iTagger v2.1 METHODS

iTagger is the production pipeline used for processing Illumina amplicon libraries at the US DOE Joint Genome Institute. This pipeline uses usearch and qiime to analyze amplicon libraries, such as 16S rRNA or fungal ITS variable regions for phylogenetic analysis. All samples to be compared should be identically constructed, sequenced, and analyzed. This software wraps popular tools to facilitate analysis of large numbers of samples and provides some minor enhancements for managing many samples and their counts and incremental clustering. Authors should credit the creators of usearch, mafft, and qiime.

READ QC:

Overlapping read pairs are merged into unpaired consensus sequences (by usearch's merge\_pairs and READ\_QC/MERGE\_MAX\_DIFF\_PCT parameter); unmerged reads are discarded. The PCR primers must be found with the correct orientation and within the expected spacing (by usearch's search\_oligodb and parameters: AMPLICON/LEN\_MEAN, AMPLICON/LEN\_STDEV, READ\_QC/PRIMERS, READ\_QC/PRIMER\_TRIM\_MAX\_DIFFS, READ\_QC/LEN\_FILTER\_MAX\_DIFFS), otherwise the read is discarded. The read quality scores are evaluated and those with too many expected errors are discarded (READ\_QC/MAX\_EXP\_ERR\_RATE). Lastly, identical sequences are dereplicated, counted, sorted alphabetically, and written to a .seqobs file for subsequent merging with other such files. Additionally, a log file is produced which records the ID and reason each filtered read was excluded.

CLUSTERING:

Samples containing less than CLUSTERING/SAMPLE\_MIN\_SIZE sequences are discarded. All remaining samples' seqobs files are combined (i.e. dereplicate identical sequences), and the sequences are sorted by decreasing abundance. Sequences are separated depending on whether they contain CLUSTERING/CENTROID\_MIN\_SIZE copies, where the low-abundance sequences are set aside and not used during clustering. The former are saved in a .fasta and matching .obs table which records the sequences per sample; the latter are saved in a .fasta file with the sequences per sample recorded in the sequence headers.

The clusterable sequences are incrementally clustered by usearch's cluster\_otus, starting at 99% identity, and increasing the radius by 1% each iteration until reaching CLUSTERING/OTU\_CLUSTERING\_PCT\_IDENTITY. The sequences are resorted by decreasing abundance between each step. After clustering, the low-abundance sequences are mapped to the cluster centroids (by usearch's usearch\_global) and are added to the OTUs' counts if they are within the prescribed percent-identity threshold, otherwise they are discarded; this step creates no new clusters. Refer to the USEARCH documentation (http://drive5.com) for a description of the usearch clustering algorithm. Note that while the tax.tsv file contains taxonomic predictions to the resolution specified by the config file parameter, CLASSIFICATION/CUTOFF, generally only the phylum level predictions are reliable, due to limitations of the reference database. Use more specific classifications with care.

CLASSIFICATION:

Cluster centroid sequences are evaluated with usearch's utax and the specified reference database (which may have been filtered). The resultant taxonomic predictions are filtered; if an OTU does not have any taxonomic classifications at the CLASSIFICATION/CUTOFF threshold, it is written to the otu.unknown.fasta and .obs files. Additionally, if the optional CLASSIFICATION/CONTAM regex is provided, any OTU with matching taxonomic classifications are filtered to the otu.contam.fasta and .obs files; this is generally used for removing chloroplast sequences from rhizosphere samples. Additionally, the optional CLASSIFICATION/KEEP\_ONLY regex allows specifying a regex that must be matched in order to be accepted, with all others being filtered as contaminant and being added to the same contam file as above; this may be used to discard all non-fungal sequences in an ITS library, for example. The accepted OTUs are found in the otu.fasta and .obs file.

MULTIPLE SEQUENCE ALIGNMENT AND PHYLOGENETIC TREE:

OTUs are aligned using MAFFT (using parameters: --maxiterate 1000 --globalpair) and a tree constructed using QIIME's make\_phylogeny.py, which produces a Newick file. It is left to the end user to generate graphical representations from this file.

DIVERSITY ANALYSES:

An OTU table, BIOM file, and QIIME-format mapping file are generated which may be used to analyze the results using QIIME tools. QIIME's core\_diversity\_analyses.py pipeline is run, using the DIVERSITY/SAMPLING\_DEPTH parameter, which generates several files under the core\_diversity\_analyses folder. Refer to the QIIME documentation for a description (http://qiime.org).

AUTHORS:

iTagger was originally written by Julien Tremblay (julien.tremblay@mail.mcgill.ca) and later developed by Edward Kirton (ESKirton@LBL.gov), to whom correspondence should be addressed. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

COPYRIGHT/LICENSE:

This software is Copyright (c) 2013 by the US DOE Joint Genome Institute but is freely available for use without any warranty under the same license as Perl itself. Refer to wrapped tools for their credits and license information.

AVAILABILITY:

The code is freely available at Bitbucket (https://bitbucket.org/berkeleylab/jgi\_itagger).

SUMMARY OF OUTPUT:

FILES GENERATED BY itaggerReadQc.pl:

ROOT/reads/\*.seqobs = sorted sequence and observations (counts)

ROOT/reads/\*.filtered = log of filtered reads

FILES GENERATED BY itaggerClusterOtus.pl:

ROOT/otu/align.txt = UTAX classification alignments

ROOT/otu/tax.tsv = UTAX classification table

ROOT/otu/otu.fasta = final OTU centroid sequences

ROOT/otu/otu.fasta.obs = final OTU centroid observations per sample table

ROOT/otu/contam.otu.fasta = filtered contaminant centroid sequences

ROOT/otu/contam.otu.fasta.obs = filtered contaminant centroid obs table

ROOT/otu/unk.otu.fasta = unclassified OTU centroid sequences

ROOT/otu/unk.otu.fasta.obs = unclassified OTU obs table

ROOT/otu/log.txt = summary report

ROOT/otu/mapping.txt = QIIME mapping file

ROOT/otu/otu.tsv = OTU table of abundances, including taxonomy

ROOT/otu/otu.biom = OTU abundance+tax table in BIOM format

ROOT/otu/msa.fasta = multiple sequence alignment of final centroids

ROOT/otu/otu.tre = phylogenetic tree of final centroids

ROOT/otu/core\_diversity\_analyses/ = folder containing QIIME core diversity analyses output

Software used:

iTagger 2.2

usearch v9.2.64\_i86linux32

mafft v7.310 (2017/Mar/17)

QIIME 1.9.1