Identification of insects in soil samples through DNA metabarcoding

Second report

Piracicaba, SP
November, 2022

Experimental procedures

On 19th September 2022, EcoMol received 70 samples of soil (Table 1). Forty-eight soil samples were stored in plastic bags (zip lock) containing 100g of silica, and the other 22 samples were stored in plastic bottles containing 120ml of EcoMol preservation buffer.

DNA was extracted using *DNeasy PowerSoil Pro kit* (Qiagen), with some modifications implemented by EcoMol. To avoid contamination, all extractions occurred in a dedicated room, under sterile conditions. Extracted DNA concentration (ng/µL) and purity (A_{260}/A_{280} ratio) were estimated using the Nanodrop 2000 spectrophotometer (Thermo Scientific) (Table 1).

Table 1. Soil samples and replicates description, their storage method (silica or EcoMol preservation buffer), and DNA concentration (ng/µL) and purity (A_{260}/A_{280}) estimated in spectrophotometer. (EM110) EcoMol project name (*) name of samples after bioinformatics (**) in the first report, DNA was eluted in 50µL of EB buffer after extraction, and then diluted before PCR1 amplifications. In this second report, DNA was eluted in 80µL of EB buffer after extraction with no further dilution before PCR amplifications, what explains differences in the concentration of extracted DNA between reports.

N	Storage	Site	Treatment	Sample	Replicate	ID_EcoMol*	[ng/µL]**	A_{260}/A_{280}
1	Silica	Edmundo	Juquia	EM110_01	1	EM110_01A	16,8	1,93
2	Silica	Edmundo	Juquia		2	EM110_01B	15,4	1,78
3	Silica	Edmundo	Juquia		3	EM110_01C	18,6	1,79
4	Silica	Everaldo	Floresta	EM110_02	1	EM110_02A	43,5	1,59
5	Silica	Everaldo	Floresta		2	EM110_02B	9,1	1,79
6	Silica	Everaldo	Floresta		3	EM110_02C	41,1	1,57
7	Silica	Everaldo	Juquira	EM110_03	1	EM110_03A	11,5	1,86
8	Silica	Everaldo	Juquira		2	EM110_03B	14,2	1,71
9	Silica	Everaldo	Juquira		3	EM110_03C	13,5	1,98
10	Silica	Everaldo	Pasto	EM110_04	1	EM110_04A	12	1,76
11	Silica	Everaldo	Pasto		2	EM110_04B	9,8	1,76
12	Silica	Everaldo	Pasto		3	EM110_04C	11,6	1,74
13	Silica	Viriato	Pasto	EM110_05	1	EM110_05A	14,9	1,68
14	Silica	Viriato	Pasto		2	EM110_05B	21	1,63
15	Silica	Viriato	Pasto		3	EM110_05C	13,1	1,74
16	Silica	Viriato	SAF 1	EM110_06	1	EM110_06A	15,1	1,75
17	Silica	Viriato	SAF 1		2	EM110_06B	19,5	1,73
18	Silica	Viriato	SAF 1		3	EM110_06C	20,9	1,69
19	Silica	Viriato	SAF 2	EM110_07	1	EM110_07A	16,6	1,83
20	Silica	Viriato	SAF 2		2	EM110_07B	16,3	1,78
21	Silica	Viriato	SAF 3		3	EM110_07C	16,7	1,66
22	Silica	Viriato	Floresta	EM110_08	1	EM110_08A	15,5	1,65
23	Silica	Viriato	Floresta		2	EM110_08B	15	1,64
24	Silica	Viriato	Floresta		3	EM110_08C	14,7	1,67
25	Silica	Idesam	SAF	EM110_09	1	EM110_09A	13,6	1,66
26	Silica	Idesam	SAF		2	EM110_09B	14,6	1,68
27	Silica	Idesam	SAF		3	EM110_09C	13,7	1,79
28	Silica	Jocelio	Juquira	EM110_10	1	EM110_10A	13,8	1,75
29	Silica	Jocelio	Juquira		2	EM110_10B	14,8	1,70
30	Silica	Jocelio	Juquira		3	EM110_10C	14,6	1,66
31	Silica	Jocelio	Pasto	EM110_11	1	EM110_11A	10,8	1,57
32	Silica	Jocelio	Pasto		2	EM110_11B	13	1,65
33	Silica	Jocelio	Pasto		3	EM110_11C	13	1,67
34	Silica	Ronaldo	SAF 1	EM110_12	1	EM110_12A	9,4	1,60
35	Silica	Ronaldo	SAF 1		2	EM110_12B	11,3	1,63
36	Silica	Ronaldo	SAF 1		3	EM110_12C	11,8	1,58

37	Silica	Ronaldo	Floresta	EM110_13	1	EM110_13A	21,3	1,77
38	Silica	Ronaldo	Floresta		2	EM110_13B	23,8	1,68
39	Silica	Ronaldo	Floresta		3	EM110_13C	21,7	1,73
40	Silica	Ronaldo	SAF 2	EM110_14	1	EM110_14A	9,6	1,49
41	Silica	Ronaldo	SAF 2		2	EM110_14B	9,5	1,43
42	Silica	Ronaldo	SAF 2		3	EM110_14C	10,7	1,51
43	Silica	Celione	Floresta	EM110_15	1	EM110_15A	20,4	1,69
44	Silica	Celione	Floresta		2	EM110_15B	21,6	1,62
45	Silica	Celione	Floresta		3	EM110_15C	22,9	1,62
46	Silica	Celione	Juquira	EM110_16	1	EM110_16A	12,7	1,52
47	Silica	Celione	Juquira		2	EM110_16B	13,7	1,53
48	Silica	Celione	Juquira		3	EM110_16C	13,7	1,49
49	Buffer	Celione	Floresta	EM110_17	1	EM110_17A	68,8	1,48
50	Buffer	Celione	Floresta		2	EM110_17B	32,9	1,67
51	Buffer	Celione	Floresta		3	EM110_17C	52,3	1,57
52	Buffer	Celione	Floresta		4	EM110_17D	31	1,66
53	Buffer	Jocelio	Pasto	EM110_18	1	EM110_18A	74,3	1,57
54	Buffer	Jocelio	Pasto		2	EM110_18B	15,9	1,65
55	Buffer	Jocelio	Pasto		3	EM110_18C	16,3	1,66
56	Buffer	Idesam	SAF	EM110_19	1	EM110_19A	44,4	1,75
57	Buffer	Idesam	SAF		2	EM110_19B	33,1	1,64
58	Buffer	Idesam	SAF		3	EM110_19C	15,5	1,65
59	Buffer	Viriato	Floresta	EM110_20	1	EM110_20A	49,4	1,63
60	Buffer	Viriato	Floresta		2	EM110_20B	24,8	1,55
61	Buffer	Viriato	Floresta		3	EM110_20C	31,3	1,54
62	Buffer	Viriato	Pasto	EM110_21	1	EM110_21A	104,7	1,4
63	Buffer	Viriato	Pasto		2	EM110_21B	14,5	1,54
64	Buffer	Viriato	Pasto		3	EM110_21C	18,9	1,61
65	Buffer	Viriato	SAF 1	EM110_22	1	EM110_22A	128,3	1,44
66	Buffer	Viriato	SAF 1		2	EM110_22B	23	1,69
67	Buffer	Viriato	SAF 1		3	EM110_22C	21	1,43
68	Buffer	Viriato	SAF 2	EM110_23	1	EM110_23A	55	1,65
69	Buffer	Viriato	SAF 2		2	EM110_23B	31,4	1,64
70	Buffer	Viriato	SAF 2		3	EM110_23C	29,8	1,48

According to customers' option, to identify insect species in the soil samples, we amplified (PCR1) a ~133bp fragment of *cytochrome oxidase I* subunit (COI gene) from the mitochondrial DNA (named here insect_R1 fragment), with MG2_LCO1490_F forward primer and the MG2_univ_R reverse primer (Gillet et al. 2015; Tournayre et al. 2020). These primers were chosen because they successfully identified a higher number of insect orders than other primer sets previously tested (first report). Both forward and reverse primer sequences received Illumina i5 and i7 pre-adaptor tails to allow sequencing in the Illumina platform. Figure 1 presents the DNA barcode length and primers positions on the COI gene. Figure 2 presents a scheme of the amplification of the COI gene with primers with forward and reverse Illumina i5 and i7 pre-adaptor tails (PCR1). We included negative controls in all PCR1 reactions. Amplification conditions were the same as described in Tournayre et al. (2020); each PCR1 reaction contained (20µL): 10X PCRBio Master Mix (PCR Biosystems Inc.), 10µM of each primer and 5µL of extracted DNA. Volume reactions were completed with sterile ultrapure water. To avoid contaminations, all amplifications were conducted in a dedicated room, under sterile conditions.

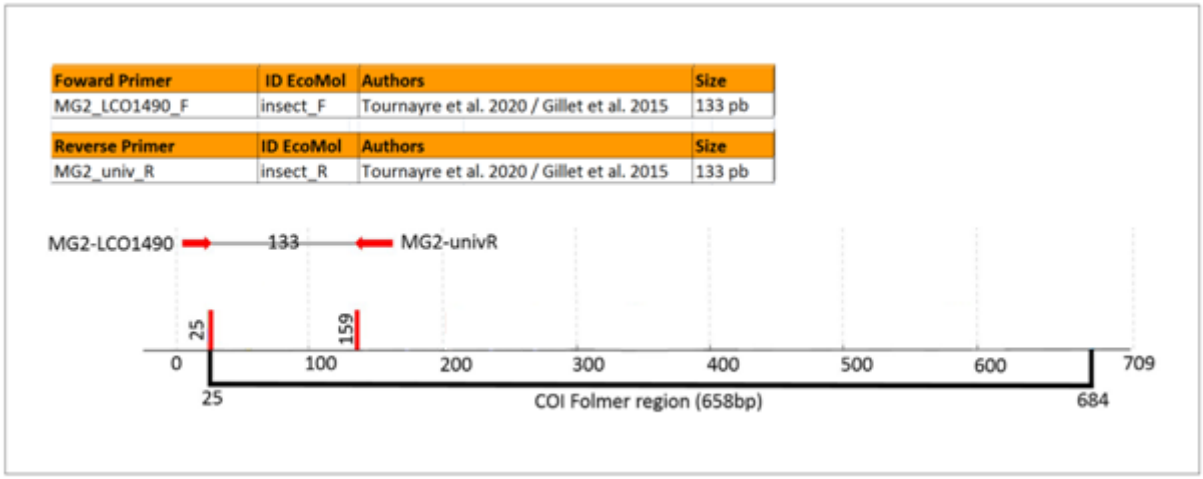

Figure 1. DNA barcode length (insect_R1 fragment) and primers positions on the COI gene. Modified from Tournayre et al. 2020.

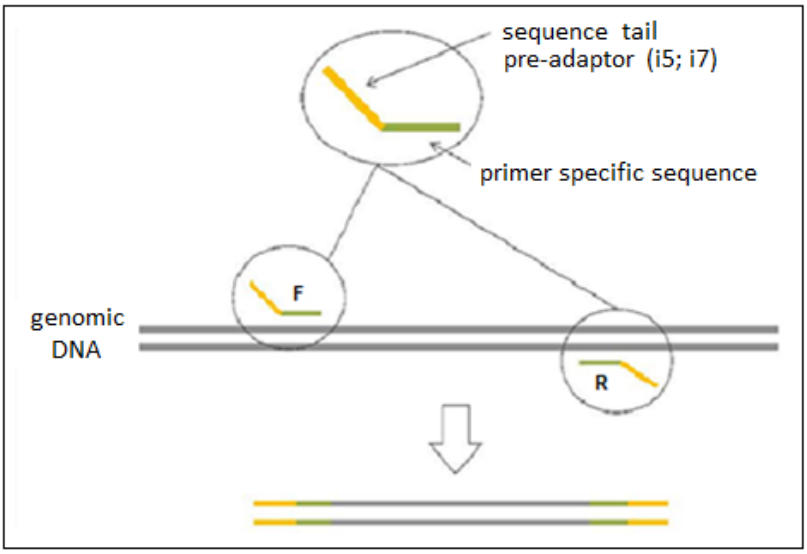

Figure 2. PCR1 scheme. Amplification of the target fragment from COI gene using specific primers (green) with complementary sequences (pre-adapters) to the adapters of the illumine platform (orange). (F) forward primer; (R) reverse primer

All PCR1 products were purified with magnetic beads (Agencourt AMPure XP® – Beckman Coulter). After purification, we conducted a second PCR (PCR2) using the *Nextera Index kit*® (Illumina) to amplify PCR1 products. In these reactions, Illumina adaptors (P5 and P7) are used as primers, as shown in Figure 3. Besides acting as primers, each adaptor has a unique 8bp index sequence that allows us to identify each sample after sequencing the pool of samples. Amplification of all PCR2 products was confirmed under UV light on 1.5 % agarose gel, as shown in Figure 4.

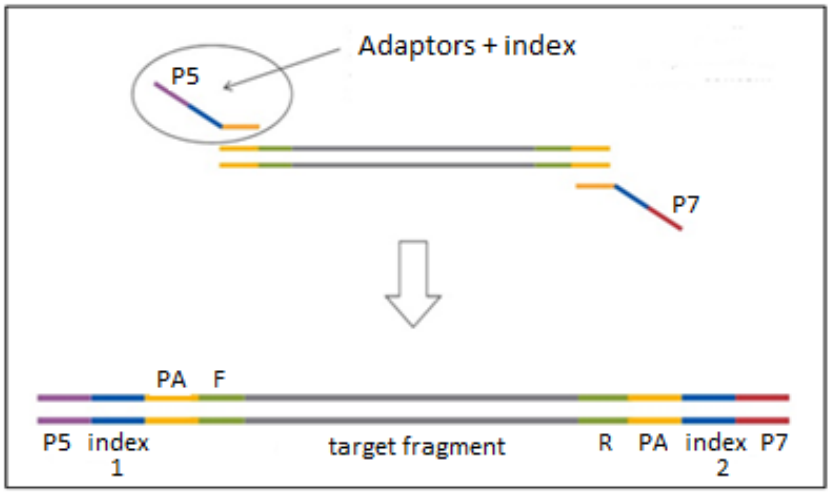

Figure 3. PCR2 scheme. Insertion of adapters P5 and P7, which make the set compatible with the Illumina flow cell, and of the index sequences (in blue), which allow the identification of each sample after sequencing. (F) forward primer; (R) reverse primer; (PA) pre-adaptors inserted in PCR1

	1	2	3	4	5	6	7	8	9
A	EM110_01A_R1	EM110_03C_R1	EM110_06B_R1	EM110_09A_R1	EM110_11C_R1	EM110_14B_R1	EM110_17A_R1	EM110_17D_R1	EM110_17B_R1
B	EM110_01B_R1	EM110_04A_R1	EM110_06C_R1	EM110_09B_R1	EM110_12A_R1	EM110_14C_R1	EM110_17C_R1	EM110_18C_R1	EM110_18B_R1
C	EM110_01C_R1	EM110_04B_R1	EM110_07A_R1	EM110_09C_R1	EM110_12B_R1	EM110_15A_R1	EM110_18A_R1	EM110_19C_R1	EM110_19B_R1
D	EM110_02A_R1	EM110_04C_R1	EM110_07B_R1	EM110_10A_R1	EM110_12C_R1	EM110_15B_R1	EM110_19A_R1	EM110_20C_R1	EM110_20B_R1
E	EM110_02B_R1	EM110_05A_R1	EM110_07C_R1	EM110_10B_R1	EM110_13A_R1	EM110_15C_R1	EM110_20A_R1	EM110_21C_R1	EM110_21B_R1
F	EM110_02C_R1	EM110_05B_R1	EM110_08A_R1	EM110_10C_R1	EM110_13B_R1	EM110_16A_R1	EM110_21A_R1	EM110_22C_R1	EM110_22B_R1
G	EM110_03A_R1	EM110_05C_R1	EM110_08B_R1	EM110_11A_R1	EM110_13C_R1	EM110_16B_R1	EM110_22A_R1	EM110_23C_R1	EM110_23B_R1
H	EM110_03B_R1	EM110_06A_R1	EM110_08C_R1	EM110_11B_R1	EM110_14A_R1	EM110_16C_R1	EM110_23A_R1	PCR1_NEG_R1	PCR2_NEG

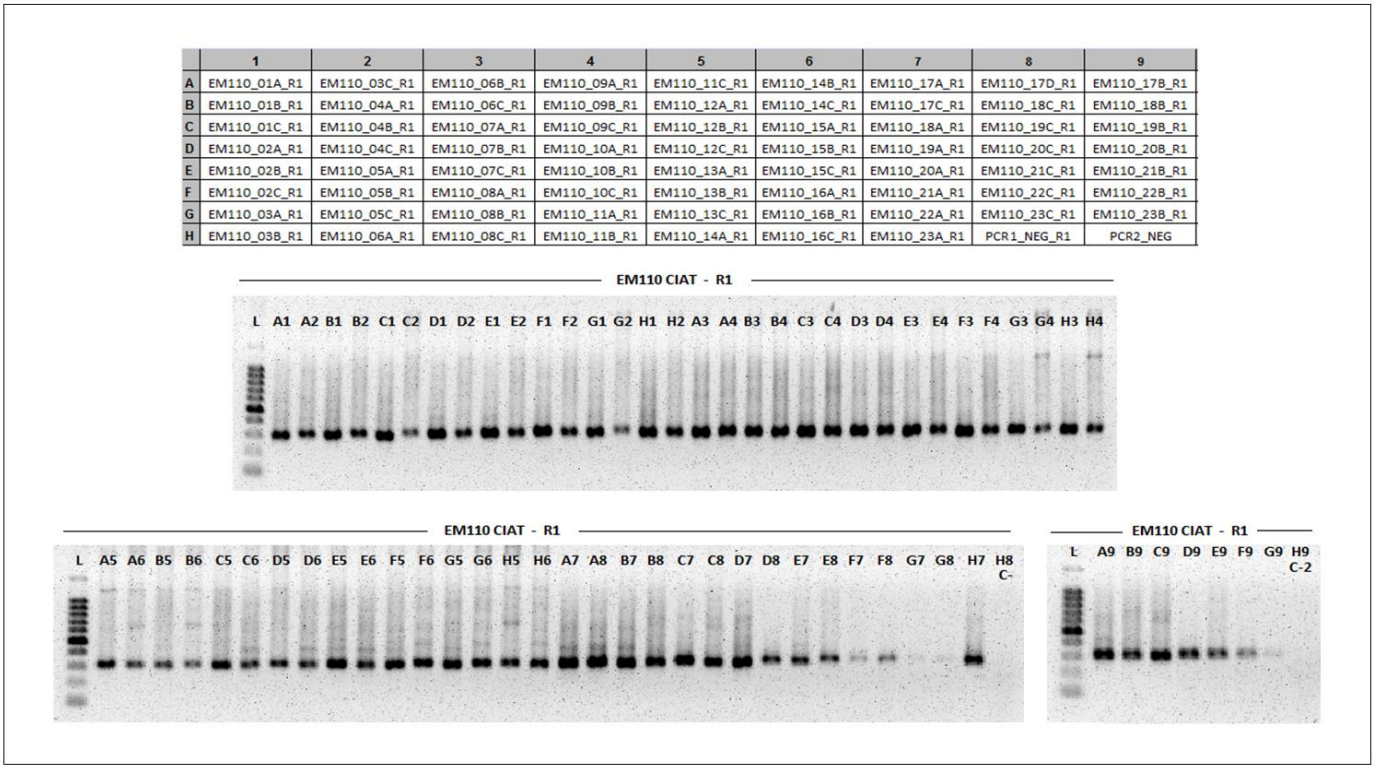

Figure 4. 1.5% agarose gel. Second PCR (PCR2) products generated after the insertion of the Illumina adaptors (P5; P7) and index sequences. The name of the samples refers to Table 1 (ID EcoMol column). (L) 100bp Ladder (Invitrogen); (EM110) EcoMol project name; (R1) insect_R1 barcode fragment; (C-) PCR1 negative control (C-2) PCR2 negative control

All samples successfully amplified after PCR2, although samples EM110_21A, 22A, 22C, 23B and 23C generated weaker amplifications (Figure 4), even after re-amplification (PCR1) of samples. We have not observed any amplification for PCR1 and PCR2 negative controls, indicating an absence of contamination among reactions. PCR2 products were then purified with magnetic beads (Agencourt AMPure XP® – Beckman Coulter), quantified using the Nanodrop 2000 spectrophotometer, normalized to 20ng/µL, and pooled together in a single tube.

By means of a real-time PCR, performed with the *KAPA Biossystems Quantification Kit* reagent (Illumina), the pool was quantified, diluted to a concentration of 2nM and again quantified to confirm the final concentration. The final solution was loaded onto the iSeq® equipment (Illumina), using the *iSeq 100 v2 300cycles* sequencing kit (2x150bp) and 30% phiX.

Bioinformatics and taxonomic assignments

The bioinformatic pipeline was organized in R (R Core Team 2022). Raw sequencing data was downloaded from *BaseSpace,* already demultiplexed into R1 and R2 read files. The FWD and REV sequences of each primer pair were detected in the reads in all possible orientations: Forward, complement, reverse & reverse complement. These primers sequences were removed from the raw reads using *Cutadapt* (Martin 2011). Then, using the *DADA2* package (Callahan 2016), cleaning procedures were carried out to remove undetermined bases (Ns) and keep only paired reads (corresponding R1 and R2). After this step, errors related to sequencing were also detected and used to identify unique R1 and R2 sequences. Corresponding R1 and R2 reads were merged by overlap (when possible) to produce merged sequences. These R1, R2 and merged sequences are called ASVs (Amplicon Sequencing Variants) and represent the complete or partial amplification products of the DNAs used in the library construction. Grouping of ASVs into OTUs was performed using SWARMv2 (Mahé *et al.*, 2015). ASVs present on Controls were matched to samples and signed as ***Possible contamination***. All ASVs were submitted to taxonomic identification using *BLASTn* against **NCBI nt** (**Nucleotide -** updated 11/2022). Using the DADA2 and P*hyloseq* packages (McMurdie and Holmes 2013), ASVs absolute and relative abundances on samples were calculated, separately for each *Read origin* class (R1, R2 and merged). Taxonomic ranks were automatically assigned using a **NCBI *Entrez Databases and Retrieval System***. ASVs belonging to Super kingdom *Bacteria* or *Archaea* were marked as ***Probable bacteria***.

Results

All the generated sequences, their abundances and the species identified according to the comparison with the sequences deposited in the reference databases are presented in the Excel file attached to this report.

We advise manual curation of the identification results before proceeding to further analyses. For this curation, one should consider:

1) **ASV size:** longer ASVs retrieve more robust identifications. Also, take into account the expected amplicon size for the R1, R2 and merged ASVs.
2) **Query coverage and Identity:** the closer to 100%, the more reliable is the identification. A summary of this metrics is calculated on **BLAST pseudo-score**, which corresponds to (query coverage * alignment identity * 100).
3) **Database representativeness:** Is the target biodiversity represented on the DB? High scoring or even perfect matches to alien species indicate that the native species could be present, but there is no reference information on the DB.
4) **Match score values and taxonomic groups:** For different taxa, a same marker region and primer sets (i.*e.,* COI-R1) can have different levels of sequence variability. There is no universal objective cutoff value to assign to Species, Genus or higher levels. These values can be defined to facilitate ecological analyses, but are experiment and target-taxa-specific.

References

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J., Holmes, S.P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 13(7): 581-583.

Gillet, F., Tiouchichine, M.-L. et al. (2015). A new method to identify the endangered Pyrenean desman (*Galemys pyrenaicus*) and to study its diet, using next generation sequencing from faeces. *Mammalian Biology - Zeitschrift Für Säugetierkunde*, 80, 505–509. https://doi. org/10.1016/j.mambio.2015.08.002

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ. Dec 10;3:e1420. doi: 10.7717/peerj.1420. PMID: 26713226; PMCID: PMC4690345.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17(1): 10–12. doi: 10.14806/ej.17.1.200.

McMurdie, P.J. and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One,* 8(4): e61217. doi: 10.1371/journal.pone.0061217.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Tournayre, O., Leuchtmann, M. et al. (2020). In silico and empirical evaluation of twelve metabarcoding primer sets for insectivorous diet analyses. *Ecology and Evolution,* 10, 6310–6332. https://doi.org/10.1002/ece3.6362

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole JR (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16): 5261-7526. doi: 10.1128/AEM.00062-07.