Supplementary Information for:Long- and short-read metabarcoding technologies reveal similar spatio-temporal structures in fungal communities

Brendan Furneaux[1], Mohammad Bahram[2,3], Anna Rosling[4], Nourou S. Yorou[5], Martin Ryberg[1]

[1]Program in Systematic Biology, Department of Organismal Biology, Uppsala University, Uppsala, Sweden

[2]Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

[3]Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia

[4]Program in Evolutionary Biology, Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden

[5]Research Unit in Tropical Mycology and Plant-Fungi Interactions, LEB, University of Parakou, Parakou, Benin

Supplementary file 1: Mapping between samples, indexed primers, and read accession numbers for short amplicons.

Supplementary file 2: Mapping between samples, indexed primers, and read accession numbers for long amplicons.

Supplementary file 3: ML tree of all long metabarcoding ASVs (5.8S–ITS2–LSU) assigned to kingdom Fungi. Thick branches represent rapid bootstrap values >90%. Taxonomic assignment by PHYLOTAX, based on assignments using RDPC, SINTAX, and IDTAXA, and Unite, Warcup, and RDP reference databases. Taxon names in blue were not fully resolved with PHYLOTAX, and are given as a list of alternative assignments, along with the number of methods which gave each assignment (max 9). Taxon names in red were reconstructed as polyphyletic. In some cases, this may be due to a poorly identified sequence "splitting" the group (e.g., the single *Aspergillus* sequence nested in *Monascus*, Aspergillaceae and Trichocomaceae sequences nested in Onygenales), but in other cases, the tree has clearly failed to reconstruct groups which are generally considered to be monophyletic based on multi-gene trees (e.g., Dothidiomycetes, Cantharellales).

List of Figures

S1 Unconstrained Principal Coordinate Analysis ordination of samples preserved with different buffers . . . 3

S2 Comparison of read counts by taxonomic groups for different preservation buffers (Xpedition vs. LifeGuard) and sampling year, for long amplicon libraries. . . . 4

S3 Comparison of read counts by taxonomic groups for different preservation buffers (Xpedition vs. LifeGuard) and sampling year, for short amplicon libraries. . . . 5

S4 Portion of the rDNA showing the 5.8S rDNA, partial SSU and LSU rDNA, and internally transcribed spacer (ITS) regions . . . 6

S5 Plot of total number of ASVs recovered vs. fraction of demultiplexed reads successfully mapped to ASVs for different rDNA regions extracted from long PacBio amplicons using `LSUx` . . . 7

S6 Phylogenetic refinement of taxonomic assignments (PHYLOTAX) . . . 8

S7 DNA concentrations after extraction and PCR, and sequencing depth along transects at the two sites Ang and Gan for the years 2015 and 2016 . . . 9

S8 Comparison of the length of denoised sequences from different length amplicons and sequencing technologies . . . 10

S9 Length distribution for each region extracted from long amplicons . . . 11

S10 Heat tree summarizing taxonomic composition of soil community . . . 12

S11 Assignment of ECM status to fungal ASVs using FUNGuild database and taxonomic assignments 13

S12 Partial Principal Coordinates Analysis (PPCoA) of fungal community using different metabarcoding approaches . . . 14

S13 Mantel correlograms for community dissimilarities and spatio-temporal distance . . . 15

S14 Comparison of the length of denoised ITS2 sequences extracted from different length amplicons and sequencing technologies . . . 16

S15 Variation in read abundance and diversity for different amplicons across the taxonomic tree . 17

S16 Kingdom-level taxonomic composition of ITS2 ASVs ≤ 140 bp . . . 18

S17 Heat tree showing variation in read abundance and diversity for different amplicons, for ECM taxa only . . . 19

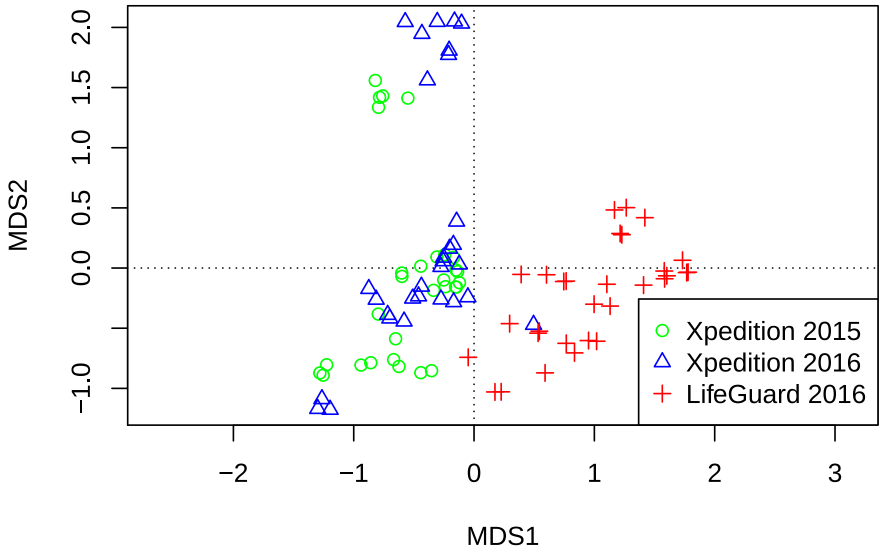

Figure S1: Unconstrained Principal Coordinate Analysis ordination of samples preserved with different buffers. For each sequencing strategy, only those sampling locations with more than 100 reads each for all three year/preservation buffer combinations were included.

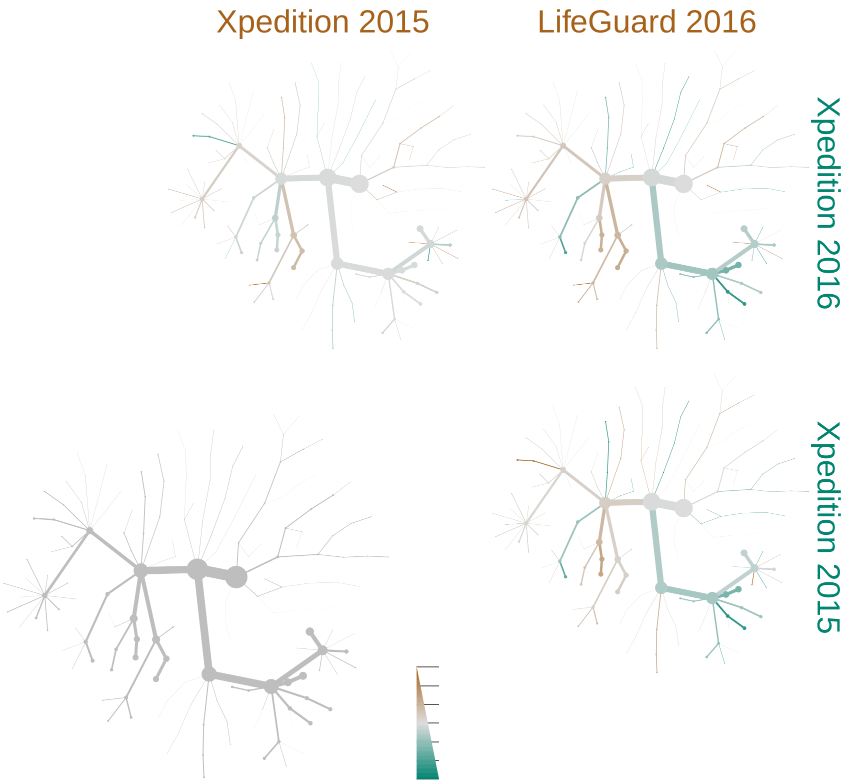

Figure S2: Comparison of read counts by taxonomic groups for different preservation buffers (Xpedition vs. LifeGuard) and sampling year, for long amplicon libraries.

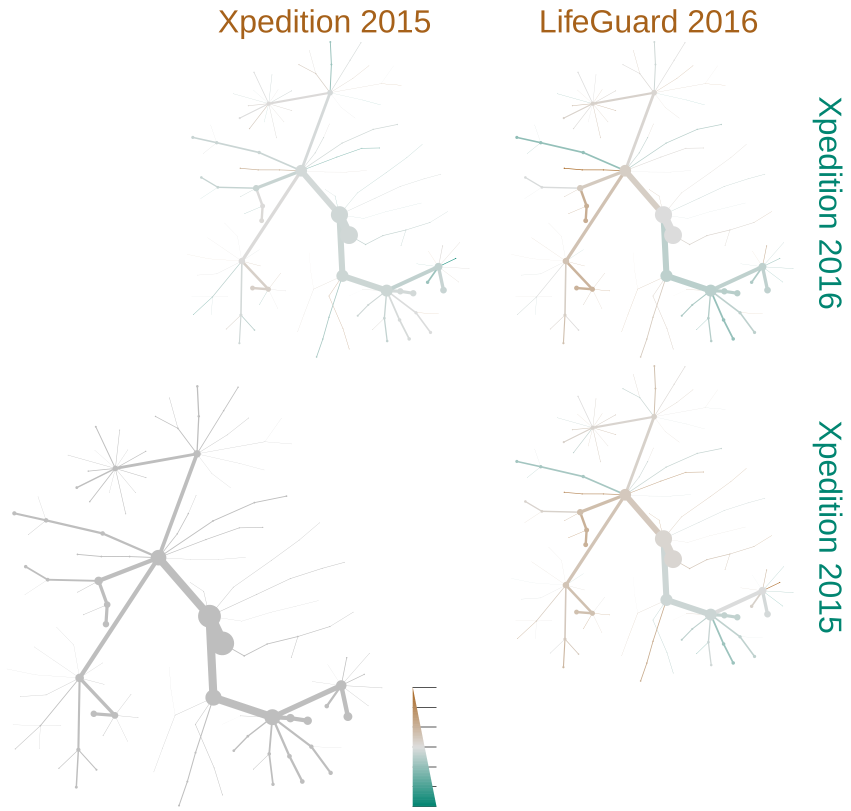

Figure S3: Comparison of read counts by taxonomic groups for different preservation buffers (Xpedition vs. LifeGuard) and sampling year, for short amplicon libraries.

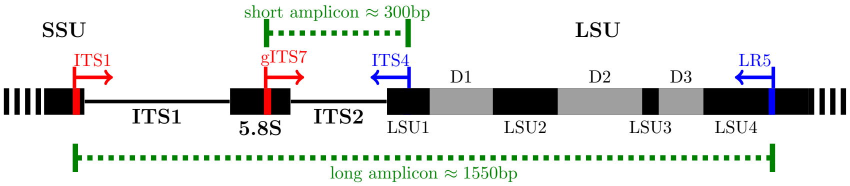

Figure S4: Portion of the rDNA showing the 5.8S rDNA, partial SSU and LSU rDNA (thick lines), and internally transcribed spacer (ITS) regions (thin black lines). D1-3 (grey) represent the first three variable regions in LSU, while LSU1-4 (black) represent the conserved regions. Primer sites used in this study are indicated in red (forward primers) and blue (reverse primers), and the resulting amplicons are shown with green braces.

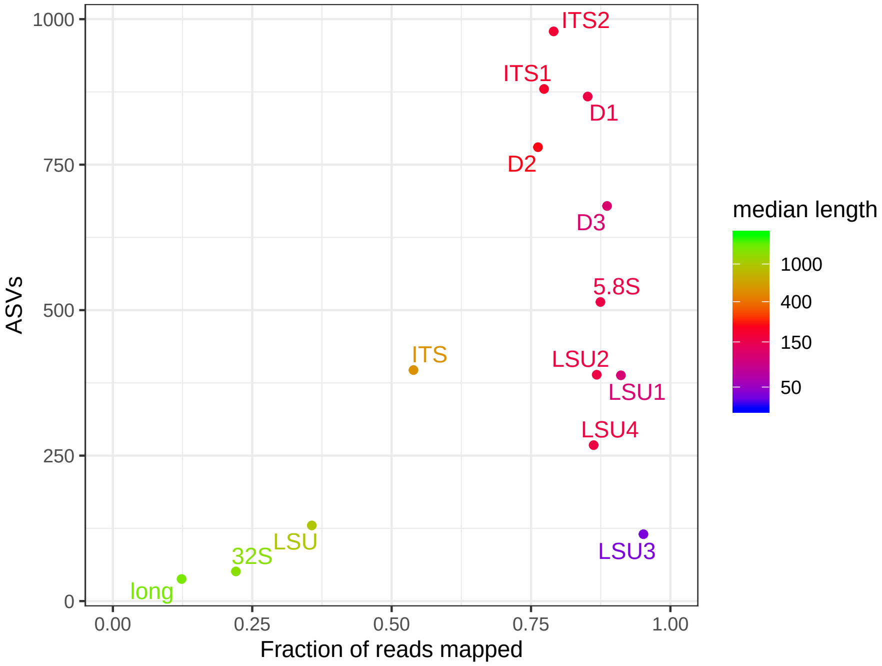

Figure S5: Plot of total number of ASVs recovered vs. fraction of demultiplexed reads successfully mapped to ASVs for different rDNA regions extracted from long PacBio amplicons using `LSUx`. Some regions were formed by concatenation of extracted regions, as follows: LSU = LSU1–D1–LSU2–D2–LSU3–D3–LSU4; ITS = ITS1–5.8S–ITS2; 32S = 5.8S–ITS2–LSU; long = ITS–LSU. For this figure, subregions were concatenated prior to quality filtering and generating ASVs with DADA2.

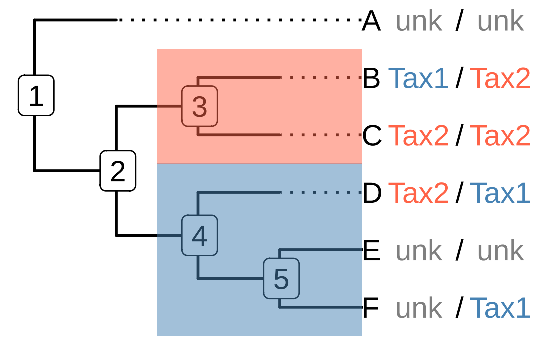

Figure S6: Phylogenetic refinement of taxonomic assignments (PHYLOTAX). In the example, a clade includes seven OTUs (A-F), which have been identified by two methods as belonging to taxa “Tax1” or “Tax2”, or are unidentified (“unk”), as shown in the tip labels. No taxon is assigned at node 1, because one of the child branches (A) is completely unassigned. No taxon is assigned at node 2, because the assignments at C and F are inconsistent. Node 3 is assigned to Tax1 because this is consistent with at least one of the assignments for both B and C. Node 4 (and thus also node 5) is assigned to Tax2 because this is consistent with at least one of the assignments for D and F, and because E is completely unassigned.

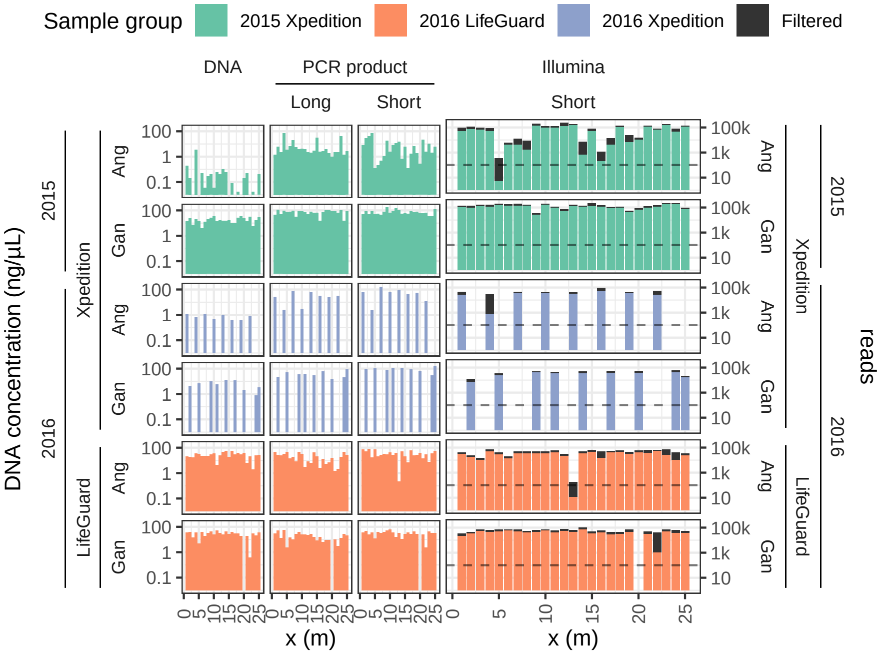

Figure S7: DNA concentrations after extraction (leftmost column) and PCR (second and third columns), and sequencing depth (right three columns) along transects at the two sites Ang and Gan for the years 2015 and 2016. Sequencing reads which were filtered out during quality control are colored dark grey. Dotted horizontal line at 100 reads indicates cutoff for inclusion in community analysis. 2016 samples were preserved using two different methods (LifeGuard, Xpedition), while 2015 samples were preserved only using Xpedition.

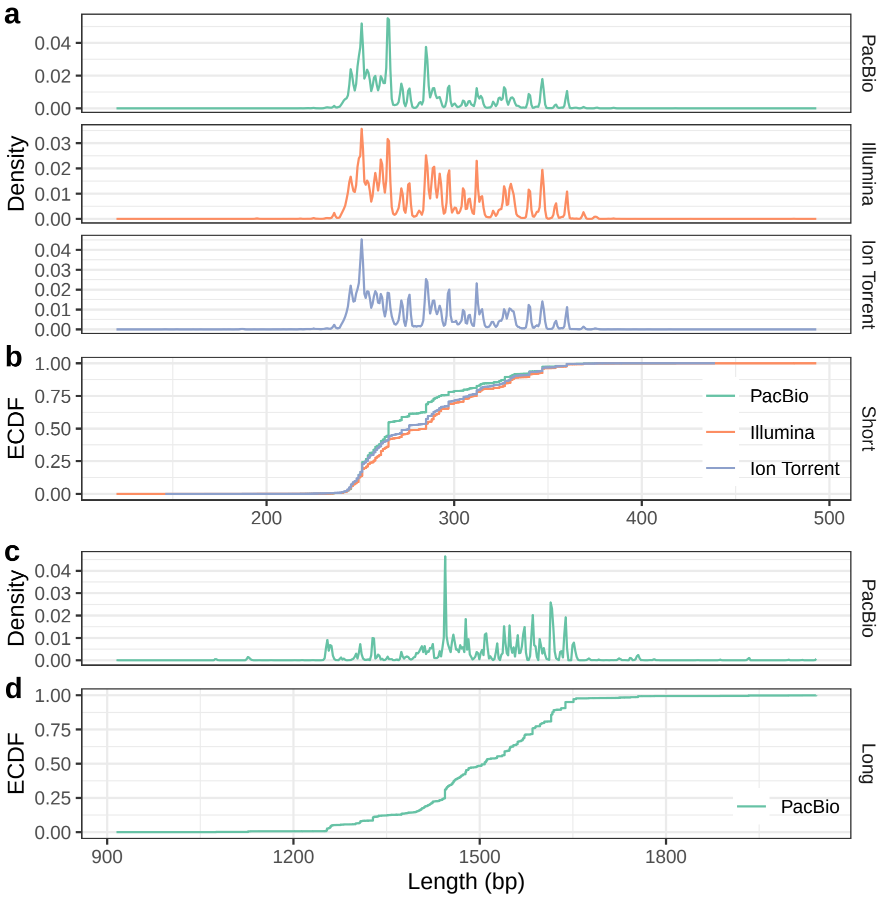

Figure S8: Comparison of the length of denoised sequences from different length amplicons and sequencing technologies. Length distribution (*a*, *c*) and empirical cumulative distribution function (ECDF; *b*, *d*) for short (*a*, *b*) and long (*c*, *d*) amplicons, respectively.

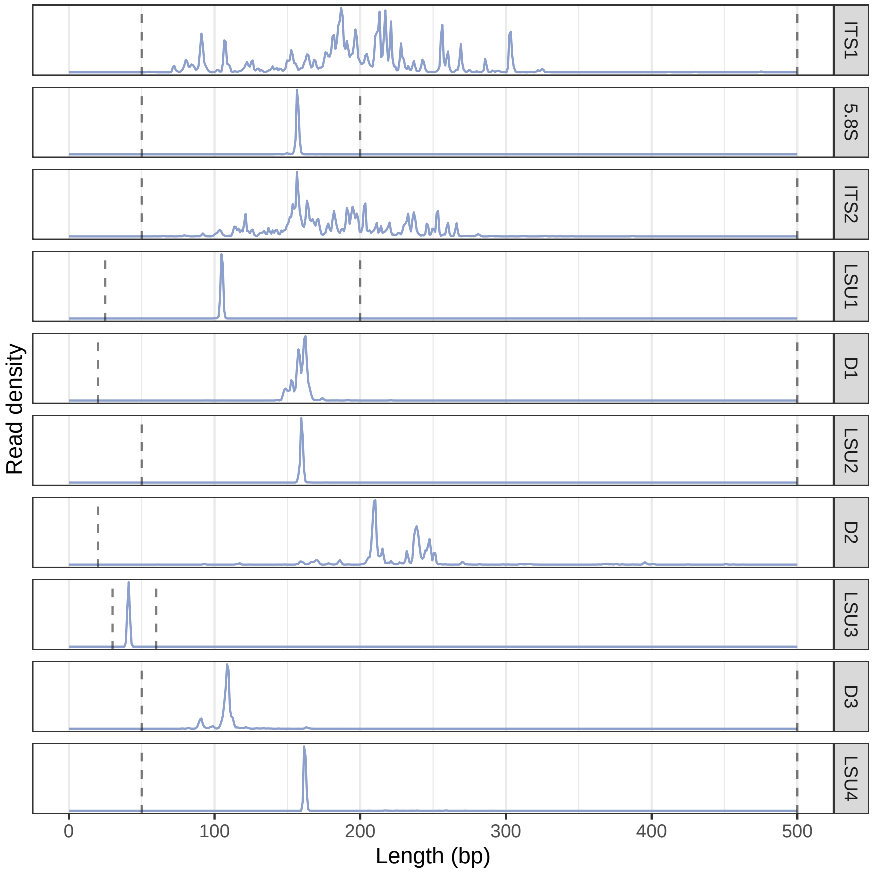

Figure S9: Length distribution for each region extracted from long amplicons. Filtering limits are shown as dashed vertical lines.

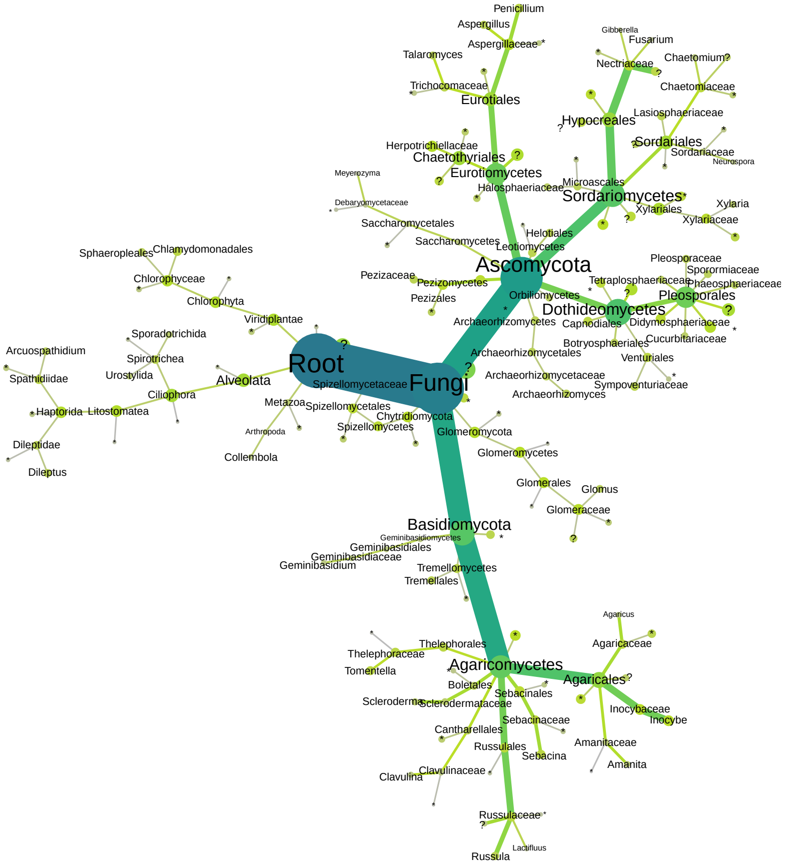

Figure S10: Heat tree summarizing taxonomic composition of soil community. Color and size of nodes represent the fractional ASV richness of that taxonomic group. Color and thickness of branches represent the fractional read abundance of all ASVs belonging to that taxonomic group. Diverse but relatively rare groups are thus shown as large, dark-colored nodes on relatively thin, light-colored branches (e.g., Dothideomycetes), while common but relatively non-diverse groups are shown as small, light-colored nodes on relatively thick, dark-colored branches (e.g., Russulales). Groups which are not represented by at least 1% of reads or 1% of ASVs in any dataset are collapsed into nodes labeled ”*“. ASVs which could not be further assigned are labelled”?”. ASV richness and read abundance are displayed as the mean across sequencing runs. Taxonomic assignment is by strict consensus of multiple ITS databases (Unite, Warcup) and assignment algorithms (RDPC, SINTAX, IDTAXA). The RDP LSU database was not included, to ensure consistent assignment between long and short amplicon libraries.

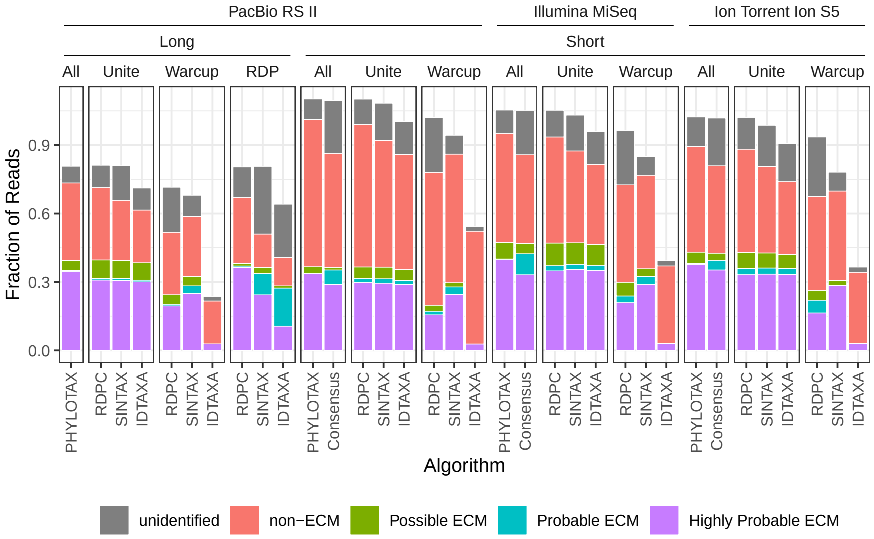

Figure S11: Assignment of ECM status to fungal ASVs using FUNGuild database and taxonomic assignments. Reads which were not identified to kingdom, or which were identified as a non-Fungi kingdom, are not included. “unidentified” denotes ASVs which were assigned to kingdom Fungi, but could not be identified to family level or could not be assigned ECM status based on the available identification. Height of each bar represents the fraction of reads represented by that bar. Results are shown for different sequencing technologies, amplicons (Long, Short), reference databases (Unite, Warcup, RDP; “All” denotes algorithms which combine results from multiple databases), and taxonomic assignment algorithms.

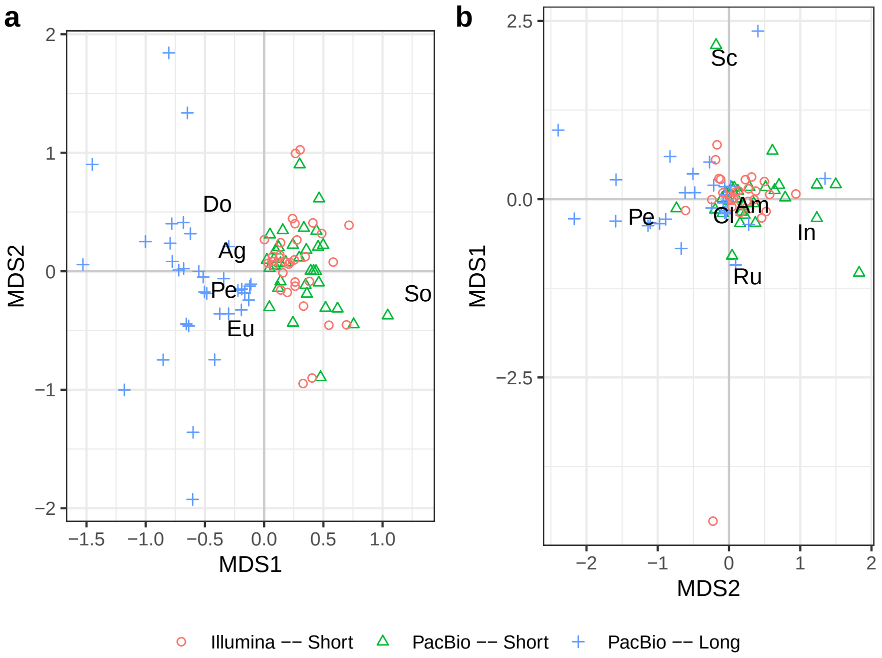

Figure S12: Partial Principal Coordinates Analysis (PPCoA) of fungal community using different metabarcoding approaches, for all fungi (**a**) and ECM fungi (**b**). Spatio-temporal variation (i.e., variation between different soil samples) has been partialled out. Only samples with at least 100 reads for all three approaches are included. ASVs are clustered by taxonomic class for all fungi, and by family for ECM fungi. Abbreviations represent the approximate direction of increase for the classes and families with the highest scores: **So**: Sordariomycetes, **Ag**: Agaricomycetes, **Pe**: Pezizomycetes, **Eu**: Eurotiomycetes, and **Do**: Dothideomycetes; **In**: Inocybaceae, **Am**: Amanitaceae, **Sc**: Sclerodermataceae, **Ru**: Russulaceae, **Pe**: Pezizaceae, and **Cl**: Clavulinaceae.

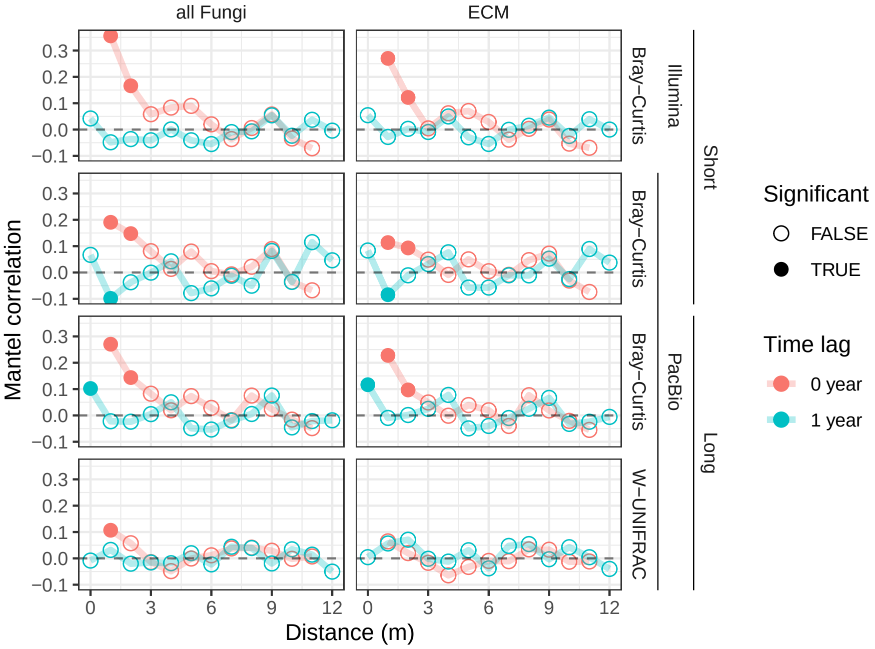

Figure S13: Mantel correlograms for community dissimilarities and spatio-temporal distance. Community dissimilarities are Bray-Curtis (top three rows) or weighted Unifrac (bottom row) dissimilarities and spatial distance, for short (top two rows) and long (bottom two rows) amplicons, and for the entire soil fungal community (left) or only ECM taxa (right). Unifrac distance was not calculated for short amplicons because it requires that the sequences are placed on a phylogenetic tree.

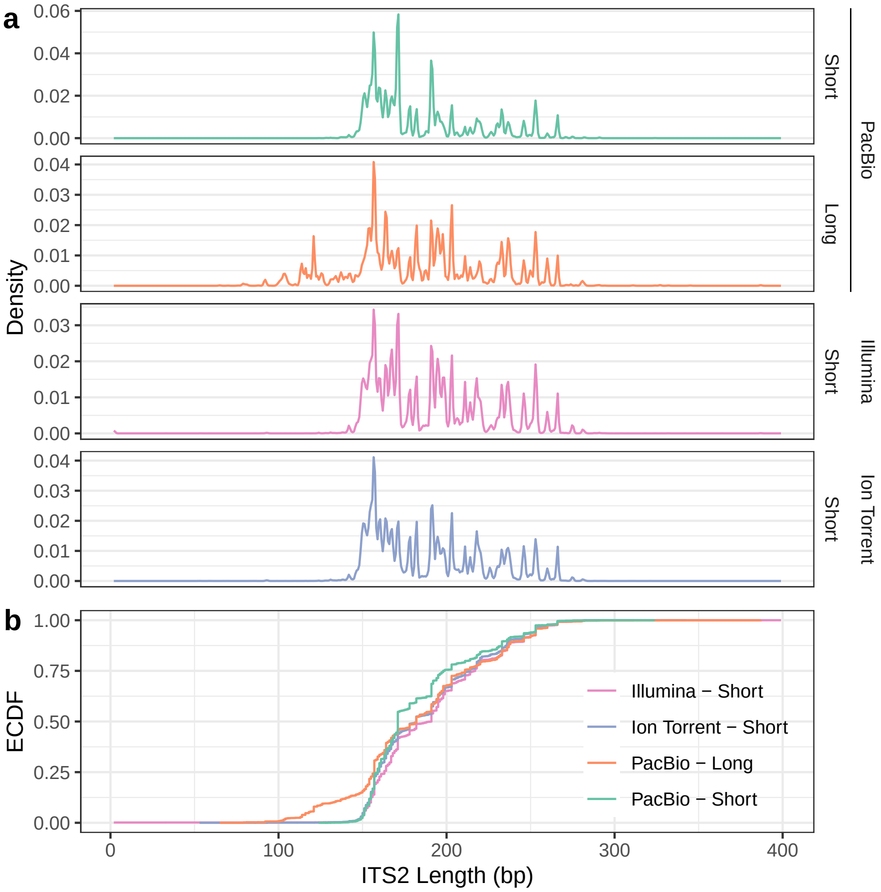

Figure S14: Comparison of the length of denoised ITS2 sequences extracted from different length amplicons and sequencing technologies. *a*) Length distribution. *b*) Empirical cumulative distribution function.

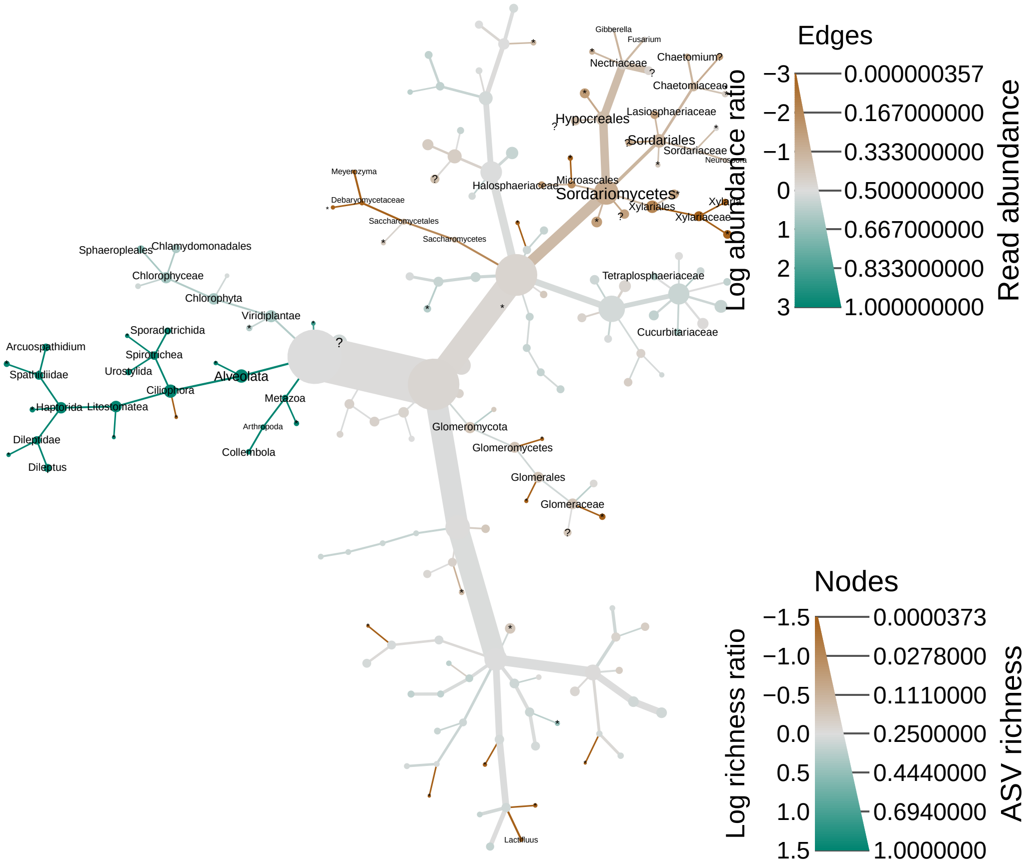

Figure S15: Variation in read abundance and diversity for different amplicons across the taxonomic tree. Color of nodes indicates variation in read abundance, while color of edges represents variation in ASV richness. In both cases red represents increased prevalence in the short amplicon library, while blue-green represents increased prevalence in the long amplicon library. Values are $\log_{10}$ of the ratio of mean values across sequencing runs for each amplicon. Groups which are not represented by at least 1% of reads or 1% of ASVs in any dataset are collapsed into nodes labeled "*". ASVs which could not be further assigned are labelled"?". Only nodes with a log read abundance ratio greater than 0.5 (abundance ratio > 3.2) are labeled. Refer to Figure S10 for complete taxon labeling.

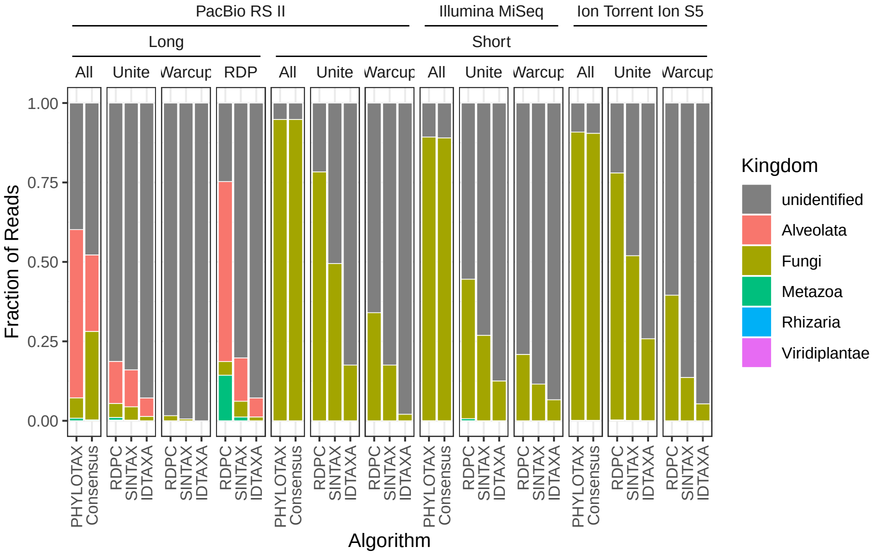

Figure S16: Kingdom-level taxonomic composition of ITS2 ASVs ≤ 140 bp. Vertical axis shows fraction of total reads ≤ 140 bp in each dataset.

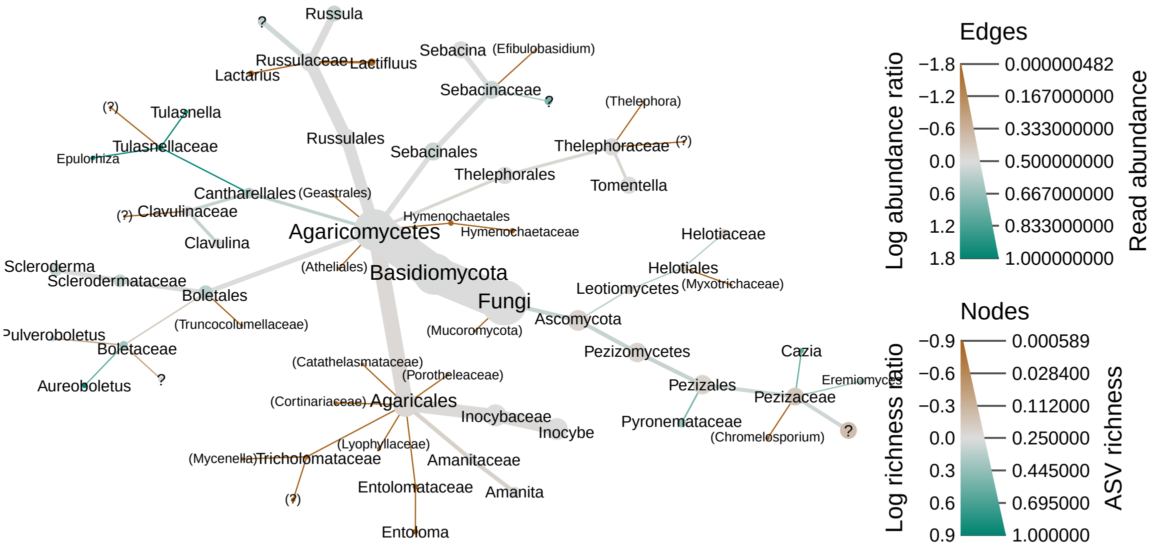

Figure S17: Heat tree showing variation in read abundance and diversity for different amplicons, for ECM taxa only. Red edges (nodes) represent increased read (ASV) count for the short amplicon library, while blue-green edges (nodes) represent increased read (ASV) count for the long amplicon library. Values are $\log_{10}$ of the ratio of mean values across all sequencing technologies for each amplicon. Groups which are not represented by at least 1% of reads or 1% of ASVs in any dataset are collapsed into nodes labeled ”*“. ASVs which could not be further assigned are labelled”?”. Taxonomic assignment is by strict consensus of multiple ITS databases (Unite, Warcup) and assignment algorithms (RDPC, SINTAX, IDTAXA). The RDP LSU database was not included, to ensure consistent assignment between long and short amplicon libraries.

List of Tables

S1 Minimum and maximum allowed lengths for each extracted region. 21

S2 Number of Amplicon Sequence Variants (ASV) and reads at different pipeline stages 22

S3 Correspondences between ASVs found by different sequencing strategies 23

S4 Correspondences between 97% OTUs found by different sequencing strategies 24

S5 Parameters for exponential distance-decay fits. Range: range at which exponential function is at 95% of its maximum value. 25

Table S1: Minimum and maximum allowed lengths for each extracted region.

Region	Min. length	Max. length
ITS1	50	500
5.8S	50	200
ITS2	50	500
LSU1	25	200
D1	20	500
LSU2	50	500
D2	20	2999
LSU3	30	60
D3	50	500
LSU4	50	500

Table S2: Number of Amplicon Sequence Variants (ASV) and reads at different pipeline stages. **Raw**: raw reads as delivered by sequencing center (CCS reads for PacBio); **Trim**: reads with primers and demultiplexing barcodes removed; **Filter (full)**: full length amplicons with a maximum of three expected errors; **LSUx**: reads with a positive CM hit for 5.8S, allowing regions to be extracted; **Filter (ITS2)**: extracted ITS2 regions with a maximum of three expected errors; **ITS2**: ASVs obtained for the ITS2 region, and reads successfully mapped to an ITS2 ASV; **short**, **ITS**, **LSU**, **long**: ITS2-based ASVs mapped to consensus ASVs for longer regions, where ”short” and ”long” denote the full-length short and long amplicons, respectively.

	PacBio RS II				Ion Torrent Ion S5		Illumina MiSeq	
	Long		Short		Short		Short	
	ASVs	reads	ASVs	reads	ASVs	reads	ASVs	reads
Raw	–	125,034	–	49,511	–	20,717,742	–	10,756,939
Trim	–	104,305	–	41,096	–	15,208,677	–	9,513,433
Filter (full)	–	44,190	–	38,742	–	12,177,705	–	7,674,712
LSUx	–	100,210	–	40,137	–	14,946,121	–	–
Filter (ITS2)	–	87,861	–	38,753	–	13,474,669	–	–
ITS2	979	82,464	564	37,215	12,960	13,416,045	4,478	6,763,654
short	–	–	562	37,211	12,240	13,413,946	4,419	6,753,567
ITS	805	81,217	–	–	–	–	–	–
LSU	777	81,087	–	–	–	–	–	–
long	708	80,287	–	–	–	–	–	–

Table S3: Correspondences between ASVs found by different sequencing strategies. Each row shows the number of ASVs (*ASVs*) shared uniquely by one or more datasets, including the fraction of total ASVs for each dataset (*ASVs frac*), the number of reads represented by those ASVs in each dataset (*reads*) and the fraction of total reads for the dataset (*reads frac*).

		PacBio RS II						Illumina MiSeq			Ion Torrent Ion S5		
		Long			Short			Short			Short		
	ASVs	ASVs frac	reads	reads frac	ASVs frac	reads	reads frac	ASVs frac	reads	reads frac	ASVs frac	reads	reads frac
	378	0.39	17k	0.21									
	14				0.02	232	0.01						
	868							0.19	165k	0.02			
	9373										0.72	1.8M	0.14
	28	0.03	682	0.01				0.01	12k	0.00			
	41	0.04	886	0.01							0.00	3.2k	0.00
	26				0.05	420	0.01	0.01	29k	0.00			
	10				0.02	148	0.00				0.00	2.4k	0.00
	2791							0.62	896k	0.13	0.22	2.1M	0.16
	20	0.02	1.1k	0.01	0.04	1.3k	0.03	0.00	149k	0.02			
	251	0.26	9k	0.11				0.06	316k	0.05	0.02	611k	0.05
	233				0.41	5.5k	0.15	0.05	703k	0.10	0.02	1.4M	0.10
	261	0.27	54k	0.65	0.46	30k	0.80	0.06	4.5M	0.66	0.02	7.5M	0.56
Total	14294	1.00	82k	1.00	1.00	37k	1.00	1.00	6.8M	1.00	1.00	13M	1.00

Table S4: Correspondences between 97% OTUs found by different sequencing strategies. Each row shows the number of OTUs (*OTUs*) shared uniquely by one or more datasets, including the fraction of total OTUs for each dataset (*OTUs frac*), the number of reads represented by those OTUs in each dataset (*reads*) and the fraction of total reads for the dataset (*reads frac*).

		PacBio RS II						Illumina MiSeq			Ion Torrent Ion S5		
		Long			Short			Short			Short		
	OTUs	OTUs frac	reads	reads frac	OTUs frac	reads	reads frac	OTUs frac	reads	reads frac	OTUs frac	reads	reads frac
	327	0.36	16k	0.20									
	1				0.00	2	0.00						
	133							0.05	8.1k	0.00			
	4054										0.60	192k	0.01
	4	0.00	56	0.00				0.00	935	0.00			
	45	0.05	950	0.01							0.01	2.6k	0.00
	2				0.00	18	0.00	0.00	3.1k	0.00			
	4				0.01	18	0.00				0.00	712	0.00
	1866							0.67	332k	0.05	0.28	919k	0.07
	242	0.26	5.1k	0.06				0.09	124k	0.02	0.04	350k	0.03
	221				0.42	4.2k	0.11	0.08	629k	0.09	0.03	1.3M	0.10
	303	0.33	60k	0.73	0.57	33k	0.89	0.11	5.7M	0.84	0.04	11M	0.79
Total	7202	1.00	82k	1.00	1.00	37k	1.00	1.00	6.8M	1.00	1.00	13M	1.00

Table S5: Parameters for exponential distance-decay fits. Range: range at which exponential function is at 95% of its maximum value.

Guild	Amplicon	Tech	Algorithm	Metric	Space Range (m)	Time Range (a)
All Fungi	Short	Illumina	Cons	Bray-Curtis	13 (7.4– 23)	-
All Fungi	Short	Illumina	PHYLO	Bray-Curtis	13 (7.4– 23)	-
All Fungi	Short	PacBio	Cons	Bray-Curtis	25 (11– 59)	-
All Fungi	Short	PacBio	PHYLO	Bray-Curtis	25 (11– 59)	-
All Fungi	Long	PacBio	PHYLO	Bray-Curtis	18 (9– 41)	3.3 (1.9–6.9)
All Fungi	Long	PacBio	PHYLO	W-UNIFRAC	478 (0–5583)	-
ECM	Short	Illumina	Cons	Bray-Curtis	13 (6.5– 26)	-
ECM	Short	Illumina	PHYLO	Bray-Curtis	13 (6.7– 26)	-
ECM	Short	PacBio	Cons	Bray-Curtis	38 (11– 66)	-
ECM	Short	PacBio	PHYLO	Bray-Curtis	38 (11– 66)	-
ECM	Long	PacBio	PHYLO	Bray-Curtis	14 (5.2– 30)	4.1 (1.9– 18)
ECM	Long	PacBio	PHYLO	W-UNIFRAC	-	-