Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics (RevMet)

Ned Peel[1,2], Lynn V. Dicks[2], Matthew D. Clark[1,3], Darren Heavens[1], Lawrence Percival-Alwyn[1], Chris Cooper[4], Richard G. Davies[2], Richard M. Leggett[1], Douglas W. Yu[2,5,6,*]

[1] Earlham Institute, Norwich Research Park, Norwich, UK

[2] University of East Anglia, Norwich Research Park, Norwich, UK

[3] Natural History Museum, London, UK

[4] University of Cambridge, Cambridge, UK

[5] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

[6] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

*Corresponding author: douglas.yu@uea.ac.uk

Running title: Characterisation of mixed pollen with RevMet

Keywords: bees, diet analysis, genome skim, metabarcoding, metagenomics, MinION, pollen, quantitative

Abstract

1. The ability to identify and quantify the constituent plant species that make up a mixed-species sample of pollen has important applications in ecology, conservation, and agriculture. Recently, metabarcoding protocols have been developed for pollen that can identify constituent plant species, but there are strong reasons to doubt that metabarcoding can accurately quantify their relative abundances. A PCR-free, shotgun metagenomics approach has greater potential for accurately quantifying species relative abundances, but applying metagenomics to eukaryotes is challenging due to low numbers of reference genomes.

2. We have developed a pipeline, RevMet (Reverse Metagenomics), that allows reliable and semi-quantitative characterization of the species composition of mixed-species eukaryote samples, such as bee-collected pollen, without requiring reference genomes. Instead, reference species are represented only by 'genome skims': low-cost, low-coverage, short-read sequence datasets. The skims are mapped to individual long reads sequenced from mixed-species samples using the MinION, a portable nanopore sequencing device, and each long read is uniquely assigned to a plant species.

3. We genome-skimmed 49 wild UK plant species, validated our pipeline with mock DNA mixtures of known composition, and then applied RevMet to pollen loads collected from wild bees. We demonstrate that RevMet can identify plant species present in mixed-species samples at

proportions of DNA ≥1%, with few false positives and false negatives, and reliably differentiate species represented by high versus low amounts of DNA in a sample.

4. RevMet could readily be adapted to generate semi-quantitative datasets for a wide range of mixed eukaryote samples. Our per-sample costs were £90 per genome skim and £60 per pollen sample, and new versions of sequencers available now will further reduce these costs.

Introduction

Pollination is a key ecosystem service; almost 90% of all flowering plant species, including 75% of food crops (mainly fruits, nuts, and vegetables), rely on animal pollination (Ollerton, Winfree, & Tarrant, 2011; Klein *et al.,* 2007). The benefits of pollinators, and pollinator-dependent plants, also include the production of medicines, biofuels, fibres, and construction materials (Potts *et al.,* 2016). There is growing concern over the decline of wild and domesticated pollinators and the resulting decrease in pollination services and crop production (Potts *et al.,* 2010; Burkle, Marlin, & Knight, 2013). These declines are thought to be caused by multiple threats acting together, including habitat loss, climate change, and the spread of diseases (Vanbergen *et al.,* 2013).

To mitigate drivers of pollinator decline, the Intergovernmental Science - Policy Platform for Biodiversity and Ecosystem Services (IPBES) has suggested three complementary strategies: (1) ecological intensification, which involves boosting agricultural production by increasing the provision of supporting ecological processes such as biotic pest regulation, nutrient cycling, and pollination (Bommarco, Kleijn, & Potts, 2013; Tittonell, 2014); (2) strengthening existing

diversified farming systems, including gardens and agroforestry, for the generation of ecosystem functions; and (3) investment in ecological infrastructure, to protect, restore, and connect natural and semi-natural habitats across agricultural landscapes, so that pollinator species can more easily disperse and find nesting and floral resources (IPBES 2016).

However, knowledge gaps limit the effectiveness of these strategies (Wood, Holland, & Goulson, 2015; Dicks *et al.,* 2013). For instance, it is still not clear which plant species are the most valuable food resources and how plant species vary in value across pollinator species, over time, and in different environmental conditions. It is also not well understood whether the addition of floral resources might draw pollinators away from pollinator-dependent crop plants (Morandin & Kremen, 2013), or whether floral enhancement will alter levels of plant-target specialism, at the levels of insect species and of individual insects, resulting in changes in pollination efficiency (Lucas *et al.*, 2018; Morales & Traveset, 2008).

Therefore, a crucial technical challenge for understanding plant-pollinator interactions is to develop a method to identify *and* quantify the species of pollen that are consumed by pollinators. Identifying and quantifying pollen has traditionally been carried out by using light microscopy to distinguish plant species by grain morphology, a labour-intensive technique that requires expert knowledge and lacks discriminatory power at lower taxonomic levels (Long & Krupke, 2016; Khansari *et al*., 2012). In contrast, high-throughput DNA sequencing now allows pollen identification without expert knowledge of pollen morphology and taxonomy.

The currently dominant sequence-based method is metabarcoding, which involves amplifying taxonomically informative marker genes from mixed samples via polymerase chain reaction (PCR) (Ji *et al.*, 2013). The resulting amplification products, known as amplicons, are sequenced, and the reads are assigned to taxonomies by matching against barcode databases, such as the Barcode of Life Data System (Ratnasingham & Hebert, 2007). Notably for plants, there is no single barcode gene that matches the resolving power of 16S rRNA for prokaryotes and Cytochrome Oxidase (CO1) for animals (Hollingsworth, Li, Van Der Bank, & Twyford, 2016). Instead, plant-related barcoding studies rely on a combination of marker genes, which include plastid regions *rbcL* and *matK* and the internal transcribed spacer (ITS) regions of nuclear ribosomal DNA (Li *et al.,* 2015; Hollingsworth *et al.,* 2016). Metabarcoding of mixed-species pollen samples can reveal the presence and absence of constituent plant species (or genera), but there are strong reasons to doubt that metabarcoding can accurately quantify their relative abundances, due to PCR amplification biases and varying copy numbers of barcode loci (Keller *et al.,* 2015; Richardson *et al.,* 2015; Sickel *et al.,* 2015; Bell *et al.,* 2017, 2018; Lamb *et al*., 2018).

In contrast to the targeted-sequencing approach of metabarcoding, 'shotgun metagenomics' involves randomly sequencing short stretches of genomic DNA from mixed samples. In standard metagenomics, these short reads ('queries') are mapped to either assembled genomes or to collections of barcode genes ('references'), which creates a requirement for large numbers of reference genomes (Sharpton, 2014) or barcodes (Zhou *et al.*, 2013), with the latter being very inefficient (Ji *et al.*, 2019). Species identification is obtained by first calculating a similarity metric between each short read and each reference sequence (e.g. % identity) and then using an

algorithm to assign each short read to the most likely reference sequence (Quince, Walker, Simpson, Loman, & Segata, 2017). The potential key advantages of shotgun metagenomics are that it can avoid the PCR-induced biases seen with metabarcoding, especially if PCR-free library preparation protocols are used (see Nayfach & Pollard, 2016; Jones *et al.,* 2015) and that by sampling across the whole genome, variation in the copy numbers of a few loci is rendered less important. However, the requirement for reference genomes means that most shotgun metagenomics studies focus on prokaryotic organisms, since large numbers of prokaryote reference genomes are available. In contrast, eukaryotes are not well represented in sequence databases and as a result have mostly been neglected in metagenomic studies (Escobar-Zepeda, De León, & Sanchez-Flores, 2015). The low numbers of reference genomes for eukaryotic species is because they are more expensive to sequence and assemble (Gilbert & Dupont, 2011).

Here we demonstrate a metagenomic pipeline for eukaryotes that avoids the need to assemble reference genomes. Instead, each reference species is represented by a 'genome skim' (Straub *et al*., 2012), which is a low-cost, low-coverage, shotgun dataset, i.e. simply a set of short reads. We use these sets of short reads to identify individual *long reads* from pollen that have been generated by sequencing mixed-species pollen loads with the Oxford Nanopore Technologies' (ONT) MinION, a nanopore sequencing device (for a review of MinION applications and performance, see Leggett & Clark, 2017). Here, we generate reference genome skims for 49 wild UK plant species, and we use them to identify and quantify plant species in two kinds of query samples: mock, mixed-plant-species DNA mixtures of known composition and mixed-species pollen samples collected from wild bees. Each of the long reads in the query samples are individually classified and we show that the proportion of long reads assigned to a plant species

is a reasonably accurate estimate of that species' frequency in a mixed-species sample, based on relative quantities of DNA. We call this pipeline Reverse Metagenomics, or RevMet, because we map reference sequences to query sequences, which is the reverse of the normal metagenomic protocol.

Methods

Sampling of bees and plant tissue

Sample collection took place in the Pensthorpe Natural Park area (52°49'23"N, 0°53'14"E) of Norfolk, UK, over four days in June and July 2016. Leaf samples were collected from all plant species with open flowers, including grasses and trees, within a 100 m radius of the collection site (n = 49 species). The 100 m radius was chosen to capture the likely area of flowering plants covered by an individual bee in a pollen-foraging bout. We assume that bees actively collecting pollen are on 'exploitation flights', defined for bumblebees by Woodgate *et al.* (2016) as single loop flights to a previously known location for the sole purpose of foraging, rather than 'exploration flights', which cover a much larger area. In the data reported by Woodgate *et al.* (2016), foraging activity on *Bombus terrestris* exploitation flights was usually constrained within a circle of radius 100 m.

Leaf tissue was preserved on dry ice in the field followed by storage at -80 °C. Foraging wild bees (n = 48: 9 *Apis mellifera*, 27 *Bombus terrestris/lucorum complex*, 12 *Bombus lapidarius*) were collected with hand nets or into falcon tubes directly from flowers and euthanized in falcon

tubes containing ethanol-soaked tissue paper. Pollen loads were scraped from bee corbiculae using a mounted needle and stored in absolute ethanol. The plant species on which each bee was foraging when collected was recorded.

Leaf tissue DNA extraction, library preparation, and Illumina sequencing

Leaf tissue from each of the 49 plant species was disrupted by bead-beating using a 4-mm stainless steel bead with a Qiagen TissueLyser II running at 22.5 Hz for 4 min, rotating the adapter sets after 2 min. DNA was extracted using the DNeasy Plant Kit (Qiagen, Hilden, Germany) following manufacturer's instructions. DNA concentrations were measured on a Qubit 2.0 fluorometer (ThermoFisher, Waltham, USA) using the dsDNA HS assay kit, and fragment size distribution was checked with a Genomic DNA Analysis ScreenTape on the TapeStation 2200 (Agilent, Santa Clara, USA).

The Earlham Institute (Norwich, UK) applied a modified version of Illumina's Nextera protocol, known as Low Input Transposase Enabled (LITE) protocol (Beier *et al.,* 2017), to generate a separate sequencing library for each leaf sample, targeting an average insert size of 500 bp. The LITE libraries were then pooled based on estimated genome sizes (Supplementary Table S1), obtained from the Royal Botanic Gardens Kew Plant DNA C-values database (Bennett and Leitch, 2012), in order to achieve 0.5x coverage of each species genome. The pooled libraries were sequenced on one lane of Illumina HiSeq 2500 in Rapid Run mode (250 bp PE).

Construction and sequencing of mock pollen samples

DNA from twelve of the 49 plant species were used to construct six mock communities. Each mock was made using 200 ng DNA in total, with species added at different proportions: 0.08% to 45.25% (Table 1). For each mock, technical-replicate pairs were prepared using ONT's (Oxford, UK) Rapid Barcoding Sequencing Kit (SQK-RBK001), following the RBK_9031_v2_revl_09Mar2017 version of the manufacturer's protocol. The 12 libraries (six mocks, duplicated) were sequenced on a single MinION R9.5 flow cell (FLO-MIN107).

Bee-collected pollen DNA extraction, library preparation, and MinION sequencing

After removing storage ethanol from the 48 bee-collected pollen loads, the pollen was disrupted with ca. five 1-mm stainless steel beads for 2 min at 22.5 Hz using a Qiagen TissueLyser II, rotating the adapter sets after 1 min. The pollen samples were resuspended in 600 µl CTAB extraction buffer (2% CTAB, 1.4 M NaCl, 20 mM EDTA, pH 8.0, 100 mM Tris-HCl pH 8.0), 0.5 µl of β-Mercaptoethanol, 4 µl of proteinase K, and vortexed for 5 s. Following a 1 hr incubation at 55 °C, the tubes were centrifuged for 6 min at 18,000 x g. The ≈500 µl of supernatant was extracted to a clean 1.5 ml tube before an equal volume of chilled (2-8 °C) Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v) was added to the lysate. The samples were vortexed for 10 s (5 x 2 s bursts), centrifuged for 5 min at 14,000 x g, and the upper aqueous phase (≈ 420 µl) was extracted by pipette and transferred into a clean 1.5 ml tube.

An equal volume of Agencourt AMPure XP beads was added to each sample, vortexed for 20 s (10 x 2 s bursts), and then incubated for 10 min at room temperature. By placing the samples

onto a magnetic tube rack for 5 min, the beads were separated from the solution, and the cleared supernatant was removed by aspiration. The beads were washed twice using the following protocol: 1 ml of 80% ethanol was added, incubated at room temperature for 30 s, and then removed, followed by air drying for ≈ 3 min. The magnetic beads were resuspended in 55 µl of EB (Elution Buffer: 10 mM Tris-HCl) and incubated at 37 °C for 10 min. The tubes were placed back onto the magnetic rack to bind the beads, and the eluted DNA (≈ 50 µl) was transferred into fresh tubes. A 1 µl aliquot of 1-in-10 diluted Qiagen RNase A was added to each DNA sample before being incubated for 30 min at 37 °C. The concentration of the eluted DNA was assessed using the dsDNA HS assay on a Qubit 2.0 fluorometer. To check the DNA for degradation, fragment size distributions were checked with a TapeStation 2200 using the Genomic DNA Analysis ScreenTape.

Finally, the extracted DNA was prepared and sequenced using the same protocol as used for the DNA mocks above, except that only one library was prepared for each sample. Twelve samples can be multiplexed using the Rapid Barcoding Sequencing Kit; we thus required four flow cells. Due to continuous software upgrades by ONT, the specific software versions of *MinKNOW* varied across runs and is recorded in the final sequence files (fast5 format), which are available from the EBI's European Nucleotide Archive (see Data accessibility).

Illumina and MinION read pre-processing

Duplicate reads were removed from the 49 plant-reference Illumina datasets using *NextClip 1.3.2* (Leggett, Clavijo, Clissold, Clark, & Caccamo, 2014), and then *cutadapt 1.10* (Martin, 2011)

was used to trim Illumina adaptors and filter out reads shorter than 100 bp. The resulting unmerged FASTQ files constitute our 49 *reference skims*.

The MinION datasets from the 12 mocks and the 48 pollen loads were basecalled and demultiplexed with *albacore 2.1.10* (ONT). The resulting FASTQ files were converted to FASTA format. We removed long reads deriving from plant organelles because they are highly conserved across plant species and in pilot tests, we observed that mapping to organellar long reads resulted in a higher rate of incorrect assignments than mapping to nuclear long reads (data not shown). NCBI Entrez (https://www.ncbi.nlm.nih.gov/sites/batchentrez) was used to download 2,583 Land Plant organelle genomes, 2,357 plastid and 226 mitochondrial. Organelle reads were identified by aligning each of the MinION datasets to the organellar genomes using *minimap2 2.7* (Li, 2018) and removed from the FASTA files. The resulting 60 (= 12 + 48) organelle-filtered FASTA files constitute our mock and pollen *query* datasets, and in the next step, we used the 49 plant reference skims to assign a taxonomy to each long read in the mock and pollen query datasets (Fig. 1c).

Taxonomic assignment of mock-sample and bee-collected pollen MinION reads

We used *bwa mem* 0.7.17 (Li, 2013) to map the Illumina reads from each of the 49 reference skims against every individual long MinION read in each of the mock and bee-collected pollen datasets. We used *SAMtools* 1.7 (Li *et al.*, 2009) to remove unmapped reads and secondary and supplementary alignments. After SAMtools indexing, the depth of mapping coverage at each long-read position was calculated using the SAMtools depth function. A python script,

percent_coverage_from_depth_file.py, was used to calculate the 'percent coverage' for each long read - defined as the fraction of nucleotide positions that were mapped to by one or more reference-skim Illumina reads. We assigned each long read to the plant species that mapped with the highest percentage coverage (Fig. 1C), unless the highest percent coverage was <15%, in which case the long read's identity was judged ambiguous and left unassigned. Additionally, for clarity of presentation, we implemented a 1% minimum-abundance filter, removing plant species represented by fewer than 1% of the total assigned long reads in each sample.

All of the bioinformatic steps for taxonomic assignment can be run on a laptop/desktop computer, but we ran the pipeline on a high performance computing cluster.

Reference-skim subsampling

To estimate a minimum recommended depth of coverage needed per reference skim, we subsampled one of the genome skims, *Knautia arvensis*, which is a major constituent species in mock mixes MM1 and MM2. We randomly subsampled this skim from its maximum of 0.65x down to 0.05x, in steps of 0.05x using a custom script. For each subsample, the whole pipeline was re-run along with the full reference skims of the other 48 plant species. The number of mock reads assigned to *Knautia arvensis*, and the number of unassigned reads, at each level of coverage was recorded. This subsampling was repeated three times (Supplementary Fig. S1).

Network construction

We constructed a pollinator-plant network diagram for the 48 wild-bee pollen samples, using the *bipartite 2.11* package (Dormann, Frund, Bluthgen, & Gruber, 2009) for the *R* statistical language (R Core Team, 2018). For presentational clarity, we only show plant species represented by more than 10% of the assigned reads in each sample.

Results

A reference set of plant genome skims

Low genome-coverage, short-read, shotgun-sequencing datasets ('reference skims') were successfully generated for all 49 plant species (Fig. 1a). After pre-processing, the mean estimated coverage was 0.6x (0.1 to 1x, details in Supplementary Table S1).

Mock DNA mixes

The six mock communities, each with two technical replicates, were sequenced on a MinION. These produced relatively short reads for nanopore sequencing, with mean length 1914 bp (longest 41,058 bp), likely due to the low mass and molecular weight of the input DNA (discussed later). After demultiplexing, 88.8% of the reads could be assigned to one of the 12 mock mixes, with the remaining reads left unclassified. Sequences originating from organellar genomes made up between 5.1% (MM4.2) to 10.2% (MM3.2) of the reads in the mocks and were removed. The remaining number of reads per mock ranged from 733 (MM2.1) to 2174 (MM4.1), mean 1347.

298

299 **Taxonomic assignment of mock-sample MinION reads**

300

301 The 49 reference skims were separately mapped to each long read in each of the 12 mock mixes,
302 and each long read was assigned to the plant species that mapped with the highest percent
303 coverage, or left unassigned if the highest coverage was <15%. In total, 65.5% of the mock reads
304 were assigned to a plant species, with 94.7% of those reads being assigned to a species known to
305 be present in that mock sample. Almost all (93.4%) of the 563 false-positive read assignments
306 were made to one species, *Ranunculus acris*, and all these assignments occurred in the mock
307 samples that contained the very closely related species *Ranunculus repens*. We return to this in
308 the Discussion. The few other false-positive assignments all occurred at a rate of less than 1% of
309 the assigned long reads in their mixes and for presentational clarity are not shown in Fig. 2. The
310 full results are in Table S2.

311

312 All of the plant species that had been added to the mock compositions at proportions ≥1% were
313 detected by our method in at least one of the two replicates, and in all cases, the frequencies of
314 long reads assigned to each plant species were reliably 'semi-quantitative', in that they
315 differentiated low- and high-abundance plant species (Fig. 2). In general, the technical replicates
316 showed a high level of repeatability, although in two of the mocks there was one species in each
317 that was detected in only one of the two replicates (*Lotus corniculatus* in MM2.1 and *Digitalis*
318 *purpurea* in MM3.2*)*. This is not too surprising, as *L. corniculatus* and *D. purpurea* were only
319 expected to be present at 3.0% and 4.6%, respectively. That said, both of these species were

consistently underrepresented across our mock data sets, which suggests that the DNA quantification may have been inaccurate prior to the creation of the mocks.

Reference-skim subsampling

As expected, the larger the reference-skim dataset size for *Knautia arvensis*, the more reads in the MM1 and MM2 mocks were assigned to this species and the fewer reads left unassigned. Importantly, the rate of increase was decelerating (Supplementary Fig. S1); over half of the MinION reads that were assigned to *Knautia arvensis* with a 0.65x genome skim could also be assigned with just a 0.1x skim, even though all the other reference skims in the mapping run were kept at their original sizes.

Taxonomic assignment of bee-collected pollen MinION reads

The 48 bee-collected pollen loads harvested from the corbiculae of three species, *Apis mellifera*, *Bombus terrestris/lucorum* complex, and *Bombus lapidarius*, yielded DNA quantities ranging from 191 to 3750 ng, and all successfully produced libraries, demonstrating that pollen carried by individual bees can provide sufficient DNA for MinION sequencing. After demultiplexing, the mean read number per pollen sample was 2430, with an average length of 2300 bp (longest 51,629 bp).

As with the 12 mocks, each of the reference skims was aligned to each long read in each of the 48 pollen samples, the long reads were either assigned to the plant species achieving the highest

percent coverage or left unassigned, and any plant species assigned fewer than 1% of the long reads in each bee-collected pollen sample was filtered out (Supplementary Table S3). In total, 49.7% of the long reads were assigned to one of the reference plant species. In 38 of the 48 bees (79.2%), pollen from the plant species on which each bee was captured was found to be present in that bee's pollen load (Supplementary Table S3).

Each of the 48 pollen loads was found to contain one or two major plant species (defined as read frequency ≥10%) (Fig. 3a). All nine of the *Apis mellifera* pollen loads contained a single major species, whereas 16 of 27 *Bombus terrestris/lucorum complex* and 6 of 12 *Bombus lapidarius* pollen loads were comprised of two major species (Fig. 3a). These differences in mean number of major species were statistically significant (*Apis mellifera* versus *Bombus terrestris/lucorum complex* (Welch's t-test, t = -6.15, df = 26, p-value < 0.0001) and versus *Bombus lapidarius* (t = -3.32, df = 11, p-value < 0.01)) (Fig. 3b). Another way of visualising the wild-bee results is as a plant-pollinator network graph (Fig. 3c). Overall, 6 of the 49 reference plant species were identified as major components in the 48 pollen loads, and the majority of bee-collected pollen samples were dominated by one plant species.

Discussion

Using light microscopy to identify plant species from pollen requires expert knowledge and is costly when applied to many samples (Khansari *et al.,* 2012). There is a need for a quick and low-cost method that can be scaled to large numbers of pollen samples. Metabarcoding is the current leading candidate, but there are concerns over its discriminatory power at lower

366 taxonomic levels, and there is good reason to believe that metabarcoding does not return reliable
367 quantitative data (Keller *et al.,* 2015; Richardson *et al.,* 2015; Sickel *et al.,* 2015; Bell *et al.,*
368 2017, 2018, Lamb *et al*., 2018). A PCR-free shotgun-metagenomics approach has greater
369 potential for providing reliable quantitative analysis with high power for resolving species.
370 However, applying shotgun metagenomics to eukaryotes is challenging due to the lack of
371 reference genomes (Gilbert & Dupont, 2011). We have developed a metagenomics method that
372 avoids the need for reference genomes. Instead, each reference species is represented by just a
373 low-cost genome skim, and we use a set of such skims to identify individual long reads from
374 pollen samples, produced by the MinION sequencer.
375
376 We evaluated our RevMet pipeline with mock DNA mixtures of known composition and then
377 applied the pipeline to pollen collected from wild bees. Our main findings are:
378
379 1) RevMet can identify plant species present in mixed-species samples at proportions of
380 DNA ≥1%, with few false positives and false negatives, and can reliably differentiate
381 species represented by high versus low amounts of DNA in a sample (Fig. 2,
382 Supplementary Table S2).
383 2) Genome skims with sequence coverage as low as 0.05x can be used for detecting
384 species presence and for estimating relative abundance in terms of DNA mass.
385 Increasing skim coverage increases detection power, at a decelerating rate
386 (Supplementary Fig. S1).
387 3) Individual pollen loads collected from wild *Apis* and *Bombus* bees yield enough DNA
388 for MinION sequencing (Supplementary Table S3) and generate plausible plant-

pollinator networks, as evidenced by the fact that (a) 56.3% of the plant species on which the bees were collected were also the dominant constituent of the corresponding pollen sample (and 79% of plant species on which the bees were collected were detected in the corresponding pollen sample) (Supplementary Table S3), and (b) pollen species richnesses and compositions were more similar within bee species than across bee species (Fig. 3).

4) Our per-plant-species cost of a reference skim was £90, and our per-pollen-sample cost was £61, including DNA extraction, library preparation, and sequencing. Sequencing costs will likely drop further, given the new Illumina NovaSeq and new MinION 'Flongle'.

Semi-quantitative species compositions. – We were able to assign roughly 65% of the mock-mix MinION reads and just under 50% of the pollen-load MinION reads to our reference plant species. Importantly, the frequencies of MinION reads that were assigned to each reference plant species were reliably *'semi-quantitative*', that is, able to differentiate low- and high-frequency plant species, based on DNA mass (Fig. 2). Within low- and high-abundance categories, accuracy was lower. For example, in mock sample MM1, *Knautia arvensis*, *Galium verum*, and *Crepis capillaris* were the three high-abundance species (each representing 30.3% of total input DNA mass each), and *Papaver somniferum*, *Anagallis arvensis*, and *Sambucus nigra* were the three low-abundance species (each representing 3.0% of total input DNA mass each). The RevMet pipeline estimated the three high-abundance frequencies at means of 34.0%, 14.7%, and 44.0%, and the three low-abundance species at 1.4%, 3.0%, and 3.0%, respectively (Supplementary Table S2).

412

413 There are at least three reasons for the remaining quantitative error. First, although we targeted
414 0.5x per reference skim, coverage still varied across species (Table S1), resulting in different
415 powers of discrimination, as shown by the experiment with subsamples of *Knautia arvensis*
416 (Supplementary Fig. S1). Fortunately, we found that even very low-depth skims of 0.05x are
417 useful for species detection and are probably still useful for differentiating rare from abundant
418 species (albeit with more error) (Supplementary Fig. S1). Genome sizes are also estimated with
419 error, so it is also helpful that the subsampling experiment suggests that detection power
420 asymptotes with higher sequencing depth (Supplementary Fig. S1), and as sequencing costs fall
421 further, we expect that the most robust protocol will be to target 1x coverage.

422

423 Second, very closely related species can generate false positives. Our reference-skim database
424 included six congener pairs, and we included two of the pairs (*Papaver* and *Ranunculus*) in the
425 mock mixes. In the case of *Papaver*, there were no *P. rhoeas* false-positives greater than the 1%
426 minimum-abundance filter in the mocks that contained *P. somniferum* (MM1 and MM6)
427 (Supplementary Table S2). In contrast, *Ranunculus acris* was regularly incorrectly assigned to
428 reads in mock mixes that contained the closely related congener *Ranunculus repens*. In fact,
429 almost all the false-positive assignments (93.4%) were to *R. acris*. In retrospect, this result is
430 expected because these two species are not easily differentiated by pollen morphology (Forup &
431 Memmott, 2005), floral morphology, or even DNA barcodes (*rbcL* (99.1% similarity), *matK*
432 (96.9%), *ITS2* (95.5%)). In other words, the RevMet results are correctly telling us that the two
433 *Ranunculus* species are closely related. We did not run negative controls through our lab
434 pipeline, because trace contaminants can only show up at low levels, if at all, in metagenomic

assays, but for production use, we encourage the addition of negative controls to provide a chance of detecting major episodes of contamination.

Third, MinION reads have relatively high error rates of roughly 5 to 10% depending on the flow cell and kit used (Leggett & Clark, 2017). Although this is dropping over time, this error rate unavoidably obscures differences between species (although not enough to confound the two *Papaver* species). We note that one of the advantages of the RevMet approach is that we use percent coverage as a predictor of species presence (Fig. 1C). Using mapped read counts alone, we observed several instances of low numbers of long reads being given false-positive assignments (data not shown). The percent-coverage filter requires many different reference-skim reads to independently identify a species before an assignment is made.

For studies of pollen collected at the colony or nest level, which sum multiple, individual foraging bouts over an entire foraging range, we recommend collecting all the flowering species within a radius of at least 1 km (Dicks et al., 2015).

Reference-skim cost. – The RevMet pipeline is relatively low cost. In our study, we generated skims for 49 plant species, with genome sizes ranging from roughly 290 Mb (*Epilobium hirsutum*) to just under 15 Gb (*Sambucus nigra*), targeting 0.5x coverage. All skims were produced on a single lane of Illumina HiSeq 2500 in Rapid Run mode (250 bp PE) at a mean coverage of 0.57x. The average cost per skim in this study was just under £90, which includes the DNA extraction, LITE library preparation, sequencing, and data QC. The per skim cost will be lower in studies with smaller eukaryotic genomes, and with Illumina's newer sequencer, the

458 NovaSeq 6000, we estimate equivalent skims will cost ~£50 (250 PE with the SP flow cell).
459 Genome-assembly campaigns will also produce numerous short-read datasets for free download.
460
461 *Long-read MinION cost*. – We used ONT's first iteration of the Rapid Barcoding Kit (RBK001),
462 which relies on transposase to randomly fragment DNA and simultaneously add barcoded
463 adapters. Longer read lengths have an increased likelihood of accurate species assignment
464 because they carry more sequence information. The two main ways to obtain longer reads with
465 transposase-based preparations are to: (1) increase the ratio of DNA to transposase e.g. by
466 increasing the input material or by heat killing a proportion of the transposase (which also lowers
467 sequencing yields); and (2) use higher molecular weight input DNA. Since the release of
468 RBK001, ONT's chemistry has evolved, and their Rapid-based kits have seen greater sequencing
469 yields. However, the recommended input for the latest iteration of the Rapid Barcoding Kit
470 (RBK004) is now higher, 400 ng of DNA per sample. That said, we anticipate that input
471 biomasses similar to those used in this study, 200 ng, will still be adequate. Also, even 400 ng is
472 achievable, as 36 of 48 of our wild-bee pollen samples yielded >400 ng (Supplementary Table
473 S3). ONT have also recently released the Flongle, which is an adapter that enables smaller and
474 cheaper (~$90) flow cells to be used on the MinION. Our results suggest that ONT's target yield
475 of 1 Gb per Flongle flow cell will be more than enough for multiplexing twelve bee-collected
476 pollen loads, reducing per-sample costs from the £61 in this study to just under £16.
477
478 *Application to pollen collected from wild bees*. – The RevMet pipeline detected consistent
479 differences in the composition of pollen loads collected by honeybees *Apis mellifera* and by the
480 two bumblebees *Bombus terrestris/lucorum* and *B. lapidarius* (Fig. 3). The low number of plant

481 species identified per pollen load is consistent with the flower constancy that is observed in a
482 range of insect pollinators, in which individuals almost exclusively visit a single flower type
483 during a given foraging trip (Grüter & Ratnieks, 2011). This method can therefore be used to
484 compare pollen collection at the individual-bee scale, across different environmental or seasonal
485 contexts and across species. We expect of course that bulk pollen samples from whole colonies
486 (such as *Apis mellifera* or *Bombus terrestris*) or from nests made by foraging solitary bees (e.g.
487 Sickel *et al.,* 2015) will reveal a higher diversity of food plants, at least for generalist bee
488 species.
489
490 As a proof of concept study, we focused on a small number of bee-collected pollen loads (n =
491 48) sampled from just one site. By generating more plant reference skims and utilising the
492 Flongle for cheaper multiplexing of pollen loads, RevMet could now be applied to compare
493 pollination networks across large-scale spatial and biogeographical gradients. We have
494 demonstrated that the RevMet pipeline can assess DNA composition from read counts. However,
495 there are other potential sources of bias that may have affected our pollen sample proportions,
496 such as the bi- or tri-cellular nature of pollen and differing ploidy levels, genome sizes, and DNA
497 extraction efficiencies. Our next step will be to test RevMet's ability to estimate pollen
498 biomasses.
499
500 The RevMet pipeline can readily be applied to a wide range of research questions. RevMet could
501 potentially be used to quantify the degree to which co-attraction of pollinators leads not to
502 benefits of increased pollinator numbers but to loss of pollination service via competition
503 (Carvalheiro *et al.,* 2014; Pornon *et al.,* 2016). Outside of pollination ecology, there is potential

for semi-quantitative assessments of many other eukaryotic species mixtures, including herbivore diets (Bhattacharyya, Dawson, Hipperson, & Ishtiaq, 2018; Kress, García-Robledo, Uriarte, & Erickson, 2015); plant-fungus interactions (Schröter *et al.,* 2018); allergenic pollen species from air samples - although this might require an additional whole-genome amplification (WGA) step (Kraaijeveld *et al.,* 2015); and algal and diatom communities (Keller *et al.,* 2015). Furthermore, due to the portability and real-time nature of the MinION platform, the method could be optimised for analysis in the field alongside sample collection.

Acknowledgements

This work was supported by the Norwich Research Park Science Links Seed Fund, the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership (grant number BB/M011216/1) and BBSRC Core Strategic Programme Grant BB/CSP17270/1 to the Earlham Institute. LVD is funded by the Natural Environment Research Council (NE/N014472/1). CC was supported by the J. Arthur Ramsay Fund, administered by the Cambridge University Zoology Department. We are grateful for the support of the Earlham Institute Genomics Pipelines group and the NBI Computing infrastructure for Science (CiS) group. We thank Iain Barr for help in fieldwork and Pensthorpe Natural Park for access.

Authors' contributions

MDC, RML, DWY, LVD, and RGD conceived and designed the study. LVD, RGD, and CC collected the samples. NP, DH, and LP performed the experiments. NP, RML, and DWY analysed the data. NP and DWY led the writing of the manuscript. All authors gave final approval for publication.

Data accessibility

The Illumina and MinION datasets are available in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under study accession PRJEB30946. RevMet scripts are available from https://github.com/nedpeel/RevMet and a tutorial using an example dataset can be found at https://revmet.readthedocs.io/en/latest/.

References

Beier, S., Himmelbach, A., Colmsee, C., Zhang, X. Q., Barrero, R. A., Zhang, Q., … Mascher, M. (2017). Construction of a map-based reference genome sequence for barley, *Hordeum vulgare L*. *Scientific Data*, **4**, 170044. https://doi.org/10.1038/sdata.2017.44

Bell, K. L., K. S. Burgess, J. C. Botsch, E. K. Dobbs, T. D. Read, and B. J. Brosi. (2018). Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. *Molecular Ecology,* **28**, 431–455. https://doi.org/10.1111/mec.14840

Bell, K. L., Fowler, J., Burgess, K. S., Dobbs, E. K., Gruenewald, D., Lawley, B., … Brosi, B. J. (2017). Applying pollen DNA metabarcoding to the study of plant-pollinator interactions. *Applications in plant sciences*, **5**, 1600124. https://doi.org/10.3732/apps.1600124

Bennett, M., & Leitch, I. (2012). Plant DNA C-values Database. Available at http://data.kew.org/cvalues/

Bhattacharyya, S., Dawson, D. A., Hipperson, H., & Ishtiaq F. (2018). A diet rich in C3 plants reveals the sensitivity of an alpine mammal to climate change. *Molecular Ecology*, **28**, 250-265. https://doi.org/10.1111/mec.14842

Bommarco, R., Kleijn, D., & Potts, S. G. (2013). Ecological intensification: Harnessing ecosystem services for food security. *Trends in Ecology and Evolution*, **28**, 230-238. https://doi.org/10.1016/j.tree.2012.10.012

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. https://doi.org/10.1126/science.1232728

Carvalheiro, L. G., Biesmeijer, J. C., Benadi, G., Fründ, J., Stang, M., Bartomeus, I., … Kunin, W. E. (2014). The potential for indirect effects between co-flowering plants via shared pollinators depends on resource abundance, accessibility and relatedness. *Ecology Letters,* **17**, 1389-1399. https://doi.org/10.1111/ele.12342

Dicks, L. V., Abrahams, A., Atkinson, J., Biesmeijer, J., Bourn, N., Brown, C., … Sutherland, W. J. (2013). Identifying key knowledge needs for evidence-based conservation of wild insect pollinators: A collaborative cross-sectoral exercise. *Insect Conservation and Diversity*, **6**, 435-446. https://doi.org/10.1111/j.1752-4598.2012.00221.x

Dicks, L. V., Baude, M., Roberts, S. P., Phillips, J., Green, M., & Carvell, C. (2015). How much flower-rich habitat is enough for wild pollinators? Answering a key policy question with incomplete knowledge. *Ecological Entomology*, **40**, 22-35. https://doi.org/10.1111/een.12226

Dormann, C. F., Frund, J., Bluthgen, N., & Gruber, B. (2009). Indices, Graphs and Null Models: Analyzing Bipartite Ecological Networks. *The Open Ecology Journal*, **2,** 7-24. https://doi.org/10.2174/1874213000902010007

Escobar-Zepeda, A., De León, A. V. P., & Sanchez-Flores, A. (2015). The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics*, **6**, 348. https://doi.org/10.3389/fgene.2015.00348

Gilbert, J. A., & Dupont, C. L. (2011). Microbial Metagenomics: Beyond the Genome. *Annual Review of Marine Science*, **3**, 347-371. https://doi.org/10.1146/annurev-marine-120709-142811

Grüter, C., & Ratnieks, F. L. (2011). Flower constancy in insect pollinators: Adaptive foraging behaviour or cognitive limitation?. *Communicative & integrative biology*, **4**, 633-636. https://doi.org/10.4161/cib.16972

Hollingsworth, P. M., Li, D. Z., Van Der Bank, M., & Twyford, A. D. (2016). Telling plant species apart with DNA: From barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **371**, 20150338. https://doi.org/10.1098/rstb.2015.0338

IPBES. (2016). The assessment report on Pollinators, pollination and food production.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., … Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **97**, 1966-1979. https://doi.org/10.1111/ele.12162

Ji, Y., T. Huotari, T. Roslin, N. Martin-Schmidt, J. Wang, D. W. Yu, & O. Ovaskainen. (2019). SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and abundances using DNA barcodes or mitogenomes. *bioRxiv*. https://doi.org/10.1101/533737

Jones, M. B., Highlander, S. K., Anderson, E. L., Li, W., Dayrit, M., Klitgord, N., … Venter, J. C. (2015). Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences*, **112**, 14024-14029. https://doi.org/10.1073/pnas.1519288112

Keller, A., Danner, N., Grimmer, G., Ankenbrand, M., von der Ohe, K., von der Ohe, W., … Steffan-Dewenter, I. (2015). Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. *Plant Biology*, **17**, 558-566. https://doi.org/10.1111/plb.12251

Khansari, E., Zarre, S., Alizadeh, K., Attar, F., Aghabeigi, F., & Salmaki, Y. (2012). Pollen morphology of *Campanula* (*Campanulaceae*) and allied genera in Iran with special focus on its systematic implication. *Flora: Morphology, Distribution, Functional Ecology of Plants*, **207**, 203-211. https://doi.org/10.1016/j.flora.2012.01.006

Klein, A. M., Vaissière, B. E., Cane, J. H., Steffan-Dewenter, I., Cunningham, S. A., Kremen, C., & Tscharntke, T. (2007). Importance of pollinators in changing landscapes for world crops. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 303-133. https://doi.org/10.1098/rspb.2006.3721

Kraaijeveld, K., De Weger, L. A., Ventayol García, M., Buermans, H., Frank, J., Hiemstra, P. S., & Den Dunnen, J. T. (2015). Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. *Mol Ecol Resour*, **15**, 8-16. https://doi.org/10.1111/1755-0998.12288

Kress, W. J., García-Robledo, C., Uriarte, M., & Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution*, **30**, 25-35. https://doi.org/10.1016/j.tree.2014.10.008

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G. & Taylor, M. I. (2018). How quantitative is metabarcoding: a meta-analytical approach. *Molecular Ecology,* **28**, 420-430. https://doi.org/10.1111/mec.14920

Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D., & Caccamo, M. (2014). Next clip: An analysis and read preparation tool for Nextera long mate pair libraries. *Bioinformatics*, **30**, 566-568. https://doi.org/10.1093/bioinformatics/btt702

Leggett, R. M., & Clark, M. D. (2017). A world of opportunities with nanopore sequencing. *Journal of Experimental Botany*, **68**, 5419-5429. https://doi.org/10.1093/jxb/erx289

Li H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv*:1303.3997.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094-3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., & Chen, S. (2015). Plant DNA barcoding: from gene to genome. *Biological Reviews of the Cambridge Philosophical Society*, **90**, 157-166. https://doi.org/10.1111/brv.12104

Long, E. Y., & Krupke, C. H. (2016). Non-cultivated plants present a season-long route of pesticide exposure for honey bees. *Nature Communications*, **7**, 11629. https://doi.org/10.1038/ncomms11629

Lucas, A., Bodger, O., Brosi, B. J., Ford, C. R., Forman, D. W., Greig, C., ... de Vere, N. (2018). Generalisation and specialisation in hoverfly (*Syrphidae*) grassland pollen transport networks revealed by DNA metabarcoding. *Journal of Animal Ecology*, **87**, 1008-1021. https://doi.org/10.1111/1365-2656.12828

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, **17**, 10. https://doi.org/10.14806/ej.17.1.200

Morales, C. L., & Traveset, A. (2008). Interspecific pollen transfer: magnitude, prevalence and consequences for plant fitness. *Critical Reviews in Plant Sciences*, **27**, 221–238. https://doi.org/10.1080/07352680802205631

Morandin, L. A., & Kremen, C. (2013). Hedgerow restoration promotes pollinator populations and exports native bees to adjacent fields. *Ecological Applications*, **23**, 829–839. https://doi.org/10.1890/12-1051.1

Nayfach, S., & Pollard, K. S. (2016). Toward Accurate and Quantitative Comparative Metagenomics. *Cell*, **166**, 1103-1116. https://doi.org/10.1016/j.cell.2016.08.007

Ollerton, J., Winfree, R., & Tarrant, S. (2011). How many flowering plants are pollinated by animals? *Oikos*, **120**, 321-326. https://doi.org/10.1111/j.1600-0706.2010.18644.x

Pornon, A., Escaravage, N., Burrus, M., Holota, H., Khimoun, A., Mariette, J., … Vidal, M. (2016). Using metabarcoding to reveal and quantify plant-pollinator interactions. *Scientific Reports*, **6**, 27282. https://doi.org/10.1038/srep27282

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 and Evolution*, **25**, 345-353. https://doi.org/10.1016/j.tree.2010.01.007

Potts, S. G., Imperatriz-Fonseca, V., Ngo, H. T., Aizen, M. A., Biesmeijer, J. C., Breeze, T. D., … Vanbergen, A. J. (2016). Safeguarding pollinators and their values to human well-being. *Nature*, **540**, 220-229. https://doi.org/10.1038/nature20588

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, **35**, 833-844. https://doi.org/10.1038/nbt.3935

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Available at http://www.R-project.org/.

706 Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The Barcode of Life Data System
707 (http://www. barcodinglife.org). *Molecular ecology notes*, **7**, 355-364.
708 https://doi.org/10.1111/j.1471-8286.2007.01678.x
709
710 Richardson, R. T., Lin, C. H., Sponsler, D. B., Quijia, J. O., Goodell, K., & Johnson, R. M.
711 (2015). Application of ITS2 Metabarcoding to Determine the Provenance of Pollen Collected by
712 Honey Bees in an Agroecosystem. *Applications in Plant Sciences*, **3**, 1400066.
713 https://doi.org/10.3732/apps.1400066
714
715 Schlumbaum A., Tensen M., & Jaenicke-Després V. (2008). Ancient plant DNA in
716 archaeobotany. *Vegetation History and Archaeobotany*, **17**, 233-244.
717 https://doi.org/10.1007/s00334-007-0125-7
718
719 Schröter, K., Wemheuer, B., Pena, R., Schoning, I., Ehbrecht, M., Schall, P., … Polle, A. (2018).
720 Assembly processes of trophic guilds in the root mycobiome of temperate forests. *Molecular*
721 *Ecology*, **28**, 348-364. https://doi.org/10.1111/mec.14887
722
723 Sickel, W., Ankenbrand, M. J., Grimmer, G., Holzschuh, A., Härtel, S., Lanzen, J., … Keller, A.
724 (2015). Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-
725 indexing approach. *BMC Ecology*, **15**, 20. https://doi.org/10.1186/s12898-015-0051-y
726
727 Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in*
728 *Plant Science*, **5**, 209. https://doi.org/10.3389/fpls.2014.00209
729
730 Straub, S. C., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., & Liston, A. (2012).
731 Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics.
732 *American Journal of Botany*, **99**, 349-364. https://doi.org/10.3732/ajb.1100335
733
734 Tittonell, P. (2014). Ecological intensification of agriculture-sustainable by nature. *Current*
735 *Opinion in Environmental Sustainability*, **8**, 53-61. https://doi.org/10.1016/j.cosust.2014.08.006
736

Vanbergen, A. J., Baude, M., Biesmeijer, J. C., Britton, N. F., Brown, M. J. F., Bryden, J., … Wright, G. A. (2013). Threats to an ecosystem service: Pressures on pollinators. *Frontiers in Ecology and the Environment*, **11**, 251-259. https://doi.org/10.1890/120126

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. https://doi.org/10.1016/j.biocon.2015.04.022

Woodgate, J. L., Makinson, J. C., Lim, K. S., Reynolds, A. M., & Chittka, L. (2016). Life-long radar tracking of bumblebees. *PloS one*, **11**, e0160333. https://doi.org/10.1371/journal.pone.0160333

Zhou, X., Y. Li, S. Liu, Q. Yang, X. Su, L. Zhou, …, Huang, Q. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2**, 4. https://doi.org/10.1186/2047-217X-2-4

Figure legends and table titles

Figure 1. RevMet pipeline overview. a) Low coverage, short-read, reference datasets were generated for 49 wild plant species. b) Bee-collected pollen loads were sequenced on a MinION, generating long read datasets. c) The 49 short-read reference datasets were separately mapped to the long-read pollen datasets, and each pollen read was assigned to the plant species that mapped with the highest percent coverage or was left unassigned if the highest coverage was <15%. d) Binned pollen reads were counted, noise was reduced by implementing a 1% minimum-abundance filter, and then the remaining bin counts were converted to percentages.

763 Figure 2. Expected vs observed mock mix compositions. Six mock plant DNA mixes, each with
764 two technical replicates, were sequenced on a MinION and the RevMet method was applied. The
765 first stacked bar of each triplet represents the expected proportions based on input DNA. The
766 second and third bar of each triplet reflect the observed MinION read assignments resulting from
767 this pipeline.
768
769 Figure 3. Bee-collected pollen compositions and plant-pollinator interactions. a) The number of
770 individual pollen loads sequenced from three different species of bee. The proportion of pollen
771 loads that contained a single major plant species are represented by green bars, while those with
772 two major plant species are shown in blue. b) Mean number of plant species per pollen load for
773 each of three different species of bee; ** $p<0.01$, *** $p<0.001$. c) Bipartite plant-pollinator
774 network. The upper bars represent individual pollen loads from three different bee species, *Apis*
775 *mellifera* (red), *Bombus terrestris/lucorum complex* (blue), and *Bombus lapidarius* (purple). The
776 lower bars (grey) represent plant species. Link width indicates the MinION read proportion of
777 each major plant species within each pollen load.
778
779 Table 1. DNA mock mix compositions.
780
781 Figure S1. Numbers of mock-mix reads assigned to *Knautia arvensis*, and declines in the number
782 of unassigned reads, at different reference-skim coverage levels. The subsampling was repeated
783 three times.
784
785 Table S1. Estimated genome sizes, read counts, and coverage for the genome skim references.

Table S2. RevMet taxonomic assignments of mock-sample MinION reads.

Table S3. RevMet taxonomic assignments of bee-collected pollen MinION reads and pollen sample information.

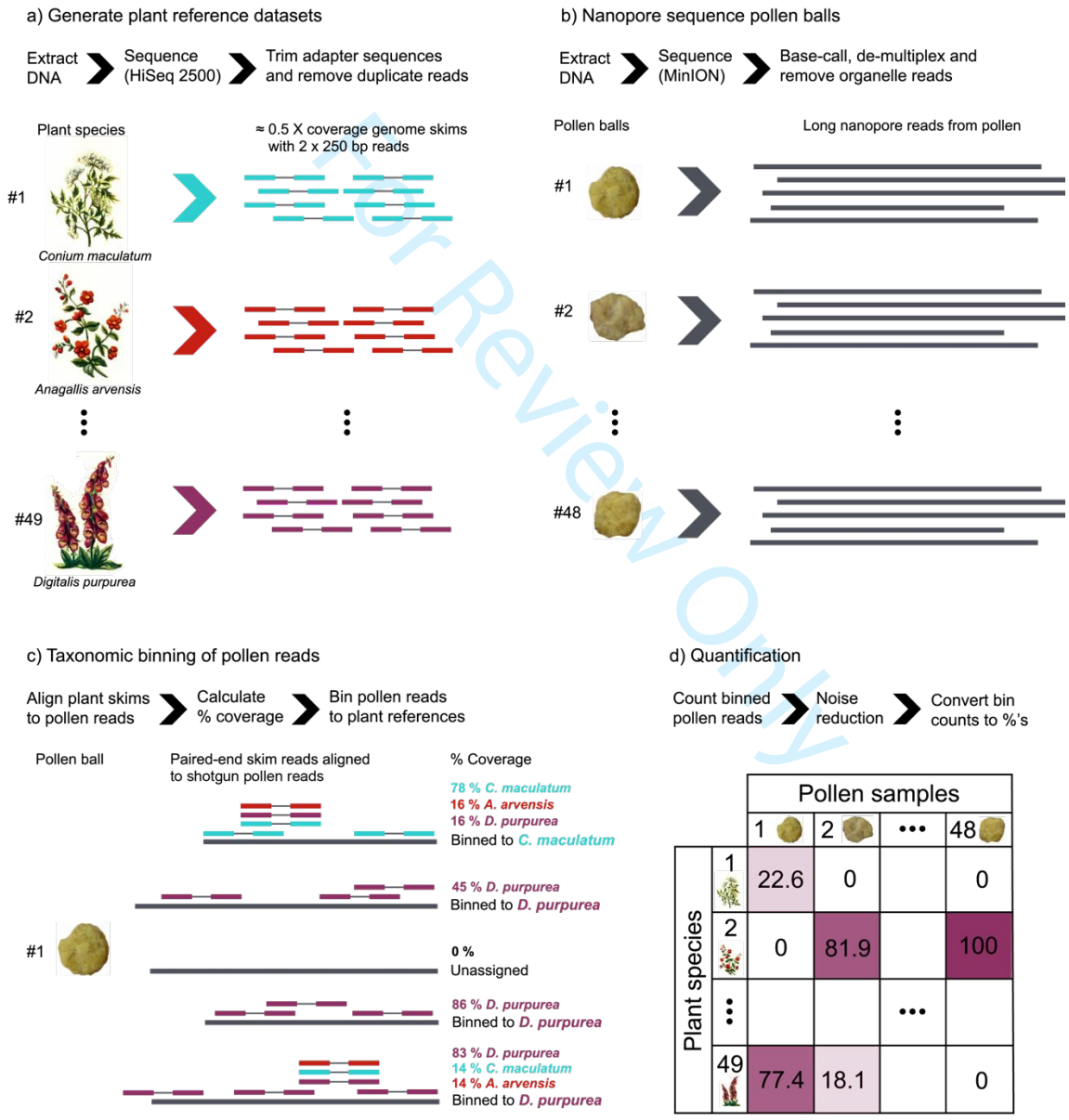


Figure 1. RevMet pipeline overview. a) Low coverage, short-read, reference datasets were generated for 49 wild plant species. b) Bee-collected pollen loads were sequenced on a MinION,

794 generating long read datasets. c) The 49 short-read reference datasets were separately mapped to
795 the long-read pollen datasets, and each pollen read was assigned to the plant species that mapped
796 with the highest percent coverage, or left unassigned if the highest coverage was <15%. d)
797 Binned pollen reads were counted, noise was reduced by implementing a 1% minimum-
798 abundance filter, and then the remaining bin counts were converted to percentages.
799

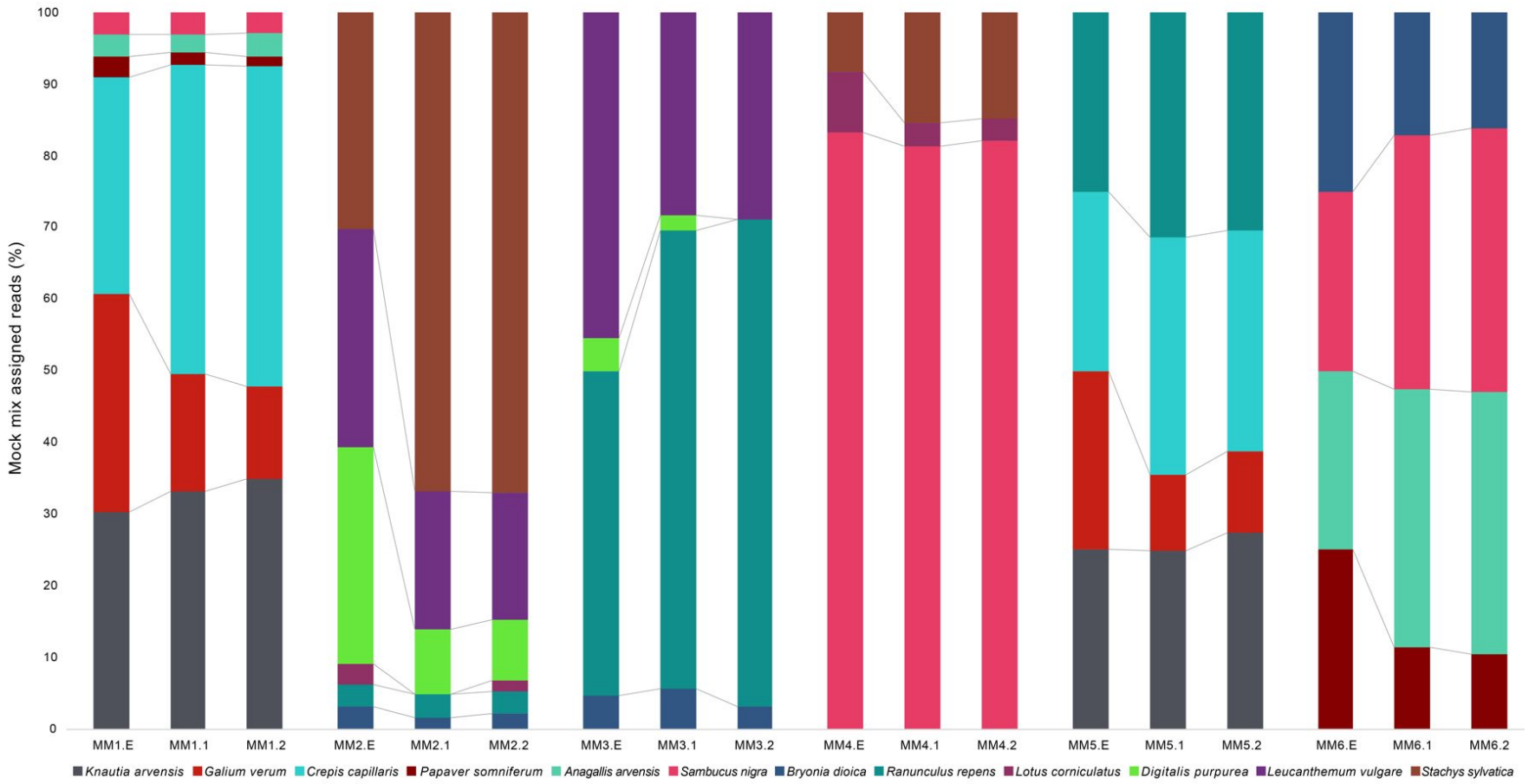


Figure 2. Expected vs observed mock mix compositions. Six mock plant DNA mixes, each with two technical replicates, were sequenced on a MinION and the RevMet method was applied. The first stacked bar of each triplet represents the expected proportions based on input DNA. The second and third bar of each triplet reflect the observed MinION read assignments resulting from this pipeline.

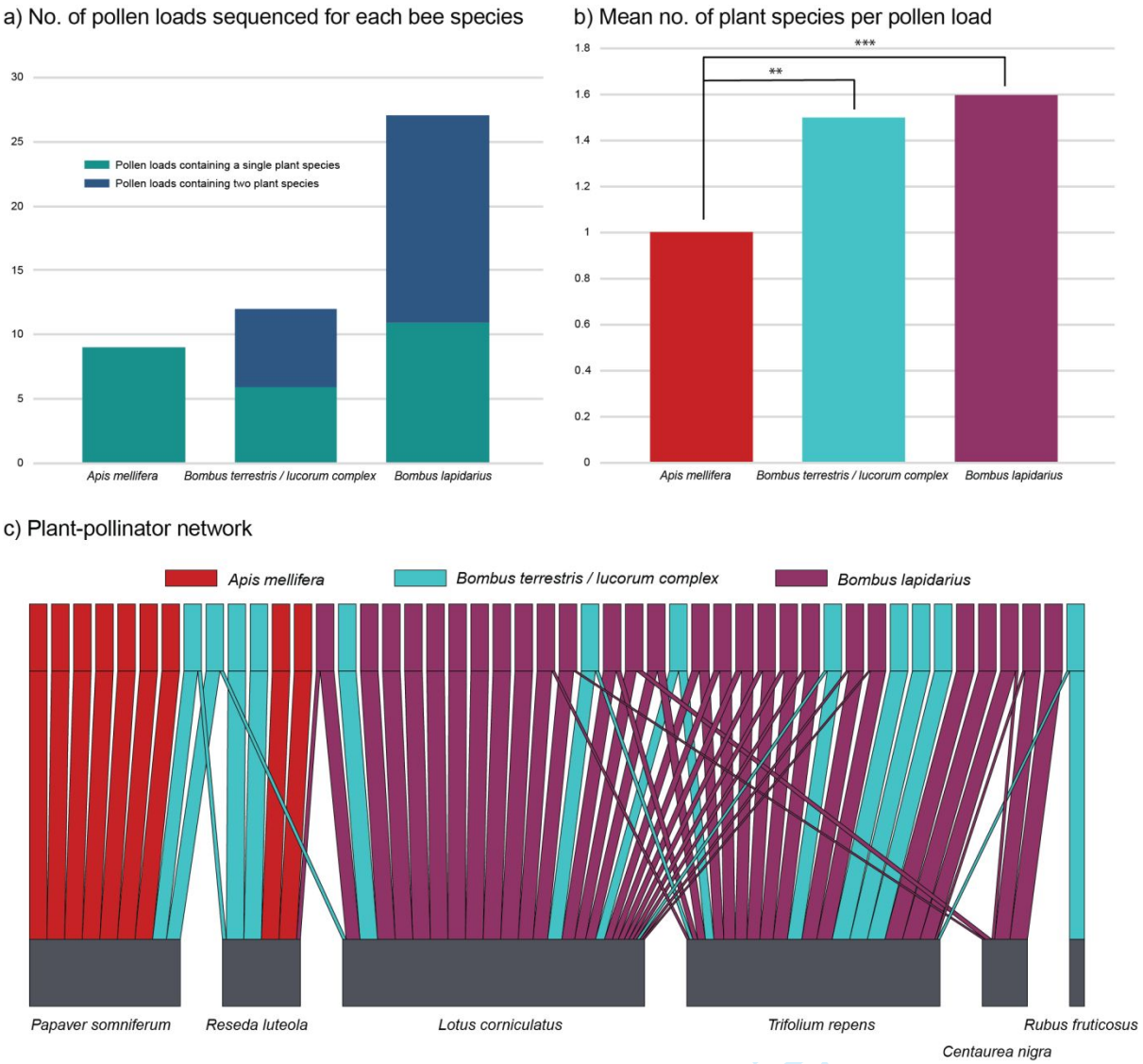


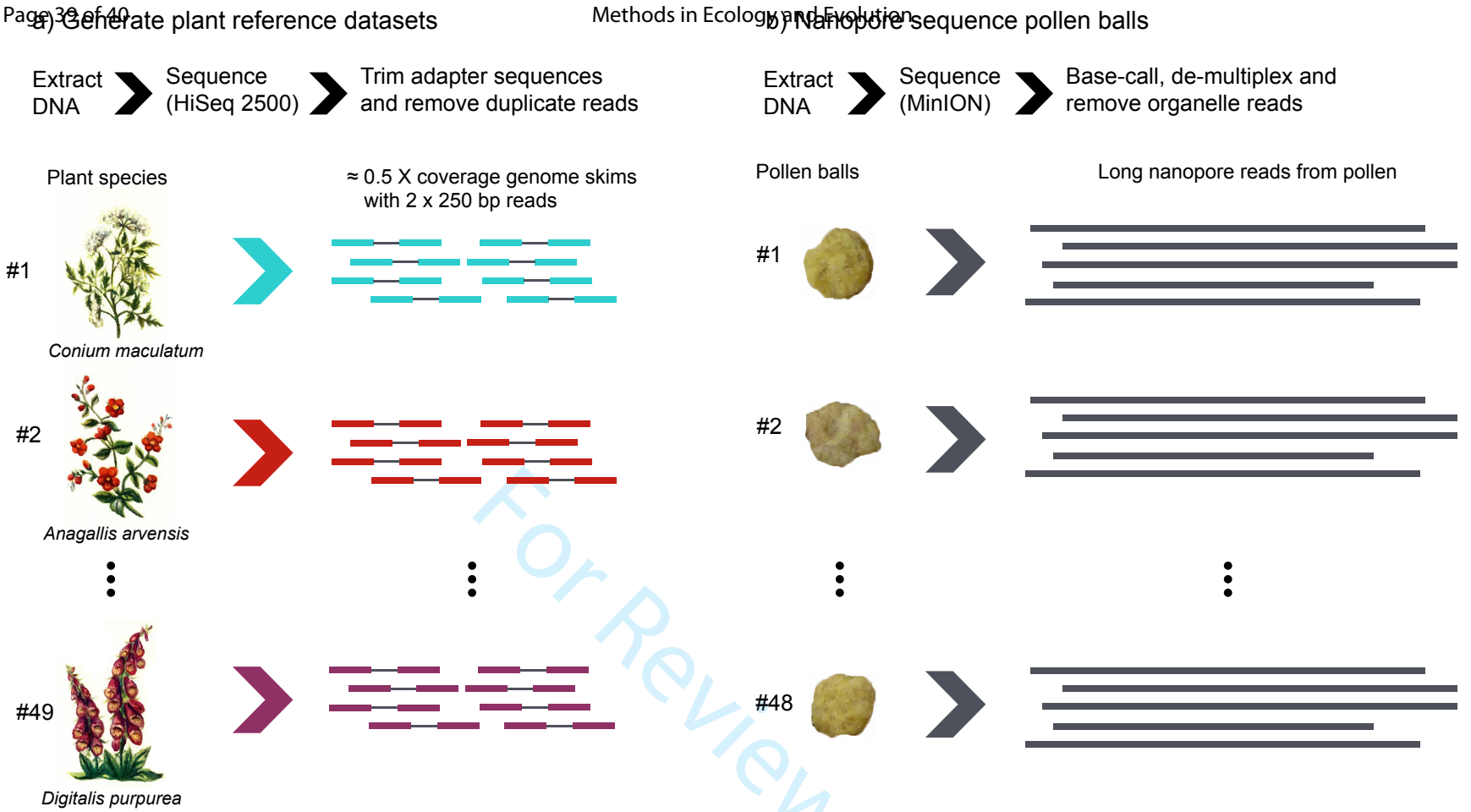
Figure 3. Bee-collected pollen compositions and plant-pollinator interactions. a) The number of individual pollen loads sequenced from three different species of bee. The proportion of pollen loads that contained a single major plant species are represented by green bars, while those with two major plant species are shown in blue. b) Mean number of plant species per pollen load for each of three different species of bee; ** $p<0.01$, *** $p<0.001$. c) Bipartite plant-pollinator network. The upper bars represent individual pollen loads from three different bee species, *Apis mellifera* (red), *Bombus terrestris/lucorum complex* (blue), and *Bombus lapidarius* (purple). The lower bars (grey) represent plant species. Link width indicates the MinION read proportion of each major plant species within each pollen load.

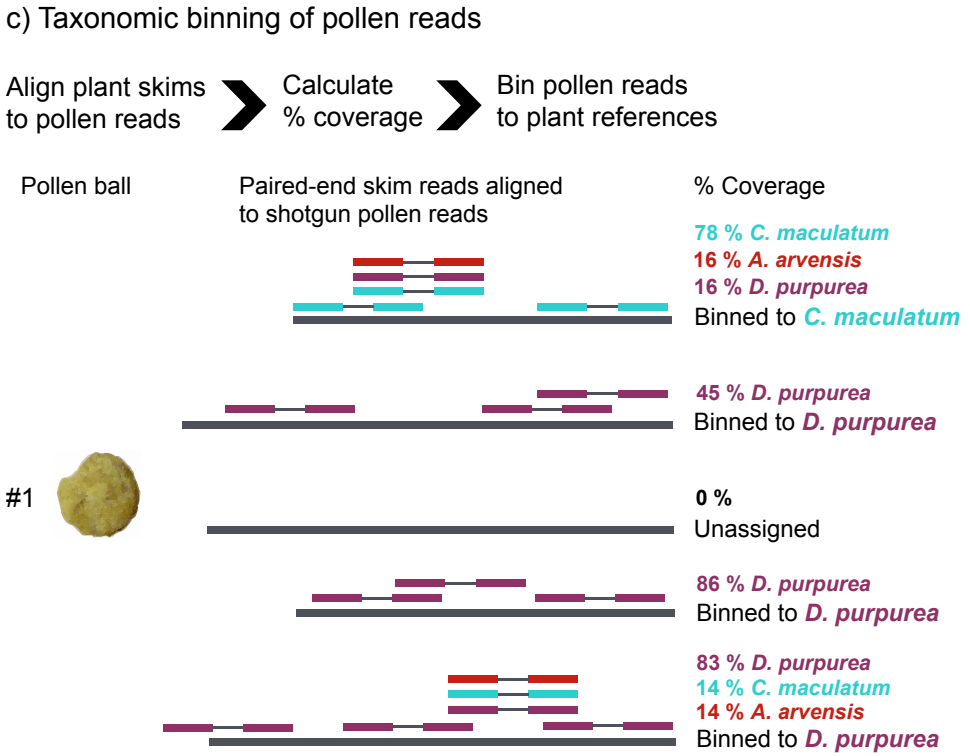
816 **Table 1**. DNA mock community compositions.

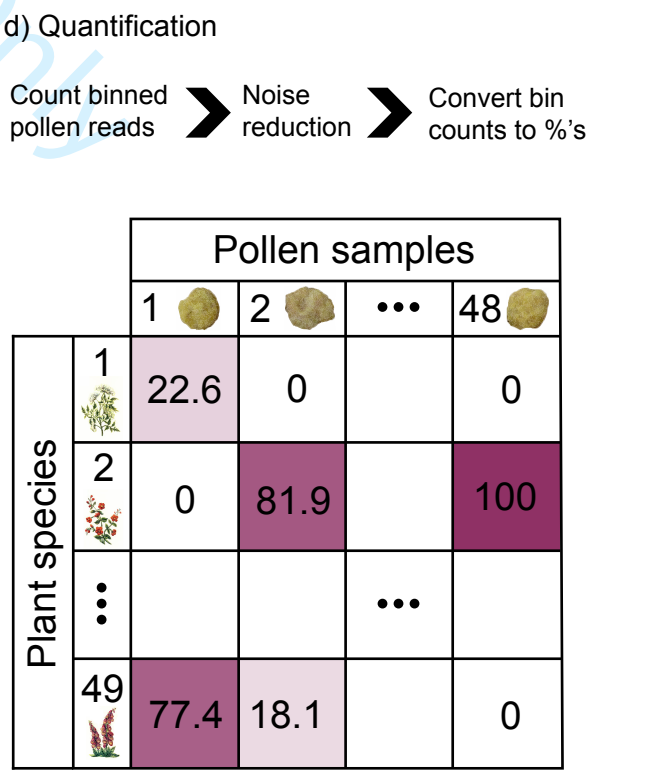
817

	Knautia arvensis	*Galium verum*	*Crepis capillaris*	*Papaver somniferum*	*Anagallis arvensis*	*Sambucus nigra*	*Bryonia dioica*	*Ranunculus repens*	*Lotus corniculatus*	*Digitalis purpurea*	*Leucanthemum vulgare*	*Stachys sylvatica*
MM1.Ratios	100	100	100	10	10	10	1	1	1	0	0	0
MM2.Ratios	0	0	0	1	1	1	10	10	10	100	100	100
MM3.Ratios	0	0	0	0	0	0	10	100	0	0	0	0
MM4.Ratios	0	0	0	1	0	1000	0	0	100	0	0	100
MM5.Ratios	100	100	100	0	1	1	1	100	0	0	0	1
MM6.Ratios	1	1	1	100	100	100	100	0	0	0	0	1
MM1.DNA (ng)	60.1	60.1	60.1	6.0	6.0	6.0	0.6	0.6	0.6	0.0	0.0	0.0
MM2.DNA (ng)	0.0	0.0	0.0	0.6	0.6	0.6	6.0	6.0	6.0	60.1	60.1	60.1
MM3.DNA (ng)	0.0	0.0	0.0	0.0	0.0	0.0	9.1	90.5	0.0	9.1	90.5	0.9
MM4.DNA (ng)	0.0	0.0	0.0	0.2	0.0	166.5	0.0	0.0	16.7	0.0	0.0	16.7
MM5.DNA (ng)	49.5	49.5	49.5	0.0	0.5	0.5	0.5	49.5	0.0	0.0	0.0	0.5
MM6.DNA (ng)	0.5	0.5	0.5	49.5	49.5	49.5	49.5	0.0	0.0	0.0	0.0	0.5
MM1.Percentages	30.0%	30.0%	30.0%	3.0%	3.0%	3.0%	0.3%	0.3%	0.3%	0.0%	0.0%	0.0%
MM2.Percentages	0.0%	0.0%	0.0%	0.3%	0.3%	0.3%	3.0%	3.0%	3.0%	30.0%	30.0%	30.0%
MM3.Percentages	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.5%	45.3%	0.0%	4.5%	45.3%	0.5%
MM4.Percentages	0.0%	0.0%	0.0%	0.1%	0.0%	83.3%	0.0%	0.0%	8.3%	0.0%	0.0%	8.3%
MM5.Percentages	24.8%	24.8%	24.8%	0.0%	0.3%	0.3%	0.3%	24.8%	0.0%	0.0%	0.0%	0.3%
MM6.Percentages	0.3%	0.3%	0.3%	24.8%	24.8%	24.8%	24.8%	0.0%	0.0%	0.0%	0.0%	0.3%

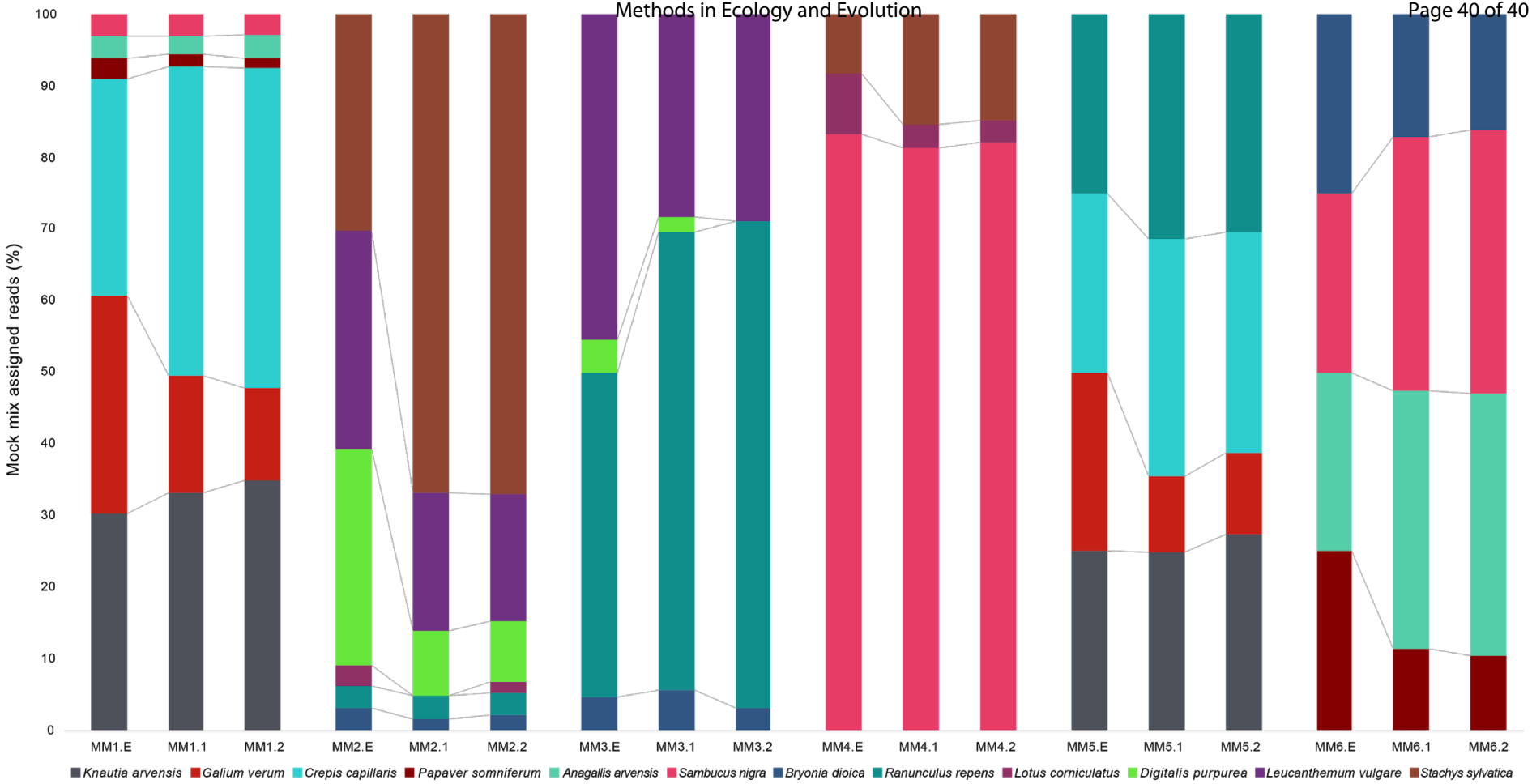
818



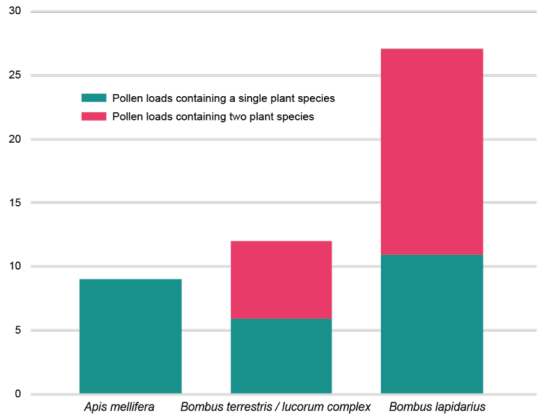




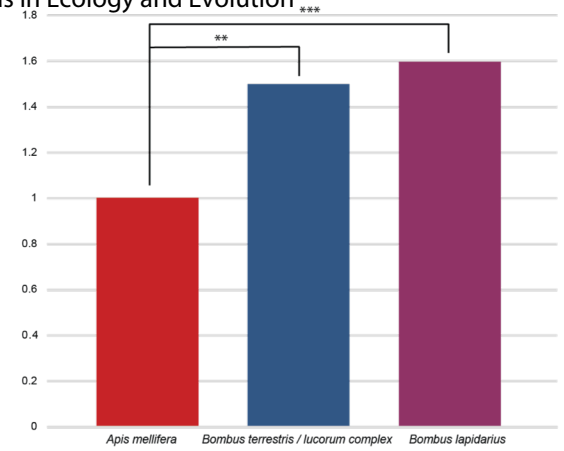
Plant species \ Pollen samples	1	2	•••	48
1	22.6	0		0
2	0	81.9		100
⋮			•••	
49	77.4	18.1		0


Mock mix assigned reads (%)
100
90
80
70
60
50
40
30
20
10
0
MM1.E
MM1.1
MM1.2
MM2.E
MM2.1
MM2.2
MM3.E
MM3.1
MM3.2
MM4.E
MM4.1
MM4.2
MM5.E
MM5.1
MM5.2
MM6.E
MM6.1
MM6.2
Knautia arvensis
Galium verum
Crepis capillaris
Papaver somniferum
Anagallis arvensis
Sambucus nigra
Bryonia dioica
Ranunculus repens
Lotus corniculatus
Digitalis purpurea
Leucanthemum vulgare
Stachys sylvatica

a) No. of pollen loads sequenced for each bee species



b) Mean no. of plant species per pollen load



c) Plant-pollinator network