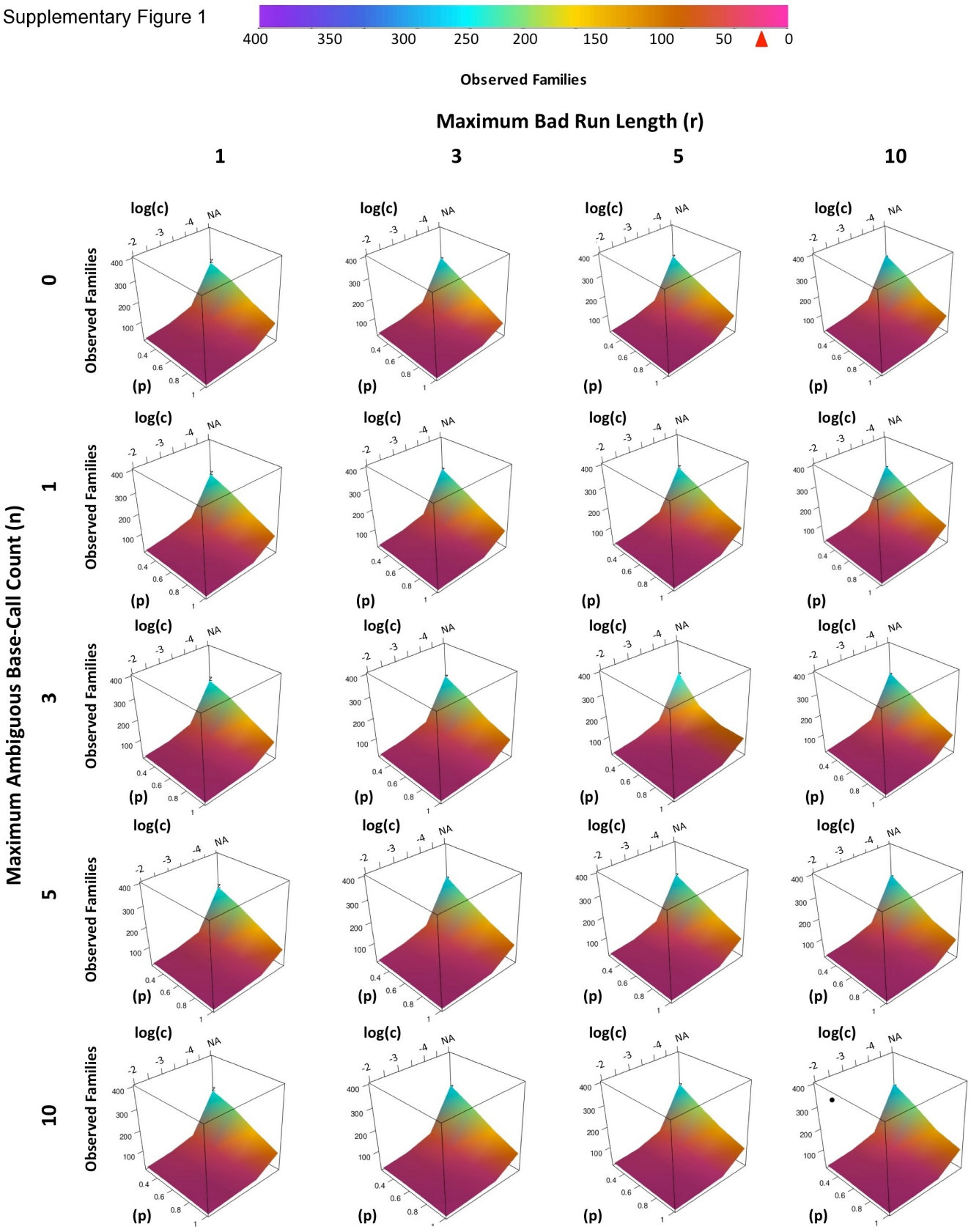
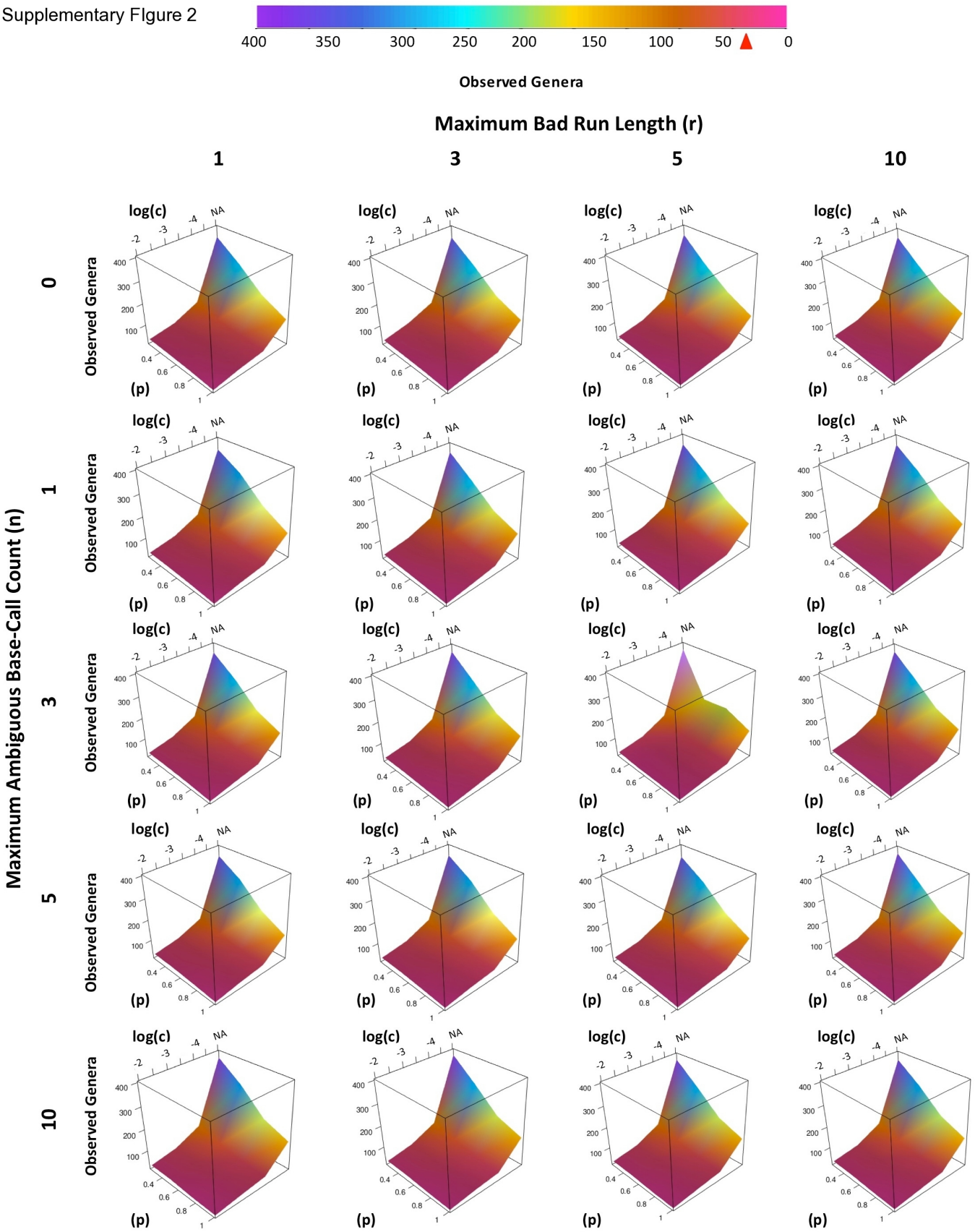
Supplementary Figure 1



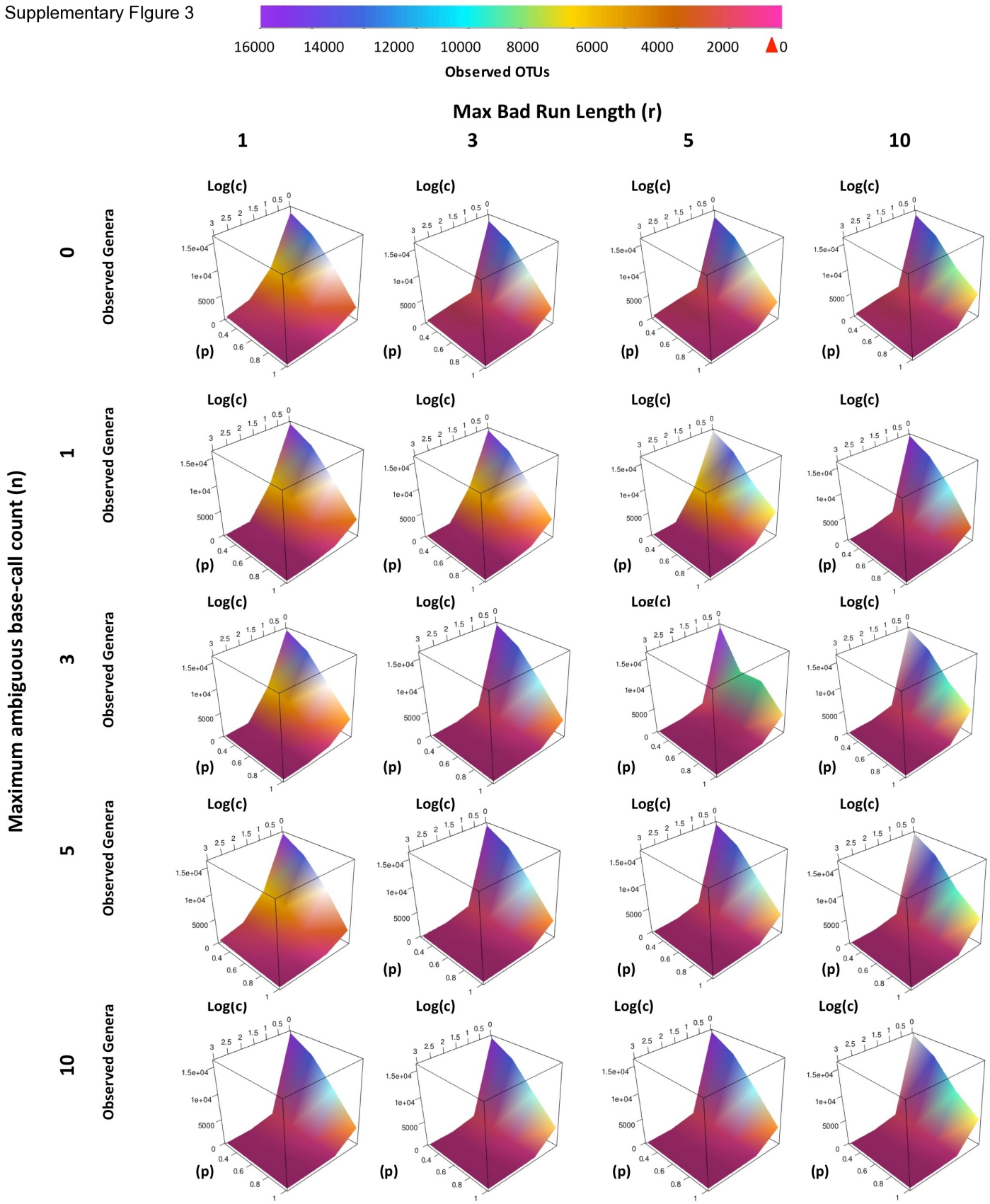
Supplementary Figure 1. α-Diversity comparison of mock community reads filtered using select quality parameter settings (dataset 1). Family-level taxon counts (z-axis) for mock communities filtered with variable split_library_fastq.py parameters, showing all cross-interactions between minimum per-read length as % of total read length (p) (x-axes), OTU abundance threshold (c) (as %; y-axes), maximum bad run length (r) (increasing values left to right), and maximum ambiguous base-call count (n) (increasing values top to bottom) at multiple OTU minimum abundance thresholds (c). Red arrow below color key indicates expected family-level taxon count.

Supplementary FIgure 2



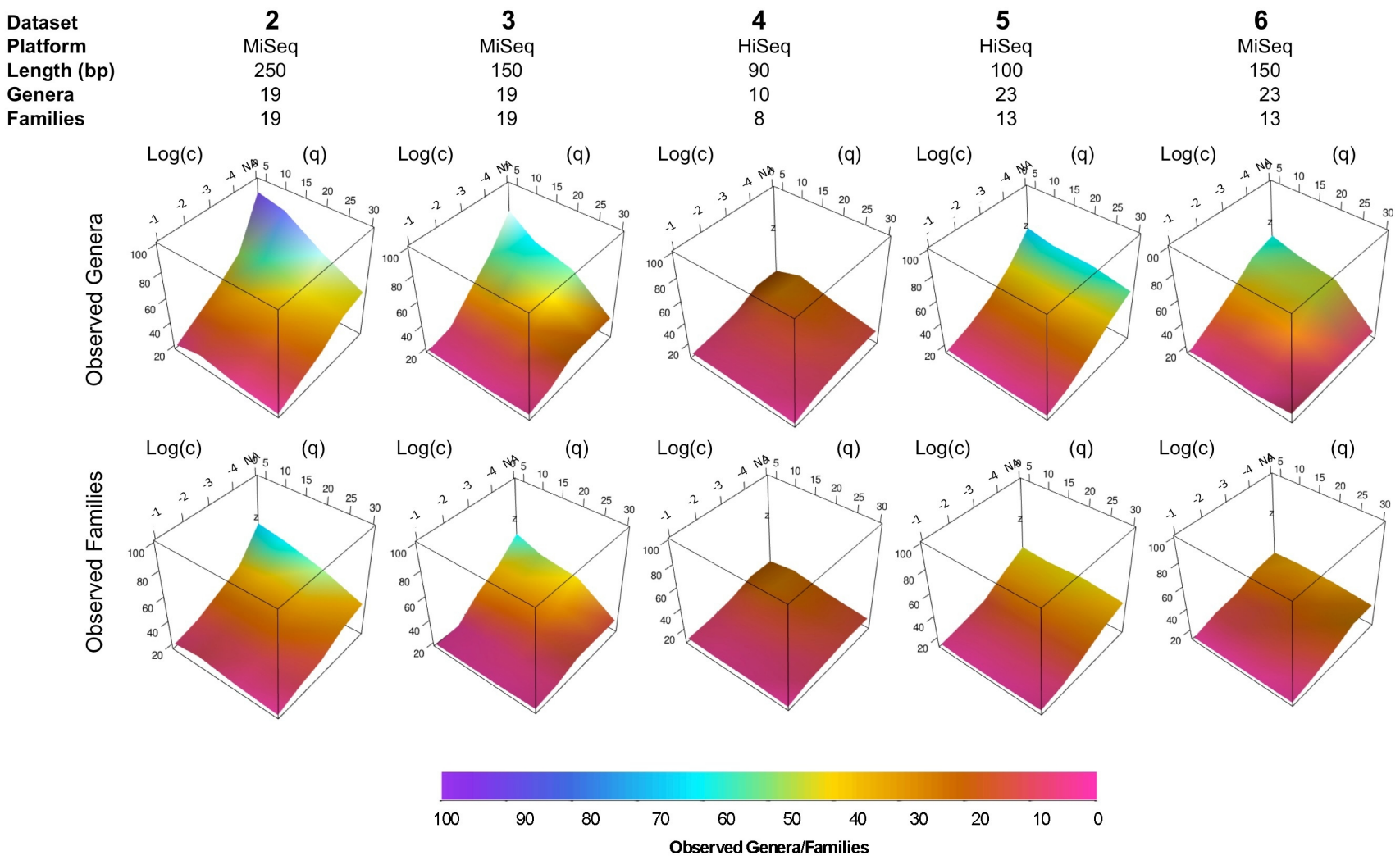
Supplementary Figure 2. α-Diversity comparison of mock community reads filtered using select quality parameter settings (dataset 1). Genus-level taxon counts (z-axis) for mock communities filtered with variable split_library_fastq.py parameters, showing all cross-interactions between minimum per-read length as % total read length (p) (x-axes), OTU abundance threshold (c) (as %; y-axes), maximum bad run length (r) (increasing values left to right), and maximum ambiguous base-call count (n) (increasing values top to bottom) at multiple OTU minimum abundance thresholds (c). Red arrow below the color key indicates expected genus-level taxon count.

Supplementary FIgure 3



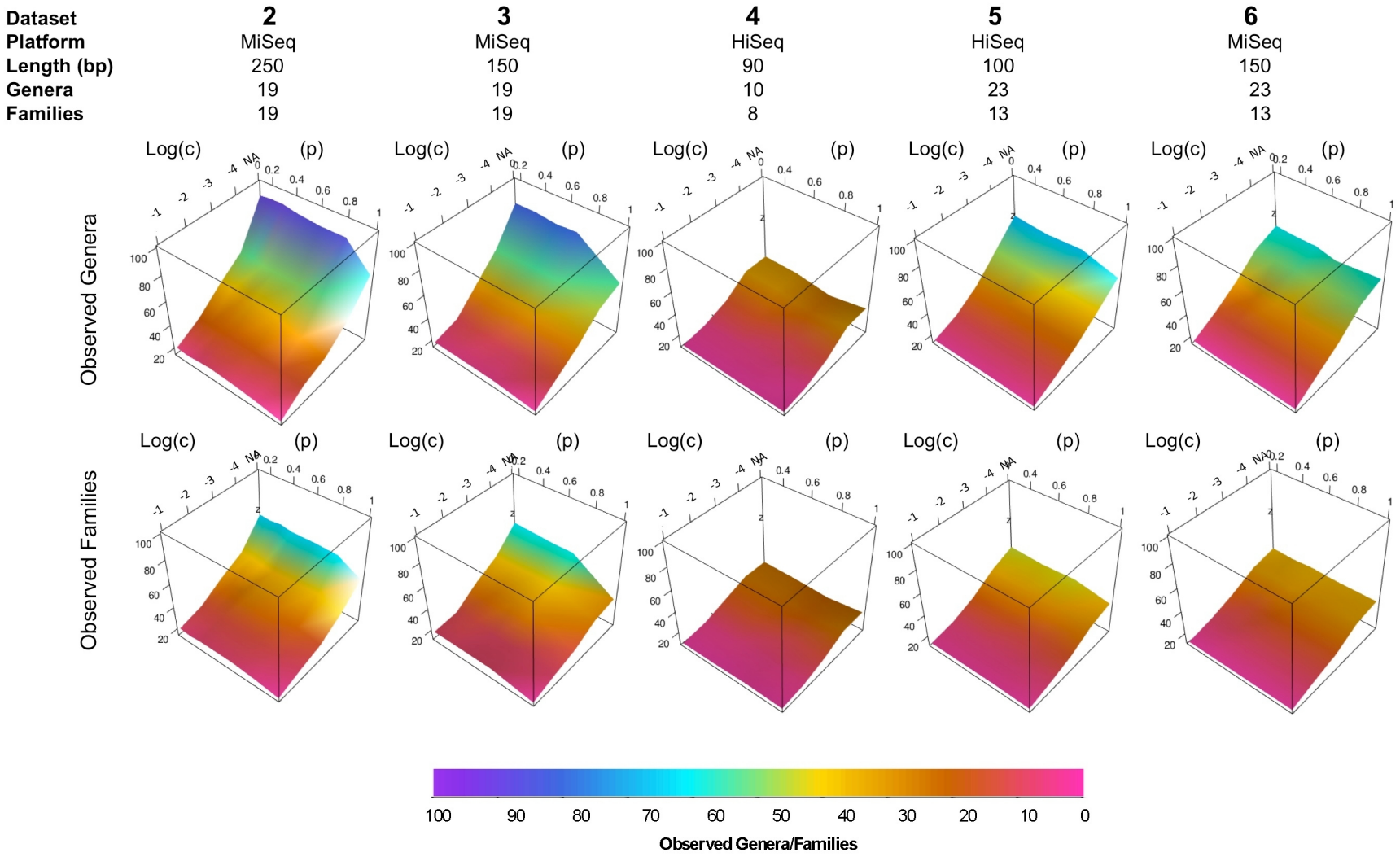
Supplementary Figure 3. α-Diversity comparison of mock community reads filtered using select quality parameter settings (dataset 1). Observed counts (z-axis) of OTUs clustered at 97% similarity for mock communities filtered with variable split_library_fastq.py parameters, showing all cross-interactions between minimum per-read length as % total read length (p) (x-axes), OTU abundance threshold (c) (as absolute values) (y-axes), maximum bad run length (r) (increasing values left to right), and maximum ambiguous base-call count (n) (increasing values top to bottom) at multiple OTU minimum abundance thresholds (c). Red arrow below the color key indicates expected species count.

Supplementary Figure 4



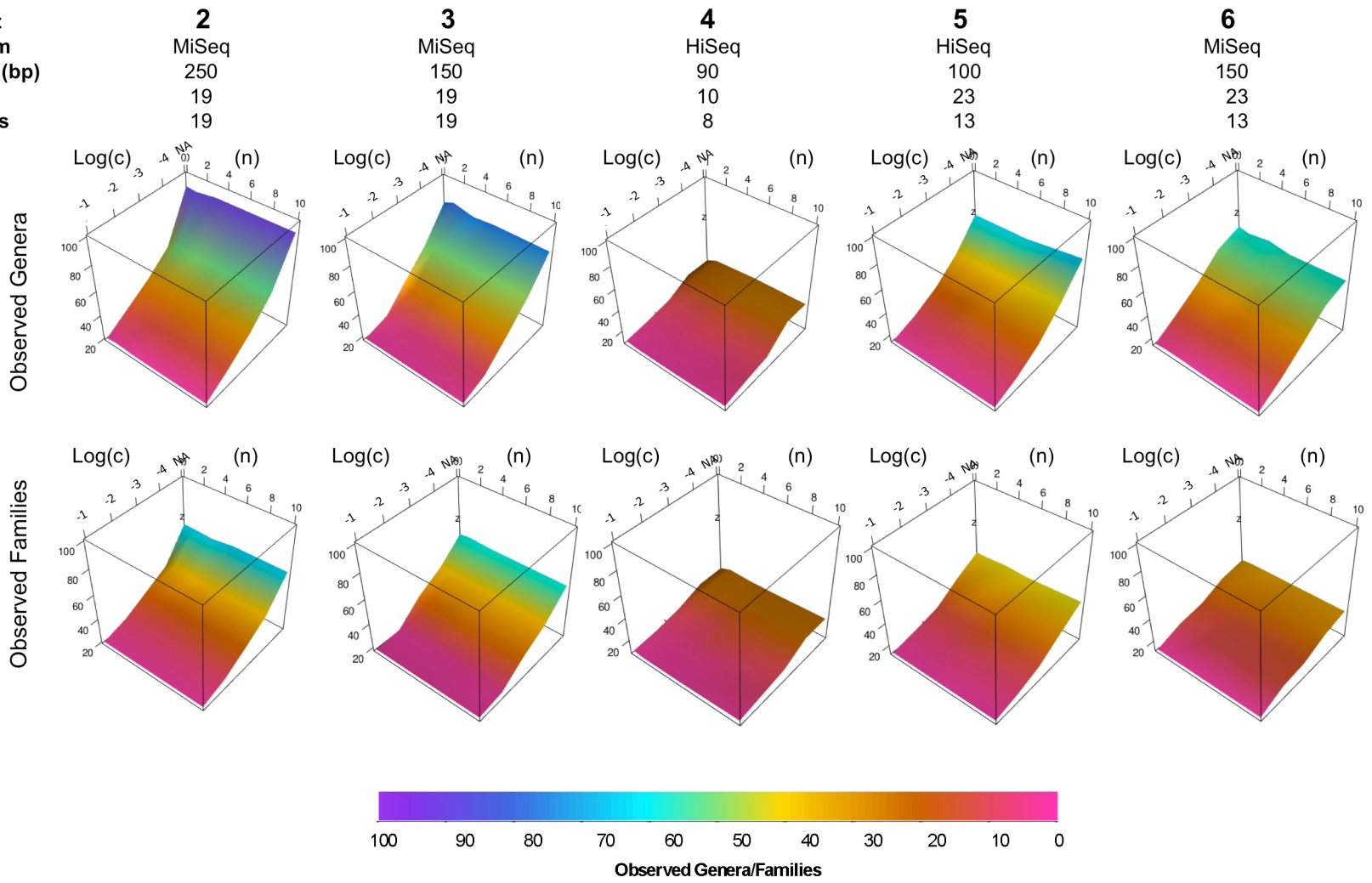
Supplementary Figure 4. Interaction of Phred score threshold (q) and OTU abundance threshold (c) (as %) on observed family and genus-level taxa (datasets 2-6). NA = no OTU abundance threshold was applied. Heading indicates platform type, read length, and expected family- and genus-level counts.

Supplementary Figure 5



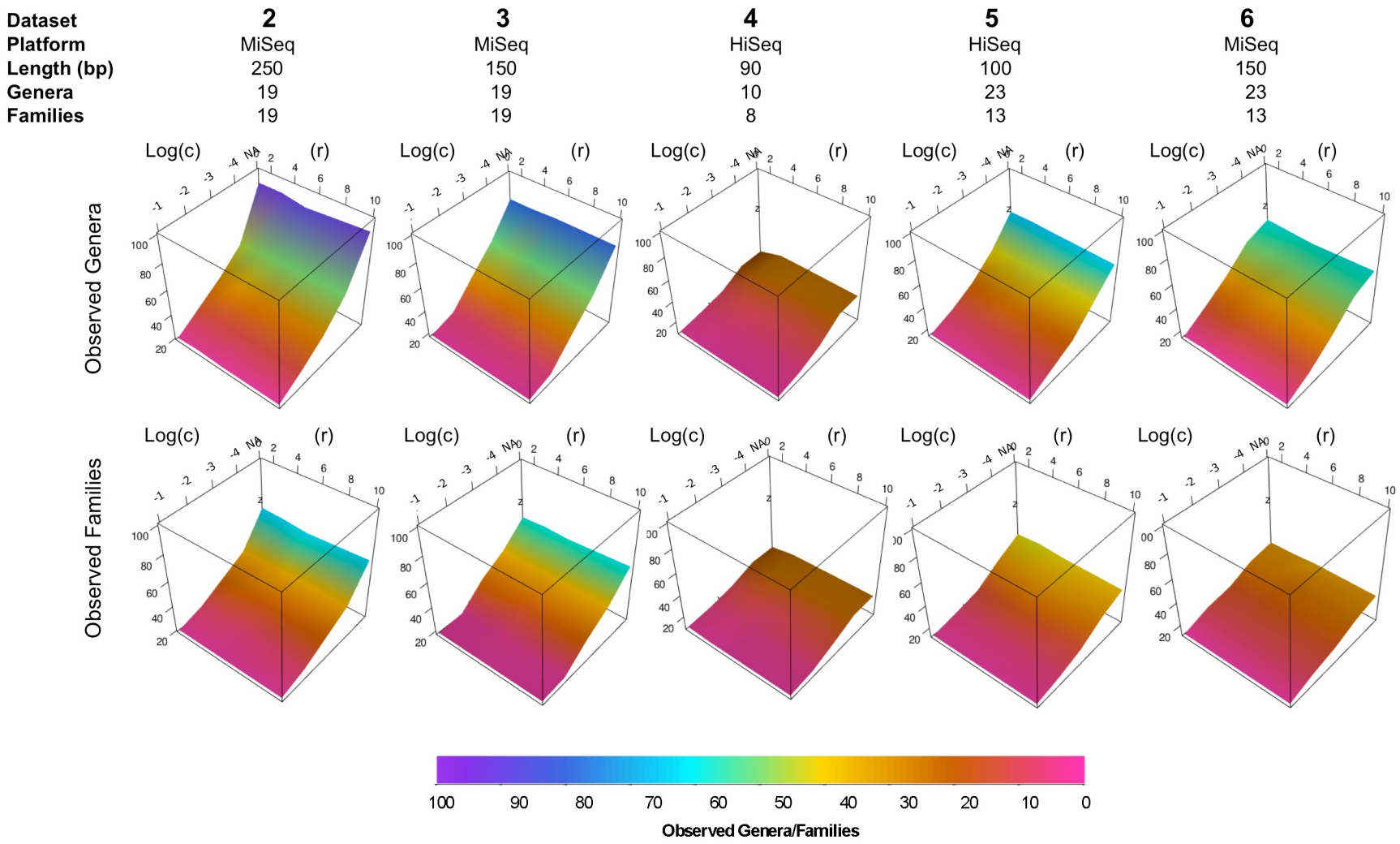
Supplementary Figure 5. Interaction of minimum per-read length (p) (as % total length) and OTU abundance threshold (c) (as %) on observed family and genus-level taxa (datasets 2-6). NA = no OTU abundance threshold applied. Heading indicates platform type, read length, and expected family- and genus-level counts.

Supplementary Figure 6

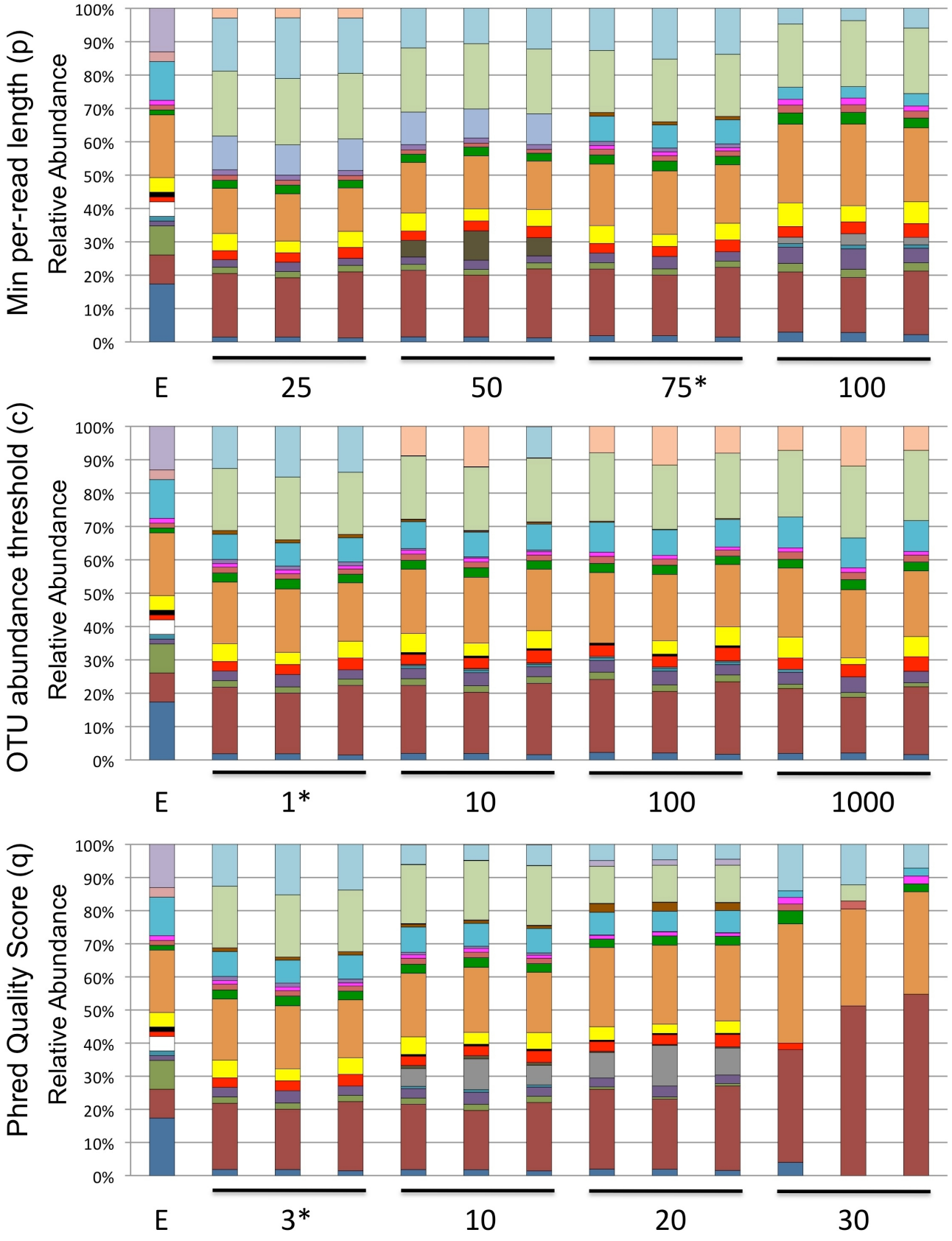


Supplementary Figure 6. Interaction of maximum ambiguous base-call count (n) and OTU abundance threshold (c) (as %) on observed family and genus-level taxa (datasets 2-6). NA = no OTU abundance threshold applied. Heading indicates platform type, read length, and expected family- and genus-level counts.

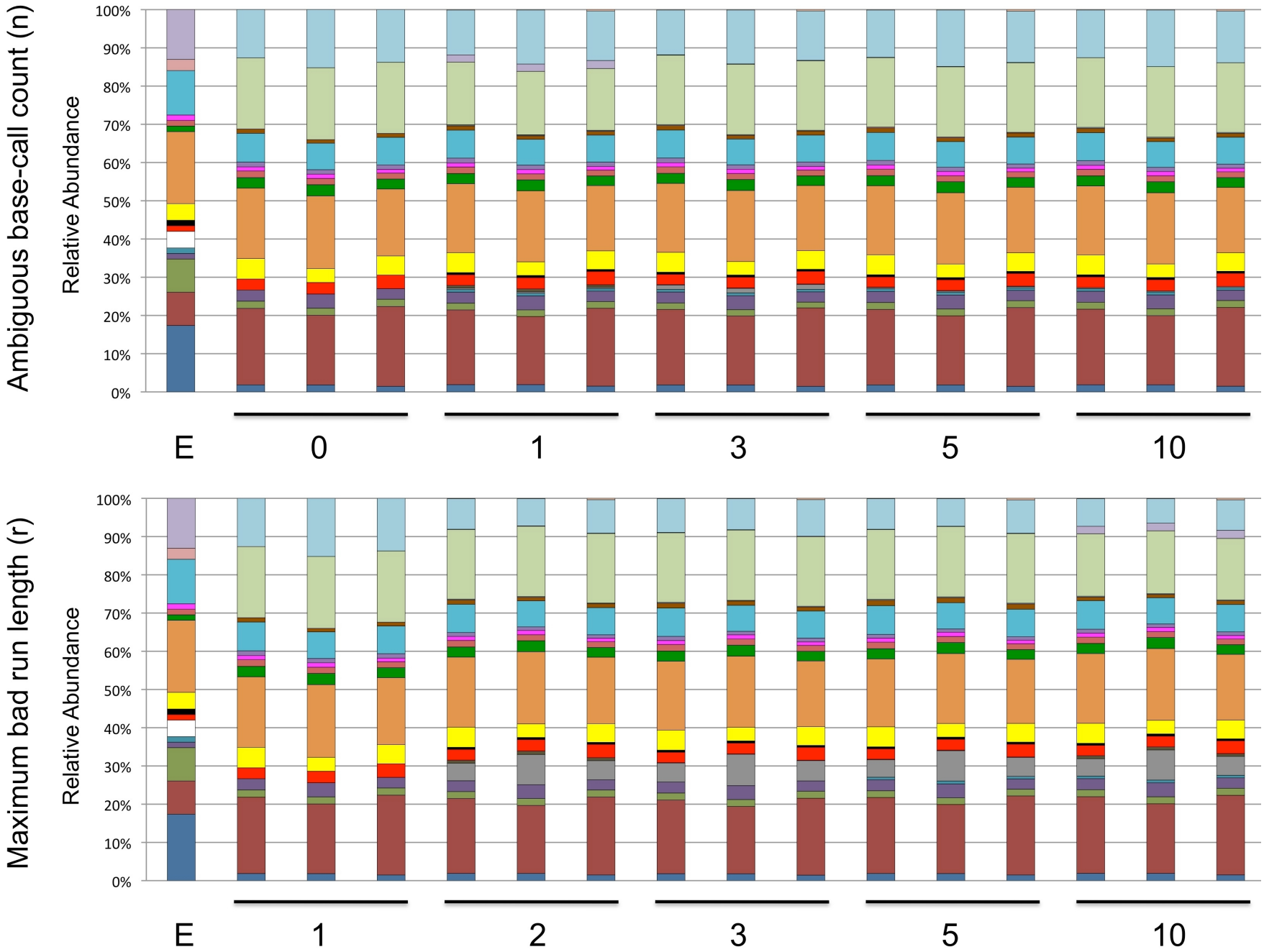
Supplementary Figure 7



Supplementary Figure 7. Interaction of maximum bad run length (r) and OTU abundance threshold (c) (as %) on observed family- and genus-level taxa (datasets 2-6). NA = no OTU abundance threshold was applied. Heading indicates type of sequencing platform used, read length, and expected family- and genus-level counts.

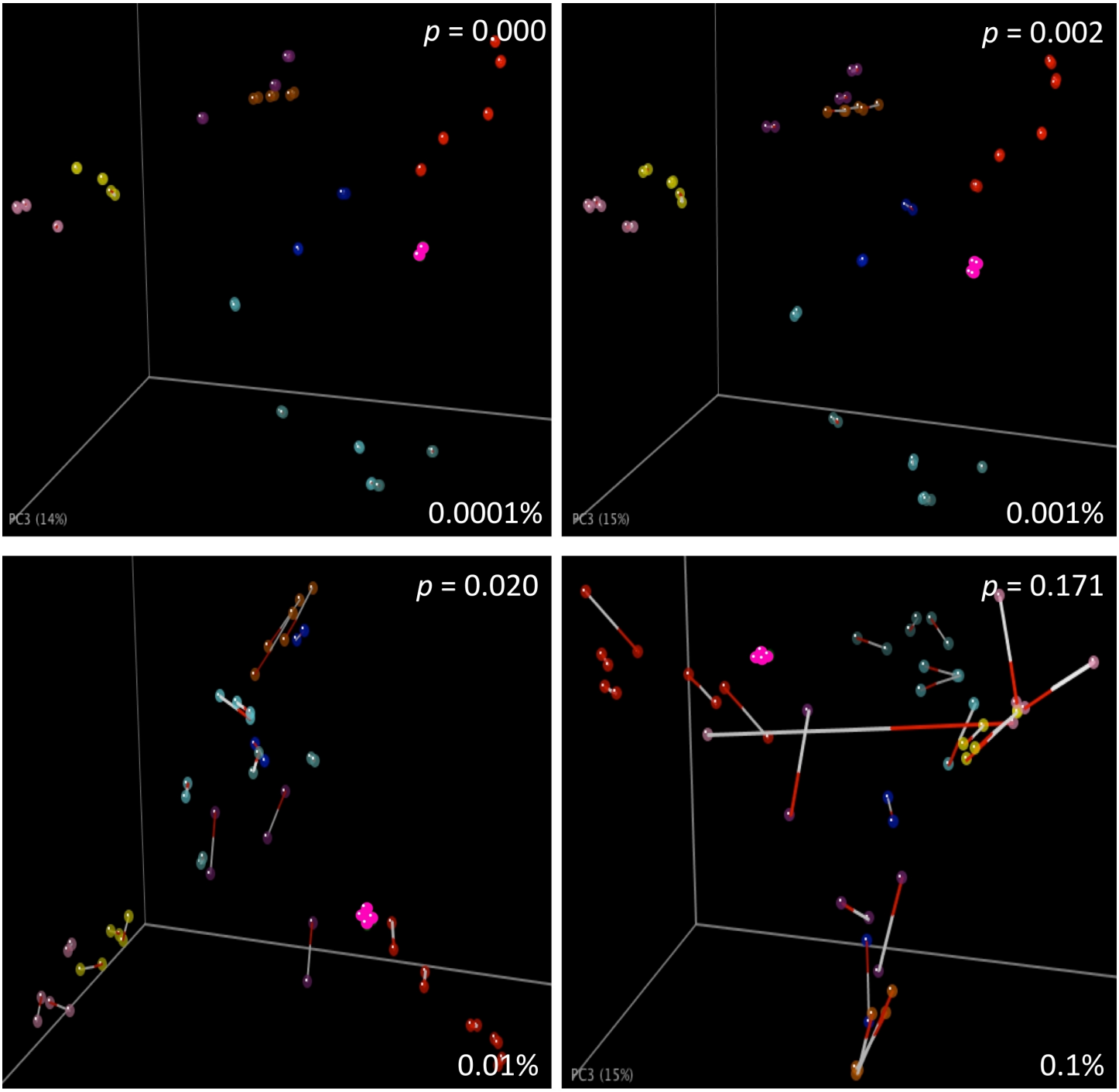


Supplementary Figure 8. Family-level taxonomic distribution of mock communities filtered using variable p (as % total length), c, and q thresholds (dataset 1). All other parameters were held constant. E, expected values. *default parameters.

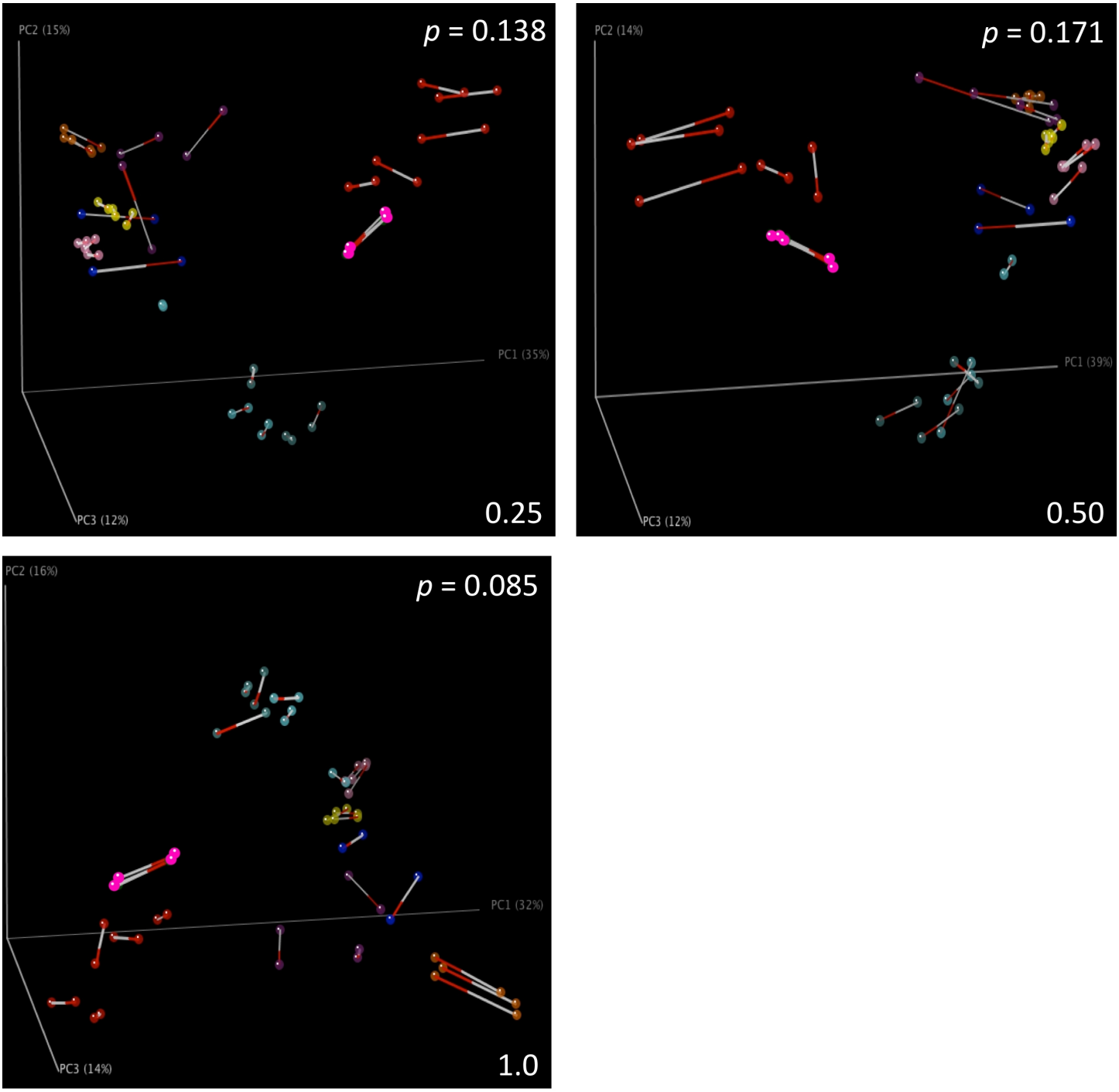


Supplementary Figure 9. Family-level taxonomic Distribution of mock communities filtered using variable n and r thresholds (dataset 1). All other parameters were held constant. E, expected values.

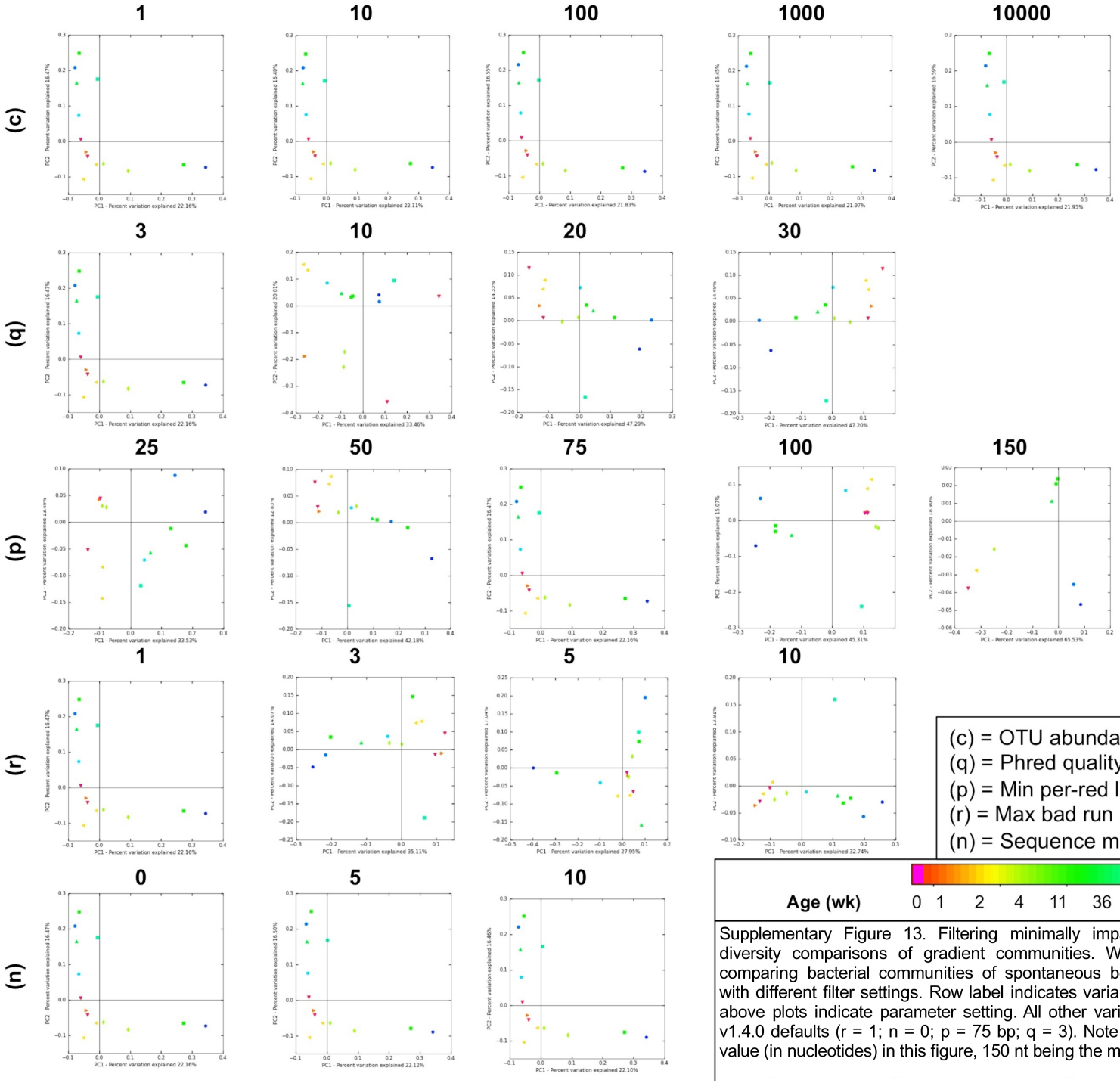
Supplementary Figure 10. Taxonomic key to Supplementary Figures 8 and 9.



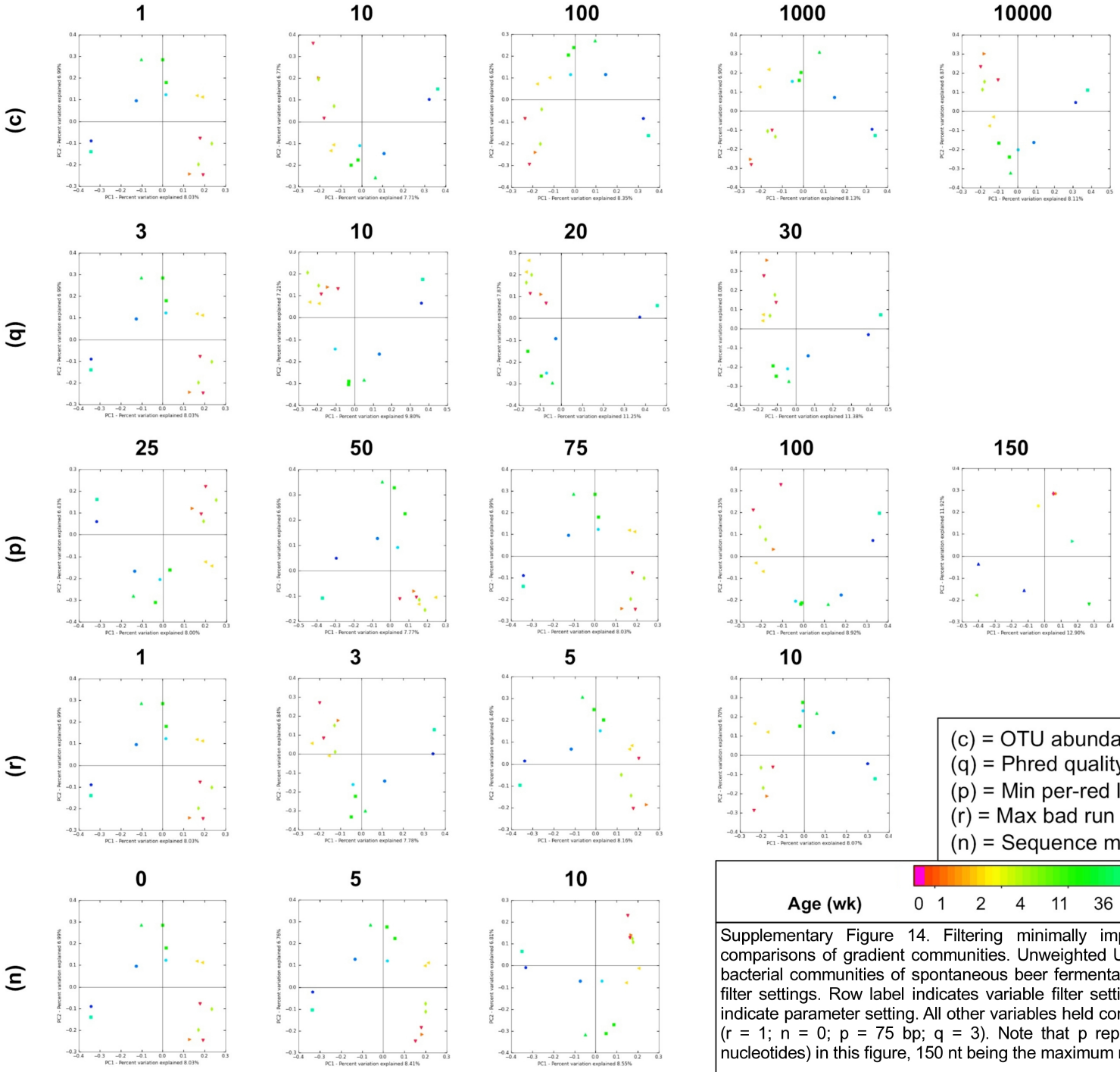
Supplementary Figure 11. Procrustes PCoA of GAIIx weighted UniFrac distance biplot comparing variation in OTU abundance threshold (c), using Illumina GAIIx data (dataset 1). Comparison of c setting listed in bottom-right corner to dataset without c filtering. Top-right corner indicates Bonferroni-corrected P-value for Procrustes goodness of fit. Red, feces; Magenta, mock community; Cyan, skin; Dark cyan, tongue; Blue, freshwater; Orange, freshwater creek; Purple, ocean; Yellow, estuary sediment; Pink, soil.



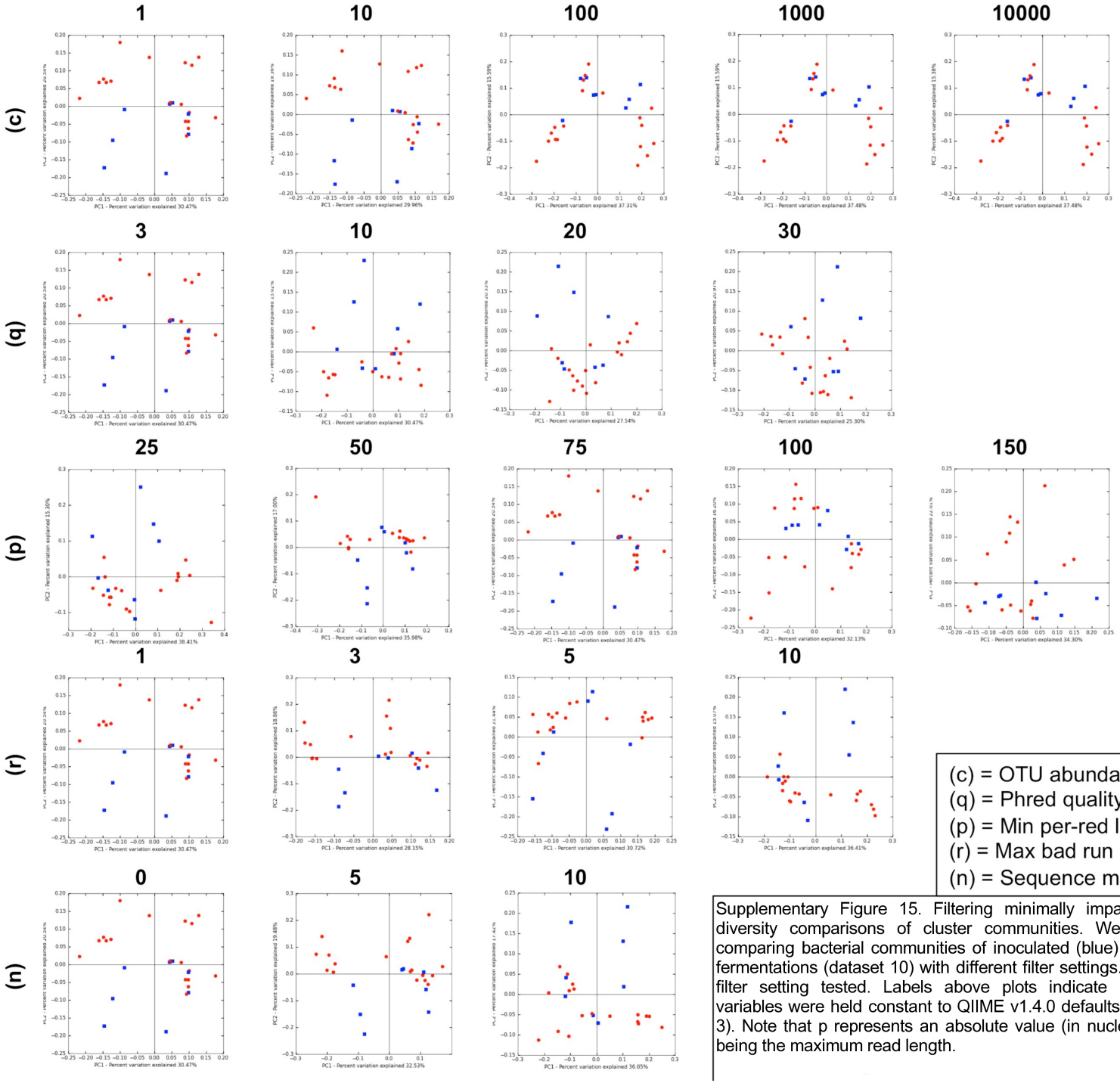
Supplementary Figure 12. Procrustes PCoA weighted UniFrac distance biplot comparing variation in minimum per-read length (p) (as % total read length), using Illumina GAIIx data (dataset 1). Comparison of setting listed in bottom-right corner to p = 0.75. Top-right corner indicates Bonferroni-corrected P-value for Procrustes goodness of fit. Red, human feces; Magenta, mock community; Cyan, human skin; Dark cyan, human tongue; Blue, freshwater; Orange, freshwater creek; Purple, ocean; Yellow, estuary sediment; Pink, soil.



Supplementary Figure 13. Filtering minimally impacts abundance-weighted β-diversity comparisons of gradient communities. Weighted UniFrac PCoA plots comparing bacterial communities of spontaneous beer fermentations (dataset 9) with different filter settings. Row label indicates variable filter setting tested. Labels above plots indicate parameter setting. All other variables held constant to QIIME v1.4.0 defaults (r = 1; n = 0; p = 75 bp; q = 3). Note that p represents an absolute value (in nucleotides) in this figure, 150 nt being the maximum read length.

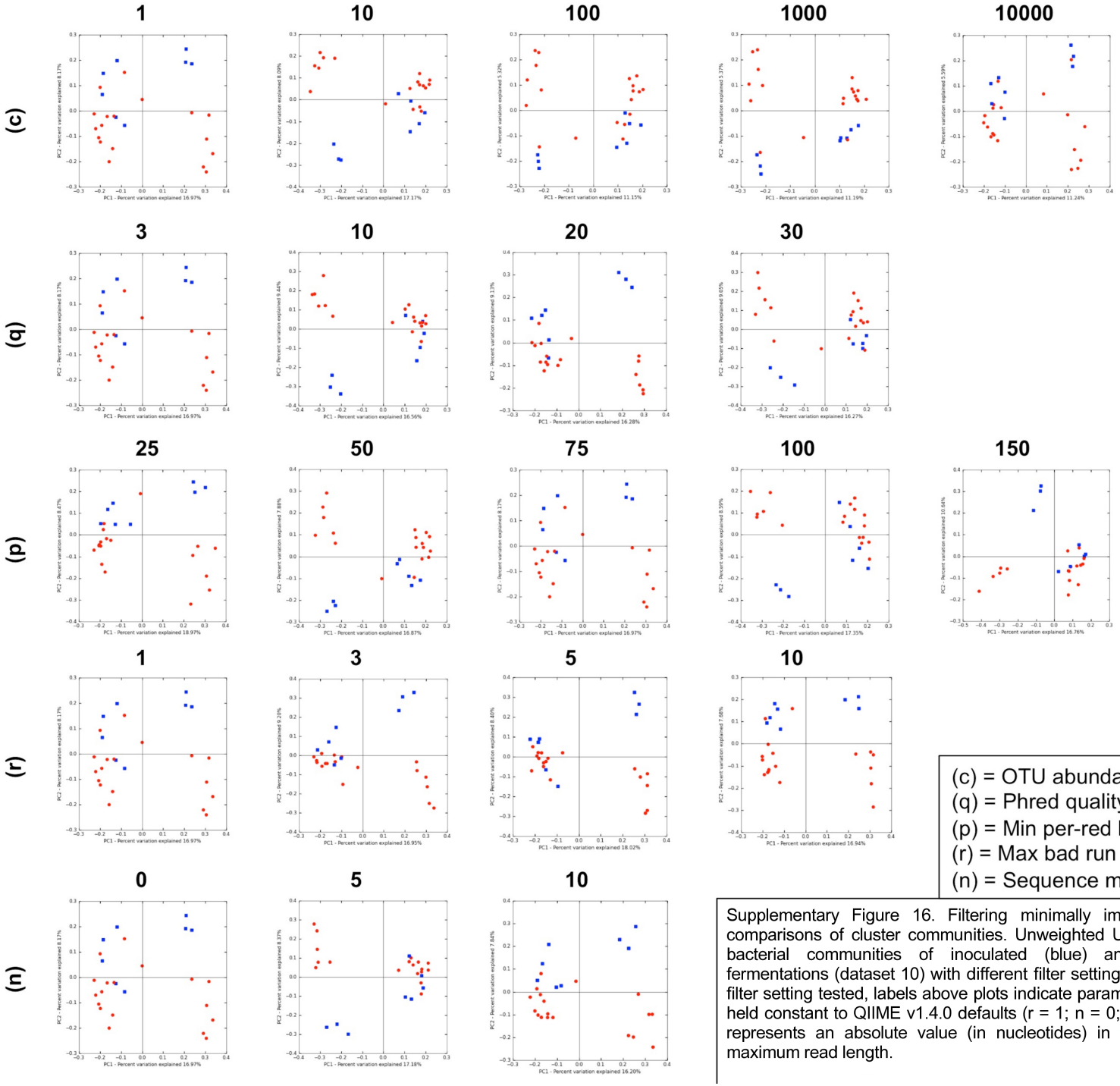


Supplementary Figure 14. Filtering minimally impacts unweighted β-diversity comparisons of gradient communities. Unweighted UniFrac PCoA plots comparing bacterial communities of spontaneous beer fermentations (dataset 9) with different filter settings. Row label indicates variable filter setting tested. Labels above plots indicate parameter setting. All other variables held constant to QIIME v1.4.0 defaults (r = 1; n = 0; p = 75 bp; q = 3). Note that p represents an absolute value (in nucleotides) in this figure, 150 nt being the maximum read length.



(c) = OTU abundance threshold
(q) = Phred quality score threshold
(p) = Min per-red length
(r) = Max bad run length
(n) = Sequence max N

Supplementary Figure 15. Filtering minimally impacts abundance-weighted β-diversity comparisons of cluster communities. Weighted UniFrac PCoA plots comparing bacterial communities of inoculated (blue) and uninoculated (red) wine fermentations (dataset 10) with different filter settings. Row label indicates variable filter setting tested. Labels above plots indicate parameter setting. All other variables were held constant to QIIME v1.4.0 defaults (r = 1; n = 0; p = 75 bp; q = 3). Note that p represents an absolute value (in nucleotides) in this figure, 150 nt being the maximum read length.



(c) = OTU abundance threshold
(q) = Phred quality score threshold
(p) = Min per-red length
(r) = Max bad run length
(n) = Sequence max N

Supplementary Figure 16. Filtering minimally impacts unweighted β-diversity comparisons of cluster communities. Unweighted UniFrac PCoA plots comparing bacterial communities of inoculated (blue) and uninoculated (red) wine fermentations (dataset 10) with different filter settings. Row label indicates variable filter setting tested, labels above plots indicate parameter setting. All other variables held constant to QIIME v1.4.0 defaults (r = 1; n = 0; p = 75 bp; q = 3). Note that p represents an absolute value (in nucleotides) in this figure, 150 nt being the maximum read length.

Supplementary Table 1. Parameter values and cross-interactions tested in this study.

phred_quality_score (q)	filter_otu_table.py min count (c)	sequence_max_n (n)	max_bad_run_length (r; default 1)																							
			1									2				3				5				10			
			min_per_read_length (p; default 75)																							
			25	50	75	90	100	110	150	200	250	25	50	75	100	25	50	75	100	25	50	75	100	25	50	75	100
default (3)	1 (default)	default (0)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		1	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		3	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		5	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		10	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	10	default (0)	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		1	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		3	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		5	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		10	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	100	default (0)	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		1	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		3	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		5	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		10	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	1000	default (0)	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		1	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		3	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		5	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		10	X	X	X		X					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
10	1 (default)	default (0)			X																						
	10				X																						
	100				X																						
	1000				X																						
	10000				X																						
20	1 (default)	default (0)			X																						
	10				X																						
	100				X																						
	1000				X																						
	10000				X																						
30	1 (default)	default (0)			X																						
	10				X																						
	100				X																						
	1000				X																						
	10000				X																						

KEY	
X	GAIIx
X	MiSeq
X	HiSeq
X	GAIIx, HiSeq, and MiSeq

Supplementary Table 2. Mock Communities Analyzed in This Study.

Dataset	Target Gene	Sequencing Platform	Read length (nt)	Mock Communities[a]	Community ID[b]	Total Families	Total Genera	Total Species	Total Strains
1	16S rRNA	GAIIx	100	3 even	A	16	29	67	67
2	16S rRNA	MiSeq	250	2 even 2 uneven	B	19	19	22	22
3	16S rRNA	MiSeq	150	2 even 2 uneven	B	19	19	22	22
4	18S rRNA	HiSeq	90	1 even	C	8	10	12	12
5	16S rRNA	HiSeq	100	1 even	D	13	23	44	48
6	16S rRNA	MiSeq	150	1 even	D	13	23	44	48

[a]Number of mock communities analyzed in this run

[b]Community ID corresponds to the four different mock communities described in Table S3-S6

Supplementary Table 3. Composition of Mock Community A.

Taxon	Even1	Even2	Even3
Actinobacteria; Actinobacteria (class); Actinobacteridae; Bifidobacteriales;Bifidobacteriaceae; Bifidobacterium;Bifidobacterium adolescentis	0.0147	0.0147	0.0147
Actinobacteria; Actinobacteria (class); Actinobacteridae; Bifidobacteriales;Bifidobacteriaceae; Bifidobacterium;Bifidobacterium angulatum	0.0147	0.0147	0.0147
Actinobacteria; Actinobacteria (class); Actinobacteridae; Bifidobacteriales;Bifidobacteriaceae; Bifidobacterium;Bifidobacterium bifidum	0.0147	0.0147	0.0147
Actinobacteria; Actinobacteria (class); Actinobacteridae; Bifidobacteriales;Bifidobacteriaceae; Bifidobacterium;Bifidobacterium breve	0.0147	0.0147	0.0147
Actinobacteria; Actinobacteria (class); Actinobacteridae; Bifidobacteriales;Bifidobacteriaceae; Bifidobacterium;Bifidobacterium catenulatum	0.0147	0.0147	0.0147
Actinobacteria; Actinobacteria (class); Actinobacteridae; Bifidobacteriales;Bifidobacteriaceae; Bifidobacterium;Bifidobacterium dentium	0.0147	0.0147	0.0147
Actinobacteria; Actinobacteria (class); Actinobacteridae; Bifidobacteriales;Bifidobacteriaceae; Bifidobacterium;Bifidobacterium gallicum	0.0147	0.0147	0.0147
Actinobacteria; Actinobacteria (class); Actinobacteridae; Bifidobacteriales;Bifidobacteriaceae; Bifidobacterium;Bifidobacterium pseudocatenulatum	0.0147	0.0147	0.0147
Actinobacteria; Actinobacteria (class); Coriobacteridae; Coriobacteriales;Coriobacterineae; Coriobacteriaceae; Collinsella;Collinsella intestinalis	0.0147	0.0147	0.0147
Actinobacteria; Actinobacteria (class); Coriobacteridae; Coriobacteriales;Coriobacterineae; Coriobacteriaceae; Collinsella;Collinsella stercoris	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides caccae	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides cellulosilyticus	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides coprocola	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides coprophilus	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides dorei	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides eggerthii	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides finegoldii	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides intestinalis	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides ovatus	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides pectinophilus	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides plebeius	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Bacteroidaceae; Bacteroides;Bacteroides stercoris	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Porphyromonadaceae; Parabacteroides;Parabacteroides johnsonii	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Porphyromonadaceae; Parabacteroides;Prevotella copri	0.0147	0.0147	0.0147
Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales;Rikenellaceae; Alistipes;Alistipes putredinis	0.0147	0.0147	0.0147
Firmicutes; Bacilli; Lactobacillales; Streptococcaceae; Streptococcus;Streptococcus infantarius	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium asparagiforme	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium bartlettii	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium celatum	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium hathewayi	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium hiranonis	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium hylemonae	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium methylpentosum	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium nexile	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium orbiscindens	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium ramosum	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium saccharolyticum-related	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium;Clostridium sporogenes	0.0147	0.0147	0.0147

Firmicutes; Clostridia; Clostridiales; Clostridiales incertae sedis; Clostridiales Family XI. Incertae Sedis; Anaerococcus;Anaerococcus hydrogenalis	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Eubacteriaceae; Anaerofustis;Anaerofustis stercorihominis	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Eubacteriaceae; Eubacterium;Eubacterium biforme	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Eubacteriaceae; Eubacterium;Eubacterium ramulus	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Butyrivibrio;Butyrivibrio crossotus	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Coprococcus;Coprococcus comes	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Dorea;Dorea formicigenerans	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Roseburia;Roseburia faecis	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Roseburia;Roseburia intestinalis	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Roseburia;Roseburia inulinivorans	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Anaerotruncus;Anaerotruncus colihominis	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Ruminococcus;Ruminococcus gnavus	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Ruminococcus;Ruminococcus hansenii	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Ruminococcus;Ruminococcus lactaris	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Ruminococcus;Ruminococcus torques	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Subdoligranulum;Subdoligranulum variabile	0.0147	0.0147	0.0147
Firmicutes; Clostridia; Clostridiales; Veillonellaceae; Mitsuokella;Mitsuokella multacida	0.0147	0.0147	0.0147
Firmicutes; Erysipelotrichi; Erysipelotrichales; Erysipelotrichaceae; Catenibacterium;Catenibacterium mitsuokai	0.0294	0.0294	0.0294
Firmicutes; Erysipelotrichi; Erysipelotrichales; Erysipelotrichaceae; Holdemania;Holdemania filiformis	0.0147	0.0147	0.0147
Firmicutes;Clostridia;Clostridiales;Peptococcaceae;Desulfitobacterium;Desulfitobacterium_hafniense	0.0147	0.0147	0.0147
Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Proteus;Proteus penneri	0.0147	0.0147	0.0147
Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae;Cedecea;Cedecea davisae	0.0147	0.0147	0.0147
Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae;Citrobacter;Citrobacter sp	0.0147	0.0147	0.0147
Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae;Edwardsiella;Edwardsiella tarda	0.0147	0.0147	0.0147
Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae;Enterobacter;Enterobacter cancerogenus	0.0147	0.0147	0.0147
Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae;Providencia;Providencia alcalifaciens	0.0147	0.0147	0.0147
Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae;Providencia;Providencia rettgeri	0.0147	0.0147	0.0147
Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae;Providencia;Providencia rustigianii	0.0147	0.0147	0.0147
Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae;Providencia;Providencia stuartii	0.0147	0.0147	0.0147

Supplementary Table 4. Composition of Mock Community B.

Taxon	Even	Staggered
Bacteria;Actinobacteria;Actinobacteria;Actinomycetales;Actinomycetaceae;Actinomyces;Actinomyces odontolyticus ATCC 17982	0.0476	0.0002
Bacteria;Actinobacteria;Actinobacteria;Actinomycetales;Propionibacteriaceae;Propionibacterium;Propionibacterium acnes DSM16379	0.0476	0.0021
Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides vulgatus ATCC 8482	0.0476	0.0002
Bacteria;Deinococcus-thermus;Deinococci;Deinococcales;Deinococcaceae;Deinococcus;Deinococcus radiodurans DSM 20539	0.0476	0.0002
Bacteria;Firmicutes;Bacilli;Bacillales;Bacillaceae;Bacillus;Bacillus cereus ATCC 10987	0.0476	0.0214
Bacteria;Firmicutes;Bacilli;Bacillales;Listeriaceae;Listeria;Listeria monocytogenes ATCC BAA-679	0.0476	0.0021
Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Staphylococcus;Staphylococcus aureus ATCC BAA-1718	0.0476	0.2143
Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Staphylococcus;Staphylococcus epidermidis ATCC 12228	0.0476	0.0214
Bacteria;Firmicutes;Bacilli;Lactobacillales;Enterococcaceae;Enterococcus;Enterococcus faecalis ATCC 47077	0.0476	0.0002
Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus;Lactobacillus gasseri DSM 20243	0.0476	0.0021
Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Streptococcus agalactiae ATCC BAA-611	0.0476	0.2143
Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Streptococcus mutans ATCC 700610	0.0476	0.0002
Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Streptococcus pneumoniae ATCC BAA-334	0.0476	0.0214
Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Clostridium;Clostridium beijerinckii ATCC 51743	0.0476	0.0214
Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Rhodobacter;Rhodobacter sphaeroides ATCC 17023	0.0476	0.0214
Bacteria;Proteobacteria;Betaproteobacteria;Neisseriales;Neisseriaceae;Neisseria;Neisseria meningitidis ATCC BAA-335	0.0476	0.0021
Bacteria;Proteobacteria;Epsilonproteobacteria;Campylobacterales;Helicobacteraceae;Helicobacter;Helicobacter pylori ATCC 700392	0.0476	0.0021
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;Escherichia coli ATCC 700926	0.0476	0.2143
Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Acinetobacter;Acinetobacter baumannii ATCC 17978	0.0476	0.0021
Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas;Pseudomonas aeruginosa ATCC 47085	0.0476	0.0214
Eukaryota;Fungi;Ascomycota;Saccharomycetes;Saccharomycetales;Incertae_sedis;Candida;Candida albicans ATCC MY-2876	0.0005	0.0002
Euryarchaeota;Methanobacteria;Methanobacteriales;Methanobacteriaceae;Methanobrevibacter;Methanobrevibacter smithii ATCC 35061	0.0476	0.2143

Supplementary Table 5. Composition of Mock Community C.

Taxon	Relative Abundance
Eukaryota;Fungi;Ascomycota;Eurotiomycetes;Onygenales;Arthrodermataceae;Trichophyton	0.0833
Eukaryota;Fungi;Ascomycota;Eurotiomycetes;Onygenales;Onygenaceae;Coccidioides;Coccidioides_immitis	0.0833
Eukaryota;Fungi;Ascomycota;Saccharomycetes;Saccharomycetales;Incertae_sedis;Candida;Candida_albicans	0.0833
Eukaryota;Fungi;Ascomycota;Saccharomycetes;Saccharomycetales;Incertae_sedis;Candida;Candida_tropicalis	0.0833
Eukaryota;Fungi;Ascomycota;Saccharomycetes;Saccharomycetales;Saccharomycetaceae;Candida;Candida_lusitaniae	0.0833
Eukaryota;Fungi;Ascomycota;Saccharomycetes;Saccharomycetales;Saccharomycetaceae;Saccharomyces;Saccharomyces_cerevisiae	0.0833
Eukaryota;Fungi;Ascomycota;Schizosaccharomycetes;Schizosaccharomycetales;Schizosaccharomycetaceae;Schizosaccharomyces;Schizosaccharomyces_pombe	0.0833
Eukaryota;Fungi;Ascomycota;Sordariomycetes;Hypocreales;Nectriaceae;Fusarium;Fusarium_oxysporum	0.0833
Eukaryota;Fungi;Basidiomycota;Tremellomycetes;Tremellales;Tremellaceae;Cryptococcus;Cryptococcus_neoformans	0.0833
Eukaryota;Fungi;Chytridiomycota;Chytridiomycetes;Rhizophydiales;Incertae_sedis;Batrachochytrium;Batrachochytrium_dendrobatidis	0.0833
Eukaryota;Fungi;Incertae_sedis;Incertae_sedis;Mucorales;Incertae_sedis;Rhizopus;Rhizopus_oryzae	0.0833
Eukaryota;Metazoa;Chordata;Craniata;Vertebrata;Euteleostomi;Mammalia;Eutheria;Euarchontoglires;Primates;Haplorrhini;Catarrhini;Hominidae;Homo;Homo sapiens	0.0833

Supplementary Table 6. Composition of Mock Community D.

Taxon	Relative Abundance
Actinobacteria;Actinobacteria;Bifidobacteriales;Bifidobacteriaceae;Bifidobacterium;bifidum;Bifidobacterium bifidum	0.0208
Actinobacteria;Actinobacteria;Bifidobacteriales;Bifidobacteriaceae;Bifidobacterium;pseudocatenulatum;Bifidobacterium pseudocatenulatum	0.0208
Actinobacteria;Actinobacteria;Coriobacteriales;Coriobacteriaceae;Collinsella;intestinalis;Collinsella intestinalis	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;cellulosilyticus;Bacteroides cellulosilyticus	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;dorei;Bacteroides dorei	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;eggerthii;Bacteroides eggerthii	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;finegoldii;Bacteroides finegoldii	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;intestinalis;Bacteroides intestinalis	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;ovatus;Bacteroides ovatus	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;thetaiotaomicron 3731;Bacteroides thetaiotaomicron 3731	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;thetaiotaomicron 7330;Bacteroides thetaiotaomicron 7330	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;thetaiotaomicron VPI-5482;Bacteroides thetaiotaomicron VPI-5482	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;uniformis;Bacteroides uniformis	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;vulgatus;Bacteroides vulgatus	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;xylanisolvens;Bacteroides xylanisolvens	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Parabacteroides;johnsonii;Parabacteroides johnsonii	0.0208
Bacteroidetes;Bacteroidia;Bacteroidales;Rikenellaceae;Alistipes;indistinctus;Alistipes indistinctus	0.0208
Firmicute;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;infantarius;Streptococcus infantarius	0.0208
Firmicute;Clostridia;Clostridiales;Clostridiaceae;Clostridium; leptum;Clostridium leptum	0.0208
Firmicute;Clostridia;Clostridiales;Clostridiaceae;Clostridium; nexile-related A2-232;Clostridium nexile-related A2-232	0.0208
Firmicute;Clostridia;Clostridiales;Clostridiaceae;Clostridium;asparagiforme;Clostridium asparagiforme	0.0208
Firmicute;Clostridia;Clostridiales;Clostridiaceae;Clostridium;hathewayi;Clostridium hathewayi	0.0208
Firmicute;Clostridia;Clostridiales;Clostridiaceae;Clostridium;nexile;Clostridium nexile	0.0208
Firmicute;Clostridia;Clostridiales;Clostridiaceae;Clostridium;saccharolyticum-related;Clostridium saccharolyticum-related	0.0208
Firmicute;Clostridia;Clostridiales;Clostridiaceae;Clostridium;sporogenes;Clostridium sporogenes	0.0208
Firmicute;Clostridia;Clostridiales;Clostridiales Family XI. Incertae Sedis;Anaerococcus;hydrogenalis;Anaerococcus hydrogenalis	0.0208
Firmicute;Clostridia;Clostridiales;Eubacteriaceae;Eubacterium;biforme;Eubacterium biforme	0.0208
Firmicute;Clostridia;Clostridiales;Eubacteriaceae;Eubacterium;eligens;Eubacterium eligens	0.0208
Firmicute;Clostridia;Clostridiales;Eubacteriaceae;Eubacterium;ventriosum;Eubacterium ventriosum	0.0208
Firmicute;Clostridia;Clostridiales;Lachnospiraceae;Blautia;hansenii;Blautia hansenii	0.0208
Firmicute;Clostridia;Clostridiales;Lachnospiraceae;Blautia;luti;Blautia luti	0.0208
Firmicute;Clostridia;Clostridiales;Lachnospiraceae;Coprococcus;comes;Coprococcus comes	0.0208
Firmicute;Clostridia;Clostridiales;Lachnospiraceae;Dorea;formicigenerans;Dorea formicigenerans	0.0208
Firmicute;Clostridia;Clostridiales;Lachnospiraceae;Dorea;longicatena;Dorea longicatena	0.0208
Firmicute;Clostridia;Clostridiales;Lachnospiraceae;Roseburia;intestinalis;Roseburia intestinalis	0.0208
Firmicute;Clostridia;Clostridiales;Ruminococcaceae;Anaerotruncus;colihominis;Anaerotruncus colihominis	0.0208
Firmicute;Clostridia;Clostridiales;Ruminococcaceae;Faecalibacterium;prausnitzii M21/2;Faecalibacterium prausnitzii M21/2	0.0208
Firmicute;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcus;gnavus;Ruminococcus gnavus	0.0208
Firmicute;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcus;lactaris;Ruminococcus lactaris	0.0208
Firmicute;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcus;torques;Ruminococcus torques	0.0208
Firmicute;Clostridia;Clostridiales;Ruminococcaceae;Subdoligranulum;variabile;Subdoligranulum variabile	0.0208
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Edwardsiella;tarda;Edwardsiella tarda	0.0208
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Enterobacter;cancerogenus;Enterobacter cancerogenus	0.0208
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;coli K12;Escherichia coli K12	0.0208
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;fergusonii;Escherichia fergusonii	0.0208
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Proteus;penneri;Proteus penneri	0.0208
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Providencia;alcalifaciens;Providencia alcalifaciens	0.0208
Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Verrucomicrobiaceae;Akkermansia;muciniphila;Akkermansia muciniphila	0.0208

Supplementary Table 7. Raw Sequence Counts and Sample Counts for Datasets Analyzed in This Study.

Dataset	Type[a]	Platform	Read Length	Samples[b]	Total[c]	Seq Count[d]
1	M/B (HAFL)	GAIIx	100	28	28	11172089
2	M	MiSeq	250	4	4	5204783
3	M	MiSeq	150	4	4	7599016
4	M	HiSeq	90	1	389	26792290
5	M	HiSeq	100	1	411	174076782
6	M	MiSeq	150	1	5	3060773
7	B (HAFL)	HiSeq	100	25	32	49757529
8	B (HAFL)	MiSeq	150	25	32	4920833
9	B (Beer)	GAIIx	150	14	240	27404059
10	B (Wine)	GAIIx	150	26	246	27836488

[a] Sample types included in run: M = mock community; B = biological sample; HAFL = host-associated and free-living communities[5,8] ; Wine = inoculated and native wine fermentation samples[10]; Beer = spontaneous American coolship ale fermentation samples[9]

[b] Number of uniquely barcoded samples analyzed from this run

[c] Total number of uniquely barcoded samples included in this sequencing run

[d] Number of raw sequences input to filtering pipeline

Supplementary Table 8. β-Diversity Comparisons of Mock Communities Between Filtration Settings.

unweighted unifrac significance test Mock Communities 1,000 Monte Carlo Iterations
p value (Bonferroni corrected)

	r1n0p75				r1n0pX			r1nXp75		rXn0p75	
	c1	**c10**	**c100**	**c1000**	**p25**	**p50**	**p100**	**n5**	**n10**	**r5**	**r10**
r1n0p75 (default)		0.605	1	1	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
r1n0p75c10	0.605		1	1	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
r1n0p75c100	1	1		1	0.055	0.385	1	1	1	0.99	1
r1n0p75c1000	1	1	1		1	1	1	1	1	1	1
r1n0p25	< 0.001	< 0.001	0.055	1		1	1	1	1	1	1
r1n0p50	< 0.001	< 0.001	0.385	1	1		1	1	1	1	1
r1n0p100	< 0.001	< 0.001	1	1	1	1		1	1	1	1
r1n5p75	< 0.001	< 0.001	1	1	1	1	1		< 0.001	1	1
r1n10p75	< 0.001	< 0.001	1	1	1	1	1	< 0.001		1	1
r5n0p75	< 0.001	< 0.001	0.99	1	1	1	1	1	1		1
r10n0p75	< 0.001	< 0.001	1	1	1	1	1	1	1	1	

weighted unifrac significance test
p value (Bonferroni corrected)

	r1n0p75				r1n0pX			r1nXp75		rXn0p75	
	c1	**c10**	**c100**	**c1000**	**p25**	**p50**	**p100**	**n5**	**n10**	**r5**	**r10**
r1n0p75 (default)		1	1	1	1	1	< 0.001	0.495	0.55	1	1
r1n0p75c10	1		1	1	1	1	0.33	0.825	0.77	1	1
r1n0p75c100	1	1		1	1	1	0.275	1	1	1	1
r1n0p75c1000	1	1	1		1	1	0.55	1	1	1	1
r1n0p25	1	1	1	1		1	1	1	1	1	1
r1n0p50	1	1	1	1	1		1	1	1	1	1
r1n0p100	< 0.001	0.33	0.275	0.55	1	1		1	1	0.605	1
r1n5p75	0.495	0.825	1	1	1	1	1		< 0.001	1	1
r1n10p75	0.55	0.77	1	1	1	1	1	< 0.001		1	1
r5n0p75	1	1	1	1	1	1	0.605	1	1		1
r10n0p75	1	1	1	1	1	1	1	1	1	1	

Supplementary Table 9. β-Diversity Comparisons of Host-associated and Free-living Communities Between Filtration Settings.

Unweighted Unifrac

Corrected Bonferroni *P* value at 1000 Monte Carlo Iterations

r1n0p25

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r1n0p50

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r1n0p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r1n0p100

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r1n0p75 c10

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r1n0p75 c100

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r1n0p75 c1000

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	0.036							
Freshwater Creek	< 0.001	0.396						
Mock	0.108	< 0.001	1					
Ocean	1	< 0.001	0.036	0.396				
Sediment Estuary	1	0.18	1	1	1			
Skin	1	< 0.001	0.072	1	1	1		
Soil	< 0.001	0.036	0.108	1	1	1	1	
Tongue	< 0.001	1	0.36	< 0.001	1	1	0.036	0.036

r1n1p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r1n5p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r1n10p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r2n0p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r5n0p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r10n0p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r1n0p75q10

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r1n0p75q20

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	< 0.001						
Mock	< 0.001	< 0.001	< 0.001					
Ocean	< 0.001	< 0.001	< 0.001	< 0.001				
Sediment Estuary	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001			
Skin	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

r1n0p75q30

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	1							
Freshwater Creek	1	1						
Mock	1	0.693	1					
Ocean	0.819	1	1	< 0.001				
Sediment Estuary	0.357	1	1	0.042	0.021			
Skin	-	-	-	-	-	-		
Soil	-	-	-	-	-	-	-	
Tongue	1	1	1	1	1	-	-	1

Weighted UniFrac

Corrected Bonferroni *P* value at 1000 Monte Carlo Iterations

r1n0p25

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	1						
Mock	1	0.036	0.108					
Ocean	< 0.001	0.9	1	< 0.001				
Sediment Estuary	< 0.001	1	1	< 0.001	< 0.001			
Skin	< 0.001	1	1	0.072	< 0.001	< 0.001		
Soil	< 0.001	0.18	0.324	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	1	1	1	1	1	1	1	0.036

r1n0p50

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	1						
Mock	1	< 0.001	0.324					
Ocean	< 0.001	1	1	< 0.001				
Sediment Estuary	< 0.001	1	1	< 0.001	0.576			
Skin	< 0.001	1	1	0.072	0.036	0.144		
Soil	< 0.001	0.288	1	< 0.001	< 0.001	0.072	< 0.001	
Tongue	0.144	1	1	1	1	1	1	1

r1n0p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	0.288	1						
Mock	1	0.864	0.828					
Ocean	< 0.001	0.54	1	< 0.001				
Sediment Estuary	< 0.001	< 0.001	1	< 0.001	< 0.001			
Skin	< 0.001	1	1	0.396	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	0.18	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	0.432	1	1	1	< 0.001	0.144	1	< 0.001

r1n0p100

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	0.756	1						
Mock	1	0.216	1					
Ocean	< 0.001	1	1	0.108				
Sediment Estuary	< 0.001	1	1	0.18	0.18			
Skin	0.036	1	1	1	0.072	0.576		
Soil	< 0.001	0.828	1	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	1	1	1	1	1	1	1	0.216

r1n0p75 c10

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	0.036							
Freshwater Creek	0.288	1						
Mock	1	1	1					
Ocean	< 0.001	1	1	< 0.001				
Sediment Estuary	< 0.001	0.18	1	< 0.001	< 0.001			
Skin	< 0.001	1	1	1	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	0.144	< 0.001	< 0.001	0.072	< 0.001	
Tongue	1	1	1	1	0.756	1	1	0.036

r1n0p75 c100

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	0.072							
Freshwater Creek	0.612	1						
Mock	1	1	1					
Ocean	< 0.001	1	1	< 0.001				
Sediment Estuary	< 0.001	0.216	1	0.036	0.036			
Skin	< 0.001	1	1	1	< 0.001	0.108		
Soil	< 0.001	< 0.001	0.36	< 0.001	< 0.001	1	< 0.001	
Tongue	1	1	1	1	1	1	1	0.252

r1n0p75 c1000

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	0.108							
Freshwater Creek	1	1						
Mock	1	1	1					
Ocean	< 0.001	1	1	0.108				
Sediment Estuary	< 0.001	1	1	0.72	1			
Skin	< 0.001	1	1	1	0.036	1		
Soil	< 0.001	< 0.001	1	< 0.001	< 0.001	1	< 0.001	
Tongue	1	1	1	1	1	1	1	0.216

r1n1p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	0.288	1						
Mock	1	0.288	1					
Ocean	< 0.001	0.036	0.288	< 0.001				
Sediment Estuary	< 0.001	0.18	1	< 0.001	< 0.001			
Skin	< 0.001	1	0.288	0.936	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	0.324	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	1	1	0.936	1	< 0.001	1	1	< 0.001

r1n5p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	0.036							
Freshwater Creek	0.468	1						
Mock	1	0.828	1					
Ocean	< 0.001	1	1	0.108				
Sediment Estuary	< 0.001	0.756	1	< 0.001	0.252			
Skin	< 0.001	0.324	0.648	0.144	< 0.001	< 0.001		
Soil	< 0.001	0.036	0.432	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	0.936	1	1	1	0.216	0.756	1	0.036

r1n10p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	0.072							
Freshwater Creek	0.36	1						
Mock	1	0.612	1					
Ocean	0.108	1	1	0.288				
Sediment Estuary	< 0.001	0.252	1	< 0.001	0.144			
Skin	< 0.001	0.072	0.792	0.036	< 0.001	< 0.001		
Soil	< 0.001	< 0.001	0.468	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	0.72	0.648	1	1	0.108	0.468	1	0.036

r2n0p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	0.036							
Freshwater Creek	< 0.001	1						
Mock	1	0.36	< 0.001					
Ocean	< 0.001	1	0.18	< 0.001				
Sediment Estuary	< 0.001	0.36	0.18	< 0.001	< 0.001			
Skin	< 0.001	1	0.108	1	< 0.001	0.036		
Soil	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	1	1	0.108	1	< 0.001	1	1	0.288

r5n0p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	1	1						
Mock	0.792	0.396	1					
Ocean	< 0.001	1	1	< 0.001				
Sediment Estuary	< 0.001	1	1	< 0.001	< 0.001			
Skin	< 0.001	1	0.792	0.648	< 0.001	0.432		
Soil	< 0.001	0.036	0.792	< 0.001	< 0.001	0.036	< 0.001	
Tongue	0.252	1	1	1	0.468	1	1	0.504

r10n0p75

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	0.252	1						
Mock	1	0.036	1					
Ocean	< 0.001	1	1	< 0.001				
Sediment Estuary	< 0.001	0.936	1	0.036	< 0.001			
Skin	< 0.001	0.864	1	1	< 0.001	1		
Soil	< 0.001	0.072	1	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	0.684	1	1	1	0.18	1	1	0.504

r1n0p75q10

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	0.036	1						
Mock	1	0.108	0.18					
Ocean	< 0.001	1	0.396	< 0.001				
Sediment Estuary	< 0.001	0.612	1	< 0.001	< 0.001			
Skin	< 0.001	1	0.18	< 0.001	< 0.001	< 0.001		
Soil	< 0.001	0.108	0.144	< 0.001	< 0.001	< 0.001	< 0.001	
Tongue	0.072	1	0.36	1	0.072	1	1	0.072

r1n0p75q20

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	0.036							
Freshwater Creek	0.18	1						
Mock	1	0.288	0.252					
Ocean	< 0.001	1	0.108	< 0.001				
Sediment Estuary	< 0.001	1	1	1	0.036			
Skin	< 0.001	1	0.18	1	< 0.001	1		
Soil	< 0.001	0.288	0.18	< 0.001	< 0.001	0.504	< 0.001	
Tongue	0.468	0.648	0.648	1	0.036	1	1	0.684

r1n0p75q30

	Feces	Freshwater	Creek	Mock	Ocean	Sediment	Skin	Soil
Freshwater	< 0.001							
Freshwater Creek	< 0.001	0.231						
Mock	0.189	0.336	0.063					
Ocean	< 0.001	1	0.252	< 0.001				
Sediment Estuary	< 0.001	1	1	< 0.001	0.084			
Skin	-	-	-	-	-	-		
Soil	-	-	-	-	-	-	-	
Tongue	0.147	1	1	1	1	-	-	1

Evaluation 1: Quality Filtration Impacts α–Diversity of Mock Communities

To determine how different filtration parameters affect estimates of microbial community composition, we measured within-sample (α) diversity of mock communities as the number of family-level and genus-level taxa detected using each filter set. Since the input community is known in this case, we can test whether a filtration strategy faithfully reconstructs the complete microbial community while detecting as few false OTUs as possible. Initially, we tested the cross-interaction of all filtering settings (Table S1) with a single GAIIx run (read length 100 nt) of three identical mock communities consisting of 67 bacterial strains mixed at equal ratios (dataset 1, Table S2, Table S3)[5, 12]. We found that α-diversity was modulated most strongly by parameters (p), (q), and OTU abundance threshold (c), while (r) and (n) had negligible impact (Figure S1-S3). High (q) settings yielded excessive filtration of sequences, resulting in a lower-than-expected family-level taxon count (16 bacterial families represented in mock samples) prior to (c) filtration (Figure 1). All primary filtration parameters except for high (q) values required secondary filtration with (c) to reach expected taxon counts, but the (c) threshold necessary was dependent upon the stringency of the initial filtration parameters. For example, a (c) value of ~0.0005% (of total filtered sequences) was adequate for elevated (p) and (q) values to reach expected taxon counts, while a threshold of ≥ 0.001% was required for lower-stringency settings. This is lower than the previously recommended[5] (c) value of ≥ 0.01% of all sequences to reach expected OTU counts using default filtration parameters in QIIME, a setting originally calibrated with this same mock community (dataset 1, Table S2). This disparity is due to the fact that the previous study was based on OTU count instead of taxon count, where multiple OTUs clustered by 97% similarity (the OTU classification set in both of these studies) may assign to the same genus-level taxon, whether this arises from minor sequencing error, PCR error, intra-genomic *rrn* copy number heterogeneity, or genotypic differences in a mixed, wild-type population. This disparity may be observed by comparison of Figure S2 (genus-level count) with Figure S3 (observed 97% OTUs); at settings retrieving the expected genus-level count, observed OTUs are still several-fold greater than expected genus (n = 29) or even species (n = 67) counts, as multiple OTUs are classified to the same genus, even after clustering at 97% similarity. Additionally, contaminating sequences from PCR reagents or growth media used to cultivate mock community members may contribute to the inherent error in any mock community comparison, and may explain a certain amount of variation between expected and observed α-diversity counts in these data.

Overall, the expected genus-level taxon counts could only be obtained with more stringent (p) and (c) settings (p ≥ 0.75, c ≥ 0.005%) than those required for obtaining expected family-level counts. This is because truncated reads in each run did not reliably reach genus level, and most of the OTUs could only be assigned to higher-level taxonomies,

even if these OTUs represented a minority population. Taking these higher-level assignments into account, however, settings yielding as many as 75 observed "genera" (Figure S3) retrieved all expected genus-level and higher-level taxa, so the same settings necessary for accurate family-level assignment ($p \geq 0.75$, $c \geq 0.005\%$) can be used to observe genus-level distribution.

Next, we tested the effect of all filtering settings (without primary filter cross-interaction) on five additional mock communities sequenced using the Illumina HiSeq, MiSeq, and GAIIx (datasets 2-6; Table S2), to confirm our original findings and determine whether platform-dependent differences and run-to-run variation impact filtering. These tests confirmed that (q) and (p) are important modulators of α-diversity on all platforms, but the influence of these settings varied in strength on a run-to-run basis, independent of read length and sequencing platform (Figure S4-S5). These tests also confirmed that (r) and (n) settings are negligible in defining taxon counts (Figures S6-S7). Overall, (c) was the single most important variable needed to reach expected taxon counts, though it varied between 0.01% to 0.0001% depending upon (q) and (p) settings. Even at constant primary filter settings, the necessary (c) setting was inconsistent between runs in a platform- and length-independent manner. Using default primary settings ($p = 0.75$, $n = 0$, $r = 1$, $q = 3$), requisite (c) fluctuated in the range of 0.005% to 0.0005%.

Given the unpredictability of run-to-run variations, giving rise to variable (c) filtering requirements, we recommend including defined mock communities in all short-amplicon sequencing runs. This is particularly pertinent for comparing samples between two or more runs, and will be a necessary component to all massive, multi-run comparisons. Even on a single-run basis, including a standard mock community will enable sensitive calibration of filtering parameters, identifying the most appropriate settings for a given sequencing run. This will avoid both losing valuable data by adhering to a conservative, arbitrarily assigned (c) setting, and admitting spurious OTUs arising from PCR and sequencing error.

Another factor that may complicate OTU threshold filtering (c) is the degree of similarity of sample types included in a single sequencing run. All filtering parameters in this study were tested against single mock communities on an individual basis. By applying an OTU abundance threshold (c) *globally* to all samples in a single sequencing run (or subset), one makes the assumption that all samples within that set contain homogeneous populations at comparable abundances. This is rarely the case, and applying a single (c) setting globally may result in the loss of rare taxa observed within a single sample or treatment group. If rare taxa are an important component of a given study or if highly divergent sample types are being compared in a single sample set, samples should be (c)-filtered *locally*. In other words, (c) should be calibrated to a single mock community sample

and that setting applied to each sample individually, such that the OTU threshold is applied as a percentage of all sequences in a given sample instead of all sequences within a single dataset. This condition could not be fully evaluated using the data in this study, but is a relevant concern in some scenarios, and should be a consideration in the experimental design of sequencing studies.

Evaluation 2: Effect of Filtration Parameters on Taxonomic Distribution

In most biological systems, we are not only concerned with how many OTUs are detected, but also the relative abundance of specific taxa comprising different communities. Thus, we were interested in whether different filtration parameters exerted an uneven impact on taxonomic distribution, favoring the detection of specific clades over others. The impact of filtration parameters on taxonomic distribution of mock communities (dataset 1) was consistent with α–diversity estimates (Figure S8-S10). The parameters (n) and (r) display little impact, but increasing these values (especially (r)) appears to yield deeper taxonomic resolution of certain taxa that were poorly differentiated in the mock communities (e.g., *Enterobacteriaceae* from *Gammaproteobacteria*), leading to closer representation of the expected distribution. However, they also increased the abundance of some taxonomically shallow OTUs (e.g., "other *Clostridiales*"). The parameters (p) and (q) had the greatest impact, with increasing (p) values leading to decreased abundance of unassigned sequences and sequences given shallow taxonomic assignment. Increasing (q) up to 20 only marginally impacted taxonomic distribution of mock communities, leading to lower abundance of unassigned sequences and greater detection of minor taxa (e.g., *Veillonellaceae*) but also of "other *Clostridiales*" and "other *Bifidobacteriales*". Increasing (q) further resulted in unacceptably high sequence filtration (97.7% of raw sequences filtered at q = 30), leading to poor replication of expected taxonomic distribution in mock communities.

Evaluation 3: Filtration Does Not Significantly Impact Within-Sample β–Diversity Comparisons

As biological observations are often drawn from between-sample (β) diversity comparisons of complex microbial communities, we were interested in how different filtration parameters impacted phylogenetic distance between the same samples. To determine whether filtration altered assessments of β-diversity between settings, we

calculated unweighted/weighted UniFrac distance[11] between mock community reads (dataset 1) filtered with different parameter settings at even sampling depth (Table S8). The only treatment resulting in a significant weighted UniFrac difference from the default setting was (p) = 1.0 (Bonferroni-corrected $p < 0.001$), indicating abundance-weighted phylogenetic differences between these settings. Variation in all parameters other than (c) resulted in significant unweighted UniFrac differences from the default settings (Bonferroni-corrected $p < 0.001$), signifying the existence of phylogenetic divergence between these settings and the default. These data suggest that phylogenetic diversity is impacted by most filtration parameters, but only among low-abundance OTUs.

Evaluation 4: Extreme Filtration Settings Degrade Between-Sample β-Diversity Comparisons

Next, we analyzed β-diversity differences between sample types within filtered sequence sets from mock, free-living, and host-associated communities (Table S9) using UniFrac[11] distance to determine how filtration impacted between-sample phylogenetic comparisons. All sample types were significantly different for all settings by unweighted UniFrac, except for (c) = 1000 and (q) = 30. As all, or most, of these communities should exhibit some phylogenetic difference (without abundance weighting), this indicates that high OTU filtration (c) and Phred quality score (q) thresholds resulted in excessive depletions of diversity, leading to the observation of complete phylogenetic parity where little should exist (e.g., between freshwater and feces). Weighted UniFrac revealed a smaller number of significant differences, primarily separating host-associated and free-living sample types for most settings, though expected patterns of significance varied with the settings applied. The most notable exceptions were greatly increased (q) and (c) settings, which led to loss of significance between sample types that should contain significant differences. As with unweighted UniFrac, this suggests that loss of diversity through extreme read filtration led to unacceptable phylogenetic convergence of these sample types. Interestingly, decreased stringency (e.g., p = 0.25) led to loss of significance between only a few sample types, and even then only between host-associated or free-living communities, retaining the expected biological result. Additionally, this demonstrates that conclusions from phylogeny-based distance metrics are robust to sequence quality, such that the inclusion of a high abundance of truncated and/or low-quality reads does not significantly impact sample clustering.

To visualize β-diversity differences between communities analyzed with the GAIIx, we constructed principle coordinate analysis (PCoA) plots based on weighted UniFrac distance and used Procrustes analysis to compare PCoA plots between filtration settings. In a

comparison of host-associated and free-living communities described by Caporaso et al[5], all settings yielded the expected pattern, separating free-living and host-associated samples, but with different degrees of similarity to the default filtration settings. Increasing (c) values resulted in increasingly divergent PCoA plots (Figure S11), and gradual loss of significance at (c) > 0.01% (Bonferroni-corrected $p_{c=1000} = 0.020$; Bonferroni-corrected $p_{c=10000} = 0.171$). All other comparisons (Figure 2; Figure S12; data not shown for (r) and (n)) yielded highly significant M^2 values ($p < 0.001$) before correction, but insignificant Bonferroni-corrected M^2 values ($p > 0.05$), suggesting that adjusting these filtration parameters alters beta-diversity assessments made by PCoA. As PCoA is commonly used to visualize β-diversity differences between microbial communities, inappropriate filtration parameters may result in misleading assessments of diversity, depending on the scenario. In this case, the same conclusions could be drawn from PCoA of most settings (host-associated and free-living community separation) and significant differences among PCoAs are not as important as this conserved biological pattern. With this goal in mind, pattern recognition was most altered by high values of (c) and (q), which blurred the boundary between host-associated and free-living communities.

Most experimental assessments of microbial communities are not concerned with differentiating highly divergent samples, but rather rely on sensitive separation of subtly different sample types to arrive at valuable conclusions. Therefore, the impact of filtering parameters on weighted and unweighted UniFrac β-diversity was also tested on closely related microbial communities whose diversity patterns exhibit either cluster or gradient effects. These communities consisted of spontaneous beer fermentations[10] and modern (inoculated with *Saccharomyces cerevisiae*) versus traditional (un-inoculated) wine fermentations[9], respectively. These spontaneous beer fermentation communities should exhibit age-dependent, gradient-type UniFrac clustering, following the transition from *Enterobacteriaceae*-rich to *Lactobacillaceae*-dominated communities during the fermentation[10]. Following the observations of our previous study[9], the traditional versus modern wine fermentations should exhibit a subtle UniFrac clustering pattern, as only a small number of bacterial taxa explain the differences between these two groups, relating to the suppression of specific bacteria by yeast inoculation. Even extreme filtering settings did not alter the expected age-dependent community shifts of spontaneous beer fermentations as assessed by weighted UniFrac PCoA (Figure S13) and unweighted UniFrac PCoA (Figure S14). Unweighted UniFrac was highly resilient to filtering setting changes, but a slight impact was seen in the weighted UniFrac data under certain conditions. Specifically, the expected community shift was blurred under the settings (q = 10) and (r = 5), but conserved by all higher and lower settings for these same parameters.

The un-inoculated and inoculated wine communities exhibit even fewer differences, which are related to minor bacterial OTUs as opposed to bulk population succession, and weakly cluster apart from each other. The delicate cluster pattern differentiating these highly similar groups with weighted UniFrac was degraded under extreme filtering settings

(Figure S15). Gradual dissolution of clusters can be observed with increasing (q) settings., (r ≥ 5), and (n = 10), but weighted UniFrac clustering is universally weak, due to a high degree of similarity between these communities. Unweighted UniFrac (Figure S16) more clearly differentiates these communities, as minor taxa are the primary OTUs differentiating the un-inoculated and inoculated wine fermentations. Under different filtering settings, unweighted UniFrac is unaffected, demonstrating the same cluster pattern throughout (Figure S16). Taken together, these data indicate that phylogeny-based diversity metrics are highly resilient to changes in filtering settings, and will yield similar observations, except in the case of extreme settings.

Evaluation 5: Filtration Settings Exhibit Similar Impact Across Illumina Platforms

Finally, we wanted to confirm that quality filtration guidelines developed on the Illumina GAIIx would be equally valid for the HiSeq2000 and MiSeq systems, ensuring that these parameters would be robust to changes in sequencing chemistry as these platforms continue to evolve. To do this, we calculated weighted UniFrac distance between "free-living" and host-associated communities sequenced on these three systems (datasets 1, 7, 8; Table S7). Procrustes weighted UniFrac PCoA biplots revealed the same trends on the HiSeq as on the GAIIx: heavily decreased (p) and increased (q) are the only factors that significantly impacted phylogenetic distance and degraded the expected separation of samples (data not shown). The MiSeq revealed the same trends, but significant differences between community types were produced even with extreme values, except for (c) > 0.1% (data not shown), indicating that these reads were more resistant to quality filtering-induced changes in β-diversity; this is because the greater read length achieved on the MiSeq (150 nt) dampened the impact of shorter, lower-quality sequences admitted by, for example, decreased (p) settings, and significance (Bonferroni-corrected $p < 0.05$) is lost by truncating the reads to 90 nt prior to OTU picking (the mean read length of HiSeq reads post-filtering) (Figure S19). This observation suggests that increased sequence length allows even greater flexibility in quality filtration. Nevertheless, the fact that all three systems reconstruct the same known biological result supports the creation of a baseline default setting that will be appropriate for all platforms, and users may adjust downstream according to sequence length and quality. We note that non-Bonferroni-corrected p-values are still significant even at these extreme settings on all platforms, and the visual separation of free-living from host-associated communities in PCoA plots is still retained; this suggests that even with these extreme parameter settings, it is unlikely that a researcher would derive different biological conclusions from the results.

Interestingly, even with extreme parameter settings on the MiSeq, the overall split between host-associated and free-living communities remained statistically significant

(data not shown), indicating that MiSeq data is more robust to quality-filtering parameters than the HiSeq 2000 and GAIIx. This result was due to the increased length of the MiSeq reads (150 bp). When the MiSeq reads were truncated to 90 bp prior to OTU picking (the median length of the truncated HiSeq2000 reads) extreme filtering settings (p = 0.25; q = 20) also resulted in a loss of statistical significance for the separation between host-associated and free-living communities (data not shown), indicating that filtering behavior is consistent across all three platforms.

Default recommendations for QIIME parameters

Our findings have informed new default parameter settings in QIIME's `split_libraries_fastq.py` and a recommended parameter for the optional `filter_otus_from_otu_table.py` script. The updated parameter defaults for split_libraries_fastq.py are (r = 3; p = 0.75; q = 3; n = 0), and the recommend parameter setting for `filter_otus_from_otu_table.py` is (`--min_count` of 0.005%). Note that the `--min_count` setting can also be passed as the `--min_otu_size` parameter to `pick_subsampled_reference_otus_through_otu_table.py`. Since `filter_otus_from_otu_table.py` is an optional script, users must call `filter_otus_from_otu_table.py` with this parameter setting to incorporate this filter.