at 1000 reads per sample vs 100 reads per sample the differences in community dissimilarity were functionally equivalent at both ecological scales (i.e. among and within habitats) (Table S1a). Somewhat surprisingly, this result indicates that the number of sequence reads per sample, at least with > 100 reads per sample, is unlikely to be a major limitation in differentiating environmental samples based on fungal community composition. When ITS1 regions from the Illumina and PacBio datasets were directly compared, we again found that the results were equivalent, both within and among habitats (Table S1b). Although both of these analyses of fungal community composition indicate that the ITS1 Illumina and PacBio datasets can perform comparably, until third generation technologies such as PacBio increase multiplexing capacity without sacrificing read depth and dramatically lower per base pair cost (currently an order of magnitude higher), we believe second generation technologies will remain a more popular choice.

Finally, to analyze the potential of the PacBio dataset to be used for phylogenetically-based community analyses, we took all of the PacBio sequences, which had been clustered and delineated into operational taxonomic units (OTUs) on the basis of the ITS1 portion of the sequences, and identified the corresponding ITS2 and LSU portions. We used this approach for the LSU portions in particular because there is currently no comprehensive and curated database for analyzing LSU sequences, as there is for ITS via UNITE (Kõljalg *et al.*, 2013). Focusing again on the mock community sample, we found that the number of chimeric sequences remained high despite applying a standard chimera-checking step in our bioinformatics pipeline. Specifically, we found that 61% and 65% of the mock species contained ITS2 or LSU sequences that did not match their expected identity as assigned by ITS1, respectively (Table 2). The number of taxonomically incorrect sequences varied across species and by gene region (Table S2) and typically spanned large phylogenetic distances (e.g. sequences belonging to *Cortinarius* sp. based on ITS1 taxonomy had corresponding ITS2 and LSU sequences matching to *Suillus*). We believe this strikingly elevated chimera rate is not due to a high number of PCR cycles (we used 28 cycles, which was lower than the minimum number used in the Tedersoo *et al.* (2018) study (range: 30–42)), but rather to the low sensitivity of chimera checking algorithms in sequence datasets characterized by very few repetitive sequence reads. Given that we had no simple way to validate which of the individual LSU sequences were correct within each environmental sample OTU, it was not possible to conduct phylogenetically-based community analyses with any confidence.

Table 2 Analysis of chimeras by gene region in the mock community sample sequenced with PacBio after quality filtering (including chimera checking)

	PacBio ITS2 Chimeric sequences (%)	PacBio LSU Chimeric sequences (%)
Species		
Amanita muscaria	5	1
Cortinarius sp.	22	23
Entoloma abortivum	20	25
Helvella dryophila	0	0
Helvella vespertina	0	0
Laccaria laccata	0	NA
Lactarius sp.	50	100
Leucopaxillus gentianeus	100	100
Paxillus cuprinus	6	13
Pholiota spumosa	0	0
Suillus americanus	18	30
Suillus granulatus	0	0
Suillus grevillei	0	0
Suillus grisellus	0	0
Suillus luteus	67	50
Thelephora terrestris	20	40
Tricholoma sp.	6	20
Xerocomus subtomentosus	75	50
Average	22	27

Chimeric sequences are defined as those not matching the expected genus within any of the 10 best matches based on BLAST of the INSD databases. NA, sequence too short to obtain confident identity.

Taken together, these more ecologically focused analyses suggest that fungal metabarcoding based on third generation technologies, as currently implemented, are inadequate compared to those of second generation technologies. While this conclusion is somewhat different from that reached in the Tedersoo *et al.* (2018) study, we agree that continued efforts to improve the metabarcoding capacity of third generation technologies are needed, as longer reads will ultimately improve both our taxonomic and ecological understanding of fungal communities. To help realize this potential in the short-term, there will need to be renewed efforts by mycologists to generate larger, curated databases for gene regions beyond ITS. Additionally, alternative bioinformatics approaches need to be applied that better capture the unique properties of third generation-based datasets. Ultimately, we believe that although advances in sequencing methodologies can generate new perspectives, the most significant advances in fungal biology will continue to come from studies prioritizing novel research questions and well-designed experiments.

Acknowledgements

We thank B. Auch for assistance with PacBio library preparation at the University of Minnesota Genomics Center, C. Fernandez for participation in, and K. Beckman for the sponsorship of, 'GenoFest', and both I. Dickie as well as K. Peay for ongoing discussions about this topic.

Peter G. Kennedy*, Lauren C. Cline and Zewei Song

Department of Plant Biology, University of Minnesota, St Paul, MN 55108, USA
(*Author for correspondence: tel +1 612 624 8519; email kennedyp@umn.edu)

References

Cline LC, Song Z, Al-Ghalith GA, Knights D, Kennedy PG. 2017. Moving beyond *de novo* clustering in fungal community ecology. *New Phytologist* **216**: 629–634.

James TY, Marino JA, Perfecto I, Vandermeer J. 2016. Identification of putative coffee rust mycoparasites via single-molecule DNA sequencing of infected pustules. *Applied and Environmental Microbiology* **82**: 631–639.

Kennedy PG, Matheny PB, Ryberg KM, Henkel TW, Uehling JK, Smith ME. 2012. Scaling up: examining the macroecology of ectomycorrhizal fungi. *Molecular Ecology* **21**: 4151–4154.

Kõljalg U, Nilsson RH, Aberenkov K, Tedersoo L, Taylor AF, Bahram M, Bates ST, Bruns TD, Bengtsson-Palme J, Callaghan TM *et al.* 2013. Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology* **22**: 5217–5277.

Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71**: 8228–8235.

Nguyen N, Smith D, Peay KG, Kennedy PG. 2015. Parsing ecological signal from noise in fungal next generation sequencing. *New Phytologist* **205**: 1389–1393.

Peay KG, Kennedy PG, Talbot JM. 2016. Dimensions of biodiversity in the Earth mycobiome. *Nature Reviews Microbiology* **14**: 434–447.

Schlaeppi K, Bender SF, Mascher F, Russo G, Patrignani A, Camenzind T, Hempel S, Rillig MC, van der Heijden MGA. 2016. High-resolution community profiling of arbuscular mycorrhizal fungi. *New Phytologist* **212**: 780–791.

Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK. 2016. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ* **4**: e1869.

Schoch CL, Siefert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, the Fungal Barcoding Consortium. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for *Fungi*. *Proceedings of the National Academy of Sciences, USA* **109**: 6241–6246.

Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, Levy A, Gies EA, Cheng J-F, Copeland A, Klenk H-P *et al.* 2016. High-resolution phylogenetic microbial community profiling. *The ISME Journal* **10**: 2020–2032.

Tedersoo L, Anslan S, Bahram M, Põlme S, Riit T, Liiv I, Kõljalg U, Kisand V, Milsson H, Hildebrand F *et al.* 2015. Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys* **10**: 1–43.

Tedersoo L, Bahram M, Põlme S, Kõljalg U, Yorou NS, Wijesundera R, Villarreal Ruiz L, Vasco-Palacios AM, Thu PQ, Suija A *et al.* 2014. Global diversity and geography of soil fungi. *Science* **346**: 1078.

Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Dhuyong 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.

Tedersoo L, Tooming-Klunderud A, Anslan S. 2018. PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *New Phytologist* **217**: 1370–1385.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Table S1 Comparison of fungal community dissimilarity among and within habitats for the Illumina and PacBio datasets

Table S2 Results of BLAST matching to the INSD databases of ITS1, ITS2 and LSU portions of the PacBio-based sequences for each of the 18 species in the mock community sample

Methods S1 Details of sampling collection, DNA extraction and amplification, bioinformatics processing and statistical analyses.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

Key words: high throughput sequencing, long read, MiSeq, PacBio, third generation.