MS. HANNAH NORMAN (Orcid ID : 0000-0002-4800-5375)

DR. CARMELO ANDUJAR (Orcid ID : 0000-0001-9759-7402)

PROF. ALFRIED VOGLER (Orcid ID : 0000-0002-2462-3718)

Article type : Resource Article

A validated workflow for rapid taxonomic assignment and monitoring of a national fauna of bees (Apiformes) using high throughput DNA barcoding

Thomas J. Creedy[1,2]*, Hannah Norman[1,3]*, Cuong Q. Tang[1,4]*, Kai Qing Chin[1], Carmelo Andujar[1,2,5], Paula Arribas[1,2,5], Rory S. O'Connor[6,7], Claire Carvell[8], David G. Notton[1], Alfried P. Vogler[1,2]

[1]Department of Life Sciences, Natural History Museum, Cromwell Rd, London, SW7 5BD, UK

[2]Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot, SL5 7PY, UK

[3]Science and Solutions for a Changing Planet DTP, Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot, SL5 7PY, UK

[4]Current address: NatureMetrics Ltd, CABI Site, Bakeham Lane, Egham, Surrey, TW20 9TY, UK

[5]Current address: Island Ecology and Evolution Research Group (IPNA-CSIC), Astrofísico Fco. Sánchez 3, 38206 La Laguna, Tenerife, Spain.

[6]Faculty of Biological Sciences, University of Leeds, Leeds, UK, LS2 9JT

[7]Current address: School of Agriculture, Policy and Development, University of Reading, Whiteknights, PO Box 237, Reading, UK, RG6 6AR

[8]NERC Centre for Ecology & Hydrology, Benson Lane, Crowmarsh Gifford, Wallingford OX10 8BB, UK

*These authors contributed equally.

Author for correspondence:

Hannah Norman, Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK

Email: hannah.norman14@imperial.ac.uk

ABSTRACT

Improved taxonomic methods are needed to quantify declining populations of insect pollinators. This study devises a high-throughput DNA barcoding protocol for a regional fauna (United Kingdom) of bees (Apiformes), consisting of reference library construction, a proof-of-concept monitoring scheme, and the deep barcoding of individuals to assess potential artefacts and organismal associations. A reference database of Cytochrome Oxidase subunit 1 (*cox1*) sequences including 92.4% of 278 bee species known from the UK showed high congruence with morphological taxon concepts, but molecular species delimitations resulted in numerous split and (fewer) lumped entities within the Linnaean species. Double tagging permitted deep Illumina sequencing of 762 separate individuals of bees from a UK-wide survey. Extracting the target barcode from the amplicon mix required a new protocol employing read abundance and phylogenetic position, which revealed 180 molecular entities of Apiformes identifiable to species. An additional 72 entities were ascribed to nuclear pseudogenes based on patterns of read abundance and phylogenetic relatedness to the

reference set. Clustering of reads revealed a range of secondary Operational Taxonomic Units (OTUs) in almost all samples, resulting from traces of insect species caught in the same traps, organisms associated with the insects including a known mite parasite of bees, and the common detection of human DNA, besides evidence for low-level cross-contamination in pan traps and laboratory procedures. Custom scripts were generated to conduct critical steps of the bioinformatics protocol. The resources built here will greatly aid DNA-based monitoring to inform management and conservation policies for the protection of pollinators.

Key words: Pollinators, community barcoding, contamination, Illumina sequencing, double dual tagging.

INTRODUCTION

Widespread declines in pollinator populations are causing concern about the future of global biodiversity and agricultural productivity (Garibaldi *et al.* 2013; Hallmann *et al.* 2017; Lever *et al.* 2014), driven by the combined effects of habitat loss, introduction of non-native and invasive species, pathogens and parasites, and various other factors contributing to environmental change (Vanbergen *et al.* 2013). Landscape effects on pollination of crops through agricultural intensification, particularly the use of monoculture crops, have led to significant changes in pollinator communities (Kennedy *et al.* 2013; Ricketts *et al.* 2008), with obvious economic implications for the agricultural sector and pollination services worth hundreds of millions of pounds in the United Kingdom alone (Potts *et al.* 2010). However, these trends in species distribution and abundance are difficult to quantify, unless solid methodologies for monitoring at regional levels can be implemented. Thus, there is an urgent need to develop strategies for large-scale and long-term systematic monitoring of pollinator populations, to better understand the impacts of declines on pollination services to crops and wild plants, and inform policy decisions and conservation efforts.

Current evidence of change in pollinator populations in the United Kingdom comes primarily from records of species occurrence submitted by volunteer recorders (e.g. Biesmeijer *et al.* 2006; Powney *et al.* 2019). While these allow for the analysis of large-scale changes in species distributions, they provide no information on abundance or local population size, and are known to be temporally and spatially biased (Isaac & Pocock 2015). Instead, pan traps have been proposed as the most effective method for systematic monitoring of bee diversity in European agricultural and grassland habitats (Westphal *et al.* 2008). Species identification is usually performed by expert taxonomists, but there is a growing need for alternative methods, in particular because the great species diversity and large quantity of specimens from mass trapping make them challenging and costly to identify (Lebuhn *et al.* 2013).

This impediment may be overcome through the use of high throughput sequencing (HTS) techniques to identify species by their DNA 'barcode' at the individual or community level. A suitable HTS-based approach for large-scale DNA barcoding could assay thousands of specimens potentially generated in the course of a pollinator monitoring scheme. The great sequencing depth of such high-throughput barcoding (HT barcoding) methodology may also reveal DNA from organisms internally or externally associated with a target specimen or as a carry-over from other specimens in the trap. Similarly, the more recent approach of 'metabarcoding', by which entire trap catches are subjected to high throughput amplicon sequencing in bulk, produces community-level species incidence data based on the mixed sequence read profile (Yoccoz *et al.* 2012). Current HTS protocols can maintain the individual information of thousands of samples using unique tags in the initial PCR prior to pooling for Illumina sequencing with secondary tags, which allows sequences to be traced back to the associated specimen (Arribas et al. 2016; Shokralla *et al.* 2015). However, crucial to these approaches is a comprehensive and validated reference set of DNA sequence data from target species to provide accurate and verifiable molecular identification.

This study lays the groundwork for the use of HTS techniques to assess diversity and abundance of mass-trapped samples of a regional-scale pollinator fauna, focused on the bees (Hymenoptera: Apiformes) of the United Kingdom. The first step in this process was the generation of a well curated reference database for the 278 species of bees known from the UK using the Cytochrome *c* Oxidase subunit I (*cox1*) barcode marker (Hebert *et al.* 2003), which provides good species discrimination in Hymenoptera (Smith *et al.* 2008). This reference set was then used to identify HTS-obtained short barcode reads from 762 bee specimens gathered as part of a pilot study for a national monitoring scheme. Morphological identifications performed in parallel provided comparisons of molecular identification against conventional methodology and allowed further refinement of the database. In addition, the deep-sequencing approach allowed the assessment of organisms associated with the target specimens, as well as cross-contamination from other species present in the traps or from specimen handling and laboratory procedures.

MATERIALS AND METHODS

Building a regional reference database

A *cox1* reference database was generated from DNA barcoding of bee species known to occur in the UK according to the list of Falk and Lewington (2015) and notes from various sources maintained by co-author DGN. Most specimens were caught by hand netting and identified by DGN, using the latest keys available at the time (Amiet *et al.* 2001, 2004, 2010; Amiet *et al.* 2007; Amiet *et al.* 2014; Bogusch & Straka 2012; Falk & Lewington 2015; Benton 2006; Mueller 2016). Identifications had to draw on these various references because the comprehensive key of Falk & Lewington (2015) became available only part way through the study, while some identifications were also cross-checked between different publications. Additional specimens were obtained using pan traps from the survey described below. The reference set included all available unique UK species as determined by morphology, with multiple specimens per species where available. These

within-species replicates allowed inclusion of specimens from across the geographical range of species, identified by different taxonomists, or belonging to species complexes. Specimen data for morphological vouchers are available at the Natural History Museum Data Portal (data.nhm.ac.uk; http://dx.doi.org/10.5519/0002965).

DNA was extracted from a hind leg using a Qiagen DNeasy Blood and Tissue Kit, after incubation at 56°C in extraction buffer (ATL and Proteinase K) overnight in a shaking incubator at 75 rpm. The complete *cox1* 'barcode region' (658 bp) was amplified using primers (BEEf TWYTCWACWAAYCATAAAGATATTGG and BEEr TAWACTTCWGGRTGWCCAAAAAATCA) newly designed based on an alignment of 84 mitochondrial genomes from 22 genera of Apiformia. PCR and sequencing using ABI dye terminator technology followed standard procedures (Supplementary Material). Sequences were deposited in BOLD (Barcode of Life Datasystem) under the project BEEEE, along with Syrphidae barcodes that were sequenced at the same time.

Sequences were aligned using the *MAFFT* v1.3. (Katoh *et al.* 2009) plugin in *Geneious*. Alignments were used for distance-based and coalescence-based species delimitation: (1) BINs (Barcode Identification Numbers) were automatically generated for sequences uploaded onto BOLD database, employing a single linkage network method (Ratnasingham & Hebert 2013). (2) The GMYC (Generalized Mixed Yule Coalescent) method for separating independent coalescent groups (Fujisawa & Barraclough 2013) was performed on phylogenetic trees constructed separately for each genus using BEAST 1.8.1 (Drummond & Rambaut 2007). For genera with only a single British representative (*Apis*, *Anthidium*, *Ceratina*, *Dasypoda*, *Macropis* and *Rophites*), GMYC was conducted after adding congeneric sequences from BOLD.

Generating a test dataset from field caught samples using HTS

The reference database was used for identification of specimens obtained through the National Pollinator and Pollination Monitoring Framework (NPPMF; Carvell *et al.* 2016). Mixed samples were collected with pan traps consisting of sets of water-filled bowls (painted UV-yellow, white and blue; after Westphal *et al.* 2008) from 14 sites across the UK, and further specimens were collected by netting along standardised transects running 200 m from each set of pan traps (Figure 1A; see Carvell *et al.* 2016 and supplementary materials for a full description of the sampling protocol). Bees (Apiformes) were separated from other taxa in the field, stored in 99% ethanol, and transferred to -20°C as soon as possible after collection. Specimens were identified morphologically by taxonomists offering commercial identification services. In total, 762 bee specimens were processed and individually sequenced. All specimens were stored in 99% ethanol and deposited as voucher specimens in the Molecular Collection Facility at the NHMUK.

DNA was extracted from individual specimens by piercing the abdomen and submerging the whole specimen in lysis solution consisting 200 ul ATL/Proteinase K buffer for 12 hours on a 56°C shaking incubator. DNA extractions were performed using either the Qiagen BioSprint 96 DNA Blood Kit or DNeasy Blood and Tissue kits applied to the lysate. Each DNA extract was PCR amplified for a 418 bp portion of the *cox1* barcode region (Andujar *et al.* 2018; supplementary materials). Amplicons for each individual were tagged using a 'double dual' PCR protocol (Shokralla *et al.* 2015) to generate unique tag combinations for each bee specimen, following the procedures of Arribas *et al.* (2016). Tags were added in the initial PCR by amplification using *cox1* primers with different 6 bp tags with a Hamming distance of 3, with a total of 8 different tagged primer sets. In all reactions, forward and reverse primers used the same tag, so that the products of tag jumping could be detected (Schnell *et al.* 2015). Amplicons generated with different primer tags were merged into 96 pools of 8 and cleaned using Agencourt AMPure XP beads (Beckman Coulter, Wycombe, UK). Secondary amplification of each pool was performed with i5 and i7 Nextera XT indices using unique

MID combinations (Illumina, CA, USA) for each of the 96 final libraries, which were then sequenced on about 50% of a flow cell of an Illumina MiSeq v.3 (2x300 bp paired-end).

Perl scripts of the custom NAPtime pipeline (www.github.com/tjcreedy/NAPtime) were used to wrap bioinformatics filtering of the raw data. The 96 libraries were demultiplexed based on XT indices using Illumina software and were further demultiplexed using NAPdemux based on the unique tags of the first-round PCR primers to separate reads into 762 paired read files. This script wraps cutadapt (Martin 2011) for large demultiplexing runs, and used the default 10% permitted mismatch to the adapter sequences (i.e. permitting no errors in the 6 bp tag used) before binning reads according to their tags. Mate pairs with only one read matching the correct tag were discarded. Read quality was reviewed using FASTQC (Andrews 2010). Following demultiplexing, the NAPmerge script was used to generate a set of full-length reads for further analysis. The script invokes cutadapt (Martin 2011), PEAR (Zhang *et al.* 2014) and *USEARCH -fastq_filter* (Edgar 2010) to remove primer sequences, assemble read pairs, and perform quality filtering respectively. Any reads not containing a correct primer sequence, and their mates, were discarded, and any merged reads with 1 or more expected errors were removed with *fastq_filter*; otherwise, wrapped software used default parameters. This process generated a pool of complete *cox1* amplicon sequences for each of the specimens.

Testing the utility of the reference dataset

From the set of reads obtained for each specimen a single putative “high-throughput barcode” (HT barcode) sequence was designated to represent the *cox1* gene of that specimen. Three methods were used to identify this HT barcode from the read mixture.

Firstly, we generated OTU (Operational Taxonomic Units) clusters and centroid sequences using *USEARCH cluster_otus*. This was performed using a metabarcoding pipeline, implemented in the NAPcluster script, which includes standard functions from the *USEARCH* suite (Edgar 2010), starting with the data output from NAPmerge (merged and quality-filtered amplicons), and comprises the following steps: (i) filtering sequences by length; (ii) dereplication and filtering by number of reads per unique sequence, to retain only sequences represented by a set minimum of reads; (iii) denoising using the *UNOISE* algorithm (Edgar *et al.* 2011); (iv) clustering of sequences according to cluster radius and generation of an output set of OTU consensus sequences, and (v) mapping of reads to OTU clusters (using *USEARCH usearch_global* and a custom .uc parser) and generation of an output table of OTU read numbers by sample. All sequences differing from 418 bp and with only 1 copy were removed in steps (i) and (ii), and *USEARCH cluster_otus* was employed for clustering with a dissimilarity threshold of 3%. The centroid of the most abundant OTU, the 'top OTU' was used as the HT barcode.

The second method for selecting the HT barcode sequence simply chooses the most frequent read for each library (see supplementary materials for details), under the assumption that the most abundant template DNA represents the target specimen. This 'top read' method is based on simple read counts of all unique sequences after quality filtering, with no length filtering. As such, it does not distinguish among variants present in the mixture, including erroneous variants resulting from PCR or sequencing errors, or true variants resulting from co-amplification of nuclear mitochondrial DNA segments (*Numts*), gut contents, internalised parasites, and cross-contamination, possibly leading to frequent incorrect sequence selection.

The third method also finds the most highly represented sequence among reads in an amplicon mix, but limits the selection to reads of the expected amplicon length of 418 bp, and validates this selection statistically and taxonomically in order to avoid these variants. The process

starts with the length filtering (in this case, rejecting any sequence not 418 bp), and groups sequences by identity (i.e. dereplication), recording the abundance of reads representing each unique sequence. Starting with the most abundant, each unique sequence is assessed for the significance of the difference in read abundance by bootstrapping. Given the total number of reads in the sample and the number of unique sequences, the probability of a sequence occurring as frequently as the most abundant sequence by chance alone is estimated using 10,000 bootstrap iterations (p-value). A p-value of 0 designates a sequence as significantly more frequent with high confidence, and less than 0.5 for low confidence, above which the entire sample is disregarded because a most abundant barcode sequence for the target specimen is not clearly defined. Finally the most abundant sequences revealed by this procedure are subjected to a *BLASTn* search against the NCBI *nt* database and the hits assessed for the focal taxon (in this case, Hymenoptera). If the most abundant read matches a different taxon, the sample was removed from further consideration. If a sequence fails the bootstrapping test, it is merged with the next most frequent sequence if their similarity is above a given threshold (99% was used here). If this merge occurs, the process restarts; if they are not sufficiently similar, the sequence is output with "low confidence" if it passes the *BLAST* test, or discarded and the process restarted if not. Sequences passing both tests are output as "high confidence". The method was implemented in a purpose-built tool, NAPselect, and the process is visualised in Supplementary Figure S1.

The success of these three methods, and the accuracy of the sequences they output, was tested by identifying the HT barcodes against the BEEEE reference collection generated above using *BLASTn* with default parameters. Only matches with >95% identity and overlap with the reference sequences of >400 bp were retained, and the match with the highest similarity was selected, using bitscore to break ties. The taxon identity of this hit was compared against the morphological identifications supplied by taxonomists. For each HT barcode selection method and taxonomic level, the number of correct molecular identifications at the genus and species levels was tallied.

Exploration of concomitant DNA in the testing dataset

The OTUs generated with the NAPcluster script (see above) allowed the exploration of co-amplified DNA from each bee specimen other than the primary *cox1* sequence, including contaminants. Specifically, the OTUs that did not match the NAPselect HT barcode sequence for the target specimen were designated as "secondary OTUs". These OTUs were searched against the NCBI *nt* database using *BLASTn*, followed by taxonomic binning using *MEGAN6* Community Edition with the weighted Lowest Common Ancestor algorithm (Huson *et al.* 2016). Any OTUs assigned to Apiformes were additionally identified using *BLASTn* (default parameters) against the BEEEE reference collection and the NAPselect HT barcodes (as above).

Numts may appear as separate OTUs in metabarcode data and add spurious OTUs to the clusters derived from the true mitochondrial copy. A tree-based filtering pipeline was used to identify *Numt*-derived OTUs based on the assumption that they are closely related to the corresponding mitochondrial copy, and are coincident across sequenced samples, while their copy number is lower. Thus, OTUs were considered derived from *Numts* if their presence completely coincided across samples with another closely related OTU that matched a BEEEE reference, and the number of reads was significantly lower in comparison.

The resulting datasets were reconfigured for various statistics and to perform downstream calculations using R (R Core Team 2018). The OTU x sample dataset was rarefied to 400 reads per sample to facilitate valid comparison between samples using the R package *vegan* (Oksanen *et al.* 2018).

Cross-contamination among samples was tested by assessing the distribution of secondary OTUs in each sample obtained from pan trapping. Only secondary OTUs that matched a (NAPselect) HT barcode from *another* sample were used in this analysis. Three sources of cross-contamination

were considered: from other individuals in the same trap, between specimens with the same PCR tag on a single plate, and between specimens with the same Nextera XT index in a single well. For each source or combination of sources, we calculated the proportion of total possible selected barcodes that were present as secondary OTUs in a sample. For example, each well in the library preparation contained 13 specimens with different XT indices: if in a set of these, each is a different HT barcode, there are 12 possible well contaminants for any one of these samples, thus a sample containing 3 other HT barcodes as secondary OTUs would have a contamination rate of 3/12 = 0.25 from well-level contamination. As a control, the rate of contamination from all possible sources together was also scored, i.e. the proportion of secondary OTUs in a sample that matched *any* HT barcodes, out of the total number of unique HT barcodes. One-sample *t* tests were used to assess if the mean contamination rate for each source or source combination was significantly greater than zero. To compare between sources against the control, the effect of source on contamination rate was fitted in a quasi-binomial ANOVA, setting the control as the reference level.

RESULTS

A reference database of UK bees

A total of 355 bee specimens were newly sequenced for the COI barcode to generate the reference set, representing 165 Linnaean species. These new sequences were compared against 1754 full-length barcode sequences obtained from the Barcode of Life Database (BOLD) for species known from the UK (Fig. 1A). The BOLD data represented 245 of the 278 UK bee species, but comprised only 14 sequences (6 species) from specimens collected in the UK. The 355 new sequences add 10 UK species (15 sequences) not represented in the BOLD dataset, and novel haplotypes for 107 further species (201 sequences). Together, the two sets include 255 bee species (92.4% of 278 species known from the UK). The missing species are either extinct (6 species), rarely introduced by accident (1 species, *Heriades rubicola*), only found in the Channel Islands (1 species, *Andrena agilissima*), listed as endangered (RDB3-RDB1) (8 species), or rare and localised (5 species), while 2 species were

only recently added (Cross & Notton 2017; Notton *et al.* 2016). When considering each of the six families separately, the greatest number of species missing from the database was in Andrenidae (9 of 69 species), followed by Apidae (4 of 76) and Halictidae (4 of 62).

Genetic variation within morphologically identified Linnaean species ranged from 0% to 5.9% (mean 0.31%, standard error ±0.04%), and interspecific variation ranged from 0% to 24.9% (mean 6.7% ±0.08%). We found that 242 (94.9%) of the *cox1*-based sequence clusters at 97% similarity mapped precisely on the Linnaean species identifications (Supplementary Figure S2). Inconsistency with the morphological species definitions were limited to five genera, *Andrena, Bombus, Colletes, Lasioglossum* and *Nomada*.

De novo species delimitation from the DNA sequences using the GMYC method were based on phylogenetic trees generated for each genus (see Fig. 2 for the genus *Nomada*). In most cases of incongruence, the GMYC either split (42 cases) or lumped (14 cases) an existing nominal species, but in rare cases the patterns of splitting and lumping were more complex (Fig. 3). The GMYC species largely agreed with the distance-based BIN network method in the extent to which nominal species were split and lumped (Fig. 2, 3). Inconsistencies of Linnaean and *cox1*-based entities were mainly due to groups of close relatives with challenging morphological identifications. Subsets of species not monophyletic with respect to each other (a requirement of the GMYC method) included: *Andrena bimaculata* - *A. tibialis*; *A. clarkella* - *A. lapponica* - *A. helvola* - *A. varians*; the recently subdivided *Colletes succinctus* species group (*C. halophilus* - *C. hederae* - *C. succinctus*) (Kuhlmann *et al.* 2007); suspected geographically confined species among the *Dasypoda hirtipes* group (Schmidt *et al.* 2015); and groupings within *Lasioglossum rufitarse*, *Nomada flava* - *N. leucophthalma* - *N. panzeri*, and *N. goodeniana* - *N. succincta* clusters.

Testing HTS data against the reference library

Illumina sequencing of 762 specimens from the UK survey resulted in an average of 9,025 read pairs per amplicon pool after demultiplexing (sd = 10,615; range = 18 - 88,241; 95% > 1,000). After read merging and stringent quality filtering, this was reduced to an average of 5,851 *cox1* sequences per specimen (sd = 6,921; range = 7 - 56,781; 87% > 1,000). Three methods were used to designate a HT barcode from these sequences for each specimen (see Materials and Methods). The NAPselect method, which validates barcode selection by statistical significance of read abundance and taxonomy, obtained a barcode for 749 individuals, failing to do so for 13 specimens (Table 1A). The latter mainly comprised libraries with very low read numbers, which were removed based on the taxonomic (no matches to Hymenoptera) or statistical (low discrimination among top abundant reads) filtering (see Supplementary Material). Given this result, we were able to leverage the wide variation in read numbers to explore the effectivity of NAPselect at different read values per sample. Figure 4 shows that while performance is poor below 500 reads per sample, the percentage of libraries producing HT barcodes based on the taxonomically validated top read reaches 90% at around 1000 reads.

Out of the barcodes chosen by NAPselect, 734 (99.7%) produced a match to sequences in the BEEEE reference set (Table 1A), while the OTU clustering and top-read methods had substantially fewer matches at 559 and 584, respectively. The number of species-level hits against the reference set for all methods was near 100%, but because the barcodes obtained by the top OTU and top read methods matched fewer reference sequences, they resulted in approx. 25% fewer specimen identifications. Across all samples, 154 unique species identifications against the BEEEE reference set were obtained for the survey samples.

Congruence of species-level molecular identifications with the species-level morphological identifications of the source specimens was high at genus level with 95-96%, with 83-86% of specimens identified as the same species with both data types (Table 1B). However, as NAPselect designated a considerably larger proportion of barcodes to species level, the absolute number of correct species identifications using this method was the highest, at 611 specimens out of the 762 sequenced (707 correct at genus level). The success rate of molecular identification differed among genera (Figure 5), in particular among the species-rich genera; for example, *Andrena* and *Bombus* produced markedly more successful identifications, whereas *Colletes* showed low success even using NAPselect (as expected because some species were inseparable by DNA; see above).

We investigated whether the lumping and splitting observed in the reference dataset was a driver of molecular misidentification by examining the proportion of correct and incorrect matches against species that were lumped and/or split in the GMYC analysis. Of the 734 HT barcode sequences generated by NAPselect that had a BLAST match to a BEEEE reference sequence, 17 were to a species that was lumped, 178 to a species that was split and 1 to a species that was both lumped and split. The proportion of correct species and genus level matches for these sets of HT barcodes was very similar to the overall rate: 76.5% of matches to lumped species and 88.2% of matches to split species were correct at the species level (94.1% and 98.8% at the genus level), and the single HT barcode matching a lumped *and* split species was correct at the species level as well.

To account for other causes of the inconsistencies in morphological versus molecular identifications, all of the 731 NAPselect HT barcodes were combined with the 335 sequences of the reference set and subjected to phylogenetic analysis (Supplementary Figure S3). The resulting tree generally grouped sequences in small clusters of close relatives that mostly correspond well with the Linnaean taxon names (similar to Fig. 2), but a total of 136 NAPselect sequences were at least

partially in conflict with the names assigned to sequences in each cluster. We found evidence for problems with both the molecular and morphological identifications that may account for most of the observed discrepancies. Library contamination at either the PCR level (within a plate of sequences) or secondary tagging (within a well) may be recognisable from a) identical sequences or clusters grouping distantly related species, or b) from the recovery of secondary OTUs with the correct molecular identification. This former inconsistency is observed in around 30 individuals, including a notable cluster assigned to *Andrena labialis* based on the reference set that contained one representative of 15 different species with exact matches to the reference haplotype (in addition to several closely related haplotypes from further species). All of these sequences were from a single plate, suggesting contamination of the tagged primer with *A. labialis* DNA. Examining secondary OTUs, 28 of the 136 mismatched samples contained secondary OTUs that matched the morphological ID for that specimen, but they represented less than 10% of the reads in 20 of these this OTU and between 10 and 40% of the reads in a further six samples.

Over 100 inconsistent records involved mixed clusters of close relatives. Notable are several species pairs of *Lasioglossum* (*L. albipes* - *L. calceatum*, *L. fulvicorne* - *L. fratellum*), *Andrena* (*A. wilkella* - *A. similis*) and *Bombus* (the *B. terrestris/lucorum* complex) whose morphological identification in the current study resulted in mixed species clusters, or whose identification differed from the name assigned to the cluster based on the reference sequence. A review of the morphological identifications of a small sample of 10 mismatched and 14 non-mismatched specimens was undertaken by DGN, and found that while the identifications for the non-mismatched samples were 100% correct, 78% of the mismatched specimens were incorrect, and the correct identification matched the molecular ID either precisely, or to species group. The majority of the 'mismatches' are down to errors in the mophological identification, and the 83-86% rate observed rises to over 97% correct molecular identifications after removal of obvious contaminants and accounting for the misidentification rate.

Exploration of concomitant DNA in the testing dataset

When OTU clustering was carried out on the entire data set of reads from all 762 samples, USEARCH within NAPcluster generated 498 OTUs, of which 263 were identified as Apiformes using BLAST/MEGAN. Out of these, the tree-based assessment of potential *Numts* identified 72 OTUs as likely *Numts*. In addition, several OTUs were reclassified as Diptera in the phylogenetic tree used for *Numt* filtering. The final count of *bona fide* OTUs identified as Apiformes was 180, of which 170 had hits to the BEEEE reference library. Apiformes thus dominated the set of OTUs, but the dataset also included 235 OTUs from across the eukaryotes, including Diptera (48 OTUs), Coleoptera (6 OTUs), and various other insects (22 OTUs). The Diptera included several species of hoverflies (Syrphidae), which were present in the traps and were sequenced in the same run as the bees. Five of the six Coleoptera OTUs were identified as common flower visitors, including three species of *Cantharis* Soldier Beetles (Cantharidae), Malachite Beetles (Malachiidae) and a Pollen Beetle (Nitidulidae), in addition to *Zophobas atratus*, a non-native species of Darkling Beetle (Tenebrionidae). There were also OTUs from organisms that associate directly with bees such as Acari (mites) and *Wolbachia* (alphaproteobacteria), as well as several flowering plants and numerous fungi and oomycetes. The Acari comprised four OTUs, of which one was identified to species, *Locustacarus buchneri,* a known tracheal parasite of bumblebees, while the others were identified only as members of the Sarcoptiformes, Crotonioidea and Parasitiformes. Finally, *Homo sapiens* DNA was detected in numerous samples. The supplementary materials and figure S4 report further details of secondary OTU community composition.

The high incidence of NAPselect barcode sequences (i.e. Apiformes) occurring as secondary OTUs raised the question about the origin of these non-target specimens in the barcoding mix. Potential sources of DNA may be carry-over from the traps, mixing of specimens during handling for taxonomic identification, errors in various DNA laboratory procedures, and errors in tag sequencing.

In general, the level of direct contamination with DNA sequences that were the HT barcode in another sample was low, but significantly greater than zero for most sources and source combinations (Supplementary Table S1). Altogether, 132 of the 180 Apiformes OTUs were recognised as secondary OTUs in at least one sample. Compared with the control, i.e. the background level of cross-contamination from any source, there was a significant increase in contamination rate for within-plate contamination and within-plate and trap contamination (Fig. 6, Supplementary Table S1), indicating that the greatest rate of contamination may have been during primary PCR. The level of cross-contamination was much lower from those samples in the same well, i.e. from secondary PCR at the library construction stage.

The low level of contamination was reflected in the pattern of cross-contamination of individual species. OTUs identified to 23 different species were each found as secondary OTUs in at least one other sample of a different species from the same trap. The most frequent of these was *Lasioglossum malachurum*, of which there were 37 specimens in the study from 21 traps. We HT barcoded 63 specimens of other species from these 21 traps, and *L. malachurum* was found in 13 of these, a rate of 20%. At trap level, the average rate for the 23 species was 7.6% (SD = 4.5). The same analysis for plates and wells showed that *Lasioglossum calceatum* was the most common cross-contamination here, being found in 7% of samples of other species sharing a plate (PCR tag) with specimens of *L. calceatum*, and 5% of samples of other species sharing a well (MID) with *L. calceatum*. 45 species cross-contaminated within plates, with a mean rate of 2.2% (SD = 1.7), and 13 species cross-contaminated within wells (mean = 2.5%, SD = 1.1).

DISCUSSION

Sample collection methodology

Cost-effective species-level identifications of bees and other insect pollinators are required to provide robust evidence for population changes and to inform land use management and conservation (Gill *et al.* 2016). This study used specimens of bees obtained through mass-collection with pan traps, which was successful in providing a wide range of species for the generation of a reference database and for testing. It should be mentioned that in the wider context of pollinator declines (Powney *et al.* 2019), and invertebrate declines in general (Hallmann *et al.* 2017), careful consideration of the use of broad-target collection methods with high collateral catches should be made (Drinkwater, Robinson and Hart, 2019), although Gezon *et al.* (2015) show that in the case of pan traps in particular, reasonable sampling does not affect long term community structure. Our study protocol used a relatively short pan trap exposure period designed to sample sufficient individuals for long-term monitoring whilst minimising catch sizes (Carvell *et al.* 2016). In this study, we demonstrate that bulk-collection methods may generate unwanted levels of cross-contamination for downstream molecular analysis, although robust bioinformatic methods can minimise the impact. More broadly, the growing use of metabarcoding as a tool for arthropod community studies allows us to take fuller advantage of the depth of data produced by mass-collecting than ever before, including the 'bycatch' of numerous other insect pollinators, mostly in the Diptera, which are taxonomically difficult and thus have not been part of conventional monitoring.

The reference database

We conducted this analysis in two stages, by first building the reference database using Sanger technology, which was then trialled for species identification using high-throughput sequencing of samples from a proof-of-concept monitoring scheme. The combined effort of new sampling and

sequencing, together with barcode data already in the BOLD database, resulted in a virtually complete set of the UK bees, with only a few rare or presumed extinct species missing. Furthermore, we expanded existing references by generating novel sequences from UK populations of widespread species. The *cox1* barcode delimited 94.9% of species in the reference database as separate entities, showing that for almost all bee species in the UK this set is sufficiently discriminatory. In the remaining cases the molecular analysis lumped the Linnaean species, as evident in the *de novo* species delimitation using the GMYC method, while an even greater proportion were shown to be split into additional GMYC groups which, however, were not incongruent with the Linnaean species.

The overall reference database comprises a mixture of UK and non-UK sequences, as many species are more widely distributed in Europe and North America. We found generally high congruence of molecular groups with the Linnaean species, which shows that the mitochondrial 'gene trees' are a good reflection of the species-level entities, as both morphological diagnostics and mitochondrial markers corroborate the species hypotheses (DeSalle *et al.* 2005). Species discrimination may be even clearer if performed with UK samples only, as the species-level differences tend to be exacerbated in local subsets (Bergsten *et al.* 2012). The UK sample contributed many new haplotypes that may add to the power of species discrimination locally. The high congruence with the BOLD database also suggests that the identifications have been correct, in some cases after secondary inspection of specimens. However, some problems remained with the reference database, which was apparent from the 136 HTS sequences with inconsistent morphological and molecular identifications. We attribute most of this failure to either contamination or misidentification of the specimens by the NPPMF taxonomists, rather than an issue with the reference set or with the NAPselect pipeline. While low rates of cross-contamination are certainly observed across the HTS dataset, there is little evidence that this was substantially higher in the mismatched specimens, observing the correct sequences as secondary reads in only a

small minority of cases and many of these at a read abundance no higher than the background rate. It appears most likely that most mismatches are due to incorrect or unclear morphological identifications, based on our small-scale re-identification. In part, these problems affected the known cases of taxonomically problematic groups, in particular in *Lasioglossum*, *Nomada* and *Bombus*, to which this study added a large number of sequences that may be useful for a refinement of the database. We note that in none of these cases did the HT barcodes add new sequence clusters beyond those already represented in the reference set. It is clear that the 86% rate of correct molecular species identifications for NAPselect (Table 1B) is an underestimate, and that most of the 136 'mismatched' HT barcodes can in fact be correctly identified through comparison to the reference set, as shown by our reassessment of a subset of these specimens. After removal of contaminants and correction for taxonomic misidentifications, the true rate of correct identification against the reference set is closer to 97%.

The reference includes DNA clusters (established by the BIN or GMYC methods) that lumped or (mostly) split the Linnaean taxa. The molecular data failed to separate a small number of species in four of the 27 genera studied ("lumped" in Fig. 3). In some instances, such as the *Colletes succinctus* species group, which shared haplotypes with *C. hederae*, morphological identification of three named species is reliable, if challenging, now that there is a key covering all UK species (Falk & Lewington 2015), and there are biological and distributional differences while c*ox1* sequences are not sufficient to delineate these species (Kuhlmann *et al.*, 2007. Similarly, the separation of the *Nomada goodeniana-succincta* group relies on subtle colour variants (Falk & Lewington 2015) and they were not separable in our analysis (Fig. 2). Additional genetic markers may be useful; e.g. the three recognised *Colletes* species lumped in *cox1* exhibit fixed differences in EF-1a and ITS (Kuhlmann 2007). Vice versa, divergent *cox1* entities (splitting) may indicate the existence of hitherto unrecognised species. For example, a divergent haplotype in *Dasypoda hirtipes* has now been associated with a morphologically differentiated, eastern European species, although it is not

part of the UK fauna (Schmidt *et al.* 2015). We have already curated the *cox1* database extensively, in particular to remove morphological identification errors, and the remaining problems affect mostly a few species of *Lasioglossum* that also accounted for most of the inconsistencies of molecular and morphological identification (Supplementary Text and Supplementary Fig. S3). In addition, the newly detected clusters may lead to the discovery of separate entities within the Linnaean species and may provide fertile ground for future morphological work. Since DNA extraction destroyed only one leg, morphological vouchers can be re-examined to refine the reference database.

Generating high throughput barcodes

High-throughput barcoding ("HT barcoding") was then used to identify species from a survey of pan traps. The methodology has great potential for sequencing mixed samples (metabarcoding) but was here applied on individual specimens to test the efficacy of this approach and our ability to confidently recover a sequence for the target specimen. We employed three methods for designating this sequence from a pool of anonymous amplicons. The most intuitive approach was to undertake a standard metabarcoding analysis using the USEARCH pipeline to designate the centroid sequence of the most highly represented OTU in each sample as the HT barcode. However, the sequence obtained with this method did not produce a BLAST hit to the reference database in 27% of cases. An alternative method was to simply select the most frequent unique sequence in the amplicon pool, analogous to the sequence that would be generated by Sanger sequencing. However, while this method also designates a barcode for every sample, these sequences are only marginally more likely to find a match to the reference database (23% did not produce a BLAST hit).

The third method, implemented in the NAPselect script, also selects the top-abundant read, but requires that this read matches a specific taxonomic group (Hymenoptera), and that the read frequency is significantly greater than frequencies of other reads, besides the requirement for exact matching the predicted sequence length. If these conditions are not met, NAPselect discards the top read and checks other reads according to descending abundance. This pipeline did not output a sequence for 13 specimens due to low read numbers or low differentiation among other abundant reads, although NAPselect generally worked very well at reasonable read numbers. The great majority of NAPselect sequences matched the reference database, and only 3.7% of specimens did not produce a sequence with a BLAST hit - a substantial improvement over the other methods (Table 1). The key improvement introduced by this script probably was that NAPselect conducts BLAST searches against GenBank and assesses the taxonomy of the hits. This method is clearly very effective, with error rates determined largely by sequencing depth issues rather than an inability to select the correct sequence.

Exploration of concomitant DNA in the testing dataset

Unlike standard metabarcoding conducted on mixed samples, the current analysis permits a precise determination of amplicons derived from single specimens. A surprising finding was the high proportion of reads attributable to secondary OTUs, and their taxonomic diversity. Specimens from the monitoring program were not substantially different from those used in Sanger sequencing to build the reference database, which produced clean base calls consistent with a single predominant PCR product. However, the primers for Illumina sequencing were designed for broad amplification of arthropods (Arribas *et al.* 2016) and probably have a wider taxonomic amplitude than the Hymenoptera-specific primers used to amplify the standard barcode region. Besides co-amplification of a broader range of associated species, this may also increase the potential for sequencing of nuclear mitochondrial DNA regions (*Numts*). Out of 509 OTUs recovered from all samples combined,

263 were identified as Apiformes initially, which greatly exceeds the number of species expected in this survey. *Numts* diverge without the constraints of coding regions and thus may deviate in length, but length filtering for the expected 418 bp fragment could not avoid these artefacts sufficiently. We therefore implemented a further filter based on the distribution of low-abundance OTUs that are co-distributed with the true mitochondrial copies. We only removed OTUs that form a clade with the presumed true copy (close matches to the reference database), under the assumption that *Numts* are of limited evolutionary persistence (Pons *et al.* 2005). Based on these criteria a total of 72 OTUs were identified as mitochondrial *Numts*. This method (and the removal of several other OTUs whose incorrect assignment was revealed with the phylogeny) reduced the total number of Apiformes OTU to 180, which is closer to the 154 species identified morphologically, in particular if OTU splitting (Fig. 3) is taken into account. The procedure for identifying OTUs likely derived from *Numts* is a novel step in the metabarcoding filtering process, however it is dependent on the availability of "true" *cox1* reference haplotypes and high variation in read abundance between target *cox1* OTUs and their putative *Numt(s*) – both situations that are common in HT barcoding studies but potentially less so in metabarcoding. Here, it proved to be a critical step preventing the overestimate of species richness frequently seen in metabarcoding studies.

Other secondary OTUs were assignable to a wide range of distantly related taxa, including highly plausible representatives of known coleopteran and dipteran pollinators attracted to flowers (and pan traps). None of the secondary OTUs belonged to species known to have been processed in the same laboratory as these samples. Instead, the detection of pollen beetles (*Meligethes*) was consistent with the presence of numerous specimens in the pan traps. Species of Diptera included the wheat stem borer *Cephus pygmeus*, a flower visitor whose larvae feed in the stems of cereal crops and wild grasses (Poaceae), and *Sarcophaga* sp. (flesh flies) that are carrion feeders or parasitoids of other invertebrates. The greatest proportion were hoverflies (Syrphidae); these were

widely present in the traps and were processed in a parallel study in the same sequencing run and thus additionally exposed to the risk of laboratory contamination as well as trap contamination. Other sequencing records were likely internal parasites, including a tracheal mite, *Locustacarus buchneri*, known to be associated with bumble bees (*Bombus* sp.), and numerous bacterial sequences. OTUs belonging to Angiospermae suggest the types of flowering plants pollinators visited, including *Caryophyllales* sp., *Cichorieae* sp., *Geraniaceae* sp. and Lamiids (a large clade of flowering plants that includes many species present in meadows). In addition, widely observed 'unknown' OTUs to which MEGAN could not confidently assign an identity may be members of taxa that were poorly represented in GenBank, or they may be chimeras or sequencing errors that escaped filtering. Yet, most secondary OTUs are plausible as true associates of the target specimens and the wider pollinator community. Thus, associated DNA can be used to detect local community composition and ecological associations, including parasites, symbionts and diet of the target (Lucas *et al.* 2018).

Cross contamination in the traps may also explain the large number of secondary OTUs assigned to Apiformes (beyond the *Numts*). The potential for DNA mixing was further increased as specimens from the same pan trap were stored together prior to morphological identification and DNA extraction. However, we find that the greatest rate of contamination may have been within a single plate, i.e. between samples with the same primer index but different library indices, which could be either due to physical mixing in the laboratory, tag-jumping (in the Nextera XT indices, not the PCR tags), or errors in index sequencing. Trap-level contamination may add to the problem, as the combined model (plate x trap) shows only marginally higher levels of contamination (Supplementary Table S1). Because the contamination within the wells was much lower, we conclude that the primary PCR using 13 different primer tags before being combined in a single Nextera XT library was not greatly affected by these problems, indicating that our approach of using

the same unique primer tags on forward and reverse strands can largely eliminate the problem of misassignment of PCR fragments. In addition, some types of contamination were less likely to be introduced during molecular lab processing, given the precautions with specimen handling and the strict protocols of the sequencing facility, in particular regarding the widely found human DNA, present in virtually every one of the specimens. As scientists using morphological and molecular methods work together, greater awareness of these issues is needed and the steps to avoid DNA contamination should be understood and implemented, such as the use of clean pans, bee nets and storage bottles, and use of latex gloves for specimen handling during morphological identification.

Conclusions

High-throughput sequencing can greatly change the approach to monitoring of pollinators, through mass identification of sequence reads against reference databases verified by taxonomic specialists (Tang *et al.* 2015; Ji *et al*. 2013). We first established the power of the *cox1* marker for species discrimination, which only left about 5% of UK species without an unequivocal identification at species level. The subsequent utilization of the database for UK bees monitoring shows high consistency with morphological identifications conducted in parallel, and accounting for the observed morphological identification errors the correct molecular identification rates exceeds 97%. However, the deep sequencing of single specimens also revealed the various pitfalls of metabarcoding. We detected surprisingly high levels of apparent mixing with other specimens from the same and other traps. In addition, we found numerous OTUs apparently contributed by *Numts*, which greatly inflate estimates of the total species diversity; they can be filtered out efficiently as their distribution 'trails' the actual mitochondrial copies, which should be a routine part of the read filtering procedure. Lastly, the widely used OTU clustering may not produce the most accurate species detection, as shown by a comparison of OTU analyses against the most abundant read in each sample (after adequate taxonomic and numerical filtering), which revealed a full identification

of the target specimen in approximately 25% more samples. Yet, if applied under stringent quality filtering, it is possible to use HTS data at the read level, i.e. to establish genotypic variation or for assignment to particular subgroups within the Linnaean species, and thus use them in the same way as data from Sanger sequencing, but scaled up by orders of magnitude. The method thus greatly increases the accuracy and speed of taxonomic identification in pollinator monitoring, at reduced cost, while also providing further information on species interactions and ecosystem composition through the secondary OTUs. The bioinformatics methodology and comprehensive barcode database can now be rolled out for the study of much larger number of specimens typically obtained by passive pan traps and can be extended to studies of pollinators in other parts of the world.

ACKNOWLEDGEMENTS

This work was funded by the UK Department of Environment, Forestry and Rural Affairs (Defra) of the UK (contract PH0521), with in-kind contributions from the NHMUK, and a fellowship of the NERC Science Solutions for a Changing Planet Doctoral Training Programme at Imperial College (to HN). The bioinformatics pipelines were developed under the iBioGen project funded by the European Commission. The NPPMF pilot pan trapping study was jointly funded by Defra and Scottish Government under project WC1101. We acknowledge the Borough of Lewisham (Blackheath), Bristol City Council (The Downs, Troopers Hill), Conservators of Wimbledon and Putney Commons, Land Trust (Greenwich Peninsula Ecology Park), National Trust (Bookham Common, Leigh Woods), Natural England (Hartslock SSSI), and Royal Borough of Greenwich (Blackheath) for permission for DGN to collect bees. Jackie Mackenzie-Dodds (NHMUK) and NPPMF staff are thanked for making the NPPMF collection available. Martin Harvey, Stuart Roberts and Ivan Wright conducted the morphological identifications of bees sampled in the NPPMF pilot study. We thank three anonymous reviewers for their comments on the manuscript.

Data accessibility

Sequence data available at BOLD under the BEEEE label. Perl scripts used for the sequence clustering and barcode selection are available at https://github.com/tjcreedy/NAPtime. Additional methods are detailed in the supplementary materials.

Author contributions

CQT, HN and APV designed the study; CQT and HN generated molecular data; TJC, CQT, KQC and HN performed data analysis. TJC developed bioinformatics tools. CC, KQC and RO collected specimens and co-ordinated morphological identifications. DGN collected specimens, identified, documented, sampled them, preserved morphological vouchers, and verified identification. PA and CA designed analytical pipelines and provided advice on project design and analysis. CQT, HN, TJC, KQC and APV wrote an initial draft of the manuscript. All authors contributed to the writing of the final draft.

Competing interests

CQT is Senior Scientist and APV is on the Science Advisory Board of NatureMetrics, a company offering commercial services in DNA-based biomonitoring.

REFERENCES

Amiet F, Herrmann M, Müller A, Neumeyer R (2001) Apidae 3 - *Halictus, Lasioglossum*. *Fauna Helvetica* **6**, 1–208.

Amiet F, Herrmann M, Müller A, Neumeyer R (2004) Apidae 4 - *Anthidium, Chelostoma, Coelioxys, Dioxys, Heriades, Lithurgus, Megachile, Osmia, Stelis*. *Fauna Helvetica* **9**, 1–273.

Amiet F, Herrmann M, Müller A, Neumeyer R (2010) Apidae 6 - *Andrena, Melitturga, Panurginus, Panurgus*. *Fauna Helvetica* **26**, 1–317.

Amiet F, Herrmann M, Müller A, Neumeyer R (2007) Apidae 5 - *Ammobates, Ammobatoides, Anthophora, Biastes, Ceratina, Dasypoda, Epeoloides, Epeolus, Eucera, Macropis, Melecta, Melitta, Nomada, Pasites, Tet*. *Fauna Helvetica* **20**, 1–356.

Amiet F, Müller A, Neumeyer R (2014) Apidae 2 - *Colletes, Dufourea, Hylaeus, Nomia, Nomioides, Rhophitoides, Rophites, Sphecodes, Systropha*. *Fauna Helvetica* **4**, 1-239.

Andujar C, Arribas P, Gray C*, et al.* (2018) Metabarcoding of freshwater invertebrates to detect the effects of a pesticide spill. *Molecular Ecology* **27**, 146-166.

Arribas P, Andujar C, Hopkins K, Shepherd M, Vogler AP (2016) Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution* **7**, 1071-1081.

Bergsten J, Bilton DT, Fujisawa T*, et al.* (2012) The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology* **61**, 851-869.

Biesmeijer JC, Roberts SPM, Reemer M*, et al.* (2006) Parallel declines in pollinators and insect-pollinated plants in Britain and the Netherlands. *Science* **313**, 351-354.

Bogusch P, Straka J (2012) Review and identification of the cuckoo bees of central Europe (Hymenoptera: Halictidae: Sphecodes). *Zootaxa*, 1-41.

Carvell C, Isaac NJB., Jitlal M, *et al.* (2016) Design and Testing of a National Pollinator and Pollination Monitoring Framework. Final summary report to the Department for Environment, Food and Rural Affairs (Defra), Scottish Government and Welsh Government: Project WC1101.

Cross I, Notton DG (2017) Small-headed Resin Bee, Heriades rubicola, new to Britain (Hymenoptera: Megachilidae). *British Journal of Entomology and Natural History* **30**, 1-6.

DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society London Series B* **360**, 1905-1916.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**, Art. 214.

Drinkwater, E. , Robinson, E. J. and Hart, A. G. (2019) Keeping invertebrate research ethical in a landscape of shifting public opinion. *Methods in Ecology and Evolution*. Accepted Author Manuscript

Edgar R (2010) *USEARCH fastq_filter, available online at https://www.drive5.com/usearch/*.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194-2200.

Falk SJ, Lewington R (2015) *Field guide to the bees of Great Britain and Ireland* Bloomsbury Publishing PLC.

Fujisawa T, Barraclough TG (2013) Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent approach: A revised method and evaluation on simulated data sets. *Systematic Biology* **62**, 707-724.

Garibaldi LA, Steffan-Dewenter I, Winfree R*, et al.* (2013) Wild pollinators enhance fruit set of crops regardless of Honey Bee abundance. *Science* **339**, 1608-1611.

Gezon, Z. J., Wyman, E. S., Ascher, J. S., Inouye, D. W. and Irwin, R. E. (2015) The effect of repeated, lethal sampling on wild bee abundance and diversity. *Methods in Ecology and Evolution* **6**, 1044-1054

Gill RJ, Baldock KCR, Brown MJF*, et al.* (2016) Protecting an ecosystem service: Approaches to understanding and mitigating threats to wild insect pollinators. In: *Ecosystem Services: From Biodiversity to Society, Pt 2* (eds. Woodward G, Bohan DA), pp. 135-206.

Hallmann CA, Sorg M, Jongejans E*, et al.* (2017) More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *Plos One* **12**.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B* **270**, 313-321.

Huson DH, Beier S, Flade I*, et al.* (2016) MEGAN Community Edition - Interactive exploration and analysis of large-scale microbiome sequencing data. *Plos Computational Biology* **12**.

Isaac NJB, Pocock MJO (2015) Bias and information in biological records. *Biological Journal of the Linnean Society* **115**, 522-531.

Ji Y, Ashton L, Pedley JM, *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via

metabarcoding. *Ecology Letters* **16**, 1245-1257

Katoh K, Asimenos G, Toh H (2009) Multiple alignment of DNA sequences with MAFFT. In: *Methods in Molecular Biology* (ed. D P), pp. 39-64. Humana Press.

Kennedy CM, Lonsdorf E, Neel MC*, et al.* (2013) A global quantitative synthesis of local and landscape effects on wild bee pollinators in agroecosystems. *Ecology Letters* **16**, 584-599.

Kuhlmann M, Else GR, Dawson A, Quicke DLJ (2007) Molecular, biogeographical and phenological evidence for the existence of three western European sibling species in the *Colletes succinctus* group. *Organisms Diversity and Evolution*, 7(2): 155-165.

Kuhlmann M (2007) Revision of the bees of the Colletes fasciatus-group in southern Africa (Hymenoptera: Colletidae). *African Invertebrates* **48**, 121-165.

Lebuhn G, Droege S, Connor EF*, et al.* (2013) Detecting insect pollinator declines on regional and global scales. *Conservation Biology* **27**, 113-120.

Lever JJ, van Nes EH, Scheffer M, Bascompte J (2014) The sudden collapse of pollinator communities. *Ecology Letters* **17**, 350-359.

Lucas A, Bodger O, Brosi BL, *et al.* (2018) Floral resource partitioning by individuals within generalised hoverfly pollination networks revealed by DNA metabarcoding. *Scientific Reports* **8**, Art. 5133

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* **17**, 10-12.

Meier R, Wong W, Srivathsan A, Foo M (2015) $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. *Cladistics*, n/a-n/a.

Notton DG, Cuong Quoc T, Day AR (2016) Viper's Bugloss Mason Bee, *Hoplitis* (*Hoplitis*) *adunca*, new to Britain (Hymenoptera, Megachilidae, Megachilinae, Osmiini). *British Journal of Entomology and Natural History* **29**, 134-143.

Oksanen J, Blanchet G, Friendly M*, et al.* (2018) *vegan: Community Ecology Package. R package version 2.5-1* available at http://cran.r-project.org/.

Pons J, Vogler AP (2005) Complex Pattern of Coalescence and Fast Evolution of a Mitochondrial rRNA Pseudogene in a Recent Radiation of Tiger Beetles. *Molecular Biology and Evolution* **22**(4), 991–1000

Potts SG, Roberts SPM, Dean R*, et al.* (2010) Declines of managed honey bees and beekeepers in Europe. *Journal of Apicultural Research* **49**, 15-22.

Powney GD, Carvell C, Edwards M, *et al.* (2019) Widespread losses of pollinating insects in Britain. *Nature Communications* **10**, Art. 1018

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. *Plos One* **8**.

Ricketts TH, Regetz J, Steffan-Dewenter I*, et al.* (2008) Landscape effects on crop pollination services: are there general patterns? *Ecology Letters* **11**, 499-515.

Schmidt S, Schmid-Egger C, Moriniere J, Haszprunar G, Hebert PDN (2015) DNA barcoding largely supports 250 years of classical taxonomy: identifications for Central European bees (Hymenoptera, Apoidea partim). *Molecular Ecology Resources* **15**, 985-1000.

Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources* **15**, 1289-1303.

Shokralla S, Porter TM, Gibson JF*, et al.* (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports* **5**.

Smith MA, Rodriguez JJ, Whitfield JB*, et al.* (2008) Extreme diversity of tropical parasitoid wasps exposed by iterative integration of natural history, DNA barcoding, morphology, and collections. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 12359-12364.

Tang M, Hardman CJ, Ji Y, *et al.* (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution* **6**(9), 1034-1043

Team RC (2018) *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/*.

Vanbergen AJ, Baude M, Biesmeijer JC*, et al.* (2013) Threats to an ecosystem service: pressures on pollinators. *Frontiers in Ecology and the Environment* **11**, 251-259.

Westphal C, Bommarco R, Carre G*, et al.* (2008) Measuring bee diversity in different European habitats and biogeographical regions. *Ecological Monographs* **78**, 653-671.

Yoccoz NG, Brathen KA, Gielly L*, et al.* (2012) DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology* **21**, 3647-3655.

Zhang JJ, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614-620.

A				Most frequent OTU	Most frequent read	NAPselect
	Of 762 total specimens:		Specimens with sequences	762	762	749
			Specimens with sequences matching reference dataset	559	584	734
	Of 761 specimens with species-level morphological identifications and 1 with genus-level identification:		Sequences with species-level molecular identification	556 (99.5%)	584 (100%)	732 (99.7%)
			Sequences with only genus-level molecular identification	3 (0.5%)	0 (0%)	2 (0.3%)

B	Morphological ID level	Molecular ID level		Most frequent OTU	Most frequent read	NAPselect sequence
	Species	Species	Total comparisons	555	583	731
			Species-level correct	471 (84.9%)	506 (86.8%)	611 (83.6%)
			Genus-level correct	528 (95.1%)	565 (96.9%)	707 (96.7%)
	Species	Genus	Total comparisons	3	0	2
			Genus-level correct	3 (100%)		2 (100%)
	Genus	Species	Total comparisons	1	1	1
			Genus-level correct	1 (100%)	1 (100%)	1 (100%)

Table 1. The recovery success of different methods of barcode selection and the rate of accurate identification of barcodes against the BEEE reference set. Table 1A. The number of sequences obtained, the number of matches to a sequence in the reference collection (a metric of accuracy of the HT barcode selection method) and proportion of those that produce a species or genus level identification (a metric of sufficiency of the reference set), respectively. Table 1B. The accuracy of identification, relative to the morphological identification of the specimen, at different levels of morphological or molecular identification. Note that the NAPselect method returned the highest absolute number of correct identifications.

Figure 1

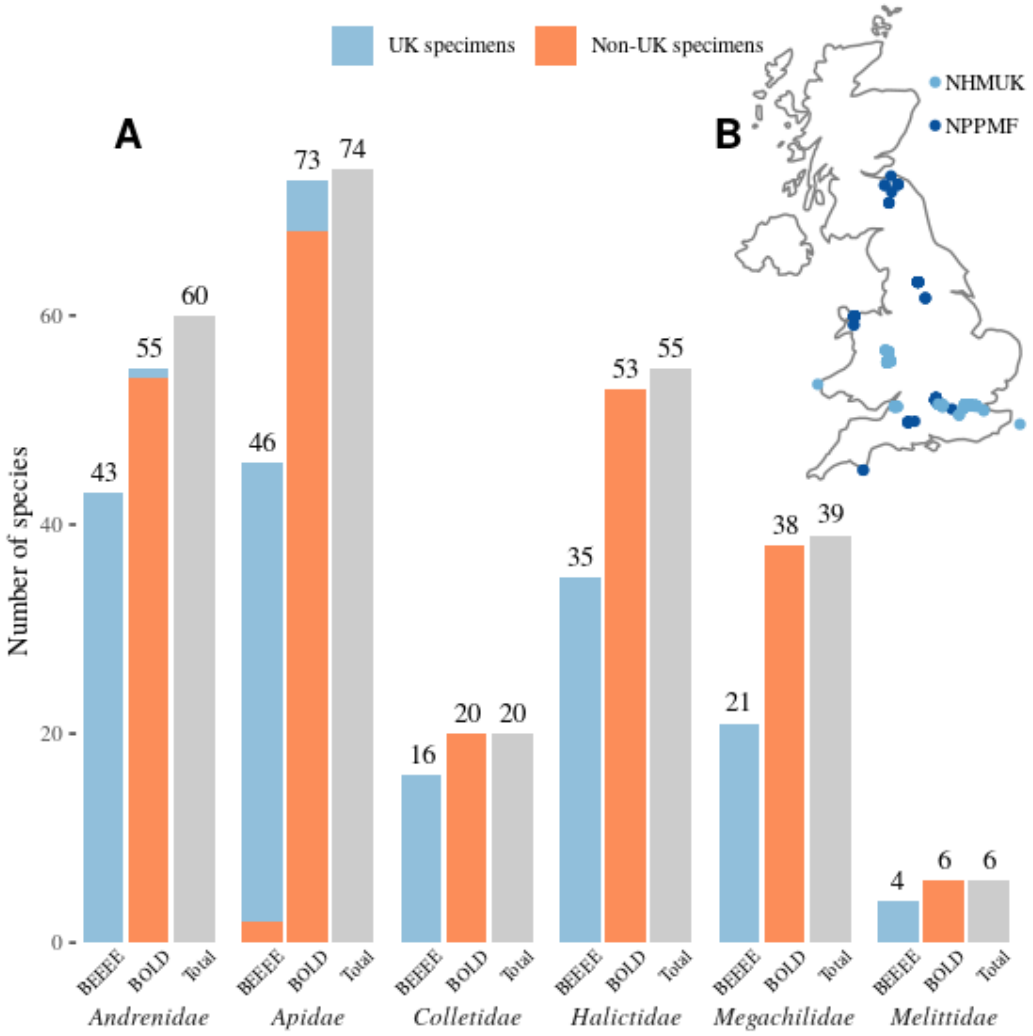

Figure 1. Specimens and species used in this study. A. The number of bee species of each family, dataset and geographical source from which sequences were compiled to form the reference collection, Column colours denote whether species from each dataset comprised any UK specimens, and numbers above bars give totals, The BEEEE columns denote the species sequenced as part of this study (165), which were compiled with existing BOLD sequences (245 species) to form the total number of species represented per family. This dataset comprises 255 of the 278 bee species in the UK. B. Sampling localities of bee specimens collected by NHMUK and the NPPMF that formed the BEEEE reference set of specimens.

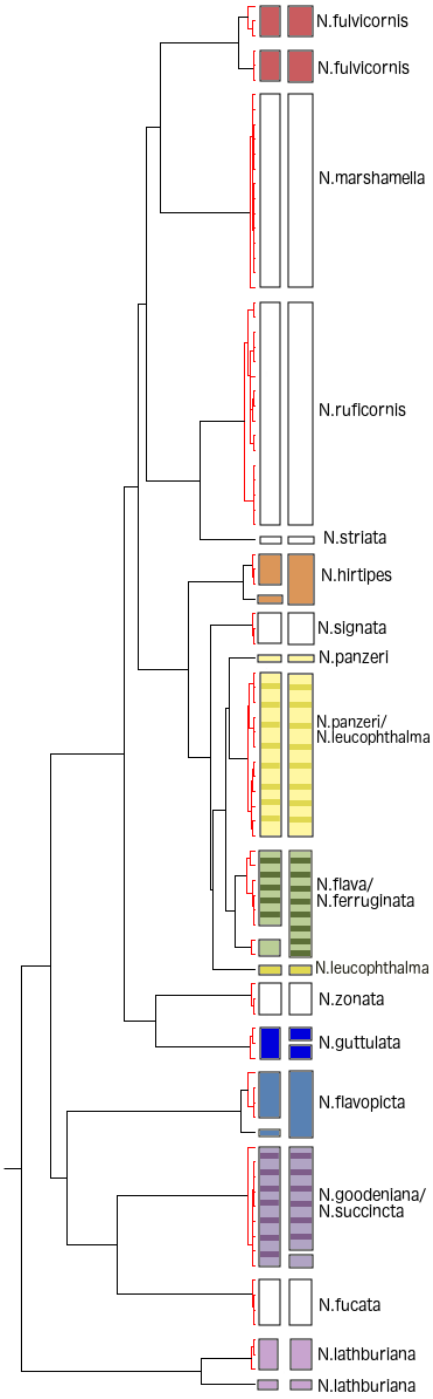

Figure 2. GMYC and BOLD analysis of a subset of the genus *Nomada*. The first column of boxes demonstrates the GMYC species, and the second column of boxes the BOLD bins. Boxes with no fill show species which are not split or lumped with other species in both the GMYC and BOLD analysis. Each colour represents a different species which is either split, lumped or both in either the GMYC or BOLD analysis, or in both. The species names are shown on the tree.

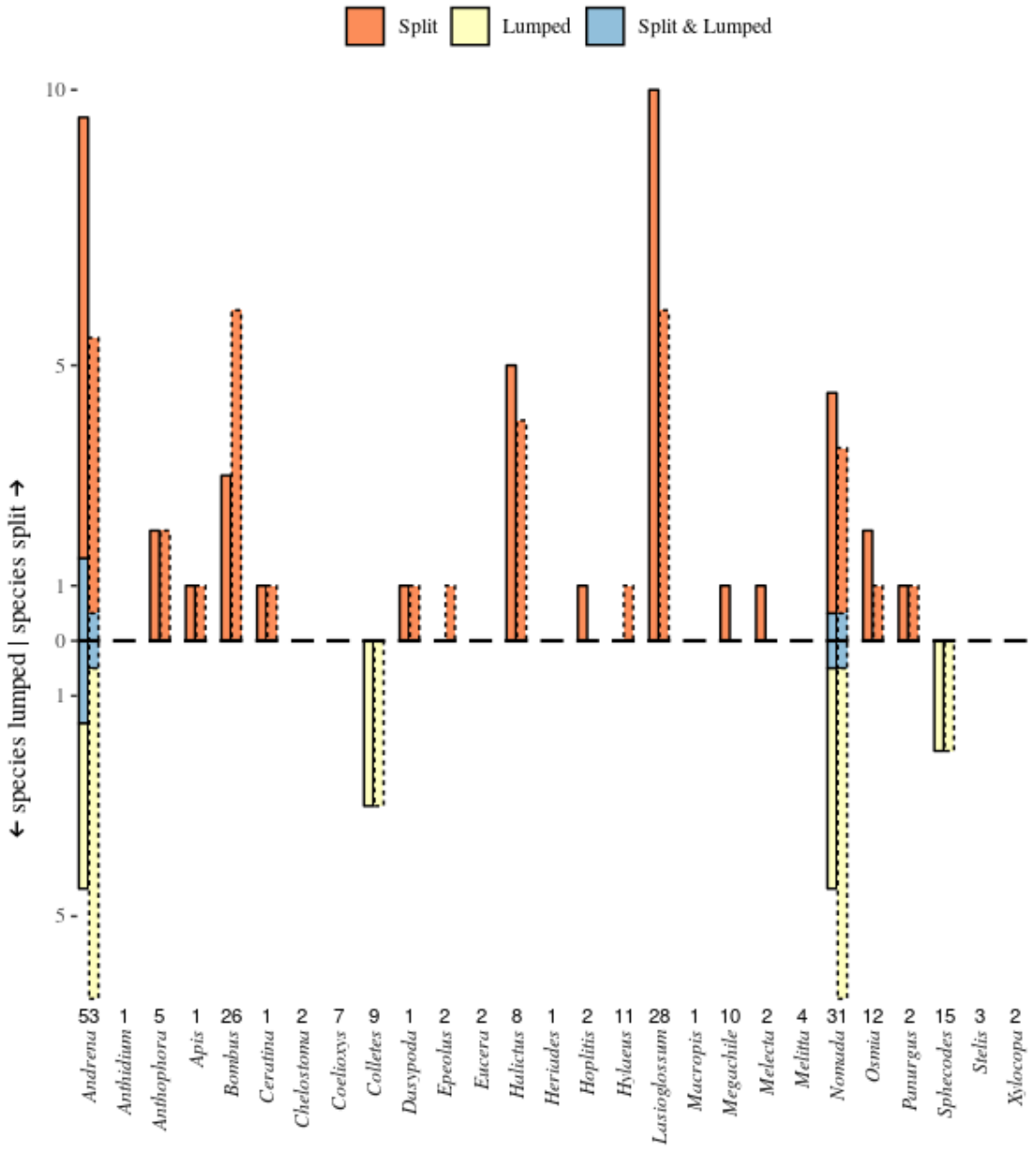

Figure 3. Congruence of species delimitation with assignment to Linnaean species, comparing the Generalised Mixed Yule Coalescent model (GMYC) (solid lines) and BOLD BIN assignments (stipled lines). Each genus is assessed separately. The number of incongruent clusters are shown, either splitting the morphospecies (orange), lump the morphospecies (yellow), or both split the morphospecies and lump those sequences with other morphospecies (blue). The total number of species in each genus is given above the genus name. Note that for many genera the morphospecies assignments were perfectly congruent with either DNA-based methods (no bars).

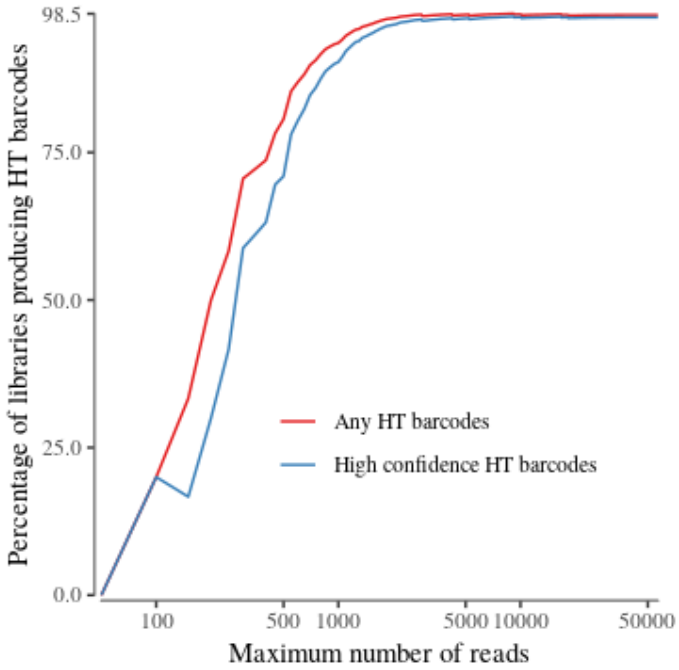

Figure 4: The percentage of amplicon pools of a given number of reads or fewer producing HT barcodes using NAPselect across the HTS dataset. Red line shows the value for any (high confidence or low confidence) HT barcode, while blue shows only high confidence HT barcodes.

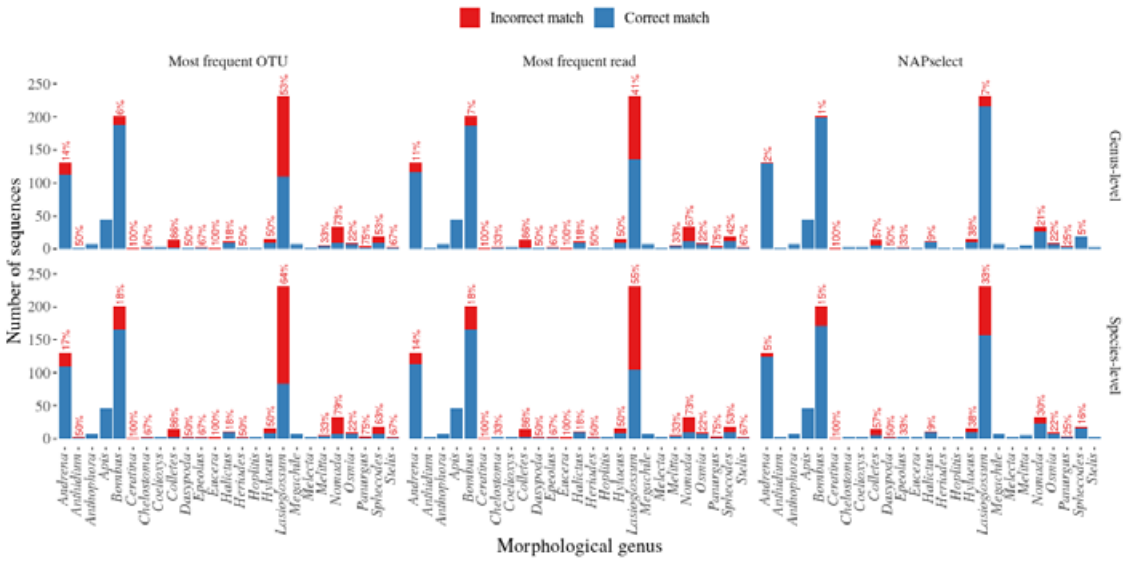

Figure 5: The proportion of molecular identification failure for different morphological species across genera. For each morphological species, we calculated the proportion of specimens for which the designated barcode failed to be correctly identified using the reference database. These values are presented here, grouped by genus and the three different barcode designation methods.

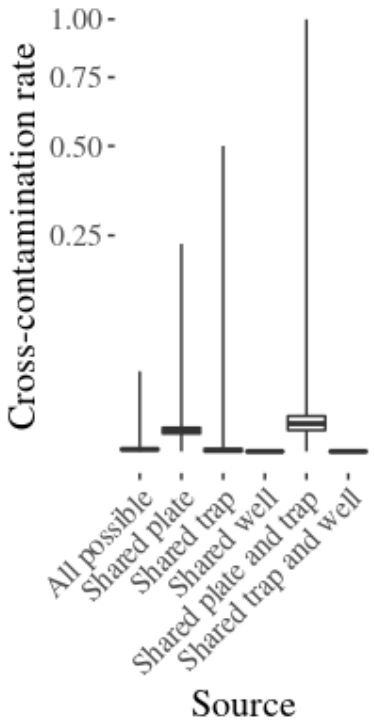

Figure 6: Box and whisker plot showing the mean and 95% confidence range of recovery rates possible cross-contamination OTUs from different sources of cross contamination. X axis shows the different sources of cross contamination, and y axis shows the proportion of possible of possible cross-contamination OTUs recovered from that source. The rate of shared OTU recovery is significantly higher when considering samples from the same plate and same plate and trap compared with a background rate of cross contamination (all possible).