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

Thomas J. Creedy^{1,2*}, Hannah Norman^{1,3*}, Cuong Q. Tang^{1,4*}, Kai Qing Chin¹, Carmelo Andujar^{1,2,5}, Paula Arribas^{1,2,5}, Rory O'Connor^{6,7}, Claire Carvell⁸, David G. Notton¹, Alfried P. Vogler^{1,2}

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

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

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

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

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

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

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

⁸NERC 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

22 **ABSTRACT (250 max.)**

23 Improved taxonomic methods are needed to quantify declining populations of insect
 24 pollinators. This study devises a high-throughput DNA barcoding protocol for a regional fauna
 25 (United Kingdom) of bees (Apiformes), consisting of reference library construction, a proof-of-
 26 concept monitoring scheme, and the deep barcoding of individuals to assess potential artefacts and
 27 organismal associations. A reference database of Cytochrome Oxidase subunit 1 (*cox1*) sequences
 28 including 92.4% of 278 bee species known from the UK showed high congruence with morphological
 29 taxon concepts, but molecular species delimitations resulted in numerous split and (fewer) lumped
 30 entities within the Linnaean species. Double tagging permitted deep Illumina sequencing of 762
 31 separate individuals of bees from a UK-wide survey. Extracting the target barcode from the amplicon
 32 mix required a new protocol employing read abundance and phylogenetic position, which revealed
 33 180 molecular entities of Apiformes identifiable to species. An additional 72 entities were ascribed to
 34 mitochondrial pseudogenes based on patterns of read abundance and phylogenetic relatedness to
 35 the reference set. Clustering of reads revealed a range of secondary Operational Taxonomic Units
 36 (OTUs) in almost all samples, resulting from traces of insect species caught in the same traps,
 37 organisms associated with the insects including a known mite parasite of bees, and the common
 38 detection of human DNA, besides evidence for low-level cross-contamination in pan traps and
 39 laboratory steps. Custom scripts were generated to conduct critical steps of the bioinformatics
 40 protocol. The resources built here will greatly aid DNA-based monitoring to inform management and
 41 conservation policies for the protection of pollinators.

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

44

45

46 INTRODUCTION

47 Widespread declines in pollinator populations are raising the alarm about the future of global
 48 biodiversity and agricultural productivity (Garibaldi *et al.* 2013; Hallmann *et al.* 2017; Lever *et al.*
 49 2014), driven by the combined effects of habitat loss, introduction of non-native and invasive
 50 species, pathogens and parasites, and various other factors contributing to environmental change
 51 (Vanbergen *et al.* 2013). Landscape effects on pollination of crops through agricultural
 52 intensification, particularly those of monoculture crops, have led to significant changes in pollinator
 53 communities (Kennedy *et al.* 2013; Ricketts *et al.* 2008), with obvious economic implications for the
 54 agricultural sector and pollination services worth hundreds of millions of pounds in the United
 55 Kingdom alone (Potts *et al.* 2010). However, these trends in species distribution and abundance are
 56 difficult to quantify, unless solid methodologies for monitoring at regional levels can be
 57 implemented. Thus there is an urgent need to develop strategies for large-scale and long-term
 58 systematic monitoring of pollinator populations, to better understand the impacts of declines on
 59 pollination services to crops and wild plants, and inform policy decisions and conservation efforts.

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

70 This study applies high throughput sequencing (HTS) techniques to assess bee diversity and
 71 abundance from mass-trapped samples, using a rapid DNA barcoding approach suitable to
 72 individually assay the thousands of specimens potentially generated in the course of a large-scale
 73 monitoring scheme. The first step in this process was the generation of a well curated reference
 74 database that links each DNA sequence to a species name, using the Cytochrome *c* Oxidase subunit I
 75 (*cox1*) barcode marker (Hebert *et al.* 2003), which provides good species discrimination in
 76 Hymenoptera (Smith *et al.* 2008). The more recent approach of ‘metabarcoding’, by which entire trap
 77 catches are subjected to amplicon sequencing in bulk, produces species incidence data based on the
 78 mixed sequence read profile (Yoccoz *et al.* 2012). Current HTS protocols can maintain the individual
 79 information of thousands of samples using unique tags in the initial PCR prior to pooling for Illumina
 80 sequencing with secondary tags, which allows sequences to be traced back to the associated
 81 specimen (Arribas *et al.* 2016; Shokralla *et al.* 2015). The great sequencing depth of this high-
 82 throughput barcoding (HT barcoding) methodology may also reveal DNA from organisms internally or
 83 externally associated with a target specimen or as a carry-over from other specimens in the trap.

84 The current study on the regional-scale pollinator fauna, focused on the bees (Hymenoptera:
 85 Apiformes) of the United Kingdom, illustrates the required steps from generating a barcode
 86 reference database for the 278 species of bees known from the UK, which was then used for the
 87 identification of samples gathered as part of a pilot study for a national monitoring scheme based on
 88 short barcode sequences obtained with HTS. Agreement between morphological and molecular
 89 identifications was assessed. In addition, the deep-sequencing approach allowed the assessment of
 90 organisms associated with the target specimens, as well as cross-contaminations from other species
 91 present in the traps or from specimen handling and laboratory procedures.

92

93

94 MATERIALS AND METHODS

95 Building a regional reference database

96 A *cox1* reference database was generated from DNA barcoding of specimens of bee species
 97 known to occur in the UK according to the list of Falk and Lewington (2015) and notes from various
 98 sources maintained by co-author DGN. Most specimens were caught by hand netting and identified
 99 by DGN, using the latest keys available at the time (Amiet *et al.* 2001, 2004, 2010; Amiet *et al.* 2007;
 100 Amiet *et al.* 2014; Bogusch & Straka 2012; Falk & Lewington 2015) (Benton 2006; Mueller 2016).
 101 Identifications had to draw on these various references because the comprehensive key of Falk &
 102 Lewington (2015) became available only part way through the study, while some identifications were
 103 also cross-checked between different publications. Specimen data for morphological vouchers are
 104 available at the Natural History Museum Data Portal (data.nhm.ac.uk)
 105 <http://dx.doi.org/10.5519/0002965>. Sequences are available at BOLD (Barcode-of-Life Datasystem)
 106 under the BEEEE project label. Additional specimens were obtained using pan traps from the survey
 107 described below. The reference set included all available unique UK species as determined by
 108 morphology, with multiple specimens per species where available. These within-species replicates
 109 allowed inclusion of specimens from across the geographical range of widely-distributed species,
 110 identified by different taxonomists and/or belonging to known species complexes.

111 DNA was extracted from a single hind leg using a Qiagen DNeasy Blood and Tissue Kit, after the
 112 specimens were incubated at 56°C in the extraction buffer (ATL and Proteinase K) overnight in a
 113 shaking incubator at 75 rpm. The complete 'barcode region' (658 bp) of *cox1* was amplified using
 114 newly designed primers (BEEf TWYTCWACWAAYCATAAAGATATTGG and BEEr
 115 TAWACTTCWGGRTGWCCAAAAAATCA), based on an alignment of 84 mitochondrial genomes from 22
 116 genera. PCR and sequencing using ABI dye terminator sequencing followed standard procedures

117 (Supplementary Material). The sequences were added to BOLD in the project BEEEE, along with
118 Syrphidae barcodes that were sequenced at the same time.

119 Sequences were aligned using the *MAFFT* v1.3. (Katoh *et al.* 2009) plugin in *Geneious*.
120 Alignments were used for molecular distance-based and coalescence-based species delimitation. (1)
121 BOLD BINs (Barcode Identification Numbers) were automatically generated for sequences that are
122 uploaded onto its database (Ratnasingham & Hebert 2013). These BINs are formed using a refined
123 single linkage network, which combines sequence similarity metrics and graph theory. (2) The GMYC
124 (Generalized Mixed Yule Coalescent) method determines species limits based on a shift in the rate of
125 branching along the root-to-tip axis of the phylogenetic tree, separating the speciation (Yule)
126 processes from fast branching rate expected within population under a neutral coalescent (Fujisawa
127 & Barraclough 2013). The analysis was done on phylogenetic trees constructed separately for each
128 genus using BEAST 1.8.1 (Drummond & Rambaut 2007), which was used as the input for the GMYC
129 analysis (Tang *et al.* 2014). The GMYC was applied to genera with only a single British representative
130 (*Apis*, *Anthidium*, *Ceratina*, *Dasypoda*, *Macropis* and *Rophites*) by adding supplementary sequences
131 from BOLD.

132 New barcode sequences for species that were already in BOLD were assessed to examine
133 whether these new sequences represented new *cox 1* haplotypes for these species. The new
134 barcodes were searched against a set of 1754 sequences downloaded from BOLD for species known
135 to exist in the UK using BLASTn on default parameters. Any sequences not 100% identical over the
136 entire length of the query or subject were designated as novel haplotypes.

137

138 **Generating a test dataset from field caught samples using HTS**

139 The reference database was used for identification of specimens obtained through the
140 National Pollinator and Pollination Monitoring Framework (NPPMF) (Carvell *et al.*, 2016). Mixed

141 samples were collected with pan traps consisting of sets of water-filled bowls (painted UV-yellow,
142 white and blue, after Westphal *et al.* 2008) from 14 sites across the UK, and further specimens were
143 collected by netting along standardised transects running 200m from each set of pan traps (Figure
144 1A, Table S1; see Carvell *et al.* 2016 and supplementary materials for a full description of the
145 sampling protocol). Bees (Apiformes) were separated from other taxa in the field, stored in 99%
146 ethanol to preserve DNA for analysis, and transferred to -20°C as soon as possible after collection.
147 Specimens were identified morphologically by expert taxonomists offering commercial identification
148 services. In total, 762 bee specimens were processed (480 bees were extracted from the pan traps,
149 and 282 specimens from the transects and further hand collecting). All specimens were stored in 99%
150 ethanol and deposited as voucher specimens in the Molecular Collection Facility at the NHMUK.

151 DNA was extracted from individual specimens by piercing the abdomen and submerging the
152 whole specimen in lysis solution consisting 180ul ATL buffer and 20ul Proteinase K for 12 hours on a
153 56°C shaking incubator. DNA extractions were performed using either the Qiagen BioSprint 96 DNA
154 Blood Kit or DNeasy Blood and Tissue kits applied to the lysate. Each DNA extract was PCR amplified
155 for a 418bp portion of the *cox1* barcode region (Andujar *et al.* 2018). Each individual amplicon was
156 tagged using a 'double dual' PCR protocol (Shokralla *et al.* 2015) to generate unique tag
157 combinations for each bee specimen, following the procedures of Arribas *et al.* (2016). Tags were
158 added in the initial PCR by amplification using tagged *cox1* primers employing different 6 bp
159 sequence combinations designed with a Hamming distance of 3, with a total of 13 different tagged
160 primer sets. In all reactions, forward and reverse primers used the same tag, so that the products of
161 tag jumping could be removed (Schnell *et al.* 2015). Amplicons generated with different primer tags
162 were pooled, and each pool was cleaned using Agencourt AMPure XP beads (Beckman Coulter,
163 Wycombe, UK), prior to secondary amplification of each pool with the i5 and i7 Nextera XT indices
164 with 96 unique MID combinations (Illumina, CA, USA) and sequencing on Illumina MiSeq v.3 (2x300
165 bp paired-end).

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

179

180 **Testing the utility of the reference dataset**

181 Three methods were used to designate a single putative “high-throughput barcode” (HT
 182 barcode) sequence representing the *cox1* gene of each specimen from the set of reads. Firstly, we
 183 employed a standard metabarcoding pipeline, implemented in the NAPcluster script, to generate
 184 OTU (Operational Taxonomic Units) clusters and centroid sequences using *USEARCH*. The script
 185 includes functions from the *USEARCH* suite (Edgar 2010), starting with the data output from
 186 NAPmerge (merged and quality-filtered amplicons), and comprises the following steps: (i) filtering
 187 sequences by length; (ii) dereplication and filtering by number of reads per unique sequence, to
 188 retain only sequences represented by a set minimum of reads; (iii) denoising using the *UNOISE*
 189 algorithm (Edgar *et al.* 2011); (iv) clustering of sequences according to cluster radius and generation
 190 of an output set of OTU consensus sequences, and (v) mapping of reads to OTU clusters (using

191 *USEARCH usearch_global* and a custom .uc parser) and generation of an output table of OTU read
 192 numbers by sample. All sequences differing from 418 bp and with only 1 copy were removed in steps
 193 (i) and (ii), and *USEARCH cluster_otus* was employed for clustering with a dissimilarity threshold of
 194 3%. The centroid of the most abundant OTU was used as the specimen barcode.

195 As sequence variants may drive up the read count of an OTU and thus obscure the haplotype
 196 of the target specimen, the second method for selecting the HT barcode sequence simply chooses
 197 the most frequent read for each library, under the assumption that error-free reads represent the
 198 most abundant template DNA and thus the target specimen. Other sequences that represent nuclear
 199 mitochondrial pseudogenes (NUMTs), gut contents, internalised parasites, and cross-contamination,
 200 or result from PCR or sequencing errors, are expected be present in lower numbers. The extraction of
 201 the most frequent read was performed using a custom perl script.

202 The third method employed a purpose-built tool, NAPselect, that finds the most highly
 203 represented sequence among reads in an amplicon mix, as above, but then statistically and
 204 taxonomically validates this selection. The process starts with steps similar to metabarcoding: firstly,
 205 filter the batch of sequence reads by length (in this case, rejecting any sequence not 418 bp), and
 206 group sequences by identity (i.e. dereplication) recording the abundance of reads representing each
 207 unique sequence. Starting with the most abundant, unique sequences are assessed one-by-one,
 208 using bootstrapping to validate the significance of the difference in read abundance, and using BLAST
 209 to assign the sequence taxonomy. Based on the total number of reads in the sample and the number
 210 of unique sequences, the probability of a sequence occurring as frequently as the most abundant
 211 sequence by chance alone is determined using 10,000 bootstrap iterations. A p-value of 0 designates
 212 a sequence as significantly more frequent with high confidence, and less than 0.5 for low confidence,
 213 above which the entire sample is disregarded because a putative barcode sequence for the target
 214 specimen is not clearly defined. The most abundant sequence is then subjected to a BLASTn search

215 against a local copy of the NCBI *nt* database and the hits assessed for presence in the focal taxon (in
216 this case, Hymenoptera). Only those sequences passing both these tests are selected.

217 The success of these three methods, and the accuracy of the sequences they output, was
218 tested by identifying the HT barcode sequences using the BEEEE reference collection of sequences
219 obtained in section 1. Each of the three putative HT barcodes for each specimen were searched for
220 matches in the BEEEE reference collection using BLASTn with default parameters. Only matches with
221 >95% identity and overlap with the reference sequences of >400 bp were retained, and the match
222 with the similarity was selected, using bitscore to break ties. The identity of this hit was compared
223 against the known morphological ID for that specimen at the genus and species level. For each HT
224 barcode selection method and taxonomic level, the number of correct molecular identifications was
225 tallied and a proportion of failure calculated.

226

227 **Exploration of concomitant DNA in the testing dataset**

228 The OTUs generated with the NAPcluster script (see above) allowed the exploration of co-
229 amplified DNA from each bee specimen other than the primary *cox1* sequence. For each sample, the
230 OTUs that did not match the NAPselect HT barcode sequence for the target specimen were
231 designated as “secondary OTUs”. We used the OTUs for this analysis, rather than the reads, to
232 reduce unnecessary complexity in the dataset. These OTUs were searched against a local copy of the
233 NCBI *nt* database using BLASTn, followed by taxonomic binning using MEGAN6 Community Edition
234 with the weighted Lowest Common Ancestor algorithm (Huson *et al.* 2016). Any OTUs assigned to
235 Apiformes were additionally identified using BLASTn against the BEEEE reference collection and
236 the NAPselect HT barcodes (above, section 3). In both cases, BLASTn employed default parameters,
237 and sequences were identified as the hit with the highest identity where identity was >95% and
238 overlap was >400 bp, with bitscore breaking ties.

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 pseudogenes if they were completely coincident across samples with another closely related OTU that did match a BEEEE reference or specimen barcode, 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 (Team 2018) packages *plyr* and *reshape2*. 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 (section 2) from *another* sample were used in this analysis. Three sources of cross-contamination were considered: contamination from other individuals in the same trap, DNA mixing between specimens with the same PCR tag on a single plate, and DNA mixing between specimens with the same Nextera XT tag in a single well. For each source or combination of sources, the total possible selected barcodes were counted, and then the proportion of those that were present as secondary OTUs in a sample was calculated. For example, each well in the library preparation contained 13 specimens tagged with different sequence identifiers: if in a set of these 13 each is a different species (different HT barcodes), there are 12 possible well contaminants for any one of these samples, and so a sample containing 3 secondary OTUs from these 12 specimen barcodes 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.

264 One-sample t tests were used to assess if the mean contamination rate for each source or source
265 combination was significantly greater than zero. To compare between sources against the control,
266 the effect of source on contamination rate was fitted in a quasi-binomial ANOVA, setting the control
267 as the reference level.

268

269 RESULTS

270 A reference database of UK bees

271

272 A total of 355 bee specimens were newly sequenced for the COI barcode to generate the
273 reference set, representing 165 Linnaean species. These new sequences were added to 1754 full-
274 length barcode sequences downloaded from the Barcode of Life Database (BOLD) for a total of 2109
275 sequences. Comparing these datasets (Fig. 1A) the BOLD data represented 245 of the 278 UK bee
276 species, but comprised only 14 sequences (6 species) from specimens collected in the UK. The 355
277 new sequences add 10 UK species (15 sequences) not represented in the BOLD dataset, and novel
278 haplotypes for 107 further species (201 sequences). The final reference set included 255 bee species
279 (92.4% of 278 species known from the UK). The missing species are either extinct (6 species), rarely
280 introduced by accident (1 species, *Heriades rubicola*), only found in the Channel Islands (1 species,
281 *Andrena agillissima*), listed as endangered (RDB3-RDB1) (8 species), or rare and localised (5 species),
282 while 2 species were only recently added (Cross & Notton 2017; Notton *et al.* 2016). When
283 considering each of the six families separately, the greatest number of species missing from the
284 database was in Andrenidae (9 of 69 species), followed by Apidae (4 of 76) and Halictidae (4 of 62).

285 Genetic variation within morphologically identified Linnaean species ranged from 0% to 5.9%
286 (mean 0.31%, standard error $\pm 0.04\%$), and interspecific variation ranged from 0% to 24.9% (mean
287 6.7% $\pm 0.08\%$). We found that 242 (94.9%) of *cox1*-based sequence clusters at 97% similarity were an
288 exact match of the Linnaean species identifications (Supplementary Figure S1). Inconsistency with
289 the morphological species definitions were limited to five genera, *Andrena*, *Bombus*, *Colletes*,
290 *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. hederæ* - *C. succinctus*) (Kuhlmann et al. 2007), suspected geographically confined species among the *Dasypoda hirtipes* group (Schmidt et al. 2015), and variation among *Lasioglossum rufitarse*, *Nomada flava* - *N. leucophthalma* - *N. panzeri*, and *N. goodeniana* - *N. succincta* clusters.

Testing HTS data against the reference library

Illumina reads generated for 762 bee specimens resulted in an average of 5851 *cox1* sequences per specimen (amplicon pool) after read merging and stringent quality filtering. 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 samples that did not produce a dominant (hymenopteran) read, while the OTU clustering and most-frequent read method produced only 559 and 584 HT barcodes, respectively (Table 1A). Out of the barcodes chosen by NAPselect, 734 (99.7%) produced a match to sequences in the BEEEE reference data (Table 1A), confirming these sequences correspond to the target specimens. This

315 proportion of hits to the reference set was equally near 100% with the other two methods, but
 316 because these produced HT barcodes for fewer specimens, they resulted in approx. 25% fewer
 317 specimen identifications. Almost all identifications were to the species level, while between 0 and 3
 318 individuals produced hits to reference sequences identified only to genus (Table 1A). Across all
 319 samples, a total of 154 unique species identifications against the BEEEE reference set were obtained.

320 Congruence of molecular identifications with the morphological identifications of the source
 321 specimens was high at genus level with 95-96%, but only 83-86% of specimens were identified as the
 322 same species with both data types (Table 1B). However, as NAPselect designated a considerably
 323 larger proportion of barcodes to species level, the absolute number of correct species identifications
 324 using this method was the highest, at 611 specimens out of the 762 sequenced (707 correct at genus
 325 level). The proportion of successful molecular identification was compared between different genera
 326 to examine whether there was a taxonomic bias in identification success. The success rate of
 327 molecular identification differed among genera (Figure 4), although this tends to be correlated with
 328 the number of species/sequences in a genus. Species-rich genera that produced markedly more
 329 successful identifications include *Andrena* and *Bombus*, whereas *Colletes* showed low success even
 330 using NAPselect (as expected because some species were inseparable by DNA; see above).

331 We investigated whether the lumping and splitting observed in the reference dataset was a
 332 driver of molecular misidentification by examining the proportion of correct and incorrect matches
 333 against species that were lumped and/or split in the GMYC analysis. Of the 734 HT barcode
 334 sequences generated by NAPselect that had a BLAST match to a BEEEE reference sequence, 17 were
 335 to a species that was lumped, 178 to a species that was split and 1 to a species that was both lumped
 336 and split. The proportion of correct species and genus level matches for these sets of HT barcodes
 337 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.

Exploration of concomitant DNA in the testing dataset

When OTU clustering was carried out on the entire data set combining the 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 pseudogenes identified 72 OTUs as likely NUMTs. In addition, several OTUs were reclassified as Diptera in the phylogenetic tree used for pseudogene 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 processed alongside the bees, but are not discussed here. 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.

Quantitative comparisons of OTU distributions across samples were conducted after rarefaction, which removed 46 samples with fewer than 400 reads, losing 15 OTUs. Rarefied samples had between 1 and 26 OTUs, with a mean of 5.9 (SD = 3.9). The majority of samples had secondary

OTUs beyond the specimen barcode sequence. Secondary OTUs contributed an average of 25.7% (SD = 33%) of the samples reads; in 238 cases, secondary OTUs contributed over 50% of the reads. The taxonomic composition of secondary OTUs (Fig. 5) showed that most samples had at least one other bee OTU, sometimes as many as 8, out of the 132 total Apiformes OTUs that were recognised as secondary OTU at least once (48 OTUs of Apiformes were only recovered as the primary OTU). Beyond Hymenoptera, high OTU numbers were contributed by Diptera, with up to 10 OTUs in a single sample, and fungi (maximum 13 OTUs in one sample and a total of 40 across all samples) (Fig. 5). However, no higher taxon was found consistently across all samples apart from the bees.

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, and errors in various DNA laboratory procedures. In general, the level of direct contamination with DNA sequences that were the primary OTU 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, and 110 of these match to one of the barcode sequences from other wells. 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 at the level of library construction, i.e. from mixing among the 96 Illumina tags. The level of cross-contamination was much lower for those samples in the same well, i.e. the combination of 13 different products from the primary PCRs conducted with a different primer tag each.

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

396

397 **DISCUSSION**

398 **The reference database**

399 Cost-effective species-level identifications of bees and other insect pollinators are required to
 400 provide robust evidence for population changes and to inform land use management and
 401 conservation (Gill *et al.* 2016). We conducted this analysis in two stages, by first building the
 402 reference database using conventional sequencing technology, which was then trialled for species
 403 identification using high-throughput sequencing of samples from a proof-of-concept monitoring
 404 scheme. The combined effort of new sampling and sequencing, together with barcode data already
 405 in the BOLD database, resulted in a virtually complete set of the UK bees, with only a few rare or
 406 presumed extinct species missing. Furthermore, we expanded existing references by generating
 407 novel sequences from UK populations of widespread species. The *cox1* barcode delimited 94.9% of
 408 species in the reference database as separate entities, showing that for almost all bee species in the
 409 UK this set is sufficiently discriminatory. In the remaining cases the molecular analysis lumped the
 410 Linnaean species, as evident in the *de novo* species delimitation using the GMYC method, while an
 411 even greater proportion were shown to be split into additional GMYC groups which, however, were
 412 not incongruent with the Linnaean species.

413 The overall reference database comprises a mixture of UK and non-UK sequences, as many
 414 species are more widely distributed in Europe and North America from which many barcodes were
 415 obtained, and the species discrimination may be even clearer if performed with UK samples only, as a
 416 local subset of intra-specific variation exacerbates the species-level differences (Bergsten *et al.*
 417 2012). Importantly, the high congruence of molecular groups with the Linnaean species also shows
 418 that the mitochondrial ‘gene trees’ are a good reflection of the species-level entities, as both
 419 morphological diagnostics and mitochondrial markers corroborate the species hypotheses (DeSalle *et*
 420 *al.* 2005), and thus the use of multiple markers for species delimitation is generally not required.

421 Finally, congruence with the BOLD database also suggests that the identifications have been correct,
422 in some cases after secondary inspection of specimens.

423 The molecular data failed to separate a small number of species in four of the 27 genera
424 studied (“lumped” in Fig. 3). In some instances, such as the *Colletes succinctus* species group,
425 morphological identification of three named species is reliable, if challenging, now that there is a key
426 covering all UK species (Falk & Lewington 2015), and there are biological and distributional
427 differences. *Cox1* sequences are not sufficient to delineate these species (Kuhlmann *et al.*, 2007), and
428 morphotaxonomy remains the most reliable method for this species group. Similarly, the separation
429 of the *Nomada goodeniana-succincta* group relies on subtle colour variants (Falk & Lewington 2015)
430 and cryptic species are likely to exist. Additional genetic markers may be useful; e.g. the three
431 recognised *Colletes* species lumped in *cox1* exhibit fixed differences in EF-1a and ITS (Kuhlmann
432 2007). Vice versa, divergent *cox1* entities (splitting) may indicate the existence of hitherto
433 unrecognised species. For example, a divergent haplotype in *Dasypoda hirtipes* has now been
434 associated with a morphologically differentiated, eastern European species that is not part of the UK
435 fauna (Schmidt *et al.* 2015). We have already curated the *cox1* database extensively, in particular to
436 remove morphological identification errors (Supplementary Text), but the new clusters may lead to
437 the discovery of separate entities within the Linnaean species and may provide fertile ground for
438 future morphological work. Since DNA extraction destroyed only one leg, morphological vouchers
439 can be re-examined, an important process in refining the reference database.

440

441 **Generating high throughput barcodes**

442 The newly created *cox1* database was then used to identify species from a survey of pan
443 traps using high-throughput barcoding (“HT barcoding”). The methodology has great potential for
444 sequencing mixed samples (metabarcoding) but was here applied on individual specimens to test the

445 efficacy of this approach and our ability to confidently recover a sequence for the target specimen.
 446 We employed three methods for designating this sequence from a pool of anonymous amplicons.
 447 The most intuitive approach was to undertake a standard metabarcoding analysis using the *USEARCH*
 448 pipeline to designate the centroid sequence of the most highly represented OTU in each sample as
 449 the HT barcode sequence. However, the sequence obtained with this method did not produce a
 450 BLAST hit to the reference database in 27% of cases. An alternative method was to simply select the
 451 most frequent unique sequence in the amplicon pool, analogous to the sequence that would be
 452 generated by Sanger sequencing. However, while this method also designates a barcode for every
 453 sample, these sequences are only marginally more likely to find a match to the reference database
 454 (23% did not produce a BLAST hit).

455 The third method, implemented in the NAPselect script, also selects the top-abundant read,
 456 but requires that this read matches a specific taxonomic group (in this case, Hymenoptera), and that
 457 the read frequency is significantly greater than frequencies of other reads. If these conditions are not
 458 met, NAPselect then discards the top read and checks other reads according to descending
 459 abundance. This pipeline did not output a sequence for 13 specimens, disregarding samples with low
 460 read numbers or low differentiation among other abundant reads. However, the majority of the
 461 remaining sequences matched the reference database, and only 3.7% of specimens did not produce
 462 a sequence with a BLAST hit - a substantial improvement over the other methods (Table 1). The key
 463 improvement introduced by this script probably was that NAPselect conducts BLAST searches against
 464 GenBank and assesses the taxonomy of the hits, which was specified to allow BLAST errors. This
 465 method is clearly very effective, with error rates determined largely by sequencing depth issues
 466 rather than an inability to select the correct sequence.

467

468

469 Exploration of concomitant DNA in the testing dataset

470 Unlike standard metabarcoding conducted on mixed samples, the current analysis permits a
 471 precise determination of amplicons derived from single specimens. A surprising finding was the high
 472 proportion of reads attributable to secondary OTUs, and their taxonomic diversity. The specimens
 473 from the monitoring program were not substantially different from those used in Sanger sequencing
 474 to build the reference database (in some cases used for both purposes), which produced clean base
 475 calls consistent with a single predominant PCR product. However, the primers for Illumina
 476 sequencing were designed for broad amplification of arthropods (Arribas *et al.* 2016) and probably
 477 have a wider taxonomic amplitude than the Hymenoptera-specific primers used to amplify the
 478 standard ‘barcode’ region. Besides co-amplification of a broader range of associated species, this
 479 may also increase the potential for sequencing of pseudogenes. Out of 509 OTUs recovered from all
 480 samples combined, 263 were identified as Apiformes initially, which greatly exceeds the number of
 481 species expected in this survey. Pseudogenes diverge without the constraints of coding regions and
 482 thus can be partially eliminated based on length differences. For example, a preliminary analysis that
 483 did not constrain the read filtering to the target length of 418 bps obtained six additional OTUs
 484 assigned to humans, all of which were confidently identified as known mitochondrial pseudogenes.
 485 However, filtering the reads to a fixed length could not avoid this problem sufficiently. We therefore
 486 implemented a further filter based on the distribution of low-abundance OTUs that are co-
 487 distributed with the true mitochondrial copies. We only removed OTUs that form a clade with the
 488 presumed true copy (close matches to the reference database), under the assumption that nuclear
 489 pseudogenes are of limited evolutionary persistence before they diverge too far from the
 490 mitochondrial ancestor and no longer are captured by the PCR primer. Based on these criteria a total
 491 of 72 OTUs were identified as mitochondrial pseudogenes. This method (and the removal of several
 492 other OTUs whose incorrect assignment was revealed with the phylogeny) reduced the total number
 493 of Apiformes OTU to 180, which is closer to the 154 species identified morphologically, in particular if

OTU splitting (Fig. 4) is taken into account. The procedure for identifying these likely pseudogene OTUs is a novel step in the metabarcoding filtering process which, to our knowledge, has been implemented here for the first time. However, it is dependent on “true” OTUs being identified by a reference collection and that there exists a high level of read variation between the set of target *cox1* OTUs and their putative pseudogenes – both situations that are common in HT barcoding studies but less so in some 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 associates of pollinator communities, which suggests carry-over of DNA with the target specimens. Extraneous insect species in the sequencing mixture mostly consisted of other known pollinators attracted to flowers (and pan traps), including various Coleoptera and Diptera. Consistent with the detection of pollen beetles (*Meligethes*), numerous specimens were observed 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 consistent with internal parasites, including a species of 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

519 the target specimens and the wider pollinator community. Thus, associated DNA can be used to
520 detect local community composition and ecological associations, including parasites, symbionts and
521 diet of the target.

522 Cross contamination in the traps may also explain the large number of secondary OTUs
523 assigned to Apiformes (beyond the pseudogenes). The potential for DNA mixing was further
524 increased as specimens from the same pan trap were stored together prior to morphological
525 identification and DNA extraction. However, we find that the greatest rate of contamination may
526 have been within a single plate, i.e. between samples with the same primer index but different
527 library indices, which could be either due to physical mixing in the laboratory, tag-jumping (in the
528 library indices, not the PCR tags), or errors in index sequencing. Trap-level contamination may add to
529 the problem, as the combined model (plate x trap) shows only marginally higher levels of
530 contamination (Supplementary Table S1). Because the contamination within the wells was much
531 lower, we conclude that the primary PCR using 13 different primer tags before being combined in a
532 single Nextera XT library was not greatly affected by these problems, indicating that our approach of
533 using the same unique primer tags on forward and reverse strands can largely eliminate the problem
534 of misassignment of PCR fragments. In addition, some types of contamination were less likely to be
535 introduced during molecular lab processing, given the precautions with specimen handling and the
536 strict protocols of the sequencing facility, in particular regarding the widely found human DNA,
537 present in virtually every one of the specimens. As scientists using morphological and molecular
538 methods work together, greater awareness of these issues is needed and the steps to avoid DNA
539 contamination should be understood and implemented, such as the use of clean pans, bee nets and
540 storage bottles, and use of latex gloves for specimen handling during morphological identification.

541

542

543 Conclusions

544 High-throughput sequencing can greatly change the approach to monitoring of pollinators,
 545 through mass identification of sequence reads against reference databases verified by taxonomic
 546 specialists. In this proof-of-concept study we used individuals, rather than bulk samples, to study the
 547 outcome of metabarcoding in greater detail. We first established the power of the *cox1* marker for
 548 species discrimination, which only left about 5% of UK species without a precise identification at
 549 species level. The subsequent utilization of the database for UK bees monitoring shows high
 550 consistency with morphological identifications conducted in parallel. However, the deep sequencing
 551 of single specimens also revealed the various pitfalls of metabarcoding. We detected surprisingly
 552 high levels of apparent mixing with other specimens from the same and other traps. In addition, we
 553 found numerous OTUs apparently contributed by pseudogenes, which greatly inflate estimates of the
 554 total species diversity; they can be filtered out efficiently as their distribution ‘trails’ the actual
 555 mitochondrial copies, which should be a routine part of the read filtering procedure. Lastly, the
 556 widely used OTU clustering may not produce the most accurate species detection, as shown by a
 557 comparison of OTU analyses against the most abundant read in each sample (after adequate
 558 taxonomic and numerical filtering), which revealed a full identification of the target specimen in
 559 approximately 25% more samples. Yet, applied under stringent quality filtering, it is possible to use
 560 high-throughput sequence data at the read level, i.e. to establish genotypic variation or for
 561 assignment to particular subgroups within the Linnaean species, and thus use them in the same way
 562 as data from Sanger sequencing, but scaled up by orders of magnitude. The method thus greatly
 563 increases the accuracy and speed of taxonomic identification in pollinator monitoring, at reduced
 564 cost, while also providing further information on species interactions and ecosystem composition
 565 through the secondary OTUs. The bioinformatics methodology and comprehensive barcode database
 566 can now be rolled out for the study of much larger number of specimens typically obtained by
 567 passive pan traps and can be extended to studies of pollinators in other parts of the world.

568

569 **ACKNOWLEDGEMENTS**

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

582

583 **Data accessibility**

584 Sequence data available at BOLD under the BEEEE label. Perl scripts used for the sequence
 585 clustering and barcode selection are available at <https://github.com/tjcreedy/NAPtime>.

586

587 **Author contributions**

588 CQT, HN and APV designed the study; CQT and HN generated molecular data; TJC, CQT, KQC
 589 and HN performed data analysis. TJC developed bioinformatics tools. CC, KC and RO collected
 590 specimens and co-ordinated morphological identifications. DGN collected specimens, identified,

591 documented, sampled them, preserved morphological vouchers, and verified identification. PA and
 592 CA designed analytical pipelines and provided advice on project design and analysis. CQT, HN, TJC,
 593 KQC and APV wrote an initial draft of the manuscript. All authors contributed to the writing of the
 594 final draft.

595

596 **Competing interests**

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

599

600

601 REFERENCES

- 602 Amiet F, Herrmann M, Müller A, Neumeyer R (2001) Apidae 3 - *Halictus*, *Lasioglossum*. *Fauna*
603 *Helvetica* **6**, 1–208.
- 604 Amiet F, Herrmann M, Müller A, Neumeyer R (2004) Apidae 4 - *Anthidium*, *Chelostoma*, *Coelioxys*,
605 *Dioxys*, *Heriades*, *Lithurgus*, *Megachile*, *Osmia*, *Stelis*. *Fauna Helvetica* **9**, 1–273.
- 606 Amiet F, Herrmann M, Müller A, Neumeyer R (2010) Apidae 6 - *Andrena*, *Melitturga*, *Panurginus*,
607 *Panurgus*. *Fauna Helvetica* **26**, 1–317.
- 608 Amiet F, Herrmann M, Müller A, Neumeyer R (2007) Apidae 5 - *Ammobates*, *Ammobatoides*,
609 *Anthophora*, *Biastes*, *Ceratina*, *Dasygoda*, *Epeoloides*, *Epeolus*, *Eucera*, *Macropis*,
610 *Melecta*, *Melitta*, *Nomada*, *Pasites*, *Tet*. *Fauna Helvetica* **20**, 1–356.
- 611 Amiet F, Müller A, Neumeyer R (2014) Apidae 2 - *Colletes*, *Dufourea*, *Hylaeus*, *Nomia*, *Nomioides*,
612 *Rhophitoides*, *Rophites*, *Sphecodes*, *Systropha*. *Fauna Helvetica* **4**, 1-239.
- 613 Andujar C, Arribas P, Gray C, *et al.* (2018) Metabarcoding of freshwater invertebrates to detect the
614 effects of a pesticide spill. *Molecular Ecology* **27**, 146-166.
- 615 Arribas P, Andujar C, Hopkins K, Shepherd M, Vogler AP (2016) Metabarcoding and mitochondrial
616 metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in*
617 *Ecology and Evolution* **7**, 1071-1081.
- 618 Bergsten J, Bilton DT, Fujisawa T, *et al.* (2012) The effect of geographical scale of sampling on DNA
619 barcoding. *Systematic Biology* **61**, 851-869.
- 620 Biesmeijer JC, Roberts SPM, Reemer M, *et al.* (2006) Parallel declines in pollinators and insect-
621 pollinated plants in Britain and the Netherlands. *Science* **313**, 351-354.
- 622 Bogusch P, Straka J (2012) Review and identification of the cuckoo bees of central Europe
623 (Hymenoptera: Halictidae: Sphecodes). *Zootaxa*, 1-41.
- 624 Carvell C, Isaac NJB., Jitlal M, *et al.* (2016) Design and Testing of a National Pollinator and Pollination
625 Monitoring Framework. Final summary report to the Department for Environment, Food and
626 Rural Affairs (Defra), Scottish Government and Welsh Government: Project WC1101.
- 627 Cross I, Notton DG (2017) Small-headed Resin Bee, *Heriades rubicola*, new to Britain (Hymenoptera:
628 Megachilidae). *British Journal of Entomology and Natural History* **30**, 1-6.
- 629 DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA
630 barcoding. *Philosophical Transactions of the Royal Society London Series B* **360**, 1905-1916.
- 631 Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC*
632 *Evolutionary Biology* **7**, Art. 214.
- 633 Edgar R (2010) *USEARCH fastq_filter*, available online at <https://www.drive5.com/usearch/>.
- 634 Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of
635 chimera detection. *Bioinformatics* **27**, 2194-2200.
- 636 Falk SJ, Lewington R (2015) *Field guide to the bees of Great Britain and Ireland* Bloomsbury Publishing
637 PLC.
- 638 Fujisawa T, Barraclough TG (2013) Delimiting species using single-locus data and the Generalized
639 Mixed Yule Coalescent approach: A revised method and evaluation on simulated data sets.
640 *Systematic Biology* **62**, 707-724.
- 641 Garibaldi LA, Steffan-Dewenter I, Winfree R, *et al.* (2013) Wild pollinators enhance fruit set of crops
642 regardless of Honey Bee abundance. *Science* **339**, 1608-1611.
- 643 Gill RJ, Baldock KCR, Brown MJF, *et al.* (2016) Protecting an ecosystem service: Approaches to
644 understanding and mitigating threats to wild insect pollinators. In: *Ecosystem Services: From*
645 *Biodiversity to Society, Pt 2* (eds. Woodward G, Bohan DA), pp. 135-206.
- 646 Hallmann CA, Sorg M, Jongejans E, *et al.* (2017) More than 75 percent decline over 27 years in total
647 flying insect biomass in protected areas. *Plos One* **12**.
- 648 Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes.
649 *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.

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.

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/>.

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.

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 CQ, Humphreys AM, Fontaneto D, Barraclough TG (2014) Effects of phylogenetic reconstruction method on the robustness of species delimitation using single-locus data. *Methods in Ecology and Evolution* **5**, 1086-1094.

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

- 698 Vanbergen AJ, Baude M, Biesmeijer JC, *et al.* (2013) Threats to an ecosystem service: pressures on
699 pollinators. *Frontiers in Ecology and the Environment* **11**, 251-259.
- 700 Westphal C, Bommarco R, Carre G, *et al.* (2008) Measuring bee diversity in different European
701 habitats and biogeographical regions. *Ecological Monographs* **78**, 653-671.
- 702 Yoccoz NG, Brathen KA, Gielly L, *et al.* (2012) DNA from soil mirrors plant taxonomic and growth form
703 diversity. *Molecular Ecology* **21**, 3647-3655.
- 704 Zhang JJ, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd
705 mergeR. *Bioinformatics* **30**, 614-620.

706

707

708

709 Tables and Figures

710

			Most frequent OTU	Most frequent read	NAPselect
A					
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%)
			Most frequent OTU	Most frequent read	NAPselect sequence
B					
Morphological ID level	Molecular ID level				
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%)

711

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

719

720

721 Figure 1

722

723

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

732

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

738

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

746

747

748

749

750 Figure 4: The proportion of molecular identification failure for different morphological species across
751 genera. For each morphological species, we calculated the proportion of specimens for which the
752 designated barcode failed to be correctly identified using the reference database. These values are
753 presented here, grouped by genus and the three different barcode designation methods. Values
754 along the x-axis show the number of morphological species in the genus, and the total number of
755 specimens.

756

757

758

759 Figure 5: The taxonomic composition of secondary OTUs in NGS barcoding of bees. Boxplot shows
 760 the average number of secondary OTUs within major taxonomic groups found in each sample. Values
 761 below boxes give the total number of OTUs for that taxon found across the dataset.

762

763

764

765

766

767

768 Figure 6: Boxplot showing the rate of recovery of all possible cross-contamination OTUs from
769 different sources of cross contamination. The rate of shared OTU recovery is significantly higher
770 when considering samples from the same plate and same plate and trap compared with a
771 background rate of cross contamination (all possible).