Appendix B: Bash script used for the bioinformatic treatment of the sequencing data and taxonomic assignations with the OBITools package

# Alignment and merging of paired-end reads using illuminapairedend. The program

# assigns an alignment score to each resulting sequence based on the phred

# quality scores and the length of the aligned regions.

illuminapairedend --fasta-output -r readsR1.fastq readsR2.fastq \

> readsR1R2.fasta

# Reads assignment using ngsfilter. The program requires a table providing the

# information regarding the primer pair and the tag combination used for each

# sample (see the OBITools documentation for more details). This step will add

# an attribute to each sequence containing the name of the corresponding sample.

# Other information can be added.

ngsfilter -t ngsfilter.tab -e 2 --nuc readsR1R2.fasta > readsR1R2\_ngsfilt.fasta

# Removal of low quality reads using obigrep (alignment scores<50, containing Ns

# or shorter than 50bp)

obigrep -s '^[acgt]+$' -l 50 -p 'score>=50' readsR1R2\_ngsfilt.fasta \

> readsR1R2\_ngsfilt\_lowqual.fasta

# Dereplication of the sequences using obiuniq (regroups every identical reads

# assigned to the same sample into one sequence and keeps the coverage information)

obiuniq -c sample readsR1R2\_ngsfilt\_lowqual.fasta \

> readsR1R2\_ngsfilt\_lowqual\_derep.fasta

# Check for the maximum sequence coverage found among sequence assigned to non-used # tag combinations (i.e. resulting from tag-switching events), using obistat. These

# sequences can be selected with obigrep thanks to an attribute added in the

# ngsfilter tab. Here, we used the value 'NU' (not used) in the ID attribute

obigrep -a 'ID:NU' readsR1R2\_ngsfilt\_lowqual\_derep.fasta | obistat -M count

# Remove sequences with a count lower than the maximum coverage found among

# sequence resulting from tag-switching events (as identified in the previous

# step, 1174 in our case)

obigrep -p 'count>1174' readsR1R2\_ngsfilt\_lowqual\_derep.fasta \

> readsR1R2\_ngsfilt\_lowqual\_derep\_min1174.fasta

# Taxonomic assignations of each sequence using ecotag. The program requires a

# list of reference sequences in fasta format annotated with GenBank taxids

# (taxid=XX) and the corresponding genbank taxonomy dump (or ecoPCR taxonomy database)

ecotag -R minicircles\_refDB.fasta -d genbank\_rXX \

readsR1R2\_ngsfilt\_lowqual\_derep\_min1174.fasta > minicircles\_ecotag.fasta

# Output assignation statistics in a tab-delimited file

obistat -c sample -c species\_name minicircles\_ecotag.fasta **\**

> minicircles\_assignation\_stat.txt

# Run the small R script provided in the supplementary material to output an identification table

Rscript Suppinfo2.R