Osorio *et al*., 2021. The corixid *Sigara (Subsigara) distincta* (Fieber, 1848) in the Pyrenees: a first record for Spain and an unsolved taxonomic puzzle. Limnetica 40-2, 2021: 375-384

SUPPLEMENTARY INFORMATION

DNA METABARCODING METHODS

Molecular laboratory

Morphologically identified and counted macroinvertebrate specimens from Lake Vilac were included in an extensive DNA metabarcoding study. When extracting complete bulk samples, large biomass rich specimens are expected to contribute significantly more DNA to the final bulk DNA isolate than small organisms with little biomass. Therefore, in order to deal with the biomass bias, we segregated taxa into two size groups and prepared bulk macroinvertebrate isolates as follows. Taxa smaller than 0.5 cm were included as whole individuals in the bulk isolate; in contrast, a pinch of tissue from the abdomen was taken from taxa larger than 0.5 cm. The resulting final isolates included organism tissue coming from at least 75% of total specimens of each taxon found in samples, that were dried overnight at room temperature on sterile filter paper sheets to remove the ethanol. Then, isolates were ground and homogenized in a mortar. Between homogenisation of each sample, the mortar and pestle were cleaned using deionized water and any remaining material was burnt off using methanol to prevent cross-contamination of samples.

Around 25 mg of homogenized powder was removed from each sample for genomic DNA analysis, and the remaining homogenized fractions were stored in a freezer at -80 ºC. The 25 mg subsample was isolated using DNeasy Blood and Tissue Kit (Qiagen, Valencia, USA) according to the manufacturer's directions. We amplified the Leray fragment of ca. 313 bp

using the primers jgHCO2198 5′-TANACYTCNGGRTGNCCRAARAAYCA-3′ (Geller *et al.*, 2013) and mlCOIintF 5′-GGWACWGGWTGAACWGTWTAYCCYCC-3′ (Leray *et al.*, 2013). Amplification of COI used AccuPower Hot start PCR PreMix (Bioneer, Seoul, South Korea), with 1 µl of each 5 µM forward and reverse 8-base tagged primers, and 1 µl of purified DNA in a total volume of 20 µl per sample. The PCR thermal regime consisted of one cycle of 10 min at 94 °C; 35 cycles of 1 min at 94 °C; 1.5 min at 50 °C; 1 min at 72 °C and a final cycle of 5 min at 72 °C in a Mastercycler (Eppendorf, Hamburg, Germany). After PCR, the quality of amplifications was assessed by electrophoresis in agarose gels. All PCR products were purified using Labopass Gel Extraction kit (Cosmogenetech, Seoul, South Korea). The final products were normalized and pooled using the PicoGreen, and the size of libraries was verified using the LabChip GX HT DNA High Sensitivity Kit (PerkinElmer, Massachusetts, USA). Samples were sequenced on an Illumina MiSeq platform (2 × 300 bp paired-end run) in a single multiplexed run, along with samples from a related research project, which were not considered in this study.

Bioinformatics

Sequence data were first demultiplexed and processed with the R wrapper script JAMP v0.60 (https://github.com/VascoElbrecht/JAMP) (Elbrecht, Vamos, Meissner, Aroviita, & Leese, 2017) on R v3.5.1 (R Core Team, 2018). Reads were paired-end merged using Usearch v11.0.667 (Edgar, 2010), and primer sequences were subsequently trimmed using Cutadapt v1.17 and default settings (Martin, 2011). Reads deviating by more than 10 bp from the expected 313 amplicon length were discarded. Usearch (Edgar & Flyvbjerg, 2015) was used to remove reads with an expected error probability of 1 or higher, and to dereplicate sequences. Then, singletons were removed, and sequences with ≥97% similarity were clustered into Molecular Operational Taxonomic Units (MOTUs) using UPARSE, that includes chimera

removal (Edgar, 2013). Pre-processed reads from all samples were dereplicated again and matched against the MOTUs with a minimum match of 97% (including singletons in this step). MOTU table obtained from MiSeq lane was processed for cross-talk detection and filtering using a custom R script based on UNCROSS2 (Edgar, 2018), which preceded an additional post-clustering curation step for removing erroneous MOTUs (LULU method) (Frøslev *et al*., 2017). Taxonomy was assigned to the remaining MOTUs running BLAST+ v2.7.1 (Camacho *et al*., 2009) on a custom DNA reference database built combining all public metazoan COI sequences from NCBI GenBank and BOLD (Ratnasingham & Hebert, 2007; Benson *et al*., 2018) (downloaded in September 2018). Conflicting taxonomy was resolved on a case-by-case basis, and a coarser taxonomic level was preferred if there was no evident correct assignment.

SUPPLEMENTARY REFERENCES

BENSON, D. A., M. CAVANAUGH, K. CLARK, I. KARSCH-MIZRACHI, J. OSTELL, K. D. PRUITT & E. W. SAYERS. 2018. GenBank. *Nucleic Acids Research*, 46 (D1): D41-D47. DOI: 10.1093/nar/gkx1094.

CAMACHO, C., G. COULOURIS, V. AVAGYAN, N. MA, J. PAPADOPOULOS, K. BEALER & T. L. MADDEN. 2009. BLAST+: architecture and applications. *BMC Bioinformatics*, 10 (1): 421. DOI: 10.1186/1471-2105-10-421.

EDGAR, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26 (19): 2460-2461. DOI: 10.1093/bioinformatics/btq461.

EDGAR, R. C. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, 10 (10): 996. DOI: 10.1038/nmeth.2604.

EDGAR, R. 2018. UNCROSS2: identification of cross-talk in 16S rRNA OTU tables. *bioRxiv* DOI: 10.1101/400762 (preprint).

EDGAR, R. & H. FLYVBJERG. 2015. Error filtering, pair assembly and error correction for next-

generation Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, 31 (21): 3476-3482. DOI: 10.1093/bioinformatics/btv401.

ELBRECHT, V., E. E. VAMOS, K. MEISSNER, J. AROVIITA & F. LEESE. 2017. Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, 8 (10): 1265-1275. DOI: 10.1111/2041-210x.12789.

FRØSLEV, T. G., R. KJØLLER, H. H. BRUUN, R. EJRNÆS, A. K. BRUNBJERG, C. PIETRONI & A. J. HANSEN. 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, 8 (1): 1188. DOI: 10.1038/s41467-017-01312-x.

GELLER, J., C. MEYER, M. PARKER, & H. HAWK. 2013. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources*, 13 (5): 851-861. DOI: 10.1111/1755-0998.12138.

LERAY, M., J. Y. YANG, C. P. MEYER, S. C. MILLS, N. AGUDELO, V. RANWEZ, J. T. BOEHM & R. J. MACHIDA. 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, 10 (1): 34. DOI: 10.1186/1742-9994-10-34.

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.

RATNASINGHAM, S. & P. D. HEBERT. 2007. BOLD: The Barcode of Life Data System (http://www. barcodinglife.org). *Molecular Ecology Resources*, 7 (3): 355-364. DOI: 10.1111/j.1471-8286.2007.01678.x.

R CORE TEAM. 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna. Austria. URL: https://www.R-project.org/.