bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NUMT dumping: validated removal of nuclear pseudogenes from mitochondrial metabarcode data

Andújar, C[1]., Creedy, T. J.[2], Arribas, P[1]., López, H[1]., Salces-Castellano, A[1]., Pérez-Delgado, A[1]., Vogler, A. P.[2,3] & B. C. Emerson[1]

1. Island Ecology and Evolution Research Group, Institute of Natural Products and Agrobiology (IPNA-CSIC), San Cristóbal de la Laguna, Spain
2. Department of Life Sciences, Natural History Museum, London, UK
3. Department of Life Sciences, Imperial College London, Ascot, UK

* Corresponding author: Carmelo Andújar. Email: candujar@ipna.csic.es; candujar@um.es;

Key words: Metazoa, NUMT, pseudogene, spurious sequences, denoising, NGS, HTS, intraspecific variation, taxonomic inflation.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

1. Metabarcoding of Metazoa using mitochondrial genes is confounded by the co-amplification of mitochondrial pseudogenes (NUMTs). Current denoising protocols have been designed to remove PCR and sequencing artefacts, but pseudogenes are not usually recognised by these procedures. Authentic mitochondrial amplicon sequence variants (ASVs), which represent the majority of reads, can be distinguished from PCR-derived errors, sequencing errors and NUMTs (non-authentic ASVs) due to their lower abundances. However, the use of simple read abundance thresholds is complicated by the highly variable DNA contribution of individuals in a metabarcoding sample.

2. We show how ASVs that survive standard denoising, but are identified as non-authentic, are consistent with expectations for NUMTs with regard to patterns of phylogenetic relatedness, read-abundance, and library co-occurrence. We then propose and demonstrate a new self-validating framework, named NUMT dumping, which allows NUMT filtering strategies to be evaluated by quantifying (i) the prevalence of non-authentic ASVs (NUMT and erroneous sequences) and (ii) the collateral effects on the removal of authentic ASVs (mtDNA haplotypes) in filtered data. We propose several filtering strategies within the NUMT dumping framework, based on the application of read-abundance thresholds, structured with regard to sequence library and phylogeny.

3. The framework was validated using mock and natural communities, both of which showed opposing trends for the removal of authentic and non-authentic ASVs, when threshold values for minimum abundance to filter out sequences were increased. Filtering can be optimized to retain less than 5% of non-authentic ASVs while retaining more than 89% of authentic mitochondrial ASVs, or complete removal of non-authentic ASV with 77% of authentic mitochondrial ASVs retained.

4. We provide a program, *NUMTdumper*, that can be used to evaluate and decide upon the most adequate metabarcoding filtering strategy for specific research objectives, providing a measure of expected prevalence of non-authentic ASVs in metabarcoding datasets. In addition, this evaluation

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

allows the user to quantify effects of taxonomic inflation when ASVs are clustered into OTUs. It improves the reliability of intraspecific genetic information derived from metabarcode data, opening the door for community-level genetic analyses requiring haplotype-level resolution.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 | INTRODUCTION

Bulk DNA amplification and high-throughput sequencing (HTS) of biological samples, known as metabarcoding (Taberlet *et al.* 2012), is becoming an established tool for the study of biodiversity (e.g., Hamady, Walker, Harris, Gold, & Knight, 2008; Yu et al., 2012). However, sequencing noise, inherent to metabarcoding, blurs the precision of metabarcoding to quantify intraspecific genetic variation. Consequently, sequencing reads are frequently grouped into Operational Taxonomic Units (OTUs) that broadly represent species, thus collapsing the fine-scale variation that would be needed to assess intraspecific diversity. Tools for filtering spurious sequences from authentic genomic sequences present in a sample, i.e. amplicon sequence variants (ASVs; Callahan, McMurdie & Holmes 2017), potentially allow direct analysis without the need for OTU clustering, resulting in improved resolution, reusability and reproducibility of metabarcoding data (Callahan, McMurdie & Holmes 2017). However, while existing denoising approaches efficiently remove sequences arising from PCR artefacts, sequencing error and chimera formation, many spurious sequences remain (e.g., Elbrecht *et al.* 2018).

Denoising methods such as UNOISE (Edgar 2016), DADA2 (Callahan *et al.* 2016) and Deblur (Amir *et al.* 2017) remove sequencing errors, assuming that erroneous sequences are very similar to authentic haplotypes, but showing lower abundances and/or lower quality scores of base calls. The use of these methods for haplotype level community analyses has been implemented primarily in bacterial metabarcoding pipelines (e.g., Caruso *et al.* 2019), although erroneous sequences partly persist even after denoising (Nearing *et al.* 2018). In Metazoa, ASV-level metabarcoding has been used for haplotype matching against reference sequences (e.g. Corse et al., 2017; Thomsen & Sigsgaard, 2019). Only one study to date explored the use of ASVs for community analysis but revealed a high proportion of unexpected sequences in mock communities with known haplotype composition (Elbrecht *et al.* 2018). Similarly, a higher than expected number of OTUs (so-

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

called OTU inflation) is common in metazoan metabarcoding (e.g., Flynn *et al.* 2015; Clare *et al.* 2016), even if denoising is followed by filtering and clustering steps (e.g., Andújar et al., 2018).

The unexpected diversity in denoised metabarcode datasets may be derived from mitochondrial insertions into the nuclear genome, i.e. "nuclear mitochondrial" sequences (NUMT; Lopez *et al.* 1994) as has been recently suggested (Elbrecht *et al.* 2018; Liu *et al.* 2019). NUMTs have been found in most eukaryotes (Bensasson *et al.* 2001; Richly & Leister 2004), and insertions may occur repeatedly and thus accumulate in the nuclear genome throughout the evolutionary history of a lineage (e.g., Hazkani-Covo, Sorek & Graur 2003; Pons & Vogler 2005). For example, more than 700 NUMTs have been identified in the human genome, ranging in size from a mere 38 bp to nearly-complete mitogenomes (Ramos *et al.* 2011), with some insertions estimated to have occurred as much as 58 million years ago (Bensasson, Feldman & Petrov 2003). Once inserted into the nuclear genome, rates of nucleotide substitution, insertion and deletion may be relatively high due to the absence of selective forces, while placement in genomic regions characterised by low mutation rates may result in the retention of pre-insertion ancestral states (Bensasson *et al.* 2001). Thus, some NUMTs will be easily detected based on frameshifts, in-frame stop codons or mutational patterns inconsistent with functional genes, but others will have no obvious features to distinguish them from a mitochondrial copy, as has been documented in DNA barcoding studies (Song *et al.* 2008; Shokralla *et al.* 2014; Creedy *et al.* 2020). Among the latter, some will be minor variants of authentic mitochondrial sequences, thus resembling sequencing errors that may be removed by existing denoising software, or clustered within OTUs when intraspecific variation is not the object of investigation. However, other divergent NUMTs will retain their functional structure as ancestrally "frozen" pseudogenes (Bensasson et al., 2001) and result in an overestimation of species-level entities and haplotype diversity in DNA barcoding and metabarcoding studies (Song *et al.* 2008).

As a first step to quantify the problem of NUMT co-amplification in metabarcoding, we use several lineages of arthropods to establish the patterns of co-occurrence, relative abundance, and

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phylogenetic relatedness of ASVs that may constitute a mixture of authentic mitochondrial haplotypes and their associated NUMTs. This analysis suggested the possibility to apply certain rules for filtering NUMTs based on read abundance relative to presumed authentic haplotypes found in the same sample or in clades of close relatives. This is exploited in a framework for optimal selection of ASV filtering, which we call NUMT dumping, which weighs the removal of presumed NUMTs and erroneous sequences against the undesirable removal of authentic haplotypes. Modifications in the criteria for filtering will alter these proportions, facilitating the identification of optimal parameters according to the characteristics of target taxa, target genes, and research objectives.

To facilitate NUMT dumping, we provide the *NUMTdumper* software that integrates several read-abundance filtering strategies designed to eliminate NUMTs. Copy number of NUMTs integrated in the nuclear genome is lower than mitochondrial genomes by a factor of 100 to 10000, depending on taxa, cell type and tissue (Bogenhagen 2012; Quiros *et al.* 2017), which should be reflected in lower relative abundances of sequence reads. However, in complex metabarcoding mixtures setting high threshold values for the minimum number of reads required to retain a particular gene copy carries the risk that rare but authentic mitochondrial haplotypes in the mixture may also be removed. Vice versa, a conservative threshold may be too low, such that many spurious sequences are not removed. Furthermore, different phylogenetic lineages may show different rates of relative NUMT incidence, so a particular threshold may remove NUMTs from one clade, while removing authentic sequences from another.

NUMTdumper integrates several read-abundance filtering strategies designed to balance the competing demands of removal of NUMTs and retention of mitochondrial reads by applying abundance thresholds not over an entire dataset, but to the sequences within individual ecological samples (frequently represented by individual sequencing libraries) and/or within lineages defined by sequence similarity or taxonomic ranks, and providing a measure of prevalence of both authentic and non-authentic sequences in the final dataset. We propose that *NUMTdumper* should be of general

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

utility to metazoan metabarcoding, allowing comprehensive evaluation and validation of optimal filtering strategies and thresholds according to specific research objectives.

2 | MATERIALS AND METHODS

2.1 | The NUMT dumping framework

The NUMT dumping rationale is based on the prevalence of NUMTs and erroneous sequences versus authentic haplotypes, which is evaluated for a range of NUMT filtering strategies (and a range of parameters or threshold values for each strategy). Filtering strategies can be designed to consider read-abundance by dataset, library, or lineage (as a proxy for phylogenetic relatedness). This evaluation allows users to decide upon the most adequate metabarcoding filtering parameters while providing a measure of the expected number of NUMTs and erroneous sequences in the final dataset. It comprises two main steps:

1. Classification of ASVs. Our starting point is a dataset of ASVs, each of which should exclusively belong to one of two categories: (i) authentic ASVs (a-ASVs) that correspond to actual mitochondrial copies, and (ii) non-authentic ASVs (na-ASVs) that are either sequencing artefacts or NUMTs. Evaluation of NUMT dumping performance relies on the ascertainment of a subset of ASVs that can be confidently considered as authentic sequences (designated as verified-authentic-ASVs; va-ASVs) or non-authentic (designated as verified-non-authentic-ASVs; vna-ASVs). va-ASVs can be identified by BLAST searches against reference databases requiring a perfect match (-perc_identity 100), under the assumption that references are valid mtDNA COI sequences. Any ASV that includes indels and/or stop codons in the translation rendering it non-functional is designated as a vna-ASV. We consider all other ASVs to be unclassified-ASVs (u-ASVs). Improvements in the confident classification of ASVs as va-ASVs or vna-ASVs, such as the availability of improved reference databases, will benefit NUMT dumping.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. Data filtering and evaluation of filtering performance. ASV datasets are subject to filtering procedures with the aim of removing na-ASVs while retaining a-ASVs. Filtering automatically generates data on the survival of va-ASV and vna-ASV. The application of a given filtering criterion (e.g. minimum number of reads by library required to retain an ASV) for a range of values for filtering parameters allows trends for the survival of va-ASV and vna-ASV to be analysed. Here we implement five different filtering strategies based on read abundances with increasing threshold values, and also evaluate the synergistic effect of using two simultaneously (see below). Additional methods and more complex models can potentially be incorporated into and evaluated within the NUMT dumping framework.

In addition to obtaining values for the survival of va-ASV and vna-ASV, for each filtering exercise the number of (i) a-ASVs in the initial ASV dataset; (ii) surviving a-ASVs in the filtered dataset; (iii) na-ASVs in the initial dataset, and; (iv) surviving na-ASVs in the filtered dataset can be estimated. These estimations are made from the known values of (i) the number of ASVs in the initial dataset and the number of ASVs retained after filtering and (ii) the proportion of retained va-ASVs and vna-ASVs, under the assumption that va-ASVs and vna-ASVs are a representative subset of all a-ASVs and na-ASVs respectively in the initial dataset. This assumption implies that (i) the ratio between the number of va-ASVs before and after filtering will be similar and can be extrapolated to the ratio between a-ASVs before and after filtering, and (ii) the ratio between the number of vna-ASVs before and after filtering will be similar and can be extrapolated to the ratio between all na-ASVs before and after filtering.

Then, we can estimate the total number of surviving a-ASVs a using the formula:

$$a \approx \frac{T - \frac{N_v}{n_v} t}{\frac{A_v}{a_v} - \frac{N_v}{n_v}}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Where A_v and a_v are the numbers of initial and surviving va-ASVs, N_v and n_v are the numbers of initial and surviving vna-ASVs and T and t are the total number of initial and surviving ASVs. From this, we can also calculate the estimated total number of surviving na-ASVs n and the total number of initial a-ASVs and na-ASVs (A and N) (formula derivation in Supplementary Materials).

To facilitate the application of NUMT dumping, the workflow has been implemented in *NUMTdumper*, written in Python3, with some aspects in R. The python code draws heavily on the *BioPython* libraries, as well as *numpy*, *itertools*, *csv* and *argparse*. The R code makes use of libraries *ape* (Paradis, Claude & Strimmer 2004), *phangorn* and *getopt*. The current version of the software (v1.0) allows the user to (i) classify ASVs as va-ASVs and vna-ASVs, (ii) apply the five alternative filtering criteria we provide with customized parameter ranges, facilitating the analysis of trends in survival of va-ASVs and vna-ASVs, and (iii) estimate the number of surviving a-ASV and na-ASVs for each filtering criterion and parameter range explored, using the formula and assumptions described above. Further details on the application of the software are provided in the Supplementary Materials and software tutorial. The software is available at https://github.com/tjcreedy/*NUMTdumper*.

2.2 | Empirical application of *NUMTdumper*

2.2.2 | Datasets

Three existing *COI* metabarcoding datasets were used, originating from: (i) 780 individually metabarcoded bees (BEE dataset; Creedy *et al.* 2020) which were used for current purposes to generate 50 *in silico* mock communities of 100 individuals drawn from a subset of 462 confidently identified specimens; (ii) 94 Coleoptera communities from soil samples from the island of Tenerife (COL dataset) (Andújar *et al.* in prep), and (iii) 48 communities of Coleoptera, Acari and Collembola from soil samples from Grazalema, southern Spain (CAC dataset) (Arribas *et al.* in prep) (Suppl. Fig.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S1). The three datasets were generated using a nested PCR approach with Nextera XT indexes. In all cases, initial amplification was performed using degenerate primers for 418 bp of the 3' end of the COI barcode region (Andújar *et al.* 2018a).

To ensure uniform treatment of datasets, raw sequence reads were re-processed following a uniform protocol including primer removal, paired end merging, quality filtering, length filtering for reads ranging between 416-420 bp (the expected 418 bp amplicon ± 2 bp), followed by denoising by library with the –unoise3 command in Usearch v11 (Edgar, 2016). The last step included chimera checking, dereplication, and removal of all singleton reads which were not considered further. Final datasets included reads surviving the cleaning and denoising steps and are referred to as ASVs. For BEE, COL and CAC datasets respectively only ASVs assigned to Apoidea, Coleoptera, and Coleoptera, Acari, or Collembola were retained. To perform this selection, we generated a reference database comprised of (i) the NCBI *nt* database (downloaded 17 June, 2018) combined with (ii) 1,011 additional reference COI sequences from Coleoptera, Acari and Collembola specimens collected in the Canary Islands and Sierra de Grazalema. For each of the three datasets, searches against the reference database were performed using the BLASTn algorithm (Altschul *et al.* 1990), with the following settings: outfmt 5, -evalue 0.001, -max_target_seqs 100. Blast results were then processed with MEGAN6 (Huson *et al.* 2016), using the weighted lowest common ancestor algorithm to assign taxonomy to ASVs.

2.2.3 | Phylogenetic assignment of va-ASVs and vna-ASVs

Phylogenetic analysis of ASVs was conducted for four lineages to explore the extent of NUMT co-amplification. We used three species within the genera *Halictus* (Hymenoptera; Apoidea) from the BEE dataset, and the genus *Cryptocephalus* (Coleoptera: Chrysomelidae) from the COL dataset. A ML tree was first estimated for all ASVs from across all libraries for each dataset, with the aim of

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

identifying the clade of ASVs corresponding to the target taxa. All relevant ASVs were extracted, but for logistical purposes related to dataset size and graphical representation, ASVs with a read-count of less than five were excluded from each library and in the case of *Cryptocephalus*, only the largest 10 libraries (>1000 reads within the *Cryptocephalus* lineage) were represented.

Classification as va-ASVs and vna-ASVs for the COL dataset was by query against the reference database described above using BLASTn (-perc_identity 100). For the BEE dataset, va-ASV were those matching the authentic haplotype that was known for all the 462 individuals included, and any ASV that included indels and/or stop codons in the translation was designated as vna-ASV. All others were considered unclassified u-ASVs. We used maximum-likelihood (ML) phylogenetic analysis to establish relationships of the ASVs within each lineage, and mapped ASV distributions and read abundances against each library using Cytoscape (Shannon *et al.* 2003). ML inferences were conducted in RAxML (Stamatakis 2006) with 100 searches for the best tree (GTR+G+I model) and 1000 bootstrap pseudoreplicates. In addition, a species delimitation analysis was conducted on the ML tree using bPTP (Zhang *et al.* 2013) on the *bPTP* web server (https://species.h-its.org/) with 100,000 generations and a burn-in of 10%.

2.2.4 | Application of *NUMTdumper*

We applied *NUMTdumper* to each dataset to evaluate the effects on the survival of a-ASVs and na-ASVs under a range of threshold values for different filtering strategies based on read-abundance of ASVs, structured with regard to sequence library and phylogeny. In all cases, thresholds were applied by library, and a given ASV was only excluded if it did not pass the read-abundance threshold in all libraries where it was present. The five filtering strategies tested were: (i) *Absolute ASV abundance by library*. ASVs were filtered setting a minimum threshold of read-abundance in a given library, with threshold values between 3 and 100 reads. (ii) *Relative ASV abundance by*

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

library. ASVs were filtered based on a minimum proportion of read-abundance relative to the total number of reads in that library, with threshold values between 0.025% to 1%. (iii) *Relative ASV abundance by library and 20% clade*. ASVs were filtered based on a minimum read proportion relative to the total number of reads within clades that are divergent by more than 20%, within which an ASV is included. Read abundance thresholds from 0.1% to 90% were explored. (iv) *Relative ASV abundance by library and 15% clade*. (v) *Relative ASV abundance by library and 26% clade*. To identify clades used for criteria (iii) to (v), a UPGMA tree of all ASVs was constructed using F84 model-corrected distances (Felsenstein & Churchill, 1996) based on a MAFFT FFT-NS-2 alignment (Katoh et al. 2002) of the ASV sequences. Clades were delimited based on a specified divergence threshold within these trees. As an alternative, *NUMTdumper* also allows for clade assignation based on taxonomic identity of the ASVs (externally designated), which allows for filtering within custom defined taxonomic ranks (e.g., when multiple phyla are present in the dataset, each ASV could be assigned to a phylum).

Potential synergy among criteria for removal of na-ASVs while maximizing the survival of a-ASVs was evaluated for the COL dataset, by applying a combination of two of the following criteria: (i) absolute ASV abundance by library, (ii) relative ASV abundance by library, and (iii) relative ASV abundance by library and 20% clade. Only ASVs surviving the separate application of both criteria were retained.

Finally, using the COL dataset we explored how ASV survival and removal affect: (i) the number of OTUs recovered under similarity thresholds for OTU clustering of 3% and 6%, (ii) the number of surviving OTUs that include one or more va-ASV and consequently can be considered as verified authentic OTUs, and (iii) the number of surviving OTUs that exclusively comprise vna-ASVs and consequently can be considered as verified non-authentic OTUs. The latter contribute to taxonomic inflation. Similarity clusters were obtained using distances estimated with the F84 model and a UPGMA tree as before.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3 | RESULTS

Raw sequence reads were subjected to uniform procedures of merging, cleaning, and denoising to establish the total ASVs for each dataset (Suppl. Fig. S1), of which a subset could be confidently classified as authentic mitochondrial (va-ASVs) against the respective reference databases (see Material and Methods) or non-mitochondrial (vna-ASV), leaving all others as unclassified (u-ASVs). The BEE dataset contained 2251 total ASVs identified as Apoidea, including 160 va-ASVs and 117 as vna-ASVs. The COL dataset yielded 1845 ASVs classified as Coleoptera, with 74 classified as va-ASVs and 228 as vna-ASVs. The CAC dataset yielded 4804 ASVs, with 712 assigned to Coleoptera (55 va-ASVs and 40 vna-ASVs), 2731 to Acari (99 va-ASVs and 92 vna-ASVs), and 1361 to Collembola (67 va-ASVs and 105 vna-ASVs).

3.1 | Phylogenetic distribution of ASVs

A subset of ASVs was assessed using phylogenetic analysis, to establish the relationships of verified authentic mitochondrial (va-ASVs) and non-mitochondrial (vna-ASVs) copies. Phylogenetic relationships and their distributions across libraries were displayed relative to their abundance in a bipartite graph (Fig. 1). The three species of BEE, *Halictus rubicundus* (5 individuals, where each individual was sequenced in a separate library), *H. tumulorum* (5 individuals), and *L. malachurum* (33 individuals) produced a total of 18, 43, and 45 ASVs, respectively, and included 2, 1 and 2 va-ASVs. In every individual, the most abundant ASV corresponded to a va-ASV (Table 1). A total of and 3, 8 and 8 vna-ASVs were identified for each species respectively, with relatively low read-counts summed across libraries (maximum read-count for a vna-ASV = 344; mean for the 19 vna-ASVs = 62) (Table 1). Fifteen of the 19 were shared across 2 or more individuals, including one vna-ASV that was shared across 24 individuals. The 82 unclassified ASVs (u-ASVs) showed relatively

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

low accumulated read-counts summed across libraries (range 10-1795; mean = 87) and 58 of 82 were shared across 2 or more individuals (Fig. 1). Species delimitation analysis on the phylogenetic tree from all ASVs with the bPTP procedure produced two candidate species for both *H. tumulorum* and *L. malachurum*, and in both cases one was exclusively composed of low abundance u-ASVs. It is worth noting that in the case of the BEE dataset all a-ASVs were known, thus all other ASVs had to be either sequencing artefacts or NUMT sequences.

For the genus *Cryptocephalus,* 6 of 118 ASVs were identified as va-ASVs, each with a high read-count summed across libraries (Fig. 2, Table 1). Several u-ASVs, closely related to the va-ASVs in the ML tree, showed similarly high read abundances, suggesting their mitochondrial origin (only a subset of the COL authentic haplotypes are known). Several libraries showed more than one high-abundance va-ASV, as expected if more than one *Cryptocephalus* species was present in a sample (libraries correspond to soil samples that sometimes contained tens of *Cryptocephalus* larvae). Thirty ASVs classified as vna-ASVs were found, all of them with low abundances (read-counts summed across libraries from 5 to 43; mean = 13). These vna-ASVs clustered together with additional low abundance u-ASVs into several clades (named C1-C8 in Figure 1D) and grades (named G1 and G2 in Figure 1D). Several of these clades were classified by bPTP as candidate species, producing 39 species in total where only 4 were expected. Despite the notable divergence of sequences inside these clades or grades from the closest va-ASV (e.g, clade C1 has a mean non-corrected p distance of 14.5% against the closest va-ASV), in many cases, vna-ASVs or closely related low abundance u-ASVs are co-amplified in two or more libraries that in addition (i) share the presence of the same *Cryptocephalus* species or (ii) share the presence of closely related species, suggesting a NUMT origin that predates speciation. As an example, libraries S20 and S92 (Fig. 1D) that share the presence of authentic mitochondrial haplotypes from *Cryptocephalus* sp. 4, also share the co-amplification of one vna-ASV and 3 closely related low abundance u-ASVs within the grade G2 and clade C8. Also, all six libraries including va-ASVs from *Cryptocephalus* sp. 3 and sp. 4 share

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

co-amplified distantly related ASVs (p-distances >15%) from clade C1 (composed of a set of low abundance u-ASVs and vna-ASVs) (Fig. 1D). In a consistent manner, libraries S29, S30, S56 and S74, that only include va-ASVs from *Cryptocephalus* sp. 1 and sp. 2 did not include any ASVs from clade C1 (Fig. 1D).

3.2 | NUMT removal efficiency of *NUMTdumper*

For all datasets, across all read abundance filtering strategies, increasing thresholds for minimum read abundance resulted in contrasting trends for the removal of va-ASVs and vna-ASVs. In general, the proportion of surviving vna-ASVs dropped quickly below 10% as thresholds were increased, at which point the percentage of surviving va-ASVs exceeded 90% (Fig. 2).

The observed values of va-ASVs and vna-ASVs, and estimated values of initial and final a-ASVs and na-ASVs (using the rationale and formula described in methods) are summarized in Table 2 and Supplementary Tables S1-S15, and represented in Figures 2 and 3. For the BEE dataset, it was possible to eliminate 99% of vna-ASVs while keeping more than 95% of the va-ASVs using filters for either an absolute or a relative ASV abundance by library. Filtering by the three variants for minimum proportion of reads by similarity cluster and library produced some recalcitrant vna-ASVs that were not removed. For COL and CAC, the filtering criteria generally allowed elimination of 90-95% of vna-ASVs while retaining 80-90% of the va-ASVs. The observed value of vna-ASVs and estimated value of final (surviving) na-ASVs always showed a strong decay reaching 0 (in the case of filtering by minimum absolute or relative ASV abundance by library) or a certain number (filtering based on relative ASV abundance by library and similarity cluster) corresponding to a fixed number of recalcitrant na-ASVs not removable with such criteria. The observed value of va-ASVs and the estimated value of final (surviving) a-ASVs showed a shallower decay with increasing threshold values.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For the BEE and COL datasets, estimates of initial a-ASVs and na-ASVs were approximately constant through increasing threshold values with the strategies of a minimum absolute or relative ASV abundance by library. In the case of BEE, estimated values for *A* and *N* (estimated number of a-ASVs and na-ASVs before filtering) approached the known true values (from Creedy et al. 2020) of *A* = 160 and *N* = 2091. Filtering by minimum absolute ASV abundance by library, we obtained *A* mean = 167 (SD = 10) and *N* mean = 2083 (SD = 10) across different thresholds while minimum relative ASV abundance by library, resulted in *A* mean = 165 (SD = 10) and *N* mean = 2086 (SD = 10). For COL, for the same two criteria, estimated values were always close to *A* = 600 (mean = 581, SD = 20; and mean = 605, SD = 27, respectively) and *N* = 1250 (mean = 1263, SD = 20; and mean = 1238, SD = 27). Estimation of initial a-ASVs and na-ASVs for the CAC dataset showed a different pattern, with a decrease in the estimated value of initial a-ASVs with increasing threshold values, from around 2,200 a-ASVs to 1,000 a-ASVs (mean = 1451, SD = 360; and mean = 1398, SD = 263, respectively for the two criteria), and an increase of initial na-ASVs from 2600 to 3800 (mean = 3245, SD = 360; and mean = 3352, SD = 263). Based on the minimum percentage of reads by similarity clusters, the estimation of final and initial a-ASVs and na-ASVs is less predictable (Fig. 3), resulting in incorrect values in the case of the mock community (BEE dataset) and stronger trends toward the increasing number of na-ASVs and decreasing number of a-ASVs with increasing threshold values in all datasets.

The simultaneous application of two filtering strategies to the COL dataset improved filtering performance (Table 3). Several of the better filtering combinations resulted in a proportion of surviving vna-ASVs between 2% and 2.6%, estimated to represent 5-6% of all ASVs surviving filtering. The same parameters retained between 82% and 88% of the va-ASVs, estimated as 94-95% of all ASVs surviving filtering. With more stringent combinations of parameters, estimated proportions of na-ASVs in the final dataset can be reduced to 0, 1%, and 2% while still retaining 77%, 80%, and 81% of a-ASVs. (Table 3).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lastly, we examined the effect of NUMT dumping on the number of OTUs. Increasing thresholds for minimum read abundance resulted in a similar trend to that found for ASVs, with contrasting results for the removal of those OTUs verified as authentic and non-authentic (Table 4). The number of surviving OTUs verified as non-authentic (exclusively formed by vna-ASVs) reduces more quickly than the number of OTUs verified as authentic. The proportion of surviving OTUs verified as authentic for both 3% and 6% clustering was very similar to the proportion of surviving va-ASVs for the three filtering criteria and all thresholds values. The proportion of surviving OTUs verified as non-authentic showed a higher rate of survival than that observed for vna-ASVs. As an example, filtering with a minimum relative ASV abundance by library of 0.009 resulted in the survival of 87.8% of va-ASVs, 86.5% of verified authentic OTUs, 4.8% of vna-ASVs, and 21.9% of verified non-authentic OTUs (OTU clustering at 3%) (Table 4). Thus, "taxonomic inflation" generated by spurious variants can be a more recalcitrant problem than removal of individual NUMTs, requiring higher threshold values for filtering, with an associated cost in the removal of rare species from the dataset.

4 | DISCUSSION

While NUMTs have long been recognised to confound barcoding with Sanger and high throughput sequencing (e.g., Song *et al.* 2008; Shokralla *et al.* 2014; Creedy *et al.* 2020), the potential impact of NUMTs on metabarcoding has only been raised in a few studies (e.g., Andújar *et al.* 2018b; Elbrecht *et al.* 2018; Liu *et al.* 2019) but never evaluated quantitatively. NUMTs are likely to be consequential because of: (i) the wide use of degenerate primers for metabarcoding (e.g. Andújar *et al.* 2018a; Elbrecht *et al.* 2019); (ii) the complexity of specimen mixtures that produce metabarcoding data, and; (iii) the sensitivity of single-molecule sequencing with HTS platforms. NUMT insertions have been documented to occur multiple times within lineages (Bensasson,

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Feldman & Petrov 2003; Hazkani-Covo, Sorek & Graur 2003; Pons & Vogler 2005; Shi *et al.* 2016), with NUMTs accumulating within genomes over time. In addition, once inserted, duplication events within the nuclear genome may contribute to the formation of NUMT families (Bensasson, Feldman & Petrov 2003; Pamilo, Viljakainen & Vihavainen 2007; Baldo *et al.* 2011), potentially resulting in hundreds of NUMTs (e.g., Ramos et al., 2011). Our datasets reveal the potential magnitude of NUMT diversity in metazoan metabarcoding. Despite thorough quality filtering for removal of sequencing artefacts (procedures that may also collaterally remove NUMTs), remaining ASVs identified as non-authentic (vna-ASVs) fit patterns of phylogenetic relatedness, read-abundance, co-occurrence and haplotype sharing across independent libraries that are expected from NUMT evolution (Fig. 1). Here we show how NUMT dumping can be used to evaluate and select customised filtering strategies according to user requirements and filtering performance, which is estimated for a subset of amplicons known to be either authentic mitochondrial haplotypes or non-functional copies with a presumed nuclear origin, and extrapolated to the full metabarcode dataset.

We have demonstrated the potential benefit of evaluating the efficiency of different filtering strategies for the removal of spurious sequences. *NUMTdumper* implements several filtering strategies based on absolute read numbers and relative read-abundances of ASVs against the total number of reads in libraries or lineages. For all strategies, opposing trends are observed for the removal of authentic mitochondrial (va-ASVs) and presumed nuclear sequences (vna-ASVs) with increasing threshold values. Rapid decay of NUMTs relative to authentic copies allows the elimination of 90-95% of vna-ASVs, while retaining 80-90% of va-ASVs. In addition, *NUMTdumper* can be used with more complex, custom made filtering models. Using paired combinations of filtering criteria, results were improved by removing 98% of vna-ASVs, while retaining 81% of the va-ASVs, and more complex filtering models may further improve these results.

After obtaining the survival ratios of both va-ASVs and vna-ASVs, *NUMTdumper* estimates the number and proportion of surviving a-ASVs and na-ASVs for each abundance threshold (Table

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2; Fig. 3), on the assumption that the subset of va-ASVs and vna-ASVs are representative of the initial number of a-ASVs and na-ASVs, respectively. This allows for the selection of thresholds based on individual acceptance criteria for the maximum number (or proportion) of na-ASVs in a given final dataset. Results from the mock community and real datasets analysed here illustrate the potential utility and issues associated with estimations of a-ASVs and na-ASVs in the initial and final (filtered) datasets. Increasing thresholds based on absolute and relative ASV abundance by library for both the BEE and the COL datasets resulted in estimates of the initial number of a-ASVs (*A*) and na-ASVs (*N)* that are approximately constant after reaching certain values, with estimates for BEE approaching the known true values. This supports the reliability of estimates. However, the CAC dataset revealed a different pattern, with a decrease in the estimated number of initial a-ASVs and an increase for initial na-ASVs with increasing threshold values. This variation in the estimated values is likely due to the violation of the assumption that va-ASVs are a representative subset of all a-ASVs ($\frac{a_v}{A_v} = \frac{a}{A}$), and thus presents a potential means to evaluate the assumption itself. To explore this further, we manipulated the subset of va-ASVs used within the COL dataset to simulate both (i) bias from a lack of low abundance va-ASVs, and (ii) bias from a lack of high abundance va-ASVs. For both types of bias we explored three intensities: strong, moderate and low (Fig. 4). Results reveal that bias generated by a lack of low abundance va-ASVs produces the pattern found for the CAC dataset, whereas bias for a lack of high abundance va-ASVs generates the opposite trend. These analyses show that the effect of bias on the estimated initial number of a-ASVs and na-ASVs increases with increasing threshold values. However, they also reveal a limited effect on the estimated number of a-ASVs and na-ASVs in the final dataset, a consequence of the low number of surviving na-ASVs with increasing thresholds (Fig. 4). All filtering strategies that use relative ASV abundance estimated within similarity clusters result in biased estimations, likely due to the prevalence of recalcitrant na-ASVs associated to clusters exclusively formed by a single or several na-ASVs. Taken together, these analyses of bias suggest that: (i) if the assumption of the ratios

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

($\frac{a_v}{A_v} = \frac{a}{A}$) is met, the correct estimation of both initial and final (surviving) numbers of a-ASVs and na-ASVs is straightforward; (ii) violation of the assumption results in predictable changes in the initial number of a-ASVs and na-ASVs, with only limited effect on the estimation of the number of a-ASVs and na-ASVs in the final dataset, and; (iii) estimates obtained from criteria where abundance is calculated within similarity clusters alone are less reliable and should not be used.

NUMTdumper complements existing denoising protocols (e.g. UNOISE, Edgar, 2016; DADA2, Callahan et al., 2016; and Deblur, Amir et al., 2017) that are designed to efficiently remove erroneous sequences derived from PCR and sequencing artefacts, but that may not efficiently remove NUMTs. Given the differences in copy number between the nuclear and the mitochondrial genomes, read abundance is an obvious parameter to filter NUMTs. However, this relationship may be imperfect due to (i) authentic haplotypes with relatively low read abundances overlapping with the abundance ranges of NUMTs, and (ii) potential amplification biases increasing the read abundance of some particular NUMT copies. This implies that: (i) it is very unlikely that a single abundance threshold can be devised for the removal of all NUMTs, while not excluding authentic haplotypes; (ii) different grouping criteria and thresholds need to be explored to minimise proportions of false positives (NUMTs retained) and false negatives (authentic haplotypes excluded); and (iii) different datasets, with different heterogeneity and amplification biases, will likely vary with regard to optimal criteria and thresholds to minimise false positives and/or false negatives. To accommodate this variation, *NUMTdumper* incorporates different abundance-based filtering strategies to evaluate performance with regard to the removal of NUMTs and retention of authentic mitochondrial sequences.

Our results also reveal that OTU clustering alone may not be sufficient to remove the effect of na-ASVs. OTUs that are identified as non-authentic can pass filtering based on read-abundance even in higher proportions than individual vna-ASVs. This highlights the problem of "taxonomic inflation" (Flynn et al. 2015), which we show can be reduced by increasing read-abundance

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

thresholds, but with the expected trade-off for the removal OTUs representing rare species (Table 4). Thus, NUMT dumping can be also used at the OTU level to evaluate filtering performance and the expected taxonomic inflation in datasets before and after filtering, optimising between taxonomic inflation and the removal of rare species.

In conclusion, our results demonstrate the potential for abundance-based removal of NUMT sequences, but also highlight the need to evaluate thresholds for each dataset according to user-defined acceptable levels of false positives and false negatives. Studies seeking data with minimal error, such as for phylogeographic (e.g., Turon et al., 2019) or population genetic analyses (e.g., Elbrecht et al., 2018), should opt for stringent thresholds, to minimise the confounding effect of NUMTs, even at the expense of removing some authentic haplotype data and rare species. For other applications, such as those based on measures of beta diversity to explore broad ecological patterns, less strict thresholds may be admissible. In addition, studies aiming to estimate richness values at haplotype or even OTU levels may consider expected biases generated by surviving NUMTs to correct data and generate estimates of a-ASVs and authentic OTUs in the initial and final (filtered) datasets. Ultimately these are decisions that can now be made and reported with the incorporation of *NUMTdumper* in analysis pipelines. *NUMTdumper* is fully compatible and complementary with current denoising methods designed for the removal of non-authentic sequences generated during the amplification and sequencing steps. Thus, *NUMTdumper* builds upon existing denoising strategies to improve the reliability of intraspecific genetic information derived from metabarcode data, opening the door for community-level genetic analyses requiring haplotype-level resolution.

ACKNOWLEDGEMENTS

CA was supported by the Spanish Ministry of Economy and Competitiveness (MINECO, Spain) (CGL2015-74178-JIN). BCE was supported by project CGL2017-85718-P (AEI, Spain/FEDER,

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

EU). PA, TJC, BCE and APV were supported by the iBioGen project funded by the H2020 European Research Council, Grant/Award Number: 810729. We extend our gratitude to the regional governments of Andalucía and Canarias (Spain) for facilitating collecting of samples, to Jesús Arribas for assistance with field sampling and Carlos Martínez for the mathematical advice.

AUTHOR CONTRIBUTIONS

CA and PA conceived and led the study; CA, PA, and TJC designed the methodology; CA, PA, HL, ASC, TPD, APV and BCE provided the data; CA and TJC analysed the data and TJC wrote the *NUMPdumper* tool. CA and BCE wrote the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Summary of the number of libraries, ASVs, va-ASVs, and vna-ASVS obtained for the three *Halictus* species and the *Crytocephalys* lineage. Read counts refers to the sum of the ASV read-abundance across all libraries where a given ASV is present.

			va-ASVs		**vna-ASVs**	
	Libraries	ASVs	n	read-counts*	n	read counts*
H. rubicundus	5	18	2	6915 (6799-7031)	3	15 (10-19)
H. tumulorum	5	43	1	8713 (8713-8713)	8	78 (11-344)
L. malachurum	33	45	2	48282,5 (4253-92312)	8	64 (11-298)
Crytocephalus	10	118	6	2223 (763-3986)	30	13 (5-43)

*mean value (minimum - maximum values)

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. Filtering performance of three filtering criteria and a range of minimum threshold values for read abundance for the COL dataset. Data for the full range of threshold values analysed, additional criteria, and datasets are provided in Supplementary Tables S1-S15.

Criterium	Threshold	Surviving ASVs (t)*	Surviving va-ASVs (av)*	Surviving vna-ASVs (nv)*	Initial a-ASVs (A)**	Surviving a-ASV (a)**	Initial na-ASVs (N)**	Surviving na-ASV (n)**	n/t ***
Pre-filtering initial values		T = 1845	A_v = 74	N_v = 228					
Absolute ASV abundance by library (Threshold = minimum threshold of read-abundance in a given library)	3	1350 (73.2%)	74 (100%)	143 (62.7%)	517.2	517.2	1327.8	832.8	0.617
	4	1139 (61.7%)	74 (100%)	105 (46.1%)	536.3	536.3	1308.7	602.7	0.529
	5	1008 (54.6%)	74 (100%)	76 (33.3%)	589.5	589.5	1255.5	418.5	0.415
	8	787 (42.7%)	68 (91.9%)	41 (18%)	615.9	566.0	1229.1	221.0	0.281
	10	706 (38.3%)	67 (90.5%)	32 (14%)	584.3	529.1	1260.7	176.9	0.251
	12	662 (35.9%)	65 (87.8%)	24 (10.5%)	605.1	531.5	1239.9	130.5	0.197
	15	598 (32.4%)	64 (86.5%)	18 (7.9%)	575.6	497.8	1269.4	100.2	0.168
	20	526 (28.5%)	57 (77%)	10 (4.4%)	612.7	472.0	1232.3	54.0	0.103
	25	482 (26.1%)	57 (77%)	5 (2.2%)	590.0	454.5	1255.0	27.5	0.057
	35	413 (22.4%)	51 (68.9%)	2 (0.9%)	583.2	401.9	1261.8	11.1	0.027
	40	400 (21.7%)	51 (68.9%)	1 (0.4%)	572.3	394.4	1272.7	5.6	0.014
	45	371 (20.1%)	48 (64.9%)	0 (0%)	572.0	371.0	1273.0	0.0	0.000
	50	355 (19.2%)	44 (59.5%)	0 (0%)	597.0	355.0	1248.0	0.0	0.000
Relative ASV abundance by library (Threshold = minimum proportion of read-abundance relative to the total number of reads in that library)	0.001	1381 (74.9%)	74 (100%)	147 (64.5%)	538.9	538.9	1306.1	842.1	0.610
	0.0015	1159 (62.8%)	74 (100%)	106 (46.5%)	563.0	563.0	1282.0	596.0	0.514
	0.002	1026 (55.6%)	74 (100%)	84 (36.8%)	548.3	548.3	1296.8	477.8	0.466
	0.0025	920 (49.9%)	72 (97.3%)	65 (28.5%)	572.8	557.3	1272.2	362.7	0.394
	0.003	832 (45.1%)	72 (97.3%)	48 (21.1%)	581.8	566.1	1263.2	265.9	0.320
	0.004	725 (39.3%)	68 (91.9%)	36 (15.8%)	569.9	523.7	1275.1	201.3	0.278
	0.005	653 (35.4%)	67 (90.5%)	27 (11.8%)	552.1	499.9	1292.9	153.1	0.234
	0.006	604 (32.7%)	65 (87.8%)	17 (7.5%)	580.3	509.7	1264.7	94.3	0.156
	0.007	563 (30.5%)	65 (87.8%)	13 (5.7%)	557.4	489.6	1287.6	73.4	0.130
	0.008	538 (29.2%)	65 (87.8%)	11 (4.8%)	540.9	475.1	1304.1	62.9	0.117
	0.009	514 (27.9%)	62 (83.8%)	10 (4.4%)	545.5	457.0	1299.5	57.0	0.111
	0.01	498 (27%)	58 (78.4%)	8 (3.5%)	578.7	453.6	1266.3	44.4	0.089
	0.015	431 (23.4%)	53 (71.6%)	3 (1.3%)	578.5	414.3	1266.5	16.7	0.039
	0.02	387 (21%)	48 (64.9%)	2 (0.9%)	579.5	375.9	1265.5	11.1	0.029
	0.025	352 (19.1%)	44 (59.5%)	1 (0.4%)	582.7	346.5	1262.3	5.5	0.016
	0.03	326 (17.7%)	42 (56.8%)	1 (0.4%)	564.5	320.4	1280.5	5.6	0.017
	0.035	302 (16.4%)	39 (52.7%)	1 (0.4%)	562.4	296.4	1282.6	5.6	0.019
	0.04	289 (15.7%)	38 (51.4%)	1 (0.4%)	551.7	283.3	1293.3	5.7	0.020
	0.045	265 (14.4%)	36 (48.6%)	0 (0%)	544.7	265.0	1300.3	0.0	0.000
	0.05	245 (13.3%)	32 (43.2%)	0 (0%)	566.6	245.0	1278.4	0.0	0.000
Relative ASV abundance by library and 20% clade (Threshold = minimum read proportion relative to the total number of reads in the 20% divergence clade where each ASV is included in that library)	0.001	1749 (94.8%)	74 (100%)	213 (93.4%)	385.8	385.8	1459.2	1363.2	0.779
	0.0025	1509 (81.8%)	74 (100%)	158 (69.3%)	750.6	750.6	1094.4	758.4	0.503
	0.005	1289 (69.9%)	74 (100%)	109 (47.8%)	779.7	779.7	1065.3	509.3	0.395
	0.01	1077 (58.4%)	72 (97.3%)	70 (30.7%)	766.6	745.9	1078.4	331.1	0.307
	0.015	960 (52%)	71 (95.9%)	50 (21.9%)	750.4	719.9	1094.6	240.1	0.250
	0.02	897 (48.6%)	70 (94.6%)	42 (18.4%)	731.4	691.9	1113.6	205.1	0.229
	0.03	808 (43.8%)	69 (93.2%)	32 (14%)	693.2	646.3	1151.8	161.7	0.200
	0.04	750 (40.7%)	68 (91.9%)	27 (11.8%)	664.0	610.1	1181.0	139.9	0.186
	0.05	708 (38.4%)	66 (89.2%)	20 (8.8%)	679.2	605.7	1165.8	102.3	0.144
	0.06	680 (36.9%)	64 (86.5%)	19 (8.3%)	673.4	582.4	1171.6	97.6	0.144
	0.09	613 (33.2%)	62 (83.8%)	14 (6.1%)	643.6	539.2	1201.4	73.8	0.120
	0.1	600 (32.5%)	62 (83.8%)	13 (5.7%)	633.7	530.9	1211.3	69.1	0.115
	0.15	529 (28.7%)	59 (79.7%)	11 (4.8%)	587.4	468.3	1257.6	60.7	0.115
	0.2	492 (26.7%)	58 (78.4%)	10 (4.4%)	555.6	435.4	1289.4	56.6	0.115
	0.25	452 (24.5%)	56 (75.7%)	10 (4.4%)	520.5	393.9	1324.5	58.1	0.129
	0.3	425 (23%)	56 (75.7%)	8 (3.5%)	499.2	377.8	1345.8	47.2	0.111
	0.35	390 (21.1%)	53 (71.6%)	7 (3.1%)	486.3	348.3	1358.7	41.7	0.107
	0.4	379 (20.5%)	53 (71.6%)	7 (3.1%)	470.2	336.8	1374.8	42.2	0.111
	0.5	355 (19.2%)	52 (70.3%)	7 (3.1%)	444.0	312.0	1401.0	43.0	0.121
	0.6	319 (17.3%)	48 (64.9%)	5 (2.2%)	444.4	288.3	1400.6	30.7	0.096
	0.7	296 (16%)	46 (62.2%)	5 (2.2%)	426.1	264.9	1418.9	31.1	0.105
	0.8	275 (14.9%)	46 (62.2%)	5 (2.2%)	391.1	243.1	1453.9	31.9	0.116
	0.9	238 (12.9%)	38 (51.4%)	5 (2.2%)	401.8	206.4	1443.2	31.6	0.133

* Observed values; ** Estimated values; *** The ratio n/t represents the estimated proportion of na-ASVs in the filtered dataset.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3. Filtering performance for a selection of pairwise combinations of filtering criteria and minimum thresholds values for read abundance for the COL dataset. Combinations are shown that minimise the number of surviving verified non-authentic ASVs (vna-ASVs) when the number of excluded verified authentic ASVs (va-ASVs) is between 0 and 17.

Excl. va-ASVs	Absolute ASV abundance by library	Relative ASV abundance by library	Relative ASV abundance by library and 20% clade	Surviving ASVs (t)**	Surviving va-ASVs (a_v)**	Surviving vna-ASVs (n_v)**	Initial a-ASVs (A)***	Surviving a-ASV (a)***	Initial na-ASVs (N)***	Surviving na-ASV (n)***	n/t ****
Pre-filtering initial values				T = 1845	A_v = 74	N_v = 228					
0	5	0.002	-	866 (46.9%)	74 (1000%)	56 (24.6%)	547	547	1298	319	0.368
2	5	0.003	-	752 (40.8%)	72 (97.3%)	36 (15.8%)	565	550	1280	202	0.269
3	5		0.015	722 (39.1%)	71 (95.9%)	28 (12.3%)	592	568	1253	154	0.213
4	5		0.02	685 (37.1%)	70 (94.6%)	24 (10.5%)	584	552	1261	133	0.194
5	5		0.035	617 (33.4%)	69 (93.2%)	15 (6.6%)	572	533	1273	84	0.136
6	-	0.0035	0.035	585 (31.7%)	68 (91.9%)	14 (6.1%)	550	505	1295	80	0.137
7	-	0.0035	0.04	572 (31%)	67 (90.5%)	14 (6.1%)	543	492	1302	80	0.140
8	5	-	0.045	535 (29%)	66 (89.2%)	11 (4.8%)	599	502	1246	33	0.062
9	8	-	0.035	556 (30.1%)	65 (87.8%)	6 (2.6%)	596	523	1249	33	0.059
10	10	-	0.035	534 (28.9%)	64 (86.5%)	5 (2.2%)	586	506	1259	28	0.052
11	8	0.008	-	516 (28%)	63 (85.1%)	5 (2.2%)	573	488	1272	28	0.054
12	10	0.008	-	512 (27.8%)	62 (83.8%)	5 (2.2%)	578	484	1267	28	0.055
13	10	-	0.045	513 (27.8%)	61 (82.4%)	5 (2.2%)	589	485	1256	28	0.055
14	20	-	0.03	469 (25.4%)	60 (81.1%)	2 (0.9%)	565	458	1280	11	0.023
15	20	0.008	-	461 (25%)	59 (79.7%)	1 (0.4%)	571	455	1274	6	0.013
17	20	0.009	-	451 (24.4%)	57 (77.0%)	0 (0%)	586	451	1259	0	0.000

* Excluded va-ASVs: values lacking (i.e., 1 and 16) represent solutions that were not found with any pairwise combinations of criteria and threshold values; ** Observed values; *** Estimated values; **** The ratio n/t represents the estimated proportion of na-ASVs in the filtered dataset.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 4. Filtering performance for the survival of OTU clusters defined at 3% and 6% similarity for three filtering criteria. Data correspond to the COL dataset and a selected range of minimum thresholds values for read abundance. In addition to the total number of surviving OTUs, the number of "Surviving verified authentic OTUs" (those that include one or more va-ASV) and "Surviving verified non-authentic OTUs" (those exclusively formed by vna-ASVs) are also shown. Number of surviving ASVs, va-ASVs and vna-ASVs are shown for comparative purposes.

Criterium	Threshold	Surviving ASVs (t)	Surviving va-ASVs (av)	Surviving vna-ASVs (nv)	OTU clustering 3%			OTU clustering 6%		
					Surviving OTUs	Surviving verified authentic OTUs	Surviving verified non-authentic OTUs	Surviving OTUs	Surviving verified authentic OTUs	Surviving verified non-authentic OTUs
Pre-filtering initial values		T = 1845	A_v = 74	N_v = 228	407	67	41	286	66	24
Absolute ASV abundance by library (Threshold = minimum threshold of read-abundance in a given library)	3	1350 (73.1%)	74 (100%)	143 (62.7%)	365 (89.6%)	67 (100%)	34 (82.9%)	259 (90.5%)	66 (100%)	20 (83.3%)
	4	1139 (61.7%)	74 (100%)	105 (46.0%)	334 (82.0%)	67 (100%)	27 (65.8%)	237 (82.8%)	66 (100%)	13 (54.1%)
	5	1008 (54.6%)	74 (100%)	76 (33.3%)	314 (77.1%)	67 (100%)	18 (43.9%)	228 (79.7%)	66 (100%)	8 (33.3%)
	8	787 (42.6%)	68 (91.8%)	41 (17.9%)	272 (66.8%)	61 (91.0%)	12 (29.2%)	208 (72.7%)	60 (90.9%)	5 (20.8%)
	10	706 (38.2%)	67 (90.5%)	32 (14.0%)	256 (62.8%)	60 (89.5%)	13 (31.7%)	195 (68.1%)	59 (89.3%)	4 (16.6%)
	12	662 (35.8%)	65 (87.8%)	24 (10.5%)	243 (59.7%)	59 (88.0%)	9 (21.9%)	191 (66.7%)	58 (87.8%)	3 (12.5%)
	15	598 (32.4%)	64 (86.4%)	18 (7.89%)	234 (57.4%)	58 (86.5%)	9 (21.9%)	187 (65.3%)	57 (86.3%)	4 (16.6%)
	20	526 (28.5%)	61 (82.4%)	10 (4.38%)	218 (53.5%)	55 (82.0%)	6 (14.6%)	175 (61.1%)	54 (81.8%)	3 (12.5%)
	25	482 (26.1%)	57 (77.0%)	5 (2.19%)	201 (49.3%)	51 (76.1%)	4 (9.75%)	167 (58.3%)	50 (75.7%)	3 (12.5%)
Relative ASV abundance by library (Threshold = minimum proportion of read-abundance relative to the total number of reads in that library)	0.001	1381 (74.8%)	74 (100%)	147 (64.4%)	365 (89.6%)	67 (100%)	37 (90.2%)	263 (91.9%)	66 (100%)	23 (95.8%)
	0.0015	1159 (62.8%)	74 (100%)	106 (46.4%)	337 (82.8%)	67 (100%)	30 (73.1%)	246 (86.0%)	66 (100%)	17 (70.8%)
	0.002	1026 (55.6%)	74 (100%)	84 (36.8%)	319 (78.3%)	67 (100%)	24 (58.5%)	235 (82.1%)	66 (100%)	13 (54.1%)
	0.0025	920 (49.8%)	72 (97.2%)	65 (28.5%)	301 (73.9%)	65 (97.0%)	22 (53.6%)	225 (78.6%)	64 (96.9%)	12 (50%)
	0.003	832 (45.0%)	72 (97.2%)	48 (21.0%)	285 (70.0%)	65 (97.0%)	18 (43.9%)	214 (74.8%)	64 (96.9%)	9 (37.5%)
	0.004	725 (39.2%)	68 (91.8%)	36 (15.7%)	266 (65.3%)	61 (91.0%)	17 (41.4%)	202 (70.6%)	60 (90.9%)	8 (33.3%)
	0.005	653 (35.3%)	67 (90.5%)	27 (11.8%)	248 (60.9%)	60 (89.5%)	15 (36.5%)	197 (68.8%)	59 (89.3%)	8 (33.3%)
	0.006	604 (32.7%)	65 (87.8%)	17 (7.45%)	233 (57.2%)	58 (86.5%)	10 (24.3%)	187 (65.3%)	57 (86.3%)	5 (20.8%)
	0.007	563 (30.5%)	65 (87.8%)	13 (5.70%)	223 (54.7%)	58 (86.5%)	9 (21.9%)	181 (63.2%)	57 (86.3%)	4 (16.6%)
	0.008	538 (29.1%)	65 (87.8%)	11 (4.82%)	215 (52.8%)	58 (86.5%)	9 (21.9%)	177 (61.8%)	57 (86.3%)	4 (16.6%)
	0.009	514 (27.8%)	62 (83.7%)	10 (4.38%)	209 (51.3%)	55 (82.0%)	8 (19.5%)	172 (60.1%)	54 (81.8%)	3 (12.5%)
	0.01	498 (26.9%)	58 (78.3%)	8 (3.50%)	202 (49.6%)	52 (77.6%)	6 (14.6%)	167 (58.3%)	51 (77.2%)	3 (12.5%)
	0.015	431 (23.3%)	53 (71.6%)	3 (1.31%)	180 (44.2%)	48 (71.6%)	2 (4.87%)	151 (52.7%)	47 (71.2%)	2 (8.33%)
	0.02	387 (20.9%)	48 (64.8%)	2 (0.87%)	167 (41.0%)	43 (64.1%)	1 (2.43%)	141 (49.3%)	42 (63.6%)	1 (4.16%)
Relative ASV abundance by library and 20% clade (Threshold = minimum read proportion relative to the total number of reads in the 20% divergence clade where each ASV is included in that library)	0.001	1749 (94.7%)	74 (100%)	213 (93.4%)	406 (99.7%)	67 (100%)	41 (100%)	286 (100%)	66 (100%)	24 (100%)
	0.0025	1509 (81.7%)	74 (100%)	158 (69.2%)	382 (93.8%)	67 (100%)	33 (80.4%)	278 (97.2%)	66 (100%)	22 (91.6%)
	0.005	1289 (69.8%)	74 (100%)	109 (47.8%)	367 (90.1%)	67 (100%)	30 (73.1%)	269 (94.0%)	66 (100%)	18 (75%)
	0.01	1077 (58.3%)	72 (97.2%)	70 (30.7%)	340 (83.5%)	65 (97.0%)	26 (63.4%)	257 (89.8%)	64 (96.9%)	16 (66.6%)
	0.015	960 (52.0%)	71 (95.9%)	50 (21.9%)	313 (76.9%)	64 (95.5%)	20 (48.7%)	241 (84.2%)	63 (95.4%)	13 (54.1%)
	0.02	897 (48.6%)	70 (94.5%)	42 (18.4%)	301 (73.9%)	63 (94.0%)	17 (41.4%)	232 (81.1%)	62 (93.9%)	10 (41.6%)
	0.03	808 (43.7%)	69 (93.2%)	32 (14.0%)	287 (70.5%)	62 (92.5%)	14 (34.1%)	227 (79.3%)	61 (92.4%)	9 (37.5%)
	0.04	750 (40.6%)	68 (91.8%)	27 (11.8%)	282 (69.2%)	61 (91.0%)	12 (29.2%)	223 (77.9%)	60 (90.9%)	7 (29.1%)
	0.05	708 (38.3%)	66 (89.1%)	20 (8.77%)	274 (67.3%)	59 (88.0%)	11 (26.8%)	219 (76.5%)	58 (87.8%)	7 (29.1%)
	0.06	680 (36.8%)	64 (86.4%)	19 (8.33%)	269 (66.0%)	57 (85.0%)	10 (24.3%)	216 (75.5%)	56 (84.8%)	7 (29.1%)
	0.075	644 (34.9%)	63 (85.1%)	16 (7.01%)	261 (64.1%)	57 (85.0%)	9 (21.9%)	211 (73.7%)	56 (84.8%)	6 (25%)
	0.09	613 (33.2%)	62 (83.7%)	14 (6.14%)	254 (62.4%)	56 (83.5%)	8 (19.5%)	208 (72.7%)	55 (83.3%)	5 (20.8%)
	0.1	600 (32.5%)	62 (83.7%)	13 (5.70%)	251 (61.6%)	56 (83.5%)	7 (17.0%)	206 (72.0%)	55 (83.3%)	4 (16.6%)

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure captions

Figure 1. Patterns of phylogenetic relatedness and library co-amplification of ASVs within selected lineages. (**A**) *Halictus rubicundus*, (**B**) *Halictus tumulorum* (**C**), *Lasioglossum malachurum*, and (**D**) *Cryptocephalus*. Graphs show ML phylogenetic trees with mapped distributions of read abundances across libraries onto each ASV. For (A), (B) and (C) each library is a single specimen, whereas in (D) each library includes a complex natural community of beetles where *Cryptocephalus* specimens were present. Phylogenetic lineages in red are the best-supported species clusters from bPTP analyses, green dashed lines highlight ASVs that are identical to a reference sequence (va-ASVs), and black dashed lines highlight ASVs with STOP codons or INDELS (vna-ASVs). Codes on nodes mark clades (C1-C8) and grades (G1-G2) exclusively formed by vna-ASVs and u-ASVs. Circles at the tips of the tree represent each ASV, with size proportional to accumulated abundance across all libraries. Circles on the right side of each graph represent the libraries, with size proportional to the library read number. Edges of the network represent the presence and abundance (line width) of each ASV within each library.

Figure 2. Proportions of va-ASVs removal (false negatives) and vna-ASVs retention (false positives) with increasing minimum abundance thresholds. Graphs represent alternative filtering criteria (up-down) and different datasets (left-right). The "X" axis corresponds to minimum threshold values based on: (**A**) absolute ASV abundance by library, (**B**) relative ASV abundance by library, (**C**) relative ASV abundance by library and within 15% similarity clades, (**D**) within 20% similarity clades, and (**E**) within 26% similarity clades. The percentage of removed va-ASVs (validated against reference sequences) is indicated with squared dots and a black line. The percentage of surviving vna-ASVs (including stop codons or indels) is indicated with circles and a red line.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3. Estimated numbers of a-ASVs and na-ASVs comprising initial and filtered ASV datasets after the application of each abundance-based filtering threshold. Graphs show trends for the estimated number of initial a-ASVs (shaded-grey), initial na-ASVs (shaded-red), retained a-ASVs (squared dots and black line), and retained na-ASVs (circles and red line) with increasing minimum abundance thresholds for alternative filtering criteria (up-down) and different datasets (left-right). The "X" axis corresponds to minimum threshold values based on: (**A**) absolute ASV abundance by library, (**B**) relative ASV abundance by library, (**C**) relative ASV abundance by library and within 15% similarity clades, (**D**) within 20% similarity clades, and (**E**) within 26% similarity clades.

Figure 4. Estimated number of a-ASVs and na-ASVs comprising the initial and filtered COL dataset after the application of thresholds for the minimum absolute ASV abundance by library, using manipulated subsets of va-ASVs. Manipulations included (i) a bias for a lack of low abundance va-ASVs (above), and (ii) a bias for a lack of high abundance va-ASVs (below; each with three bias intensities: strong, moderate and low). Graphs show estimations of initial a-ASVs (shaded-grey), initial na-ASVs (shaded-red), retained a-ASVs (squared dots and black line), and retained na-ASVs (circles and red line) with increasing minimum abundance thresholds. The black and red dotted lines represent, respectively, a-ASVs and na-ASVs estimations using the full set of va-ASVs.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–10.

Amir, A., Daniel, M., Navas-Molina, J., Kopylova, E., Morton, J., Xu, Z.Z., Eric, K., Thompson, L., Hyde, E., Gonzalez, A. & Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *American Society for Microbiology*, **2**, 1–7.

Andújar, C., Arribas, P., Gray, C., Bruce, C., Woodward, G., Yu, D.W. & Vogler, A.P. (2018a). Metabarcoding of freshwater invertebrates to detect the effects of a pesticide spill. *Molecular Ecology*, **27**, 146–166.

Andújar, C., Emerson, B.C., Arribas, P., Yu, D.W. & Vogler, A.P. (2018b). Why the COI barcode should be the community DNA metabarcode for the metazoa. 3968–3975.

Baldo, L., De Queiroz, A., Hedin, M., Hayashi, C.Y. & Gatesy, J. (2011). Nuclear-mitochondrial sequences as witnesses of past interbreeding and population diversity in the jumping bristletail mesomachilis. *Molecular Biology and Evolution*, **28**, 195–210.

Bensasson, D., Feldman, M.W. & Petrov, D.A. (2003). Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *Journal of Molecular Evolution*, **57**, 343–354.

Bensasson, D., Zhang, D.X., Hartl, D.L. & Hewitt, G.M. (2001). Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends in Ecology and Evolution*, **16**, 314–321.

Bogenhagen, D.F. (2012). Mitochondrial DNA nucleoid structure. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, **1819**, 914–920.

Callahan, B.J., McMurdie, P.J. & Holmes, S.P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, **11**, 2639–2643.

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

Caruso, V., Song, X., Asquith, M. & Karstens, L. (2019). Performance of microbiome sequence inference methods in environments with varying biomass. *mSystems*, **4**, 1–19.

Clare, E.L., Chain, F.J.J., Littlefair, J.E. & Cristescu, M.E. (2016). The effects of parameter choice

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. *Genome*, **59**, 981–990.

Corse, E., Meglécz, E., Archambaud, G., Ardisson, M., Martin, J.F., Tougard, C., Chappaz, R. & Dubut, V. (2017). A from-benchtop-to-desktop workflow for validating HTS data and for taxonomic identification in diet metabarcoding studies. *Molecular Ecology Resources*, **17**, e146–e159.

Creedy, T.J., Norman, H., Tang, C.Q., Qing Chin, K., Andujar, C., Arribas, P., O'Connor, R.S., Carvell, C., Notton, D.G. & Vogler, A.P. (2020). A validated workflow for rapid taxonomic assignment and monitoring of a national fauna of bees (Apiformes) using high throughput DNA barcoding. *Molecular Ecology Resources*, **20**, 40–53.

Edgar, R. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*, 081257.

Elbrecht, V., Braukmann, T.W.A., Ivanova, N.V., Prosser, S.W.J., Hajibabaei, M., Wright, M., Zakharov, E. V, Hebert, P.D.N. & Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, **7:e7745**.

Elbrecht, V., Vamos, E.E., Steinke, D. & Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ*, **6:e4644**.

Flynn, J.M., Brown, E. a., Chain, F.J.J., MacIsaac, H.J. & Cristescu, M.E. (2015). Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, **5**, 2252–2266.

Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J. & Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods*, **5**, 235–237.

Hazkani-Covo, E., Sorek, R. & Graur, D. (2003). Evolutionary dynamics of large Numts in the human genome: Rarity of independent insertions and abundance of post-insertion duplications. *Journal of Molecular Evolution*, **56**, 169–174.

Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.J. & Tappu, R. (2016). MEGAN community edition - Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Computational Biology*, **12**, 1–12.

Liu, M., Clarke, L.J., Baker, S.C., Jordan, G.J. & Burridge, C.P. (2019). A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology*, **45**, 373–385.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lopez, J. V., Yuhki, N., Masuda, R., Modi, W. & O'Brien, S.J. (1994). Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of Molecular Evolution*, **39**, 174–190.

Nearing, J.T., Douglas, G.M., Comeau, A.M. & Langille, M.G.I. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, **6**, e5364.

Pamilo, P., Viljakainen, L. & Vihavainen, A. (2007). Exceptionally high density of NUMTs in the honeybee genome. *Molecular Biology and Evolution*, **24**, 1340–1346.

Paradis, E., Claude, J. & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Pons, J. & Vogler, A.P. (2005). Complex pattern of coalescence and fast evolution of a mitochondrial rRNA pseudogene in a recent radiation of tiger beetles. *Molecular Biology and Evolution*, **22**, 991–1000.

Quiros, P.M., Goyal, A., Jha, P. & Auwerx, J. (2017). Analysis of mtDNA/nDNA ratio in mice. *Current protocols in mouse biology*, **7**, 47.

Ramos, A., Barbena, E., Mateiu, L., del Mar González, M., Mairal, Q., Lima, M., Montiel, R., Aluja, M.P. & Santos, C. (2011). Nuclear insertions of mitochondrial origin: Database updating and usefulness in cancer studies. *Mitochondrion*, **11**, 946–953.

Richly, E. & Leister, D. (2004). NUMTs in sequenced eukaryotic genomes. *Molecular Biology and Evolution*, **21**, 1081–1084.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**, 2498–504.

Shi, H., Dong, J., Irwin, D.M., Zhang, S. & Mao, X. (2016). Repetitive transpositions of mitochondrial DNA sequences to the nucleus during the radiation of horseshoe bats (Rhinolophus, Chiroptera). *Gene*, **581**, 161–169.

Shokralla, S., Gibson, J.F., Nikbakht, H., Janzen, D.H., Hallwachs, W. & Hajibabaei, M. (2014). Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, **14**, 892–901.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.157347; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Song, H., Buhay, J.E., Whiting, M.F. & Crandall, K. a. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 13486–91.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, **22**, 2688–90.

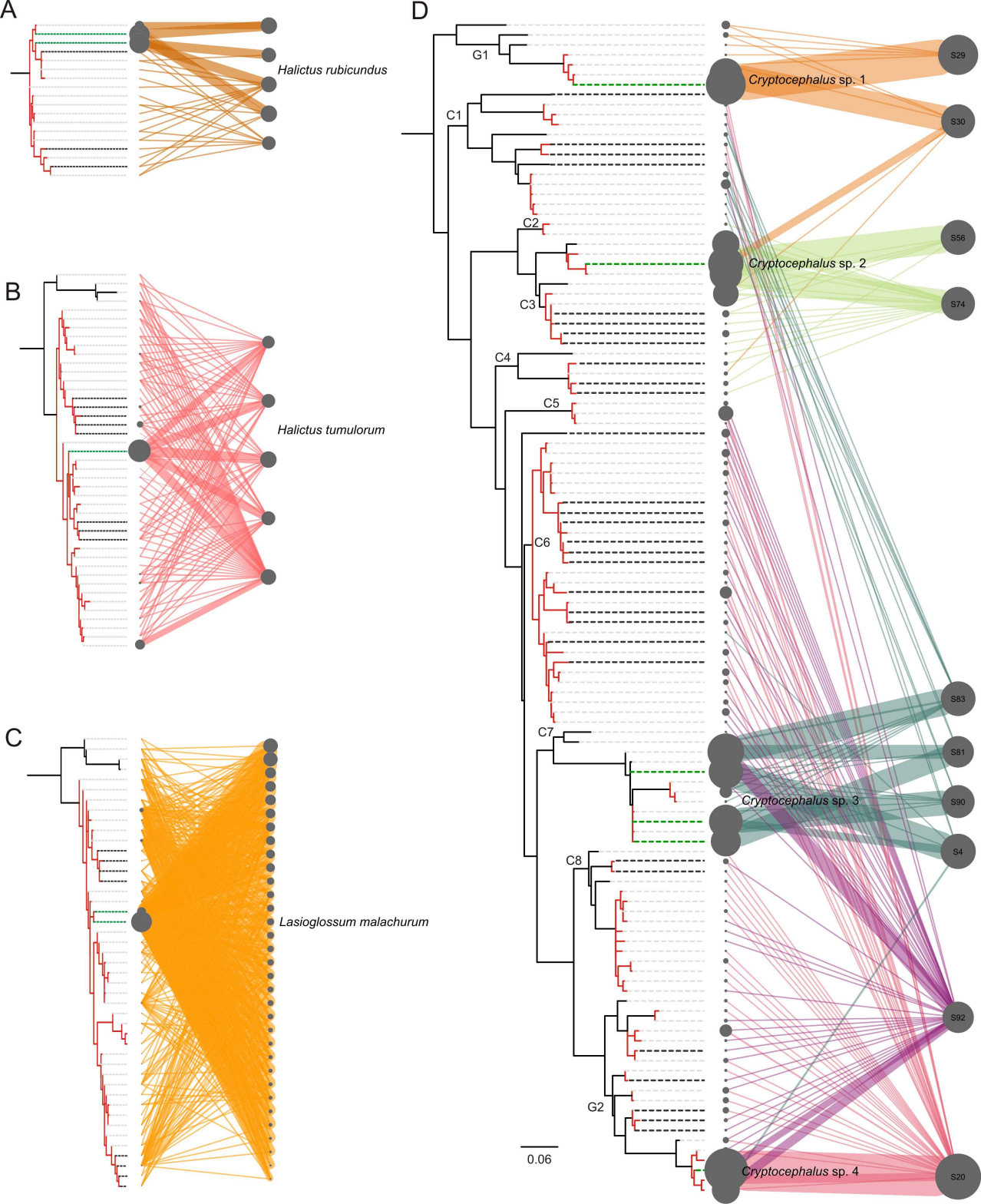
Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–50.

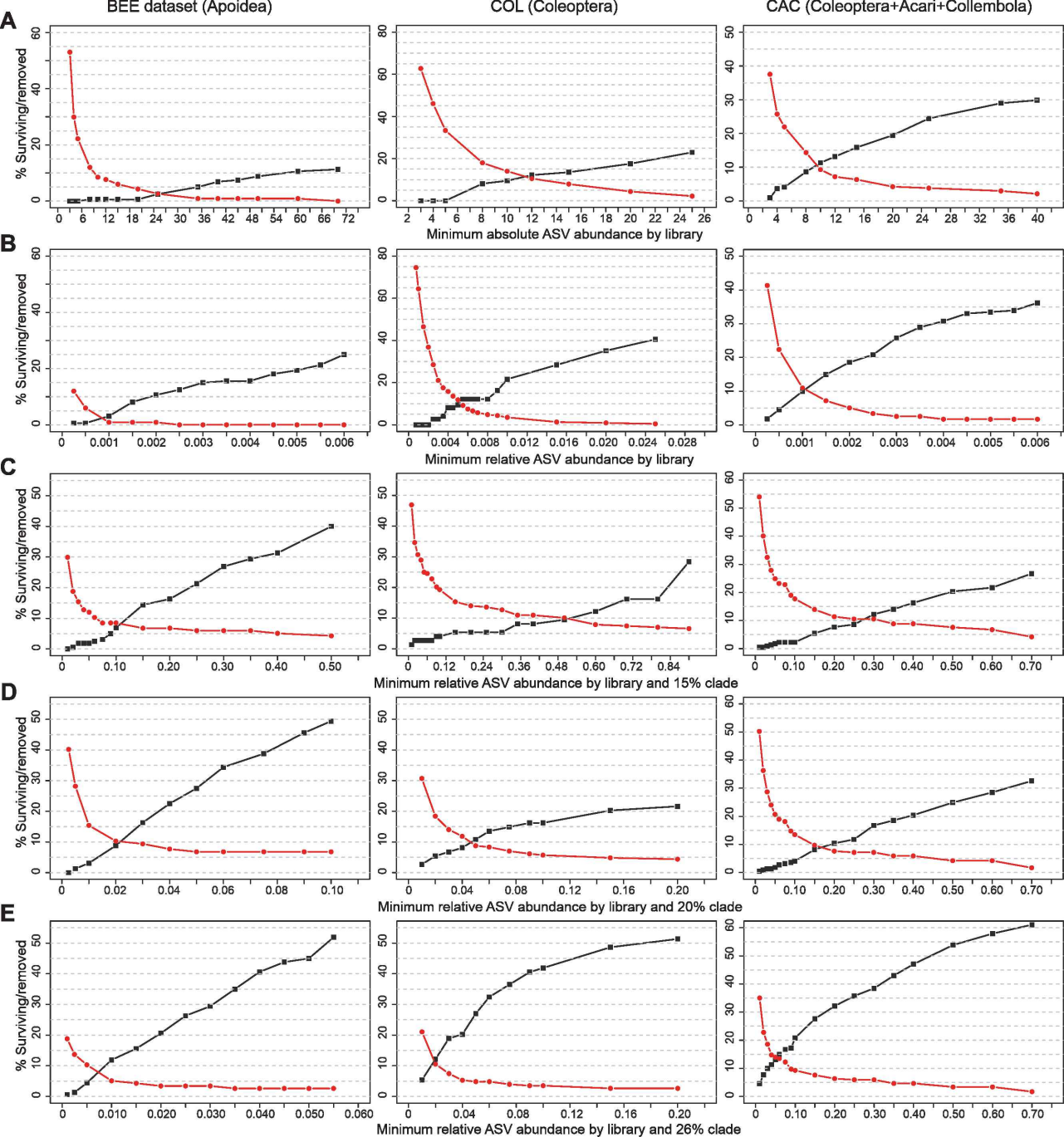
Thomsen, P.F. & Sigsgaard, E.E. (2019). Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods. *Ecology and Evolution*, **9**, 1665–1679.

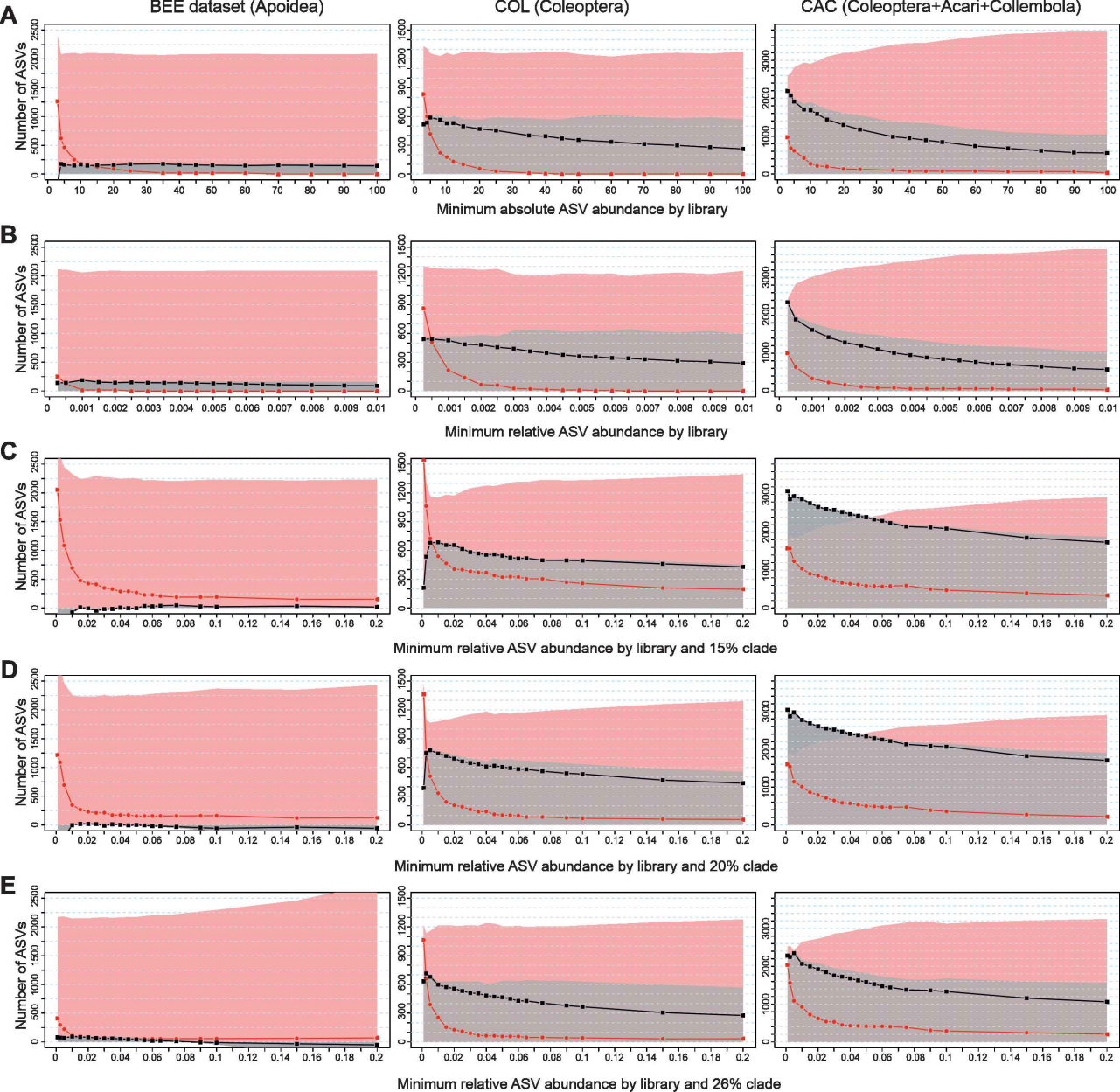
Turon, X., Antich, A., Palacín, C., Præbel, K. & Wangensteen, O.S. (2019). From metabarcoding to metaphylogeography: separating the wheat from the chaff. *bioRxiv*, 629535.

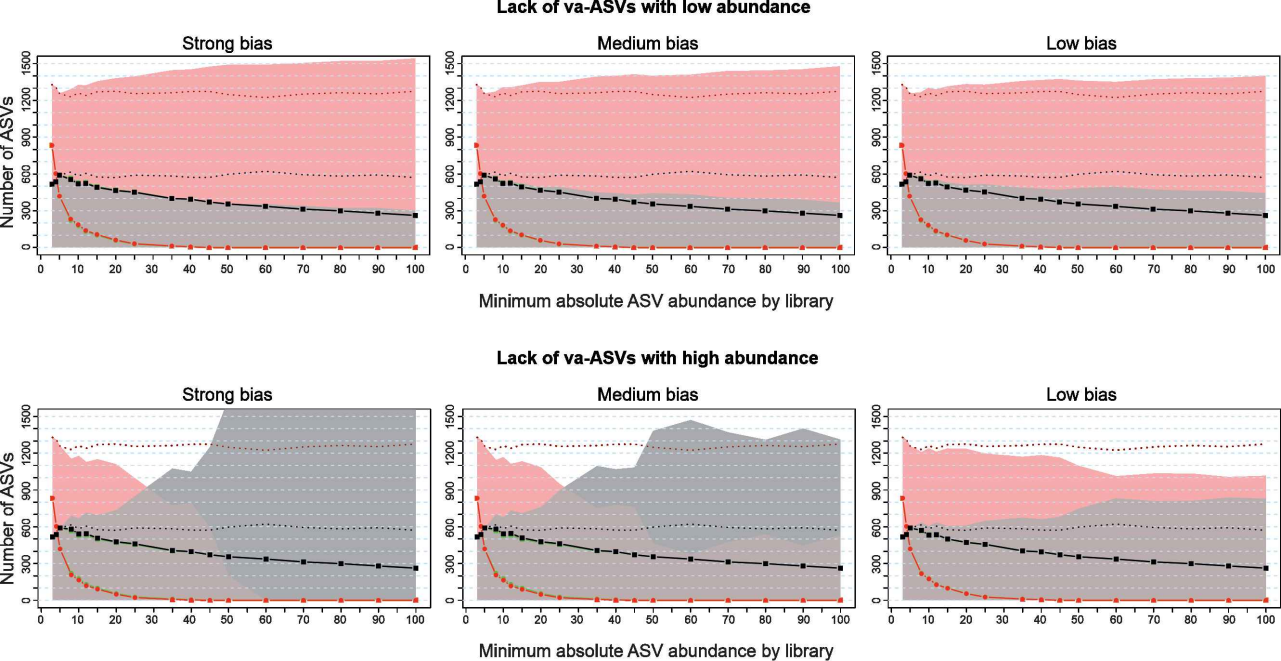
Yu, D., Ji, Y., Emerson, B., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zhang, J., Kapli, P., Pavlidis, P. & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics (Oxford, England)*, **29**, 2869–76.


A
Halictus rubicundus
B
Halictus tumulorum
C
Lasioglossum malachurum
D
G1
C1
C2
C3
C4
C5
C6
C7
C8
G2
Cryptocephalus sp. 1
Cryptocephalus sp. 2
Cryptocephalus sp. 3
Cryptocephalus sp. 4
S29
S30
S56
S74
S83
S81
S90
S4
S92
S20
0.06


BEE dataset (Apoidea)
COL (Coleoptera)
CAC (Coleoptera+Acari+Collembola)
A
B
C
D
E
% Surviving/removed
Minimum absolute ASV abundance by library
Minimum relative ASV abundance by library
Minimum relative ASV abundance by library and 15% clade
Minimum relative ASV abundance by library and 20% clade
Minimum relative ASV abundance by library and 26% clade


BEE dataset (Apoidea)
COL (Coleoptera)
CAC (Coleoptera+Acari+Collembola)
A
B
C
D
E
Number of ASVs
Minimum absolute ASV abundance by library
Minimum relative ASV abundance by library
Minimum relative ASV abundance by library and 15% clade
Minimum relative ASV abundance by library and 20% clade
Minimum relative ASV abundance by library and 26% clade


Lack of va-ASVs with low abundance
Strong bias
Medium bias
Low bias
Number of ASVs
0
300
600
900
1200
1500
0
10
20
30
40
50
60
70
80
90
100
Minimum absolute ASV abundance by library
Lack of va-ASVs with high abundance
Strong bias
Medium bias
Low bias
Number of ASVs
Minimum absolute ASV abundance by library