High-throughput monitoring of wild bee diversity and abundance via mitogenomics

Min Tang[1]†, Chloe J. Hardman[2]†, Yinqiu Ji[3]†, Guanliang Meng[1], Shanlin Liu[1], Meihua Tan[1,4], Shenzhou Yang[1], Ellen D. Moss[5], Jiaxin Wang[3], Chenxue Yang[3], Catharine Bruce[6], Tim Nevard[7,8], Simon G. Potts[2], Xin Zhou[1]* and Douglas W. Yu[3,6]*

[1]*China National GeneBank, BGI-Shenzhen, Shenzhen, Guangdong 518083, China;* [2]*Centre for Agri Environmental Research, School of Agriculture Policy and Development, University of Reading, Reading RG66AR, UK;* [3]*State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China;* [4]*University of Chinese Academy of Sciences, Beijing 100094, China;* [5]*School of Biological, Biomedical and Environmental Sciences, University of Hull, Hull HU67RX, UK;* [6]*School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR47TJ, UK;* [7]*Conservation Grade Ltd., St Neots, Cambridgeshire PE196TY, UK; and* [8]*Charles Darwin University, Darwin, NT NT0909, Australia*

Summary

1. Bee populations and other pollinators face multiple, synergistically acting threats, which have led to population declines, loss of local species richness and pollination services, and extinctions. However, our understanding of the degree, distribution and causes of declines is patchy, in part due to inadequate monitoring systems, with the challenge of taxonomic identification posing a major logistical barrier. Pollinator conservation would benefit from a high-throughput identification pipeline.

2. We show that the metagenomic mining and resequencing of mitochondrial genomes (mitogenomics) can be applied successfully to bulk samples of wild bees. We assembled the mitogenomes of 48 UK bee species and then shotgun-sequenced total DNA extracted from 204 whole bees that had been collected in 10 pan-trap samples from farms in England and been identified morphologically to 33 species. Each sample data set was mapped against the 48 reference mitogenomes.

3. The morphological and mitogenomic data sets were highly congruent. Out of 63 total species detections in the morphological data set, the mitogenomic data set made 59 correct detections (93·7% detection rate) and detected six more species (putative false positives). Direct inspection and an analysis with species-specific primers suggested that these putative false positives were most likely due to incorrect morphological IDs. Read frequency significantly predicted species biomass frequency ($R^2 = 24·9\%$). Species lists, biomass frequencies, extrapolated species richness and community structure were recovered with less error than in a metabarcoding pipeline.

4. Mitogenomics automates the onerous task of taxonomic identification, even for cryptic species, allowing the tracking of changes in species richness and distributions. A mitogenomic pipeline should thus be able to contain costs, maintain consistently high-quality data over long time series, incorporate retrospective taxonomic revisions and provide an auditable evidence trail. Mitogenomic data sets also provide estimates of species counts within samples and thus have potential for tracking population trajectories.

Key-words: agri-environment schemes, biodiversity and ecosystem services, farmland biodiversity, genome skimming, Hymenoptera, metabarcoding, metagenomics, mitogenomes, neonicotinoids, pollination

Introduction

Safeguarding wild bee populations and their pollination services is a policy priority (DEFRA 2014; Gilbert 2014) because wild bees play a keystone role in the pollination of wild plants and cultivated crops and thereby help to maintain biodiversity and food production (Breeze *et al.* 2011; Garibaldi *et al.*

2013). However, pollinators are threatened by habitat loss, pesticides, climate change and disease (Potts *et al.* 2010; Goulson *et al.* 2015), and evidence exists of declines in wild pollinators and insect-pollinated plants (Biesmeijer *et al.* 2006; Cameron *et al.* 2011; Burkle, Marlin & Knight 2013; Ollerton *et al.* 2014), but these analyses use historical records, which suffer from unequal survey effort and geographical bias.

Our understanding of bee population trajectories and responses to conservation interventions could be improved with systematic, comprehensive and auditable monitoring methods (Goulson *et al.* 2015). An important motivation for

*Correspondence authors: E-mails: dougwyu@gmail.com; xinzhou@genomics.cn
†These authors contributed equally.

this work is Lebuhn *et al.*'s (2012) calculation that 200 sampling sites are needed to have a > 90% chance to detect an annual population decline of ≥2% over a 5-year span. Lebuhn *et al.* estimated that each site would generate 3120 bees per year (pooling 26 biweekly collections), resulting in $3120_{bees} \times 200_{sites} \times 2_{yrs1\&5} = 1\cdot25$ million bees that need to be identified to species. The total cost was estimated to be US$2 million, assuming that the bees could be identified at a rate of <2 min per specimen. This assumption is possibly heroic, given that wild bee species richness ranges from hundreds to thousands of species per country, many of them difficult to separate morphologically (Schmidt *et al.* 2015).

A large-scale bee monitoring programme would therefore benefit from a high-throughput identification pipeline that produces reliable species-level identifications and estimates of species abundances, is able to add taxa and to incorporate taxonomic revisions (including to already-processed samples), is robust to sample contamination and staff turnover and is auditable by independent parties. A pipeline that uses high-throughput DNA sequencing can in principle meet these requirements.

It is now feasible to assemble large numbers of mitochondrial genomes (mitogenomes), even from species pools (Gillett *et al.* 2014; Tang *et al.* 2014; Andújar *et al.* 2015; Crampton-Platt *et al.* 2015; Gómez-Rodríguez *et al.* 2015). Mitogenomes can be thought of as super-DNA-barcodes, opening the possibility of 'mitogenomics', which we define as the application of bacterial metagenomic methods to the former bacteria now living symbiotically inside eukaryotes. We present a mitogenomics pipeline that shotgun-sequences total DNA from bulk-bee samples and conducts taxonomic binning against a reference library of bee mitogenomes (Fig. 1). In our study, bees were first identified morphologically, allowing us to conduct three tests. We asked (1) whether the morphological and mitogenomic data sets detected the same bee species and (2) whether read frequencies could estimate species biomass frequencies. (3) We also conducted community analyses and asked whether the two data sets clustered samples similarly and extrapolated similar estimates of overall bee diversity. We compare and contrast with the output from a metabarcoding pipeline, and finally, we discuss the relative merits of mitogenomics, metabarcoding, quantitative PCR and individual barcoding.

Materials and methods

SAMPLING

Bees were collected as part of a study assessing the effectiveness of agri-environment schemes for pollinators. Sampling took place in four landscapes in southern England: Chilterns North, Chilterns South, Hampshire Downs and Low Weald. Three farms per landscape were sampled, each in a different agri-environment scheme. 'Entry-Level Stewardship' is a government-funded agri-environment scheme and covered 65% of England's agricultural land in October 2013 (Natural England 2012). 'Conservation Grade' is a land-sparing protocol allocating at least 10% per farm area to wildlife habitat, but allowing some chemical inputs (www.conservationgrade.org, accessed 19 January
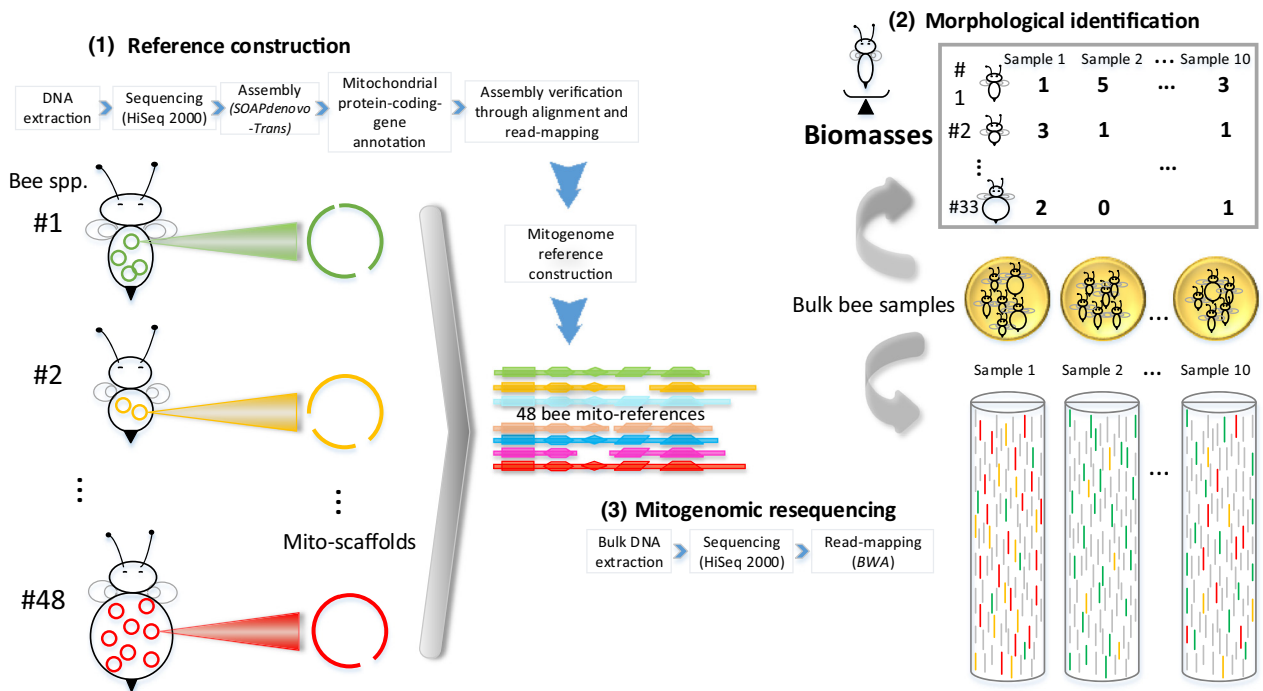
Fig. 1. Mitogenomic resequencing pipeline. (1) Reference mitogenomes were assembled from 48 bee species. (2). The 204 bee individuals in 10 bulk samples were morphologically identified to 33 bee species. (3) Total DNA from the same 10 samples was shotgun-sequenced (the 'resequencing' step), and the reads were bioinformatically mapped to the reference mitogenomes, generating Table 1. Note that the vast majority of the output in step 3 was nuclear genome reads, which were discarded.

2015). 'Organic' is a land-sharing approach, with bans on synthetic chemicals. There were three sampling rounds between 30 April and 23 August 2012. Pan-trap sampling was used because it is considered the most effective method for sampling bee diversity in European agricultural and grassland habitats and has particular advantages for solitary bees (Westphal, Bommarco & Carré 2008). Pan traps were plastic bowls painted with UV paint to form triplicate sets of blue, white and yellow. Each pan trap was half-filled with water to which a couple drops of liquid dish soap were added to reduce surface tension. Pan traps were left for 24 h, after which bees were collected and frozen. Bees were defrosted, dried and pinned, then identified to species using the keys of Else (2000) for solitary bees and Prŷs-Jones & Corbet (2011) for bumblebees. After identification, the bees were returned to the freezer. One individual from each of the 48 most abundant bee species (half the total species richness of 97 species) and the ten largest samples (3–11 species and 13–51 bees per sample) were selected for mitogenomic analysis and shipped in alcohol-filled tubes to the Kunming Institute of Zoology, China. In one of the ten samples, a single individual of a rare species, *Bombus rupestris*, was present, but since we did not make its reference mitogenome, we omitted this individual from further analysis. All other species in the samples were included in the 48 reference species. For each of the 48 reference bee species, a female not used to make the reference mitogenome was measured for its intertegular (between wing plate) distances, which is correlated with thorax volume, and thus with biomass (Cane 1987).

MITOGENOME ASSEMBLY

Genomic DNA was extracted from the thorax and legs of each of the 48 reference species following Ivanova, deWaard & Hebert (2006), avoiding the rest of the body to minimise bacterial DNA. A library with an insert size of 200 bp was prepared from each specimen following manufacturer's instruction and sequenced at 2·5 Gb depth and 100 bp PE on an Illumina HiSeq2000 at BGI-Shenzhen, China. Raw reads were filtered with a Perl script that removes reads containing adaptor contamination (with >15 bp matched to the adaptor sequence), poly-Ns (>5 bp Ns) or >1% error rate (>10 bp bases with quality score <20) following Zhou *et al.* (2013) and Tang *et al.* (2014). *De novo* assemblies for each bee were generated using *SOAPdenovo-Trans* (-K 61) (Xie *et al.* 2014), and scaffolds encoding mitochondrial proteins (mitoscaffolds) were annotated using a custom Perl script described by Zhou *et al.* (2013) with a 774 species reference data base of arthropod mitogenomes (Tang *et al.* 2014), allowing us to remove nuclear mitochondrial insertions (numts). Mitoscaffolds were used to construct bee mitogenome references, which were manually corrected and checked following Tang *et al.* (2014). Each of the 13 mitochondrial protein-coding genes extracted from the mitoscaffolds, together with reference protein-coding gene sequences from 6 bees (*Apis cerana, Apis florea, Apis mellifera, Bombus ignitus, Bombus hypocrite sapporensis* and *Melipona bicolor*), was globally aligned with CLUSTALW 2.1 (Thompson, Higgins & Gibson 1994) and ensured for correct translation frames with MEGA6 (Tamura *et al.* 2013), allowing us to correct the number of Ns generated during scaffolding of the paired-end reads. The original reads were then mapped onto the mitoscaffolds with BWA 0.6.2 (Li & Durbin 2009) to identify regions with exceptionally low or zero coverage relative to adjacent regions, and these problematic sites were confirmed or corrected using the *mpileup* command of SAMTOOLS 0.1.19 (Li *et al.* 2009). Five bee species (*Bombus pratorum, Lasioglossum laevigatum, L. lativentre, L. xanthopus* and *L. leucozonium*) with relatively poorly assembled mitogenomes were selected for additional sequencing of the remaining limited genomic DNA to improve their assemblies. We pooled the 5 species, prepared a single library of insert size of 500 bp,

and sequenced at 2 Gb depth and 300 bp PE on a MiSeq at the Kunming Institute of Zoology. Metagenomic mitoscaffolds from four bee species were assembled as previously described for the HiSeq mitogenome assembly and recovered by BLAST against their HiSeq mitoscaffolds (*L. laevigatum* sequences were not found), and the longest mitoscaffolds matching by at least 98% identity were used to improve the assemblies.

MITOGENOMIC RESEQUENCING

From each of the 10 bulk samples, the bees were homogenised in a FastPrep-24 (MP Biomedicals, Santa Ana, CA, USA), total DNA was extracted using Qiagen DNeasy Blood & Tissue Kits (Hilden, Germany), and 5 μg was used for 250-bp insert-size library construction and sequenced at 5–6 Gb depth and 100 bp PE on a HiSeq2000 at BGI-Shenzhen, China. After data filtering, clean reads from each sample were uniquely mapped using BWA onto the 48 reference mitogenomes at high stringency: 100% read coverage at 99% identity.

For species with incomplete mitogenomes, the number of mapped reads per species and sample was divided by (achieved_mitogenome_length/16000 bp) to derive a normalised read number. Finally, because each reference bee species had been separately sequenced, we could calculate the percentage of reads that were mitochondrial in origin, and we divided the read number per species per sample by this percentage to try to correct for species-level differences in mitonuclear ratio.

PCR-BASED METABARCODING

We used aliquots of the same DNA extracted from the 10 bulk samples for mitogenomic resequencing and amplified from each a 319-bp COI fragment, a subunit of the standard COI barcode region. The forward primer was LepF (5′ ATTCAACCAATCATAAAGATATTGG 3′), and the reverse primer (mlCOIintBeeR, 5′ GGDGGRTA-WANDGTTCANCCHGTHCC 3′) was modified from mlCOIintR (Leray *et al.* 2013), based on 160 bee COI reference sequences downloaded from GenBank. To build Illumina-ready PCR amplicons, we attached the standard Illumina HP10 or HP11 sequencing primers, an 8-bp index sequence, a 0- to 5-bp 'heterogeneity spacer' to the 5′ end of LepF, and mlCOIintBeeR (Fig. S3), following Fadrosh *et al.* (2014). Each sample was amplified in three independent reactions and pooled. PCRs were performed in 20 μL reaction volumes containing 2 μL of 10X buffer, 1·5 mM MgCl$_2$, 0·2 mM dNTPs, 0·2 μM each primer, 0·6 U Hot Start Taq DNA polymerase (TaKaRa Biosystems, Dalian, China) and approximately 60 ng of genomic DNA. We used a thermocycling profile of 94°C for 3 min: 35 cycles of 94°C for 1 min, 46°C for 1 min and 72°C for 90 s; with a final extension of 72°C for 7 min. PCR products were visualised on 2% agarose gels, gel-purified using the Qiagen QIAquick PCR purification kit, quantified using the QuantiT Pico-Green dsDNA Assay kit (Invitrogen, Grand Island, New York, USA), pooled and sequenced on a 300-bp PE Illumina MiSeq run at the Kunming Institute of Zoology. The raw reads were denoised with BLUE 1.1.2 (-k 25 -g 370) (Greenfield *et al.* 2014), and paired reads were merged in FLASH 1.2.10 (-m 10 -M 300) (Magoč & Salzberg 2011). The merged reads were split by sample, and the primer sequences and low-quality reads were removed in the QIIME 1.8.0 environment (Caporaso *et al.* 2010) with the script *split_libraries.py* (-l 330 -L 400 -H 9 -M 4 -b 8 -r -z truncate_remove -t –reverse_primer_mismatches 4). Only merged reads with a length of 319 bp were retained, using USEARCH's 7.0.1090 (Edgar 2010) *sortbylength* command (-minseqlength 319 -maxseqlength 319). These retained reads were clustered into unique sequences in
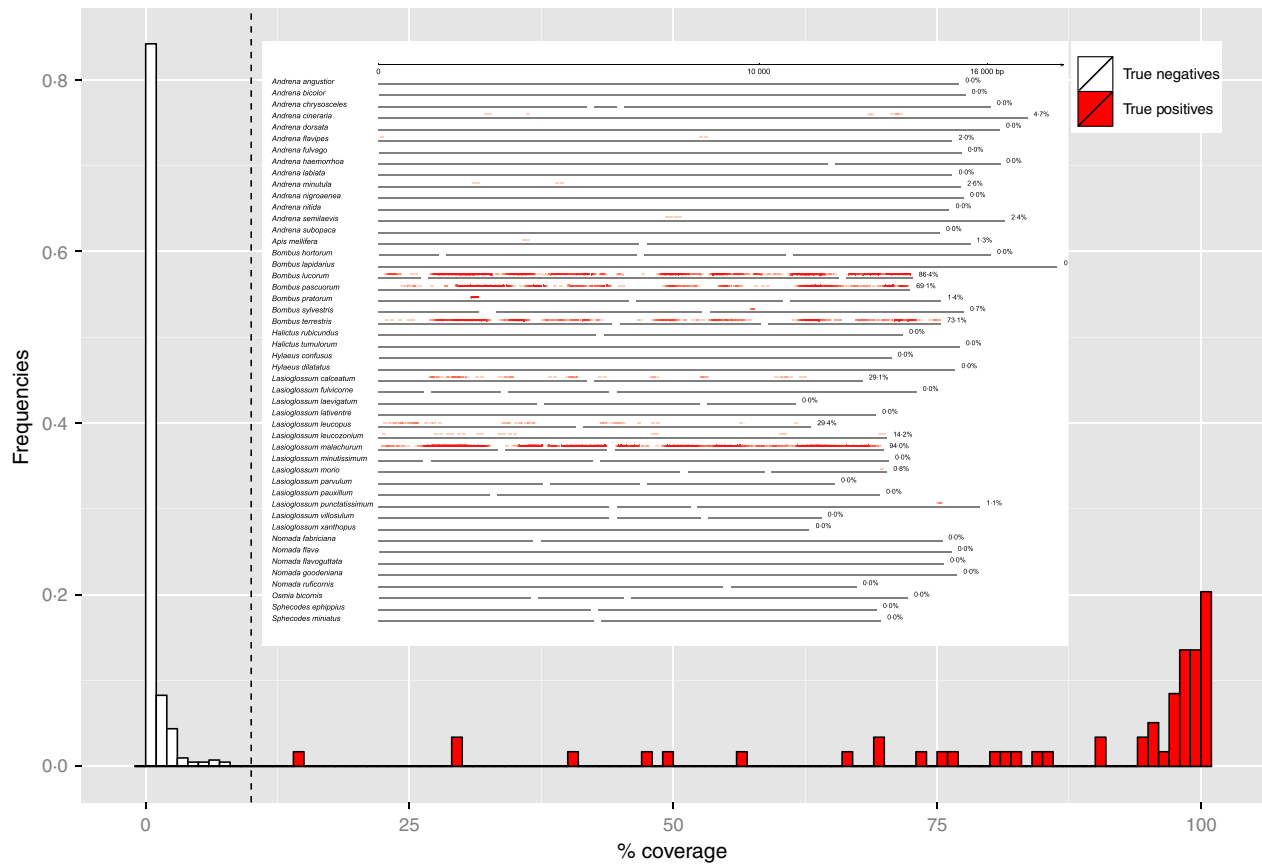
USEARCH with the *derep_fulllength* command, and USEARCH's *UCHIME* function (Edgar *et al.* 2011) was used to perform *de novo* and reference-based chimera detection and removal, the latter method using the COI sequences of the 48 reference mitogenomes. The remaining sequences were clustered at 98% similarity in CROP 1.33 (Hao, Jiang & Chen 2011), producing 468 OTUs, which were assigned taxonomies using the naïve Bayesian classifier (Wang *et al.* 2007) against the 48 COI sequences from the mitogenome data set, using a cut-off value of 0·8. RDP assigned species-level taxonomies to 288 of the OTUs, with 232 assigned at a confidence level >0·9. We merged OTUs assigned to the same species and discarded those not assigned to species.

ANALYSIS

To test whether read number can estimate bee biomasses, we first z-transformed the normalised read numbers (see Mitogenomic resequencing) and the summed biomasses per sample and then ran a linear regression between read numbers and summed biomasses by species and sample in R's 3.1.2 (R Core Team 2015) base package. Error variance increased with read number, so we re-analysed using generalised least squares regression in the R package *nlme* 3.1-118 (Pinheiro *et al.* 2014), and an exponential covariate term (*varExp*) for the biomass frequencies largely removed the heteroscedasticity.

To compare bee communities in the morphological, mitogenomic and metabarcode data sets, we conducted non-metric multidimensional (NMDS) analyses using *metaMDS( )* in the R package *vegan* 2.2-0 (Oksanen *et al.* 2014). Community dissimilarities were calculated with the *vegdist(method = 'Jaccard')* function from presence/absence (*binary = TRUE*) and quantitative (*binary = FALSE*) data (read and biomass frequencies). Finally, community correlation was estimated with *vegan*'s Procrustes test: *protest(symmetric = TRUE)*. We also used *vegan*'s *specpool( )* function to extrapolate total species richness.

Results

MITOGENOME ASSEMBLY

Mitochondrial reads accounted for 0·005–1·319% of each species' total reads. Of the 48 mitogenomes, 40 were completely assembled with all 13 expected protein-coding genes, and the other 8 contained 11 or 12 protein-coding genes (Fig. S1). Mean coverage across all mitogenomes was 224X (range 18·6X-1855·3X).

SPECIES DETECTION

A total of 204 bees were morphologically identified to 33 species (Table 1). Read coverage per mitogenome was bimodally distributed (Fig. 2), and within mitogenomes, reads mapped approximately evenly (Fig. 2 inset). In order to calculate species-detection statistics, we classified species as present if read coverage was greater than 10% (see Fig. 2). Using this threshold, mitogenomic resequencing successfully made 59 correct detections out of the 63 species-sample combinations in the morphological data set (93·7% detection rate for 'true positives', mean read coverage 86·7%, range 14·0–100%) and correctly designated 411 species-sample combinations as absent ('true negatives', mean 0·4%, range 0–7·7%) (Table 1). Four species-sample combinations in the morphological data set were not detected by mitogenomics (putative 'false negatives', mean 0·15% read coverage, range 0–0·6%), and 6 species-sample combinations were detected that were not in the morphological data set (putative 'false positives', mean 56·5%, range

Table 1. Bee counts, biomasses and mitogenomic resequencing read numbers subdivided by sample (columns) and bee species (rows). To facilitate comparison of samples across the three data sets, each sample (column) is formatted so that the largest number is reddest, descending to light pink. Discrepancies between the morphological data sets (bee counts and biomasses) and the mitogenomic data set are indicated in green (possible false negatives) and blue (possible false positives) in the mitogenomic data set. See Table S3 for the metabarcoding results

Fig. 2. Frequency histograms of read coverages from the true-negative and true-positive detections in the mitogenomics pipeline. The dashed line at 10% is the threshold used to calculate species-detection statistics. Inset: A map of read coverages on the 48 mitogenomes from sample HD_CG_1, showing the 6 true positives (*Bombus pascuorum*, *B. terrestris*, *Lasioglossum calceatum*, *L. leucopus*, *L. leucozonium* and *L. malachurum*) plus *Bombus lucorum*, a putative false positive that was confirmed by species-specific PCR (Fig. S6).

12·9–89·9%). Profiling success was 97·9% = (59 + 411)/(59 + 411 + 4+6) (Gómez-Rodríguez *et al.* 2015).

PCR-based metabarcoding was more error-prone, with 11 false negatives and 49 false positives, compared to 53 true positives (Table S2). Profiling success was 87·5%. Many of the false positives in the metabarcode data set were represented by low read numbers, but the distributions of false-positive and true-positive read numbers overlapped (Table S2).

BIOMASS FREQUENCY VS. READ FREQUENCY

A priori, larger bees should make up a larger fraction of the total DNA in a sample, and therefore, each species' mitochondrial-read frequency should correlate positively with its biomass frequency. We normalised read number per species and sample by each species' mitogenome size and ratio of mitochondrial to nuclear DNA, which were obtained from our 48 reference bee specimens. Both a linear model and a generalised least squares model to correct for heteroscedasticity indeed found that read-number frequencies could predict biomass frequencies ($P < 0.001$, $R^2 = 24.9\%$, statistical details in Fig. 3). Not correcting for mitogenome size and/or

mitonuclear ratio reduced explained variance by a few percentage points ($R^2 = 21.0\%$ for uncorrected reads; $R^2 = 21.2\%$ for reads corrected only for mitogenome size). PCR-based metabarcoding failed to find a biomass–read–number relationship ($P = 0.237$; Fig. S4).

COMMUNITY ANALYSIS

Comparisons of the morphological and mitogenomic data sets resulted in highly significantly correlated site clusters (statistical details in Fig. 4), with clear groupings by site and region. In contrast, the morphological and metabarcoding ordinations were less similar, and for the presence/absence data, non-significantly correlated (Fig. S5).

The Chao2 estimator extrapolated similar total species diversities from the two data sets (morphological: 56.3 ± 15.9 SE; mitogenome: 47.9 ± 11.8 SE; Welch's *t*-test, $t_{\text{d.f.} = 16.6} = 0.42$, $P = 0.68$). The metabarcoding data set extrapolated a lower total species diversity (36.0 ± 7.6 SE) due to a lower incidence of singleton species from the many false positives, although given the large standard errors, this extrapolation was also non-significantly different from the morphological data set ($t_{\text{d.f.} = 12.9} = 1.15$, $P = 0.27$).
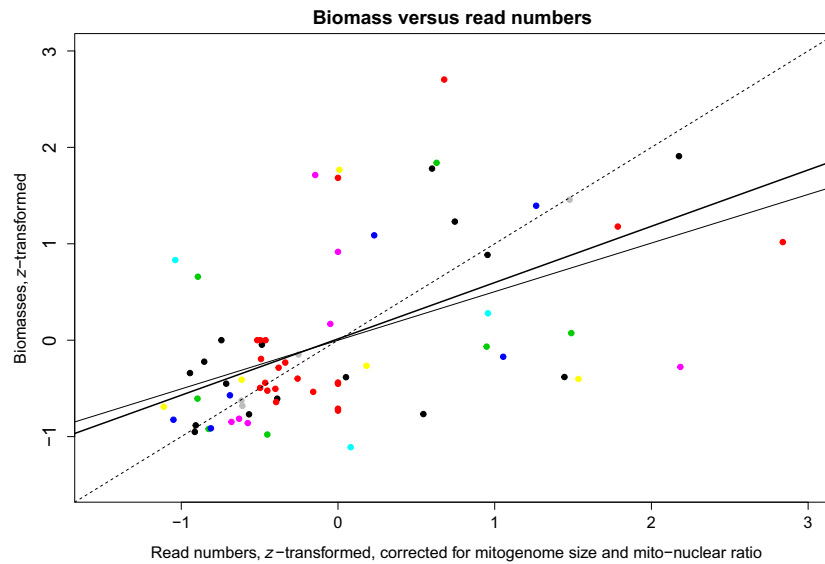
Biomass versus read numbers

Fig. 3. Scatterplot of biomasses versus read numbers. Each data point is one bee species in one sample (samples indicated by colours). The biomass and read numbers were z-transformed to correct for different sample sizes. The dashed line is the 1:1 line. If all points were on this line, there would be no error in converting from reads to biomass, and thus from biomass to counts (given a species-typical biomass). The thick solid line is the generalised least squares (GLS) regression (read_freqs ~ 0·0137 + 0·5840*biomass_freqs), and the thin solid line is the linear regression. Both regressions are highly significant ($P = 0\cdot0001$), and the linear regression returns an R^2 of 24·9%. Conducting the same regression analysis but using metabarcode-read frequency produced a non-significant GLS regression ($P = 0\cdot237$, Fig. S4).
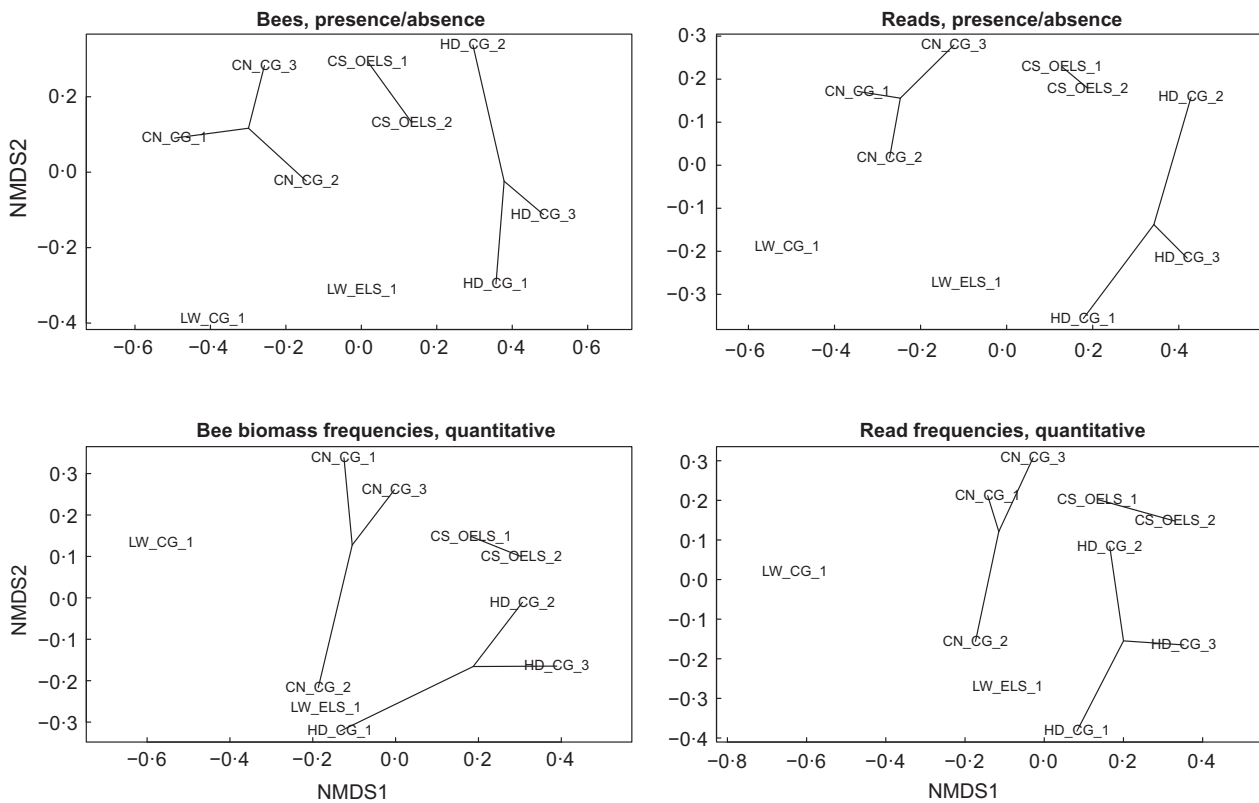
Fig. 4. Community analyses. Lines connect samples from the same farm. In the left hand column are the results using the morphological data set (Bee biomass frequencies). In the right-hand column are the results using the mitogenomic data set (Read frequencies). The top row uses presence/absence. The bottom row uses biomass and read frequencies (quantitative). In general, the morphological and mitogenomic data sets (comparing left with right) organise the samples highly similarly (procrustes $r_{presence/absence} = 0\cdot981$, $P = 0\cdot001$; $r_{quantitative} = 0\cdot966$, $P = 0\cdot001$; 9999 permutations). Samples from the same farm and locations tend to cluster together. CN = Chilterns North; CS = Chilterns South, HD = Hampshire Downs, and LW = Low Weald. CG = Conservation Grade farm; OELS = Organic+Entry-Level-Stewardship farm; ELS = Entry-Level-Stewardship farm. See Fig. S5 for the metabarcoding result.

Inspection of the mitogenome data set (Table 1) suggested that all but one of the 10 putative false positives and negatives could reasonably be ascribed to errors in the morphological data set.

Mitogenomic resequencing detected *Bombus lucorum* in two samples (HD_CG_1 and LW_ELS_1, Table 1), but morphology did not. In those samples, *Bombus terrestris* had also been detected by both morphology and mitogenomes, and the workers of these two species are very difficult to tell apart. We designed *B. lucorum*-specific PCR primers and successfully detected *B. lucorum* in both samples, and we did not detect *B. lucorum* in any other sample (Fig. S6), which suggests that *B. lucorum* in the mitogenomic data set was not a false positive.

We were unable to design PCR tests for the other discrepancies, but in the case of *Andrena subopaca*, which mitogenomics did not detect in two samples (CN_CG_1 and CN_CG_2, green cells, Table 1), multiple individuals of other *Andrena* species were in the same samples. Similarly, mitogenomics did not detect two *Lasioglossum* species in two samples (CN_CG_2 and LW_ELS_1, green cells, Table 1), but both samples had other *Lasioglossum* species that were not detected in the morphological data set (blue cells). *Lasioglossum* species can be difficult to differentiate on the basis of morphology.

Finally, in CN_CG_2, the morphological data set contained only one *Lasioglossum* bee (*L. calceatum*), but the mitogenome data set detected two *Lasioglossum* species. In this case, either mitogenomics truly threw up a false positive, or DNA from two *Lasioglossum* species was in the sample, and mitogenomics detected them both.

Discussion

We show that mitogenomic resequencing of bulk samples and mapping against a reference data base provides a reliable and high-throughput method for identifying bee species (Table 1). The output is very suitable for community analysis (Fig. 4) and occupancy modelling, which will allow tracking of changes in species richness and distributions, two of Lebuhn *et al.*'s (2012) three proposed metrics. We note that a higher detection threshold than the 10% we used would have little effect on our classification success, as all 'true positive' detections but one had coverages ≥29% (Fig. 2).

Mitogenomic resequencing also successfully recovers quantitative information on biomass frequencies (Zhou *et al.* 2013; Gómez-Rodríguez *et al.* 2015; Paula *et al.* 2015; Srivathsan *et al.* 2015), although currently, the biomass–read relationship is heteroscedastic and noisy (Fig. 3). Sources of noise include measurement error, bee biomasses that vary across individuals of the same species, especially in social species with workers and queens (Richards & Packer 1996); mitochondrial DNA densities that vary across individuals, tissues within individuals, life spans (Veltri, Espiritu & Singh 1990) and species (this study); and noise introduced during DNA extraction, library construction, sequencing, quality control and read matching. Finally, with an incomplete reference data base, small numbers of reads could be matched to an incorrect species, even when only unique mappings are accepted, as we did.

Nonetheless, even a noisy relationship can be used for tracking the population trajectories of hundreds of bee species at a time. A sample's total bee biomass can be measured before DNA extraction, and after sequencing, the biomass frequencies per species can be converted to absolute biomasses, which can then be converted to counts using species-specific estimates of biomass per worker bee. Count data produced by mitogenomics will thus contain non-process (observation) error (Hilborn & Mangel 1997), and the cost of this error is the need for more samples to achieve the same statistical power in detecting population declines. We note that Lebuhn *et al.*'s (2012) simulation did not model taxonomic identification error, which would add a similar sort of error.

In the future, when we have sufficient bee mitogenomes to act as an unbiased reference set, capture-enrichment techniques (Avila-Arcos *et al.* 2011) could be employed to increase the proportion of raw mitochondrial reads from the current ~1% to >40% (thus using more of the sequencer output), and we hypothesise that this will reduce heteroscedasticity and increase explained variance in the biomass–read relationship (Fig. 3), by ensuring that high-biomass species in a sample are represented accordingly. Importantly, low-biomass species are more reliably represented by fewer reads (Fig. 3), which suggests that low-abundance species will be identified as such. This is of particular importance, given that low-abundance species are arguably of greater conservation concern.

In sum, mitogenomics has high potential for allowing monitoring programmes (DEFRA 2014) to track pollinator populations and to assess and target appropriate conservation interventions. Mitogenomics pipelines possess institutional advantages desired in an identification pipeline. Automated taxonomic identification, even for cryptic species, should contain cost inflation, maintain consistently high quality data over long time series and provide an auditable evidence trail, since data sets can be independently analysed at any stage of the bioinformatic pipeline, and, at extra cost, parallel samples can be taken and processed independently. Moreover, taxonomic revisions and new taxa, such as pests, their predators and other pollinators, can be incorporated at any time by (re-)mapping old and new sequencing data sets against new reference data sets. The mitogenomic pipeline is scalable to more species and larger samples, as we are relying on software and sequencers designed for whole-genome scale resequencing. The skills needed to carry out a mitogenomics pipeline (non-destructive DNA extraction, running bioinformatic scripts) are easily learned, with the other steps able to be outsourced to sequencing centres.

A key advantage of mitogenomics is the opportunity to do away with PCR, which reduces laboratory workload, sequence error and contamination risk and therefore results in lower rates of false-positive and false-negative species detections relative to metabarcoding (Table 1 vs. Table S2). We observe that despite the fact that the bees in this study were handled for morphological identification and were thus exposed to more cross-contamination than would be the case in a pure molecu-

lar study, read-coverage values of the true negatives and true positives did not overlap in the mitogenomic data set (Fig. 2), but they did overlap in the metabarcoding data set (Table S2). We conclude that contaminants are inherently easier to detect and omit in a mitogenomic pipeline than in a metabarcoding pipeline. This is a crucial feature in a large-scale, long-term monitoring programme where it is impossible to guarantee that collecting and sorting apparatus has always been correctly cleaned between samples.

It is worth emphasising that our 48-species reference data set deliberately included species that were abundant overall in the landscape but not present in our ten-sample morphological data set, and thus, we did not expect to detect these species. With the exception of *Bombus lucorum*, which our PCR test suggests was indeed present, these non-expected species were not detected. This suggests that a synoptic reference data base *per se* will not produce false positives.

It is even possible to run a mitogenomics pipeline without whole mitogenomes. Gómez-Rodríguez *et al*. (2015) have shown it possible to map against a reference data base of only standard, 5′-end COI DNA barcodes. However, the advantages of mitogenomes are that the larger target makes resequencing more efficient (Zhou *et al*. 2013; Tang *et al*. 2014; Gómez-Rodríguez *et al*. 2015; Paula *et al*. 2015), and mitogenomes provide more resolved phylogenetic information (Gillett *et al*. 2014; Andújar *et al*. 2015; Crampton-Platt *et al*. 2015). Mapping to mitogenomes also increases detection confidence, because only species that are truly present in a sample will produce DNA reads that map across the whole mitogenome (Fig. 2 inset), but any stray PCR amplicons will only map to a single locus. The many thousands of bee species that have been collected for standard DNA barcoding can therefore be used for resequencing, and in the future, these specimens can be used as DNA sources for mitogenome assembly.

ALTERNATIVE PIPELINES

Metabarcoding

The big advantage of metabarcoding is that, with appropriate controls and filtering, it can estimate beta and alpha diversity from bulk samples in which taxa are not well characterised and there is no reference data base, such as with meiofauna (Fonseca *et al*. 2010), environmental DNA (Yoccoz *et al*. 2012) and novel locations (Ji *et al*. 2013). The cost is that PCR endpoint read numbers are not reliable estimates of starting DNA concentrations (Fig. S4) (Amend, Seifert & Bruns 2010; Yu *et al*. 2012), due to inherent stochasticity and since each nucleotide mismatch between primer and primer region can result in a ten-fold drop in amplification (Piñol *et al*. 2014). Amplification bias is what makes it difficult to identify contaminants, because contaminant tissue might match primers better than some of the truly present taxa. We think this is why some of our false-positive read numbers are greater than some of the true-positive read numbers (Table S2). Stochasticity in end-point PCR read numbers can also play a role in amplifying contaminants.

Another challenge for metabarcoding is primer design. We used fusion primers with heterogeneity spacers to make a separate library for each sample, which prevents tag jumping (Schnell, Bohmann & Gilbert 2015) and to increase sequence entropy, which improves sequence quality, but fusion primers are longer and thus somewhat less likely to amplify species, which might have contributed to the greater number of false negatives relative to the mitogenomics data set.

qPCR/ddPCR

qPCR (quantitative PCR) and ddPCR (droplet-digital PCR) can quantify species-specific DNA concentrations (Doi *et al*. 2015), given properly designed primers and probes, and samples and primer sets can be multiplexed. However, these systems have not yet, to our knowledge, been applied to bulk samples (although they are widely used for environmental DNA (Ficetola *et al*. 2015)), and it remains unclear whether this approach can be scaled up to hundreds or thousands of species. Moreover, adding taxa would require re-amplification of all samples.

Massively parallel barcoding

Surprisingly, the most competitive alternative to mitogenomics could be individual-based DNA barcoding, in which (portions of) individual bees are separately extracted, amplified and sequenced in parallel using tagged amplicons on Illumina sequencers, at an estimated cost of ≤US\$1·5 per specimen (Meier *et al*. 2015; Shokralla *et al*. 2015). Clearly, this method would generate the best count data. However, following Shokralla *et al*.'s (2015) estimate of seven hands-on hours per 1000 specimens, the 1·25 million bees estimated by Lebuhn *et al*. (2012) would require ~50 person-months. With mitogenomics, we estimate that 500 samples can be extracted per person-month, meaning that ($26_{weeks} \times 200_{sites} \times 2_{yrs1\&5}=$) 10 400 samples would require ~21 person-months before sending to a sequencing centre for library prep and sequencing. In the more seasonal UK, 16 weeks of sampling in 200 sites might generate around 192 000 bees (CQ Tang, pers. comm.), amounting to ~7·7 person-months for barcoding and ~12·9 person-months for mitogenomics.

However, barcoding costs scale with specimen number, but mitogenomic costs scale with sample number. Thus, if we include non-bee taxa (e.g. flies, mites), the mitogenomics workload would not increase, except for the one-time cost of assembling additional mitogenomes, whereas the individual-barcoding workload could increase many-fold.

In conclusion, a mitogenomics approach provides reliable species detection and information on abundance from bulk samples and also provides important institutional advantages: robustness to contamination, the ability to add taxa cheaply and retrospectively, a low skills requirement from staff and the ability to audit samples. Other DNA-based pipelines do have their advantages, and the choice of which to use will depend on study scale, on-going improvements in genomics technology,

the number of samples that can be pooled per library and the importance of accuracy in specimen counts.

Data accessibility

The genomic data sets (48 individual bees and 10 bulk samples) have been deposited in GenBank (SRA264601). The 48 mitogenome assemblies and the metabarcoding data and command histories have been deposited in the Dryad depository: http://datadryad.org/resource/doi:10.5061/dryad.r7122 (Tang *et al.* 2015).

References

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

Andújar, C., Arribas, P., Ruzicka, F., Crampton-Platt, A., Timmermans, M.J.T.N. & Vogler, A.P. (2015) Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Molecular Ecology*. doi:10.1111/mec.13195 [Epub ahead of print].

Avila-Arcos, M.C., Cappellini, E., Romero-Navarro, J.A., Wales, N., Moreno-Mayar, J.V., Rasmussen, M. *et al.* (2011) Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Scientific Reports*, **1**, 74.

Biesmeijer, J.C., Roberts, S.P.M., Reemer, M., Ohlemuller, R., Edwards, M., Peeters, T. *et al.* (2006) Parallel declines in pollinators and insect-pollinated plants in Britain and the Netherlands. *Science*, **313**, 351–354.

Breeze, T.D., Bailey, A.P., Balcombe, K.G. & Potts, S.G. (2011) Pollination services in the UK: how important are honeybees? *Agriculture Ecosystems & Environment*, **142**, 137–143.

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**, 1611–1615.

Cameron, S.A., Lozier, J.D., Strange, J.P., Koch, J.B., Cordes, N., Solter, L.F. & Griswold, T.L. (2011) Patterns of widespread decline in North American bumble bees. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 662–667.

Cane, J.H. (1987) Estimation of bee size using intertegular span (Apoidea). *Journal of the Kansas Entomological Society*, **60**, 145–147.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.

Crampton-Platt, A., Timmermans, M.J.T.N., Gimmel, M.L., Kutty, S.N., Cockerill, T.D., Khen, C.V. & Vogler, A.P. (2015) Soup to tree: the phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Molecular Biology and Evolution*. doi:10.1093/molbev/msv111 [Epub ahead of print].

DEFRA. (2014) *The National Pollinator Strategy: for bees and other pollinators in England*. Retrieved from www.gov.uk/government/uploads/system/uploads/attachment_data/file/370199/pb14221-national-pollinator-strategy.pdf

Doi, H., Uchii, K., Takahara, T., Matsuhashi, S., Yamanaka, H. & Minamoto, T. (2015) Use of droplet digital PCR for estimation of fish abundance and biomass in environmental DNA surveys. *PLoS ONE*, **10**, e0122763.

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

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Else, G.R. (2000) *Draft keys to solitary bees of Great Britain. Handbook of the Bees of the British Isles*.

Fadrosh, D.W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R.M. & Ravel, J. (2014) An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, **2**, 6.

Ficetola, G.F., Pansu, J., Bonin, A., Coissac, E., Giguet Covex, C., De Barba, M. *et al.* (2015) Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, **15**, 543–556.

Fonseca, V.G., Carvalho, G.R., Sung, W., Johnson, H.F., Power, D.M., Neill, S.P. *et al.* (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, **1**, 98.

Garibaldi, L.A., Steffan-Dewenter, I., Winfree, R., Aizen, M.A., Bommarco, R., Cunningham, S.A. *et al.* (2013) Wild pollinators enhance fruit set of crops regardless of honey bee abundance. *Science*, **339**, 1608–1611.

Gilbert, N. (2014) 'Life on Earth' project gets under way. *Nature*, **510**, 455.

Gillett, C.P.D.T., Crampton-Platt, A., Timmermans, M.J.T.N., Jordal, B.H., Emerson, B.C. & Vogler, A.P. (2014) Bulk *de novo* mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, **31**, 2223–2237.

Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M.J.T.N., Baselga, A. & Vogler, A.P. (2015) Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12376 [Epub ahead of print].

Goulson, D., Nicholls, E., Botías, C. & Rotheray, E.L. (2015) Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. *Science*, **347**, 1255957.

Greenfield, P., Duesing, K., Papanicolaou, A. & Bauer, D.C. (2014) Blue: correcting sequencing errors using consensus and context. *Bioinformatics*, **30**, 2723–2732.

Hao, X., Jiang, R. & Chen, T. (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics*, **27**, 611–618.

Hilborn, R. & Mangel, M. (1997) *The Ecological Detective*. Princeton University Press, Princeton, New Jersey, USA.

Ivanova, N.V., deWaard, J.R. & Hebert, P.D.N. (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes*, **6**, 998–1002.

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A. *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Lebuhn, G., Droege, S., Connor, E.F., Gemmill-Herren, B., Potts, S.G., Minckley, R.L. *et al.* (2012) Detecting insect pollinator declines on regional and global scales. *Conservation Biology*, **27**, 113–120.

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T. & Machida, R.J. (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**, 34.

Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Magoč, T. & Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

Meier, R., Wong, W., Srivathsan, A. & Foo, M. (2015) $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. *Cladistics*. doi:10.1111/cla.12115 [Epub ahead of print].

Natural England. (2012) Land Management Update. Retrieved October 2013, from https://www.gov.uk/environmental-stewardship

Oksanen, J., Blanchet, F.G., Kindt, R., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. & Wagner, H. (2014) *vegan*: Community Ecology Package. R package version 2.0-10. http://CRAN.R-project.org/package=vegan.

Ollerton, J., Erenler, H., Edwards, M. & Crockett, R. (2014) Extinctions of aculeate pollinators in Britain and the role of large-scale agricultural changes. *Science*, **346**, 1360–1362.

Paula, D.P., Linard, B., Andow, D.A. & Sujii, E.R. (2015) Detection and decay rates of prey and prey symbionts in the gut of a predator through metagenomics. *Molecular Ecology*. doi:10.1111/1755-0998.12364 [Epub ahead of print].

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. (2014) *nlme*: Linear and Nonlinear Mixed Effects Models. R package version 3.1-118, http://CRAN.R-project.org/package=nlme.

Piñol, J., Mir, G., Gomez Polo, P. & Agustí, N. (2014) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12355 [Epub ahead of print].

Potts, S.G., Biesmeijer, J.C., Kremen, C., Neumann, P., Schweiger, O. & Kunin, W.E. (2010) Global pollinator declines: trends, impacts and drivers. *Trends in Ecology & Evolution*, **25**, 345–353.

Prŷs-Jones, O.E. & Corbet, S.A. (2011) *Bumblebees*. Pelagic Publishing Ltd, Exeter, UK.

R Core Team. (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria URL http://www.R-project.org/.

Richards, M.H. & Packer, L. (1996) The socioecology of body size variation in the primitively eusocial sweat bee, *Halictus ligatus* (Hymenoptera: Halictidae). *Oikos*, **77**, 68.

Schmidt, S., Schmid Egger, C., Morinière, J., Haszprunar, G. & Hebert, P.D.N. (2015) DNA barcoding largely supports 250 years of classical taxonomy: identifications for Central European bees (Hymenoptera, Apoidea partim). *Molecular Ecology Resources*. doi: 10.1111/1755-0998.12363. [Epub ahead of print].

Schnell, I.B., Bohmann, K. & Gilbert, M.T.P. (2015) Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12363 [Epub ahead of print].

Shokralla, S., Porter, T.M., Gibson, J.F., Dobosz, R., Janzen, D.H., Hallwachs, W., Golding, G.B. & Hajibabaei, M. (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, **5**, 9687.

Srivathsan, A., Sha, J.C.M., Vogler, A.P. & Meier, R. (2015) Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). *Molecular Ecology Resources*, **15**, 250–261.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, **30**, 2725–2729.

Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S. *et al.* (2014) Multiplex sequencing of pooled mitochondrial genomes-a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, **42**, e166.

Tang, M., Hardman, C., Ji, Y., Meng, G., Tan, M., Yang, S. *et al.* (2015) Data from: high-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*. doi:10.5061/dryad.r7122.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Veltri, K.L., Espiritu, M. & Singh, G. (1990) Distinct genomic copy number in mitochondria of different mammalian organs. *Journal of Cellular Physiology*, **143**, 160–164.

Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

Westphal, C., Bommarco, R. & Carré, G. (2008) Measuring bee diversity in different European habitats and biogeographical regions. *Ecological Modelling*, **78**, 653–671.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S. *et al.* (2014) SOAP-denovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, **30**, 1660–1666.

Yoccoz, N.G., Bråthen, K.A., Gielly, L., Haile, J., Edwards, M.E., Goslar, T. *et al.* (2012) DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, **21**, 3647–3655.

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**, 613–623.

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L. *et al.* (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2**, 4.

Supporting Information

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

Fig. S1. The 48 reference mitogenomes, color-coded for the 13 protein-coding genes and the rDNA+Control Region.

Fig. S2. Mapping of reads (red line segments) on reference mitogenomes (black lines).

Fig. S3. Metabarcoding primers.

Fig. S4. Scatterplot of Biomasses versus Metabarcoding Read numbers.

Fig. S5. Community analyses for Metabarcoding data.

Fig. S6. PCR test for *Bombus lucorum*.

Table S1. DNA and assembly quality information for the 48 reference bee species and the 10 bulk samples.

Table S2. Bee counts, biomasses, and metabarcoding read numbers, subdivided by sample (columns) and bee species (rows).

Appendix S1. Read map (see Fig. S2) for CN_CG_1.

Appendix S2. Read map (see Fig. S2) for CN_CG_2.

Appendix S3. Read map (see Fig. S2) for CN_CG_3.

Appendix S4. Read map (see Fig. S2) for CS_OELS_1.

Appendix S5. Read map (see Fig. S2) for CS_OELS_2.

Appendix S6. Read map (see Fig. S2) for HD_CG_1.

Appendix S7. Read map (see Fig. S2) for HD_CG_2.

Appendix S8. Read map (see Fig. S2) for HD_CG_3.

Appendix S9. Read map (see Fig. S2) for LW_CG_1.

Appendix S10. Read map (see Fig. S2) for LW_ELS_1.

High-throughput monitoring of wild bee diversity and abundance via mitogenomics: Supplemental information

Min Tang[*,1]

Chloe J. Hardman[*,2]

Yinqiu Ji[*,3]

Guanliang Meng[1]

Shanlin Liu[1]

Meihua Tan[1,4]

Shenzhou Yang[1]

Ellen D. Moss[5]

Jiaxin Wang[3]

Chenxue Yang[3]

Catharine Bruce[6]

Tim Nevard[7,8]

Simon G. Potts[2]

Xin Zhou[†,1]

Douglas W. Yu[†,3,6]

[1] China National GeneBank, BGI-Shenzhen, Shenzhen, Guangdong 518083 China

[2] Centre for Agri Environmental Research, School of Agriculture, Policy and Development, University of Reading, Reading RG66AR UK

[3] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Kunming, Yunnan 650223 China

[4] University of Chinese Academy of Sciences, Beijing 100094 China

[5] School of Biological, Biomedical and Environmental Sciences, University of Hull,   Hull HU67RX UK

[6] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR47TJ UK

[7] Conservation Grade, Ltd., St Neots, Cambridgeshire PE196TY UK

[8] Charles Darwin University, Darwin NT0909 Australia

[*] These authors contributed equally

[†]Co-corresponding authors: dougwyu@gmail.com; xinzhou@genomics.cn

Figure S1. The 48 reference mitogenomes, color-coded for the 13 protein-coding genes and the rDNA+Control Region. Bolded genomes were supplemented with MiSeq data.

Figure S2. Mapping of reads (red line segments) on reference mitogenomes (black lines). Sample shown is CN_CG_1. All ten true positive detections were indicated by a uniform distribution of reads along the corresponding reference mitogenome, with coverage ≥ 40.4%. All true negatives were indicated by low coverage (<5.3%). The one putative false negative (*Andrena subopaca*) was matched by no reads.
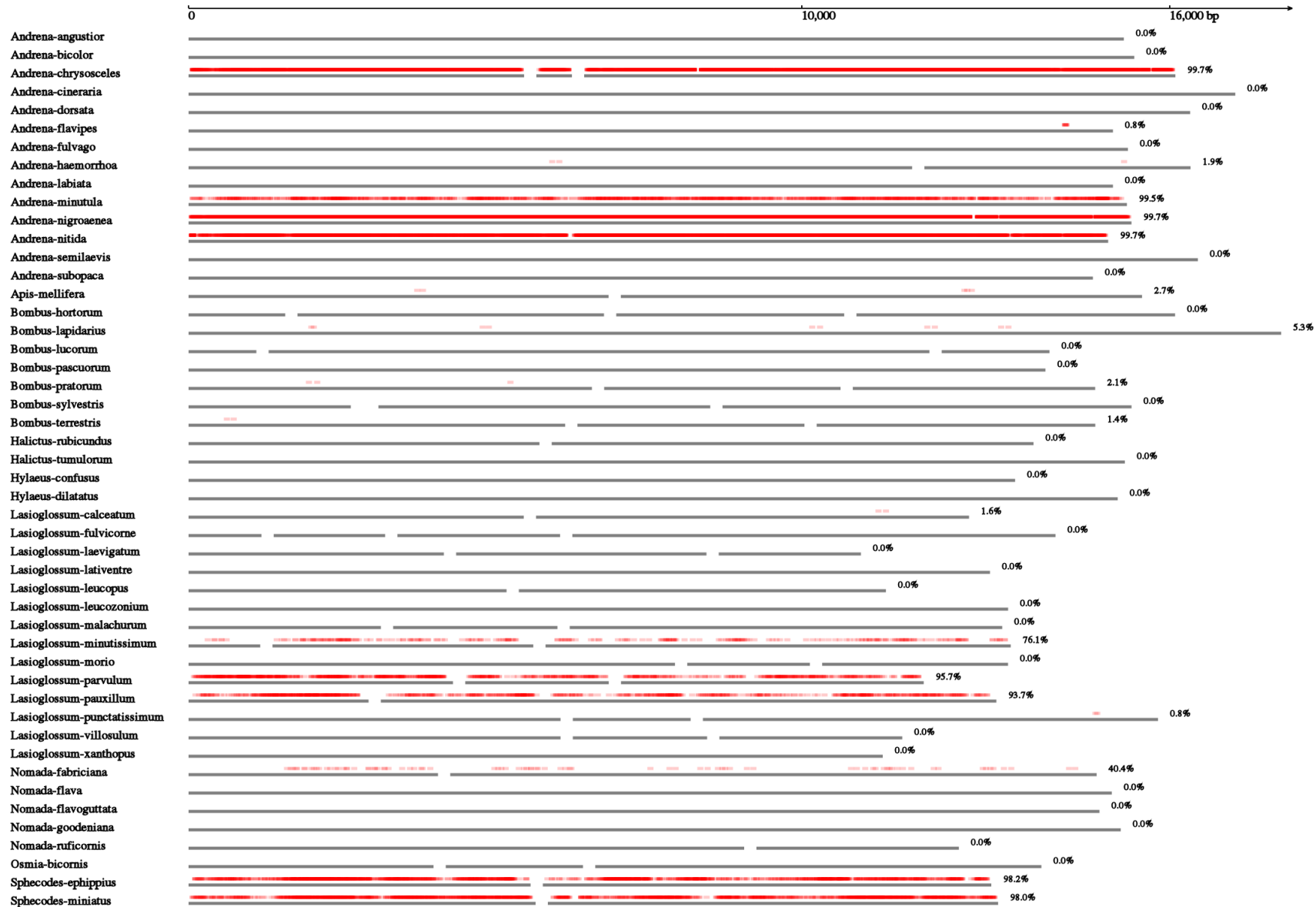
Figure S3. Metabarcoding primers.

	5' HP10/HP11 + index + *spacer* + LepF/ mlCOIintBeeR
LepF-index1	CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT TCAGTGCG ATTCAACCAATCATAAAGATATTGG
LepF-index2	CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT GTAGCAGA *T* ATTCAACCAATCATAAAGATATTGG
LepF-index3	CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT ATTCACAG *GT* ATTCAACCAATCATAAAGATATTGG
LepF-index4	CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT ATTCCATA *CGA* ATTCAACCAATCATAAAGATATTGG
LepF-index5	CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT TGGCCGAT *ATGA* ATTCAACCAATCATAAAGATATTGG
LepF-index6	CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT ATGCATAC *TGCGA* ATTCAACCAATCATAAAGATATTGG
mlCOIintBeeR-index1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT TCAGTGCG GGDGGRTAWANDGTTCANCCHGTHCC
mlCOIintBeeR-index2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT GTAGCAGA *T* GGDGGRTAWANDGTTCANCCHGTHCC
mlCOIintBeeR-index3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT ATTCACAG *GT* GGDGGRTAWANDGTTCANCCHGTHCC
mlCOIintBeeR-index4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT ATTCCATA *CGA* GGDGGRTAWANDGTTCANCCHGTHCC
mlCOIintBeeR-index5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT TGGCCGAT *ATGA* GGDGGRTAWANDGTTCANCCHGTHCC
mlCOIintBeeR-index6	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT ATGCATAC *TGCGA* GGDGGRTAWANDGTTCANCCHGTHCC

Table S4. Bee counts, biomasses, and metabarcoding read numbers, subdivided by sample (columns) and bee species (rows). To facilitate comparison of samples across the three datasets, each sample (column) is formatted so that the largest number is reddest, descending to light pink. Discrepancies between the morphological (Bee counts and Biomass) datasets and the mitogenomic dataset are indicated in green (possible false negatives) and blue (possible false positives) in the mitogenomic dataset.

Bee counts

Bee species	CN_CG_1	CN_CG_2	CN_CG_3	CS_OELS_1	CS_OELS_2	HD_CG_1	HD_CG_2	HD_CG_3	LW_CG_1	LW_ELS_1
Andrena angustior										
Andrena bicolor			2							
Andrena chrysosceles	5									
Andrena cineraria				1		2				
Andrena dorsata				1						
Andrena flavipes									5	
Andrena fulvago										
Andrena haemorrhoa		1	1	1	5					
Andrena labiata										
Andrena minutula	9	2							4	
Andrena nigroaenea	3	1	4	2		2				
Andrena nitida	1		3	3	2					
Andrena semilaevis								1		
Andrena subopaca	2	1								
Apis mellifera		1	1						1	1
Bombus hortorum									1	
Bombus lapidarius		1								
Bombus lucorum										
Bombus pascuorum						1				
Bombus pratorum			1							
Bombus sylvestris										
Bombus terrestris	2			2						4
Halictus rubicundus							2			
Halictus tumulorum							1			
Hylaeus confusus										
Hylaeus dilatatus										
Lasioglossum calceatum		1		10	10	2	13	15		5
Lasioglossum fulvicorne										
Lasioglossum laevigatum										
Lasioglossum lativentre										
Lasioglossum leucopus				2		2				
Lasioglossum leucozonium				1						
Lasioglossum malachurum				10	4					
Lasioglossum minutissimum	4									
Lasioglossum morio										
Lasioglossum parvulum	14		6							
Lasioglossum pauxillum	5								2	3
Lasioglossum punctatissimum										1
Lasioglossum villosulum										
Lasioglossum xanthopus										
Nomada fabriciana	1									
Nomada flava										
Nomada flavoguttata										
Nomada goodeniana		1								
Nomada ruficornis		1								
Osmia bicornis										
Sphecodes ephippius	3	1								
Sphecodes miniatus	4									
Species count	11	11	7	6	3	6	5	4	5	5

Estimated bee biomasses (mg)

Bee species	CN_CG_1	CN_CG_2	CN_CG_3	CS_OELS_1	CS_OELS_2	HD_CG_1	HD_CG_2	HD_CG_3	LW_CG_1	LW_ELS_1
Andrena angustior										
Andrena bicolor			22.4							
Andrena chrysosceles	51.5									
Andrena cineraria				23.8		47.6				
Andrena dorsata				11.9						
Andrena flavipes									90.7	
Andrena fulvago										
Andrena haemorrhoa		19.8	19.8	19.8	99.0					
Andrena labiata										
Andrena minutula	41.5	9.2							18.4	
Andrena nigroaenea	81.3	27.1	108.4	54.2		54.2				
Andrena nitida	29.4		88.3	88.3	58.9					
Andrena semilaevis								4.7		
Andrena subopaca	10.0	5.0								
Apis mellifera		36.4	36.4						36.4	36.4
Bombus hortorum									137.1	
Bombus lapidarius		113.9								
Bombus lucorum										
Bombus pascuorum						73.1				
Bombus pratorum			76.5							
Bombus sylvestris										
Bombus terrestris	314.3			314.3						628.5
Halictus rubicundus							23.8			
Halictus tumulorum							4.9			
Hylaeus confusus										
Hylaeus dilatatus										
Lasioglossum calceatum		8.1		81.4	81.4	16.3	105.8	122.1		40.7
Lasioglossum fulvicorne										
Lasioglossum laevigatum										
Lasioglossum lativentre										
Lasioglossum leucopus				5.0		5.0				
Lasioglossum leucozonium				10.3						
Lasioglossum malachurum				62.7	25.1					
Lasioglossum minutissimum	6.0									
Lasioglossum morio										
Lasioglossum parvulum	46.4		19.9							
Lasioglossum pauxillum	14.3								5.7	8.6
Lasioglossum punctatissimum										3.6
Lasioglossum villosulum										
Lasioglossum xanthopus										
Nomada fabriciana	5.9									
Nomada flava										
Nomada flavoguttata										
Nomada goodeniana		14.0								
Nomada ruficornis		7.9								
Osmia bicornis										
Sphecodes ephippius	16.3	5.4								
Sphecodes miniatus	7.4									
Species count	11	11	7	6	3	6	5	4	5	5

Metabarcoding read numbers

11 False negatives and 49 false positives (CROP OTU *de novo* clustering)

Bee species	CN_CG_1	CN_CG_2	CN_CG_3	CS_OELS_1	CS_OELS_2	HD_CG_1	HD_CG_2	HD_CG_3	LW_CG_1	LW_ELS_1
Andrena angustior										
Andrena bicolor			124							
Andrena chrysosceles	6392				2					
Andrena cineraria				8		138				
Andrena dorsata				100						
Andrena flavipes		2		76					2	141703
Andrena fulvago										
Andrena haemorrhoa		255	811	282	11808					
Andrena labiata										
Andrena minutula	4257	25			3				1685	
Andrena nigroaenea	161089	1221	119699	140690	225	61	119397	100	162	150
Andrena nitida	7346		4671	11686	113468	8	33	285		2
Andrena semilaevis				13	2				31912	40
Andrena subopaca										
Apis mellifera	4	828	43997	1	39	30	104	19	4239	91524
Bombus hortorum										
Bombus lapidarius	74	224169	64			271	15		4	34
Bombus lucorum			2	2		8				142
Bombus pascuorum						2				
Bombus pratorum										
Bombus sylvestris										
Bombus terrestris		88				84	4	2	2	857
Halictus rubicundus							880			
Halictus tumulorum							1085			
Hylaeus confusus										
Hylaeus dilatatus										
Lasioglossum calceatum	14		123	1492	10256	58	8809	141597	242	2175
Lasioglossum fulvicorne										
Lasioglossum laevigatum										
Lasioglossum lativentre										
Lasioglossum leucopus									146	
Lasioglossum leucozonium							3			226
Lasioglossum malachurum	7	126	57	3	12171	2	147412	192		2
Lasioglossum minutissimum										
Lasioglossum morio										
Lasioglossum parvulum										
Lasioglossum pauxillum	2								2	
Lasioglossum punctatissimum										
Lasioglossum villosulum										
Lasioglossum xanthopus										
Nomada fabriciana	58									
Nomada flava										
Nomada flavoguttata										
Nomada goodeniana		132								
Nomada ruficornis		36								
Osmia bicornis										
Sphecodes ephippius	8	2								
Sphecodes miniatus	87									
Species count	12	11	11	8	8	11	10	9	12	9

Figure S5. Scatterplot of Biomasses versus Metabarcoding Read numbers. Each data point is one bee species in one sample. Colours indicate the 10 samples. The biomass and read numbers were z-transformed to correct for different sample sizes. The dashed line is the 1:1 line. If all points were on this line, there would be no error in converting from reads to biomass, and thus from biomass to counts (given a species-typical biomass). The solid line is the non-significant Generalised Least Squares (GLS) regression (p=0.237).
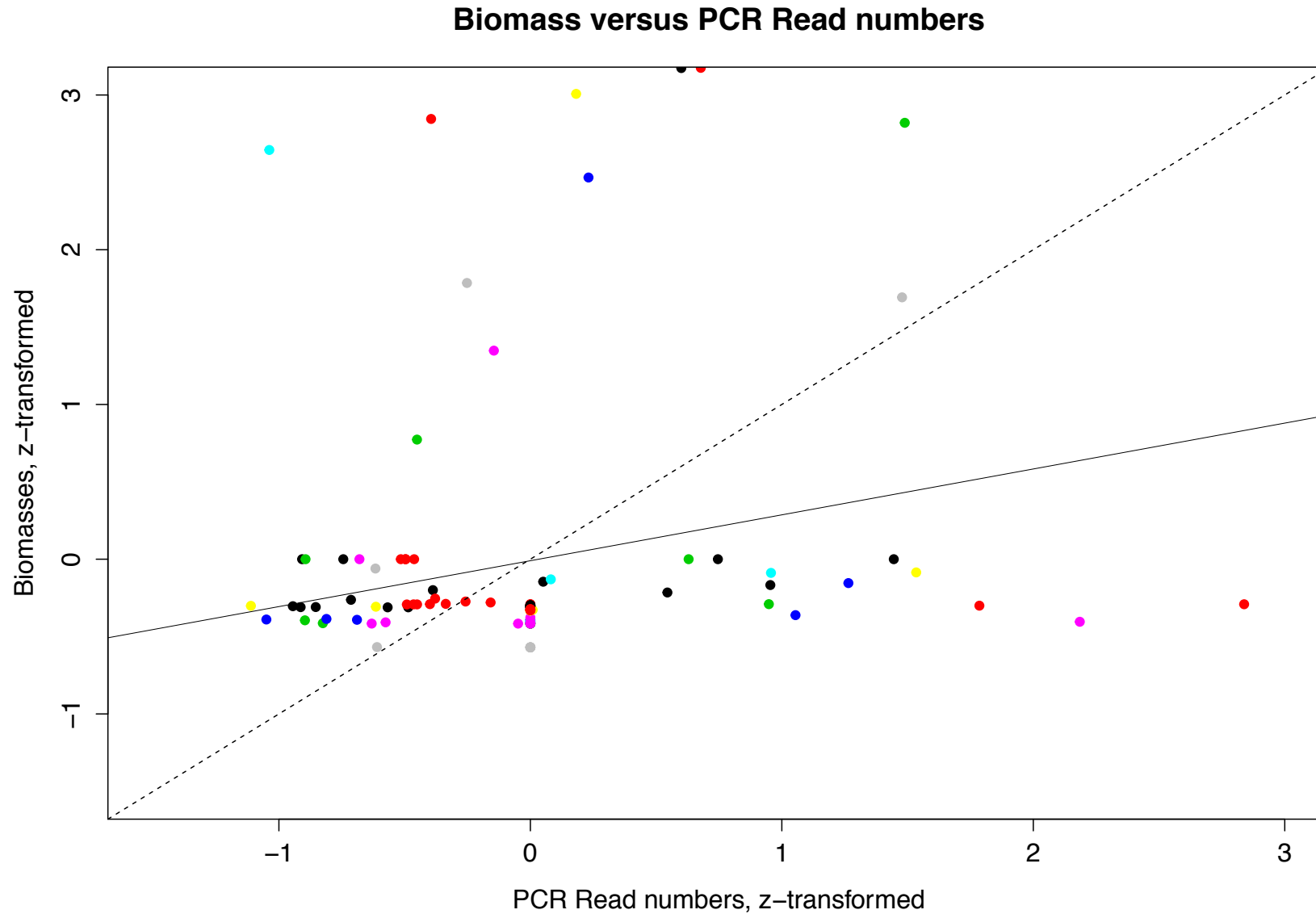
Figure S6. Community analyses. Lines connect samples from the same farm. In the left hand column are the results using the morphological dataset (Bee biomass frequencies). In the right-hand column are the results using the metabarcoding dataset (Read frequencies). The top row uses Presence/Absence data. The bottom row uses biomass and read frequencies (Quantitative). The morphological and metabarcoding datasets (comparing left with right) do not organise the samples similarly in the presence/absence datasets (Procrustes $r_{presence/absence}= 0.424$, $p= 0.39$, 9999 permutations) but do show a correlation with the quantitative datasets, which is due to true positive detections being having higher numbers of reads, in general (Table S1; $r_{quantitative}= 0.889$, $p=0.001$). Samples from the same farm (connected with line segments) and geographic locations tend to cluster together. Codes for geographic locations are: CN = Chilterns North; CS = Chilterns South, HD = Hampshire Downs, and LW = Low Weald. Codes for farm management are: CG = Conservation Grade farm; OELS = Organic + Entry-Level-Scheme farm; ELS = Entry-Level-Scheme farm.
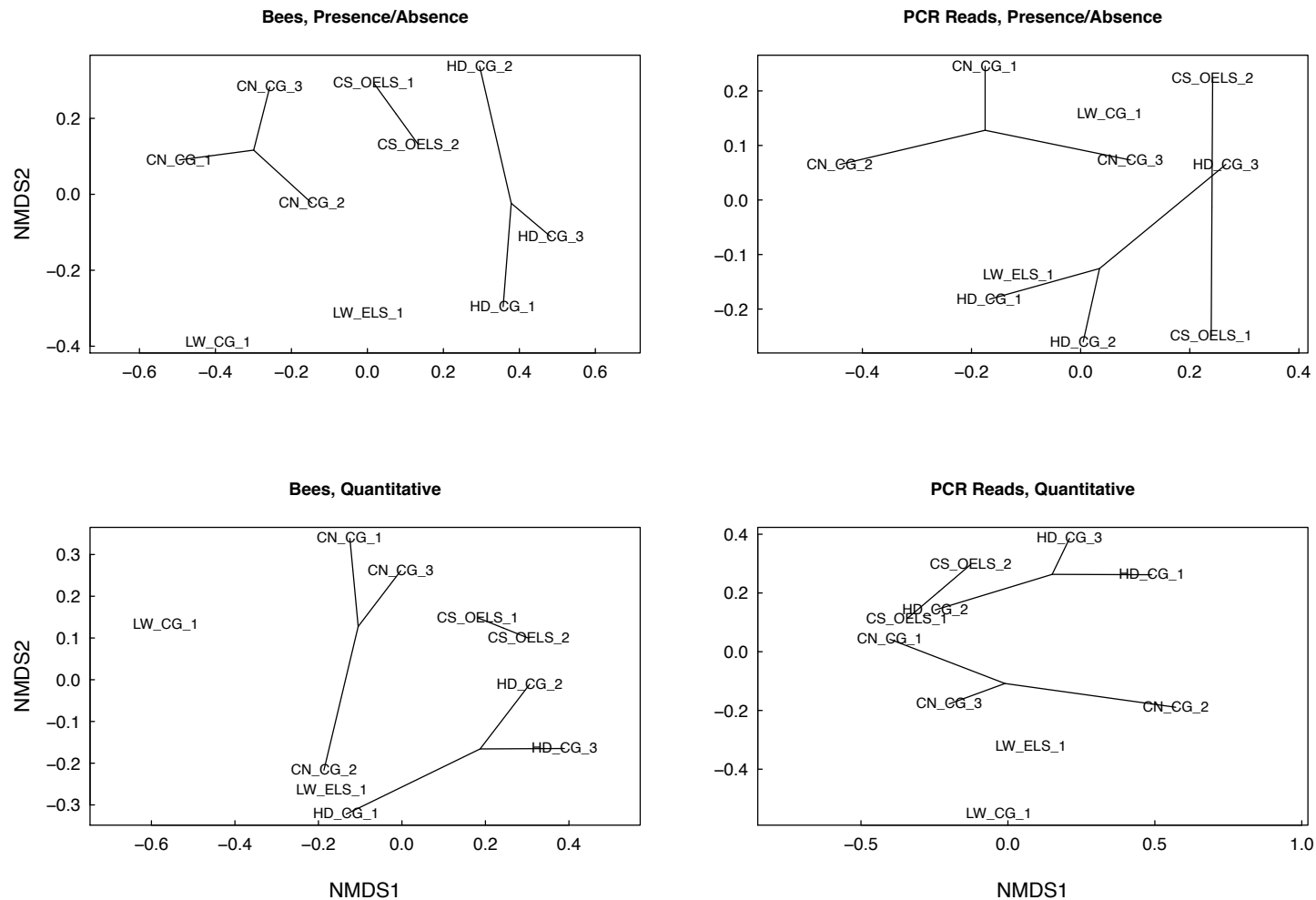
Figure S7. The mt-genome sequences of 48 reference bee species were used to design specific primers for *Bombus lucorum* (LucorumF: 5' GGTTCATCAATAAGTTTAC 3' and LucorumR: 5' TATGCTCGTGTATCTACA 3'). The target fragment length is 787 bp (not including primers). The 33 UK bee species that were detected in the mitogenome dataset (including *Bombus lucorum*) and the 10 mixed samples were tested with the *B. lucorum*-specific primers. PCRs were performed in 30 μL reaction volumes containing 3 μL of 10× buffer, 1.5 mM MgCl₂, 0.2 mM dNTPs, 0.2 μM each primer, 0.9 U ExTaq DNA polymerase (TaKaRa Biosystems, Dalian, China) and approximately 10 ng of genomic DNA. We used a thermocycling profile of 95 °C for 2 min; 35 cycles of 95 °C for 15 sec, 52 °C for 15 sec, 72 °C for 1 min; and a final extension of 72 °C for 10 min. PCR products were visualized on 2% agarose gels.

The two samples for which the mitogenomic dataset detected *Bombus lucorum* are coded here as M6 and M7, and these two samples are the only two that produced PCR product.

Gel figures of the PCR products:
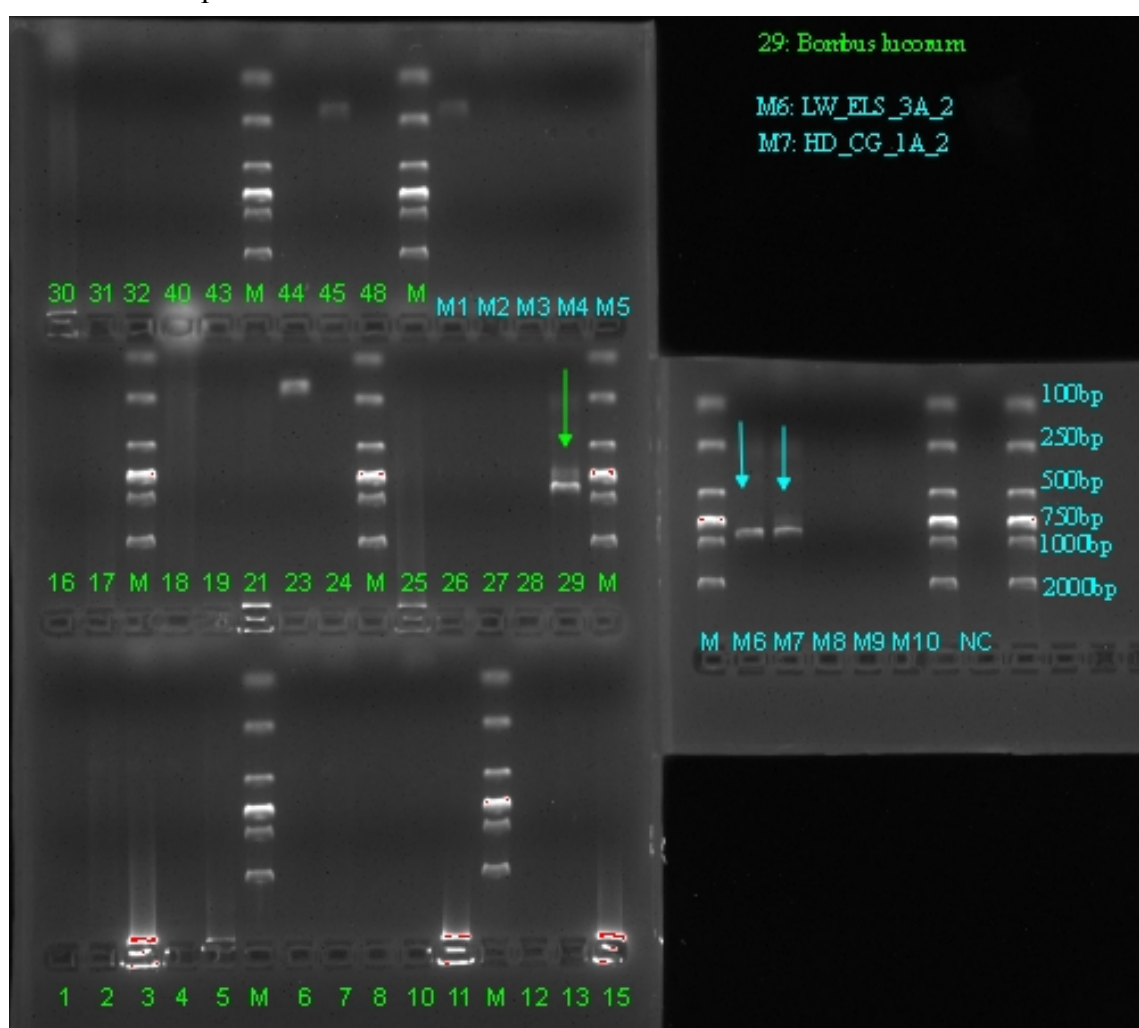
Label key:
1 - *Lasioglossum calceatum*
2 - *Lasioglossum malachurum*
3 - *Andrena nitida*
4 - *Andrena nigroaenea*
5 - *Andrena cineraria*
6 - *Andrena haemorrhoa*

7 - *Andrena bicolor*
8 - *Andrena minutula*
10 - *Bombus lapidarius*
11 - *Lasioglossum leucopus*
12 - *Andrena dorsata*
13 - *Bombus pascuorum*
15 - *Andrena chrysosceles*
16 - *Lasioglossum pauxillum*
17 - *Lasioglossum parvulum*
18 - *Bombus hortorum*
19 - *Andrena flavipes*
21 - *Halictus rubicundus*
23 - *Lasioglossum minutissimum*
24 - *Bombus terrestris*
25 - *Sphecodes ephippius*
26 - *Halictus tumulorum*
27 - *Nomada goodeniana*
28 - *Nomada ruficornis*
29 - *Bombus lucorum*
30 - *Bombus pratorum*
31 - *Andrena subopaca*
32 - *Andrena semilaevis*
40 - *Nomada fabriciana*
43 - *Lasioglossum leucozonium*
44 - *Sphecodes miniatus*
45 - *Lasioglossum punctatissimum*
48 - *Apis mellifera*
M1 – CN_CG_1
M2 – CN_CG_2
M3 – CN_CG_3
M4 – CS_OELS_1
M5 – CS_OELS_2
M6 – LW_ELS_1 (working code: LW_ELS_3A_2)
M7 – HD_CG_1 (working code: HD_CG_1A_2)
M8 – HD_CG_2
M9 – HD_CG_3
M10 – LW_CG_1
NC – negative control
M – DL2000 marker (TaKaRa Biosystems, Dalian, China)