|  |  |
| --- | --- |
| **TITLE of SUB-AGREEMENT** | Subagreement ECOMOL CIAT under CALPSE (S180 |
| **SUB-AGREEMENT NUMBER** | C23PAL197 |
| **IMPLEMENTING PARTNER** | Ecomol Consultoria e Projetos LTDA-ME |
| **AUTHOR OF THE REPORT** | Priscilla M.S. Villela, Renata A. Miotto and Heron Hilário |
| **DATE SUBMITTED** | 15/09/2023 |
| **REPORTING DATES (DD/MM/YY-DD/MM/YY)** | 21/06/2023 – 15/09/2023 |
| **TYPE OF REPORT** | Technical Report |
| **ABSTRACT** (maximum 200 words) | This first report presents preliminary results of the project. It describes laboratory procedures for DNA extraction from the first 39 collected soil samples, construction of DNA metabarcoding library with specific primers for the identification of insect species, library next generation sequencing and bioinformatics. All the generated sequences, their abundances and the species identified according to the comparison with sequences in the reference databases are presented in the Excel file attached to this report. |
| **KEYWORDS** | Country/region: Amazon - Brazil  Crop(s):  Subject: Molecular identification of insect species in soil samples |

# 



**Identification of insects in soil samples through DNA metabarcoding**

First report - 2023

Piracicaba, SP, BR

August 2023

**Samples and laboratory procedures**

On 21th June 2023 and on 20th July 2023, EcoMol received 12 and 27 samples of soil, respectively (Table 1). Soil samples were stored in plastic bags (*ziplock*) containing 100g or 200g of silica (one or two 100g silica bags, depending on soil moisture). All samples contained approximately 40mL of soil, excluding samples F1, F2 and F3 from the first batch of samples, in which each contained only 20mL of soil.

**Table 1.** Soil samples and extracted DNA concentration (ng/µL) and purity (A260/A280, in nm unit) estimated in spectrophotometer Nanodrop. (EM135) EcoMol project name (c1) first field work (c2) second field work) (\*) name of samples after bioinformatics (\*\*) samples containing 20mL of soil, while the others contained 40mL.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **N** |  | **Sample/ID CIAT** | **Replicate** | **ID EcoMol\*** | **[ng/μL]** | **A260/280** |
| 1 | First field work | S1 | 1 | EM135c1\_S1 | 17,7 | 1,79 |
| 2 | S2 | 2 | EM135c1\_S2 | 16,4 | 1,79 |
| 3 | S3 | 3 | EM135c1\_S3 | 17,9 | 1,81 |
| 4 | V1 | 1 | EM135c1\_V1 | 11,1 | 1,73 |
| 5 | V2 | 2 | EM135c1\_V2 | 10,9 | 1,94 |
| 6 | V3 | 3 | EM135c1\_V3 | 11,8 | 1,8 |
| 7 | R1 | 1 | EM135c1\_R1 | 22,7 | 1,79 |
| 8 | R2 | 2 | EM135c1\_R2 | 20,1 | 1,76 |
| 9 | R3 | 3 | EM135c1\_R3 | 35,3 | 1,66 |
| 10 | F1\*\* | 1 | EM135c1\_F1 | 10,2 | 1,84 |
| 11 | F2\*\* | 2 | EM135c1\_F2 | 14,7 | 1,86 |
| 12 | F3\*\* | 3 | EM135c1\_F3 | 13,9 | 1,74 |
| 13 | Second field work | Tbio3\_1\_CF\_R1 | 1 | EM135c2\_Tbio3\_1\_CF\_R1 | 33,2 | 1,93 |
| 14 | Tbio3\_1\_CF\_R2 | 2 | EM135c2\_Tbio3\_1\_CF\_R2 | 33,6 | 2,01 |
| 15 | Tbio3\_1\_CF\_R3 | 3 | EM135c2\_Tbio3\_1\_CF\_R3 | 29,5 | 1,96 |
| 16 | Tbio3\_1\_I\_R1 | 1 | EM135c2\_Tbio3\_1\_I\_R1 | 11 | 1,92 |
| 17 | Tbio3\_1\_I\_R2 | 2 | EM135c2\_Tbio3\_1\_I\_R2 | 15,9 | 1,89 |
| 18 | Tbio3\_1\_I\_R3 | 3 | EM135c2\_Tbio3\_1\_I\_R3 | 14,6 | 1,86 |
| 19 | Tbio3\_1\_F\_R1 | 1 | EM135c2\_Tbio3\_1\_F\_R1 | 21,3 | 1,80 |
| 20 | Tbio3\_1\_F\_R2 | 2 | EM135c2\_Tbio3\_1\_F\_R2 | 15,9 | 2,18 |
| 21 | Tbio3\_1\_F\_R3 | 3 | EM135c2\_Tbio3\_1\_F\_R3 | 21,8 | 1,97 |
| 22 | Tbio3\_2\_I\_R1 | 1 | EM135c2\_Tbio3\_2\_I\_R1 | 15,7 | 1,9 |
| 23 | Tbio3\_2\_I\_R2 | 2 | EM135c2\_Tbio3\_2\_I\_R2 | 17,6 | 1,88 |
| 24 | Tbio3\_2\_I\_R3 | 3 | EM135c2\_Tbio3\_2\_I\_R3 | 12,5 | 1,68 |
| 25 | Tbio3\_2\_CF\_R1 | 1 | EM135c2\_Tbio3\_2\_CF\_R1 | 22,1 | 1,90 |
| 26 | Tbio3\_2\_CF\_R2 | 2 | EM135c2\_Tbio3\_2\_CF\_R2 | 13,5 | 1,88 |
| 27 | Tbio3\_2\_CF\_R3 | 3 | EM135c2\_Tbio3\_2\_CF\_R3 | 16,5 | 2,07 |
| 28 | Tbio3\_2\_F\_R1 | 1 | EM135c2\_Tbio3\_2\_F\_R1 | 12,1 | 1,93 |
| 29 | Tbio3\_2\_F\_R2 | 2 | EM135c2\_Tbio3\_2\_F\_R2 | 14,4 | 2,14 |
| 30 | Tbio3\_2\_F\_R3 | 3 | EM135c2\_Tbio3\_2\_F\_R3 | 15,2 | 2,06 |
| 31 | Tbio3\_3\_I\_R1 | 1 | EM135c2\_Tbio3\_3\_I\_R1 | 13,6 | 1,96 |
| 32 | Tbio3\_3\_I\_R2 | 2 | EM135c2\_Tbio3\_3\_I\_R2 | 7,9 | 2,26 |
| 33 | Tbio3\_3\_I\_R3 | 3 | EM135c2\_Tbio3\_3\_I\_R3 | 11,6 | 1,94 |
| 34 | Tbio3\_3\_F\_R1 | 1 | EM135c2\_Tbio3\_3\_F\_R1 | 17,6 | 1,82 |
| 35 | Tbio3\_3\_F\_R2 | 2 | EM135c2\_Tbio3\_3\_F\_R2 | 19,2 | 1,87 |
| 36 | Tbio3\_3\_F\_R3 | 3 | EM135c2\_Tbio3\_3\_F\_R3 | 37,5 | 1,63 |
| 37 | Tbio3\_3\_CF\_R1 | 1 | EM135c2\_Tbio3\_3\_CF\_R1 | 16,8 | 1,9 |
| 38 | Tbio3\_3\_CF\_R2 | 2 | EM135c2\_Tbio3\_3\_CF\_R2 | 18,6 | 1,89 |
| 39 | Tbio3\_3\_CF\_R3 | 3 | EM135c2\_Tbio3\_3\_CF\_R3 | 10,3 | 2,17 |

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 (A260/A280 ratio) were estimated using the Nanodrop 2000 spectrophotometer (ThermoScientific) (Table 1).

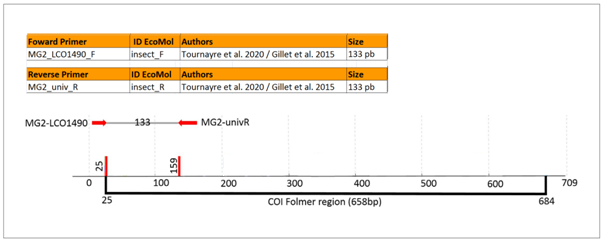
To identify insect species in the soil samples, we amplified (PCR1) a ~133bp fragment of *cytochrome oxidase I* subunit (COI gene) from the mitochondrial DNA (named here as insect\_R1 fragment), with MG2\_LCO1490\_F foward primer and the MG2\_univ\_R reverse primer (Gillet et al. 2015; Tournayre et al. 2020). Figure 1 presents this DNA barcode length and primers positions on the COI gene.

When visiting EcoMol on June 2023, Wendy Francesconi demonstrated interest in identifying mammal species in forest soil samples by means of DNA metabarcoding sequencing. To do this, we amplified F1, F2 and F3 DNA samples (together) with the 12v5\_Foward and 12SV5\_Reverse primers described by Riaz et al. (2011). These pair of primers amplify a ~98pb fragment of the 12S gene from the mitochondrial DNA (named here as 12SV5 fragment) and were designed to identify vertebrate species in general. This molecular marker is widely used in the scientific literature and is routinely used at EcoMol in its services.

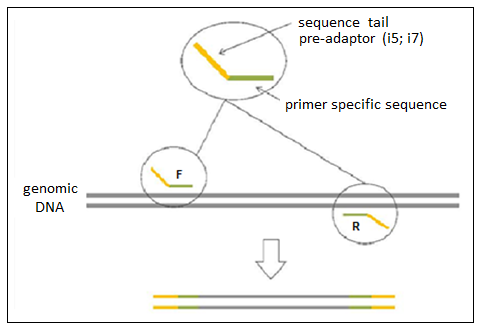
Both forward and reverse primer sequences (for insect\_R1 fragment and 12SV5 fragment amplification) received illumina i5 and i7 pre-adaptor tails to allow sequencing in the illumina plataform. Figure 2 presents a scheme of the amplification of a target gene with primers with foward and reverse illumina i5 and i7 pre-adaptor tails (PCR1).

We included negative controls in all PCR1 reactions. Amplification conditions for the insect\_R1 fragment 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. Amplification conditions for the 12SV5 fragment were the same as described in Riaz et al. (2011); each PCR1 reaction contained (20μL): 10X PCRBio Master Mix (PCR Biosystems Inc.), 0,3μM of each primer, 10μM of ‘BlockHomo’ primer (a blocking primer with a 3’ extremity modification against human DNA amplification) 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. PCR1 products are shown in Figure 3.

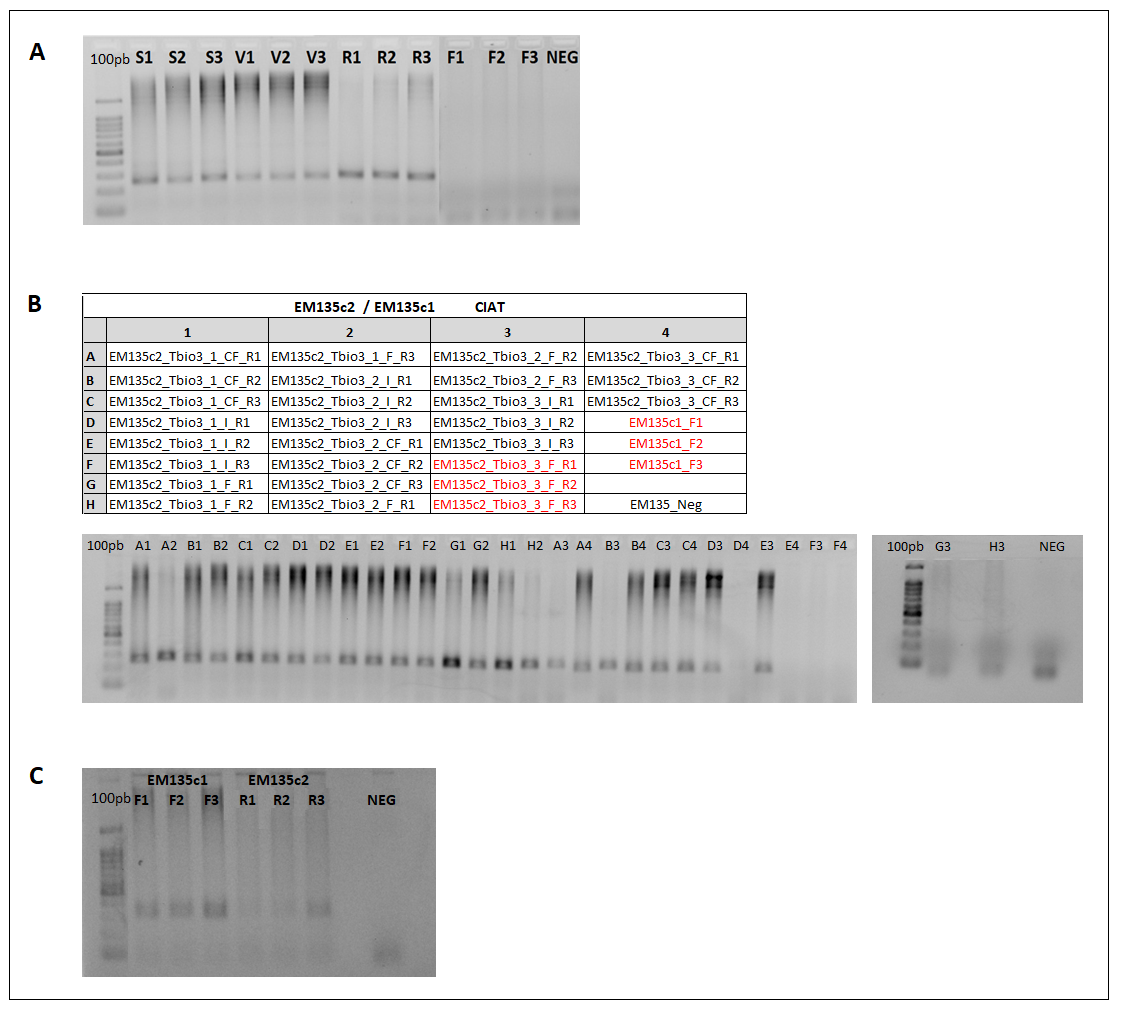
Amplifications of samples F1, F2 and F3 (first field work; EM135c1 in Table 1), and samples R1, R2 and R3 (second field work; EM35c2 in Table 1) failed after two tries (Figures 3A and 3B), so we purified the DNA with magnetic beads (Agencourt AMPure XP® – Beckman Coulter) to eliminate any inhibitors that may prevent amplification. After DNA purification, all the six samples have successfully amplified (Figure 3C).



**Figure 1.** DNA barcode length (insect\_R1 fragment) and primers positions on the COI gene. Modified from Tournayre et al. 2020.

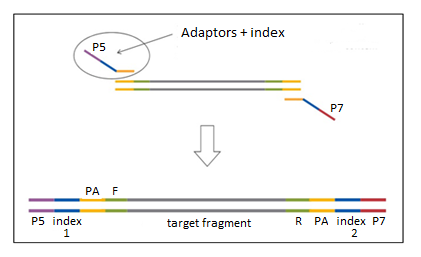


**Figure 2**. PCR1 scheme. Amplification of the target fragment using specific primers (green) with complementary sequences (pre-adapters tails) to the adapters of the illumina platform (orange). (F) forward primer; (R) reverse primer

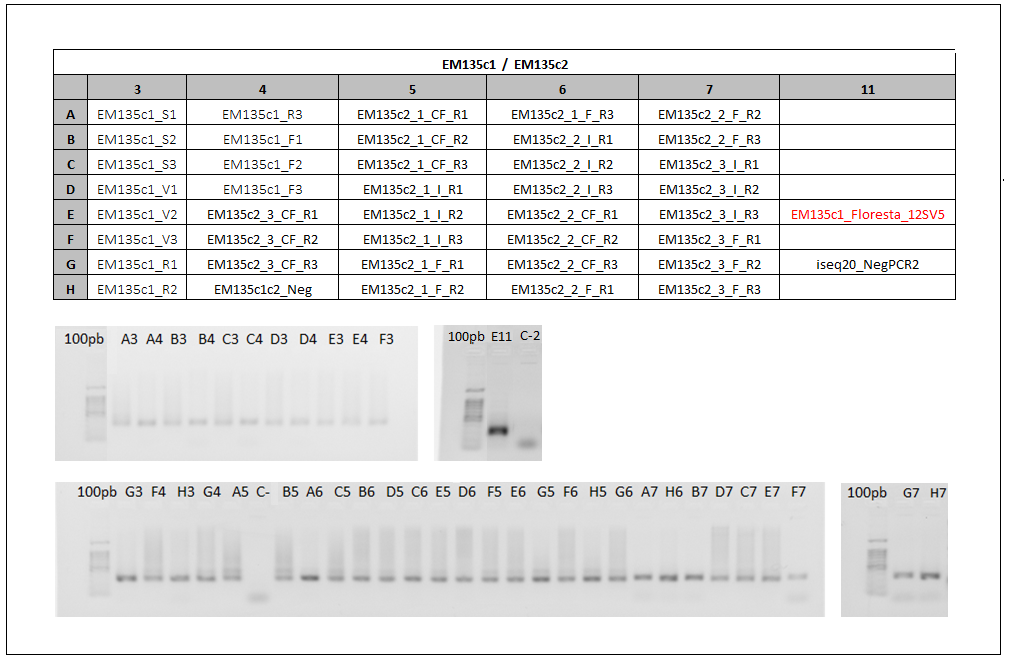


**Figure 3.** 1.5% agarose gel. First PCR (PCR1) products generated after amplification with specific primers (insect\_R1 fragment). (A) samples collected during the first field work, in which amplification of samples F1, F2 and F3 have failed. (B) samples collected during the second field work, in which amplifications of samples R1, R2 and R3 have also failed, as the second try of samples F1, F2 and F3. Legends in the agarose gel correspond to the position of the samples in the plate drawn above. (C) amplification after DNA purification with magnetic beads. In red: samples that went to DNA purification with magnetic beads. (100pb) 100bp Ladder (Invitrogen)

All PCR1 products were then 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 this reaction, illumina adaptors (P5 and P7) are used as primers, as shown in Figure 4. 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 5.



**Figure 4.** PCR2 scheme.Insertion of P5 and P7 adapters, which turns the set compatible with the illumina flow cell, and insertion of the index sequences (in blue), which allows the identification of each sample after sequencing. (F) forward primer; (R) reverse primer; (PA) pre-adaptors tails inserted in PCR1

****

**Figure 5**. 1.5% agarose gel. Second PCR (PCR2) products generated after the insertion of illumina adaptors (P5; P7) and index sequences. Legends in the agarose gel correspond to the position of the samples in the plate drawn above. (100pb) 100bp Ladder (Invitrogen); (EM135c1\_Floresta\_12SV5) 12SV5 fragment; (C- and EM135c1c2\_Neg) PCR1 negative control; (C-2 and iseq20\_NegPCR2) PCR2 negative control

All samples successfully amplified after PCR2. 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**

Raw sequencing data was downloaded from BaseSpace already demultiplexed into R1 (forward) and R2 (reverse) read files. The R1 and R2 sequences generated for each pair of primers were matched in all possible orientations: Forward, complement, reverse & reverse complement. By using the Cutadapt (Martin 2011) and DADA2 (Callahan 2016) programs, we carried out several cleaning procedures. The pipeline was organized in R (R Core Team 2020). Reads were submitted to the removal of sequences containing undefined bases (Ns), as well as the removal of the primer sequences. After this step, errors related to sequencing were also identified and removed. When possible, the remaining R1 and R2 sequences were combined by overlapping to generate merged sequences. From the remaining sequences of this quality control, unique sequences present in the samples were identified (R1, R2 and merged). These sequences are called ASVs (Amplicon Sequencing Variants) and represent the amplification products of the DNAs used in the construction of the library, which were submitted to taxonomic identification. Using the DADA2 and phyloseq packages (McMurdie and Holmes 2013), a count relative to its frequency in the library was assigned to each of the AVSs, as well as a classification based on the association with sequences from the NCBI and BOLD SYSTEMS public reference banks (https://www.ncbi.nlm.nih.gov/; https://www.boldsystems.org/). We made these associations using the classifier developed by Wang et al. (2007).

**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.

Before any ecological analysis, we suggest that researchers submit the identification results to a manual curation. In this curation, the following should be considered:

1) The length of the ASV: Longer ASVs retrieve more robust IDs. In addition, the expected amplicon size for R1, R2, and merged ASVs must be taken into account;

2) Coverage and identity: the closer to 100%, the more reliable the identification. A summary of these metrics was calculated into the BLAST 'pseudo-score', which corresponds to (query coverage \* alignment identity \* 100);

3) Database representativeness: is the target biodiversity represented in the database? High scores or even perfect matches with alien species indicate that native species may be present in the study area, but there is no reference information in the database;

4) Correspondence between score values and taxonomic groups: for different taxa, the same marker and primer sets may have different levels of sequence variability. There is no universal objective cutoff value to assign to species, genus, or higher taxonomic levels. These values can be defined to facilitate ecological analyses, but are generally experiment and target taxon specific. Thus, a manual curation of the results may consider the similarity between the sequences, the distribution of the species and, when possible, the probable or already confirmed occurrence in the study area. To define the final taxonomic classification at each sampling point, we suggest the following criteria:

* Species level: identity (sequence similarity with the reference sequence in the database) above 98% (ID > 98%) and relative abundance (number of reads generated in relation to the total generated for the sample) higher than 0.02 (Relative abundance > 2%);
* Genus level: sequence identity above 95% (ID > 95%);
* Family and Order Levels: sequence identity with the database reference sequence lower than 95% and higher than 80% (80% < ID < 95%).

**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

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 (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

# Riaz T et al. (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Research; 39(21): e145.

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.