Supplementary Figures

Schematic of OTU and DADA2 approaches towards amplicon sequencing errors.

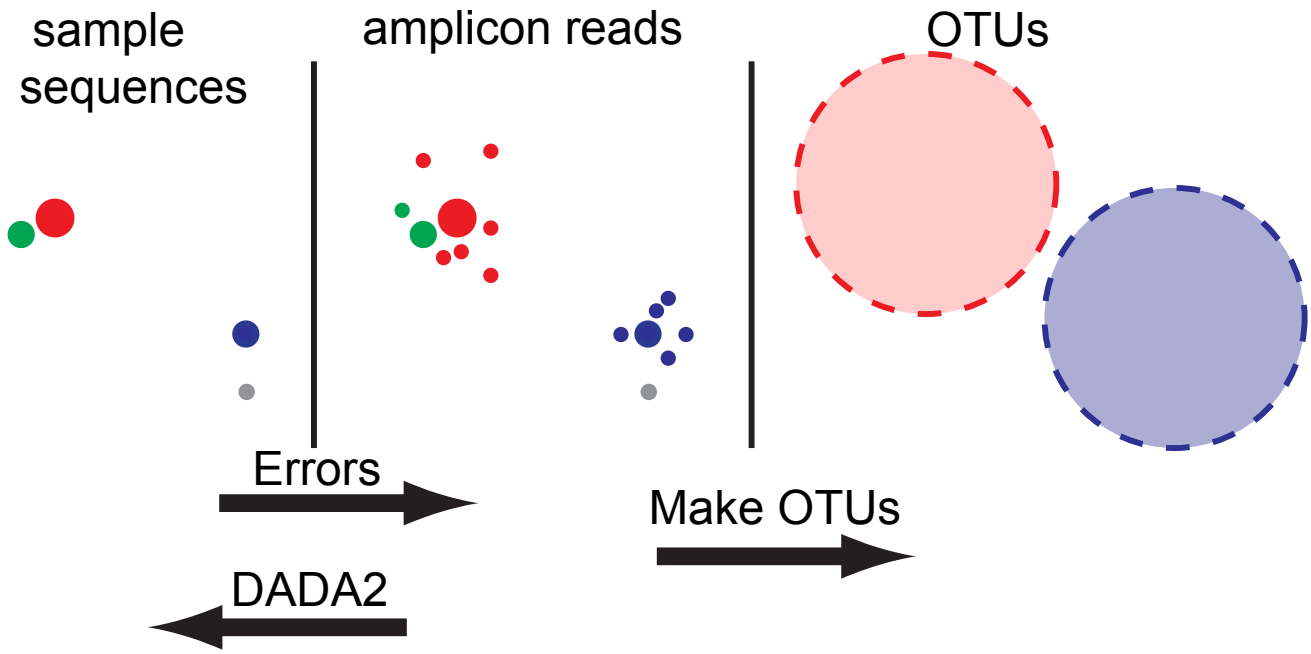


Figure 1. Circles represent identical sets of sequencing reads with size scaled by abundance and color corresponding to the true error-free sequence (there are four distinct sequences in the sample: red, green, blue and grey). Errors are introduced by amplicon sequencing from the left to the middle part of the diagram. OTU methods guard against false positive inferences by lumping similar sequences together. DADA2 uses a statistical model of amplicon errors to infer the underlying sample sequences directly, and thus tries to denoise the data from the middle to the left.

The output sequences inferred from the Balanced forward reads for UPARSE, DADA2, MED, mothur (average-linkage) and QIIME (uclust).

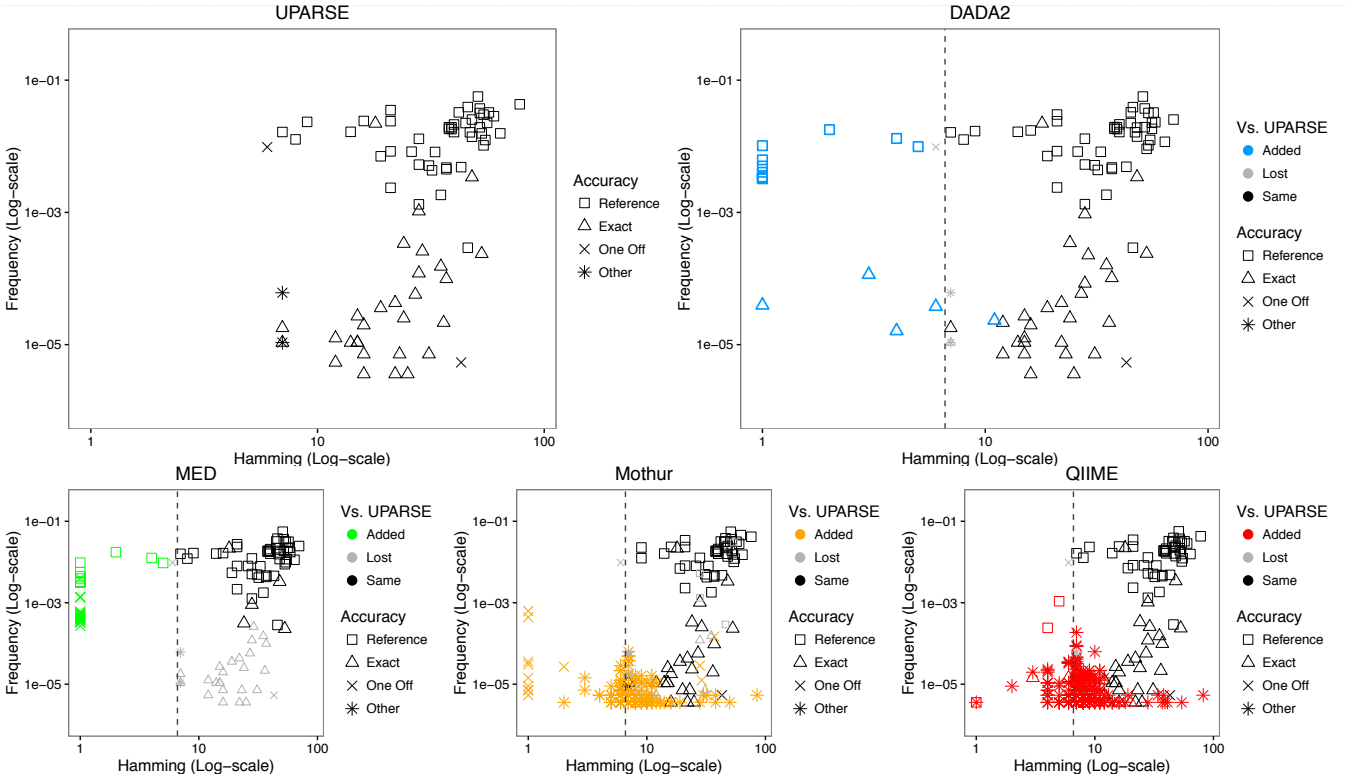


Figure 2. Frequency of output sequences from the Balanced forward dataset is plotted on the y-axis. Hamming distance from each sequence to its nearest more-abundant neighbor is plotted on the x-axis. UPARSE is used as a baseline to which the outputs of the other methods are compared. Algorithms largely concur (black) in identifying sequences that are abundant and very different from other sample sequences. However, DADA2 detects additional variation (blue) relative to UPARSE, especially within UPARSE's OTU radius (dashed line). MED also detects some fine-scale variation (green), but at the cost of a significant number of false positives, typically One Offs that are 1-away from a more abundant correct sequence, and MED does not detect low abundance sequences (grey). Mothur (orange) and QIIME (red) both report a large number of additional spurious sequences, although most are relatively low frequency.

The output sequences inferred from the HMP forward reads by UPARSE, DADA2, MED, mothur (average-linkage) and QIIME (uclust).

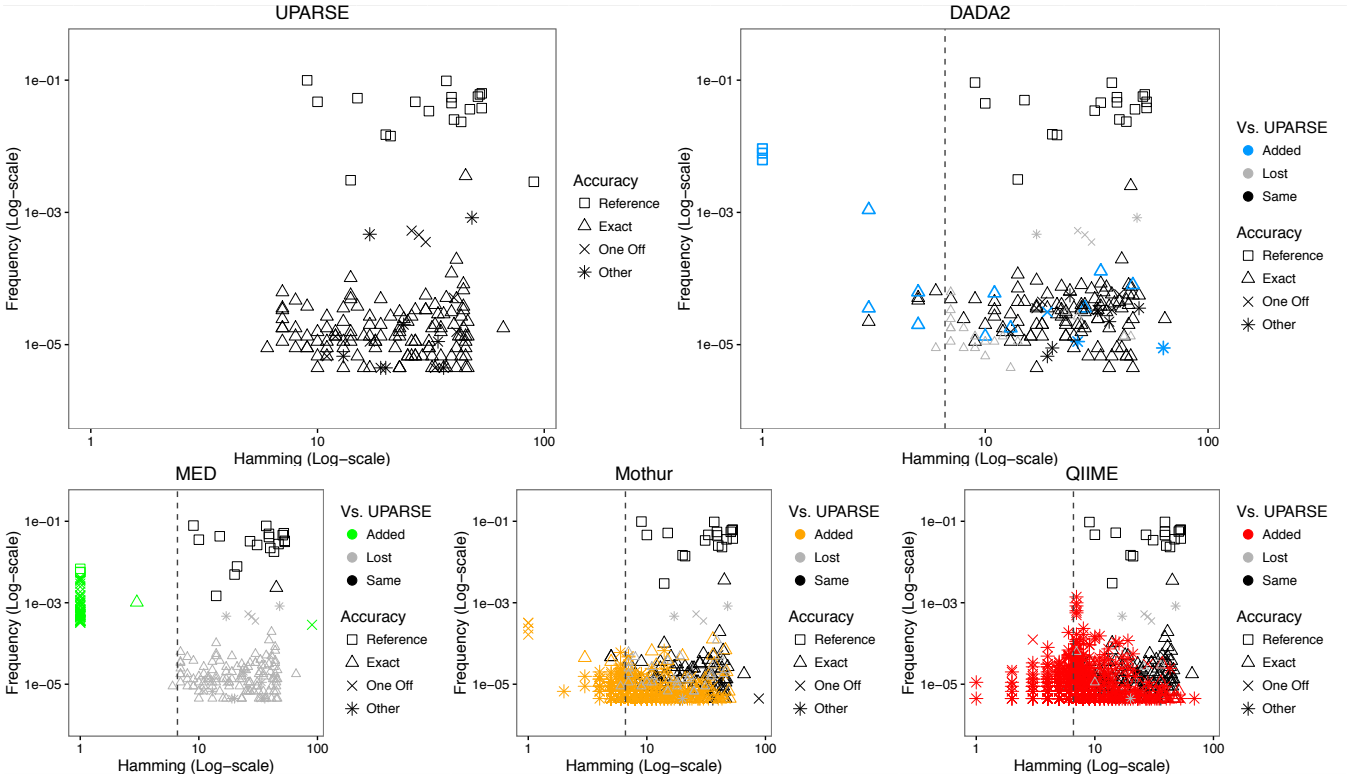


Figure 3. Frequency of output sequences from the HMP forward dataset is plotted on the y-axis. Hamming distance from each sequence to its nearest more-abundant neighbor is plotted on the x-axis. UPARSE is used as a baseline to which the outputs of the other methods are compared. Algorithms largely concur (black) in identifying sequences that are abundant and very different from other sample sequences. However, DADA2 detects additional variation (blue) relative to UPARSE, especially within UPARSE's OTU radius (dashed line). MED also detects some fine-scale variation (green), but at the cost of a significant number of false positives, typically One Offs that are 1-away from a more abundant correct sequence, and MED does not detect low abundance sequences (grey). Mothur (orange) and QIIME (red) both report a large number of additional spurious sequences, although most are relatively low frequency.

The output sequences inferred from the Extreme forward reads by UPARSE, DADA2, MED, mothur (average-linkage) and QIIME (uclust).

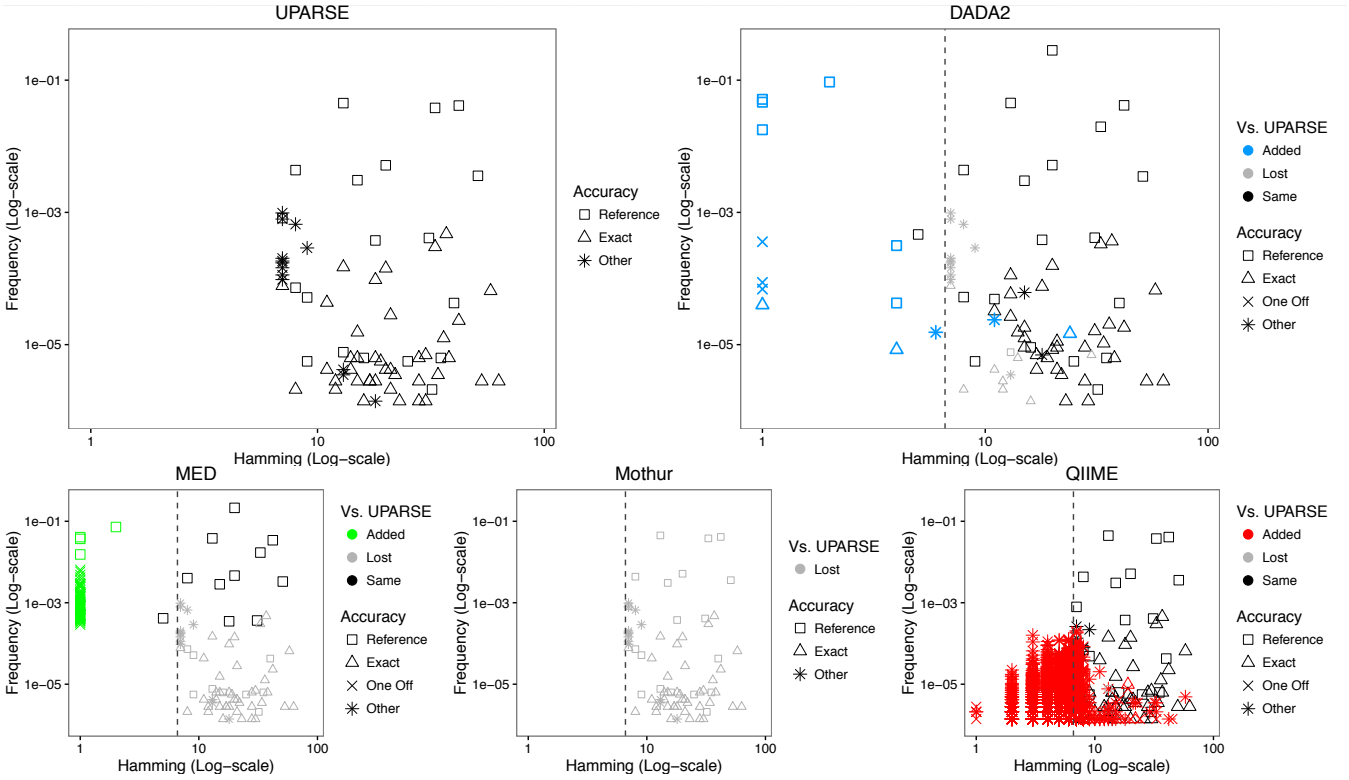


Figure 4. Frequency of output sequences from the Extreme forward dataset is plotted on the y-axis. Hamming distance from each sequence to its nearest more-abundant neighbor is plotted on the x-axis. UPARSE is used as a baseline to which the outputs of the other methods are compared. Algorithms largely concur (black) in identifying sequences that are abundant and very different from other sample sequences. However, DADA2 detects additional variation (blue) relative to UPARSE, especially within UPARSE's OTU radius (dashed line). MED also detects some fine-scale variation (green), but at the cost of a significant number of false positives, typically One Offs that are 1-away from a more abundant correct sequence, and MED does not detect low abundance sequences (grey). QIIME (red) reports a large number of additional spurious sequences, although most are relatively low frequency. Mothur failed to complete on this dataset due to the size of its calculated distance matrix.

The output sequences inferred from the Balanced merged reads by UPARSE, DADA2, MED, mothur (average-linkage) and QIIME (uclust).

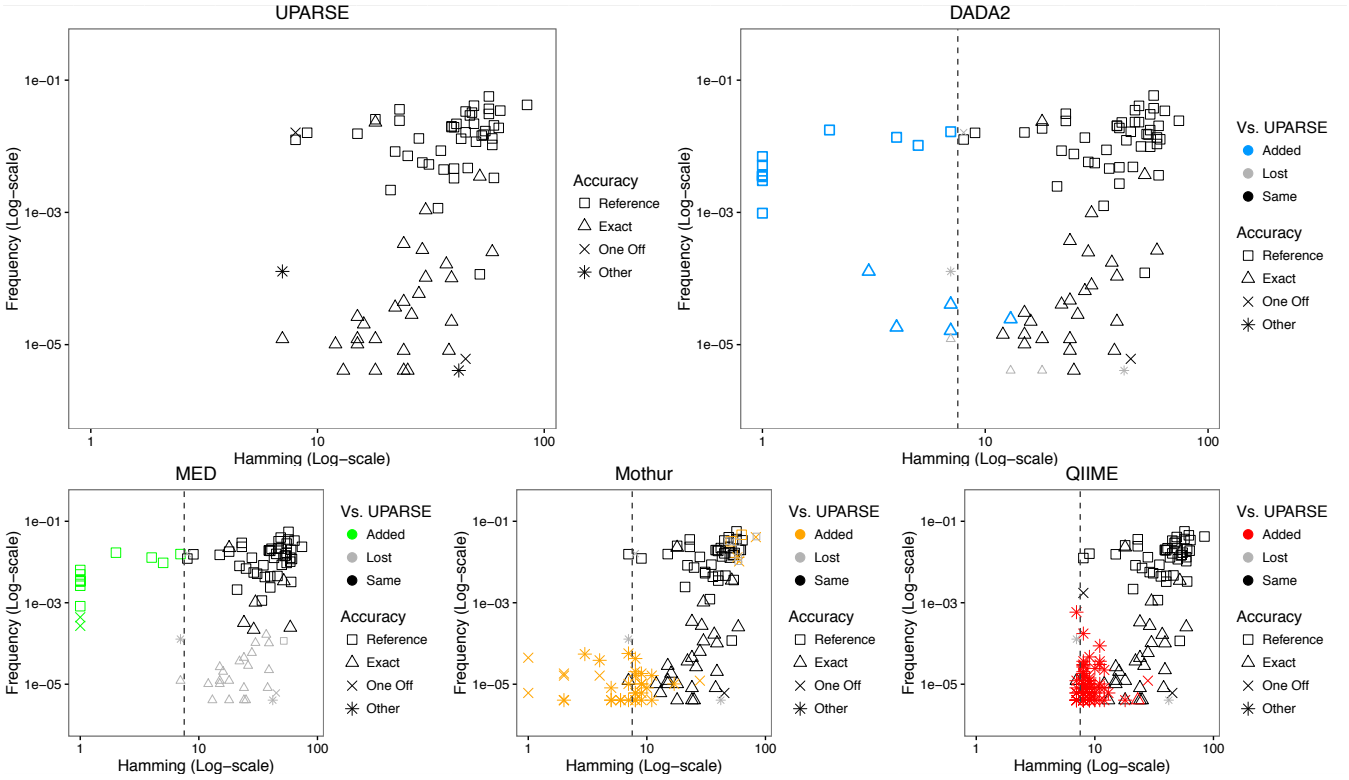


Figure 5. Frequency of output sequences from the Balanced merged dataset is plotted on the y-axis. Hamming distance from each sequence to its nearest more-abundant neighbor is plotted on the x-axis. UPARSE is used as a baseline to which the outputs of the other methods are compared. Algorithms largely concur (black) in identifying sequences that are abundant and very different from other sample sequences. However, DADA2 detects additional variation (blue) relative to UPARSE, especially within UPARSE's OTU radius (dashed line). MED also detects some fine-scale variation (green), but at the cost of some false positives, typically One Offs that are 1-away from a more abundant correct sequence, and MED does not detect low abundance sequences (grey). Mothur (orange) and QIIME (red) both report a significant number of additional spurious sequences, although most are relatively low frequency.

The output sequences inferred from the HMP merged reads by UPARSE, DADA2, MED, mothur (average-linkage) and QIIME (uclust).

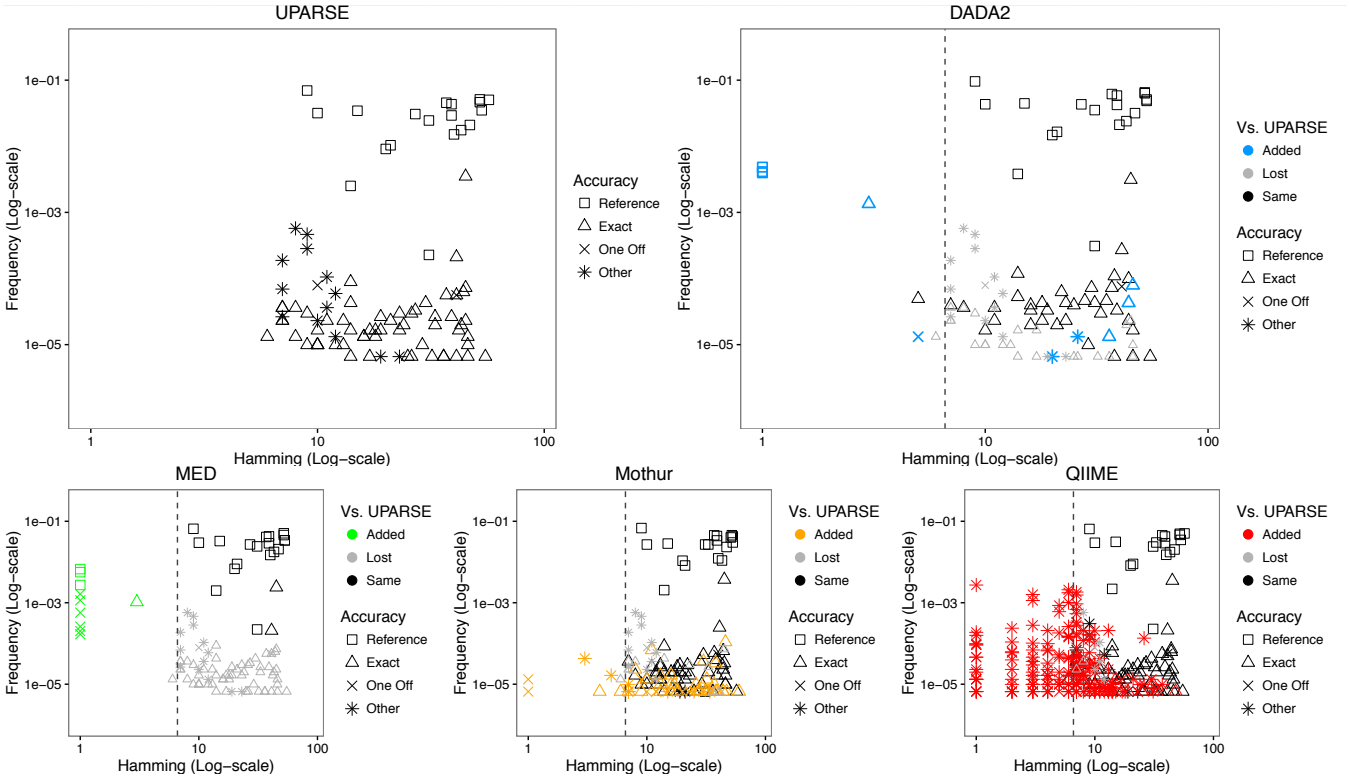


Figure 6. Frequency of output sequences from the HMP merged dataset is plotted on the y-axis. Hamming distance from each sequence to its nearest more-abundant neighbor is plotted on the x-axis. UPARSE is used as a baseline to which the outputs of the other methods are compared. Algorithms largely concur (black) in identifying sequences that are abundant and very different from other sample sequences. However, DADA2 detects additional variation (blue) relative to UPARSE, especially within UPARSE's OTU radius (dashed line). MED also detects some fine-scale variation (green), but at the cost of some false positives, typically One Offs that are 1-away from a more abundant correct sequence, and MED does not detect low abundance sequences (grey). Mothur (orange) and QIIME (red) both report a significant number of additional spurious sequences, although most are relatively low frequency.

The output sequences inferred from the Extreme merged reads by UPARSE, DADA2, MED, mothur (average-linkage) and QIIME (uclust).

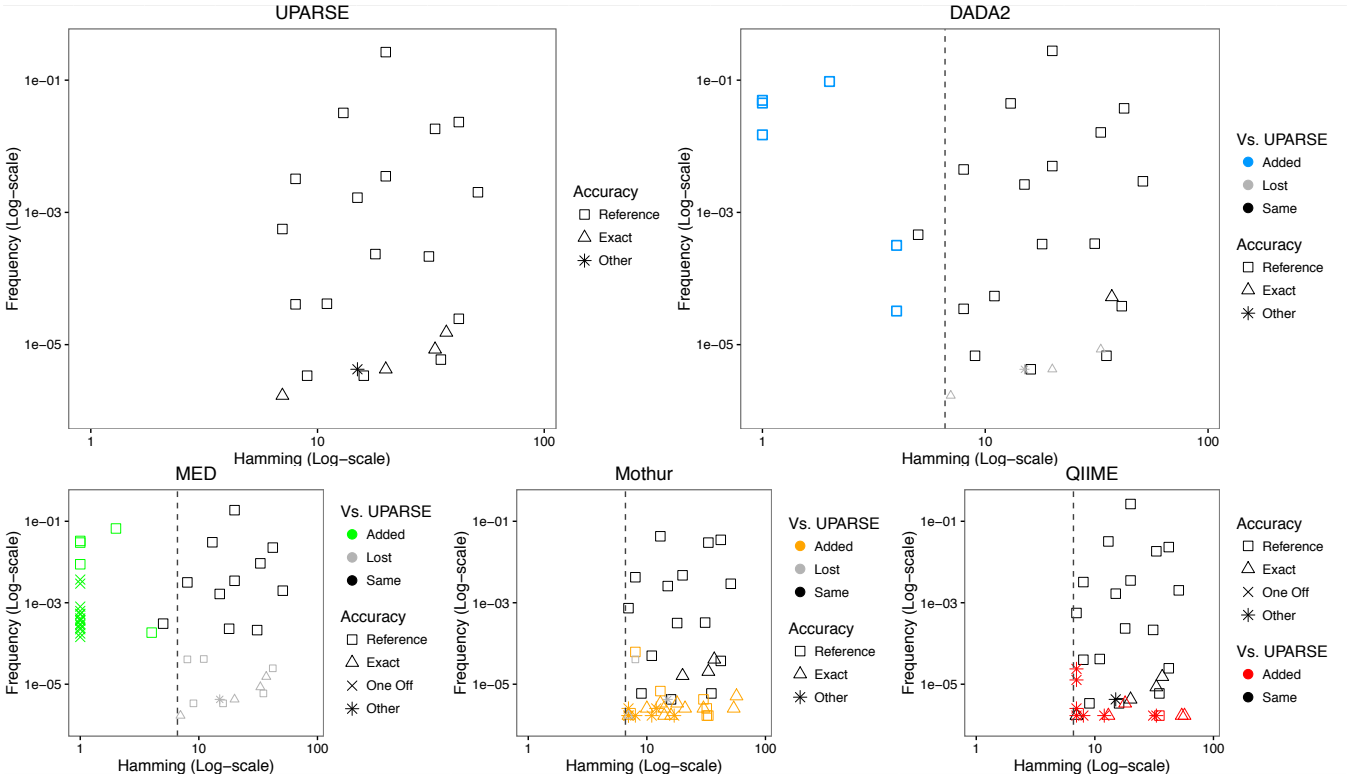


Figure 7. Frequency of output sequences from the Extreme merged dataset is plotted on the y-axis. Hamming distance from each sequence to its nearest more-abundant neighbor is plotted on the x-axis. UPARSE is used as a baseline to which the outputs of the other methods are compared. Algorithms largely concur (black) in identifying sequences that are abundant and very different from other sample sequences. However, DADA2 detects additional variation (blue) relative to UPARSE, especially within UPARSE's OTU radius (dashed line). MED also detects some fine-scale variation (green), but at the cost of some false positives, typically One Offs that are 1-away from a more abundant correct sequence, and MED does not detect low abundance sequences (grey). Mothur (orange) and QIIME (red) both report some additional low-frequency spurious sequences.

Illumina Miseq error rates as a function of quality.

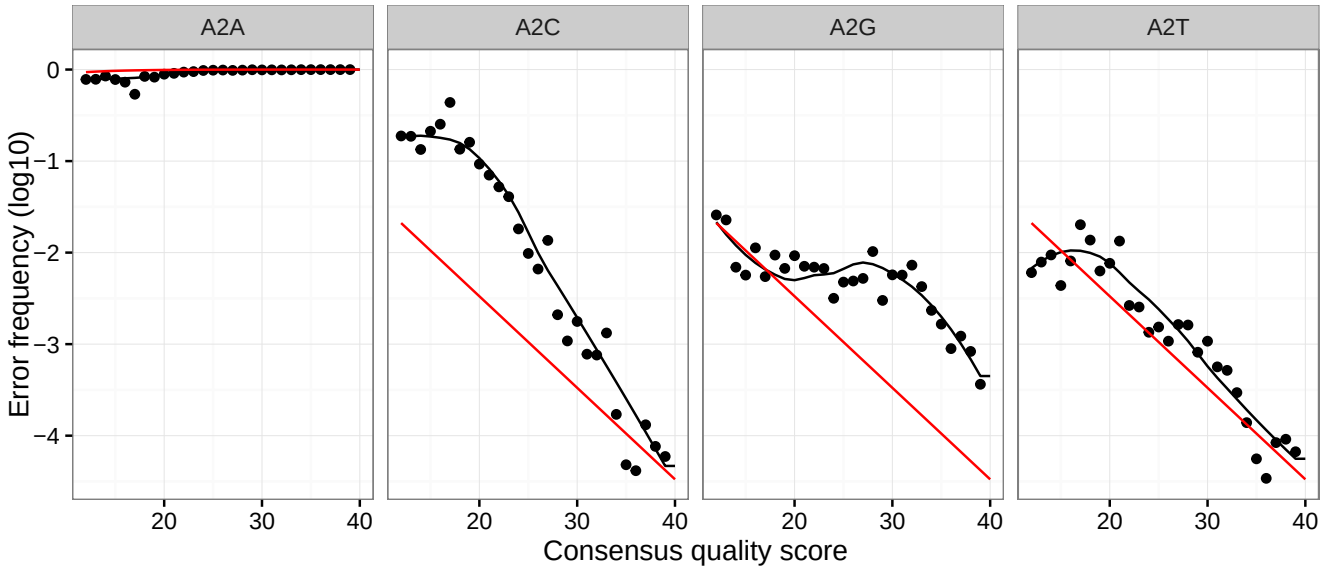


Figure 8. The forward-read error rates observed in the 142 pooled samples from MacIntyre 2015 are shown for the case where the correct base is an A. The x-axis shows the quality score; the y-axis the frequency of the specified transition. Dots show the observed frequencies, the black line the error model inferred by DADA2 using its default loess fitting, and the red line the expected rates given the nominal definition of the quality score: $Q = -10log_{10}(p_{err})$. Illumina quality scores are quite informative about substitution error rates, but systematic deviations from the expected rates are observed. This plot was generated by the `plotErrors` function in the DADA2 R package.

Supplementary Tables

	Direction	Reads	Quality Scores 1Q	Mean	3Q	Filtered Reads Merged	Forward Only
Balanced	Forward	593868	37	35.92	38	492082	557946
	Reverse	593868	34	33.52	39	492082	
HMP	Forward	613352	31	32.34	38	303293	449269
	Reverse	613352	16	28.7	37	303293	
Extreme	Forward	2082062	32	33	38	1178835	1431321
	Reverse	2082062	23	29.33	37	1178835	

Table 1. Sequencing summary of the Balanced, HMP and Extreme test datasets.

Strain	Greengenes taxonomy	Dilution group	Tag
Bacteroides cellulosilyticus DSM 14838	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Bacteroidaceae	10^-1	t__70164
Bacteroides eggerthii BEI HM-210	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Bacteroidaceae	10^-5	t__91265
Bacteroides fragilis ATCC 23745	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Bacteroidaceae	10^-3	t__84566
Bacteroides massiliensis JCM 12982	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Bacteroidaceae	10^-4	t__38326
Bacteroides ovatus DSM 1896	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Bacteroidaceae	10^-0	t__74296
Bacteroides thetaiotaomicron DSM 2079	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Bacteroidaceae	10^-3	t__50907
Bacteroides uniformis DSM 6597	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Bacteroidaceae	10^-2	t__89266
Bacteroides vulgatus DSM 1447	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Bacteroidaceae	10^-0	t__21615
Barnesiella intestinihominis DSM 21032	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__[Barnesiellaceae]	10^-5	t__21316
Clostridium celatum JCM 1394	k__Bacteria,p__Firmicutes,c__Clostridia,o__Clostridiales,f__Clostridiaceae	10^-3	t__31700
Clostridium cocleatum DSM 1551	k__Bacteria,p__Firmicutes,c__Erysipelotrichi,o__Erysipelotrichales,f__Erysipelotrichaceae	10^-2	t__56149
Clostridium methylpentosum DSM 5476	k__Bacteria,p__Firmicutes,c__Clostridia,o__Clostridiales,f__Ruminococcaceae	10^-5	t__68506
Clostridium phytofermentans ATCC 700394	k__Bacteria,p__Firmicutes,c__Clostridia,o__Clostridiales,f__Lachnospiraceae	10^-5	t__73805
Clostridium xylanovorans DSM 12503	k__Bacteria,p__Firmicutes,c__Clostridia,o__Clostridiales,f__Lachnospiraceae	10^-1	t__70731
Coprococcus comes ATCC 27758	k__Bacteria,p__Firmicutes,c__Clostridia,o__Clostridiales,f__Lachnospiraceae	10^-2	t__41496
Eubacterium rectale DSM 17629	k__Bacteria,p__Firmicutes,c__Clostridia,o__Clostridiales,f__Lachnospiraceae	10^-4	t__53720
Howardella ureilytica DSM 15118	k__Bacteria,p__Firmicutes,c__Clostridia,o__Clostridiales,f__91otu452	10^-5	t__23462
Parabacteroides distasonis JCM 13400	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Porphyromonadaceae	10^-5	t__21798
Parabacteroides distasonis JCM 13401	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Porphyromonadaceae	10^-1	t__1086
Parabacteroides merdae DSM 19495	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Porphyromonadaceae	10^-2	t__33431
Parabacteroides sp. D13 BEI HM-77	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Porphyromonadaceae	10^-5	t__91131
Paraprevotella clara DSM 19731	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__[Paraprevotellaceae]	10^-4	t__81974
Prevotella buccalis ATCC 35310	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Prevotellaceae	10^-5	t__52712
Prevotella copri DSM 18205	k__Bacteria,p__Bacteroidetes,c__Bacteroidia,o__Bacteroidales,f__Prevotellaceae	10^-5	t__30223
Roseburia intestinalis DSM 14610	k__Bacteria,p__Firmicutes,c__Clostridia,o__Clostridiales,f__Lachnospiraceae	10^-4	t__27000
Roseburia inulinivorans DSM 16841	k__Bacteria,p__Firmicutes,c__Clostridia,o__Clostridiales,f__Lachnospiraceae	10^-3	t__29181
Ruminococcus gnavus ATCC 29149	k__Bacteria,p__Firmicutes,c__Clostridia,o__Clostridiales,f__Lachnospiraceae	10^-5	t__30356

Table 2. Composition of Extreme mock community. Tags were included in the id lines of the fasta reads from each amplified strain.

		Forward			Merged		
		Total	Chimeric	Non-chimeric	Total	Chimeric	Non-chimeric
Balanced	DADA2	0	0	0	0	0	0
	UPARSE	2	0	2	2	1	1
	MED	0	0	0	0	0	0
	Mothur	165	102	63	28	10	18
	QIIME	290	185	105	93	71	22
HMP	DADA2	8	1	7	2	0	2
	UPARSE	8	0	8	13	10	3
	MED	0	0	0	0	0	0
	Mothur	605	58	547	10	4	6
	QIIME	1118	182	936	191	110	81
Extreme	DADA2	4	0	4	0	0	0
	UPARSE	13	0	13	1	0	1
	MED	0	0	0	0	0	0
	Mothur	*	*	*	7	0	7
	QIIME	3100	5	3095	8	0	8

Table 3. Classification of Other sequences as chimeric or non-chimeric. The uparse-ref algorithm was used to classify the Other sequences output by each method as chimeric or non-chimeric by comparison to the references sequences for each mock community. The same parameter setting for uparse-ref were used as in [9].

Supplementary Notes

Supplementary Note 1: The limits of DADA2's sensitivity to rare variants

DADA2 relies on repeated observations of a sequence to distinguish biological variants from sequencing error. Thus, in order to identify a biological variant, there must be at least 2 error-free reads of that sequence variant present in the data, with more required if the variant is close to another sample sequence. Given a sequencing depth D, and a fraction of error-free sequences f (this varies greatly, but $f \sim 0.5$ is not uncommon for Illumina Miseq 2x250 forward reads), DADA2 will not effectively identify variants with frequencies at or below $2/Df$, as such variants are unlikely to produce at least two error-free reads.

This limitation is illustrated by the 4 reference strains DADA2 failed to detect in the forward reads of the Extreme dataset. Strains in this dataset were separately PCR-ed and tagged with an indexing barcode, so we can use that information (encoded in the id line of the fastq file) to examine these false negatives in detail. Each strain that DADA2 failed to identify in the forward reads is listed below. NN refers to the nearest neighbor, i.e. the most similar sequence that was identified by DADA2. Max abundance is the maximum abundance among the unique sequences present in each strain's reads:

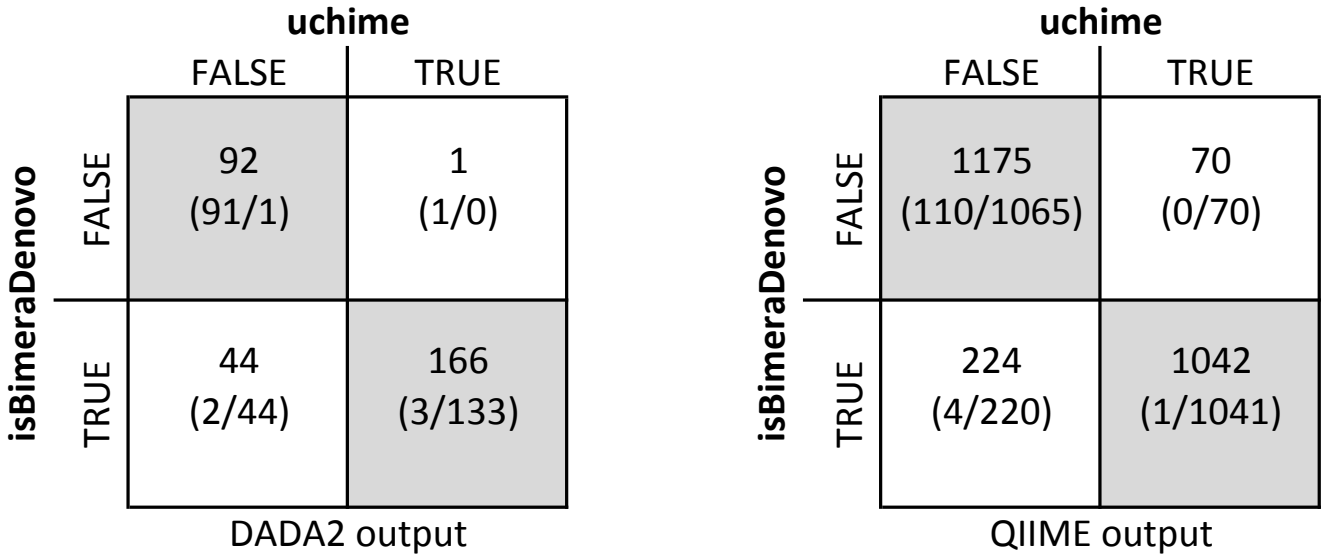
Strains missed by DADA2	Total Reads	Max Abundance	Hamming to NN	NN Abundance	Tag
Prevotella buccalis	5	1	51	9	t__52712
Clostridium methylpentosum DSM 5476	5	1	25	13	t__68506
Clostridium phytofermentans ISDg	11	2	15	59564	t__73805
Parabacteroides sp. D13	3	2	1	28242	t__91131

Prevotella buccalis and *Clostridium methylpentosum* cannot be identified by DADA2 because the reads from those strains are all unique singletons. The *Parabacteroides sp. D13* reads include a doubleton sequence, but it is 1-away from another real variant with abundance 28242, and it is not possible to statistically differentiate sequences at such low abundance so near to another sequence at such high abundance from errors. Finally, *Clostridium phytofermentans* has a doubleton that is 15-away from another inferred sequence with $\sim$ 60k reads. This is on the boundary of DADA2's sensitivity under the default sensitivity parameter $\Omega_A = 10^{-40}$, and if this parameter is made less conservative DADA2 will detect this sequence. However, here we used default parameters, including the conservative default Ω_A.

Supplementary Note 2: Comparing uchime and isBimeraDenovo

The common chimera identification algorithms, and uchime in particular, were designed under the assumption that sensitivity on nearby chimeric variants was not very important because nearby variants were likely to be joined into the same OTU anyway. However, this assumption is violated by DADA2, as DADA2 distinguishes variants that differ by as little as one nucleotide. To overcome this shortcoming we developed a simple new algorithm (isBimeraDenovo) that can identify chimeras at any separation (Methods).

To test the sensitivity and specificity of isBimeraDenovo we processed the Balanced forward reads with DADA2 and with QIIME/uclust without any chimera filtering, and then compared the results of isBimeraDenovo and uchime on the output of each method. The rows (isBimeraDenovo) and columns (uchime) of these 2x2 tables correspond to whether a sequence was identified as non-chimeric (FALSE) or chimeric (TRUE). The entries indicate the number of corresponding output sequences or OTUs, and in parentheses are the number of those sequences that exactly (matched, did not match) a reference sequence or nt (Methods).

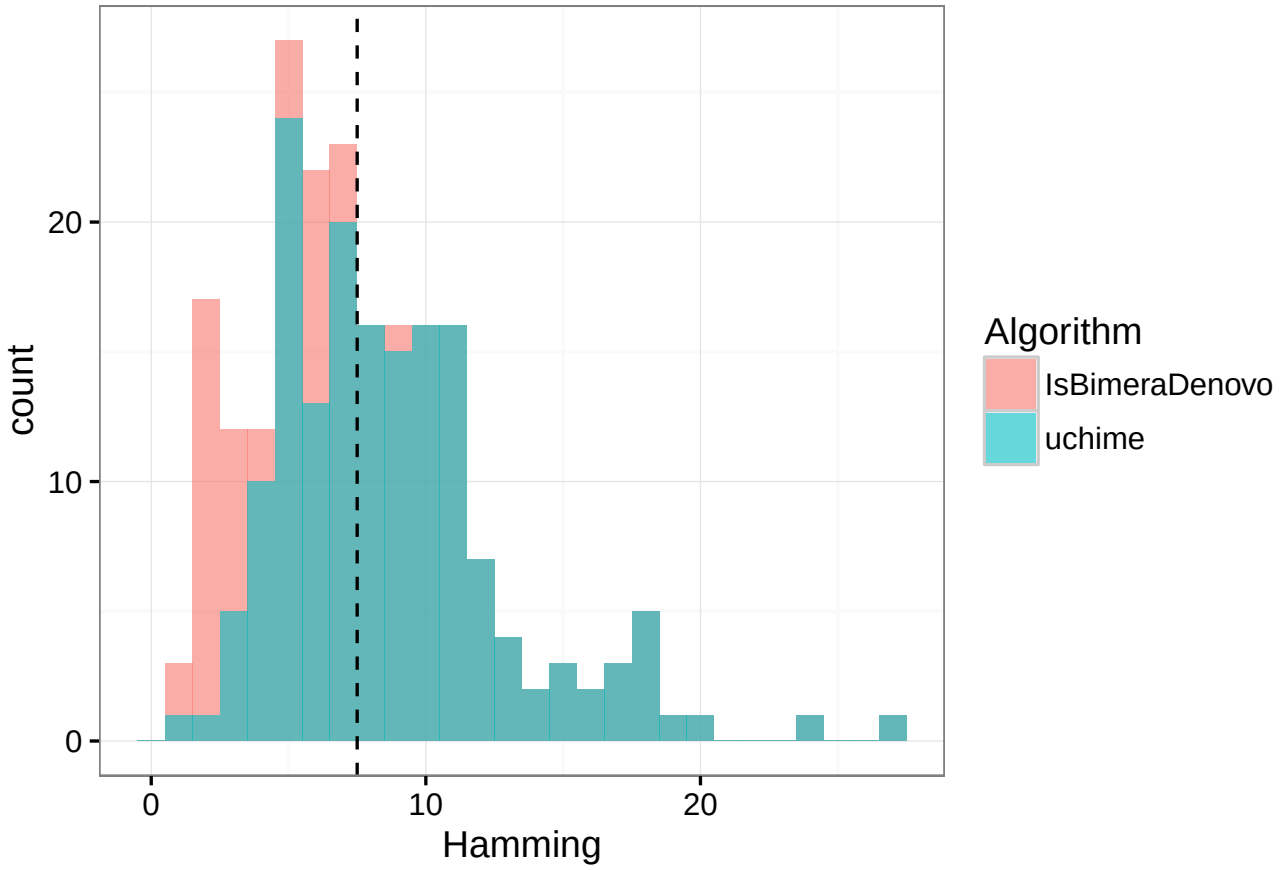


When applied to the DADA2 output sequences, isBimeraDenovo and uchime mostly agreed: 166/211 output sequences flagged be either algorithm were flagged by both. The main difference between methods were the 44 sequences flagged as chimeric by DADA2 but not by uchime. Of those 44 sequences, 42/44 did not match a reference sequence or nt, consistent with these putative chimeras being spurious variants. Furthermore, the 1 sequence that uchime alone flagged as chimeric was an exact match, consistent with this variant being a real sequence that isBimeraDenovo correctly left unflagged.

The results on the output from QIIME/uclust were less clear. While the algorithms again agreed on most chimera calls, each algorithm flagged a significant number of sequences that the other did not, and both sets of algorithm-specific chimera calls appeared to consist primarily of spurious sequences. isBimeraDenovo

flagged 224 sequences that uchime did not, of which 220/224 were not matches, while uchime flagged 70 sequences that isBimeraDenovo did not, of which 70/70 were not matches.

For a closer look at the characteristics of the chimeras identified by isBimeraDenovo in the DADA2 output, we plotted a histrogram of the sequences flagged by each algorithm as a function of the hamming distance between the flagged sequences and the nearest more abundant output sequence.



The histograms for each method are overlaid on top of each other, and the uchime bars are completely overlapped by the isBimeraDenovo bars. The excess area of the isBimeraDenovo bars shows that the additional variants it identified relative to uchime were nearby other output sequences, and in fact were almost entirely within 3% of a more abundant sequence (dashed line).

In total, these results show that isBimeraDenovo is more sensitive to nearby chimeras than uchime while having a similarly high level of specificity. This makes isBimeraDenovo particularly well-suited for exact sample inference methods like DADA2 which distinguish closely related sequences. The evidence for the utility of isBimeraDenovo with fuzzier OTU methods was more mixed. The increased sensitivity of

isBimeraDenovo to nearby variants was useful for QIIME/uclust as well, but isBimeraDenovo failed to identify some chimeras with residual errors and chimeras whose "parent sequences" were subsumed into another OTU.