MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2015) **15**, 1289–1303 doi: 10.1111/1755-0998.12402

Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies

IDA BÆRHOLM SCHNELL,*†[1] KRISTINE BOHMANN*‡[1] and M. THOMAS P. GILBERT*§

**Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350 Copenhagen K, Denmark, †Center for Zoo and Wild Animal Health, Copenhagen Zoo, 2000 Frederiksberg, Denmark, ‡School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK, §Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia 6102, Australia*

Abstract

Metabarcoding of environmental samples on second-generation sequencing platforms has rapidly become a valuable tool for ecological studies. A fundamental assumption of this approach is the reliance on being able to track tagged amplicons back to the samples from which they originated. In this study, we address the problem of sequences in metabarcoding sequencing outputs with false combinations of used tags (tag jumps). Unless these sequences can be identified and excluded from downstream analyses, tag jumps creating sequences with false, but already used tag combinations, can cause incorrect assignment of sequences to samples and artificially inflate diversity. In this study, we document and investigate tag jumping in metabarcoding studies on Illumina sequencing platforms by amplifying mixed-template extracts obtained from bat droppings and leech gut contents with tagged generic arthropod and mammal primers, respectively. We found that an average of 2.6% and 2.1% of sequences had tag combinations, which could be explained by tag jumping in the leech and bat diet study, respectively. We suggest that tag jumping can happen during blunt-ending of pools of tagged amplicons during library build and as a consequence of chimera formation during bulk amplification of tagged amplicons during library index PCR. We argue that tag jumping and contamination between libraries represents a considerable challenge for Illumina-based metabarcoding studies, and suggest measures to avoid false assignment of tag jumping-derived sequences to samples.

Keywords: chimeras, diversity assessment, environmental DNA, metabarcoding, second-generation sequencing, tag jumping

Received 22 November 2014; revision accepted 2 March 2015

Introduction

Metabarcoding is characterized by identification of multiple taxa in environmental samples through sequencing of amplicons generated with universal primers, for example in soil, water, faeces or bulk samples of entire organisms (e.g. Deagle *et al.* 2009; Pegard *et al.* 2009; Rasmussen *et al.* 2009; Soininen *et al.* 2009; Valentini *et al.* 2009; Bohmann *et al.* 2011). Today, this approach is increasingly applied in ecological studies where it has been used in biodiversity (e.g. Ficetola *et al.* 2008; Chariton *et al.* 2010) and diet studies (e.g. Deagle *et al.* 2009; Soininen *et al.* 2009; Valentini *et al.* 2009), environmental monitoring (e.g. Jerde *et al.* 2011; Nathan *et al.* 2014) and reconstruction of past ecosystems (e.g. Haile *et al.* 2009; Sønstebø *et al.* 2010), to name but a few. This range of questions now being addressed highlights the importance of ensuring that results are reliable.

To exploit second-generation sequencing capacities by simultaneously sequencing amplicons from many samples, extracts are typically PCR-amplified using generic primers to which 5′-nucleotide tags are added (see Fig. 1) (Binladen *et al.* 2007). As the tags leave a PCR-specific mark on the amplicons, amplicons can be pooled for sequencing and bioinformatically traced back to the samples from which they originated. Although originally developed for the first commercially available second-generation sequencing platform, the Roche/454 FLX (Binladen *et al.* 2007; as implemented in e.g. Soininen *et al.* 2009; Valentini *et al.* 2009; Deagle *et al.* 2009; Pegard *et al.* 2009), subsequent studies have explored parallel sequencing of tagged amplicons on the Ion Torrent (e.g. Piñol *et al.* 2014; Murray *et al.* 2013) and, due to its very

Correspondence: Kristine Bohmann,
E-mail: kristinebohmann@gmail.com

[1]These authors contributed equally to this work.

high output, the Illumina series of platforms (e.g. Shehzad *et al.* 2012; Quéméré *et al.* 2013).

Preparation of amplicon pools for Illumina sequencing typically requires a so-called library build (Fig. 1) (Meyer & Kircher 2010). Here, amplicons are often blunt-ended and adapters are ligated to the amplicons followed by a library index PCR amplification (Fig. 1) (e.g. Hope *et al.* 2014). This process may result in a hurdle to metabarcoding studies that, to our knowledge, have not yet been formally explored – the generation of artefactual sequences in which amplicons carry different tags than originally applied (tag jumps). Although sequences with false tag combinations are easily identified and excluded, if this issue is not accounted for in the experimental set-up, tag jumps that create sequences with false, but already used tag combinations may introduce significant levels of misidentifications and artificially inflate diversity in the samples (Fig. 2).

Several metabarcoding studies have noted the occurrence of false combinations of used tags in the sequencing output when sequencing on the Roche/454 sequencing platform (Blaalid *et al.* 2013, 2014; Carew *et al.* 2013; Davey *et al.* 2013, 2014; Lindner *et al.* 2013; Botnen *et al.* 2014) and on the Illumina sequencing platform (Esling *et al.* 2015). This has also been acknowledged in non-metabarcoding Illumina-based studies. Specifically, Kircher *et al.* (2011) reported a significant fraction of sequences occurring with false index combinations derived from target-captured, pooled then PCR-amplified double-indexed shotgun libraries. They identified three major causes: (i) mixed clusters, (ii) sporadic cross-contamination of primers carrying different indices and (iii) chimeras which occurred solely in experiments where sequencing libraries from multiple samples were amplified in bulk. Kircher *et al.* (2011)'s set-up is transferable to metabarcoding studies where pools of tagged amplicons are bulk-amplified during library index PCR (Fig. 1), and therefore, these three causes might explain

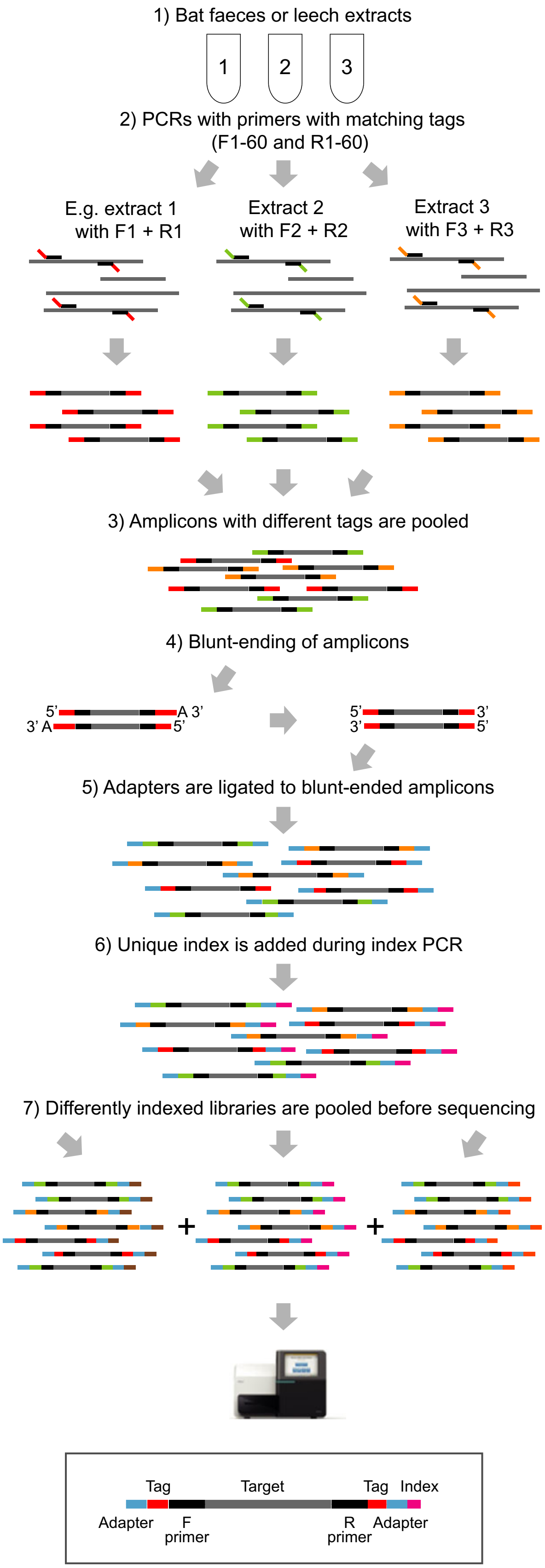
1) Bat faeces or leech extracts
1
2
3
2) PCRs with primers with matching tags (F1-60 and R1-60)
E.g. extract 1 with F1 + R1
Extract 2 with F2 + R2
Extract 3 with F3 + R3
3) Amplicons with different tags are pooled
4) Blunt-ending of amplicons
5'
A 3'
3' A
5'
5'
3'
3'
5'
5) Adapters are ligated to blunt-ended amplicons
6) Unique index is added during index PCR
7) Differently indexed libraries are pooled before sequencing
+
+
Tag
Target
Tag
Index
Adapter
F primer
R primer
Adapter

Fig. 1 Experimental overview. (1) Bat faecal and leech extracts are (2) amplified with COI insect generic and 16S mammal-generic primers, respectively. Both primer sets are 5′-tagged with 7-8 nucleotide sequences to obtain 60 uniquely tagged forward and 60 uniquely tagged reverse primers. Amplifications are carried out with primers with matching tags, that is F1-R1, F2-R2, etc. (3) Amplicons with different tags are pooled, and library-build steps are performed on each amplicon pool. (4) A Taq-based DNA polymerase creates 3′-A-overhangs on the tagged amplicons, which are removed using T4 DNA Polymerase during the blunt-end step in the library-build protocol. (5) Adapters are ligated. (6) Indexing every library with a different index during library index PCR enables (7) pooling of different amplicon libraries and parallel sequencing, here on an Illumina MiSeq platform. F = tag on forward primer, R = tag on reverse primer.

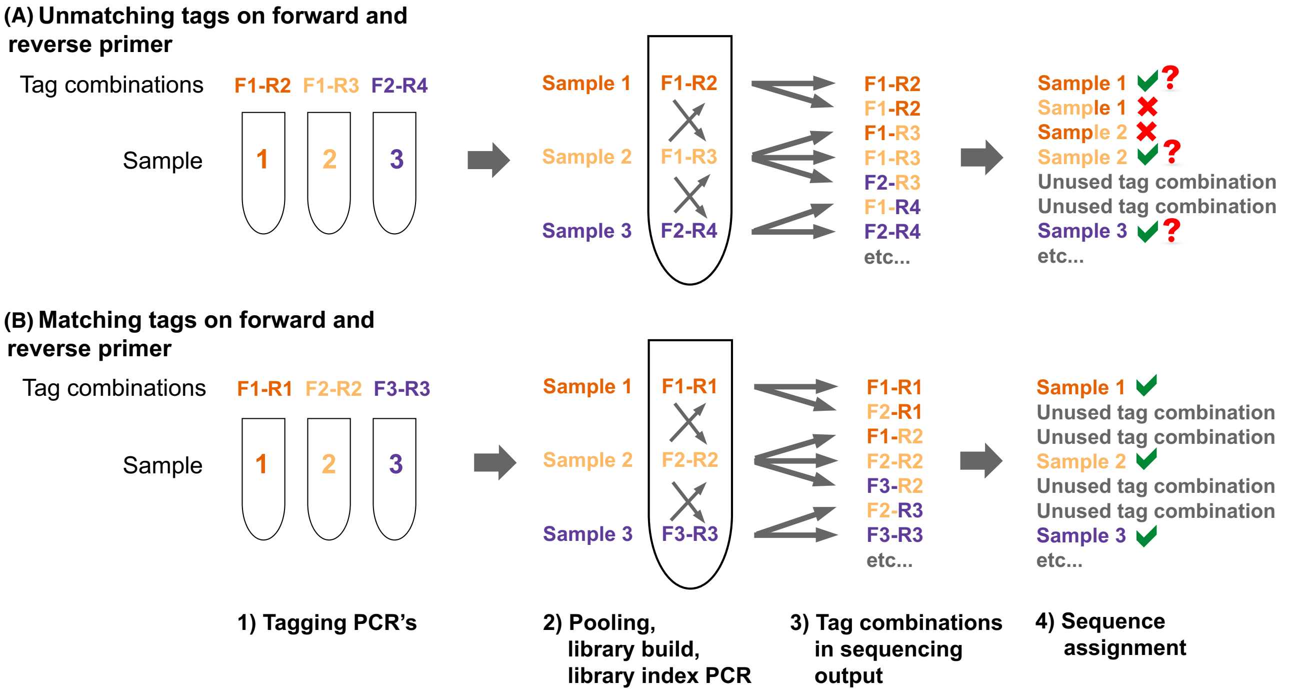

Fig. 2 Consequences of tag jumping when using (A) unmatching or (B) matching tags. Tag jumps occurring during library build or library index PCR (2) can cause sequences with false tag combinations to occur in the sequencing output (3 and 4). If amplifications are carried out with matching tags (B), however, false assignments of sequences to samples will be reduced (4). F = tag on forward primer, R = tag on reverse primer.

the occurrence of tag jumps between tagged amplicons (see Box 1).

Mixed clusters, the so-called bleeding, occur during sequencing and are PCR product colonies on the flowcell derived from more than one template molecule (Kircher *et al.* 2011), which could cause the occurrence of sequences with false combinations of used tags. Following Kircher *et al.* (2011), false combinations of tags in the sequencing output can also be caused by cross-contamination with tagged primers (Box 1). Their third proposed cause is chimera formation. Chimeric sequences are the product of two or more different molecules joining together. During PCR, the majority of chimeric amplicons are thought to be created as a result of incomplete primer extension during the elongation phase of the PCR cycle (Fig. 3) (Meyerhans *et al.* 1990; Wang & Wang 1997; Judo *et al.* 1998; Shin *et al.* 2014). Chimera formation is most likely to happen when closely related sequences are amplified in the same reaction (e.g. Judo *et al.* 1998; Smyth *et al.* 2010), as is the case when tagged amplicons from different samples are pooled and amplified during library index PCR (Fig. 3).

An alternate explanation for the appearance of sequences with false tag combinations in the sequencing output, as at least demonstrated on the Roche/454 sequencing platform, is offered by van Orsouw *et al.* (2007). Specifically, they documented that switched tags are a consequence of T4 DNA polymerase activity in the blunt-ending step of the 454 library build protocol (Margulies *et al.* 2005; van Orsouw *et al.* 2007) (Fig. 4). As blunt-ending using T4 DNA polymerase is part of many library build protocols, in metabarcoding studies with the purpose of blunt-ending amplicons with 3′-A-overhangs created by the PCR polymerase (e.g. Kozarewa *et al.* 2009; Meyer & Kircher 2010), this could also partly explain the occurrence of tag jumps in metabarcoding studies on Illumina sequencing platforms (Figs 1 and 4).

Even though metabarcoding has proven its worth in ecological studies many times over, we are, however, still perfecting this approach. This is emphasized by the multitude of articles recently appearing contributing to the ‘best practice’ of metabarcoding on second-generation sequencing platforms (see e.g. Coissac 2012; Coissac *et al.* 2012; Pompanon *et al.* 2012; Taberlet *et al.* 2012a,b; De Barba *et al.* 2014 for overviews of challenges, limitations, and guidelines). However, only little emphasis has been placed on tag jumping during an Illumina metabarcoding workflow and the resulting problem of misidentifying sequences to samples (Esling *et al.* 2015). To address this, we use data from two different metabarcoding studies, including different primer sets and tags, to

Box 1 Overview of where tag contamination and jumping can arise during a typical metabarcoding workflow, and how to minimize false assignment of sequences to samples, based, among others, on Wang & Wang (1996); Qiu *et al.* (2001); Acinas *et al.* (2005); van Orsouw *et al.* (2007); Lahr & Katz (2009); Meyer & Kircher (2010); Kircher *et al.* (2011); Coissac (2012); Shapiro & Hofreiter (2012); De Barba *et al.* (2014); Esling *et al.* (2015) as well as personal experiences and observations

Overview of where tag contamination and jumping can arise

Primer synthesis and handling

- Errors and contamination between tagged primers and index primers

During tagging PCR

- Cross-contamination between samples and primers during set-up

After tagging PCR

- Cross-contamination by tagged amplicons prior to library build

During pooling of tagged amplicons and library build

- Contamination of tagged amplicons
- T4 DNA polymerase-induced tag jumping during blunt-end step

During library index PCR

- Cross-contamination between libraries and index primers during set-up of library index PCRs
- Chimera formation (tag jumping) between tagged amplicons

During sequencing

- Mixed clusters ('bleeding')

Recommendations to minimize false assignments of sequences to samples

- Design tags with a high number of minimum differences
- Order tagged primers and index primers in smaller separate batches to minimize cross-contamination
- Work in separate pre- and post-PCR laboratories
- Use matching tags at both extremities in the tagging PCR
- Do PCR replicates of the tagging PCRs with different tag combinations
- Increase elongation time, make index PCR replicates with reduced template concentration and reduce number of cycles in the library index PCR to minimize chimera formation
- Be careful and minimize handling of primers, amplicons and libraries
- Sequence a subset of extraction blanks, PCR blanks from the tagged PCRs, library blanks and library index PCR blanks to allow estimations of tag contamination and tag jump levels
- Only trust sequences that occur across several PCR replicates

(i) document the occurrence and extent of tag jumping in metabarcoding studies on an Illumina sequencing platform, (ii) suggest possible causes of tag jumping, (iii) investigate the magnitude of cross-contamination by tagged primers and tagged amplicons and (iv) demonstrate a laboratory set-up to minimize sequence-to-sample misidentifications.

Materials and methods

We used data from two metabarcoding studies to investigate tag jumping: metabarcoding of vertebrate DNA in leech stomach contents and of insect DNA in bat faecal extracts. PCR amplification of metabarcoding markers in leech stomach contents and bat faecal extracts were carried out following modifications of methods described in Bohmann *et al.* (2011) and Hope *et al.* (2014). An overview can be found in Fig. 1. Primers were 5′ nucleotide tagged (Binladen *et al.* 2007) to yield a set of 60 unique forward and 60 unique reverse primers for each of the two primer sets. Tags were 7–8 nucleotides in length, containing a minimum of two (leech diet study) and three (bat diet study) nucleotide mismatches between tags. Tag sequences were different between the two studies. Each sample was amplified with matching tags drawn from the 60 alternate forward and reverse primers (e.g. F1-R1, F2-R2, etc.) to ensure tag jumps could be identified (Fig. 1).

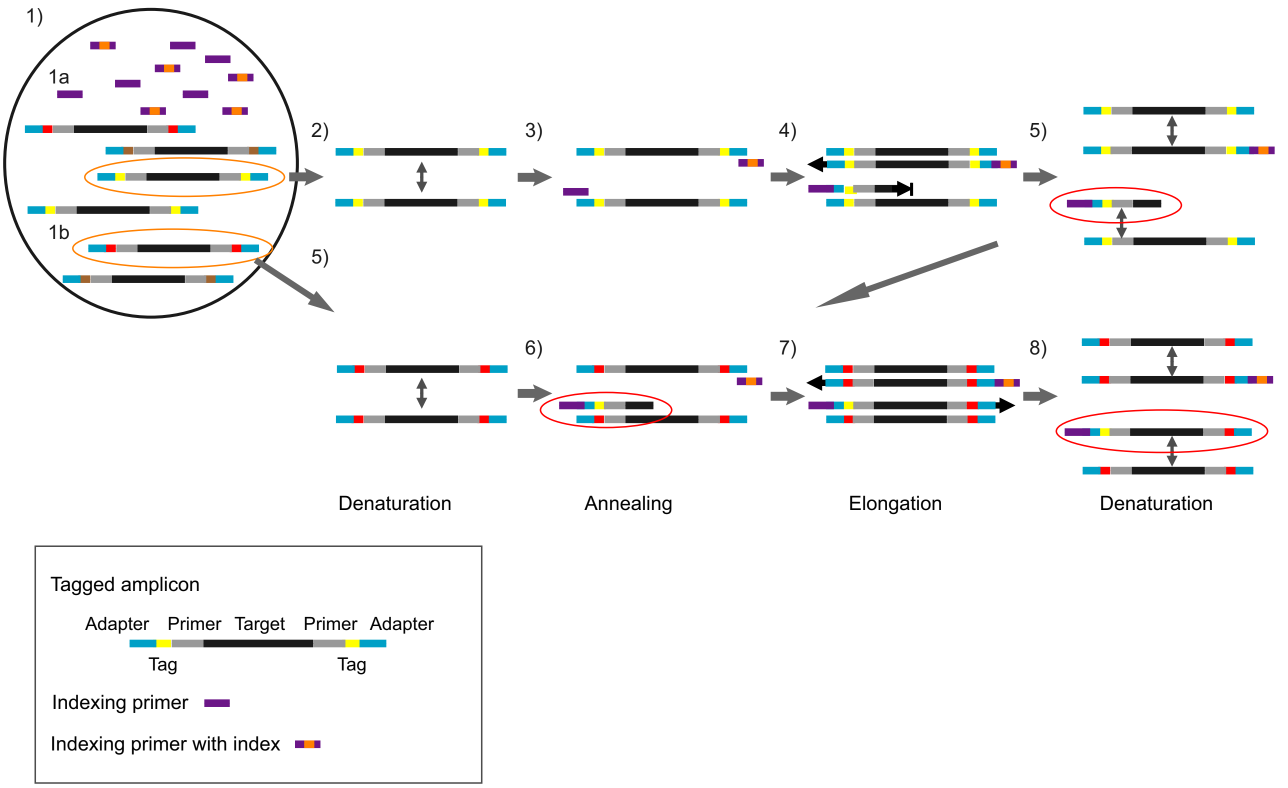

Fig. 3 Chimeric sequences with false tag combinations (tag jumps) can be created in metabarcoding studies where a library index PCR is carried out on a pool of tagged amplicons. During library index PCR, indexing primers (1a) are added to tagged amplicons (1b). After denaturing (2) and annealing of primers (3), incomplete primer extension (4) will enable the partially extended strand to act as an extended primer and compete with 'real' primers during the next cycle's annealing step. Here, it can bind to a template derived from a similar sequence but with a different tag (6). During the subsequent extension (7) and denaturing step (8), a template for a chimeric sequence consisting of two differently tagged amplicons is formed.

For the leech diet study, a ca. 95-bp (excluding primers) 16S fragment was amplified using mammal-generic primers (Taylor 1996), while in the bat diet study, insect-generic primers were used to amplify a 157-bp (excluding primers) COI mini-barcode fragment (Zeale *et al*. 2010). In the bat diet study, qPCR screening was carried out on all extraction blanks and on dilution series of a subset of the sample extracts. This enabled (i) contamination screening within extraction blanks, (ii) optimal cycle number determination for the following library amplification PCRs and (iii) assurance that PCR inhibitory substances, copurified with the DNA, would not distort the results of the following amplification of prey DNA (Bustin *et al*. 2009; Shapiro & Hofreiter 2012) (Appendix S1, Supporting information). To support our recommendations for tag design, we investigated whether tags delayed amplifications. Specifically, DNA from nine bulk insect extracts was qPCR-amplified using both tagged and untagged COI primers (Zeale *et al*. 2010). The primers were 5′ tagged with 8 nucleotide sequences on each extremity (Appendix S2, Supporting information).

All PCRs were set up in a dedicated pre-PCR laboratory to minimize the risk of contamination. A negative PCR control was included for every three to five (leech diet study) and eight (bat diet study) samples. Five μl of each PCR was visualized on a 2% agarose gel. All negative controls appeared negative. Despite no negative controls showing identifiable amplicons, we included 15 PCR replicates of extraction and PCR blanks from the bat diet study in the Illumina library construction and sequencing.

Only PCR products with different tags were pooled before library build to enable sequencing of many PCR replicates in parallel, while being able to track the tagged PCR products back to the correct PCR replicate (Fig. 1). A total of 166 PCR products were included from the leech diet study (representing 101 leech digest pools), and 215 PCR products (representing 163 samples and 15 extraction/PCR blanks) from the bat diet study. PCR products were pooled at approximately equimolar ratios as determined by gel band strength, with the exception of the PCR-amplified extraction and PCR blanks, from which 5 μL was added. In total, six library pools were made from the leech diet study, each consisting of 19–35

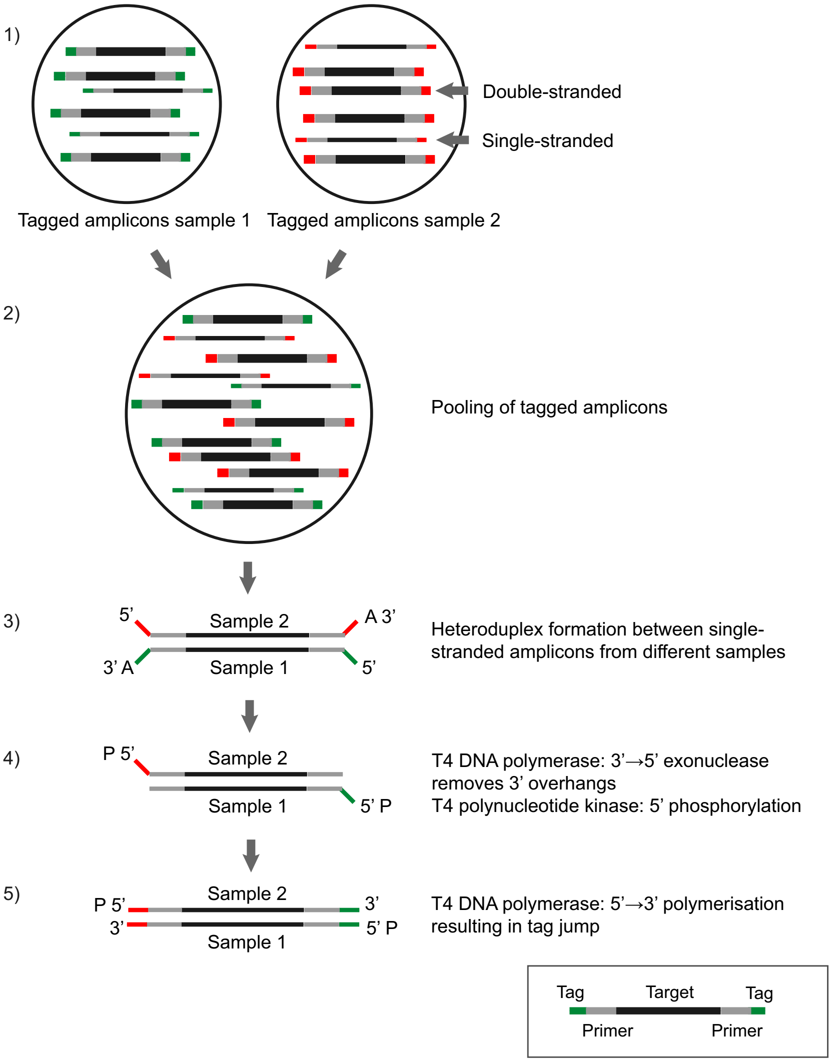
1)
Double-stranded
Single-stranded
Tagged amplicons sample 1
Tagged amplicons sample 2
2)
Pooling of tagged amplicons
3)
5'
Sample 2
A 3'
3' A
Sample 1
5'
Heteroduplex formation between single-stranded amplicons from different samples
4)
P 5'
Sample 2
Sample 1
5' P
T4 DNA polymerase: 3'→5' exonuclease removes 3' overhangs
T4 polynucleotide kinase: 5' phosphorylation
5)
P 5'
Sample 2
3'
3'
5' P
Sample 1
T4 DNA polymerase: 5'→3' polymerisation resulting in tag jump
Tag
Target
Tag
Primer
Primer

Fig. 4 Tag jumps can occur when a library is built on a pool of tagged amplicons and a T4 DNA polymerase blunt-ending step is included. In the example, a Taq-based DNA polymerase has created A-overhangs on amplicons. After the tagging PCR, individual reactions might contain single-stranded amplicons (1). When tagged amplicons are pooled before library build (2), heteroduplexes might form between semi- or fully complementary single strands of amplicons carrying different tags (3). During the blunt-end step, T4 DNA polymerase removes the tags at the 3′-ends because they are single-stranded overhangs (4). Finally, T4 DNA polymerase extends the 3′-ends using the opposite strand as template resulting in tag jumping (5). Modified from van Orsouw *et al*. 2007.

positive PCR products. Five library pools were made from the bat diet study, each consisting of 35–45 positive PCR products and 1–6 blanks.

PCR pools were converted into Illumina sequencing libraries, using the NEBNext DNA Library Prep Master Mix Set for 454 (#E6070) (NEB, Ipswich, MA, USA) although using blunt-end Illumina adapters (Meyer & Kircher 2010) in place of Roche/454 FLX adaptors. Libraries were subjected to library index PCR in 50 μL reactions using AmpliTaq Gold (Applied Biosystems, Foster City, CA) (leech diet study) or Platinum Taq DNA polymerase High Fidelity (Life Technologies, Carlsbad, CA, USA) (bat diet study). Libraries were indexed with different reverse indices with at least three nucleotide differences. The library index PCRs were in 50 μL reactions in which each library was amplified (10 and 12 cycles for the leech and bat diet study, respectively) in three separate reactions each containing 5 μl library. After the PCR, the three 50 μl replicates of each library were pooled. Each leech pool was sequenced 150 bp paired-end on ca. 7% of a MiSeq flowcell, and each bat diet pool was sequenced 250 bp paired-end on ca. 10% of a MiSeq flowcell. Full details of extractions, PCR amplifications and sequencing preparation are given in Appendix S1 (Supporting information).

After sequencing of the 11 libraries from both studies, adapters and consecutive stretches of N's and low-quality bases from both the 5′- and 3′-ends were trimmed using default settings in ADAPTERREMOVAL (version 1.1-fixed) (Lindgreen 2012), although with a minimal read length of 25 bp. Using a customized Perl script, paired reads were merged if the overlap was 100% identical. For both the bat and the leech diet study, there was full overlap of paired reads. Trimmed and merged sequences were sorted based on tag combinations using GENEIOUS (version 6.1.6, created by Biomatters). Only sequences with perfect match to both forward and reverse primer sequences were kept for further analysis, sorted in Geneious based on tag combinations and assigned to one of five categories based on their tag combination (Box 2, Table 1). As all extracts used in both studies were amplified using primers with matching tags on each extremity, sequences with tag combinations assigned to the category 'Used, Matching' were expected to occur, while all other combinations of tags should not exist in the libraries. Two heat maps (one from each study) were created to visualize the occurrence of sequences with different tag combinations (Fig. 5) (Appendix S3, Supporting information). The proportion of sequences with tag combinations 'Used, Nonmatching', 'One used/One unused', 'Unused, Matching' or 'Unused, Nonmatching' (Box 2) was calculated to investigate the cause of their presence, and the observed distributions of sequences assigned to these different combinations of tags were compared to the expected distributions using paired *t*-tests (Table S4 in Appendix S3, Supporting information).

As presented in the Introduction, we hypothesized the presence of sequences with incorrect tag combinations and unused tags to be caused by (i) contamination, (ii) chimera formation, (iii) T4 DNA polymerase activity and/or (iv) mixed clusters. The presence of unused tags could be caused by cross-contamination either by (i) tagged primers creating a random pattern of tag combinations and/or (ii) tagged amplicons resulting in sequences with unused matching tag combinations

Box 2 Definitions of used and unused tags, tags with errors, and an overview of the five categories for combinations of tags on the extremities of sequences

Used tags
Tags used in the specific library

Unused tags
Tags that belong to the 60 designed tags for each study, but were not used in the specific library

Tags with errors
Tags that do not have a perfect match to any of the 60 tags used in each study

Used, Matching
Identical tags in both ends of the sequence. Tags have been used in the library

Used, Nonmatching
Different tags in the ends of the sequence. Tags have been used in the library

One used/One unused
Different tags in both ends of the sequence. One tag has been used in the library, the other has not

Unused, Matching
Identical tags in both ends of the sequence. Tags have not been used in the library

Unused, Nonmatching
Different tags in the ends of the sequence. Tags have not been used in the library

Table 1 Mean ± SD of sequences assigned to five different categories of tag combinations (see overview in Box 2) in libraries in the leech and the bat diet study, and the percentage of sequences belonging to each category of the total amount of sequences that had perfect match to existing tags and primers. See numbers for individual libraries in Appendix S4 (Supporting information), Table S1

		Used, Matching		Used, Nonmatching		One used/One unused		Unused, Matching		Unused, Nonmatching	
Library	No. tags used in pool	No. seqs.	%	No. seqs.	%	No. seqs.	%	No. seqs.	%	No. seqs.	%
Leech diet	28 ± 6.1	519 862 ± 129 258	97.1 ± 0.9	13 988 ± 6540	2.6 ± 1.0	213 ± 272	0.04 ± 0.06	926 ± 853	0.20 ± 0.23	216 ± 181	0.04 ± 0.03
Bat diet	43 ± 3.2	1 115 197 ± 204 180	97.8 ± 0.1	24 032 ± 4451	2.1 ± 0.2	570 ± 310	0.05 ± 0.02	264 ± 111	0.02 ± 0.01	2 ± 2	0.00 ± 0.00

(Box 1). Chimera formation was investigated by comparing distribution of sequences belonging to the five categories based on their tag combination before and after chimera removal. If sequences with 'Used, Nonmatching' tag combinations were caused by chimera formation during library index PCR (Fig. 3), then the chimera removal programs should eliminate sequences from this category. To identify chimeras, all merged sequences within each library were collapsed into unique sequences prior to chimera removal using USEARCH, prefix dereplication (Edgar 2010). Chimera removal was performed using UCHIME denovo (Edgar *et al*. 2011) on sequences within each library. Sequences not identified as chimeras with a perfect match to both forward and reverse primer were sorted in Geneious based on tag combinations and assigned to one of the five categories based on their tag combination (Box 2, Table 1).

The occurrence of sequences with errors in tags was calculated to investigate reliability of tagging amplicons in one end versus double-tagging. Details of calculations of expected distribution of tag combinations, statistics, heat maps and statistics on errors in tag sequences are explained in Appendix S3, Supporting information.

Results

Sequencing outputs and initial sorting

On average, ca. 800 000 and 1 700 000 paired raw reads were produced in each library in the leech and bat diet study, respectively. Of these, an average of 76.2% and 78.5% was 100% identical and merged. Of the merged reads, an average of 87.9% and 87.2% had a perfect match to both primers and existing tags at both extremities, for the leech and bat diet study, respectively (Table S1 in Appendix S3, Supporting information).

Sorting sequences based on tag combinations

As expected, the majority of sequences were found in the category 'Used/Matching' (Table 1). However, a large number of sequences were observed which carried false combinations of tags, mainly tags that had been used in the respective library, and few that had not. Also, some sequences carried combinations of unused, matching tags (Fig. 5).

Investigating the cause of used tags in false combinations

A relatively large number of sequences had false combinations of tags used in the specific library (Table 1, Fig. 5). As presented in Fig. 6, the proportion of observed sequences with these used, but nonmatching tags, was significantly higher in both studies compared to what would be expected if they had been caused by contamination. The observed proportion of sequences with one used and one unused tag were significantly lower than the expected in both studies (Appendix S3, Supporting information).

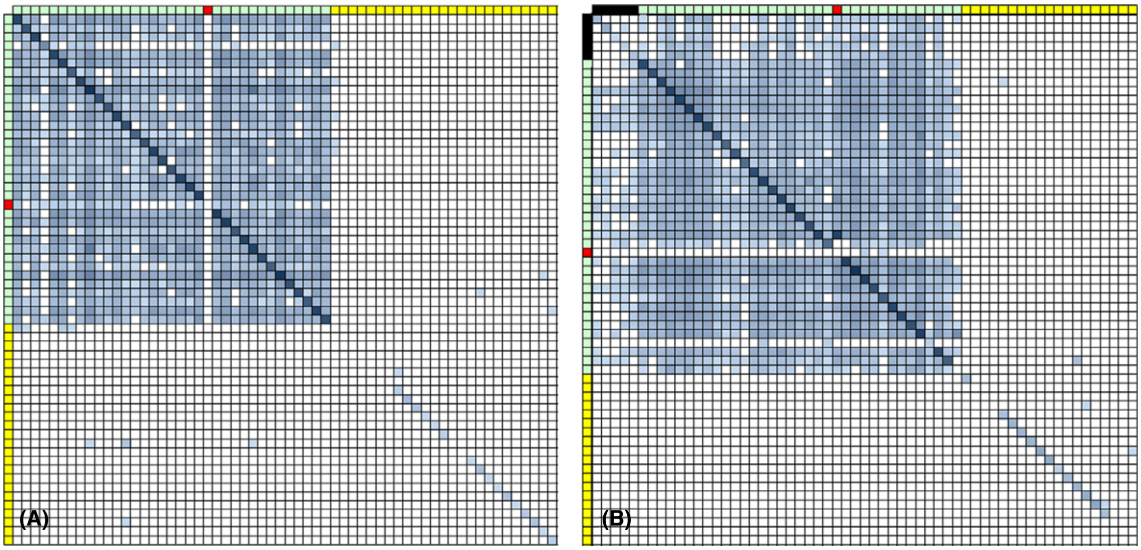

Fig. 5 Heat maps with (log-transformed) counts for all tag combinations in the A) leech diet study (library LD1) and B) bat diet study (library BD1). Forward tags on the horizontal and reverse tags on the vertical axis, with used tags (green), unused tags (yellow) and tags used to amplify negative controls or PCR blanks (black). Tags marked with red in A) belong to a used tagged primer that did not work, and B) indicate a wrong sequence on an ordered tagged primer (Appendix S3, Supporting information). Darker colours indicate higher sequence counts.

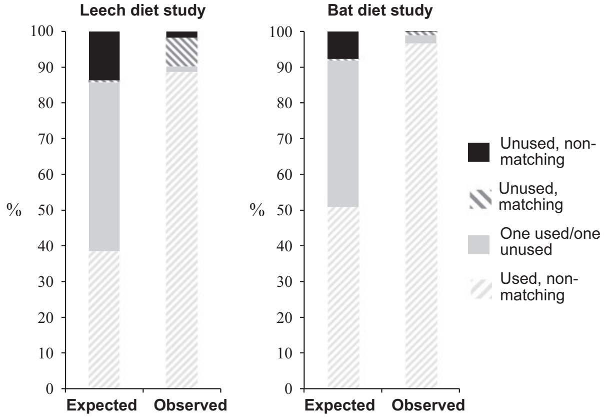

Fig. 6 Averages across libraries of expected and observed distribution of sequences in the four categories where tag combinations are incorrect (Table S4 in Appendix S3, Supporting information). In both studies, the observed frequency of sequences with used but nonmatching tag combinations was significantly higher than would be expected if they had been caused by random cross-contamination of tagged primers. Observed ratios in the three other categories were significantly lower except for 'Unused, Matching' in the leech diet study, which were not significantly different from the expected (Table S5 in Appendix S3, Supporting information).

Chimera formation and tag jumps

If sequences with 'Used, Nonmatching' tag combinations are caused by chimera formation during library index PCR (Fig. 3), then the chimera removal program should eliminate sequences from this category. Note, however, that the chimera program might not identify all chimeras (see Discussion). After chimera removal, many unique sequences with 'Used, Nonmatching' tag combinations were still present. Across libraries, chimera removal on average reduced the number of unique sequences in the category 'Used, Nonmatching' by 8.6% in the leech diet study and 14.2% in the bat diet study (Table 2). The number of unique sequences assigned to the expected tag combinations, 'Used, Matching', were also reduced by chimera removal, by 1% and 2.5% for the leech and bat diet study, respectively, which can be attributed to removal of chimeras which arose during the initial tagging PCR (Table 2). UCHIME removed a higher fraction of chimeras within sequences with 'Used, Nonmatching' tag combinations compared with sequences with 'Used, Matching' tag combinations (Paired *t*-test; leech diet study: $t = 13.7$, d.f. = 5, $P < 0.0001$, bat diet study: $t = 18.3$, d.f. = 4, $P < 0.0001$).

Sources of contamination

Sequencing output from all libraries included tags that were not intentionally used in the initial PCRs (Fig. 5 and 7). In both the leech and bat diet study, the observed number of sequences with 'Unused, Matching' tag combinations was significantly higher ($P < 0.0001$, Table S7 in Appendix S3, Supporting information) than expected if random contamination was the cause (Fig. 7). For sequences with 'Unused, Nonmatching' combinations of tags, the opposite pattern was observed. Here, the expected ratio was significantly higher ($P < 0.0001$, Appendix S3, Supporting information) than the observed (Fig. 7).

Errors in tags

Table 3 shows the mean (± SD) of sequences containing zero sequencing errors in tags, one error in the F-tag, one error in the R-tag and one error in both tags (including the average for each of the two different studies). Occurrence of sequences with one error in both tags (0.04% in the leech diet study and 0.01% in the bat diet study) was significantly lower than the frequency of sequences with an error in a tag on either extremity (Table S7 in Appendix S3, Supporting information).

Tags delay PCR amplification

The mean Ct values of reactions with tagged and untagged primers were significantly different with an increase

Table 2 Mean ± SD of unique sequences with tag combinations belonging to the category 'Used, Matching' and 'Used, Nonmatching' in leech and bat libraries, before and after removal of chimeric sequences with UCHIME. See numbers for individual libraries in Appendix S4 (Supporting information), Table S2

	Number unique sequences before chimera removal		Number unique sequences after chimera removal		
Study	Used, Matching (correct tag combinations)	Used, Nonmatching (assumed tag jumps)	Used, Matching (correct tag combinations)	Used, Nonmatching (assumed tag jumps)	% Used, Nonmatching (assumed tag jumps) removed by UCHIME
Leech diet	4479 ± 513	1314 ± 694	3775 ± 1712	1200 ± 632	8.6 ± 1.3
Bat diet	61 709 ± 9127	7923 ± 1410	60 094 ± 8445	5275 ± 2585	14.2 ± 1.8

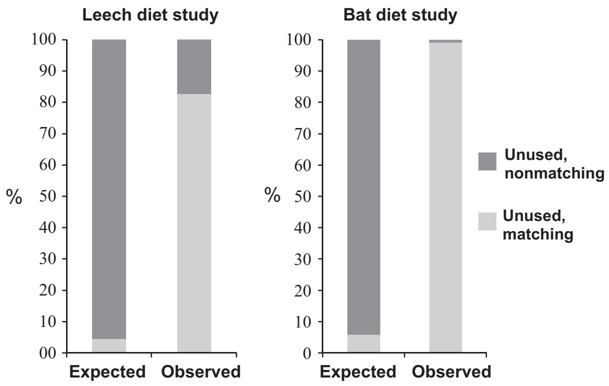

Fig. 7 Investigating the cause of contamination. The expected and observed distribution of sequences with unused tags where tags are either matching (indicating amplicon contamination) or nonmatching (indicating tagged primer contamination). The observed ratios in the two studies were significantly different from the expected ratios (Appendix S3, Supporting information) indicating that contamination by tagged amplicons is the main source of contamination.

of 4.42 cycles (95% CI, 3.74 to 5.10), t(8) = 15.02, $P < 0.0001$) when tagged primers were used (Appendix S2, Supporting information).

Discussion

Tag jumping in metabarcoding studies

Documenting tag jumping on an Illumina sequencing platform. Metabarcoding studies rarely report on the occurrence of sequences with false tag combinations in the sequencing output and how these, and the likelihood of false assignment of sequences to samples, are dealt with bioinformatically. In the current study, on average, 2.6% and 2.1% of sequences, which had a perfect match to both primers and existing tags from the leech and bat diet study, respectively, had false combinations of used tags (Table 1). In the following, we discuss the four previously suggested mechanisms, which might explain the occurrence of these sequences: (i) sporadic cross-contamination of primers carrying different tags occurring during primer synthesis or handling (Kircher *et al.* 2011), (ii) chimera formation during index PCR (Fig. 3) (Kircher *et al.* 2011), (iii) tag jumping caused by T4 DNA polymerase activity during the blunt-end step of the library build (Fig. 4) (van Orsouw *et al.* 2007) and (iv) mixed clusters on the Illumina flowcell (Kircher *et al.* 2011).

i) As shown in Fig. 6, the observed ratio of sequences with false tag combinations of used tags was significantly higher than would be expected if the main cause was cross-contamination of tagged primers. This pattern is also visualized in the heat maps (Fig. 5).

ii) High chimera rates have been associated with highly similar sequences (e.g. Wang & Wang 1997; Shin *et al.* 2014). Therefore, studies where a library index PCR is performed on a pool of differently tagged amplicons prior to sequencing will be prone to chimera formation as the sequences, after having undergone adapter ligation, consist of identical adapters and primers, and only differ in the tag sequences (Fig. 3). Furthermore, the amplified prey sequences might, depending on which marker is targeted and how closely related the sequenced taxa are, only differ with a couple of nucleotides.
The chimera removal program identified a significantly higher fraction of chimeras among unique sequences with false combinations of used tags than in unique sequences with used combinations of tags (used matching) (Table 2). This indicates that some sequences with false combinations of used tags are caused by chimera formation during library index PCR. Given the high similarity of sequences in metabarcoding studies, the most commonly formed chimeras are the most difficult to computationally detect (Dunshea *et al.* 2008; Smyth *et al.* 2010). Therefore, the numbers of identified chimeras might be underestimated.

iii) While the formation of chimeras between tagged amplicons during library index PCR creates recombinant sequences consisting of different templates, sequences with T4 DNA polymerase-mediated tag jumps will, in most instances, only have the tags switched and will therefore not be detected by chi-

Table 3 Mean ± SD of sequences with different levels of errors in tag sequences and their fraction of merged sequences. Only sequences with perfect match to both forward and reverse primers are included. See numbers for individual libraries in Appendix S3 (Supporting information) Table S3

	One error in F-tag		One error in R-tag		One error in both tags	
Library	Number sequences	%	Number sequences	%	Number sequences	%
Leech diet	13 210 ± 4600	2.2 ± 0.5	11 494 ± 3274	1.9 ± 0.2	261 ± 73	0.04 ± 0.01
Bat diet	13 761 ± 2977	1.0 ± 0.1	13 737 ± 2098	1.0 ± 0.1	141 ± 23	0.01 ± 0.00

mera removal programs (Fig. 4). Thus, sequences with false tag combinations of used tags, which were not identified as chimeras, might be caused by this mechanism.

iv) The observed tag jumps can be caused by the occurrence of mixed clusters on the sequencing flowcell, the so-called 'bleeding' (Kircher *et al.* 2011). In a similar set-up, but where double-indexed shotgun libraries were pooled before capture, amplification and sequencing, Kircher *et al.* 2011 documented that the majority of false assignments could be explained by this. In the current study, we rule out bleeding as a major cause of tag jumping because jumps mainly consisted of tags used within specific libraries and not tags used in the flowcell (Fig. 5).

Based on the above, we conclude that of these possible explanations, the majority of the observed tag jumps must be caused by T4 DNA polymerase activity during the blunt-end step of the library build (Fig. 4) and chimera formation during library index PCR (Fig. 3) (van Orsouw *et al.* 2007; Kircher *et al.* 2011).

Implications of tag jumping. Regardless of the origin, if measures to minimize and identify tag jumps are not taken, they can cause incorrect assignment of sequences to samples (Esling *et al.* 2015). Thereby metabarcoding studies face the risk of artificially inflating diversity in the samples and/or wrongly assign taxa to samples. On the Roche/454 sequencing platform, Carew *et al.* (2013) found that all species that were incorrectly identified as being present at a study site were already in the experiment. Also on the Roche/454 sequencing platform, Lindner *et al.* (2013) identified contaminant sequences in some data sets, which represented species from other data sets. Apart from this, mixed-template chimeras might be mistaken as (rare) OTUs, which can also artificially inflate diversity in the samples.

Contamination

The heat maps in Fig. 5 both show a conspicuous 'tail' of sequences carrying unused but matching tags. This not only indicates occurrence of contamination by tagged amplicons across libraries, but also that some of the sequences, which would bioinformatically be assigned to samples because they carry used tags in used combinations, could, in fact, be contamination by tagged amplicons (see Appendix S3, Supporting information for a discussion on levels of cross-contamination between libraries). The sequenced negative controls in the bat diet study show that despite running both qPCRs and gels on negative controls, some sequences carrying the tag combinations used for these negative controls still occur in the sequencing output (Fig. 5B). These occur at similar frequencies as sequences with matching tags, which were not used in pools, further indicating that amplicon contamination can cause sequences to be falsely assigned.

We found that the main source of contamination in the two metabarcoding data sets was tagged amplicons as opposed to tagged primers (Fig. 7). Cross-contamination by tagged amplicons can happen during handling of amplicons before and during library build and library index PCR (Box 1).

We attribute unmatching combinations of unused tags to contamination by tagged primers or to tag jumping between contaminating tagged amplicons (Table 1 & Fig. 7). Cross-contamination by tagged primers could happen during primer synthesis and, despite carefulness, during handling in the laboratory (Kircher *et al.* 2011) (Box 1).

Errors in tags

Several causes including errors during primer synthesis, PCR errors and sequencing errors can result in incorrect tag sequences. However, the main cause cannot be determined based on these data sets. We expect tag sequence reliability to be optimized in the current study as (i) tags were 7–8 nucleotides long and designed with a minimum of two and three nucleotide distances between tags for the leech and bat diet study, respectively, (ii) matching tags were used for amplifications, (iii) sequencing length was chosen to allow full overlap of paired reads and (iv) bioinformatically, only 100% identical paired reads were merged.

We advocate using tags with as high as possible number of mismatches between tags. Furthermore, in agreement with several other authors, as occurrence of sequences with one error in both tags in this study was low (Table 3), we advocate tagging amplicons at both extremities (e.g. Carlsen *et al.* 2012; Coissac 2012) and extend this to the use of matching tags on both extremities, as identical errors then need to occur on each extremity in order for a sequence to be falsely assigned.

Suggestions for future metabarcoding studies

Experimental set-up. As only some of the observed tag jumps were removed by the chimera removal program (Table 2), we conclude that one cannot solely rely on bioinformatic processing post sequencing to remove errors, but that it is important to implement experimental set-up measures to minimize tag jumps. In Box 1, we give an overview of where tag contamination and jumps can arise during a typical metabarcoding workflow for sequencing on an Illumina sequencing platform, and which measures can be taken to minimize false assignment of sequences to samples. To sum up, (i) using matching tags, (ii) doing PCR replicates, (iii) minimize and only carefully handle tagged amplicons, and (iv) incorporate negative controls at all steps and sequence a subset, even if negative, are the main focuses. Regarding (i), we note that tags do not necessarily have to be identical at each extremity; what is important is that each tag is only used once in each library.

Matching tags on both extremities of sequences enables easy detection of the majority of possible tag jumps. However, using matching tags does not safeguard for the fact that occasionally chimeras can be composed of more than two true sequences (Quince *et al.* 2011), and although we believe such occurrences are rare, they could give rise to new sequences with matching tags already used in the library. Another potential source of 'false' sequences with matching tag pairs is tagged amplicon contamination. As a large fraction of existing tags are used in each of the pooled libraries, contamination with tagged amplicons from other libraries cannot always be detected. However, if PCR replicates are made of all samples using different, but matching tags, it is possible to bioinformatically discard sequences not present in several replicates. Thereby, both chimeras formed by two or more sequences and amplicon contamination should be minimized.

Designing tags. Given the observed decrease in mean Ct values for qPCR-amplified samples when using tagged primers instead of untagged primers (4.42 cycles, Appendix S2, Supporting information), we expect a trade-off between tag length and efficiency of amplification. Therefore, we argue that use of tags that are longer than necessary are not advisable, as long as sequences are tagged at both extremities. For thorough information about tag design, see Coissac (2012).

Illumina sequencing platforms employ a sequencing-by-synthesis approach, in which DNA fragments are attached to a flowcell, then amplified multiple times to create a cluster of identical fragments. At each sequencing cycle, a fluorescent-labelled base is incorporated into each fragment in the cluster, and images of the flowcell surface are captured (http://technology.illumina.com/technology/next-generation-sequencing/sequencing-technology.html). Coordinates for each cluster are determined in the first four sequencing cycles, but in low-complexity libraries, such as amplicon-based libraries, similar bases in the cluster calling cycles interfere with mapping of cluster coordinates, and generation of sequencing data in the subsequent cycles (http://blog.genohub.com/nextseq-hiseq-or-miseq-for-low-diversity-sequencing/; Krueger *et al.* 2011). When sequencing tagged amplicons, the first four bases in the tags are the first to be sequenced. Therefore, the base composition of tags used in a library should be balanced in all positions to create as much complexity as possible. Furthermore, as the quality of the first four sequenced bases are typically of lower quality than the following bases sequenced, tagging at both extremities and sequencing paired-end with full overlap and only merging paired reads that are 100% identical will decrease false assignments. Additionally, use of tags with different lengths, as in the current study, creates a shift between reads, which increases complexity at each base position in the entire sequencing length, thus decreasing the need to spike libraries with other DNA sources, such as PhiX, before sequencing (http://support.illumina.com/sequencing/sequencing_instruments/miseq/questions.html; Krueger *et al.* 2011).

Exploring other solutions to minimize or avoid that tag jumps cause false assignments

Accounting for tag jumping in the laboratory set-up increases both cost and workload. Therefore, solutions to avoid tag jumps should be investigated.

Omitting the blunt-ending step. An obvious solution to avoid tag jumps caused by T4 DNA polymerase during library build (Fig. 4) is simply to omit this step. Many metabarcoding studies use Taq-based DNA polymerases, such as AmpliTaq Gold, in the tagging PCR, which creates amplicons with 3′-A-overhangs (Rittié & Perbal 2008). If following a protocol such as Meyer & Kircher (2010), these have to be removed during a blunt-ending step in order to ligate blunt-end adapters (Fig. 1). However, when sequencing tagged amplicons on a Roche/454 sequencing platform, van Orsouw *et al.* (2007) demonstrated that using tagged primers with a 5′-phosphate, omitting the blunt-ending step and modifying the GS 20 adapters by adding a 5′-T nucleotide to allow T/A ligation with A-overhangs, they were able to reduce the occurrence of tag jumps from 0.1 to 16% to <0.00025% of reads per run.

Another solution to omit the blunt-ending step while using blunt-end Illumina adapters could be to use tagged primers with 5′-phosphate and a PCR polymerase, which creates blunt-ended amplicons in the tagging PCR, for example Pfx DNA polymerase (www.lifetechnologies.com/), Phusion DNA polymerase, Q5 DNA polymerase (www.neb.com) and Pfu DNA polymerase (www.thermoscientificbio.com). To our knowledge, the effect of this has not been investigated.

Avoid or reduce chimera formation. Reducing the number of cycles, increasing the extension time during library index PCR, decreasing template concentration and using proofreading enzymes have been found to reduce chimera formation (e.g. Wang & Wang 1996; Qiu *et al.* 2001; Acinas *et al.* 2005; Lahr & Katz 2009). To estimate the minimum number of cycles required in the library index PCR, a qPCR can be carried out on the libraries before library index PCR (Meyer & Kircher 2010; Shapiro & Hofreiter 2012). In addition, library index PCRs can be run in several reactions with lower template concentrations (see Materials and Methods) (Box 1). Chimeras are, of course, also formed during the initial tagging PCR, but as these chimeras are not the focus of this study, we just note that recommendations to reduce chimeras in the library index PCR apply to the tagging PCR as well and we recommend qPCR to estimate the minimum number of cycles (as well as check for inhibition) across samples (Shapiro & Hofreiter 2012).

One obvious solution to reduce chimera formation is to avoid bulk amplification of pools of tagged amplicons. Illumina offers PCR-free library preparation kits, which will eliminate tag jumps caused by chimera creation between tagged amplicons (Fig. 3), but not tag jumps created by T4 DNA polymerase activity (Fig. 4). Likewise, Vo & Jedlicka (2014) propose a library build protocol without a postligation PCR enrichment step to avoid chimera formation between tagged amplicons, but this also includes a T4 DNA polymerase blunt-ending step. Another option is amplifying DNA extracts with primers not only containing tags but also sequencing adapters and indices (fusion primers). However, long primers might be associated with decrease in PCR efficiency (Appendix S2, Supporting information). Another alternative to avoid chimera-based tag jumping during library index PCR could be emulsion PCR (Nakano *et al.* 2003; Zinger *et al.* 2012). During emulsion PCR, each template is amplified separately within a microdroplet, which prevents incompletely elongated primers in annealing to foreign templates. To our knowledge, the effect of the above suggestions in minimizing tag jumping still needs investigation.

Conclusions

The phenomenon of tag jumps in metabarcoding studies have only been scarcely reported. Although sequences with unused tags are easily identified and excluded, tag jumps that create sequences with false, but already used tag combinations can cause incorrect assignment of sequences to samples and artificially inflate diversity. Therefore, to document and investigate tag jumping in metabarcoding studies on Illumina sequencing platforms, we sequenced tagged amplicons from two different metabarcoding studies. In the two studies, we found that an average of 2.6% and 2.1% of sequences with a perfect match to primers and existing tags had false tag combinations of used tags (tag jumps). A chimera removal program (UCHIME) was able to remove an average of 8.6% and 14.2% of these unique sequences in the two studies. Whether the remaining sequences are chimeras, but not recognized as such, or if they are created by, for example, T4 DNA polymerase activity when pools of tagged amplicons are blunt-ended during library build remains unclear. However, the high occurrence of these sequences, even after chimera removal, emphasizes that bioinformatics is not sufficient to avoid false assignment of sequences to samples. Furthermore, we found contamination by tagged amplicons across libraries to be a bigger issue than cross-contamination of tagged primers, both of which could lead to false assignment of sequences to samples. Based on our findings, we suspect that tag jumping and tagged amplicon contamination is a considerable problem in metabarcoding studies on Illumina sequencing platforms, and advise future studies to take measures to minimize and identify these, so as to avoid false assignment of sequences to samples, thus increasing trustworthiness of the drawn conclusions.

Acknowledgements

We thank Stine Richter, Sebastian Emil Schultze, Martin Nielsen and Amanda Pittmann for laboratory assistance in generating leech and bat diet data for use in other studies, and David Hill for letting us use the generated bat diet data for this additional study. We thank Morten Rasmussen for development of Perl script for merging and Marie Lisandra Zepeda Mendoza for modifications, and Christian Carøe for fruitful discussions and comments on the original manuscript. We thank the Danish National High-Throughput DNA Sequencing Centre for their invaluable advice and technical assistance with regard to generation of the MiSeq data, especially Kim Magnussen for explanations and discussions about the Illumina platform sequencing technology. Finally, we acknowledge the Natural History Museum of Denmark and Danish National Research Centre grant DNRF94 for funding this research.

References

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology*, **71**, 8966–8969.

Binladen J, Gilbert MTP, Bollback JP *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, **2**, e197.

Blaalid R, Kumar S, Nilsson RH *et al.* (2013) ITS1 versus ITS2 as DNA metabarcodes for fungi. *Molecular Ecology Resources*, **13**, 218–224.

Blaalid R, Davey ML, Kauserud H *et al.* (2014) Arctic root-associated fungal community composition reflects environmental filtering. *Molecular Ecology*, **23**, 649–659.

Bohmann K, Monadjem A, Lehmkuhl Noer C *et al.* (2011) Molecular diet analysis of two African free-tailed bats (Molossidae) using high throughput sequencing. *PLoS ONE*, **6**, e21441.

Botnen S, Vik U, Carlsen T *et al.* (2014) Low host specificity of root-associated fungi at an Arctic site. *Molecular Ecology*, **23**, 975–985.

Bustin SA, Benes V, Garson JA *et al.* (2009) The MIQE Guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*, **55**, 611–622.

Carew ME, Pettigrove VJ, Metzeling L, Hoffmann AA (2013) Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology*, **10**, 45.

Carlsen T, Aas AB, Lindner D *et al.* (2012) Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecology*, **5**, 747–749.

Chariton AA, Court LN, Hartley DM, Colloff MJ, Hardy CM (2010) Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment*, **8**, 233–238.

Coissac E (2012) OLIGOTAG: a program for designing sets of tags for next-generation sequencing of multiplexed samples. *Methods in Molecular Biology (Clifton, N.J.)*, **888**, 13–31.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, **21**, 1834–1847.

Davey ML, Heimdal R, Ohlson M, Kauserud H (2013) Host-and tissue-specificity of moss-associated Galerina and Mycena determined from amplicon pyrosequencing data. *Fungal Ecology*, **6**, 179–186.

Davey ML, Kauserud H, Ohlson M (2014) Forestry impacts on the hidden fungal biodiversity associated with bryophytes. *FEMS Microbiology Ecology*, **90**, 313–325.

De Barba M, Miquel C, Boyer F *et al.* (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources*, **14**, 306–323.

Deagle BE, Kirkwood R, Jarman SN (2009) Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, **18**, 2022–2038.

Dunshea G, Barros NB, Wells RS *et al.* (2008) Pseudogenes and DNA-based diet analyses: a cautionary tale from a relatively well sampled predator-prey system. *Bulletin of Entomological Research*, **98**, 239–248.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Edgar RCR, Haas BJB, Clemente JCJ, Quince CC, Knight RR (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Esling P, Lejzerowicz F, Pawlowski J (2015) Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, doi:10.1093/nar/gkv107.

Ficetola GF, Miaud C, Pompanon F, Taberlet P (2008) Species detection using environmental DNA from water samples. *Biology Letters*, **4**, 423–425.

Haile J, Froese DG, MacPhee RDE *et al.* (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**, 22352–22357.

Hope PR, Bohmann K, Gilbert MTP *et al.* (2014) Second generation sequencing and morphological faecal analysis reveal unexpected foraging behaviour by *Myotis nattereri* (Chiroptera, Vespertilionidae) in winter. *Frontiers in Zoology*, **11**, 39.

Jerde CL, Mahon AR, Chadderton WL, Lodge DM (2011) "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conservation Letters*, **4**, 150–157.

Judo MS, Wedel AB, Wilson C (1998) Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Research*, **26**, 1819–1825.

Kircher M, Sawyer S, Meyer M (2011) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, **40**, e3.

Kozarewa I, Ning Z, Quail MA *et al.* (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, **6**, 291–295.

Krueger F, Andrews SR, Osborne CS (2011) Large scale loss of data in low-diversity illumina sequencing libraries can be recovered by deferred cluster calling. *PLoS ONE*, **6**, e16607.

Lahr DJG, Katz LA (2009) Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques*, **47**, 857–866.

Lindgreen S (2012) ADAPTERREMOVAL: easy cleaning of next-generation sequencing reads. *BMC Research Notes*, **5**, 337.

Lindner DL, Carlsen T, Henrik Nilsson R *et al.* (2013) Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. *Ecology and Evolution*, **3**, 1751–1764.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, **2010**, pdb.prot5448.

Meyerhans A, Vartanian J-P, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids Research*, **18**, 1687–1691.

Murray DC, Haile J, Dortch J *et al.* (2013) Scrapheap Challenge: a novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Scientific Reports*, **3**, doi:10.1038/srep03371.

Nakano M, Komatsu J, Matsuura S *et al.* (2003) Single-molecule PCR using water-in-oil emulsion. *Journal of Biotechnology*, **102**, 117–124.

Nathan LR, Jerde CL, Budny ML, Mahon AR (2014) The use of environmental DNA in invasive species surveillance of the Great Lakes commercial bait trade. *Conservation Biology*, doi:10.1111/cobi.12381.

van Orsouw NJ, Hogers RCJ, Janssen A *et al.* (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE*, **2**, e1172.

Pegard A, Miquel C, Valentini A *et al.* (2009) Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from feces. *Journal of Agricultural and Food Chemistry*, **57**, 5700–5706.

Piñol J, San Andrés V, Clare EL, Mir G, Symondson WOC (2014) A pragmatic approach to the analysis of diets of generalist predators: the use of next-generation sequencing with no blocking probes. *Molecular Ecology Resources*, **14**, 18–26.

Pompanon F, Deagle BE, Symondson WOC *et al.* (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, **21**, 1931–1950.

Qiu X, Wu L, Huang H *et al.* (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Applied and Environmental Microbiology*, **67**, 880–887.

Quéméré E, Hibert F, Miquel C *et al.* (2013) A DNA metabarcoding study of a primate dietary diversity and plasticity across its entire fragmented range. *PLoS ONE*, **8**, e58971.

Quince C, Lanzén A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.

Rasmussen M, Cummings LS, Gilbert MTP *et al.* (2009) Response to Comment by Goldberg et al. on "DNA from Pre-Clovis Human Coprolites in Oregon, North America". *Science*, doi:10.1126/science.1167672.

Rittié L, Perbal B (2008) Enzymes used in molecular biology: a useful guide. *Journal of Cell Communication and Signaling*, **2**, 25–45.

Shapiro BA, Hofreiter M (eds) (2012) *Ancient DNA: Methods and Protocols*. Humana Press.

Shehzad W, Riaz T, Nawaz MA *et al.* (2012) Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (*Prionailurus bengalensis*) in Pakistan. *Molecular Ecology*, **21**, 1951–1965.

Shin S, Lee TK, Han JM, Park J (2014) Regional effects on chimera formation in 454 pyrosequenced amplicons from a mock community. *Journal of Microbiology*, **52**, 566–573.

Smyth RP, Schlub TE, Grimm A *et al.* (2010) Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene*, **469**, 45–51.

Soininen EM, Valentini A, Coissac E *et al.* (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.

Sønstebø JH, Gielly L, Brysting AK *et al.* (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Molecular Ecology Resources*, **10**, 1009–1018.

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012a) Environmental DNA. *Molecular ecology*, **21**, 1789–1793.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012b) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.

Taylor PG (1996) Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Molecular Biology and Evolution*, **13**, 283–285.

Valentini A, Miquel C, Nawaz MA *et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular Ecology Resources*, **9**, 51–60.

Vo ATE, Jedlicka JA (2014) Protocols for metagenomic DNA extraction and Illumina amplicon library preparation for faecal and swab samples. *Molecular Ecology Resources*, **14**, 1183–1197.

Wang GC, Wang Y (1996) The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology (Reading, England)*, **142**(Pt 5), 1107–1114.

Wang GC, 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**, 4645–4650.

Zeale MRK, Butlin RK, Barker GLA, Lees DC, Jones G (2010) Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. *Molecular Ecology Resources*, **11**, 236–244.

Zinger L, Gobet A, Pommier T (2012) Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, **21**, 1878–1896.

I.B.S., K.B. and M.T.P.G. conceived the study. I.B.S. and K.B. participated in data generation, and I.B.S. participated in data analyses. K.B. participated in drafting of explanatory figures. K.B. wrote the original manuscript. I.B.S., K.B. and M.T.P.G. participated in edits and comments.

Data Accessibility

Sequencing data available from the Dryad Digital Repository, DOI:10.5061/dryad.64687. Information on tag combinations in the sequencing outputs and errors in tags are presented in Table 1, Fig. 5 and Appendices S3 and S4 (Supporting information).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Details of extractions, amplifications and preparations for sequencing.

Appendix S2 Details of qPCR amplifications carried out with tagged and untagged primers and statistics on differences in Ct values.

Appendix S3 Calculations of expected distribution of tag combinations and statistics, heat maps and statistics on errors in tag sequences.

Appendix S4

Table S1 Number of sequences assigned to different categories of tag combinations.

Table S2 Number of unique sequences with tag combinations belonging to the category 'Used, Matching' and 'Used, Nonmatching', before and after removal of chimeric sequences.

Table S3 Number of sequences with different levels of errors in tag sequences and their fraction of merged sequences.