Received: 19 January 2022 | Revised: 7 October 2022 | Accepted: 17 November 2022

DOI: 10.1111/1755-0998.13745

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

The topological nature of tag jumping in environmental DNA metabarcoding studies

Saul Rodriguez-Martinez[1] | Jonatan Klaminder[1] | Marina A. Morlock[1] | Love Dalén[2,3] | Doreen Yu-Tuan Huang[1]

[1]Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden

[2]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden

[3]Centre for Palaeogenetics, Svante Arrhenius väg 20C, Stockholm, Sweden

Correspondence
Saul Rodriguez-Martinez, Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden.
Email: saul.rguezm@gmail.com

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: P2BEP2-188256; Swedish research council, Grant/Award Number: 2017-04548

Handling Editor: Carla Martins Lopes

Abstract

Metabarcoding of environmental DNA constitutes a state-of-the-art tool for environmental studies. One fundamental principle implicit in most metabarcoding studies is that individual sample amplicons can still be identified after being pooled with others—based on their unique combinations of tags—during the so-called demultiplexing step that follows sequencing. Nevertheless, it has been recognized that tags can sometimes be changed (i.e., tag jumping), which ultimately leads to sample cross-talk. Here, using four DNA metabarcoding data sets derived from the analysis of soils and sediments, we show that tag jumping follows very specific and systematic patterns. Specifically, we find a strong correlation between the number of reads in blank samples and their topological position in the tag matrix (described by vertical and horizontal vectors). This observed spatial pattern of artefactual sequences could be explained by polymerase activity, which leads to the exchange of the 3′ tag of single stranded tagged sequences through the formation of heteroduplexes with mixed barcodes. Importantly, tag jumping substantially distorted our data sets—despite our use of methods suggested to minimize this error. We developed a topological model to estimate the noise based on the counts in our blanks, which suggested that 40%–80% of the taxa in our soil and sedimentary samples were likely false positives introduced through tag jumping. We highlight that the amount of false positive detections caused by tag jumping strongly biased our community analyses.

KEYWORDS
a-DNA, detection limits, e-DNA, false positive, index hopping, sample crosstalk

1 | INTRODUCTION

Metabarcoding (Taberlet et al., 2012) is a technique that enables identification of multiple taxa in bulk samples from one single sequencing run. It is frequently applied for environmental monitoring (Fernandez et al., 2018; Jeunen et al., 2021), paleo-reconstructions (Clarke et al., 2019; Haile et al., 2009), ecological research (Chariton et al., 2010; Ficetola et al., 2008; Lacoursière-Roussel et al., 2018) and assessment of diets (Deagle et al., 2009; Jo et al., 2020; Soininen et al., 2009; Valentini et al., 2009) among other uses. Successful recovery of specific DNA from complex environmental matrices such as water, faeces, soils, sediments or bulk tissue samples makes

metabarcoding a widely-applicable tool for environmental and ecological studies (Bohmann et al., 2011; Deagle et al., 2009; Pegard et al., 2009; Schnell et al., 2015; Soininen et al., 2009; Valentini et al., 2009).

To facilitate the simultaneous analysis of multiple samples during sequencing, each DNA amplicon from individual samples is labelled (tagged) by adding a pair of short but unique oligonucleotide sequences (hereafter referred to as tags) to the 5′ end of both primers in a PCR reaction (Binladen et al., 2007), allowing the tagged amplicons to be pooled for sequencing. These tags leave specific marks in the amplicons that can be identified during the later bioinformatics (demultiplexing) step and used to backtrack the sequence to its original sample. This tagging technique is a well stablished approach that has been applied on high throughput screening platforms (Quéméré et al., 2013; Shehzad et al., 2012).

Tags are assumed to be immobile, but it has been shown that tags can indeed switch during the library preparation step (Carøe & Bohmann, 2020;Esling et al., 2015; Schnell et al., 2015). At this time, amplicons go through an end-repairing reaction, after which the Illumina indexes and adapters are ligated to both ends of the sequences (Meyer & Kircher, 2010). During the library preparation stage, artefactual (nonoriginal) sequences carrying a different combination of tags can be generated by tag jumping (Schnell et al., 2015). In fact, studies have detected amplicons with incorrect tags following processing through the Illumina sequencing platform, which can lead to amplicons being attributed to incorrect samples if the tag combination formed matches that of an actual sample (Esling et al., 2015; Schnell et al., 2015). The result of this tag jumping is a sequence being misidentified as belonging to an amplicon from another sample, causing an effect similar to sample cross-contamination. The tag jumping described above should not be confused with "index hopping", an issue previously presented by van der Valk et al. (2020), who showed that indexes added to amplicons can move between libraries during sequencing and affect the results of studies with low DNA yields (van der Valk et al., 2020). Nevertheless, similar to index hopping, tag jumping can introduce sequences that artificially inflate species diversity in samples.

Four causes of tag jumping have been previously proposed: (1) mixed clusters, also called cluster bleeding (Kircher et al., 2012); (2) cross-contamination of primers with different tags, either during synthesis or laboratory handling (Kircher et al., 2012); (3) chimera amplification during library enrichment (Kircher et al., 2012), and (4) the use of DNA polymerase with 3′-5′ exonuclease activity during end repair (Schnell et al., 2015; van Orsouw et al., 2007). However, their respective contribution(s) to the problem remain unexplored. In this study we evaluate the latter of the suggested tag jumping explanations illustrated in Figure 1, mediated by a polymerase (usually T4 DNA polymerase) with 3′-5′ exonuclease activity during end repair (Schnell et al., 2015; van Orsouw et al., 2007). These polymerases are commonly used during library preparation (of adapter ligation protocols) for Illumina in the end-repair step (Meyer & Kircher, 2010). The applied polymerase can digest nonmatching DNA nucleotides at the 3′-end of the DNA strand and extend that same end using the complementary strand as a template. During this process, single-stranded tagged sequences of two different amplicons can interact, forming a heteroduplex with mixed tags (van Orsouw et al., 2007). In a heteroduplex, the activity of the commonly used T4 DNA polymerase can switch the 3′ tag at each side, creating artefactual sequences that carry a combination of tags formed by the 5′-tag of each interacting strand (van Orsouw et al., 2007). If the newly formed tag combination is shared by another sample, it cannot be separated from 'true' sample sequences during the demultiplexing step.

Within a tag matrix, where primers are arranged to generate unique tag combinations for all samples using both forward (rows) and reverse primers (column), tag jumping (if occurring as described in Figure 1) is constrained to produce misidentified amplicons along horizontal or vertical lines in the original tags matrix. That is, artefactual combinations carrying the 5′ tag of each strand interacting during the end-repair step will share the same row/column in the matrix as the two interacting strands. If chimera amplification during a second PCR can be excluded as an explanation for tag jumps, artefactual sequences forming "cross-like" patterns in tagging matrices are a unique characteristic of a tag jumping processes mediated by T4 DNA polymerase as described by Schnell et al. (2015). That is, other suggested processes such as cluster bleeding and/or the cross-contamination of primers do not generate cross-shaped patterns related to sample positioning within the tagging matrix. To date, this unique topological signature of heteroduplex-driven tag jumping has not been explored.

In this study, we evaluated topological patterns in four metabarcoding libraries by both inspecting findings of rare taxa (i.e., taxa present in few samples) as well by assessing artefactual sequences detected through our extensive use of sample blanks. Here, the rationale for our focus on rare taxa was that their absence from most samples simplified visualization of eventual cross-like patterns. We applied a topological model accounting for a sample's position in the tagging matrix (column × row) to provide the first quantitative mapping of samples affected by a tag jumping process that behaves as proposed by Schnell et al. (2015). We hypothesized that tag jumping is a process potent enough to substantially alter the findings and outcome of metabarcoding studies applied to soil and sediment samples.

2 | MATERIALS AND METHODS

2.1 | DNA extraction

We extracted DNA from two sets of samples: First, lake sediments from Lake Grosssee (Flumserberg, Switzerland) in 47°04′43.8" N 9°14'47.1" E at 1619 m.a.s.l. Sediment cores were recovered in 2017 from the central basin of Lake Grosssee with a percussion piston-scoring system (Uwitec, Austria). Cores were split lengthwise and aligned to a composite of 7.73 m as described in

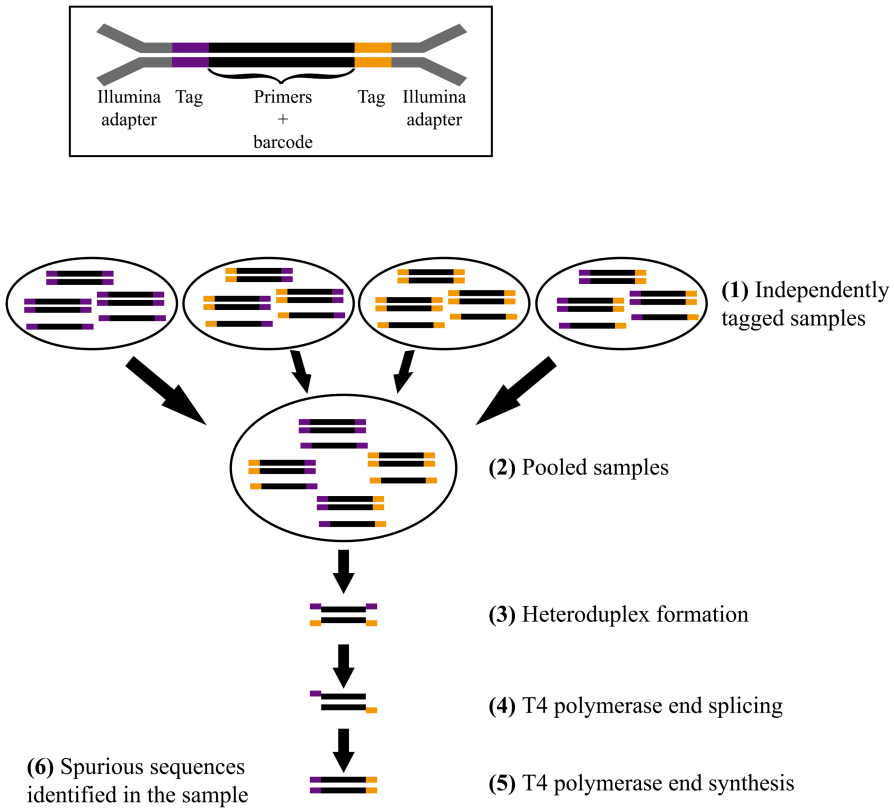

FIGURE 1 Within the rectangle, a simple scheme of a double stranded DNA fragment. Below, six different steps illustration of how tag jumping as described in Schnell et al. (2015) affect a library. (1) four different samples (indicated with black circles) are tagged with reverse and forward primers ('tag') combinations (purple and yellow bars). Here, all amplicons have a certain level of single stranded DNA. (2) prior to sequencing, amplicons are pooled in a single library containing all four samples (black circle). (3) during the after following library preparation, single stranded sequences interact to form a heteroduplex. (4) during the end-repair step the formed heteroduplex is exposed to T4 polymerase. The 3′-5′ exonuclease activity of this polymerase can splice the non-matching 3′ ends. (5) the 3′ end is synthesized again matching the complementary strand. The switching of tags between the ends results in a new tag combination for the heteroduplex (6). The sequences that formed will be identified as a sample if the new combination match that of an actual sample. This artefactual sample will consist of strands derived from two different samples.

Glaus (2018) and sampled at 14, 132, 232, 235, 352 and 529 cm of composite depth. These sedimentary samples were analysed in libraries labelled MAM1, MAM2 and PLAN. The second set of samples originated from a sedimentary soil profile developing about 10 km from Abisko Scientific Research Station (68°21′17.0″ N; 18°48′54″ E) where both soils and earthworms have previously been described in detail (Wackett et al., 2018). Soil samples were taken at 30, 90 and 220 cm depth below the surface and analysed in the library WORM.

For each depth, three subsamples (extraction replicates, 0.5 g of material each) were extracted using the DNeasy Power Soil kit (Qiagen). DNA extractions were carried out in a dedicated ancient DNA laboratory at Umeå University that is isolated from other modern DNA laboratories and has a positive air pressure system accompanied by a HEPA air filter system. We followed protocols to avoid contamination during the extraction process (Paijmans et al., 2019), including using extensive personal protective gear (facemask, gloves, and clean suit). Multiple extraction blanks and control amplifications were processed alongside all samples to screen for potential contamination.

2.2 | DNA metabarcoding with tagged primers and sequencing

All prepolymerase chain reaction (PCR) preparations for DNA metabarcoding were performed in the ancient DNA laboratory to minimize possible contamination. On a 96-well plate, tagged primers were arranged in a matrix. We arranged at least one PCR blank (with PCR mix and water, no primer, no DNA) per row and per column to assess the extent of tag jumping. Each extract, including those from extraction blanks, had 2–3 independent PCR replicates (i.e., with individual tag combinations; sequences and number of replicates detailed in Data S1) in the same column of a 96-well plate; all the extracts were randomly distributed on the PCR plate. We also included at least six PCR negatives (with PCR mix, water, and primers, but no DNA) for each plate and one PCR positive with DNA extracted from *Picea abies* (library PLAN), *Bos taurus* (libraries MAM1 and MAM2) and an earthworm molecularly identified as *Aporrectodea caliginosa* (WORM).

Libraries PLAN (chloroplast DNA), MAM1 (16S mitochondrial DNA), MAM2 (16S mitochondrial DNA), and WORM (16S

mitochondrial DNA) included amplicons amplified by the trnL g/h (Taberlet et al., 2007), MamP007F/R (Giguet-Covex et al., 2014), 16Smam-1/−2 (Taylor, 1996), and EW-B/−E (Bienert et al., 2012) primers, respectively.

DNA amplification for the PLAN library was carried out in 20 µl reactions containing 1x PCR buffer (Qiagen), 2 mM $MgCl_2$, 0.2 mM dNTPs (Qiagen), 0.4 µM trnL g/h primers, 1.25 U HotStartTaq DNA polymerase (Qiagen), 4 µg bovine serum albumin (BSA, Thermo Fisher Scientific), and 1 µl DNA template. PCR was used for the enzyme activation at 95°C for 15 min and then 45 cycles consisting of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s, followed by final elongation at 72°C for 10 min.

DNA in library MAM1 was amplified in a 25 µl reaction containing 1x PCR buffer (Qiagen), 3 mM $MgCl_2$, 0.2 mM dNTPs (Qiagen), 0.2 µM of each primer, 0.625 U HotStartTaq DNA polymerase (Qiagen), 4 µg BSA (Thermo Fisher Scientific), 2 µM human DNA blocker (Giguet-Covex et al., 2014), and 2 µl DNA template. PCR was used for the enzyme activation at 95°C for 15 min and then 45 cycles consisting of 95°C for 30 s, 50°C for 30 s, and 72°C for 30 s, followed by final elongation at 72°C for 10 min.

The amplification of DNA in library MAM2 was prepared in a 25 µl reaction containing 1x PCR buffer (Qiagen), 2 mM $MgCl_2$, 0.2 mM dNTPs (Qiagen), 0.4 µM of each primer, 0.625 U of HotStartTaq DNA polymerase (Qiagen), 2 µg BSA (Thermo Fisher Scientific), 2 µM human DNA blocker (Boessenkool et al., 2012) and 2 µl DNA template. PCR was used for the enzyme activation at 95°C for 15 min and then 45 cycles consisting of 95°C for 30 s, 60°C for 30 s, and 72°C for 30 s, followed by final elongation at 72°C for 10 min.

The amplification of DNA in library WORM was done in a 25 µl reaction containing 1x PCR buffer (Qiagen), 1.5 mM $MgCl_2$, 0.2 mM dNTPs (Qiagen), 0.2 µM of each primer, 0.625 U HotStartTaq DNA polymerase (Qiagen), 4 µg BSA (Thermo Fisher Scientific), and 2 µl DNA template. PCR was used for the enzyme activation at 95°C for 15 min and then 45 cycles consisting of 94°C for 30 s, 50°C for 30 s, and 72°C for 30 s, followed by final elongation at 72°C for 10 min.

After PCR, 5 µl of all PCR negatives and blanks were checked on a 2% agarose gel. We observed no visible bands on the gel, confirming no considerable contamination from PCR processing. Select samples and the PCR positive control were also checked to confirm successful DNA amplification. Amplicons were further processed only if all negatives were clean and most samples confirmed the success of the preceding PCRs. After gel visualization, the libraries were pooled to equivolumes and purified. MAM1, MAM2 and PLAN were purified with the QIAquick PCR purification kit (Qiagen) following the manufacturer's protocol. The library WORM presented nonspecific amplification of DNA, so we further purified the library by selecting DNA of expected lengths from the gel, followed by gel purification using QIAquick Gel Extraction kit (Qiagen).

Library preparation was finished at external facilities (Novogene (UK) Company Limited) following the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB no. E7645) without the second PCR step. This was done following New England Biolabs instructions to use NEBNext Multiplex Oligos for Illumina with the NEBNext Ultra II DNA Library Prep Kit for Illumina, which essentially transforms it into the PCR-free version. The libraries were then sequenced on an Illumina platform, either HiSeq 4000 (WORM) or NovaSeq 6000 (MAM1, MAM2, and PLAN).

2.3 | Bioinformatics

Data quality was checked with fastQC (Andrews, 2010) and all subsequent steps from raw data to assignment were carried out using OBITools3 (Boyer et al., 2016) through the following data processing pipeline: (1) Reads were merged using the function obi alignpairedend that uses no parameters. Only aligned reads were kept, using the function obi grep with the parameter -a mode: alignment. (2) Aligned reads were then demultiplexed with the function obi ngsfilter, allowing for no mismatches on the tags with the parameter −e 0. (3) The file was dereplicated by removing strictly identical sequences with obi uniq, keeping only the counts in each sample. (4) Quality filtering of sequences shorter that 40 base pairs (bps) or less than 10 counts in the whole library were removed using obi grep. We also removed additional low-repeated sequences by hand after assignment. (5) PCR and sequencing errors were eliminated with the function obi clean. The data kept were the head sequences (-H option) with no variants, presenting a count greater than 5% of their own count (−r 0.05). (6) Taxonomic assignment was performed with the obi ecotag function to a 97% similarity threshold with the best match. The database, composed of the relevant genomic region for each primer pair and the whole kingdom, plants for PLAN and animals for the other libraries. These databases were downloaded from NCBI by searching the genome region and by filtering out the taxonomy down to the kingdom of interest. The 16 S gene database for animals was downloaded on the 9 December 2020 and the trnL gene region database for plants on 14 January 2021. Low repeated molecular operational taxonomic units (MOTUs) with <100 read counts for the whole library were removed while all other assignments were grouped to the deepest taxonomic level. Note that this data set, which includes very low read counts (<10 in a sample), was used for regression analysis in the context of tag jumping, because even low numbers of reads may be explained by the effects of tag jumping. Additionally, low read count samples (<10 counts) were removed prior to clustering analysis and analysis of similarities, following standard practice for curation protocols.

2.4 | Testing the validity of the T4 DNA polymerase model

To test the validity of the T4 DNA polymerase tag jumping model proposed by Schnell et al. (2015), we initially relied on a visual inspection of how reads of rare taxa were distributed within the tag matrices. The rationale behind this is that for rare taxa, tag jumping patterns between samples (column-wise or row-wise) should be more easily detectible, given that these species are expected

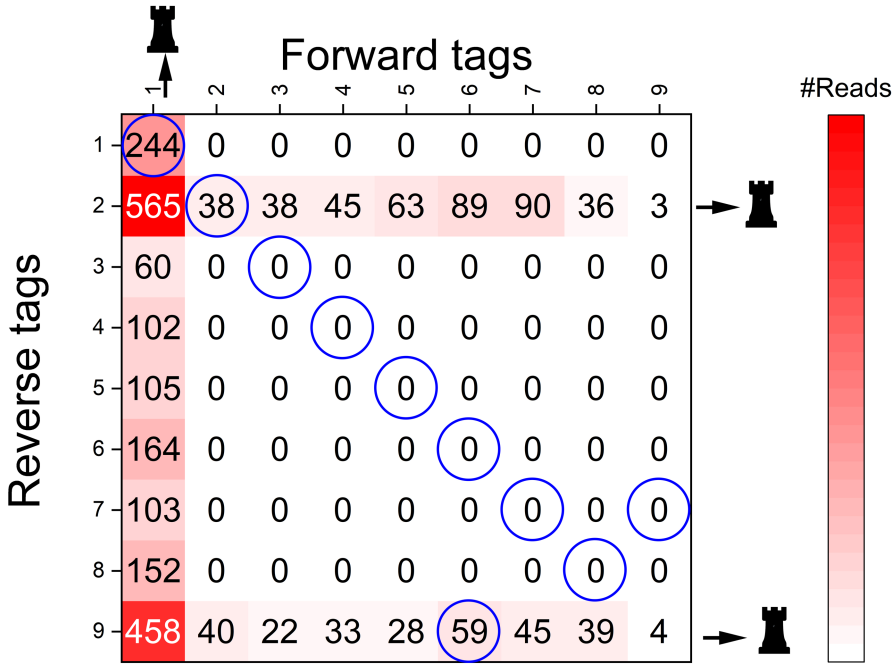

FIGURE 2 An example of the distribution of reads for a rare species (Lepus sp.) in one of our sedimentary DNA libraries (MAM1). Reads are shown as a function of the topologic space of the tagging matrix where forward primers are shown as rows and reverse primers as columns. Blue circles represent PCR-blanks. Samples were originally randomly distributed within the tagging matrix, and the observed linear distribution of reads within the matrix (as indicated by the rook symbol) is a pattern to be expected from T4 polymerase-mediated tag jumping.

to be absent in most samples. We followed the distribution of rare taxa within the tags matrix and searched for cross-like patterns in abundance emerging along horizontal (row) and vertical (column) directions.

To statistically test to what extent a samples' topological position could be related to artefactual sequences formed by tag jumps, we performed a regression analysis using a topological descriptor modelling the heteroduplex formation (reads in the row × reads in the column) and the number of reads in our blanks after which residuals were checked for normality. All regression statistics were performed in MATLAB (MathWorks). To further scrutinize the impact of tag jumping, sequenced species data from library MAM2 was imported into the program Primer-e and transformed into presence-absence data. The latter data set was used to build a Jaccard's similarity matrix, where we clustered samples using the group average method with a cophenetic correlation coefficient of 0.8967. In addition, we conducted a multidimensional similarity analysis (ANOSIM) where we used PCR and extract replicates as groups to assess within group variance and test for significant ($p < .05$) differences between pairwise comparisons.

3 | RESULTS

All results of sequenced data can be found in Table S1, but a library specific summary follows below.

For the library MAM1 the achieved sequencing depth was 16,567,225 reads, of which 3,699,003 reads were grouped into 2221 MOTUs. In this library, 9.6% of the reads were found in blanks, with an average of 16,631 reads (range: 53–35,305). In comparison, experimental (sediment and soil) samples constituted 81.6% of all reads with an average of 29,304 reads (range: 10,780–82,923). In total, 18 genera were identified. On average 13 genera were identified in experimental samples (range: 9–17) while 12 were detected in blanks (range: 2–15).

In MAM2 13,703,604 reads were sequenced, of which 273,3421 reads were assigned to 898 MOTUs. In this library, only 0.6% of the reads were found in blanks, with an average of 2195 (range: 17–4035), while 98% were found in experimental samples with an average of 67,892 reads per sample (range: 1241–122,734). Twelve taxa were found in MAM2 with an average nine taxa in each experimental sample (range: 6–10) and 8 taxa in blanks (range: 5–10).

For library PLAN the sequencing depth was 12,269,751 reads, of which 6,978,779 reads were assigned in 4883 MOTUs. About 1.7% of the reads were found in blanks, with an average of 6162 reads per blank (range: 192–27,389), while 89.7% were found in experimental samples (range: 15,028–132,014). A total of 44 taxa were detected. On average 40 (range: 35–44) taxa was found in the experimental samples and 33 in blanks (range: 18–40).

Finally, library WORM had a total of 11,598,479 reads, of which 6,015,348 reads could be assigned to 961 different MOTUs. In this library, 3% of the reads were found in blanks, with an average of 25,014 reads per blank (range: 8347–65,677). In comparison, 96% of the reads originated from experimental samples, with an average of 119,586 reads per sample (range: 15,028–540,308). For library WORM, nine taxa were found in total with an average, seven taxa were detected in experimental samples (range: 5–9) and six in blanks (range: 5–6).

In line with the proposed heteroduplex tag jumping model (Figure 1), we found counts of rare species forming horizontal and vertical patterns in the tags matrix as exemplified in Figure 2. High absolute numbers of reads (and relative number of reads) were systematically found in samples originally placed within the same column and/or rows within the tag matrix. Here, rare species reads

were always highest at the point of intersection, where horizontal and vertical tagging patterns crossed. Importantly, replicates positioned in other places around the tag matrix did not show positive detections.

Through the four libraries, 0.6% to 9.6% of the total reads detected in the PCR blanks had tag combinations that did not match any sample. In all four data sets we examined, the total number of reads observed in the PCR blanks could be largely explained by a linear regression taking into account the topological position (row × column) of the PCR blank and the total number of reads associated with each of the tags (Table S2, Figure 3a–d). This relationship between tag combination and counts in the PCR blanks was consistent through all four libraries, even when separated by taxa. For many of the regression lines, the coefficient of determination (R^2) was above 0.8, and the correlations were significant ($p < .05$) for 67 out of 82 taxa. Despite the relatively small percentage of reads in the blanks, all the samples contained a high number of taxa with counts within the regression confidence interval and, therefore, indistinguishable from the baseline tag jumping effect (Figure 4).

The regressions predicting the tag jumping "noise" based on the number of artefactual reads in the PCR blanks and their matrix position allowed us to calculate the expected number of artefactual reads, given the topological position of each sample. The estimated proportion of reads derived from tag jumping relative to the total number of measured reads was used to group detections into four categories of uncertainty: (1) Unreliable detections, that is, taxa in which the total number of reads was within the upper 95% confidence interval of the regression predicting number of artefactual reads per sample as a function of matrix position; (2) uncertain detections, taxa in which total counts were only ≤50% higher than those predicted to be artefactual sequences (inferred using the 95% confidence boundaries of the regression analyses); (3) reliable detections, that is, samples with counts ≥50% higher than the predicted artefactual sequences; and (4) detections outside the limits, that is, taxa where their position within the tag matrix were not covered by the positions of the PCR blanks used to derive the regressions. Our classification of detections found that a considerable proportion of the taxa was classified as unreliable (Figure 4). That proportion varied from 40% in library PLAN to as high as 80% in library WORM.

In line with these high numbers of unreliable detections, we found that sample similarities were more dependent on their tag combination than on their shared background (i.e., sourcing from the same sample), as highlighted by our cluster analysis (Figure 5). Surprisingly, sample replicates and blanks showed no clear pattern of similarity. Some few replicates that showed similarities with another replicate also shared a similar tag (Figure 5a). Pairwise comparison of replicate groups suggested that there were no significant ($p > .05$) differences between most groups (i.e., when samples were

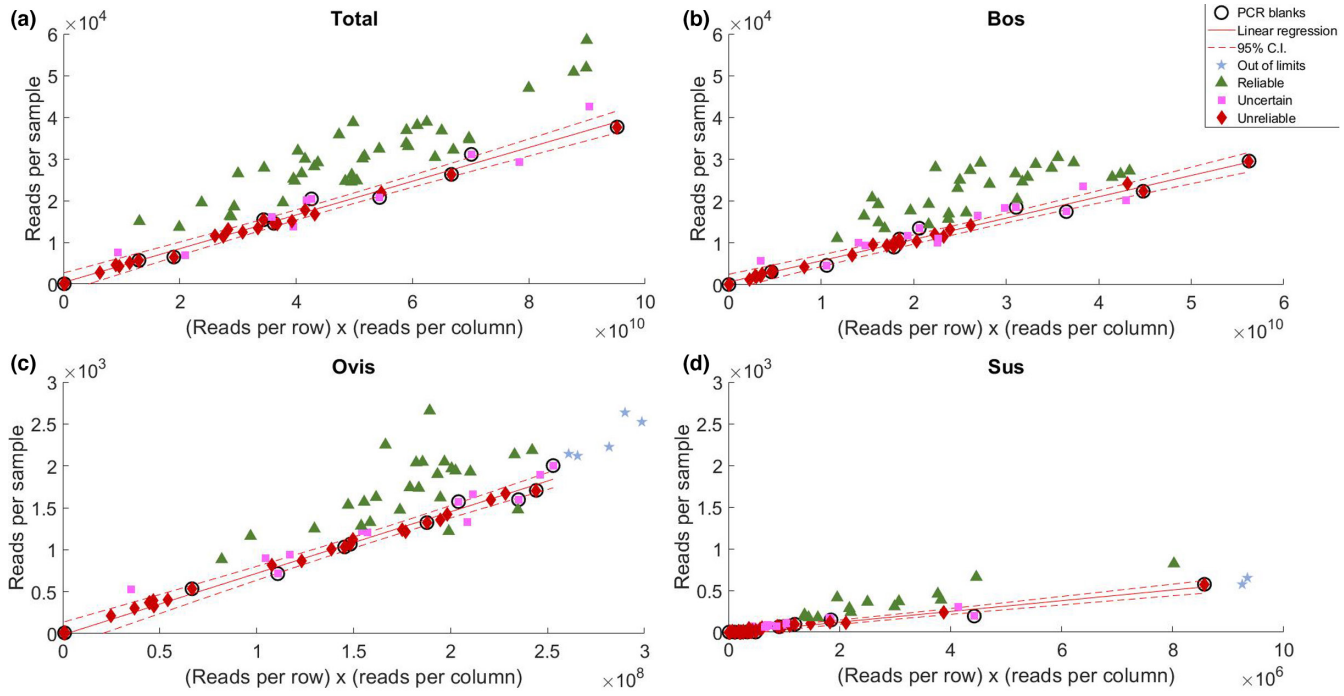

FIGURE 3 Linear regressions (red line) between a metric describing the numbers of reads and their position in the tagging matrix and number of assigned read per sample. Data is shown for all taxa combined (panel a), for Bos (panel b), for Ovis (panel c) and for Sus (panel d). Red dotted lines denote the 95% confidence interval around each regression line. Symbols indicate how samples were classified according to our four categories of uncertainties: Unreliable samples (red diamonds), uncertain samples (pink squares), reliable samples (green symbols) and samples judged outside the limits of the regression (blue stars). Black circles enclose the PCR blanks that were used to construct the regression. Equations for the regression lines reads: Upper left panel (Total reads = 4.383*10−7*X − 301.76); upper right panel (Bos reads = 5.248*10−7*X + 38102.2); lower left panel (Ovis reads = 7.208*10−6*X − 13.025); and lower right panel (Sus reads =6.6327*10−5*X − 8.407). Not axis have been cut and scaled to better show the regressions.

similar). In fact, we could only find significant differences between two replicate groups despite their vastly different sample origins.

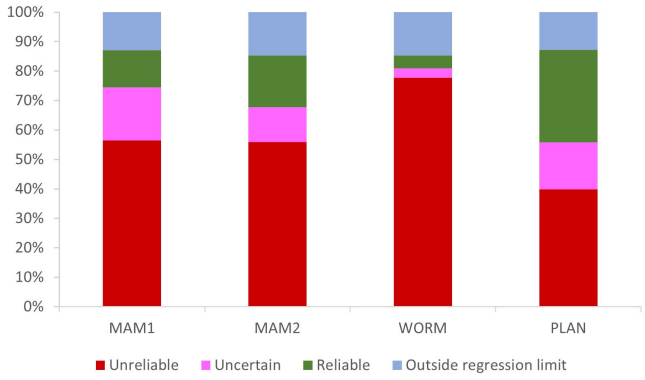

FIGURE 4 Categories of sample uncertainties for each of the four studied libraries (MAM1, MAM2, WORM, and PLAN). Bars show the percentage of taxa, averaged per sample for each library (taxa in category/taxa in sample). Categories representing levels of uncertainty (shown in colours) are based on the predicted number of artefactual sequenced counts using the regressions in Figure 3 and the actual number of reads in the sample (see result section for a detailed description on how categories were calculated).

Instead, similarities between clusters were primarily explained by their sharing of tags (Figure 5b).

When removing detections classified as unreliable or uncertain (categories 1 and 2), some taxa were strongly affected by this filtering. For example, sequences of *Cervus* in library MAM1 were indistinguishable from the predicted tag jumping noise in most of the replicates and were removed from 72 of 81 occurrences (Figure 6). Others, such as *Capra* or *Sus*, experienced dramatic shifts in distribution after filtering. When focusing only on replicates classified as reliable (category 3), *Capra* or *Sus* were only found to be present in a few samples, while the unfiltered data suggested a nearly ubiquitous distribution of these rare taxa across samples. Nevertheless, distribution of some taxa remained relatively unaffected by our filtering approach. For example, *Bos* or *Ovis* distributions with the samples remained largely intact, even after removing the unreliable samples. In MAM1, each taxon was on average found to be reliable in only every ninth sample—a frequency that illustrates the strong impacts that tag jumping imposes on the data set (see Table S3 for a complete breakdown by taxa in all libraries). We also note that the PCR blanks, which did not contain DNA or primers and should therefore be free from any sequences, became clean after our topology-based filtering. In contrast to the original data set, the analysis of similarity in the denoised data proves a significant difference between groups

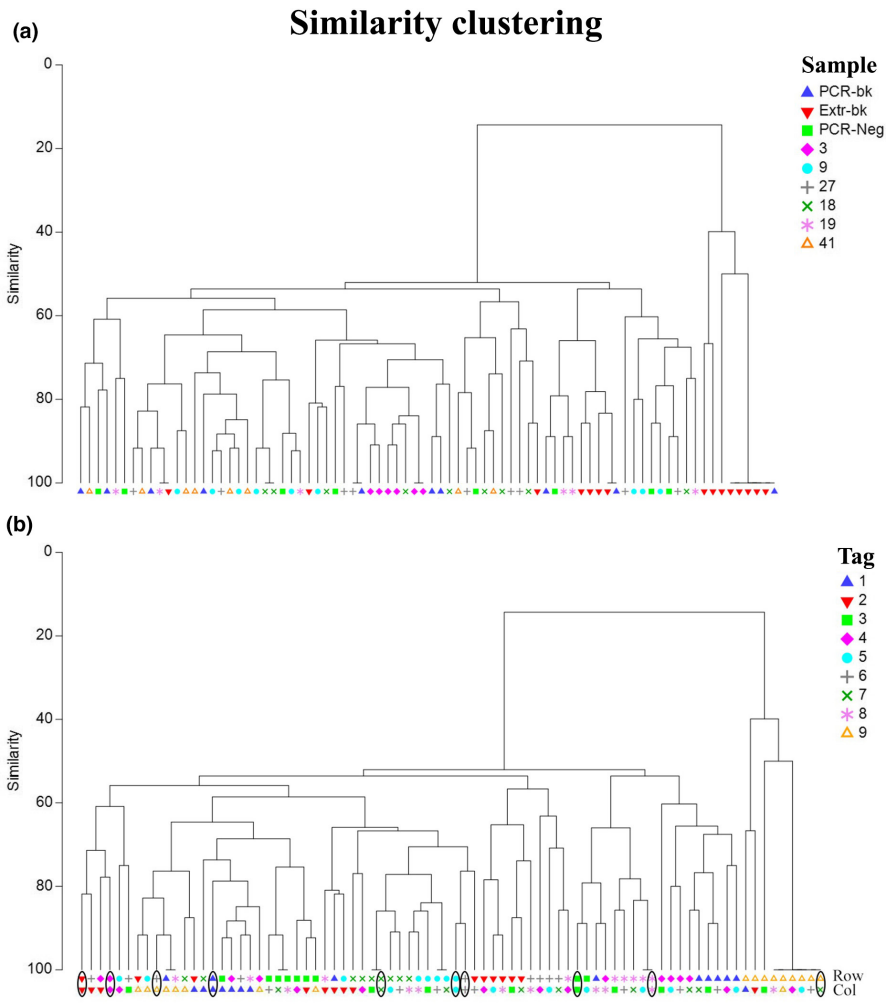

FIGURE 5 Cluster dendrogram constructed from the Jaccard's similarities matrix for the analysed samples. Black branches represent clusters identified in the analysis to share similarities (cophenetic correlation coefficient = 0.8967). Note the black ellipses marking PCR blanks combinations and thus, samples that should be free from taxa. In panel (a) replicates share the same coloured symbol. Here, the spread of coloured symbols at the base of the dendrogram illustrate that most replicates are dissimilar to each other. In panel (b), tags combination used for each samples are shown at the base. Note how samples identified as being similar to each other.

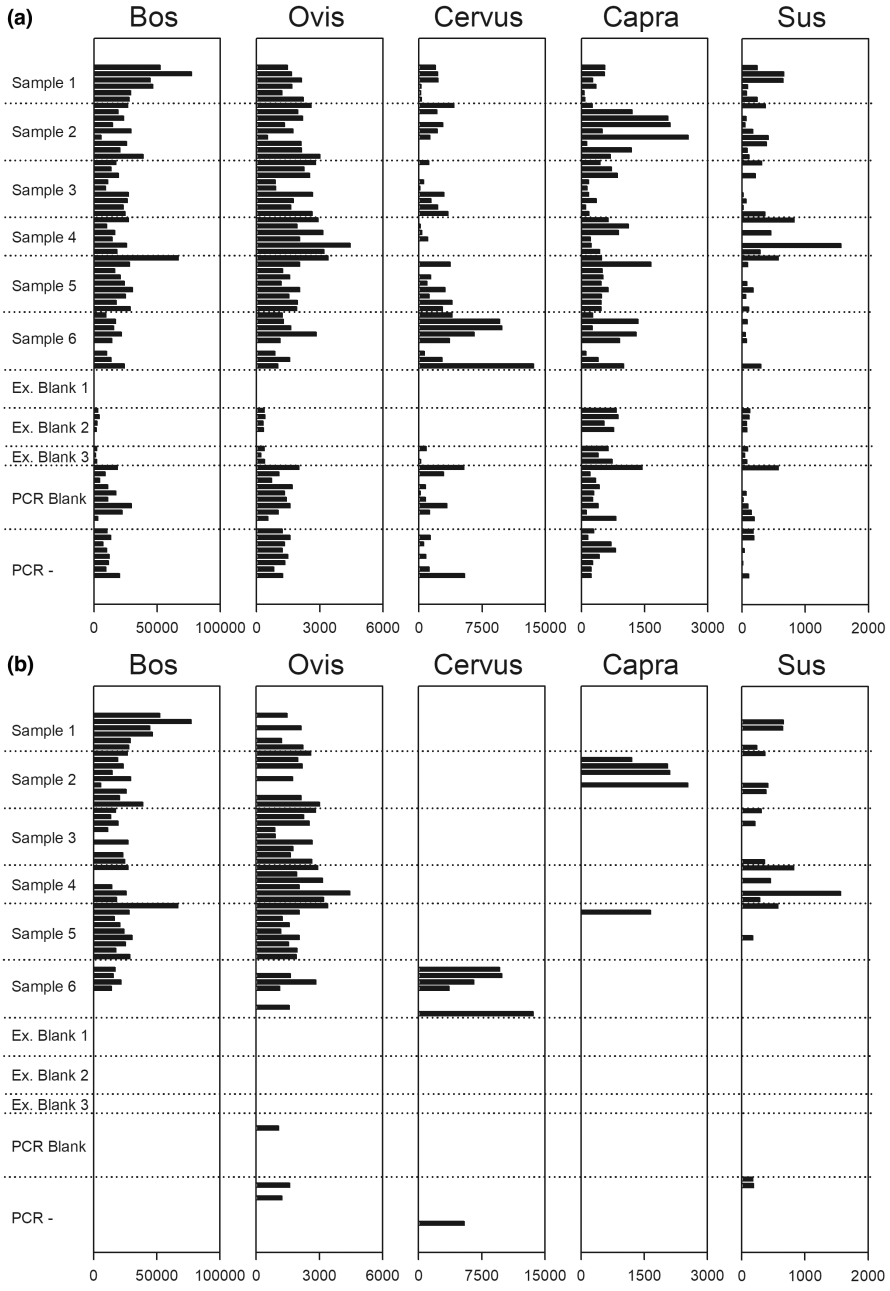

FIGURE 6 Comparison of sequence data from library MAM2 before (a) and after (b) filtering away samples predicted to be strongly affected by tag jumping according our topologic based model. Here, samples that were removed were part of the categories classified as unreliable and uncertain (see result section). Note that some taxa (such as Cervus or Sus) were present across many samples in the unfiltered dataset (a) but had well constrained distribution after filtering away samples strongly affected by tag jumping (b).

of replicates, with nine of 15 comparisons being significant in the pairwise analysis.

4 | DISCUSSION

The T4 DNA polymerase-driven tag jumping model, first proposed by van Orsouw et al. (2007) and further developed by Schnell et al. (2015), can explain most of the observed noise in our data. This model predicts how artefactual sequences are distributed along columns or rows of the tag matrix, and is also consistent with the strong positive correlation between reads per row and column and the number of detected sequences present in the blanks. Given that Schnell et al. (2015) point towards the activity of the T4 DNA polymerase as one of the main drivers of tag jumps, and that this polymerase was also used in our end-repair step, it appears that our results strongly support the previous inferences made by Schnell et al. (2015).

Metabarcoding and high-throughput sequencing technologies have repeatedly proven their importance and utility in environmental studies. Nevertheless, there is an urgent need to improve the underlying methods and account for tag jumping effects, as has been previously suggested (Esling et al., 2015; Schnell et al., 2015; Zinger et al., 2019). In that sense, the reliability of positive detections is obviously a key element. As we show here, tag jumping can produce major distortions of sequenced data and—if not accounted for—can lead to incorrect interpretations of the data. The latter is evident from the striking contrast between our topology-based filtering and the original data sets (Figure 6a,b). Hence, we seemingly validate our main hypothesis that the artefacts

introduced by tag jumps are potent enough to substantially distort the conclusions of metabarcoding studies. Clearly, there is an urgent need to recognize impacts of tag jumping in environmental DNA data sets and to understand its underlying mechanism(s). Regarding the former, it is important to highlight that the use of multiple replicates cannot rule out false detections induced by tag jumping, as it is possible that several replicates could present as false positives if they share one of the tags. While tag jumping may have grave implications for environmental DNA studies, our findings suggest that tag jumping generates a predictable pattern that not only provides routes for detecting affected data and reducing its impacts, but also offers key insights into the mechanism underlying the process.

Importantly, tag jumping should not be regarded as a minor effect that can be removed by simply raising the detection limit based on the number of reads observed in laboratory blanks. This is especially true considering that most studies use few controls (i.e., only 1 or 2 control blanks), which means they cannot fully capture the topological patterns we identified in our data sets that strongly suggest tag jumps are unevenly distributed across the tagging matrix. In other words, assuming that artefactual sequences present in one and/or a few blank samples are representative of all samples within the tagging matrix is too simplistic and might generate flawed data. We also show that the overall impact of tag jumping on a data set is an artificially inflated similarity between samples sharing a tag (e.g., Figure 5), as well as a distortion of taxa presence (e.g., Figure 6). Clearly, our data proves that PCR-free indexing methods, which have been suggested to reduce tag jumping (Esling et al., 2015), are not enough to assure that samples are not affected by tag jumps.

It was only after "denoising" our data sets with the algorithms we developed that differences between individual sample groups became significant. That is, tag jumping initially masked the community differences between groups in the original sequenced data. This finding suggesting that tag jumping increased false positive detections to the point that our sample communities were no longer different indicates that this process can strongly influence interpretations of metabarcoding data. It is important to note that tag jumping does not necessarily generate homogenizing effects, as differences between groups could also be artificially inflated—the direction and magnitude of the effects will ultimately depend on the arrangement of samples in the tagging matrix. For example, sample replicates could be placed in the same row/column of the tagging matrix, which would cause jumps to artificially lower between-replicate variance and thereby increase the likelihood of finding statistical differences between these samples and others.

Intriguingly, we also found that the impact of tag jumping varied substantially between our four libraries. These library-specific differences may be the result of differences in the amplified barcoding regions. Previous studies have shown that chimera rates are higher when barcodes share a high degree of similarity (Shin et al., 2014; Wang & Wang, 1997). As similarities between barcodes modulate the interaction between DNA strands, we argue that barcode-specific variation can also be expected in tag jumping. In addition, small differences in sample post-processing, GC content, time of processing, or temperature during purification and/or storage prior to sequencing may also contribute to variable levels of tag jumping between libraries. Additional research is needed to fully evaluate effects of these factors on the tag jumping process.

4.1 | Current practices and the way forward

There are generally three methods for library preparation: (1) One-step PCR, (2) two-step PCR and (3) adapter ligation (Bruce et al., 2021). Importantly, one-step PCR and two step PCR (including Phusion primers) are not likely affected by the tag jumping process as shown to affect libraries in our study. However, adapter ligation is sensitive to the tag jumping. According to Illumina, this latter technique has been cited in almost 10,000 publications. Whether these studies were affected by tag jumping or not is unknown, but there is a rationale for expecting a large number of them may have been affected by tag jumping.

We suggest that adequately reporting tag jumping, and, ideally, preventing these processes is important for all future metabarcoding studies that use adapter ligation. This is particularly relevant for environmental DNA studies where the validity of a positive detection within complex DNA templates cannot be verified through comparison with in situ monitoring data (for example, when used in palaeoecological reconstructions where reconstructed taxa cannot always be validated against historical documentations). We suggest four practices that can be used to reduce the impacts arising from tag jumping (see Data S1 for more details):

1. **Twin-tagging**. Using each tag only once eliminates all cross contamination mediated by heteroduplex driven processes. Examples of the use of this tagging technique can be found for example in "An appetite for pests: Synanthropic insectivorous bats exploit cotton pest irruptions and consume various deleterious arthropods" (Cohen et al., 2020) or "Amazonian mammal monitoring using aquatic environmental DNA" (Coutant et al., 2021). However, this tagging strategy substantially increases analytical costs and workloads.
2. **Tagsteady protocol**. This alternative library preparation method circumvents end-repair and second PCR effectively eliminating tag jumping (Carøe & Bohmann, 2020).
3. **Use of a systematic blank system within the tags matrix**. By using multiple blanks embedded in the matrix in a way that every column and every row has at least one PCR blank, it is possible to characterize tag jumping and filter out affected samples. That is, by using regressions predicting the baseline noise as shown in our study. While this approach cannot completely remove tag jumping from data sets, it can substantially improve the reliability of results and sharpen data curation procedures.
4. **Skip the end-repair step**. Polymerase activity during the end repair step seems to explain the occurrence of tag jumping in our data. By removing the end-repair step, the tag jumping issue is probably reduced, although the impact on DNA sequencing beyond tag jumping is unknown.

4.2 | Conclusions

Artefactual sequences arising due to tag jumps leave behind a topological signature in metabarcoding data sets. This signature can be traced back to each samples' original position in the tag matrix, wherein the abundance of artefactual sequences form a "cross-like" pattern. In this study, where chimera from second-step PCR can be ruled out as a possible mechanism behind tag jumping, the topological pattern is consistent with a process caused by a 3′-end digestion occurring through the activity of the T4 DNA polymerase. We highlight that tag jumping is a major concern for metabarcoding studies: the degree of false positivity in our data sets was high enough to cause different sample communities to become statistically indistinguishable from one another. The high number of false positives in our PCR blanks strongly suggests that when attempting to curate data sets, the contaminating influence from tag jumping cannot be avoided by setting a threshold for reliable data based on the number of reads found in the blanks– a procedure commonly done in the published literature. While multiple sample replicates can be used to detect the effects of tag jumping, increasing the number of replicates cannot, by itself, solve the problem. Only adequate laboratory protocols and careful inspection of the data can guarantee data unaffected by tag jumping.

ACKNOWLEDGEMENTS

We thank the Swedish research council for founding JK (Dnr 2017-04548), the Swiss National Science Foundation for founding MAM (grant no. P2BEP2-188256) and four anonymous reviewers that helped us to improve our study.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at [https://doi.org/10.5061/dryad.tdz08kq33].

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Dryad at https://doi.org/10.5061/dryad.tdz08kq33.

ORCID

Saul Rodriguez-Martinez https://orcid.org/0000-0001-9791-405X

REFERENCES

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Available online. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bienert, F., De Danieli, S., Miquel, C., Coissac, E., Poillot, C., Brun, J.-J., & Taberlet, P. (2012). Tracking earthworm communities from soil DNA. *Molecular Ecology*, *21*(8), 2017–2030. https://doi.org/10.1111/j.1365-294X.2011.05407.x

Binladen, J., Gilbert, M. T., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*, *2*(2), e197. https://doi.org/10.1371/journal.pone.0000197

Boessenkool, S., Epp, L. S., Haile, J., Bellemain, E., Edwards, M., Coissac, E., Willerslev, E., & Brochmann, C. (2012). Blocking human contaminant DNA during PCR allows amplification of rare mammal species from sedimentary ancient DNA. *Molecular Ecology*, *21*(8), 1806–1815. https://doi.org/10.1111/j.1365-294X.2011.05306.x

Bohmann, K., Monadjem, A., Lehmkuhl Noer, C., Rasmussen, M., Zeale, M. R., Clare, E., … Gilbert, M. T. (2011). Molecular diet analysis of two african free-tailed bats (molossidae) using high throughput sequencing. *PLoS One*, *6*(6), e21441. https://doi.org/10.1371/journal.pone.0021441

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). OBITOOLS: A UNIX -inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182. https://doi.org/10.1111/1755-0998.12428

Bruce, K., Blackman, R., Bourlat, S. J., Hellström, A. M., Bakker, J., Bista, I., Bohmann, K., Bouchez, A., Brys, R., Clark, K., Elbrecht, V., Fazi, S., Fonseca, V., Hänfling, B., Leese, F., Mächler, E., Mahon, A. R., Meissner, K., Panksep, K., … Deiner, K. (2021). *A practical guide to DNA-based methods for biodiversity assessment* (pp. 54–57). Pensoft Publishers. https://doi.org/10.3897/ab.e68634

Carøe, C., & Bohmann, K. (2020). Tagsteady: A metabarcoding library preparation protocol to avoid false assignment of sequences to samples. *Molecular Ecology Resources*, *20*(6), 1620–1631. https://doi.org/10.1111/1755-0998.13227

Chariton, A. A., Batley, G. E., & Hardy, C. M. (2010). Improving aquatic ecological assessments. *Integrated Environmental Assessment and Management*, *6*(1), 187–188. https://doi.org/10.1002/ieam.22

Clarke, C. L., Edwards, M. E., Brown, A. G., Gielly, L., Lammers, Y., Heintzman, P. D., Ancin-Murguzur, F. J., Bråthen, K.-A., Goslar, T., & Alsos, I. G. (2019). Holocene floristic diversity and richness in Northeast Norway revealed by sedimentary ancient DNA (*sed* a DNA) and pollen. *Boreas*, *48*(2), 299–316. https://doi.org/10.1111/bor.12357

Cohen, Y., Bar-David, S., Nielsen, M., Bohmann, K., & Korine, C. (2020). An appetite for pests: Synanthropic insectivorous bats exploit cotton pest irruptions and consume various deleterious arthropods. *Molecular Ecology*, *29*(6), 1185–1198. https://doi.org/10.1111/mec.15393

Coutant, O., Richard-Hansen, C., de Thoisy, B., Decotte, J., Valentini, A., Dejean, T., Vigouroux, R., Murienne, J., & Brosse, S. (2021). Amazonian mammal monitoring using aquatic environmental DNA. *Molecular Ecology Resources*, *21*(6), 1875–1888. https://doi.org/10.1111/1755-0998.13393

Deagle, B. E., Kirkwood, R., & Jarman, S. N. (2009). Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, *18*(9), 2022–2038. https://doi.org/10.1111/j.1365-294X.2009.04158.x

Esling, P., Lejzerowicz, F., & Pawlowski, J. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, *43*(5), 2513–2524. https://doi.org/10.1093/nar/gkv107

Fernandez, S., Rodriguez, S., Martinez, J. L., Borrell, Y. J., Ardura, A., & Garcia-Vazquez, E. (2018). Evaluating freshwater macroinvertebrates from eDNA metabarcoding: A river Nalon case study. *PLoS One*, *13*(8), e0201741. https://doi.org/10.1371/journal.pone.0201741

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*(4), 423–425. https://doi.org/10.1098/rsbl.2008.0118

Giguet-Covex, C., Pansu, J., Arnaud, F., Rey, P.-J., Griggo, C., Gielly, L., Domaizon, I., Coissac, E., David, F., Choler, P., Poulenard, J., & Taberlet, P. (2014). Long livestock farming history and human landscape shaping revealed by lake sediment DNA. *Nature Communications*, *5*(1), 3211. https://doi.org/10.1038/ncomms4211

Glaus, N. (2018). *Holocene flood events and sediment flux in alpine lake Grosssee (Flumserberg, Switzerland)*. University of Bern Master Thesis.

Haile, J., Froese, D. G., Macphee, R. D., Roberts, R. G., Arnold, L. J., Reyes, A. V., Rasmussen, M., Nielsen, R., Brook, B. W., Robinson, S., Demuro, M., Gilbert, M. T., Munch, K., Austin, J. J., Cooper, A., Barnes, I., Möller, P., & Willerslev, E. (2009). Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(52), 22352–22357. https://doi.org/10.1073/pnas.0912510106

Jeunen, G.-J., Lipinskaya, T., Gajduchenko, H., Golovenchik, V., Moroz, M., Rizevsky, V., Semenchenko, V., & Gemmell, N. (2021). Environmental DNA (eDNA) metabarcoding surveys extend the range of invasion for non-indigenous freshwater species in Eastern Europe. [Preprint]. *Ecology*. https://doi.org/10.1101/2021.05.16.444374

Jo, H., Choi, B., Park, K., Kim, W.-S., & Kwak, I.-S. (2020). First gut content analysis of 4th instar midge larvae (Diptera: Chronomidae) In large-scale weirs using a DNA meta-barcoding approach. *International Journal of Environmental Research and Public Health*, *17*(8), 2856. https://doi.org/10.3390/ijerph17082856

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, *40*(1), e3. https://doi.org/10.1093/nar/gkr771

Lacoursière-Roussel, A., Howland, K., Normandeau, E., Grey, E. K., Archambault, P., Deiner, K., Lodge, D. M., Hernandez, C., Leduc, N., & Bernatchez, L. (2018). eDNA metabarcoding as a new surveillance approach for coastal Arctic biodiversity. *Ecology and Evolution*, *8*(16), 7763–7777. https://doi.org/10.1002/ece3.4213

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb prot5448. https://doi.org/10.1101/pdb.prot5448

Paijmans, J. L. A., González Fortes, G., & Förster, D. W. (2019). Application of solid-state capture for the retrieval of small-to-medium sized target loci from ancient DNA. In B. Shapiro, A. Barlow, P. D. Heintzman, M. Hofreiter, J. L. A. Paijmans, & A. E. R. Soares (Eds.), *Ancient DNA* (pp. 129–139). Springer New York. https://doi.org/10.1007/978-1-4939-9176-1_14

Pegard, A., Miquel, C., Valentini, A., Coissac, E., Bouvier, F., François, D., Taberlet, P., Engel, E., & Pompanon, F. (2009). Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from feces. *Journal of Agricultural and Food Chemistry*, *57*(13), 5700–5706. https://doi.org/10.1021/jf803680c

Quéméré, E., Hibert, F., Miquel, C., Lhuillier, E., Rasolondraibe, E., Champeau, J., Rabarivola, C., Nusbaumer, L., Chatelain, C., Gautier, L., Ranirison, P., Crouau-Roy, B., Taberlet, P., & Chikhi, L. (2013). A DNA metabarcoding study of a primate dietary diversity and plasticity across its entire fragmented range. *PLoS One*, *8*(3), e58971. https://doi.org/10.1371/journal.pone.0058971

Schnell, I. B., Bohmann, K., & Gilbert, M. T. (2015). Tag jumps illuminated—Reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*(6), 1289–1303. https://doi.org/10.1111/1755-0998.12402

Shehzad, W., Riaz, T., Nawaz, M. A., Miquel, C., Poillot, C., Shah, S. A., Pompanon, F., Coissac, E., & Taberlet, P. (2012). Carnivore diet analysis based on next-generation sequencing: Application to the leopard cat (Prionailurus bengalensis) in Pakistan. *Molecular Ecology*, *21*(8), 1951–1965. https://doi.org/10.1111/j.1365-294X.2011.05424.x

Shin, S., Lee, T. K., Han, J. M., & Park, J. (2014). Regional effects on chimera formation in 454 pyrosequenced amplicons from a mock community. *Journal of Microbiology*, *52*(7), 566–573. https://doi.org/10.1007/s12275-014-3485-6

Soininen, E. M., Valentini, A., Coissac, E., Miquel, C., Gielly, L., Brochmann, C., Brysting, A. K., Sønstebø, J. H., Ims, R. A., Yoccoz, N. G., & Taberlet, P. (2009). Analysing diet of small herbivores: The efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in Zoology*, *6*, 16. https://doi.org/10.1186/1742-9994-6-16

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C., & Willerslev, E. (2007). Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, *35*(3), e14. https://doi.org/10.1093/nar/gkl938

Taylor, P. G. (1996). Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Molecular Biology and Evolution*, *13*(1), 283–285. https://doi.org/10.1093/oxfordjournals.molbev.a025566

Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E., Coissac, E., Pompanon, F., Gielly, L., Cruaud, C., Nascetti, G., Wincker, P., Swenson, J. E., & Taberlet, P. (2009). New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: The trnL approach. *Molecular Ecology Resources*, *9*(1), 51–60. https://doi.org/10.1111/j.1755-0998.2008.02352.x

van der Valk, T., Vezzi, F., Ormestad, M., Dalen, L., & Guschanski, K. (2020). Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Molecular Ecology Resources*, *20*(5), 1171–1181. https://doi.org/10.1111/1755-0998.13009

van Orsouw, N. J., Hogers, R. C., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., van der Poel, H., van Oeveren, J., Verstegen, H., & van Eijk, M. J. (2007). Complexity reduction of polymorphic sequences (CRoPS): A novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One*, *2*(11), e1172. https://doi.org/10.1371/journal.pone.0001172

Wackett, A. A., Yoo, K., Olofsson, J., & Klaminder, J. (2018). Human-mediated introduction of geoengineering earthworms in the Fennoscandian arctic. *Biological Invasions*, *20*(6), 1377–1386. https://doi.org/10.1007/s10530-017-1642-7

Wang, G. C., & Wang, Y. (1997). Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Applied and Environmental Microbiology*, *63*(12), 4645–4650. https://doi.org/10.1128/AEM.63.12.4645-4650.1997

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., de Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., … Taberlet, P. (2019). DNA metabarcoding-need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*(8), 1857–1862. https://doi.org/10.1111/mec.15060

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Rodriguez-Martinez, S., Klaminder, J., Morlock, M. A., Dalén, L., & Huang, D.-T. (2023). The topological nature of tag jumping in environmental DNA metabarcoding studies. *Molecular Ecology Resources*, *23*, 621–631. https://doi.org/10.1111/1755-0998.13745