Next Generation Pollinator Identification using High Throughput Sequencing (PH0521)

Supplementary Text 1. User Guide to high-throughput sequencing.

Authors: Cuong Q. Tang[1,*], David G. Notton[1], Hannah Norman[1,2], Alfried P. Vogler[1,2]

[1]Department of Life Sciences, The Natural History Museum, London, UK

[3]Department of Life Sciences, Imperial College London, London, UK

*present address: Nature Metrics Ltd, Bakeham Lane, CABI, Egham, Surrey, TW20 9TY

ct@naturemetrics.co.uk

S1. High throughput sequencing technology

S1.1 What is it?

Since 2005, advances in high throughput sequencing technology have revolutionised biological science (Shokralla et al. 2012). High throughput sequencing (otherwise known as next generation sequencing) utilizes massively parallel sequencing to generate millions of sequences amounting to thousands of megabases of information in a single stroke. This is opposed to Sanger sequencing (i.e. the technology used for DNA barcoding), which produces a single sequence per reaction.

High throughput sequencing has many applications, for example individual DNA samples can be sequenced to build genomes (Kocher et al. 2013), mixed DNA samples can be sequenced simultaneously to gain a genomic perspective (metagenomics) of whole communities (e.g. Andújar et al. 2015; Gómez-Rodríguez et al. 2015; Linard et al. 2015; Crampton-Platt et al. 2015), or specific markers can be sequenced (metabarcoding) to gain a more targeted perspective of communities (Yu et al. 2012; Bik, Porazinska, et al. 2012; Creer et al. 2010; Bohmann et al. 2014; Hänfling et al. 2016). These techniques have shed light on many previously inaccessible environments (e.g. deep sea - Bik, Sung, et al. 2012) and organisms (e.g. marine meiofauna - Fonseca et al. 2014; gut bacteria - Meeus et al. 2015), and are being tested for their use with actual biomonitoring (see Pawlowski et al. 2016).

S1.2 How is it better than DNA barcoding?

DNA barcoding using conventional Sanger sequencing has been championed as a means to accurately and rapidly identify species (Hebert et al. 2003), it is limited by the per specimen cost, and as such restricted to small scale monitoring programs (Stein et al. 2014). Further, conventional sequencing requires very high concentrations of high quality DNA template in order to be successful (Polz & Cavanaugh 1998), completely miss out intra-individual variation (Shokralla et al. 2015), and is prone to co-amplification of contaminant organisms (e.g. *Wolbachia* - Smith et al. 2012) that can leave the resulting data near-useless.

High throughput sequencing (HTS) produces such a massive amount of data that a great number of samples can be processed simultaneously. Consequently, there is a tipping point where the expected per-specimen time and costs associated with identifying species is lower with these new technologies (Shokralla et al. 2015; Meier et al. 2015; Stein et al. 2014), a fact well documented in recent literature (e.g. Woodward et al. 2013; Baird & Hajibabaei 2012; Gill, Katherine C.R. Baldock, et al. 2016). However, the starting cost of running a high throughput sequencing run is much greater than a Sanger sequencing run and so these technologies are only cost effective when there is a high enough number of specimens. The sequencing depth of these technologies means that the researcher can successfully analyse low concentration and low quality DNA. *Post hoc* bioinformatic analyses can be used to identify erroneous sequences that plague DNA barcoding studies (i.e. nuclear mitochondrial pseudogenes [NuMTs] and *Wolbachia*). Plus, high throughput sequencing can answer questions that Sanger sequencing technology cannot, for example deeper sequencing of individuals can be used to ascertain diet and trophic associations (e.g. Pompanon et al. 2012; Gibson et al. 2014; Keller et al. 2015; Walker et al. 2016).

As an example of how HTS can be used to identify species from mixed samples, Tang et al. (2015) were able to use mitochondrial metagenomic techniques to identify mock communities of bees. Importantly they were able to correctly recognise 62 species where morphology was only able to correctly identify 53 (e.g. correctly identifying *B. lucorum* workers as opposed to lumping with *B. terrestris*).

The greater than exponential rate of consumables cost drop in DNA sequencing over the last 30 years (a trend that will likely continue given the newer sequencers currently being developed) and the high capacity and read length of new sequencing machines means that it is finally cost effective and feasible to adopt these technologies for biomonitoring programmes.

S1.3 Illumina barcoding

The most common uses of high throughput sequencing technology for biodiversity assessment are from a whole-community perspective (i.e. metabarcoding, metagenomics, mitochondrial metagenomics). The data output from these high throughput sequencing runs is a huge list of sequence reads, these may be specific markers (metabarcoding) or fragmented genomic DNA (shotgun sequencing – metagenomics). These techniques *sensu stricto* cannot easily link specimens to sequences, which might be undesirable from the perspective of some ecological managers and conservationists given the way that some ecological and conservation metrics are currently adopted (i.e. needing diversity and abundance information).

To address this need for individual-based data that can easily slot into established ecological frameworks, we test a new technique which allows for the delimitation of specific specimens as opposed to pools of specimens (e.g. Shokralla et al. 2014; Shokralla et al. 2015; Meier et al. 2015). This technique, which we call **Illumina barcoding**, has been developed at the NHMUK by Paula Arribas, Carmelo Andújar and Alfried Vogler. It is similar to a method described recently by Meier et al. (2016). To maintain specimen separation from DNA extraction through to the sequencing output requires the use of oligonucleotide tags, which can be designed and added to either primers (pre-amplification) or libraries (post-amplification). By combining both uniquely tagged primers with unique library indices it is feasible and cost effective to individually tag hundreds (if not thousands) of specimens to be sequenced, which can be used to generate very detailed diversity and abundance information from highly complex mixed-samples. This method is a high throughput version of standard DNA barcoding. A more detailed review of the different molecular techniques and how to choose and use them are presented in Section 9.

Here we apply Illumina barcoding to individually tag a total of 1,248 specimens to be sequenced on a single sequencing run on a Illumina MiSeq platform. Once sequenced these reads will be identified to species-level using the reference library compiled as part of Objective 1. This will generate a species list for each sampling site and date and this can be compared directly to the species lists produced by the consultant taxonomists as part of the National Pollinator and Pollination Monitoring Framework (donors of many of the samples; Carvell et al. 2016). This comparison will allow us to validate the use of high throughput sequencing for samples akin to those collected as part of a potential future monitoring scheme in the UK.

S2. User guide for high throughput sequencing

S2.1 Introduction

The idea of identifying animal species using a universal stretch of the genome (DNA barcode) has been attributed to the seminal work of Paul D. N. Hebert (2003) although genomic approaches for DNA typing was introduced much earlier for forensics (e.g. Higuchi et al. 1988) and standardised for other groups using different loci (e.g. Yeast - Kurtzman 1994; Bacteria - Wilson 1995). The identification of species relies on the idea that the target sequence is found in a broad range of species, that sequences within a species will cluster more discreetly than those between species, i.e. that intraspecific divergences will be less than interspecific divergences (the barcode gap), and that this gap makes all animal species diagnosable. DNA barcoding has been used for the rapid assessment of diversity (e.g. Ward et al. 2005; Yu et al. 2012; Hebert et al. 2013), to identify larval stages (e.g. Webb et al. 2006), and for forensics (e.g. Dawnay et al. 2007). DNA barcodes are analogous to retail barcodes in that a species' genetic code at a standardised locus can be used as a species identifier (Hebert, Cywinska, et al. 2003). The process of DNA barcoding involves the amplification of a standard genetic region (COI for animals - Hebert, Cywinska, et al. 2003; *matK* and *rbcL* for plants - CBOL Plant Working Group 2009; ITS for fungi - Schoch et al. 2012), the comparison of this sequence to a database (Ratnasingham & Hebert 2007), and some degree of clustering based on nucleotide similarity (Hebert, Cywinska, et al. 2003).

How these DNA barcodes are obtained varies between different techniques and these tools can be tailored to the question at hand. The choice of these techniques depends on three things: 1) scale, 2) expense, and 3) resolution. Other than these three considerations, and perhaps more important, is the focal question; different techniques should be tailored to answer the question at hand. Here an overview is provided of the different techniques, the considerations to be made before choosing which technique to use (summarised in a flow chart), and finally a protocol to adopt including considerations on collecting, sample storage, and laboratory techniques.

S2.2 Different types of HTS

S2.2.1 Sanger sequencing

Sanger sequencing is the longest standing technique introduced here and is the process by which ~5 million of specimens have been uploaded onto BOLD (Barcode of Life Data systems). This technique does not involve using high throughput sequencing and so a much smaller output is reached using this method, but it has the benefit that long sequence reads (700 bp) at very low nucleotide error rate are produced. The process involves the separate processing of each individual through DNA extraction, PCR, sequencing and bioinformatic processing (Figure 9A). The end product is DNA barcode sequence associated independently with a specimen, and if the identity of the specimen is unknown then it can be retrospectively identified by comparing the focal sequence to a reference database.

1. Specimens are **sorted** into individual microtubes.
2. The investigator must choose how to extract the DNA from the specimen. The DNA extraction method can be non-destructive, semi-destructive or destructive. Non-destructive: Pierce the specimen (as would be done for pinning purposes) and place the whole specimen in the extraction buffer, DNA from within the specimen is extracted while the external tissues are undamaged and can be used for morphological taxonomy. Semi-destructive: Subsample part of the specimen, a single hind leg for insects is very common practise as the morphological characters can be investigated on the other hind leg. Here the specimen is not wetted, but the removed tissue can be pulverised to facilitate more DNA extraction. Destructive: This involves pulverising the whole specimen and extracting DNA from the homogenate. This is necessary if abundance/biomass estimates are required, for a quantitative DNA extraction.
3. **DNA extraction** was performed using a Qiagen DNeasy Blood and Tissue kit following the manufacturer's protocol.

4. Genetic **marker choice** will depend on what the target taxa is. Typically animals are barcoded for COI, plants are barcoded for *matK, rbcL* and increasingly ITS, fungi are barcoded for ITS and microbial diversity is assessed using 16S.
5. For insects, **PCR amplification** of COI was performed following the Canadian Centre for DNA Barcoding standardised techniques and reagents.
6. PCR amplicons are **purified** using Qiagen purification spin columns.
7. Purified PCR products are used in the **sequencing reaction** using BigDye Terminator v3.1 Cycle Sequencing Kit following the manufacturer's protocols.
8. Sequence reactions are purified before Sanger sequencing on an ABI 3770 automated sequencer.
9. Sequences are imported into **sequence editing software**, **assembled**, **checked** and **edited**.
10. Sequences can be **identified** by comparing them to existing sequences either by BLASTing them against the NCBI GenBank database, uploading to BOLD, or by locally BLASTing them against a *de novo* database. GenBank is not curated for taxonomy, while BOLD and your own reference library are likely to be more reliable. On the other hand, GenBank and BOLD are going to be more expansive than your own database.

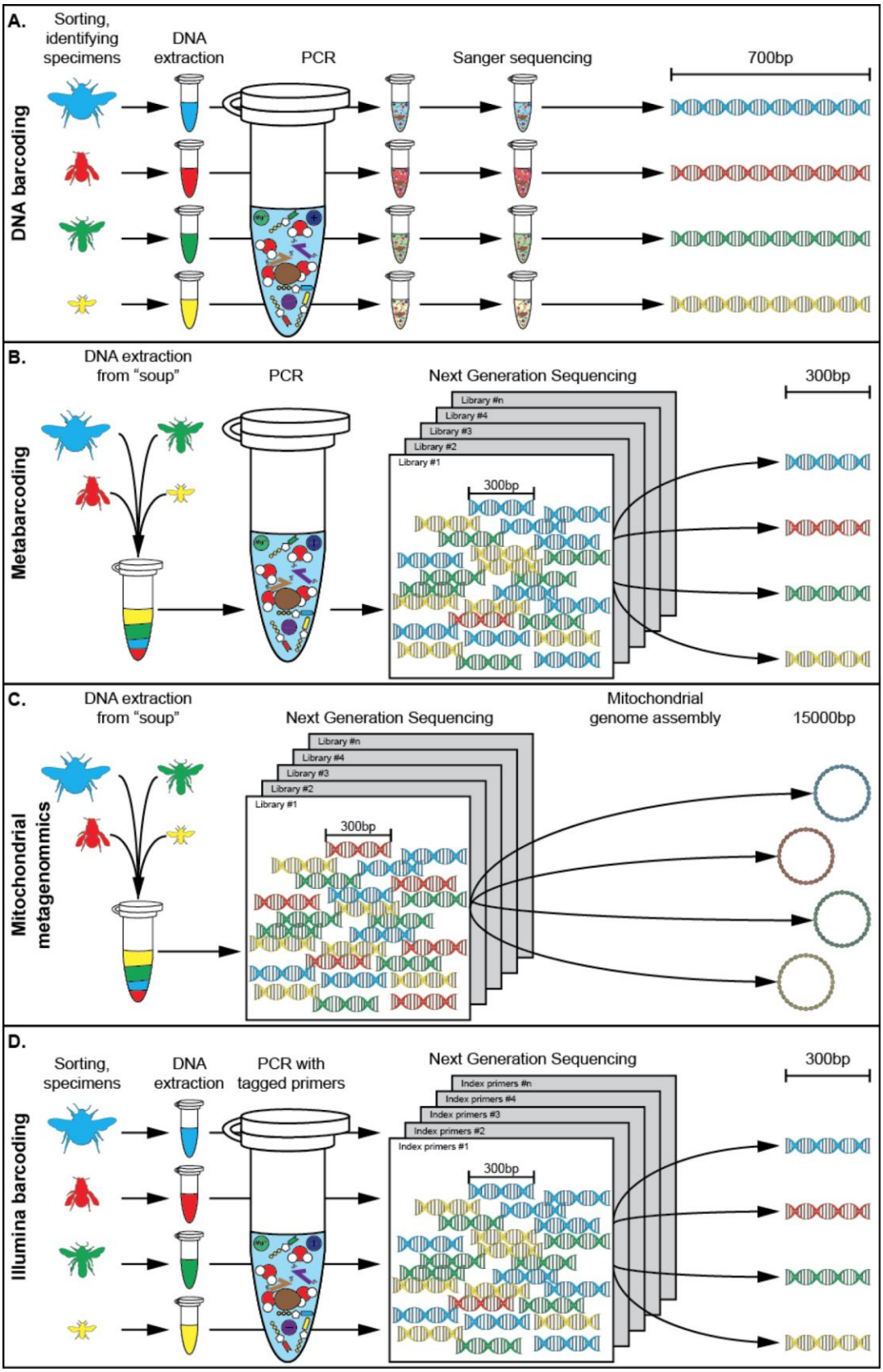
A.
Sorting, identifying specimens
DNA extraction
PCR
Sanger sequencing
700bp
DNA barcoding
B.
DNA extraction from "soup"
PCR
Next Generation Sequencing
300bp
Library #n
Library #4
Library #3
Library #2
Library #1
300bp
Metabarcoding
C.
DNA extraction from "soup"
Next Generation Sequencing
Mitochondrial genome assembly
15000bp
Library #n
Library #4
Library #3
Library #2
Library #1
300bp
Mitochondrial metagenommics
D.
Sorting, specimens
DNA extraction
PCR with tagged primers
Next Generation Sequencing
300bp
Index primers #n
Index primers #4
Index primers #3
Index primers #2
Index primers #1
300bp
Illumina barcoding

Figure 9. An overview of the four different molecular techniques described. A. Sanger sequencing: individuals are sorted and separate PCR and sequencing reactions are performed for each individual. The output is long sequences for each individual. B. Metabarcoding: DNA is extracted from a bulk. The target marker is amplified from the bulk DNA using primers with Illumina adaptors. Libraries are prepared for each bulk sample and sequenced on a next generation sequencer. The output, after bioinformatic processing, is a list of 300 bp sequences for each library, which can be concatenated with other sequences to form a list of 420 bp COI sequences. C. Mitochondrial metagenomics: DNA is extracted from a bulk. Libraries are prepared for each bulk DNA and shotgun sequenced on a next generation sequencer. The sequences are then assembled into mitochondrial genomes. The output, after bioinformatic processing, is a list of mitochondrial genomes for each sample. D. Illumina barcoding: individuals are sorted and separate PCRs are performed for each individual. Uniquely indexed primers are used every 96 individuals. PCR amplicons are then pooled and prepared into libraries such that each individual has been doubly tagged by primer indexes and library indexes. The DNA is once again pooled and sequenced on a next generation sequencer. The output, after bioinformatic processing, is list of 420 bp sequences for each individual specimen.

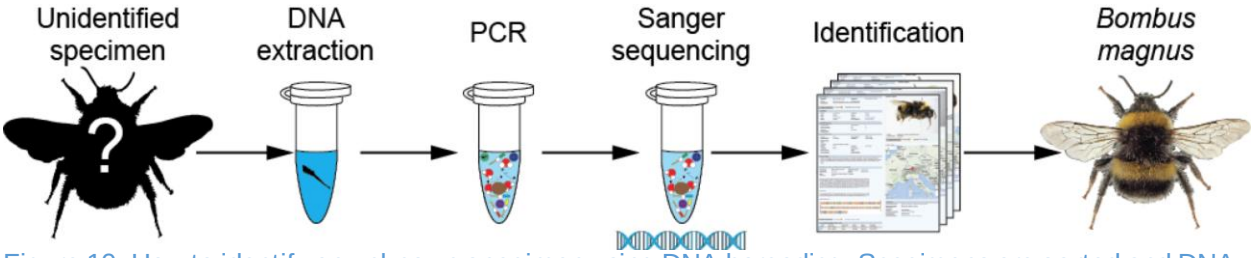

Figure 10. How to identify an unknown specimen using DNA barcoding. Specimens are sorted and DNA extracted from each individual. A target marker is PCR amplified for each DNA and sequenced using Sanger technology. The sequences are edited and quality controlled and then compared to a DNA reference library. The closest match to the sequence is used as it identification.

The benefits of this technique are first and foremost that it is relatively easy and that it requires equipment typically found in a molecular lab (e.g. a PCR machine). Sequences of up to 1,100 bp can be generated, which is currently substantially longer than those produced through Illumina systems (the most commonly used next generation sequencer). This longer sequence length provides more data to identify sequences and thus the reliability of these taxonomic placements is high. Because each specimen is handled separately for each process this allows one to link sequence to specimen and also a count after the sequences are used to identify the specimens.

Several practical limitations have been suggested that reduce the efficacy of COI-based DNA barcoding. Firstly, Sanger sequencing produces a single usable sequence per reaction, and so any intra-individual heterogeneity is either essentially missed (Shokralla et al. 2015) or turned into ambiguous regions of the sequence, resulting in poor quality sequence reads. Secondly, co-amplification and sequencing of pseudogenes (NuMTs - Song et al. 2008; Thalmann et al. 2004; Cristiano et al. 2012) or contaminant sequences (e.g. *Wolbachia* - Smith et al. 2012) can lead to misrepresentative focal barcodes. Similarly, introgressive hybridisation between species, the movement of whole genes, can lead to misrepresentative focal barcodes (Chase et al. 2005). Thirdly, the linear relationship between number of specimens and the cost of processing them make large scale investigations unfeasible and herculean in cost.

DNA barcoding was a transformative step in biodiversity science (Cristescu 2014); and field is now "on the brink of irrelevance" owing to the rapid development of high throughput sequencing technologies and the associated bioinformatic pipelines, computational infrastructure and experimental designs (Taylor & Harris 2012). The following three techniques involve the use of high throughput sequencing and each has their own advantages and considerations: in turn I will discuss metabarcoding, mitochondrial metagenomics and Illumina barcoding.

S2.2.2 Metabarcoding

Metabarcoding is similar to DNA barcoding using Sanger sequencing except that DNA from a community of specimens is the base input rather than a single individual. Using high throughput sequencing to generate many thousands times more output than Sanger sequencing, whole communities can be sequenced and characterised without the need to sort out individuals or clone PCR amplicons. The first step in the process involves extracting DNA from a community, which may be a trap of insects (e.g. Malaise traps - Yu et al. 2012; Ji et al. 2013), gut contents (e.g. fish guts - Leray et al. 2013), sediments (Giguet-Covex et al. 2014), honey (Richardson et al. 2015), leaf litter (Yang et al. 2014), soil cores (Andújar et al. 2015), etc. Secondly, the DNA is PCR amplified for a region of interest (depending on the Kingdom of interest, i.e. animals, plants, fungi, bacteria, etc.) in triplicate. Third, the pooled replicate PCR products are transformed into libraries with the appropriate adapters for the high throughput sequencer, further these libraries can be indexed such that different samples can be multiplexed and separated later on. These pooled libraries are then sequenced together generating millions of sequence reads. These sequence reads are bioinformatically quality controlled and filtered, and separated into the various different indexed libraries. Sequences from each library are then clustered together based on sequence similarity, and identified against a reference database. The end product is a list of operational taxonomic units (OTUs), which are a cluster of closely related sequences (i.e. 99% similar over the length of the sequence). A separate list is produced for each library.

Using the NPPMF sample as an example, each pan trap (n ≈ 280), could have been examined separately.

1. The entire contents of each pan trap would have been homogenised and DNA extracted (destructive DNA extraction). At the NHMUK we use liquid nitrogen to **embrittle** the samples and a separate ultra clean pestle and mortar (bleached, HCl acid treated, autoclaved) to **homogenise** each sample.
2. The resulting dust from the grinding is used as a base for a Qiagen DNeasy Blood and Tissue **DNA extraction** following the manufacturer's protocols. Or if enough samples are produced then by BioSprint 96.
3. Three separate dilutions (1, 1/10, 1/100) of each DNA extract is used for **PCR amplification** using modified short COI primers (BF and FolDegenR) and non-proof reading Taq (i.e. Takara).
4. The success of the PCR reactions is determined by **gel electrophoresis** using a 2% TAE gel run at 90V for 40 minutes.
5. The replicate PCR reactions are **pooled**, and **purified** using a low volume of AMPure XP beads. The protocol was modified according to the Illumina 16S rDNA size selection protocol.
6. Purified PCR products are once again checked by **gel electrophoresis** to confirm that primer dimers have been removed.
7. **Library preparation** involved a secondary PCR is performed using these purified PCR products to add the N5 and N7 indexes. These Nextera indexes are unique for each sample. Such that if 280 pan traps were samples, then 280 unique library indexes would be required.
8. These secondary PCRs are once again checked by **gel electrophoresis**.
9. Each library is **quantified** using Nanodrop and **pooled** into one pool in equimolar concentrations.
10. The final pool is **quality** checked and **quantified** by quantitative PCR and by BioAnalyser.
11. If the correct size fragments are present then between 2 and 3 pM of the final pool are loaded onto a flowcell to be **sequenced** on the MiSeq using 2×300 bp chemistry.
12. Once the run is complete, the libraries are **demultiplexed**, the poor quality sequences are **filtered** out, and then each library is **clustered** into OTUs, and each OTU is **identified** by comparing it to a reference database.

Metabarcoding is the most commonly used high throughput sequencing technique and has a rich catalogue of literature and bioinformatic pipelines. The ability to multiplex multiple samples allows for a very large scale. The Nextera library preparation kit we use allows for 96 indexes, but the latest version (v2) allows for 384 different combinations of indexes and therefore 384 samples can be multiplexed at once. The single DNA extraction for a pool is both an advantage and a disadvantage. On the one hand, a single DNA extraction for pools that can include 100s of individuals makes the process much cheaper, indeed the gold standard DNA extractions we perform are among the most costly processes in the whole workflow. On the other hand, there is no accountability for individual specimens. If DNA of a certain

species does not extract, amplify, or sequence then there is no way to track this down, failures are masked by successes in pooled processes.

It is becoming increasingly apparent that certain taxa amplify more successfully than others (Polz & Cavanaugh 1998), this is due to primer bias. Universal primers are not completely universal, and will have a higher affinity to certain taxa compared to others (Clarke et al. 2014), as such in pooled DNA extractions certain taxa will not amplify as well as others or in some cases not at all. This leads to a potential for false negatives, moreover, these false negatives are very difficult to actually spot without pre-sorting the samples (which defeats one of the main advantages of metabarcoding – that sorting is unnecessary). In theory if every single taxa in a pool amplified equally as well, then one would be able to use the number of sequences pertaining to each species as a measure of abundance (specifically if the number of sequences was calibrated with body sizes of those species). However, primer bias among taxa corrupts the sequence read – abundance relationship and so makes abundance measures using metabarcoding unreliable (Elbrecht & Leese 2015).

S2.2.3 Mitochondrial metagenomics

Mitochondrial metagenomics is a relatively new technique for assessing species diversity and potentially abundance. The first applications of the technique were Zhou et al. (2013) and Gillett et al. 2014 (2014), and the method is thoroughly described by Crampton-Platt et al. (2015). Metagenomics does not involve any PCR amplification, instead the entirety of the DNA extract is shotgun sequenced, which is different to the targeted sequencing of the other methods (i.e. metabarcoding and Illumina barcoding). Again, by utilising high through sequencing, millions of reads are generated, but here the sequences are bioinformatically stitched back together to form whole genomes. Mitochondrial metagenomics specifically filters out the mitochondrial genomes, which are already in high copy number owing to the makeup of animal genomes (typically ~1% of the genome). By pooling together multiple sources of DNA, or by extracting the DNA of whole communities, multiple mitochondrial genomes can be simultaneously sequenced and stitched together (hence the “meta” part of the method’s name).

The first step of the process is the **DNA extraction**, again, like metabarcoding, DNA from whole communities is extracted. Previous examples of this technique have included DNA extracted from whole pan traps (Tang et al. 2015), or from pooling DNA extracted from individuals (Gillett et al. 2014; Gómez-Rodríguez et al. 2015; Timmermans et al. 2015; Zhou et al. 2013; Andújar et al. 2015; Crampton-Platt et al. 2015). The sampling technique is very robust, whereby DNA from any bulk sample can be extracted and processed in the same way; i.e. freshwater samples, malaise traps, pitfall traps, soil pits, canopy fogging, etc.

The DNA is then prepared into TruSeq **libraries**, where the DNA is sheared into smaller inserts sizes (typically 800 bp). Separate libraries can be made with unique index tags for each sample, such that samples can be **multiplexed** and sequenced in the same sequencing run and bioinformatically separated afterwards. These libraries are then sequenced on a MiSeq.

After **sequencing** is complete, poor quality sequences are **filtered** out and various assembly software is used to generate long contigs. Mitochondrial genomes are typically ~15kb long and can be circularised. Various **assembly** software used include Celera, IDBA-UD, and Newbler. Once assembled these contigs from these various softwares are **concatenated** into “supercontigs”. These supercontigs are then **identified** using bait sequences, i.e. COI or CYTB sequences either generated *a priori* or from genetic repositories.

The end product is a list of mitochondrial genomes made per sample and identified based on bait sequences. A key advantage of this method is that whole mitochondrial genomes are generated and the massive length of these (×22 more data per specimen than Sanger, and ×37 more than metabarcoding and Illumina barcoding) results in far more data to identify species. Further, these genomes are especially good when trying to identify new species where data is not readily available. For example, these mitogenomes can be phylogenetically placed in a mitogenome phylogeny to accurately identify new species (Crampton-Platt et al. 2015).

1. The entire contents of each pan trap would have been homogenised and DNA extracted (destructive DNA extraction). At the NHMUK we use liquid nitrogen to **embrittle** the samples and a separate ultra clean pestle and mortar (bleached, HCl acid treated, autoclaved) to **homogenise** each sample.
2. The resulting dust from the grinding is weighed and the appropriate amount of extraction buffers (from Qiagen) are used to lyse and extract the DNA from the mixture. The extraction lysate is then used as the basis for Qiagen DNeasy Blood and Tissue **DNA extraction** following the manufacturer's protocols.
3. The DNA is then **sheared** into 800 bp insert size and **library preparation** follows the TruSeq Nano manufacturer's protocol. A separate library is generated from each sample.
4. Each library is **quantified** using Nanodrop and **pooled** into one pool in equimolar concentrations.
5. The final pool is **quality** checked and **quantified** by quantitative PCR and by BioAnalyser.
6. If the correct size fragments are present then between 2 and 3 pM of the final pool are loaded onto a flowcell to be **sequenced** on the MiSeq using 2×300 bp chemistry.
7. Once the run is complete, the libraries are **demultiplexed**, the poor quality sequences are **filtered** out. The mitochondrial sequences are retained by filtering out the nuclear sequences against a reference database of mitochondrial genomes.
8. Each mitochondrial library is **assembled** into larger contigs using Celera, IDBA-UD, and Newbler.
9. The contigs are then **concatenated** into supercontigs by assembling them in Geneious.
10. The identity of each mitochondrial genome is then determined by blasting the COI or CYTB segments of each genome to a reference database.

With a reference database of mitogenomes, whole communities can be readily and accurately identify by read matching (Figure 11). Here sequences from bulk DNA extracts can be matched onto an existing mitogenome reference database and the presence of specific species can be determined based on correct matching to those references. Moreover, the number of sequence reads, because there is no associated primer bias, can be used as a basis for abundance measures. Given that the amount of DNA extracted from an individual is predominantly determined by the body size of that individual, the number of sequences reads can be calibrated against biomass such that the abundance of those individuals can be calculated (Tang et al. 2015; Gómez-Rodríguez et al. 2015; Zhou et al. 2013).

1. With a complete mitochondrial reference database, the investigator can determine the species composition making up a bulk DNA extraction by read matching.
2. Here shotgun sequencing of bulk DNA extraction can be filtered and quality checked in the same way that the mitogenomes were generated. But here, instead of assembling the mitogenomes, the raw sequence reads can be matched against the reference genomes. The identity of these raw sequences can be determined in this way.
3. Further the number of sequences per species can be counted.

If the average biomass for the species is known than these can be calibrated such that the number of individuals can be determined afterwards (Tang et al. 2015).

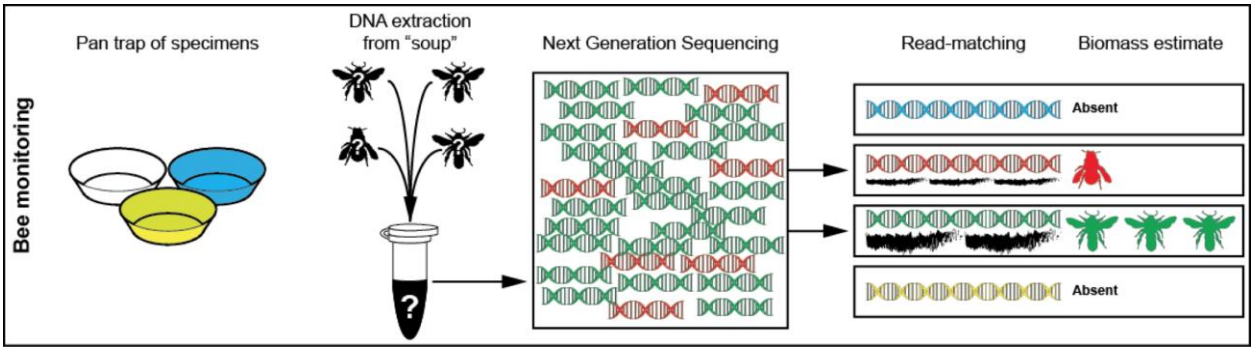

Figure 11. Read matching of mitochondrial genomes to identify specimens and estimate abundance. With a mitochondrial genome reference database, the investigator can take the raw output of a sequencing run and read match it against a reference database. If the sequences assemble with any particular species in the reference database, above a certain threshold, then their presence in the bulk sample can be confirmed. The amount of sequence reads coming from a particular species should relate to how much DNA that particular species had in the original pool. Using biomass estimate for each species it should be possible to calibrate the amount of sequence reads pertaining to that species with the number of individuals of that species in the bulk sample.

S2.2.4 Illumina barcoding

Illumina barcoding is the newest technique of those described here. Developed by Carmelo Andujar and Paula Arribas at the NHMUK, the method is analogous to DNA barcoding but utilising the massive capacity of high throughput sequencing.

1. Specimens are **sorted** into individual microtubes.
2. DNA extraction is non-destructive: Pierce the specimen (as would be done for pinning purposes) and place the whole specimen in the extraction buffer, DNA from within the specimen is extracted while the external tissues are undamaged and can be used for morphological taxonomy.
3. **DNA extraction** was performed using a Qiagen DNeasy Blood and Tissue kit following the manufacturer's protocol.
4. **Primers** were designed with unique tags and Illumina adaptors (See Table 1). For the NPPMF samples, 13 different primer sets were used such that for every 96 individuals a different set of primers were used. The primers amplify 420 bp from the 3' end of the the typical DNA barcode region. The primer comprises 26 bp of the priming region, a unique 6 bp index tag (with at least 3 bp difference between the different tags), and the Illumina adaptor sequences.
5. **PCR amplification** of COI was performed following reagents and cycling conditions outlined by Andujar and Arribas (pers. comm.).
6. **Gel electrophoresis** was used (2% TAE gel run at 90V for 40 minutes) to check the success of the PCR reaction.
7. PCR amplicons were pooled resulting in 96 different pools, each pool comprised of amplicons which were amplified with different primer sets. Therefore in the NPPMF example, 13 PCR products were in each pool. 3 µl of each PCR were pooled, this level can be altered such that weaker reactions should be pooled in higher volumes.
8. PCR amplicons are **purified** using AMPure XP beads, this method is the best for removing primer dimers, which if not removed are very problematic for the downstream processes. The amount of beads used to clean each reaction should be at ×0.8 of the volume of the PCR reaction.
9. The purified PCR products should be checked on a gel to determine if all of the primer dimers have been removed, and if not these will need to be re-purified.
10. **Library preparation** involved a secondary PCR is performed using these purified PCR products to add the N5 and N7 indexes. These Nextera indexes are unique for each pool. Here 96 different pools can be indexed using a single Nextera library preparation kit.
11. These secondary PCRs are once again checked by **gel electrophoresis**.
12. Each library is **quantified** using Nanodrop and **pooled** into one pool in equimolar concentrations.

13. The final pool is **quality** checked and **quantified** by quantitative PCR and by BioAnalyser.
14. If the correct size fragments are present then between 2 and 3 pM of the final pool are loaded onto a flowcell to be **sequenced** on the MiSeq using 2×300 bp chemistry.
15. Once the run is complete, the libraries are **demultiplexed**, the poor quality sequences are **filtered** out, and then each library is **clustered** into OTUs, and each OTU is **identified** by comparing it to a reference database.

The Illumina-generated sequences are a product of thousands of forward (R1) and reverse (R2) sequence reads, which account for intraspecific heterogeneity, pseudogenes, and contaminants. DNA barcoding by Sanger sequencing, on the other hand, is a product of a single sequencing reaction, which as described in the DNA barcoding section, is prone to sequencing error, sequence interpretation error, and potentially erroneous species identities amounting to contaminants or pseudogenes. Another key advantage of this technique is that the specimens are kept separate and as such (if non-destructive methods are used for DNA extraction) can be kept as voucher specimens. Furthermore, by keeping the individuals separate other aspects illuminated by their DNA can be studied. For example, microbiota, pollen, pathogens, diet can be investigated (Gill, K. C. R. R. Baldock, et al. 2016).

The main disadvantages of this method is the cost and time expense incurred from sorting individuals and extracting DNA from each of them, which may be prohibitively expensive with large samples. For example, a pan trap comprising of 100 individuals would require 100 DNA extractions, while metabarcoding or mitochondrial metagenomics would require only one.

S2.3 Considerations

S2.3.1 Data

The length of reads that each method provides is different. The greater the length of the focal sequence, the more data there is to reliably match and identify it to other reference sequences. For COI, the output from Sanger sequencing produces a single read of ~700 bp. High throughput sequencers as of today are limited to sequencing fragments of 300 bp, with paired end sequencing this can increase to 600 bp. For additional reliability, a 420 bp fragment of the 5' end of COI were amplified for both Illumina barcoding and metabarcoding, with paired end sequencing and concatenation of both forward (R1) and reverse (R2) directions this results in the full 420 bp sequence with a ~180 bp overlapping region in the middle (Figure 2E). Mitochondrial metagenomics relies on the assembly of shotgun sequences. DNA from the same species has a higher affinity that to those between species and so will assemble into species specific mitogenomes. Mitochondrial genomes are ~15kb and include 13 genes (including the full COI, not just the barcoding region) and 22 tRNAs. (roughly ×21 longer that DNA barcodes, and ×36 longer than barcodes coming from metabarcoding and Illumina barcoding).

Another consideration is the number of sequences that are produced per specimen. Sanger sequencing only produces a single sequence and any intra-individual heterogeneity is missed and contaminant sequences and NuMTs are more troublesome. High throughput sequencing produces many hundreds of sequences per individual specimen and so concatenation of these produces more reliable data with heterogeneity, but also the ability to bioinformatically choose the "true" barcode (Shokralla et al. 2015). Further the higher throughput allows for the recovery of low abundance sequences within each individual. Shokralla et al. (2015) noted that Sanger sequencing sequence was only ~55%, while that for the high throughput sequencing method they used (analogous to Illumina barcoding) being ~97%.

High throughput methods are more sensitive to low concentration and low quality DNA and thus are more likely to successfully generate sequences. While Sanger sequences produce longer sequences than the metabarcoding and Illumina barcoding approaches described here, they are only a single sequence and thus do not encompass heterogeneity and prone to sequence false barcodes like contaminants and NuMTs. The short sequences produced metabarcoding and Illumina barcoding are high quality with high coverage and enough to identify specimens to species-level. The massive mitogenomes produced by mitochondrial metagenomics are more than enough to identify species but also excellent for phylogenomic placement of new or unknown species (e.g. Crampton-Platt et al. 2015).

S2.3.2 Abundance

For ecological purposes, abundance is an important piece of information that links with function; for example, for pollination, total pollination depends on both species identity and the abundance of those pollinators (Figure 12). If abundance is a requirement of the study then certain methods are more appropriate than others: Sanger sequencing and Illumina barcoding can provide accurate abundance measures simply because they both require a lengthy sorting process to start with. Mitochondrial metagenomics has the potential for garnering abundance from pooled samples with no sorting or PCR amplification necessary. For this purpose however, mitochondrial metagenomics has had mixed results (e.g. Zhou et al. 2013; Tang et al. 2015; Gómez-Rodríguez et al. 2015) and would require thorough tests and method development before becoming a viable means to determine abundance from bulk samples. For example, no structured assessment the relationship between abundance and sequence read abundance has been performed, with all of the literature performing this test being a secondary proof of concept of the method rather than a test (e.g. Zhou et al. 2013; Tang et al. 2015; Gómez-Rodríguez et al. 2015). Furthermore, focussing on mitochondria limits the investigation to ~1% of the genome and would benefit from mitochondrial enrichment (Liu et al. 2015).

S2.3.3 Link to specimen

Only DNA barcoding and Illumina barcoding can utilise non-destructive or semi-destructive DNA extractions, while metabarcoding and mitochondrial metagenomics methods can be performed on individually extracted DNAs, the definitive link between specimen and sequence becomes a blurred by the end. Illumina barcoding and Sanger sequences are the only methods that can fully accommodate voucher specimens and the link to sequence.

Sanger and Illumina barcoding are better for tracking successes and failures, because the sequences link directly to specimens, if a sequence is missing then it is possible to determine which specimen it is. On the other hand, bulk extraction methods (metabarcoding and mitochondrial metagenomics) loose this link between specimens and so failed sequencing reactions are masked by the successful one. It is very difficult to determine if a particular specimen has not sequenced through the metabarcoding or mitochondrial metagenomics (false negatives), this issue is magnified if particular species are recalcitrant to sequencing because then whole taxa would be missing from the analysis.

S2.3.4 Scale

Scale and costs increase together, the bigger the scale the more these processes will cost, but the different technique have different trajectories. Sanger sequencing for example should only be used for smaller scale samples because the cost in time and expense of handling individuals separately will increase near linearly with scale. Each step of this process is reasonably expensive (our costs per sample = £9.75) and there is no chance of multiplexing samples to reduce the costs.

The other three methods, by incorporating high throughput sequencing technologies, are able to sequence many thousands more specimens. Moreover, because these methods produce multiple sequences per individual, intra-individual heterogeneity can be assessed and contaminants and NUMTs are easily distinguishable.

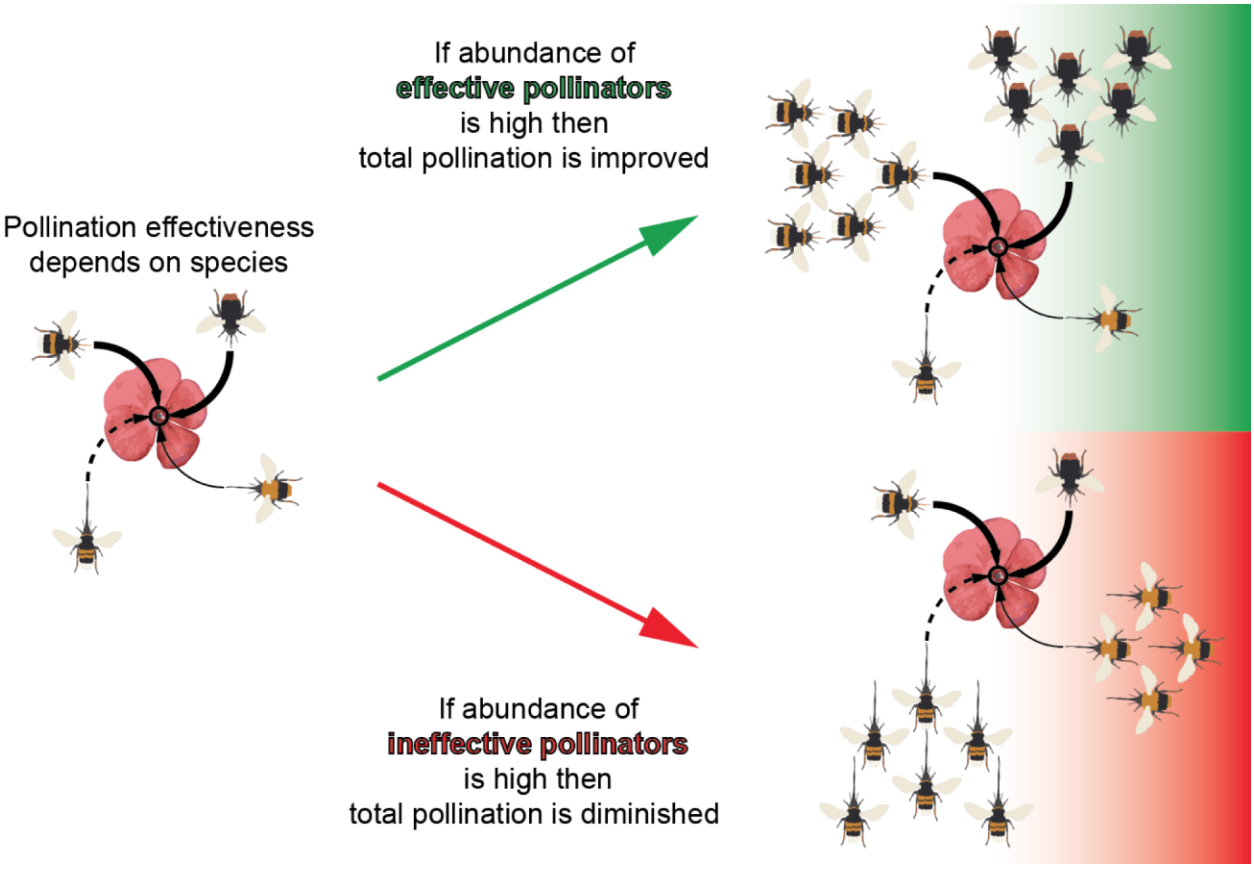

Figure 12. Pollinator effectiveness depends on both the efficiency of the pollinator but also the abundance of that pollinator. Pollination is most effective in the presence of efficient pollinators in high abundance. A high abundance of inefficient pollinators will lead to ineffective pollination. The high abundance of inefficient pollinators may be sufficient for the plant, this is also true for systems with a the low abundance of highly efficient pollinators.

Illumina barcoding is high throughput DNA barcoding, it still requires individual sorting of specimens, DNA extraction and individual PCRs, but with two levels of indexing it is possible to separate sequences coming from the high throughput sequencer back to the voucher specimens. By utilising uniquely indexed primers it is possible to pool as many PCR products as you have unique primers. In this study all of the specimens amounted to 1,248 individuals and so only 13 primer sets (1,248 ÷ 96 wells) were required, but many more can be multiplexed. One can expect 20gb of data from a high throughput sequencer, and with this 48 plates worth of specimens (4,608 individuals) can be sequenced with 10X coverage at a cost of £3.31 per sample. The vast majority of this cost is for DNA extraction (£2.80), where vastly cheaper (not gold standard) methods are available. For 4,608 individuals the total cost would be £15,252, the same endeavour using Sanger sequencing would cost £44,928. The tipping point at which Illumina barcoding becomes cheaper than Sanger sequencing is ~718 individuals, which would cost approximately £7,000 to sequence either way. Of course these estimates are based on all reactions working perfectly. Sanger sequencing is notoriously fickle and will require high concentrations of DNA at each step in the process, while high throughput sequencing methods are much more forgiving and can work with much lower concentrations of DNA, resulting in fewer repeated reactions.

Under the different scenarios outlined by Carvell et al. (2016), the high level of sampling makes identifying them using Illumina barcoding cheaper than Sanger sequencing, irrespective of the level of sensitivity required (Table 6).

Table 6. Expected consumable costs of molecular identification (Sanger, Illumina barcoding, metabarcoding, mitochondrial metagenomics) from different NPPMF scenarios employed to detect change in pollinator abundance at different levels of sensitivity. DB = DNA barcoding, IB = Illumina barcoding, MBC = Metabarcoding, MMG = Mitochondrial metagenomics.

# Sites	Annual change	Total catch	DB	IB	MBC	MMG
145	3.5%	9,425	£92K	£6.4K	£3.9K	£69K
75	3.5%	4,875	£48K	£4.8K	£1.9K	£20K
45	7%	2,925	£29K	£3.2K	£1.8K	£12K
20	7%	1,300	£13K	£3.1K	£1.7K	£6K

Mitochondrial metagenomics and metabarcoding significantly reduce the number of DNA extractions required as the DNA is extracted from bulk samples rather than sorted individuals. These methods are the most cost effective when abundance and diversity within samples is high. Metabarcoding is the most suitable method when the number of sample and the abundance and diversity within those is high. For example, the method has been used successfully to assess tropical samples (Crampton-Platt et al. 2015), in soils (Andújar et al. 2015).

While mitochondrial metagenomics can be used against bulk DNA extractions, only few studies take advantage of this (e.g. Tang et al. 2015), others have extracted DNA from individuals and pooled the DNA rather than the individuals (e.g. Crampton-Platt et al. 2015; Gillett et al. 2014; Gómez-Rodríguez et al. 2015; Andújar et al. 2015). Each bulk sample would be processed in the same way as for metabarcoding. The number of specimens pooled in each bulk is the concern, it is unknown how many false negatives are produced by each bulk sampling method, but fewer specimens in each pool is likely to produce fewer false negatives. There may be an upper limit to the number of specimens placed in each pool but the tipping point at which more false negatives are produced is unknown. A reasonable maximum number of specimens that should be added to a single MiSeq run would be 500 (Crampton-Platt pers. comm.). This number of specimens is based on the number that is required to obtain full mitochondrial assemblies for each species in the pool, but many more specimens can be pooled if read matching against a reference dataset. With a full reference database, the cost of read matching is considerably cheaper than assembling the mitochondrial genomes.

S2.3.5 Recommendations

The choice of method will depend on the aforementioned considerations (Table 7). Probably the most telling factor is the scale of the study. Smaller studies (here 700 individuals being the tipping point in cost) can more feasibly be sequenced using Sanger technology. However, this recommendation does not take into account the added costs associated with having to repeat Sanger sequencing, which is notoriously more fickle and less sensitive than the high throughput technologies. Further Sanger sequencing cannot assess intra-individual heterogeneity and, because it provides only a single sequence (as opposed to the concatenation of hundreds of reads) is more badly affected by contaminant amplicons and NuMTs. For example, all of the *Hylaeus communis* (n = 3) specimens sequenced as part of the reference library here, and their repeats, returned as some form of *Wolbachia* contamination: the likely cause of which is higher primer affinity for *Wolbachia* in this species. While abundance and a link to specimen can be attained through Sanger sequencing, the method should only really be performed when the sample size is sufficiently low to discount the much greater benefits of using high throughput sequencing.

Within the high throughput sequencing methods there is another level of scale. Illumina barcoding can feasibly be done for 1000's of specimens in terms of time, but the cost of individual DNA extractions becomes decidedly large at the top end. This gold standard of sequencing may cost more in terms of time and expense, but benefits from being able to identify individuals, detect false negatives, give a definitive abundance, and provide a DNA resource for each individual that can be used for further analyses such as

diet (through gut contents analysis), foraging behaviour (through pollen analysis), health (by assessing parasites and pathogens), and even microbial diversity.

For these NPPMF samples the initial plan was to use a combination of metabarcoding and mitochondrial metagenomics, but the type of samples made these methods less appropriate. These methods should be reserved for massive scale experimental designs where the number of samples is high and the diversity and abundance within them is large. For example these methods would have both been appropriate for the NPPMF samples if there was no pre-sorting of the pan traps and all of the by catch was assessed as well as the bees and hoverflies, the total catch would have equated to 280 pan traps with well over 20,000 specimens (calculated based on the 1,660 specimens counted from one Summer round and 15 pan traps). Having said that, for mitochondrial metagenomics, fewer specimens can be pooled on a single MiSeq run to fully assemble the species' genomes. Many more specimens can be pooled and sequenced if read matching against a fully comprehensive mitogenome dataset.

After scale, both abundance and link to specimen are two considerations of importance. If a link to specimen is important, then neither metabarcoding nor mitochondrial metagenomics should not be used as the link is effectively lost when DNA is pooled or bulk extracted. The extra work and expense incumbent with Illumina barcoding result in a definitive link between specimen and sequence. Similarly, this sorting allows for definitive abundance measures as well. Estimating abundance from metabarcoding experiments is ill advised (Amend et al. 2010; Tang et al. 2015; Gómez-Rodríguez et al. 2015; Elbrecht & Leese 2015). While it is potentially possible to obtain abundance measure from mitochondrial metagenomics, the technique and the calibrations are still their infancy, moreover this measure of abundance is a correlation and requires that no specimens drop out from the pipeline. Therefore if abundance and a link from sequence to specimen is important than Illumina barcoding should be used.

Table 7. Summary of the important factors to consider when deciding which method is appropriate for your investigation.

Method	Scale	Data	Abundance	Link to specimen
DNA barcoding	Small	Single read – ~700 bp	Yes	Yes
Illumina barcoding	Medium	Hundreds of reads – 420 bp	Yes	Yes
Metabarcoding	Medium – very large	Hundreds of reads – 420 bp	No	No
Mitochondrial metagenomics	Medium	Hundreds of assembled reads – 15 kb	Potentially	No

Next Generation Pollinator Identification using High Throughput Sequencing (PH0521)

Supplementary Text 2. Literature references used in this report

Authors: Cuong Q. Tang[1,*], David G. Notton[1], Hannah Norman[1,2], Alfried P. Vogler[1,2]

[1]Department of Life Sciences, The Natural History Museum, London, UK

[3]Department of Life Sciences, Imperial College London, London, UK

*present address: Nature Metrics Ltd, Bakeham Lane, CABI, Egham, Surrey, TW20 9TY

ct@naturemetrics.co.uk

References

Ahrens, D., Monaghan, M.T. & Vogler, A.P., 2007. DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae). *Molecular Phylogenetics and Evolution*, 44(1), 436–49.

Allen-Wardell, G. et al., 1998. The potential consequences of pollinator declines on the conservation of biodiversity and stability of food crop yields. *Conservation Biology*, 12, 8–17.

Amend, A.S., Seifert, K.A. & Bruns, T.D., 2010. Quantifying microbial communities with 454 pyrosequencing: Does read abundance count? *Molecular Ecology*, 19(24), 5555–5565.

Amiet, F. et al., 2010. Apidae. 6 - *Andrena*, *Melitturga*, *Panurginus*, *Panurgus*. *Fauna Helvetica*, 26, 1–317.

Amiet, F. et al., 2001. Apidae 3 - *Halictus*, *Lasioglossum*. *Fauna Helvetica*, 6, 1–208.

Amiet, F. et al., 2004. Apidae 4 - *Anthidium*, *Chelostoma*, *Coelioxys*, *Dioxys*, *Heriades*, *Lithurgus*, *Megachile*, *Osmia*, *Stelis*. *Fauna Helvetica*, 9, 1–273.

Amiet, F. et al., 2007. Apidae 5 - *Ammobates*, *Ammobatoides*, *Anthophora*, *Biastes*, *Ceratina*, *Dasypoda*, *Epeoloides*, *Epeolus*, *Eucera*, *Macropis*, *Melecta*, *Melitta*, *Nomada*, *Pasites*, Tet. *Fauna Helvetica*, 20, 1–356.

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

Andújar, C. et al., 2015. Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Molecular Ecology*, 24(14), 3603-3617.

Austen, G. E., Bindemann, M., Griffiths, R. A. & Roberts, D. L. 2016. Species identification by experts and non-experts: comparing images from field guides. *Scientific Reports* 6, Article number: 33634.

Baird, D.J. & Hajibabaei, M., 2012. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, 21(8), 2039–2044.

Ball, S. & Morris, R., 2015. *Britain's Hoverflies: A Field Guide* 2nd ed., Woodstock: Princeton University Press.

Benton, T., 2006. *Bumblebees: The natural history & identification of the species found in Britain*, London: Collins.

Biesmeijer, J.C. et al., 2006. Parallel declines in pollinators and insect-pollinated plants in Britain and the Netherlands. *Science*, 313, 351–354.

Bik, H.M., Sung, W., et al., 2012. Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Molecular Ecology*, 21, 1048–1059.

Bik, H.M., Porazinska, D.L., et al., 2012. Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, 27(4), 233–243.

Birky, C.W. et al., 2010. Using population genetic theory and DNA sequences for species detection and identification in asexual organisms. *PLoS ONE*, 5(5), e10609.

Blaxter, M.L., 2004. The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, 359, 669–79.

Bogusch, P. & Straka, J., 2012. Review and identification of the cuckoo bees of central Europe (Hymenoptera: Halictidae: *Sphecodes*). *Zootaxa*, 3311, 1–41.

Bohmann, K. et al., 2014. Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution*, 29(6), 358–367.

Bossert, S. 2015. Recognition and identification of bumblebee species in the *Bombus lucorum*-complex (Hymenoptera, Apidae) – A review and outlook. *Deutsche Entomologische Zeitschrift*, 62(1), 19-28.

Brown, S.D.J. et al., 2012. SPIDER: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources*, 12(3), 562–5.

Burkle, L.A., Marlin, J.C. & Knight, T.M., 2013. Plant-pollinator interactions over 120 years: loss of species, co-occurrence, and function. *Science*, 339, 1611–5.

Carolan, J.C. et al., 2012. Colour patterns do not diagnose species: quantitative evaluation of a DNA barcoded cryptic bumblebee complex. *PloS one*, *7*(1), e29251.

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

CBOL Plant Working Group, 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 106(31), 12794–7.

Chamberlain, S., 2014. bold: Interface to Bold Systems API. https://github.com/ropensci/bold.
Chapin III, F.S. et al., 1997. Biotic Control over the Functioning of Ecosystems. *Science*, 277, 500–504.
Chariton, A.A. et al., 2010. Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment*, 8(5), 233–238.
Chase, M.W. et al., 2005. Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, 360(1462), 1889–95.
Clarke, L.J. et al., 2014. Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular ecology resources*, 14(6), 1160–70.
Collins, R.A. et al., 2012. Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. *Methods in Ecology and Evolution*, 3(3), 457–465.
Collins, R.A. & Cruickshank, R.H., 2012. The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, 13(6), 969–975.
Crampton-Platt, A. et al., 2015. Soup to tree: the phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Molecular Biology and Evolution*, 32(9), 2302-2316.
Creer, S. et al., 2010. Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology*, 19(Supplement 1), 4–20.
Cristescu, M.E., 2014. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, 1–6.
Cristiano, M.P., Fernandes-Salomão, T.M. & Yotoko, K.S.C., 2012. Nuclear mitochondrial DNA: An Achilles' heel of molecular systematics, phylogenetics, and phylogeographic studies of stingless bees. *Apidologie*, 43, 527–538.
Cross, I. & Notton, D. G. in press. Small-headed resin bee, *Heriades rubicola*, new to Britain (Hymenoptera: Megachilidae). British Journal of Entomology and Natural History.
Danforth, B., 2007. Bees. *Current Biology*, 17(5), 156–161.
Danforth, B.N., 1999. Phylogeny of the Bee Genus Lasioglossum (Hymenoptera : Halictidae) Based on Mitochondrial COI Sequence Data. *Systematic Entomology*, 24(4), 377–393.
Danforth, B.N., Mitchell, P.L. & Packer, L., 1998. Mitochondrial DNA differentiation between two cryptic *Halictus* (Hymenoptera: Halictidae) species. *Annals of the Entomological Society of America*, 91(4), 387–391.
Dawnay, N. et al., 2007. Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic science international*, 173(1), 1–6.
Defra, 2014. *The National Pollinator Strategy: for bees and other pollinators in England*,
Dinca, V. et al., 2011. Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe. *Proceedings of the Royal Society B: Biological Sciences*, 278(1704), 347–55.
Doczkal, D. 2000. Description of *Cheilosia ranunculi* spec. nov. from Europe, a sibling species of *C. albitarsis* Meigen (Diptera, Syrphidae). *Volucella* 5, 63-78.
Drummond, A.J. & Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7, 214.
Eilers, E.J. et al., 2011. Contribution of pollinator-mediated crops to nutrients in the human food supply. *PLoS ONE*, 6(6).
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. *PeerJ*, (Preprint).
Elphick, C.S. 2008. How you count counts: the importance of methods research in applied ecology. *Journal of Applied Ecology*, 45, 1313–1320.
Else, G. R., Bolton, B. & Broad, G. R. 2016. Checklist of British and Irish Hymenoptera – aculeates (Apoidea, Chrysidoidea and Vespoidea). *Biodiversity Data Journal*, 4, 1–188.
Falk, S.J. & Lewington, R., 2015. *Field guide to the bees of Great Britain and Ireland*, London, UK: Bloomsbury Publishing Plc.
Fonseca, V.G. et al., 2014. Metagenetic analysis of patterns of distribution and diversity of marine meiobenthic eukaryotes. *Global Ecology and Biogeography*, 23, 1293–1302.
Fontaine, C. et al., 2006. Functional diversity of plant-pollinator interaction webs enhances the persistence of plant communities. *PLoS Biology*, 4(1), 0129–0135.

Frankie, G.W. et al., 2002. Monitoring: an Essential Tool in Bee Ecology and Conservation. In K. P & I. F. VL, eds. *Pollinating Bees - The Conservation Link Between Agriculture and Nature*. Brasilia: Ministry of Environment, 187–198.

Free, J., 1993. *Insect Pollination of Crops.*, Academic Press, London.

Fujisawa, T. & Barraclough, T.G., 2013. Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent (GMYC) approach: a revised method and evaluation on simulated datasets. *Systematic Biology*, 62(5), 707–724.

Garibaldi, L.A. et al., 2013. Wild pollinators enhance fruit set of crops regardless of honey bee abundance. *Science*, 339, 1608–1611.

Gezon, Z.J. et al., 2015. The effect of repeated, lethal sampling on wild bee abundance and diversity. *Methods in Ecology and Evolution*, 6(9), 1044-1054.

Gibson, J.F. et al., 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 of the United States of America*, 111(22), 8007–12.

Giguet-Covex, C. et al., 2014. Long livestock farming history and human landscape shaping revealed by lake sediment DNA. *Nature communications*, 5, 3211.

Gill, R.J., Baldock, K.C.R., et al., 2016. Protecting an ecosystem service: approaches to understanding and mitigating specific threats to wild insect pollinators. *Advances in Ecological Research*, 54, 135–206.

Gill, R.J., Baldock, K.C.R.R., et al., 2016. Protecting an ecosystem service: approaches to understanding and mitigating specific threats to wild insect pollinators. *Advances in Ecological Research*, 54, 135–206.

Gill, R.J., Baldock, K.C.R., et al., 2016. Protecting an Ecosystem Service: Approaches to Understanding and Mitigating Threats to Wild Insect Pollinators. *Advances in Ecological Research*, 54, 135-206.

Gillett, C.P.D.T. et al., 2014. Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, 31(8), 2223–2237.

Glasel, J.A., 1995. Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios. *BioTechniques*, 18(1), 62–63.

Gómez-Rodríguez, C. et al., 2015. Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, 6(8), 883–894.

Gonzalez, V.H., Griswold, T. & Engel, M.S., 2013. Obtaining a better taxonomic understanding of native bees: Where do we start? *Systematic Entomology*, 38(4), 645–653.

Goulson, D. et al., 2015. Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. *Science*, 2010, 1–16.

Goulson, D. & Williams, P., 2001. *Bombus hypnorum* (L.)(Hymenoptera: Apidae), a new British bumblebee? *British Journal of Entomology and Natural History*, (1973), 1–3.

Hänfling, B., et al. 2016. Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, 25(13), 3101-3119.

Hebert, P.D.N. et al., 2013. A DNA “Barcode Blitz”: Rapid Digitization and Sequencing of a Natural History Collection. *PLoS ONE*, 8(7).

Hebert, P.D.N., Cywinska, A., et al., 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270(1512), 313–321.

Hebert, P.D.N. et al., 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America*, 101(41), 14812–7.

Hebert, P.D.N., Ratnasingham, S. & DeWaard, 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, S96-9.

Higuchi, R. et al., 1988. DNA typing from single hairs. *Nature*, 332, 543–546.

Illumina Inc., 2013. 16S Metagenomic Sequencing Library. *Illumina.com*.

Ji, Y. et al., 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology letters*, 16(10), 1245–57.

Jukes, A. 2016. *Andrena nigrospina* (Thomson, 1872) and *Andrena pilipes* (Fabricius, 1781): where we are now. BWARS Newsletter, 2016 (Autumn), 15-17.

Katoh, K., Asimenos, G. & Toh, H., 2009. Multiple alignment of DNA sequences with MAFFT. *Methods in Molecular Biology*, 537, 39–64.
Kearns, C.A. & Inouye, D.W., 1998. Endangered Mutualisms: The Conservation of Plant-Pollinator Interactions. *Annual Review of Ecology and Systematics*, 29, 83–112.
Kearse, M. et al., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647–1649.
Keller, A. et al. 2015. Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. *Plant Biology*, 17, 558–66.
Kirby-Lambert, C. 2016. *Nomada alboguttata* Herrich-Schäffer, 1839 new to the British Isles and *Nomada zonata* Panzer, 1798 first record for mainland Britain. BWARS Newsletter, 2016 (Autumn), 29-30.
Kleijn, D. et al., 2015. Delivery of crop pollination services is an insufficient argument for wild pollinator conservation. *Nature Communications*, 6(May), 7414.
Kocher, S.D. et al., 2013. The draft genome of a socially polymorphic halictid bee, *Lasioglossum albipes*. *Genome biology*, 14(12), R142.
Krüger, E. 1951. Phänoanalytische Studien an einigen Arten der Untergattung *Terrestribombus* O. Vogt (Hymen. Bomb.). I. Teil. *Tijdschrift voor Entomologie*, 93(1950): 141-197.
Kuhlmann, M. et al., 2007. Molecular, biogeographical and phenological evidence for the existence of three western European sibling species in the *Colletes succinctus* group (Hymenoptera: Apidae). *Organisms Diversity and Evolution*, 7(2), 155–165.
Kurtzman, C.P., 1994. Molecular taxonomy of the yeasts. *Yeast*, 10(13), 1727–1740.
Lebuhn, G. et al., 2013. Detecting Insect Pollinator Declines on Regional and Global Scales. *Conservation Biology*, 27(1), 113–120.
Lebuhn, G. et al., 2015. Evidence-based conservation: reply to Tepedino et al. *Conservation Biology*, 29(1), 283–285.
Leray, M. et al., 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in zoology*, 10(1), 34.
Linard, B. et al., 2015. Metagenome skimming of insect specimen pools: potential for comparative genomics. *Genome Biology and Evolution*, 7(6), 1474–1489.
Liu, S. et al., 2015. Mitochondrial capture enriches mito-DNA 100 folds enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, 16(2), 470-479.
Løken, A. 1973. Studies on Scandinavian bumble bees (Hymenoptera, Apidae). *Norsk entomologisk Tidsskrift*, 20, 1-218.
Lovett, G.M. et al., 2007. Who needs environmental monitoring? *Frontiers in Ecology and the Environment*, 5(5), 253–260.
Magnacca, K.N. & Brown, M.J.F., 2012. DNA barcoding a regional fauna: Irish solitary bees. *Molecular Ecology Resources*, 12(6), 990–998.
Martins, A.C., Goncalves, R.B. & Melo, G. a R., 2013. Changes in wild bee fauna of a grassland in Brazil reveal negative effects associated with growing urbanization during the last 40 years. *Zoologia*, 30(2), 157–176.
Meeus, I. et al., 2015. 16S rRNA amplicon sequencing demonstrates that indoor-reared bumblebees (*Bombus terrestris*) harbor a core subset of bacteria normally associated with the wild host. *PLoS ONE*, 10(4), 1–15.
Meier, R. et al., 2015. $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. , 1–11.
Meier, R. et al., 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, 55(5), 715–28.
Michener, C.D., 2007. *The bees of the world*, Baltimore, Md, USA: Johns Hopkins University Press.
Nichols, J.D. & Williams, B.K., 2006. Monitoring for conservation. *Trends in Ecology and Evolution*, 21(12), 668–673.
Nieto, A. et al., 2014. *European Red List of Bees*, Luxembourg: Publication Office of the European Union.
Notton, D.G., Tang, C.Q. & Day, A.R. 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.
Novacek, M.J., 2008. Engaging the public in biodiversity issues. *Proceedings of the National Academy of Sciences of the United States of America*, 105 Suppl., 11571–11578.

Obrist, M.K. & Duelli, P., 2010. Rapid biodiversity assessment of arthropods for monitoring average local species richness and related ecosystem services. *Biodiversity and Conservation*, 19(8), 2201–2220.
Ollerton, J. et al., 2014. Extinctions of aculeate pollinators in Britain and the role of large-scale agricultural changes. *Science*, 346(6215), 1360–1362.
Orford, K.A., Vaughan, I.P. & Memmott, J., 2015. The forgotten flies: the importance of non-syrphid Diptera as pollinators. *Proceedings of the Royal Society B: Biological Sciences*, 282(20142934).
Packer, L. et al., 2009. DNA barcoding and the mediocrity of morphology. *Molecular Ecology Resources*, 9 Suppl s1, 42–50.
Padial, J.M., Miralles, A., De la Riva, I.J., Vences, M. 2010. The integrative future of taxonomy. *Frontiers in Zoology*, 7, 16.
Paradis, E., Claude, J. & Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20(2), 289–290.
Pawlowski, J,, Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J., & Esling, P. 2016. Protist metabarcoding and environmental biomonitoring: Time for change. *European Journal of Protistology*, 55(A), 12-25.
Paxton, R. et al., 2015. The bee-all and end-all. *Nature*, 521, S57–S59.
Pekkarinen, A. 1979. Morphometric, colour and enzyme variation in bumblebees (Hymenoptera, Apidae, *Bombus*) in Fennoscandia and Denmark. *Acta zoologica fennica*, 158: 60.
Polz, M.F. & Cavanaugh, C.M., 1998. Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, 64(10), 3724–3730.
Pompanon, F. et al., 2012. Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, 21, 1931–1950.
Potts, S.G. et al., 2010. Global pollinator declines: Trends, impacts and drivers. *Trends in Ecology and Evolution*, 25(6), 345–353.
Proctor, M., Yeo, P. & Lack, A., 1997. The Natural History of Pollination. *Ecology*, 78(1), 327–328.
Puillandre, N. et al., 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, 21(8), 1864–1877.
Pywell, R.F. et al., 2012. Wildlife-friendly farming benefits rare birds, bees and plants. *Biology Letters*, 8, 772–775.
R Core Team, 2014. *R: A language and environment for statistical computing. R Core Team*, Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org.
Rader, R. et al., 2015. Non-bee insects are important contributors to global crop pollination. *Proceedings of the National Academy of Sciences*, 113(1), 146–151.
Rasmont, P., Scholl, A., Dejonghe, R., Obrecht, E., Adamski, A. 1986. Identification and variability of males of the genus *Bombus latreille* sensu-stricto in Western and Central-Europe (Hymenoptera, Apidae, Bombinae). *Revue Suisse De Zoologie*, 93(3), 661-682.
Ratnasingham, S. & Hebert, P.D.N., 2013. A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. *PLoS ONE*, 8(7), e66213.
Ratnasingham, S. & Hebert, P.D.N., 2007. BOLD : The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, 7, 355–364.
Raupach, M.J. et al., 2014. Building-up of a DNA barcode library for true bugs (insecta: hemiptera: heteroptera) of Germany reveals taxonomic uncertainties and surprises. *PloS one*, 9(9), e106940.
Richardson, R.T. et al., 2015. Application of ITS2 Metabarcoding to Determine the Provenance of Pollen Collected by Honey Bees in an Agroecosystem. *Applications in Plant Sciences*, 3(1), 1–6.
Rotheray, G.E., 1993. *Colour guide to hoverfly larvae (Diptera, Syrphidae)*, Derek Whiteley.
Rundlöf, M. et al., 2015. Seed coating with a neonicotinoid insecticide negatively affects wild bees. *Nature*, 0, 1–7.
Saunders, M.E., Luck, G.W. & Gurr, G.M., 2015. Keystone resources available to wild pollinators in a winter-flowering tree crop plantation. *Agricultural and Forest Entomology*, 17, 90–101.
Saunders, M.E., Luck, G.W. & Mayfield, M.M., 2013. Almond orchards with living ground cover host more wild insect pollinators. *Journal of Insect Conservation*, 17(5), 1011–1025.
Schindel, D.E. & Miller, S.E., 2005. DNA barcoding a useful tool for taxonomists. *Nature*, 435(7038), 17.
Schmieder R. & Edwards R., 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27 863-864.

Schmidt, S., Schmid-egger, C. & Ere, M., 2015. DNA barcoding largely supports 250 years of classical taxonomy: identifications for Central European bees (Hymenoptera, Apoidea partim. *Molecular Ecology Resources*, 15(4), 985–1000.

Schoch, C.L. et al., 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, 109(16), 6241–6.

Senapathi, D. et al., 2015. The impact of over 80 years of land cover changes on bee and wasp pollinator communities in England. *Proceedings of the Royal Society B: Biological Sciences*, 282, 20150294.

Sheffield, C.S. et al., 2009. DNA barcoding a regional bee (Hymenoptera: Apoidea) fauna and its potential for ecological studies. *Molecular ecology resources*, 9 Suppl s1, 196–207.

Shokralla, S. et al., 2015. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, 5(9687), 1–7.

Shokralla, S. et al., 2014. Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, 14(5), 892–901.

Shokralla, S. et al., 2012. Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, 21(8), 1794–805.

Sirohi, M.H. et al., 2015. Diversity and abundance of solitary and primitively eusocial bees in an urban centre: a case study from Northampton (England). *Journal of Insect Conservation*, 19(3), 487–500.

Smith, M.A. et al., 2012. Wolbachia and DNA barcoding insects: patterns, potential, and problems. *PLoS ONE*, 7(5), e36514.

Song, H. et al., 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United States of America*, 105(36), 13486–13491.

Ståhls, G. et al., 2009. COI barcodes for identification of *Merodon* hoverflies (Diptera, Syrphidae) of Lesvos Island, Greece. *Molecular Ecology Resources*, 9, 1431–1438.

Steffan-Dewenter, I., Potts, S.G. & Packer, L., 2005. Pollinator diversity and crop pollination services are at risk. *Trends in Ecology and Evolution*, 20(12), 651–653.

Stein, E.D. et al., 2014. Is DNA barcoding actually cheaper and faster than traditional morphological methods: Results from a survey of freshwater bioassessment efforts in the United States? *PLoS ONE*, 9(4), e95525.

Tang, C.Q. et al., 2014. Effects of phylogenetic reconstruction method on the robustness of species delimitation using single locus data. *Methods in Ecology and Evolution*, 5, 1086–1094.

Tang, C.Q. et al., 2012. The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 16208–16212.

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

Tautz, D. et al., 2003. A plea for DNA taxonomy. *Trends in Ecology & Evolution*, 18(2), 70–74.

Taylor, H.R. & Harris, W.E., 2012. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, 12(3), 377–88.

Teletchea, F. 2010. After 7 years and 1000 citations: Comparative assessment of the DNA barcoding and the DNA taxonomy proposals for taxonomists and non-taxonomists. *Mitochondrial DNA*, 21(6), 206–226.

Tepedino, V.J. et al., 2015. Documenting bee decline or squandering scarce resources. *Conservation Biology*, 29(1), 280–282.

Thalmann, O. et al., 2004. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. *Molecular Ecology*, 13(2), 321–335.

Timmermans, M.J.T.N. et al., 2015. Rapid assembly of taxonomically validated mitochondrial genomes from historical insect collections. *Biological Journal of the Linnean Society*.

Tkalcu, B. 1974. Eine Hummel-Ausbeute aus dem Nepal-Himalaya (Insecta, Hymenoptera, Apoidea, Bombinae). *Senckenbergiana biologica* 55, 311- 349.

Tylianakis, J.M., 2013. The global plight of pollinators. *Science*, 339, 1532–1533.

Vanbergen, A.J. et al., 2014. *Status and value of pollinators and pollination services*,

Vanbergen, A.J. & the Insect Pollinators Initiative, 2013. Threats to an ecosystem service: Pressures on pollinators. *Frontiers in Ecology and the Environment*, 11, 251–259.

Vogt, O. 1911. Studien über das Artproblem. 2. Mitteilung. Über das Variieren der Hummeln. 2. Teil. (Schluss). *Sitzungsberichte der Gesellschaft naturforschender Freunde zu Berlin*, 1911, 31-74.
Walker, F.M. et al. 2016. Species from feces: Order-wide identification of Chiroptera from guano and other non-invasive genetic samples. *PLoS One*, 11, e0162342.
Ward, R.D. et al., 2005. DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, 360(1462), 1847–57.
Webb, J.M. et al., 2012. A DNA barcode library for North American Ephemeroptera: progress and prospects. *PLoS ONE*, 7(5), e38063.
Webb, K.E. et al., 2006. DNA barcoding: a molecular tool to identify Antarctic marine larvae. *Deep Sea Research Part II: Topical Studies in Oceanography*, 53(8–10), 1053–1060.
Will, K.W., Mishler, B.D., Wheeler, Q.D. 2005. The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology*, 54, 844–851.
Wilson, K.H., 1995. Molecular biology as a tool for taxonomy. *Clinical Infectious Diseases*, 20(Supplement 2), S117–S121.
Wood, T.J., Holland, J.M. & Goulson, D., 2015. Pollinator-friendly management does not increase the diversity of farmland bees and wasps. *Biological Conservation*, 187, 120–126.
Woodward, G., Gray, C. & Baird, D.J., 2013. Biomonitoring for the 21 st Century : new perspectives in an age of globalisation and emerging environmental threats. *Limnetics*, 32(2), 159–174.
Yang, C. et al., 2014. Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators*, 46, 379–389.
Yeates, D.K. et al. 2011. Integrative taxonomy, or iterative taxonomy? *Systematic Entomology*, 36, 209–217.
Yu, D.W. et al., 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3(4), 613–623.
Zhang, J. et al., 2014. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5), 614–620.
Zhou, X. et al., 2013. Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, 2(1), 4.