

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

First report

Piracicaba, SP

July, 2022

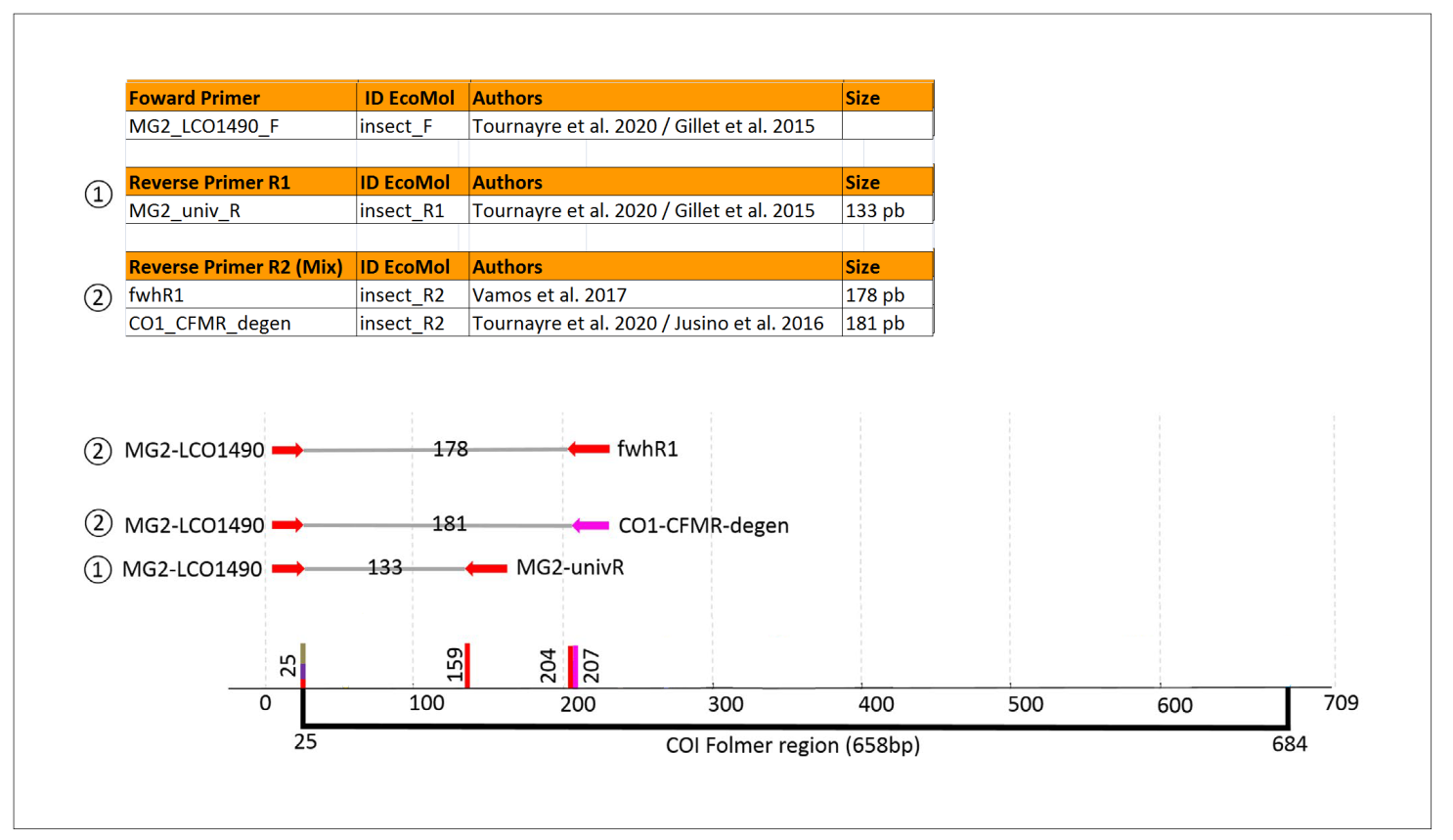
On 27th June 2022, EcoMol received 78 samples of soil from two distinct sampling sites/objectives: 54 samples for 'volume testing' and 24 samples from 'Horta' site (Table 1). Thirty nine soil samples were stored in plastic bags (ziplock) containing 100g of silica, and the other 39 samples were stored 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 (A260/A280 ratio) were estimated using the Nanodrop 2000 spectrophotometer (Thermo Scientific) (Table 1).

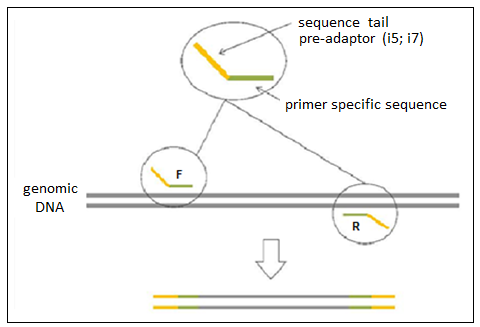
**Table 1.** Groups of soil samples ('volume testing' and 'Horta' site) and replicates, their volumes, storage method (silica or EcoMol preservation buffer), and concentration (ng/µL) and purity (A260/A280) estimated in spectrophotometer.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Volume Test** | | | | | | | |
| **N** | **Storage** | **Volume (300mL each)** | **Replicate** | **Replicate Vol.** | **EcoMol ID** | **[ng/μL]** | **A260/A280** |
| 1 | Silica | A | 1 | 40 mL | EM91S\_ A1 | 12.2 | 1.9 |
| 2 | Silica | A | 2 | 40 mL | EM91S\_A2 | 9.8 | 1.7 |
| 3 | Silica | A | 3 | 40 mL | EM91S\_A3 | 4.7 | 1.7 |
| 4 | Silica | B | 1 | 40 mL | EM91S\_B1 | 10.1 | 1.8 |
| 5 | Silica | B | 2 | 40 mL | EM91S\_B2 | 8.4 | 2 |
| 6 | Silica | B | 3 | 40 mL | EM91S\_B3 | 15.4 | 1.8 |
| 7 | Silica | C | 1 | 40 mL | EM91S\_C1 | 8.7 | 1.7 |
| 8 | Silica | C | 2 | 40 mL | EM91S\_C2 | 7.4 | 1.9 |
| 9 | Silica | C | 3 | 40 mL | EM91S\_C3 | 7.7 | 1.8 |
| 10 | Silica | D | 1 | 40 mL | EM91S\_D1 | 8.7 | 2 |
| 11 | Silica | D | 2 | 40 mL | EM91S\_D2 | 10.4 | 1.8 |
| 12 | Silica | D | 3 | 40 mL | EM91S\_D3 | 8.8 | 1.9 |
| 13 | Silica | E | 1 | 40 mL | EM91S\_E1 | 11.5 | 1.7 |
| 14 | Silica | E | 2 | 40 mL | EM91S\_E2 | 8.2 | 1.8 |
| 15 | Silica | E | 3 | 40 mL | EM91S\_E3 | 7.8 | 1.9 |
| 16 | Silica | F | 1 | 40 mL | EM91S\_F1 | 14.5 | 1.7 |
| 17 | Silica | F | 2 | 40 mL | EM91S\_F2 | 11.4 | 1.8 |
| 18 | Silica | F | 3 | 40 mL | EM91S\_F3 | 10.9 | 1.7 |
| 19 | Silica | G | 1 | 40 mL | EM91S\_G1 | 8.7 | 1.6 |
| 20 | Silica | G | 2 | 40 mL | EM91S\_G2 | 9 | 1.8 |
| 21 | Silica | G | 3 | 40 mL | EM91S\_G3 | 14.5 | 1.7 |
| 22 | Silica | H | 1 | 40 mL | EM91S\_H1 | 6.9 | 1.7 |
| 23 | Silica | H | 2 | 40 mL | EM91S\_H2 | 21.9 | 1.6 |
| 24 | Silica | H | 3 | 40 mL | EM91S\_H3 | 5.7 | 2 |
| 25 | Silica | I | 1 | 40 mL | EM91S\_I1 | 9.3 | 1.7 |
| 26 | Silica | I | 2 | 40 mL | EM91S\_I2 | 11.6 | 1.7 |
| 27 | Silica | I | 3 | 40 mL | EM91S\_I3 | 12.9 | 1.7 |
| 28 | Buffer | A | 1 | 40 mL | EM91b\_A1 | 31.9 | 1.7 |
| 29 | Buffer | A | 2 | 40 mL | EM91b\_A2 | 37.6 | 1.6 |
| 30 | Buffer | A | 3 | 40 mL | EM91b\_A3 | 28 | 1.7 |
| 31 | Buffer | B | 1 | 40 mL | EM91b\_B1 | 56.2 | 1.7 |
| 32 | Buffer | B | 2 | 40 mL | EM91b\_B2 | 53.3 | 1.7 |
| 33 | Buffer | B | 3 | 40 mL | EM91b\_B3 | 60.7 | 1.7 |
| 34 | Buffer | C | 1 | 40 mL | EM91b\_C1 | 34.9 | 1.7 |
| 35 | Buffer | C | 2 | 40 mL | EM91b\_C2 | 30.8 | 1.6 |
| 36 | Buffer | C | 3 | 40 mL | EM91b\_C1 | 32.3 | 1.7 |
| 37 | Buffer | D | 1 | 40 mL | EM91b\_D1 | 65.5 | 1.6 |
| 38 | Buffer | D | 2 | 40 mL | EM91b\_D2 | 71.6 | 1.7 |
| 39 | Buffer | D | 3 | 40 mL | EM91b\_D3 | 77.2 | 1.7 |
| 40 | Buffer | E | 1 | 40 mL | EM91b\_E1 | 41.8 | 1.6 |
| 41 | Buffer | E | 2 | 40 mL | EM91b\_E2 | 20.2 | 1.7 |
| 42 | Buffer | E | 3 | 40 mL | EM91b\_E3 | 34.3 | 1.6 |
| 43 | Buffer | F | 1 | 40 mL | EM91b\_F1 | 47.7 | 1.6 |
| 44 | Buffer | F | 2 | 40 mL | EM91b\_F2 | 54.3 | 1.6 |
| 45 | Buffer | F | 3 | 40 mL | EM91b\_F3 | 50.6 | 1.6 |
| 46 | Buffer | G | 1 | 40 mL | EM91b\_G1 | 38 | 1.7 |
| 47 | Buffer | G | 2 | 40 mL | EM91b\_G2 | 26.5 | 1.6 |
| 48 | Buffer | G | 3 | 40 mL | EM91b\_G3 | 42.2 | 1.6 |
| 49 | Buffer | H | 1 | 40 mL | EM91b\_H1 | 42 | 1.6 |
| 50 | Buffer | H | 2 | 40 mL | EM91b\_H2 | 45.6 | 1.7 |
| 51 | Buffer | H | 3 | 40 mL | EM91b\_H3 | 41.6 | 1.6 |
| 52 | Buffer | I | 1 | 40 mL | EM91b\_I1 | 32.2 | 1.6 |
| 53 | Buffer | I | 2 | 40 mL | EM91b\_I2 | 23.1 | 1.6 |
| 54 | Buffer | I | 3 | 40 mL | EM91b\_I3 | 30.1 | 1.6 |
| **Horta sampling** | | | | | | | |
| **N** | **Storage** | **Site (1.5L Each)** | **Replicate** | **Replicate Vol.** | **EcoMol ID** | **[ng/μL]** | **A260/A280** |
| 1 | Silica | Intervention / syntropic | 1 | 40 mL | EM91S\_Sy1 | 33.9 | 1.6 |
| 2 | Silica | Intervention / syntropic | 2 | 40 mL | EM91S\_Sy2 | 20 | 1.6 |
| 3 | Silica | Intervention / syntropic | 3 | 40 mL | EM91S\_Sy3 | 6.5 | 1.8 |
| 4 | Silica | Floresta / control | 1 | 40 mL | EM91S\_Co1 | 7.6 | 1.9 |
| 5 | Silica | Floresta / control | 2 | 40 mL | EM91S\_Co2 | 9.8 | 1.8 |
| 6 | Silica | Floresta / control | 3 | 40 mL | EM91S\_Co3 | 8.5 | 1.7 |
| 7 | Silica | Vizinho / Counterfactual | 1 | 40 mL | EM91S\_CF1 | 16.6 | 1.7 |
| 8 | Silica | Vizinho / Counterfactual | 2 | 40 mL | EM91S\_CF2 | 30.5 | 1.6 |
| 9 | Silica | Vizinho / Counterfactual | 3 | 40 mL | EM91S\_CF3 | 11.5 | 1.9 |
| 10 | Silica | Restauration | 1 | 40 mL | EM91S\_Re1 | 7.1 | 1.8 |
| 11 | Silica | Restauration | 2 | 40 mL | EM91S\_Re2 | 14.2 | 1.6 |
| 12 | Silica | Restauration | 3 | 40 mL | EM91S\_Re3 | 6.9 | 1.5 |
| 13 | Buffer | Intervention / syntropic | 1 | 40 mL | EM91b\_Sy1 | 44.3 | 1.7 |
| 14 | Buffer | Intervention / syntropic | 2 | 40 mL | EM91b\_Sy2 | 42.8 | 1.7 |
| 15 | Buffer | Intervention / syntropic | 3 | 40 mL | EM91b\_Sy3 | 30.4 | 1.6 |
| 16 | Buffer | Floresta / control | 1 | 40 mL | EM91b\_Co1 | 39.7 | 1.7 |
| 17 | Buffer | Floresta / control | 2 | 40 mL | EM91b\_Co2 | 38.1 | 1.7 |
| 18 | Buffer | Floresta / control | 3 | 40 mL | EM91b\_Co3 | 38.5 | 1.7 |
| 19 | Buffer | Vizinho / Counterfactual | 1 | 40 mL | EM91b\_Cf1 | 47.5 | 1.6 |
| 20 | Buffer | Vizinho / Counterfactual | 2 | 40 mL | EM91b\_Cf2 | 45 | 1.6 |
| 21 | Buffer | Vizinho / Counterfactual | 3 | 40 mL | EM91b\_Cf3 | 40.5 | 1.6 |
| 22 | Buffer | Restauration | 1 | 40 mL | EM91b\_Re1 | 77.1 | 1.5 |
| 23 | Buffer | Restauration | 2 | 40 mL | EM91b\_Re2 | 59.4 | 1.6 |
| 24 | Buffer | Restauration | 3 | 40 mL | EM91b\_Re3 | 61.4 | 1.6 |

To identify insect species in the soil samples, we amplified (PCR1) two portions of the *cytochrome oxidase I* subunit (COI gene) from the mitochondrial DNA with two sets of primers, as described by Tournayre et al. (2020). For the first barcode amplification, we used the MG2\_LCO1490\_F foward primer and the MG2\_univ\_R reverse primer (Gillet et al. 2015; Tournayre et al. 2020) to amplify a ~133bp fragment of the COI gene (named here insect\_R1 fragment); for the -second barcode amplification, we used the same MG2\_LCO1490\_F foward primer and a mix containing the fwhR1 reverse primer (Vamos et al. 2017) together with the CO1\_CFMR\_degen reverse primer (Jusino et al. 2016; Tournayre et al. 2020) to amplify a ~180pb fragment of the COI gene (named here insect\_R2 fragment). These set of primers were chosen because they, apparently, successfully amplify a higher number of insect orders than other primer sets designed to amplify COI barcodes (Tournayre et al. 2020). All primer sequences received foward or reverse illumina i5 and i7 pre-adaptor tails to allow sequencing in the illumina plataform. Figure 1 presents DNA barcode lengths and primer positions on the COI gene. Figure 2 presents a scheme of the amplification of the COI gene with primers with foward and reverse illumina i5 and i7 pre-adaptor tails (PCR1). We included positive and 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 (F and R1 or R2) 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.

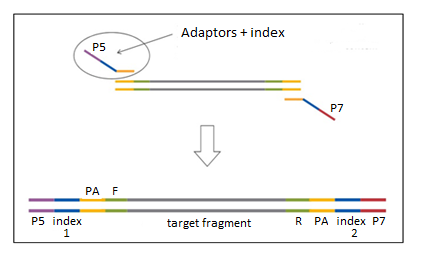


**Figure 1.** DNA barcode lengths and selected primer positions on the COI gene. (1) and (2) are the two sets of primers used. Modified from Tournayre et al. 2020.

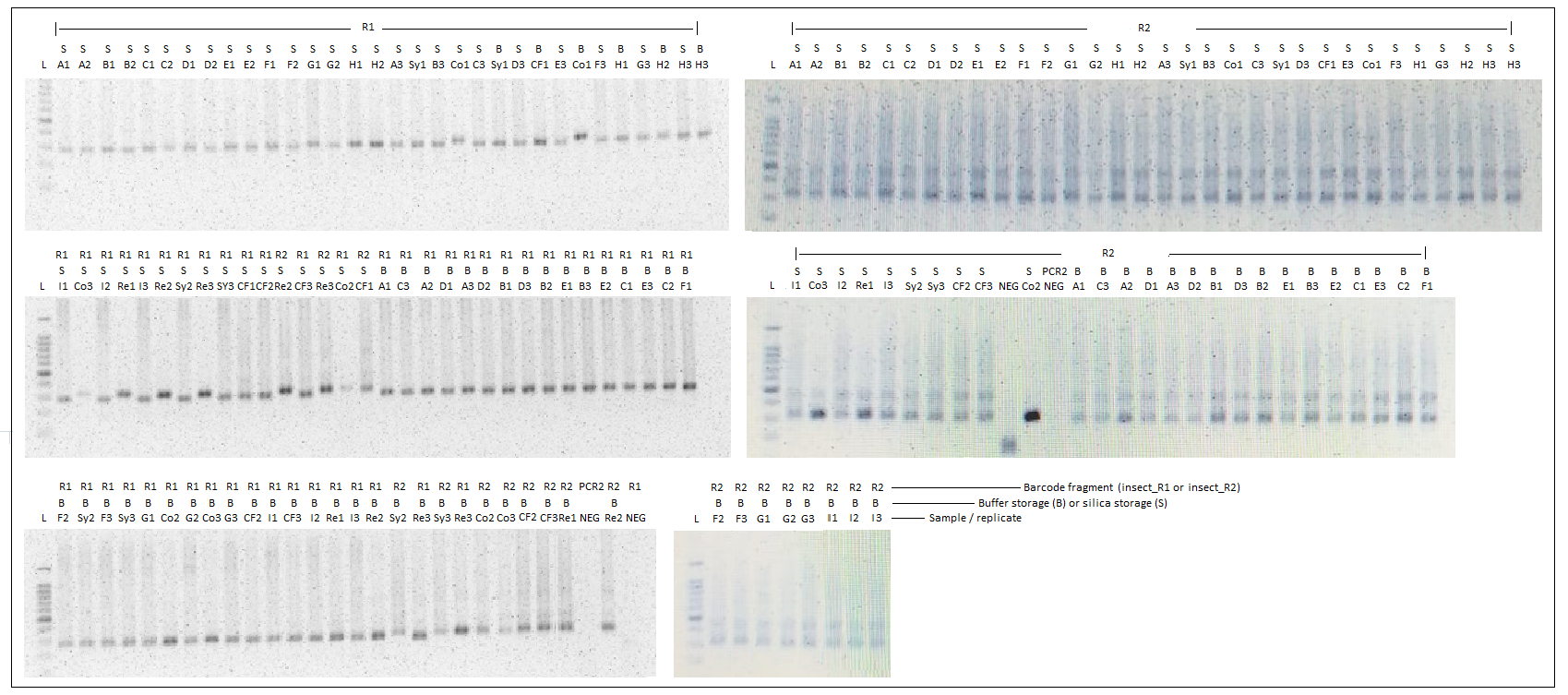


**Figure 2**. PCR1 scheme. Amplification of the target fragment of the COI gene using specific primers (green) with complementary sequences (pre-adapters) to the adapters of the illumina platform (orange). (F) forward; (R) reverse

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. (F) forward; (R) reverse; (PA) pre-adaptors inserted in PCR1



**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 samples refers to Table 1 (ID EcoMol column). (L) 100pb Ladder (Invitrogen); (B) Buffer storage; (S) Silica storage; (R1) insect\_R1 barcode fragment; (R2) insect\_R2 barcode fragment; (R1 NEG) PCR1 negative control for insect\_R1 barcode fragment; (R2 NEG) PCR1 negative control for insect\_R2 barcode fragment; (PCR2 NEG) PCR2 negative control

All samples successfully amplified after PCR2 and 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 20% phiX.

**Bioinformatics and taxonomic assignments**

Raw sequencing data was downloaded from BaseSpace already demultiplexed into Forward and Reverse read files. The Forward and Reverse 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. From the remaining sequences of this quality control, unique sequences present in the samples were identified. 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.

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%);
* Gender 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

Jusino, M. A., Banik, M. T. et al. (2018). An improved method for utilizing high-throughput amplicon sequencing to determine the diets of insectivorous animals. *Molecular Ecology Resources*, *19*(1), 176–190. https://doi.org/10.1111/1755-0998.12951

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/

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

Vamos, E., Elbrecht, V., Leese, F. (2017). Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcoding and Metagenomics*, 1, e14625. https://doi.org/10.3897/mbmg.1.14625

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.