Open OTU picking and assignment methods

Each of the four amplicons was submitted to the same workflow, separately, to pick OTUs and assign read abundance to a Sample-by-OTU matrix. This workflow followed a similar conceptual outline to that advocated in the QIIME open reference OTU picking pipeline [1], with the following differences: a) USEARCH 64 bit v8.0.1517 was employed directly; b) reference OTUs were not initially assigned via a round of closed reference picking, instead *de novo* OTUs were picked (OTUs were classified later); c) in order make compute time manageable for *de novo* picking, OTUs were initially picked on the numerically dominant sequences only (sequences with > 6 representatives across the full dataset); d) instead of randomly picking sequences that failed to be recruited to OTUs for subsequent clustering, all sequences with >2 representatives were used. The workflow proceeded via the following commands, where "filename" indicates the input (*.FASTA) filename:

usearch64 -derep_fulllength filename.fasta -fastaout filename_unique.fasta -sizeout -slots 1000000001

usearch64 -sortbysize filename_unique.fasta -fastaout filename_unique_sorted10.fasta -minsize 6

usearch64 -cluster_otus filename_unique_sorted4.fasta -otus filename_unique_sorted6_otus.fasta –uparseout filename_unique_sorted6.uparse –relabel

The .uparse file was used to select chimeras, which were then clustered into "chimeric OTUS" using the usearch fast cluster method. OTU's and chimeric OTUs were joined to give an OTU database to map against ("filename_otus_chimeras.fasta").

usearch64 –usearch_global filename.fasta -db filename_otus_chimeras.fasta -strand plus -id 0.97 -uc filename_unique_sorted6_readmap.uc -maxaccepts 10 -top_hit_only -maxrejects 256

Sequences mapping to chimeras were discarded, unmapped, nonchimeric sequences were clustered to form additional OTU's following the method above. OTU sequences from both rounds of de novo OTU clustering were concatenated to give the final OTU database, against which reads were mapped (usearch_global as above) to produce the OTU table.

Sequences were identified using GreenGenes(13-5), UNITE(v7.0) or SILVA123 using the rdp clsassifer [4] as implemented in MOTHUR at 60 % probability.

The final sample-by-OTU data matrix and taxonomy file were created by discarding sequences not identified as belonging to the correct lineage (*i.e.*, bacteria, archaea, fungi, eukaryotes), unidentified at the phylum level, or having < 50 sequences across all samples in the database.

These final curation steps were guided by analysis of mock community samples

1. Rideout JR, He Y, Navas-Molina JA, Walters WA, Ursell LK, Gibbons SM et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ. 2014;2:e545. doi:10.7717/peerj.545.
2. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Meth. 2013;10(10):996-8. doi:10.1038/nmeth.2604

http://www.nature.com/nmeth/journal/v10/n10/abs/nmeth.2604.html#supplementary-information.
3. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460-1. doi:10.1093/bioinformatics/btq461.

4. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 2007;73(16):5261-7. doi:10.1128/aem.00062-07.