Species identification of Swedish mosquitoes through DNA metabarcoding

Tobias Lilja[1], Johan A. A. Nylander[2], Karin Troell[1] and Anders Lindström[1]

[1]SVA, National Veterinary Institute, Dept of Microbiology, Sweden, 751 89 Uppsala.

[2]BILS/Dept. of Bioinformatics and Genetics, Swedish Museum of Natural History, Box 40007, Stockholm, Sweden.

Corresponding author: Tobias.lilja@sva.se

First published online 16th February 2017

Abstract: DNA-barcoding utilises a fragment of the mitochondrial cytochrome oxidase subunit 1 (COI) gene to identify most animal species. Using next generation sequencing (NGS), this method can be further developed into metabarcoding processes that allow the simultaneous identification of several species from a mixed sample. We created a database of COI sequences of 27 mosquito species collected in Sweden, and combined our data with 27 additional sequences from GenBank to cover the taxa recently documented in Sweden and to include possible invasive taxa. Comparisons show that COI metabarcoding reliably identifies 41 of 54 species and the remainder to species group. Using three independent primer pairs along the COI gene, we further developed this barcoding approach to simultaneously identify Swedish mosquitoes in communities using NGS and quantify relative abundance of each mosquito species in the sample, using bioinformatics methods. We tested the accuracy of the metabarcoding method using communities assembled from morphologically identified mosquitoes, revealing 80% positive identification rate and the estimates of population structure which reflects the input sample. We conclude that metabarcoding is useful as a high throughput identification technique and for the quantification of species. ***Journal of the European Mosquito Control Association* 35: 1-9, 2017**

Keywords: Culicidae, metabarcoding, COI, next generation sequencing, vectors, surveillance

Introduction

Mosquitoes are capable of transmitting a wide range of pathogens, such as parasites, bacteria or viruses. Mosquito-borne infections are also among the most important new and emerging diseases globally (Gubler 2002, ECDC 2012). In Europe, several exotic vectors have established populations and are expanding their range (ECDC 2012). There has also been autochthonous outbreaks of exotic diseases such as dengue, chikungunya, Usutu and West Nile fever, and vector-borne diseases are responsible for nearly a third of the recorded emerging infectious disease events in the last decades (ECDC 2014). The opportunities for invasive vector species as well as their associated pathogens to become established in regions of Europe are increasing through changes in climate, travel and global trade. The vector competence for any pathogen varies between mosquito species and it is of utmost importance to monitor the distribution of these vector taxa. This distribution data can be used to predict the spread of vector-borne diseases during future outbreaks, and to focus mosquito control programmes to areas where vector species are present. Control of invasive species is also dependent on efficient systems for surveillance to be effective (ECDC 2012).

Monitoring programmes, often using carbon dioxide emitting traps, can easily collect thousands of mosquitoes per night. The morphological identification of mosquitoes requires expert training, is time consuming and often fails to identify damaged specimens or distinguish between cryptic or isomorphic taxa. In addition, if traps are subsampled, there is a risk that rare species are missed. When surveillance is based on ovitraps or collection of mosquito larvae, morphological identification may be challenging. For these reasons, molecular techniques to identify mosquito vectors in field samples have been developed. For example, mosquitoes as well as other insects have recently been identified through use of MALDI-TOF mass-spectroscopy (Campbell 2005, Kaufmann et al. 2011, Yssouf et al. 2014), and many animal species can be identified by sequencing the mitochondrial cytochrome oxidase subunit 1 (COI) gene which has been established as the biological barcode. The publicly available Barcode of Life Database (BOLD) (Ratnasingham & Hebert 2007) hosts COI sequences from many animal species with morphologically identified voucher specimens.

For most countries, a complete COI barcode reference databases of the mosquito fauna is not available, but for Belgium all 24 species found during a 2-year survey were barcoded (Versteirt et al. 2015). Several studies of mosquito barcodes show that the intraspecific variation was smaller than the interspecific variation in most cases (Cywinska et al. 2006, Versteirt et al. 2015). This allowed species level identification using COI for 22 of the species collected in Belgium, with the exception of *Aedes annulipes* and *Aedes cantans* that could not be resolved by this method. For the 49 reported mosquito species from Sweden (Lundström et al. 2013), some species had already been sequenced for COI (Engdahl et al. 2014) but no comprehensive database was available for the COI sequences of Swedish mosquitoes. In the BOLD database, COI sequences were available for 40 species from other countries, but 9 species are not yet represented.

Although identifying individual samples using MALDI TOF mass spectrometry or DNA barcoding are undoubtedly

useful, they are not convenient and economic for large-scale surveillance activities, which can include thousands of mosquitoes. In order to minimise sample handling, next generation sequencing (NGS) can be employed with barcoding primers, which is sometimes referred to as metabarcoding (Taberlet et al. 2012). NGS has been used for several arthropod biodiversity studies (Hajibabaei et al. 2012, Yu et al. 2012, Zhou et al. 2013, Gibson et al. 2014), and it has been shown that sequencing of COI amplicons from mixed samples can facilitate identification of most morphologically identifiable species in the sample, through comparison to reference data. However, due to mismatches in the primer binding site or differences in GC content, making some species harder to PCR-amplify than others, it is believed that primers used may potentially select for some species over others. For these reasons the quantification of results has not been a priority but all identified species have been considered equally abundant in the sample.

One way to circumvent primer bias has been to use several primer pairs and combine the results (Gibson et al. 2014). Another way has been to use a PCR-free method where a crude mtDNA extract was purified by centrifugation and directly prepared for NGS (Zhou et al. 2013). A PCR-free approach was evaluated for a study of bee diversity where mitogenome references were created for 48 species of bees. These references were then used to map whole genome sequence reads from community samples of bees to allow species identification and quantification (Tang et al. 2015).

It is of high interest to be able to estimate both the presence and abundance of mosquito species for use in bio-surveillance activities. We here describe a molecular species identification method to enable bulk sample analysis of Swedish mosquitoes that is also able to predict relative abundance of the identified species.

Materials and Methods

Sequencing of reference specimens

Individual adult female mosquitoes were morphologically identified by a trained entomologist using morphological keys (Becker et al.. 2010), with taxa names used herein following the most recent classification proposed by Wilkerson et al. (2015). Three mosquito legs were used for DNA isolation by homogenization in 30 µl Prepman Ultra (Life Technologies, Carlsbad, CA, USA) (Ander et al. 2013). The sample was then lysed at 100°C for 10 min. Tissue debris was removed by centrifugation at 12,000 x g for 2 min after which 20 µl of the supernatant were transferred to a fresh tube and used as template in PCR reactions. The COI region was amplified using two previously published primer pairs, including the universal barcoding primers, LCO1490 (GGTCAA CAAATCATAAAGATATTGG) and HCO2198 (TAAAC TTCAGGGTGACCAAAAAATCA) (Folmer et al. 1994), and GB_1358_83F (ACTCAAGAAAGAGG TAAAAAGGAAAC) and TL2-N-3014R (TCCAATGCACTAA TCTGCC-ATATTA) (Engdahl et al. 2014) which amplify the 5'-part and the 3'-part of the gene correspondingly. PCR reactions were performed using AmpliTaq DNA polymerase (Invitrogen, Thermo Fischer Scientific, Waltham, MA, USA).

PCR products were purified with JETQUICK PCR Purification Spin Kit (Genomed, Löhne, Germany) and sequenced using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Invitrogen, Thermo Fischer Scientific, Waltham, MA, USA) in both directions using the original PCR primers. The product was run on an ABI 3100 sequencer (Applied Biosystems, Thermo Fischer Scientific, Waltham, MA, USA). Sequences were assembled using BioEdit Sequence Alignment Editor.

Sequences covering the 5´-part and the 3´-part of COI were concatenated into one sequence. Depending on availability, we sequenced several specimens for each species where possible: *Aedes cantans* (n=3), *Ae. caspius* (n=1), *Ae. cataphylla* (n=2), *Ae. cinereus* (n=4), *Ae. communis* (n=3), *Ae. dorsalis* (n=3), *Ae. flavescens* (n=1), *Ae. geniculatus* (n=6), *Ae. hexodontus* (n=1), *Ae. leucomelas* (n=1), *Ae. nigrinus* (n=2), *Ae. punctor* (n=5), *Ae. refiki* (n=1), *Ae. pullatus* (n=1), *Ae. rossicus* (n=4), *Ae. rusticus* (n=2), *Ae. vexans* (n=34); *Anopheles algeriensis* (n=4), *An. beklemishevi* (n=4), *An. claviger* (n=3), *An. messeae* (n=2); *Culex pipiens* (n=2); *Culiseta alaskaensis* (n=1), *Cs. bergrothi* (n=1), *Cs. morsitans* (n=2), *Cs. ochroptera* (n=2), *Cs. subochrea* (n=1). A reference database for Swedish mosquitoes was created from the sequenced specimens along with COI gene sequences from the BOLD database. All sequences have been deposited in GenBank (accession numbers KP942677 - KP942777).

Metabarcoding

To test if the frequency of mosquitoes of each species in a mixed sample could be determined using a metabarcoding approach(fig. 1), eight batches were prepared, of which six contained 100 morphologically identified mosquitoes of different species and two contained mosquitoes that had been identified by Sanger sequencing to evaluate mistakes due to morphological misidentification (Supplementary Table 1). Mock community samples were assembled to simulate true trap catches with one or a few dominating species and some rare species. The sample size of 100 mosquitoes was chosen in order to create communities where a rare species could make up 1% of the sample. Mosquitoes that had been individually Sanger-sequenced were missing three legs each but were otherwise prepared similarly to the other batches.

Mosquito collection samples were prepared into DNA samples enriched for mtDNA (mitochondrial DNA) as previously described (Zhou et al. 2013). Each batch of mosquitoes was homogenized in 5ml MS buffer (210 mM mannitol, 70 mM sucrose, 5 mM TrisHCl, 1 mM EDTA), aliquoted into six labelled 1.5 ml tubes and centrifuged at 1,300 g for 2 min at 4°C. The supernatant was transferred to new tubes and further centrifuged at 17,000 g for 30 min at 4°C. This procedure will enrich the mitochondrial portion of the sample. The supernatant was removed, and pelleted mitochondria were lysed using 40 µl mitochondrion lysis buffer (0.15 M NaCl, 10 mM TrisHCl, 1 mM EDTA, with 5% SDS and 0.5 mg/ml proteinase K) per tube and incubated at 56°C for 15 min. DNA was then purified using JETQUICK PCR Purification Spin Kit. DNA purified from each batch of mosquitoes was used as template in PCR reactions using primer pairs LCO1490/ HCO2198 (referred to as LCO-primers), GB1310_29F(GAAGGAGTTTGATCAGGAATAGT)/ GB_1960_1936R(TCCTCCTCCAATAGGGTCAAAGAA) (Engdahl et al. 2014) (referred to as COI5-primers) and GB_1358_83F/TL2-N-3014R (referred to as COI3-primers) (fig. 1A) The resulting fragments as well as the purified DNA from each batch were then processed for Illumina sequencing by the Nextera XT DNA Library Preparation Kit (Illumina Inc., San Diego, CA, USA), which randomly cuts the DNA into fragments onto which sequencing adapters are ligated. The resulting DNA fragments were inspected by Bioanalyzer, indicating fragment lengths of 500-700 bp, and subsequently sequenced using the MiSeq Reagent Kit v3 (600-cycle) (Illumina Inc.) for paired end sequencing. Illumina sequencing

of several libraries in the same sequencing run was allowed by use of Illumina index adapters. Mean sequencing depth per amplicon was 175,000 reads (105 megabases). For the PCR-free shotgun approach mean sequencing depth was 2.1 million reads (1,26 gigabases).

Fastq sequences resulting from the batches were trimmed to remove primer sequences using the fastx-toolkit-0.0.14 (Assaf Gordon, http://hannonlab.cshl.edu/fastx_toolkit/) and used directly as single reads for mapping to the reference sequences using Bowtie2 (Langmead et al. 2012). Bowtie2 was run using default settings and reporting only aligned reads. During mapping, Bowtie2 calculates a quality score, AS, to describe how well the read matches the reference. In the default end-to-end mode, the AS score is the sum of penalties for mismatches and gaps in the alignment. This quality score does not compensate for differences in length between reads by default. Reads mapping to the reference sequences longer than 200 bp and with a quality score AS>-30 were counted, and the proportion of reads matching each species of the reference database were presented as a measure of relative abundance of each species in the sample. Several different cut-offs for quality and length were tested but did not give similarly good approximations of the mosquito communities (data not shown). Reads that did not match any species were discarded in the analysis.

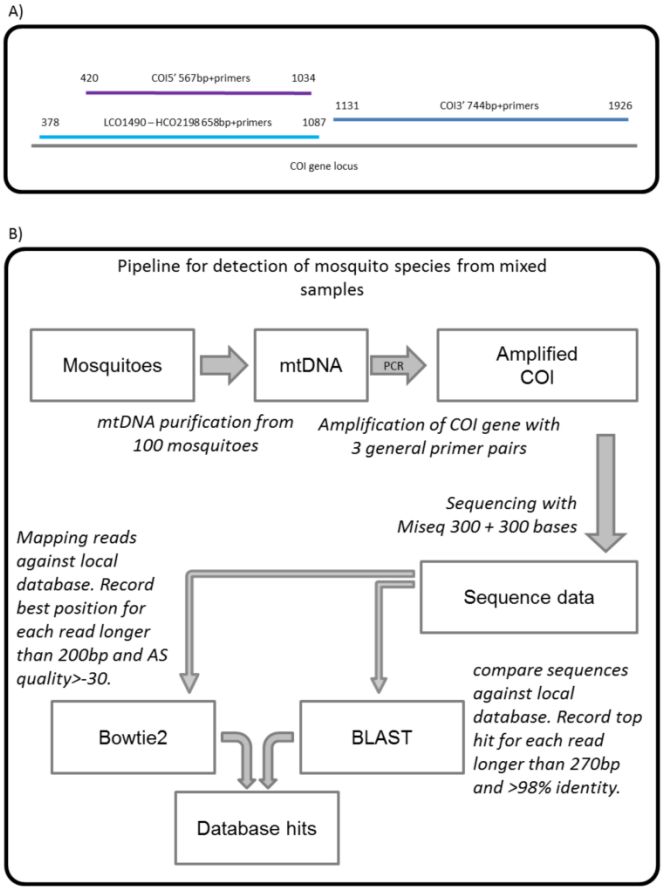

Figure 1: (A) Illustration of the COI loci and fragments produced by the primers used. (B) Outline of the metabarcoding workflow used.

The pair end fastq files were merged into long single reads using Pear (Zhang et al. 2014) with the fastx-toolkit to remove primer sequences, and were converted into FASTA format. Sequence reads produced were compared to the mosquito database by stand-alone BLASTn reporting top hits only. Hits longer than 270 bp, with more than 98% identity to the reference, were counted and the proportion of reads matching each species of the reference database were presented as a measure of relative abundance of each species in the mosquito pool. Other cut-offs for length and identity were tested but did not result in as good approximations of the mosquito communities as the values subsequently used (data not shown). To combine results for several primer pairs, an average between the results for each primer pair was calculated for each species. The resulting populations were then compared to the population structure of the mock community samples to evaluate how well the method could recreate the samples in regards to species presence and abundance.

Results

In order to facilitate mosquito surveillance in Sweden, we produced a database of COI sequences of 48 mosquito species recently reported in Sweden as well as additional mosquito species currently posing big problems as invasive species in other European countries, such as *Ae. albopictus*, *Ae. aegypti*, *Ae. koreicus*, *Ae. japonicus*, *Ae. triseriatus* and *Ae. atropalpus*. For 27 species, that were at the time the project started not available in the BOLD database, new barcode sequences were produced using morphologically identified mosquitoes, but for other Swedish species we relied on barcode sequences present in the BOLD database. At the time of manuscript writing, COI barcodes were not publicly available for nine of these taxa, however barcodes for eight of these (*An. algeriensis*, *Ae. refiki*, *Ae. cyprius*, *Ae. detritus*, *Cs. alaskaensis*, *Cs. subochrea*, *Ae. nigrinus* and *Ae. leucomelas*) were obtained during this study. The final missing taxa is *Ae. geminus*, which was reported in older records of Swedish mosquitoes but has not been recently substantiated (Lundström et al. 2013).

The interspecific variation, based on the 5´-part of COI available for all species range from 0.002-0.192 differences per site (Supplementary Table 2), with the most divergent species being *An. maculipennis* and *Cq. richiardii* (0.192) and the most similar being *Ae. hexodontus* and *Ae. punctor* (0.002).

We also estimated the intraspecific variation by sequencing 34 *Ae. vexans* mosquitoes from eight different locations. Our results show that the divergence in the COI locus in this species is large (pairwise distance 0.000 to 0.063 differences per site), and the sequences are divided into two clearly separate groups based on the estimated phylogeny (Supplementary Fig. 1). Despite the large divergence, all *Ae. vexans* sequences grouped together when compared to COI sequences from all 54 Swedish and possible invasive species included in this study.

In our dataset of Swedish mosquitoes, only 0.7% of pairs showed lower than 2% difference. In barcoding of arthropods using COI, a similarity higher than 98% has been considered necessary for assigning specimens to the same species, and in dipterans only 3% of tested congeneric pairs had a lower divergence (Hebert et al. 2003). The variation in the COI gene is sufficient to distinguish 41 of all 54 Swedish and possible invasive mosquitoes to species level, but between some closely related species the variation within the COI gene is too low to assign an unknown sample to a species to. In these cases, more sequence data is needed to get sufficient resolution. The species for which the COI sequences are insufficient are (pairwise distance in parentheses): *Ae. intrudens* and *Ae. diantaeus* (0.01), *Ae. cataphylla* and *Ae. leucomelas* (0.012), the *Aedes punctor*-group (*Ae. hexodontus*, *Ae. punctor*, *Ae. punctodes* (0.002, 0.009 and 0.009)), the *Aedes annulipes*-group (*Ae. cantans*, *Ae. annulipes*, *Ae. excrucians* (0.01, 0.014 and 0.008)) and the closely

related *Ae. cinereus* and *Ae. rossicus* (0.004)). Furthermore, morphological methods cannot distinguish adult females of *Ae. punctor* and *Ae. punctodes* or *Ae. cinereus* and *Ae. geminus*, making it difficult to acquire voucher specimens in order to sequence other genes.

We further compared our species discrimination with the BINs (clusters approximating taxa) available through the BOLD database and found that our distinctions matched those of BINs available in the database, such that species that we could not distinguish using the barcoding 5-part of COI were also part of the same BINs in the BOLD database.

Eight sets of morphologically determined species were assembled into mock communities of known composition in order to test the validity of the metabarcoding method for mosquito identification. Sequence reads from amplicon based sequencing using three primer pairs covering COI compared with the generated COI reference database of Swedish and possible invasive mosquitoes resulted in detection of most species but also in low levels of false detections of species not included in the sample (Fig. 2). Even though the primers used are well established, have been used to amplify COI from a wide range of species, and have been placed in conserved positions, there was some diversity in the primer binding positions (Supplementary Fig. 3). We tested the efficiency of each primer pair individually and combined the results of all three primer pairs to avoid primer specific bias against certain species. Results from the LCO pair (LCO1490 and HCO2198) underestimates mosquitoes from *Ae cinereus*, *Ae. rossicus* and *Ae. vexans* but overestimates the presence of *Cs. bergrothi*, while results using the COI5 primer pair (GB1310_29F and GB_1960_1936R) overestimates specimens of the subgenus *Aedes* and underestimates *Cs. bergrothi*. Results from the COI3 primer pair (GB_1358_83F and TL2-N-3014R) has an overrepresentation of the *Ae. annulipes*-group and underrepresentation of the *Ae. punctor*-group. Each primer pair alone failed to identify more species than when the results were combined (Fig. 2). The Pearson correlation between the input fractions and resulting fractions of species in all eight community samples were also lower for single primer pairs than for the combination of all three pairs (Figs. 4A and B). Although no single primer pair was as good as the combination of all three pairs, the three primer pairs tested differ in how well sequence reads from the amplicons represent the sampled mosquito batches, where the COI3 pair seem to be better than the LCO pair and COI5 pair (Fig. 2).

The comparison of the two bioinformatic methods for assigning sequence reads to mosquito species, BLAST (Altschul et al. 1990) and Bowtie2 (Langmead et al. 2012), basically shows that the two bioinformatics methods produced very similar results (Pearson correlation 0.99) (Figs. 2-4). The speed of the two algorithms used was compared on several datasets and the Bowtie2 pipeline was five times faster on average than the BLAST pipeline, but the difference was more pronounced on large datasets than on datasets with fewer reads.

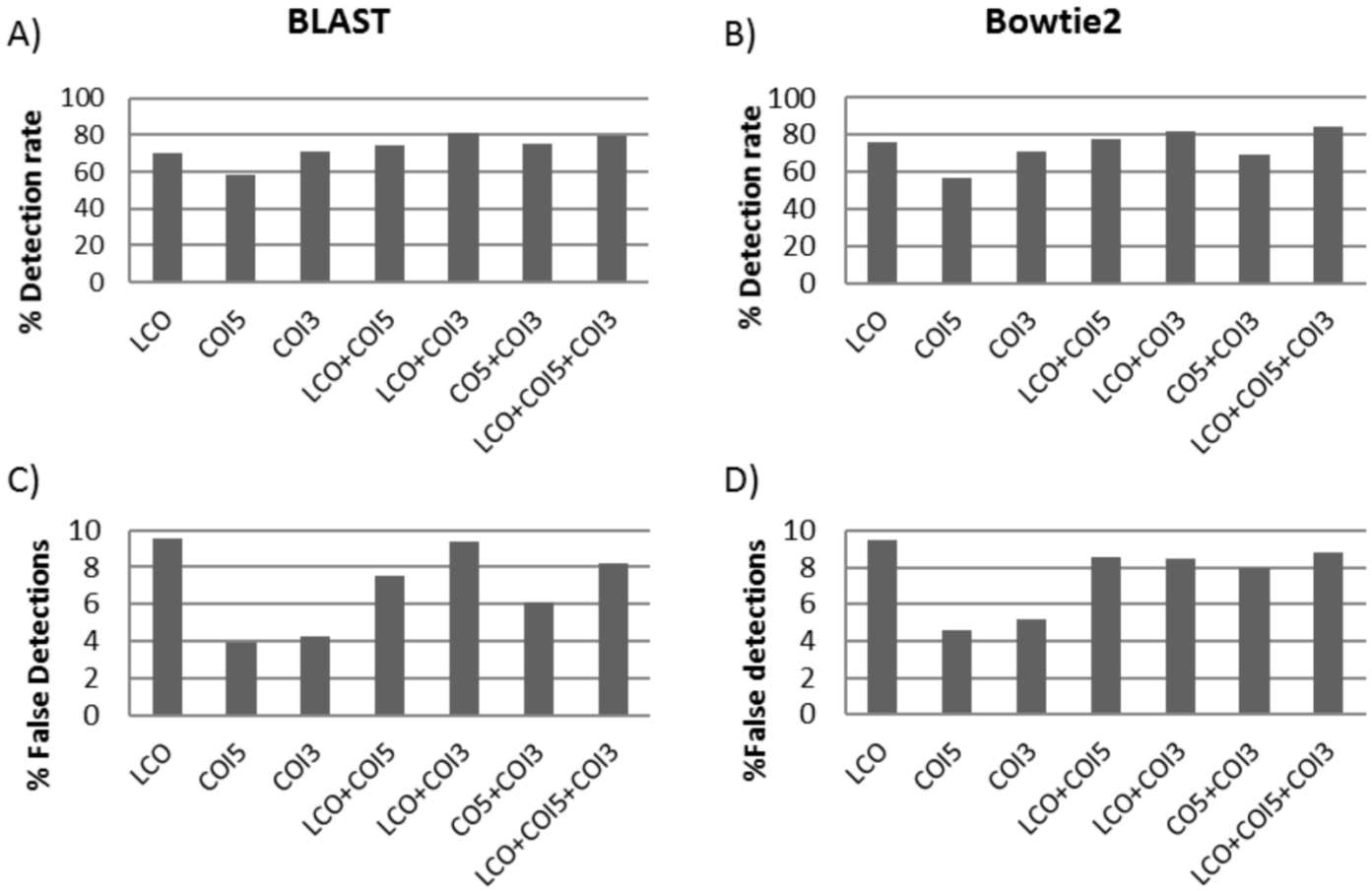

Figure 2: Results of metabarcoding method on known populations. Detection of mosquito species in known populations using single primer pairs and combinations of primer pairs using (A) the BLAST based data analysis and (B) the Bowtie2 based data analysis. Figures 2C and 2D show false positives of mosquito species in known populations using single primer pairs and combinations of primer pairs.

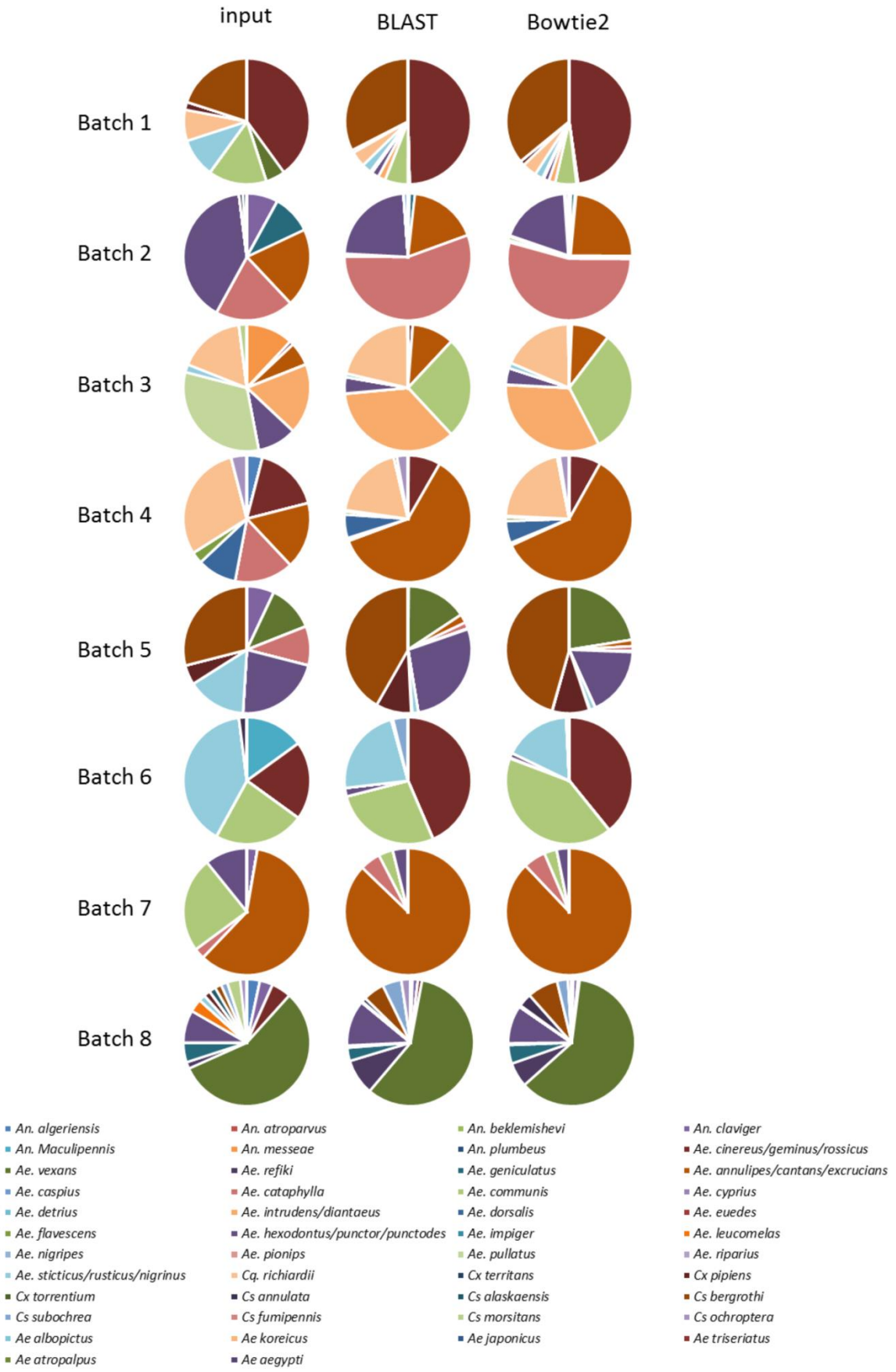

Figure 3: Graphic representations of the population structure of known population input and results for the BLAST based data analysis and the Bowtie2 based data analysis.

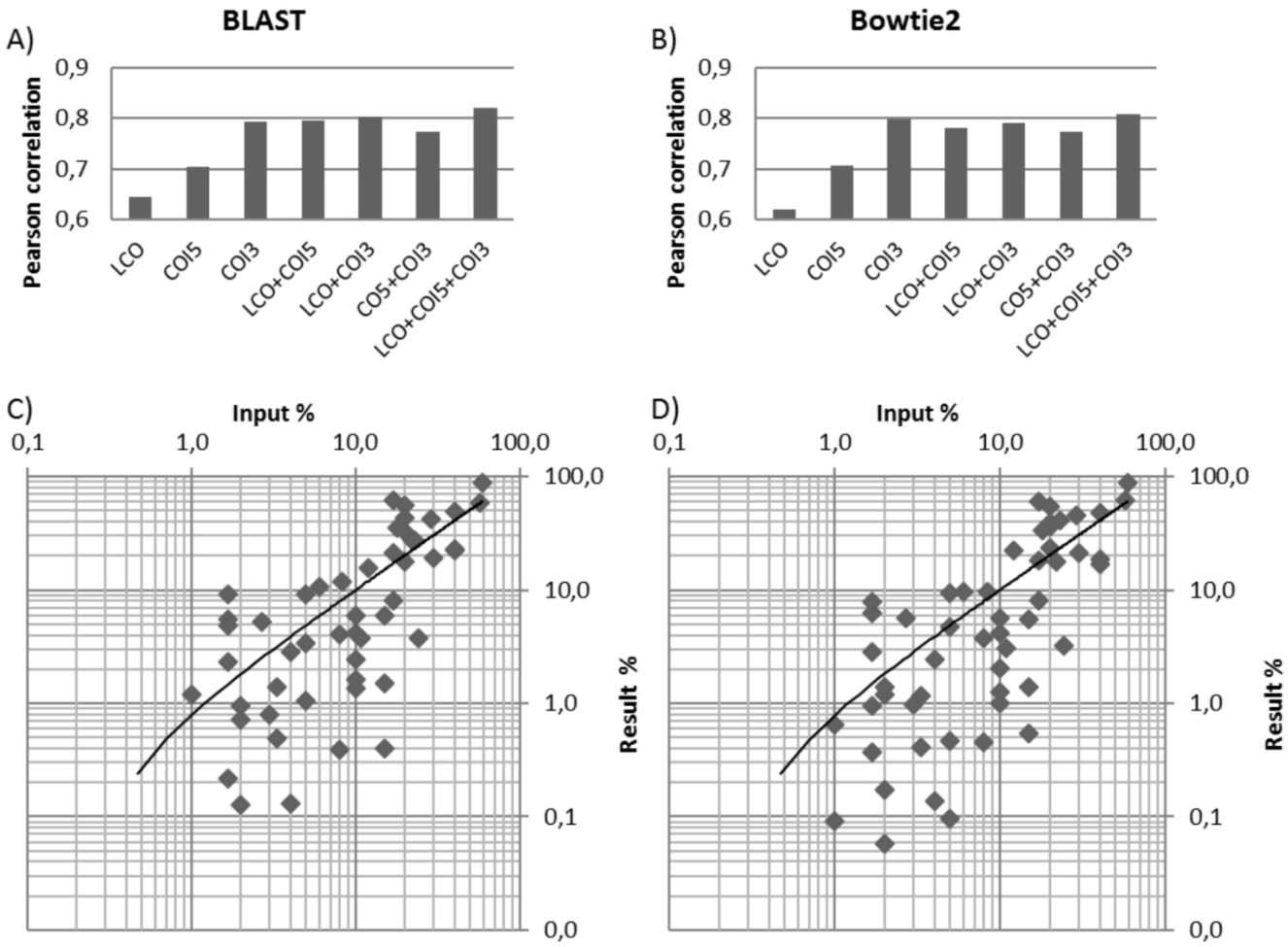

Figure 4: Pearson correlation between known population input and results for single primer pairs and combinations of primer pairs using (A) the BLAST based data analysis and (B) the Bowtie2 based data analysis. Proportion of each species in the input sample plot

By combining the results from all three primer pairs, some of the bias overestimating some species while underestimating others could be balanced out, and the combined results fit the input data better than any of the single-primer results, reaching a Pearson correlation of 0.82 for the BLAST method and 0.81 for the Bowtie2 method (Figs. 3 and 4). The resulting estimated population structures are thus similar to the morphologically determined input in regards to represented species and estimation of the dominant species, even if the proportion for each species in a community sample differs.

Discussion

In this study, we generated new COI barcode sequences for 27 mosquito species, adding to the reference database available for identifying the mosquitoes of Sweden, building up on a previous effort which sequenced 14 of the most common mosquitoes in Sweden (Engdahl et al. 2014). Phylogenetic analysis of the species present in Sweden show that the COI gene is sufficient for identification to species level in most cases and work well as a general identification method. However, the COI 5´-marker is unable to identify 13 species with high confidence. Even when the COI 3´-part is included, the COI fragment is insufficient to distinguish some closely related species as observed already by Engdahl et al. (2014) for *Ae. intrudens* and *Ae. diantaeus* as well as for *Ae. cantans* and *Ae. annulipes*. Also morphological identification of adult females is unable to distinguish ten of the species. For certain species, another molecular marker would thus be helpful for reliable identification. Previously, Indian mosquitoes have been analysed by barcoding of the COI marker, and 61 out of 63 morphologically determined species could be identified, with one pair of closely related species having too similar sequences to be separated (Kumar et al. 2007). Also in the genus *Culicoides*, there are species that cannot be separated by the COI marker, and the presence of cryptic species complicates matters where the sequences are divergent but specimens cannot be distinguished morphologically (Ander et al. 2013). However, with any identification system there will be limits to the resolution, so the COI 5´-marker is still a reasonable marker for population scale examinations for mosquitoes.

For metabarcoding methods, where many individuals are analysed together, the number of individuals that can be pooled without loss of detection of rare species depend on how sensitive the method is. This is to a large extent dependent on sequencing depth. For our metabarcoding protocol, 100 mosquitoes per mock community were used to allow detection of single individuals from a rare species. A field sample may be divided into several community samples before DNA preparation in order to minimise the risk of losing rare species from the result.

Metabarcoding has been used to assay biodiversity of many environments but has not been used to make a quantitative estimate on abundance of the different species. Zhou et al. (2013) showed that with their PCR-free method there was a correlation between biomass of a specimen and sequencing volume of that species. Recently, others have shown that also

PCR-based metabarcoding can be quantitative (Diaz-Real et al. 2015, Elbrecht et al. 2015). However, it is clear that primer bias is a problem when assaying more diverse community samples (Elbrecht et al. 2015).

We evaluated the diversity in the primer binding positions for primers and species where sequences of the primer binding positions were available. There is some diversity in the primer binding positions that might affect primer binding and thus PCR efficiency. For further studies, use of degenerate primers might accommodate this issue. To compensate for this primer bias we used three primer pairs and combined the results to increase both detection of species and the correlation between the input population structure and the result. We tested the correlation between the number of mosquitoes of a certain species in the sample and the amount of sequences matching that species and found that sequencing volume can be a good approximation for species quantification. Studies looking at primer binding site conservation across insects (Clarke et al. 2014, Deagle et al. 2014) have noticed that finding primers that amplify a broad repertoire of species may be harder in COI than in mitochondrial 16S and 12S DNA. Many of these markers are shorter than the COI markers we have used and also showed lower taxonomic resolution on closely related *Anopheles* species than the longer COI markers (Clarke et al. 2014), and would thus not be suitable for metabarcoding of mosquitoes. In a more narrowly defined dataset such as ours where all specimens of interest are of the same family the challenge to find primers that amplify all species is less problematic, especially when several primer pairs are combined.

In an attempt to avoid primer bias, we tested direct sequencing of mtDNA extracted from community samples of mosquitoes following a previously published protocol (Zhou et al. 2013). However, this approach required deeper sequencing as the mtDNA was not enriched enough compared with nuclear DNA in our preparations (data not shown). Our community samples were sequenced to an average depth of 2.1 million reads per sample which resulted in only 23 reads on average mapped to the COI regions of all species, whereas sequences matching mosquito rRNA as well as bacterial genomes were well represented. In order to be able to detect rare species in a community sample, sequencing depth would need to be at least 100fold deeper. Also according to Zhou et al. (2013), the reported fraction of reads mapping to mitochondria is only 0.53% of total reads which is in the same order of magnitude as our results. To use this method and match reads to a database of COI sequences requires sequencing to a very large depth. A PCR-free method used for determining bee diversity (Tang et al. 2015) required very deep sequencing even with the complete mitochondrial genomes available for mapping.

Full mitochondrial genomes for each species requires a thorough sequencing effort as well as access to well-identified voucher specimens, especially when intraspecific variation is taken into consideration. Further developments of methods to more stringently isolated mtDNA from insects as well as cost reduction for NGS may make amplification-free methods more attractive for mosquito surveillance. One method to isolate mtDNA in an effective way is mtDNA-capture by hybridization to probes covering mitochondrial genes from many species (Liu et al. 2015). This method could enrich mtDNA by 100fold but would, however, need to be tested for any bias it might introduce, depending on the conservation of mitochondrial genes between the species of interest. So, while a PCR-free metabarcoding method may be the most quantitative it is currently not economically feasible for surveillance.

In the field of metabarcoding, many bioinformatic methods have been used to analyse the data (Hajibabaei et al. 2012, Yu et al. 2012, Ji et al. 2013, Zhou et al. 2013), depending on the sequencing method. In this case, we wanted to use well-established primers for the amplification PCRs. However, this resulted in fragments that were too long for direct amplicon sequencing on the Illumina MiSeq system with 300 bp pair end method. We instead prepared sequencing libraries using the Illumina Nextera XT Kit to create fragments that could be sequenced. This approach resulted in sequences covering the complete amplified fragments but made it impossible to use clustering programmes that required all sequences to cover the same fragment. We instead tested two well-established bioinformatics methods, Bowtie2 and BLAST, to match sequences against the same sequence database. The similarity of the results indicates that for metabarcoding the accuracy of the result rather relies on primer design being unbiased and sequencing method used creating sufficiently long reads to allow good matches to distinguish also closely related species than on the bioinformatics method used. In our case, the Bowtie2 method was much faster and still produced a similar result as the slower BLAST-based method.

False detection of mosquito species from the metagenomics data can result from the presence of certain species with stretches of similar DNA, such that read sequences may match the wrong species, even though they have less than 98% identity over the whole COI region. To avoid misidentifications, reads as long as possible should be used. To further improve correct attribution of species, the reference database could be optimised by having longer references available for all species and from more samples from each species to account for intraspecific variation.

Results not corresponding between sequence-based and morphological identification can also be attributed to morphological misidentifications. In our data there was one example of such a mistake leading to mosquitoes included in the mixed batch #3 as *Ae. pullatus* (32% of the batch) most likely being the source of sequences matching *Ae. communis* in the analysis results (32% and 26% in Bowtie2 and BLAST analysis, respectively). To avoid further mistakes, batches #7 and 8 were assembled from mosquitoes that had all been individually identified by Sanger-sequencing of the COI 5´-region. The results from these samples also had a higher Pearson correlation to the input than results from the other community samples (Supplementary Fig. 3).

In conclusion, we have developed a batch identification of mosquitoes that identify 41 of 54 Swedish plus potential invasive mosquitoes to species level and to subgenus/species-group level in the remaining cases. The metabarcoding method can quantify the proportion of each species in the population sample to a high degree. Further advances to identify all mosquitoes to species level and improve quantification may be possible by testing other primer combinations. Since morphological identification is more time consuming and cannot handle damaged specimens there is a niche for an efficient method to identify large samples of mosquitoes. However, for surveillance purposes such a method needs to be affordable. Even though the running costs of NGS has dropped dramatically during the last decade it is still an expensive method. By using amplicon-based sequencing a lot of data can be generated from a single NGS run, making the cost per sample much lower than Sanger sequencing. The processing time per sample is also reduced making the method easier to

scale up. The use of metabarcoding is an economically viable method for large scale monitoring of mosquito species allowing many samples to be studied in a short period of time and also allows identification of larvae and damaged specimens.

Acknowledgements

This work was financed by EMIDA-VICE project (Tobias Lilja and Anders Lindström), Swedish Board of Agriculture and Formas project 2014-1556, MOBOZO (Tobias Lilja). The authors thank Eric Blomgren for morphological identification of mosquitoes used in the paper and reviewers for their comments that helped improve the paper.

Author Contributions

TL performed the laboratory work, the bioinformatic analysis and wrote the manuscript. JAAN contributed to the bioinformatic analysis and to the manuscript. KT initiated the project and contributed to the manuscript. AL initiated the project, provided previously collected mosquito samples and contributed to the manuscript.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403-410.

Ander, M., Troell, K. & Chirico, J. (2013) Barcoding of biting midges in the genus *Culicoides*: a tool for species determination. *Medical and Veterinary Entomology*, **27**(3), 323-331.

Becker, N., Zgomba, M., Boase, C., Madon, M., Dahl, C. & Kaiser, A. (2010) Mosquitoes and their control, 2nd Edition. Heidelberg, Springer.

Campbell, P.M. (2005) Species differentiation of insects and other multicellular organisms using matrix-assisted laser desorption/ionization time of flight mass spectrometry protein profiling. *Systematic Entomology*, **30**(2), 186-190.

Clarke, L. J., Soubrier, J., Weyrich L.S. & Cooper, A. (2014) Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, **14**(6), 1160-1170.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F. & Taberlet, P. (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, **10**(9), e20140562.

Diaz-Real, J., Serrano, D., Piriz A. & Jovani, R. (2015) NGS metabarcoding proves successful for quantitative assessment of symbiont abundance: the case of feather mites on birds. *Experimental and Applied Acarology*, **67**(2), 209-218.

ECDC (2012) Guidelines for the surveillance of invasive mosquitoes in Europe. Stockholm, ECDC.

ECDC (2014) Guidelines for the surveillance of native mosquitoes in Europe. Stockholm, ECDC.

Elbrecht, V. & Leese, F. (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS One*, **10**(7), e0130324.

Engdahl, C., Larsson, P., Näslund, J., Bravo, M., Evander, M., Lundström, J.O., Ahlm, C. & Bucht, G. (2014) Identification of Swedish mosquitoes based on molecular barcoding of the COI gene and SNP analysis. *Molecular Ecology Resources*, **14**(3), 478-488.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**(5), 294-299.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., Hallwachs, W. & Hajibabaei, M. (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences*, **111**(22), 8007-8012.

Gubler, D.J. (2002) The global emergence/resurgence of arboviral diseases as public health problems. *Archives of Medical Research*, **33**(4), 330-342.

Hajibabaei, M., Spall, J.L., Shokralla, S. & van Konynenburg, S. (2012) Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology*, **12**(1), e28.

Hebert, P.D., Ratnasingham, S. & de Waard, J.R. (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, **270** (Suppl 1), S96-99.

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P.M., Woodcock, P. & Edwards, F.A. (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**(10), 1245-1257.

Kaufmann, C., Ziegler, D., Schaffner, F., Carpenter, S., Pflüger, V. & Mathis, A. (2011) Evaluation of matrix-assisted laser desorption/ionization time of flight mass spectrometry for characterization of *Culicoides nubeculosus* biting midges. *Medical and Veterinary Entomology*, **25**(1), 32-38.

Kumar, N.P., Rajavel, A., Natarajan, R. & Jambulingam, P. (2007) DNA barcodes can distinguish species of Indian mosquitoes (Diptera: Culicidae). *Journal of Medical Entomology*, **44**(1), 1-7.

Langmead, B. & Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**(4), 357-359.

Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., Zhang, H., Misof, B., Kjer, K.M. & Tang, M. (2015) Mitochondrial capture enriches mito-DNA 100fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, **16**(2), 470-479.

Lundström, J.O., Schäfer, M.L., Hesson, J.C., Blomgren, E., Lindström, A., Wahlqvist, P., Halling, A., Hagelin, A., Ahlm, C. & Evander, M. (2013) The geographic distribution of mosquito species in Sweden. *Journal of the European Mosquito Control Association*, **31,** 21-35.

Ratnasingham, S. & Hebert, P.D. (2007) BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular Ecology Notes*, **7**(3), 355-364.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**(8), 2045-2050.

Tang, M., Hardman, C.J., Ji, Y., Meng, G., Liu, S., Tan, M., Yang, S., Moss, E.D., Wang, J. & Yang, C. (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, **6**(9), 1034-1043.

Versteirt, V., Nagy, Z., Roelants, P., Denis, L., Breman, F., Damiens, D., Dekoninck, W., Backeljau, T., Coosemans, M. & Van Bortel, W. (2015) Identification of Belgian mosquito species (Diptera: Culicidae) by DNA barcoding. *Molecular Ecology Resources*, **15**(2), 449-457.

Wilkerson, R.C., Linton, Y.-M., Fonseca, D.M., Schultz, T.R., Price, D.C. & Strickman, D.A. (2015) Making mosquito taxonomy useful: a stable classification of tribe Aedini that balances utility with current knowledge of evolutionary relationships. *PloS One*, **10**(7), e0133602.

Yssouf, A., Parola, P., Lindström, A., Lilja, T., L'Ambert, G., Bondesson, U., Berenger, J.-M., Raoult, D. & Almeras, L. (2014) Identification of European mosquito species by MALDI-TOF MS. *Parasitology Research*, **113**(6), 2375-2378.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**(4), 613-623.

Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, **30**(5), 614-620.

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., Tang, M., Fu, R., Li, J. & Huang, Q. (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2**(1), e4.