MBMG
Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 6: 293–303
DOI 10.3897/mbmg.6.89753

Research Article

Assessing the lysis of diverse pollen from bulk environmental samples for DNA metabarcoding

Madison A. Moore[1], **Melissa K.R. Scheible**[2], **James B. Robertson**[3], **Kelly A. Meiklejohn**[2]

1 *North Carolina State University, College of Agriculture and Life Sciences, Dept. of Plant and Microbial Biology, 112 Derieux Place, Raleigh, NC 27607 USA*

2 *North Carolina State University, College of Veterinary Medicine, Dept. of Population Health and Pathobiology, 1060 William Moore Drive, Raleigh, NC, 27607 USA*

3 *North Carolina State University, College of Veterinary Medicine, Office of Research, 1060 William Moore Drive, Raleigh, NC, 27607 USA*

Corresponding author: Kelly A. Meiklejohn (kameikle@ncsu.edu)

Academic editor: Birgit Gemeinholzer | **Received** 30 June 2022 | **Accepted** 8 September 2022 | **Published** 23 September 2022

Abstract

Pollen is ubiquitous year-round in bulk environmental samples and can provide useful information on previous and current plant communities. Characterization of pollen has traditionally been completed based on morphology, requiring significant time and expertise. DNA metabarcoding is a promising approach for characterizing pollen from bulk environmental samples, but accuracy hinges on successful lysis of pollen grains to free template DNA. In this study, we assessed the lysis of morphologically and taxonomically diverse pollen from one of the most common bulk environmental sample types for DNA metabarcoding, surface soil. To achieve this, a four species artificial pollen mixture was spiked into surface soils collected from Colorado, North Carolina, and Pennsylvania, and subsequently subjected to DNA extraction using both the PowerSoil and PowerSoil Pro Kits (Qiagen) with a heated incubation (either 65 °C or 90 °C). Amplification and Illumina sequencing of the internal transcribed spacer subunit 2 (*ITS2*) was completed in duplicate for each sample (total n, 76), and the resulting sequencing reads taxonomically identified using GenBank. The PowerSoil Pro Kit statistically outperformed the PowerSoil Kit for total DNA yield. When using either kit, incubation temperature (65 °C or 90 °C) used had no impact on the recovery of DNA, plant amplicon sequence variants (ASVs), or total plant *ITS2* reads. This study highlighted that lysis of pollen in bulk environmental samples is feasible using commercially available kits, and downstream DNA metabarcoding can be used to accurately characterize pollen DNA from such sample types.

Key Words

DNA extraction, DNA metabarcoding, internal transcribed spacer 2, pollen, surface soils

Introduction

Seed plants, which account for >90% of land plants, produce pollen grains which vary in shape size, aperture, and morphology (Shivanna and Tandon 2014). For seed plants to reproduce, pollen, which contains male gametes (*i.e.*, plant sperm), must reach the female reproductive structure (known as pollination). The mechanisms by which plants achieve pollination differ and largely dictate the amount of pollen produced; species that rely on wind and water produce much more pollen than insect-pollinated species (Shivanna and Tandon 2014). While most pollen is produced during spring and fall, it is ubiquitous in the environment year-round and can easily travel 300–400 m from the source plant (Shivanna and Tandon 2014). Pollen can remain well preserved and biologically viable in the environment for decades due to its impenetrable cell wall (Bryant et al. 1990; Mildenhall et al. 2006). Given these features, it is possible to morphologically identify pollen present in compromised yet diverse samples (*e.g.*, dust, aged soils), to obtain valuable information on the source plant community and relative species abundance.

The current "gold standard" for identifying pollen is via microscopic examination of grain morphology where distinguishing features can permit genus-level identification (Mildenhall et al. 2006; Shivanna and Tandon 2014). Palynologists identify a representative number of grains from a sample to determine the source seed plant community with this information being useful in disciplines such as paleoecology (*e.g.*, understanding previous plant ecosystems [Delcourt et al. 1983; Sugita 2007]), forensics (*e.g.*, associating the suspect with crime scene [Mildenhall 2006; Mildenhall et al. 2006; Wiltshire et al. 2015]), or food safety and regulatory compliance (*e.g.*, identifying honey origin and legitimacy [Kenjerić et al. 2008]). Despite the utility of pollen for characterizing seed plant communities, several limitations have hindered its broad application: the amount needed for comprehensive analysis varies (*e.g.*, 0.5 g for peat soil vs. >60 g for sandy soil sediments) and morphological identification is very time consuming and can only be performed by highly trained experts (Bryant et al. 1990; Mildenhall et al. 2006). Progress has been made to use instrumentation to complete morphological identification of pollen, either via an automated trainable pollen location and classifier system (Holt et al. 2011) or flow cytometry and deep learning (Dunker et al. 2020); however both methods still require some level of pollen preparation, which is tedious and requires specialized training.

With advances in sequencing technologies, researchers have assessed the reliability and accuracy of DNA-based approaches for characterizing pollen (Kraaijeveld et al. 2015; Bell et al. 2016, 2019). Nuclear, mitochondrial and plastid DNA found within the cytoplasm of pollen grains is largely protected by sporopellenin depositions in the cell wall (Sassen 1964; Jackson 1987), making it amenable to isolation for molecular analyses. Whilst damage to pollen DNA due to exposure to extreme environmental conditions (*e.g.*, UV radiation, heavy metals, dehydration, alkaloids and naturally occurring carcinogens) is possible (Jackson 1987; Taylor and Jonsson 2004), previous studies have reported that sufficient DNA remains for characterization using highly sensitive next generation sequencing (NGS) approaches (*e.g.*, Sønstebø et al. 2010; Jørgensen et al. 2012a, b; Kraaijeveld et al. 2015; Richardson et al. 2015a, b; Bell et al. 2017). DNA metabarcoding is the most commonly used NGS approach for pollen characterization from bulk environmental sample types (*e.g.*, soil, dust, water, feces *etc.*) and involves amplification of short yet informative regions of the genome from different pollen grains concurrently. After sequencing, identification is achieved by comparing the unknown sequences to a reference database of sequences generated from known taxa.

DNA metabarcoding offers several advantages over traditional morphological identification of pollen from bulk environmental samples. 1) Increased taxonomic resolution, as most land plants can be identified to genus level at minimum, but often down to the species level (70–90% of cases [Lang et al. 2019; Peel et al. 2019]). 2) Statistically relevant characterizations are possible from small amounts of bulk material. For example, DNA isolation kits typically require 100–250 mg of soil for input, whereas grams of soil can be needed for morphological characterization of pollen from specific soil types (Chen et al. 2010; Young et al. 2014; Cheng et al. 2015). 3) Pollen present in low quantities that may be missed when identifying a set number of grains using morphology, are more easily detected using bulk DNA analysis (Fahner et al. 2016). 4) Higher throughput and shorter processing time, as sample barcoding permits multiple samples to be pooled and analyzed simultaneously.

Numerous studies have successfully implemented DNA metabarcoding to characterize pollen from diverse sample types, including ancient sediments, soil, insects and air filters (*e.g.*, Sønstebø et al. 2010; Jørgensen et al. 2012a, b; Kraaijeveld et al. 2015; Richardson et al. 2015a, b; Bell et al. 2017). The results of these previous studies have, however, been based on a key assumption; the extraction method used successfully lysed all pollen grains present in a sample releasing DNA for downstream analysis. Notably, if all pollen grains are not lysed or successfully released DNA is subsequently sheared by excess chemical or mechanical digestion, the DNA metabarcoding results would not accurately represent the pollen community present in the sample. Seminal work completed by Simel and colleagues (1997) on single source pollen grains identified that microbead maceration was the most effective method (out of the eight tested) at releasing high-quality DNA suitable for downstream molecular analysis. Subsequently, many currently commercially available plant and bulk environmental sample DNA extraction kits include a microbead maceration step to ensure lysis of DNA from pollen and other challenging tissue types (*e.g.*, seeds, bark, waxy cuticles, spores *etc.*). The extraction kits used for lysis of pollen from bulk environmental samples in published studies such as dust, surface soil, honey and corbiculae pollen, have ranged from commercially available kits (*e.g.*, DNeasy Plant Mini Kit [Qiagen, Hilden, Germany], NucleoMag Kit [Macherey-Nagel, Düren, Germany], PowerMax Soil Extraction Kit [Qiagen]) to traditional CTAB methods (*e.g.*, Zhou et al. 2007; Niemeyer et al. 2017; Leontidou et al. 2018; Manivanan et al. 2018; Peel et al. 2019). A recent study examined the optimal incubation time for pollen grain lysis for downstream DNA metabarcoding, however only pure pollen grain mixtures (not reflective of bulk environmental samples) were used for testing (Swenson and Gemeinholzer 2021). Bulk environmental samples contain various other components and chemicals which may interact (positively or negatively) with isolation of DNA from pollen grains. Given this, there is a need to assess the effectiveness of pollen lysis in bulk environmental samples most commonly subjected to DNA metabarcoding.

To address this pivotal gap, this study focused on assessing the lysis of morphologically (*i.e.*, size, shape, aperture) and taxonomically diverse pollen for one of the most common bulk environmental sample types for DNA metabarcoding, surface soil. To achieve this, surface soil collected from three states in the continental U.S. representing various geological and climate features was spiked with a known four-taxa artificial pollen mixture and subjected to DNA isolation using two commercially available soil extraction kits. The impact of heated incubation (65 °C or 90 °C) on the lysis of pollen was assessed for each sample using both kits in duplicate (total n, 76). The internal transcribed spacer subunit 2 (*ITS2*) was amplified in duplicate for each sample, with duplicates pooled prior to library preparation and sequencing using Illumina chemistry. Downstream data analysis focused on assessing variation in the recovery of the baseline plant community along with known spiked pollen taxa, to identify the optimal method for pollen lysis.

Materials and methods

Pollen samples

Mixed corbiculae pollen granules collected from North America were used in this study. In a sterile biosafety cabinet, corbiculae pollen granules were initially sorted by eye into groups according to color. A second round of sorting was performed under a stereomicroscope (Fisher Scientific, Hampton, NH) to confirm that each group contained pollen of the same color and possessed similar morphological features. Each group of colored corbiculae pollen was subsequently treated as a different species. The following steps were completed to carefully remove the nectar, sugars, wax and other compounds associated with corbiculae pollen without lysing individual grains: 1) 1 mL of sterile water was added to ~1 cm^3 corbiculae pollen in a 2 mL microcentrifuge tube, 2) the tube was incubated at 600 rpm for 30 minutes at room temperature, and 3) excess liquid was removed using a pipette and washed corbiculae pollen was allowed to air dry at room temperature in a fume hood for 21 days prior to storage at -20 °C until use. To confirm the taxonomic identity of the corbiculae pollen (n, 4 colored groups) the following was completed: 1) a subsample was ground using a disposable mortar and pestle and the DNA subsequently isolated following the manufacturers protocol for the Qiagen DNeasy Plant Mini Kit (Qiagen), 2) a 590 bp region of *rbcL* was amplified and bidirectionally Sanger sequenced as outlined in Meiklejohn et al. (2018), and 3) after removal of the primer sequences, the consensus was searched against GenBank. Genus level identifications were only possible for two out of four taxa; *Symphyotrichum* spp. (Asteraceae) and *Trifolium* spp. (Fabaceae). Granular pollen from *Populus tremuloides* (Salicaceae; Sigma Aldrich, St Louis, MO) and *Zea mays* (Poaceae; Carolina Biological Supply, Burlington, NC) was purchased for use in the study. As the *Z. mays* pollen was suspended in liquid by the manufacturer, it was washed three times with sterile water and allowed to air dry for two days in a fume hood prior to use.

Soil samples

Approximately 100 g of surface soil (top 1–10 cm) were collected during October 2019 from three locations with differing geology and climate: 1) Erie, Colorado, 2) Cary, North Carolina, and 3) Laboratory, Pennsylvania. Each sample was initially collected into a plastic zip-lock bag. Once the samples reached the laboratory, they were immediately transferred into separate sterile food-grade foil tins and allowed to air dry inside a fume hood. Once dry, each soil sample was sieved three times through a sterile food-grade metal kitchen sieve, in order to remove any large debris (*i.e.*, plant and insect fragments) and homogenize the soil prior to downstream analysis.

Pollen spiked soil samples

To determine which isolation method could robustly lyse pollen, the artificial pollen mixture contained pollen with diverse morphological features (*i.e.*, size, shape, aperture) from four different orders. A four-taxa artificial pollen mixture consisting of both dry granular (*P. tremuloides* [Salicaceae] and *Z. mays* [Poaceae]) and dry corbiculae pollen (*Symphyotrichum* spp. [Asteraceae] and *Trifolium* spp. [Fabaceae]) was created. To assess sensitivity (or limit of detection), the relative abundance of pollen from each taxa varied in the artificial mixture as follows: approximately 0.4% – Poaceae, 9.6% – Salicaceae, 40.5% – Asteraceae, and 49.5% – Fabaceae. The weight of pollen added to the mixture from each taxa, determined using an analytical scale (AG104, Mettler Toledo, Columbus, OH), was as follows: *P. tremuloides* – 0.395 g, *Z. mays* – 0.0179 g, *Symphyotrichum* spp. – 1.67 g, and *Trifolium* spp. – 2.045 g. The artificial pollen mixture was spiked into separate 5 g subsamples of each surface soil 1) at a concentration (*i.e.*, number of grains/g of soil) which mimicked the reported naturally occurring concentration of pollen in soils collected from North Carolina (Russell 1993), Pennsylvania (Kelso 1994) and Colorado (Maher 1972) (herein referred to as 'spiked normal'), and 2) at one fifth the naturally occurring concentration (herein referred to as 'spiked partial'). Spiked soils were carefully mixed using a sterile 1000 µL pipette tip to ensure even homogenization of pollen throughout the soil without causing premature pollen lysis. The weight of a single *Z. mays* pollen grain (Stanley and Linskens 1974) was used to calculate the weight of the artificial pollen mixture to be spiked into each soil subsample (Table 1). Table 2 outlines the estimated number of pollen grains for each taxa spiked into the soil samples, along with the estimated number of pollen grains present in a 100 mg subsample used for DNA isolation.

Table 1. Reported naturally occurring concentrations of pollen in soils from North Carolina (NC), Pennsylvania (PA) and Colorado (CO) used to calculate the weight of artificial pollen mixture to be spiked into subsamples. * denotes that the weight (g) of a single *Zea mays* pollen grain (2.47E-07; Stanley and Linskens 1974) was used to calculate the weight of pollen naturally occurring in each soil type.

	Pollen concentration (grains/cm^3)	Pollen weight/gram*	Weight (g) artificial pollen mixture spiked into 5 g soil subsample	
			Spiked normal	*Spiked partial*
Erie, CO	200,000 (Maher 1972)	0.049	0.25	0.049
Cary, NC	117,500 (Russell 1993)	0.029	0.15	0.029
Laboratory, PA	6,000 (Kelso 1994)	0.0015	0.0074	0.0015

Table 2. Estimated number of pollen grains for each of the four taxa (rounded to the nearest single grain) spiked into 5 g subsamples of soil from Colorado (CO), North Carolina (NC) and Pennsylvania (PA). Number given in parentheses indicates estimated number of grains in a 100 mg subsample used for DNA isolations, providing complete homogenization after pollen spike. * denotes commercially purchased dry granular pollen, ^ denotes washed corbiculae pollen.

	Spiked normal			*Spiked partial*		
	CO	*NC*	*PA*	*CO*	*NC*	*PA*
Zea mays *	54 (1)	32 (0)	2 (0)	11 (0)	6 (0)	0 (0)
Populus tremuloides *	1,184 (24)	704 (14)	38 (1)	237 (5)	141 (3)	8 (0)
Symphyotrichum spp. ^	5,009 (100)	2,978 (60)	159 (3)	1,002 (20)	596 (12)	32 (1)
Trifolium spp. ^	6,131 (123)	3,646 (73)	195 (4)	1,226 (25)	729 (15)	39 (1)

DNA isolation

Pollen present in both unspiked (baseline sample) and spiked (both normal and partial) North Carolina, Pennsylvania and Colorado soils were lysed and the DNA subsequently isolated using two different commercial soil DNA isolation kits: DNeasy® PowerSoil® Kit (Qiagen) and the DNeasy® PowerSoil® Pro Kit (Qiagen). These kits were chosen for use in this study, as they 1) are reported by the manufacturer to yield highly pure DNA, 2) use a patented Inhibitor Removal Technology to remove compounds which negatively impact downstream DNA analysis (*i.e.*, humic acid associated with soil), and 3) have been used in some other studies for isolating pollen from diverse sample types (Niemeyer et al. 2017). A 100 mg subsample of soil was used as input for each DNA isolation, and was transferred to isolation tubes using a sterile disposable spatula (VWR International, Radnor, PA). The impact of a 30 minute heated incubation at either 65 °C or 90 °C immediately after the addition of C1 solution (PowerSoil) or CD1 solution (PowerSoil Pro) on pollen lysis, was assessed (subsequently referred to as 'method(s)'). Two additional modifications to the manufacturer protocols, suggested for challenging sample types, were implemented across all isolations: 1) 20 µL of Proteinase K (20 mg/mL; Qiagen) was added to solution C1 or CD1, and 2) samples were subjected to bead-beating for 15 minutes at maximum speed using a Vortex Adapter (Qiagen), immediately after the heated incubation. Duplicate isolations were completed for each soil sample and method (total n, 72). Isolated DNA from either kit was eluted in 100 µL of solution C6 and stored at -20 °C when not in use. The total genomic DNA yielded was quantified using the Qubit™ 3 Fluorometer (Invitrogen, Carlsbad, CA) and the Qubit™ HS DNA Assay Kit (Invitrogen). All isolations of a given method were completed in a single batch (n, 18) and processed along-side a reagent blank. Reagent blanks were carried through the remainder of the workflow (n, 4).

DNA amplification, library preparation and sequencing

A ~350 bp fragment of the nuclear *ITS2* was chosen for DNA metabarcoding in this study. The rationale behind only using a single nuclear marker (as opposed to a combination of nuclear and plastid markers commonly used in plant DNA metabarcoding) was threefold: 1) internal testing demonstrated that the primers chosen for use can successfully amplify DNA from the four taxa in the artificial pollen mixture if present (results not shown), 2) *ITS2* sequences for the four taxa included in the artificial pollen mixture are in GenBank for comparisons, and 3) there are more overall *ITS2* sequences available for comparison on GenBank than *trnL* (~84,000 vs 59,000, respectively [as of 20/7/2022]). Amplification of *ITS2* was completed using ITS2F (5'-ATGCGATACTTGGTGTGAAT -3'; Chen et al. 2010) and ITSp4 (5'- CCGCTTAKTGATATGCTTAAA-3'; Cheng et al. 2015) following the optimized reaction and cycling conditions outlined in Timpano et al. (2020). Briefly, duplicate amplifications for each sample were completed and consisted of 2 µL of isolated DNA in a 25 µL total reaction volume (Timpano et al. 2020). All amplifications were performed on a Veriti 96-Well Thermal Cycler (Applied Biosystems, Foster City, CA). Duplicate amplicons were pooled (total 50 µL) and post-amplification cleanup, quantification and subsequent library preparation was completed as outlined in Timpano et al. (2020). Each library (total n, 76) was individually quantified using the KAPA Library Quantification kit (KAPA Biosystems, a Roche Company; Wilmington, MA) on the QuantStudio 5 System (ThermoFisher Scientific, Waltham, MA), and an appropriate volume combined in a single DNA LoBind tube (Eppendorf, Hamburg, Germany) to create an approximately equi-molar library pool. The concentration of the final library pool was verified using the KAPA Library Quantification kit (KAPA Biosystems). The pooled library was prepared as described in the Denature and Dilute Libraries Guide (Document # 15039740 v10) with a 50% PhiX spike (based on recommendations from the manufacturer for low diversity library pools on the MiniSeq) and final loading concentration of 1.4 pM. This pool was subsequently sequenced on a single run of the Illumina MiniSeq using the MiniSeq System Mid-Output Kit (Illumina, San Diego, CA; 1 X 300 bp).

Data analysis

Raw sequence data were processed and analyzed on the NC State University High Performance Cluster as follows 1) Cutadapt (v2.10) (Martin 2011) was used to trim primer sequences from raw reads (error rate of 0.11), 2) quality filtering, trimming of reads and identification of amplicon sequence variants (ASVs) was completed using default parameters of the DADA2 (v1.18.0) pipeline (Callahan et al. 2016), 3) resulting ASVs were searched against GenBank's nucleotide database using the remote command line *blastn* (v2.10.1) with the top 10 matches recorded, and 4) ASVs with matches meeting strict *blastn* criteria (>90% sequence coverage, >95% sequence identity, e-value of <0.001) were identified, and subjected to the program taxize (v0.2.2) (Chamberlain and Szöcs 2013) to obtain detailed taxonomic information. ASVs identified across reagent blanks were removed from all samples prior to data analysis.

To strike a balance with respect to informational content and ease of interpretation, resulting statistical analyses focused on families in which both the total number of reads and ASVs were >1%. A total of nine families met both of these criteria: Asteraceae (daisies, sunflowers), Brassicaceae (mustards, cabbages), Caryophyllaceae (carnations), Fabaceae (legumes, peas, beans), Juglandaceae (walnuts), Poaceae (grasses), Rosaceae (roses), Salicaceae (willows, poplar) and Ulmaceae (elms) (Suppl. material 1).

Principal components analyses were conducted to examine discriminatory ability of ASV read abundances between sample kit, method and location, with data then being plotted against the first two principal components. One observation was removed from 5-family considerations due to excessive influence on the model. For examining variability between duplicates, log-scale differences for each pair of duplicates were calculated for each of the nine key families. When examining whether a) pollen spikes were successful and b) the resulting sequence reads were recovered in the expected ratios, the difference between the averaged spiked duplicates (both partial and normal) and the averaged unspiked duplicates were obtained for the four target taxa at the genus level (*i.e.*, ASVs assigned to *Populus*, *Zea*, *Symphyotrichum* or *Trifolium*). If spiked taxa were increased then the one-sample Hotelling's T^2 with 2 numerator and 10 denominator degrees of freedom was used to compare the isometric log-transformed observed ratios to the expected ratios. *Zea* was not considered for this analysis as all spiked samples saw reduced *Zea* compared to the unspiked. Statistical *t*-tests and Pearson correlation analyses were completed using JMP Pro, Version 16.0.0 (SAS Institute Inc., Cary, NC). The *MVTests* (Bulut 2019), 'compositions' (van den Boogaart et al. 2022), *ggplot2* (Wickham 2016), *ggfortify* (Tang et al. 2016; Horikoshi and Tang 2018), and *FactoMineR* (Lê et al. 2008) packages as well as base R (R Core Team 2022) (Version 4.1.3) were also used to complete Hotelling's T^2, principal components analysis, and summaries of inter-duplicate differences. Tableau Desktop v2022.1 (Tableau Software Inc., Seattle, WA) was also used to visualize data.

Data availability

The final dataset of ASVs used in analyses are available in FigShare (10.6084/m9.figshare.20377146).

Results and discussion

DNA isolation

When using the PowerSoil Pro Kit for DNA isolations, significantly (p <0.0001 [t(41)=-7.62]) higher DNA quantities were obtained over the original PowerSoil Kit regardless of the incubation temperature; 28.5 ± 15.8 ng/µL *vs.* 7.08 ± 5.25 ng/µL, respectively. The incubation temperature used did not significantly impact the DNA yield with either kit; p = 0.242 (t(31)=-1.19; PowerSoil) and p = 0.634 (t(27)=-0.48; PowerSoil Pro) (Suppl. material 2). Whilst the main steps in the manufacturer's protocol for the PowerSoil and PowerSoil Pro kits are similar, it is safe to assume that some of the buffers/solutions used between the kits differ given they have different names (*e.g.*, C1 *vs.* CD1, and C2 *vs.* CD2). As Qiagen does not disclose buffer/solution composition, it is not possible to identify any specific chemical differences between kits that may have contributed to the varied DNA yield. Aside from a cost difference (PowerSoil ~$6/sample, PowerSoil Pro ~$7.50 sample) the beads differ between kits; the PowerSoil Pro Kit contains both 0.1 mm and 0.5 mm ceramic beads, whereas the PowerSoil Kit contains only 0.7 mm garnet beads. A recent published study examined the impact of bead size, type and lysis time on the rupture of pollen grains for downstream DNA metabarcoding (Swenson and Gemeinholzer 2021). The authors of that study reported that 1.4 mm and 2.8 mm beads were the most effective at causing lysis of pollen grains, and subsequently increased DNA quantities were recovered when >95% of pollen grains were lysed. Additionally, Swenson and Gemeinholzer (2021) cautioned that excessive lysis of pollen grains (either prolonged bead beating [>15 minutes] or chemical lysis [>2 hours]) can negatively impact downstream DNA metabarcoding results; DNA yields may be higher but sequencing results can be diminished due to shearing of exposed DNA. In our study, both the bead beating and chemical lysis steps used were on the lower range of these times.

Library preparation and sequencing metrics

In this study, we observed an overall weak positive correlation between DNA yield and resulting library yield (r = 0.46). Notably, the DNA quantification method used in this study (Qubit™ HS DNA Assay Kit) quantifies all double stranded DNA present in the sample, regardless of source or length. While the quantity of only plant DNA isolated from each sample would have provided a more accurate and useful comparison, no commercially available plant-specific DNA quantification kits currently exist. After processing data through the DNA metabarcoding pipeline, a total of 5,746 ASVs encompassing 2,286,926 reads were recovered across all samples. When ASVs which a) did not match to sequences derived from a plant specimen (kingdom, Viridiplantae; n, 4,172), and b) were present in the reagent blanks (n, 28) were excluded, a total of 1,574 ASVs encompassing 878,771 reads across all samples remained for downstream statistical analyses. Notably, the vast majority of excluded ASVs were those for which taxonomic classification even at the highest level (superkingdom) was not obtained (given as 'unknown' in taxize output; n, 3,667). The average (± standard deviation) of the total ASVs was 75.9 ± 31.7 (range 2–149), which related to 12,205 ± 5,100 (range 33–33,467) reads per sample (Suppl. material 2). The taxonomic classification of the final dataset of 1,574 ASVs spanned 54 plant families belonging to 30 plant orders.

Comparison of extraction kits and methods

To assess the overall effect of extraction kit (PowerSoil or PowerSoil Pro kit) and incubation temperature (65 °C or 90 °C) on the plant community recovered, soils not spiked (*i.e.*, baseline samples) with the four-taxa artificial pollen mixture (n, 24 [12 samples in duplicate]) were initially evaluated. At a broad level, no statistical difference was noted in the number of total reads (p = 0.1904 [t(21)=1.35]) or ASVs (p = 0.1102 [t(21)=1.66]) recovered between the two kits (PowerSoil or PowerSoil Pro). Incubation temperature did not have a statistical impact on the recovery of total reads (p = 0.666 [t(10)=-0.445] and p = 0.428 [t(10)=-0.825] for PowerSoil and PowerSoil Pro, respectively) or total ASVs (p = 0.249 [t(9)=-1.233] and p = 0.944 [t(10)=0.072] for PowerSoil and PowerSoil Pro, respectively) with either kit. To compare the differences in taxonomic composition between kits, methods and locations, a PCA using the *ITS2* ASV read counts was completed (Fig. 1). The similarity between extraction kits (and tested incubation temperatures) for a single location is apparent, given samples from Colorado (circles), North Carolina (squares) and Pennsylvania (diamonds) are mostly clustered together in two-dimensional space (Fig. 1). This result shows that the baseline plant community for a single location is recovered consistently, regardless of the kit and method used.

A comparison between the two extraction methods was also completed using the spiked soils (partial and normal; n, 48) by focusing only on key families which were not included in the artificial pollen mixture (n, 5; Brassicaceae, Caryophyllaceae, Juglandaceae, Rosaceae, Ulmaceae). After excluding reads assigned to one of the families of the spiked pollen taxa (Asteraceae, Fabaceae, Poaceae and Salicaceae; ~83% of total reads), 128,965 reads were available for comparisons. No statistical difference was noted in the number of total reads (p = 0.3145 [t(68.3]=1.01]) or total ASVs (p = 0.0916 [t(61.0]=1.71]) recovered for the five key families between the two kits. When comparing the kits for each soil location separately, a statistical difference in the number of total reads and ASVs was only observed for the Pennsylvania soil samples (p = 0.0266 [t(21.0]=2.38] and p = 0.0028 [t(21.1)=3.37], respectively). To compare differences between kits, methods and locations, a PCA using ASV read counts for the five families was completed (Fig. 2). The vast majority of samples were clustered together in two-dimensional space, reflecting that kit and incubation temperature had little impact on recovered plant composition (Fig. 2). Soils from Pennsylvania did show greater separation, likely due to recovering a larger number of ASVs and reads from Rosaceae and Ulmaceae (Fig. 2).

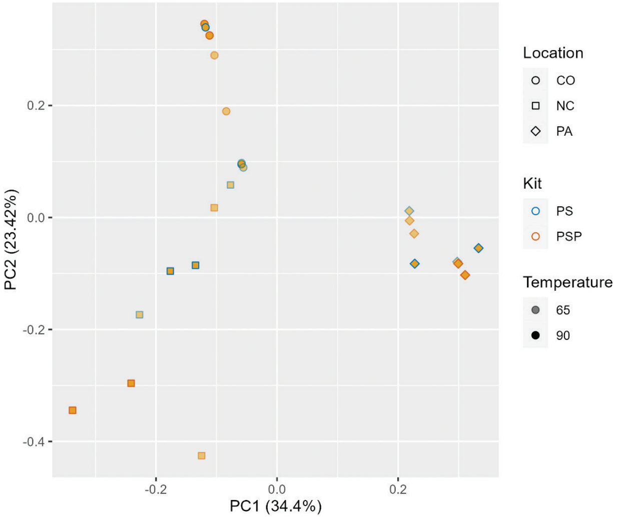

Figure 1. Principal component analysis of read counts for amplicon sequence variants belonging to one of nine key plant families in unspiked (baseline) soil samples collected from Colorado (CO; circles), North Carolina (NC; squares) and Pennsylvania (PA; diamonds) determined by *ITS2* DNA metabarcoding. The outline color of the shape denotes the kit (blue = PS [PowerSoil]; or orange = PSP [PowerSoil Pro]) and the color intensity denotes incubation temperature (light, 65 °C; dark, 90 °C). Axes represent the first and second principal components with percent variance explained in parentheses.

Lysis of spiked taxa

To compare the efficiency of the two different extraction kits on lysing the pollen spiked into the soil samples, ASVs which returned a high-quality match to the same genus of the four spiked pollen taxa were identified. A total of 151 ASVs across all samples were

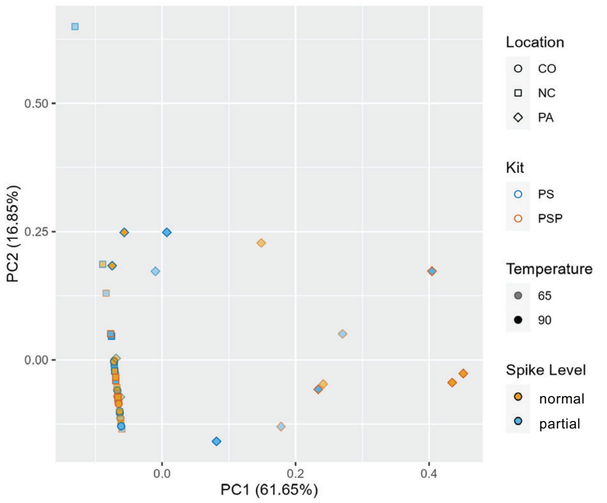

Figure 2. Principal component analysis of read counts for amplicon sequence variants belonging to five key plant families in spiked soil samples (partial and normal) collected from Colorado (CO; circles), North Carolina (NC; squares) and Pennsylvania (PA; diamonds) determined by *ITS2* DNA metabarcoding. The outline color of the shape denotes the kit (blue = PS [PowerSoil]; or orange = PSP [PowerSoil Pro]), fill color of the shape denotes spike level (blue = partial; orange = normal) and the color intensity denotes incubation temperature (light, 65 °C; dark, 90 °C). Axes represent the first and second principal components with percent variance explained in parentheses.

assigned to these four genera with breakdown as follows: *Zea* (Poaceae) – n, 0; *Symphyotrichum* (Asteraceae) – n, 6; *Populus* (Salicaceae) – n, 21; and *Trifolium* (Fabaceae) – n, 124. While these three genera only represent 9.6% of total ASVs, they encompass 46.3% of all total reads (406,904). A statistically significant difference in the total number of reads assigned to the spiked genera was observed between unspiked and partially spiked samples (p = <0.0001 [t(29.3]=-10.2]), along with unspiked and normal spiked samples (p = <0.0001 [t(26.9]=-7.66]).

The design of this study allows the limit of detection to be evaluated based on the compositional differences between each of the spiked taxa in the artificial mixture. For the comparisons described herein, we are using read count as a proxy for taxa abundance, given numerous previous pollen DNA metabarcoding studies have reported a positive correlation between sequence reads and relative abundance (Keller et al. 2015; Richardson et al. 2015a; Pornon et al. 2017; Rojo et al. 2019; Baksay et al. 2020). Notably for *Z. mays*, the calculated number of grains present in 100 mg subsamples was <1 (Table 2). As expected, no ASVs were assigned to the genus *Zea* in any of the samples. Two other published studies have reported issues with the detection of *Zea* in artificial pollen mixtures, primarily attributed to high GC content which can interfere with amplification (Bell et al. 2019; Swenson and Gemeinholzer 2021). To confirm this was not the reason for a lack of *Zea* ASVs in this study, separate amplification of *Z. mays* DNA was performed using ITS2F/ITSp4 and an amplicon of the expected size was obtained (results not shown). Thus, the lack of *Zea* ASVs in this study reflects the absence of *Z. mays* pollen in soil subsamples subjected to DNA isolation, given amplification was previously successful and >30 *ITS2* sequences from *Zea* species are available in GenBank for comparisons.

Table 3. Observed spiked-in proportions of three genera included in the artificial pollen mixture, based on read counts (*Zea* excluded). The expected proportions for each genus are given in the column headers (adjusted for the exclusion of *Zea*). Only samples for which there was an increase in read count when compared to the appropriate unspiked samples are reported. Abbreviations are as follows: CO, Colorado; NC, North Carolina; PA, Pennsylvania; PS, PowerSoil kit; PSP, PowerSoil Pro kit.

	State	Kit	Temp (°C)	*Populus* (9.6%)	*Symphyotrichum* (40.7%)	*Trifolium* (49.7%)
Spiked normal	CO	PS	65	5.2	6.9	87.9
	NC	PS	65	16.7	19.3	63.9
	NC	PS	90	38.5	60	1.5
	NC	PSP	65	30.1	30.2	39.7
	PA	PS	65	31.7	39.1	29.3
	PA	PS	90	35.9	35.7	28.4
	PA	PSP	65	16.4	18.3	65.3
	PA	PSP	90	22.1	20.7	57.2
Spiked partial	CO	PS	65	9.2	0.1	90.7
	NC	PS	65	28.9	65.7	5.4
	PA	PS	65	19.6	34.9	45.5
	PA	PS	90	36.9	39.8	23.3
	PA	PSP	65	36.4	31.4	32.2
	PA	PSP	90	42.1	35.5	22.4

Across all spiked samples (n, 48), the proportion of reads assigned to the remaining three genera were as follows: 0.04% – *Symphyotrichum*, 28.04% –*Populus*, and 71.92% – *Trifolium*. These results do not correspond with the proportions of each of these species spiked into the artificial pollen mixture. The spiked samples did not consistently have higher reads for each of the known spiked genera; only eight normal spiked and six partially spiked samples had an increase in reads when compared to the appropriate unspiked sample. In the cases where there was an increase in read count for any of the known spiked genera (except *Zea*), the proportions were significantly different from the expected for both the normal spiked samples (p = <0.001, T^2 1101.5 on 2 and 10 degrees of freedom) and partially spiked samples (p = <0.001, T^2 270.2 on 2 and 10 degrees of freedom). The recovered proportions for those samples are given in Table 3. For the normal spiked samples, three out of four samples from Colorado had lower *Symphyotrichum* reads than the unspiked sample and one North Carolina sample had lower *Trifolium* reads than the unspiked sample. In the partially spiked samples, three out of four Colorado samples had lower *Symphyotrichum* reads compared to the unspiked sample and three out of four North Carolina samples had reduced *Trifolium* reads compared to the unspiked sample. All spiked samples had increased

Populus. High quantities of *Trifolium* ASVs/reads were expected, given it was the most abundant species in the artificial pollen mixture (~49.5%). Data recovered from *Symphyotrichum* was significantly less than expected. Given a) published studies have not reported any primer mismatches or bias with *ITS2* and members of the Asteraceae family (Moorhouse-Gann et al. 2018; Timpano et al. 2020) and b) DNA isolation, amplification (*ITS2*, *rbcL* and *trnL*) and analysis via Sanger sequencing of the *Symphyotrichum* pollen prior to combining in the artificial mixture was successful (indicating DNA was lysed and of sufficient quality and quantity), low recovery is likely attributed to the incomplete homogenization. Corbiculae pollen was used for *Symphyotrichum* and after completing the washing protocol pollen was the consistency of wet sand. Given this, some clumping of *Symphyotrichum* pollen was likely such that it may not have been recovered in a 100 mg subsample, causing the taxon to drop out of sequencing results completely. Conversely, reads assigned to *Populus* were higher than expected (~9.6% *vs*. 28.04%) and likely reflects the more straight-forward and uniform homogenization of granular pollen in surface soil. Other published studies have also reported differences between the expected and observed proportions of known species in artificial pollen mixtures (*e.g.*, Hawkins et al. 2015; Kraaijeveld et al. 2015; Richardson et al. 2015a, b; Bell et al. 2019; Macgregor et al. 2019; Swenson and Gemeinholzer 2021) possibly due to biases including unequal starting material, genome duplication, and primer annealing.

When examining the logarithmic fold change in read count for spiked taxa between spiked and unspiked sample pairs (Fig. 3), it is evident that lysis and downstream sequencing of spiked pollen was successful with both kits and incubation temperatures (*i.e.*, a log-scale difference reflects an increase in recovered reads for spiked samples). Notably in all but two cases, samples processed using the PowerSoil Pro kit showed greater similarity to the unspiked sample pair (lower log-scale differences; Fig. 3), indicating possibly that lysis was less effective than when using the PowerSoil Kit. Arguably, a more straightforward approach to assess the impact of lysis conditions and extraction kits on pollen lysis, would have been to spike surface soils with an artificial pollen mixture that only contained species not present in the surface soils. This was not feasible in this study for two reasons: 1) very few vendors sell single source pollen, and thus ensuring pollen diversity (taxonomically and morphologically) would have been challenging, and 2) morphological identification of pollen and other plant materials (*i.e.*, seeds, root/leaf fragments *etc*.) in the three soil samples would have been required prior to the experimental set up, which is outside our area of expertise.

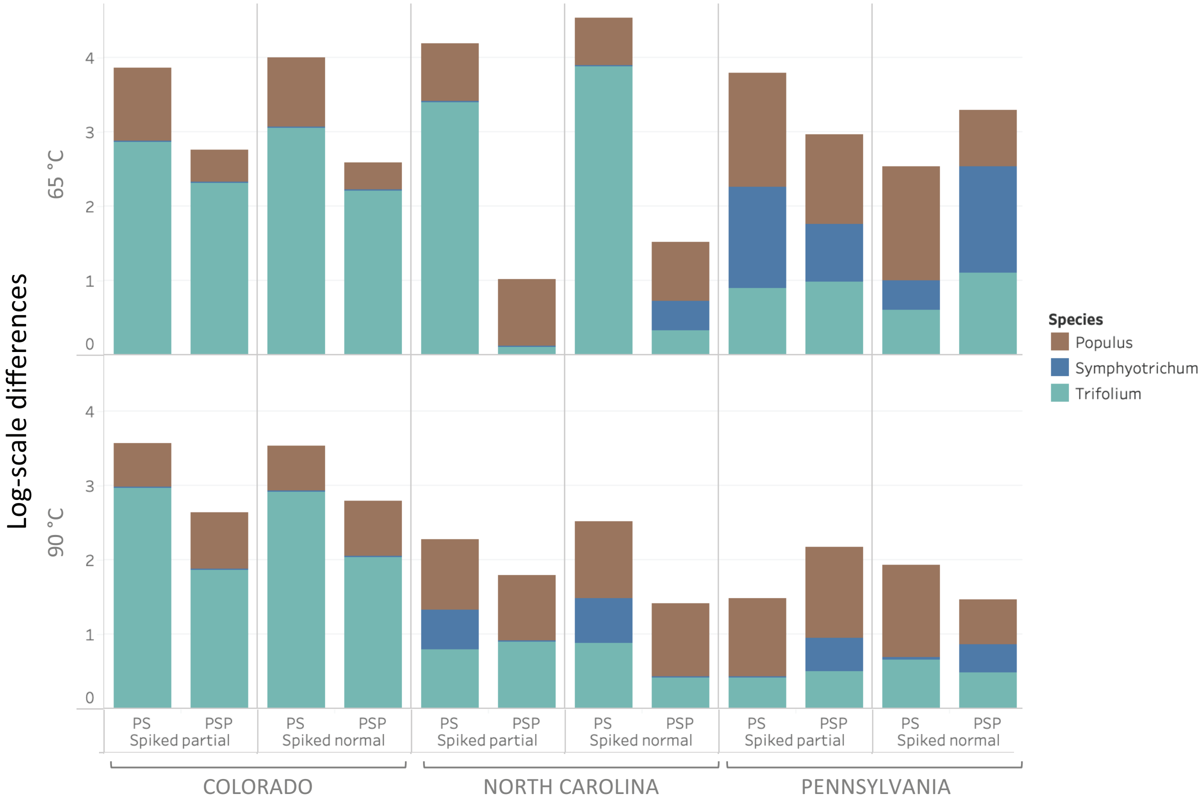

Figure 3. Logarithmic fold change in read count for the spiked taxa (*Populus*, *Symphyotrichum*, *Trifolium*) between spiked (partial and normal) and unspiked pairs. Data are separated by incubation temperature (65 °C top, 90 °C bottom). Abbreviations are as follows: PS, PowerSoil kit; PSP, PowerSoil Pro kit.

Conclusions

This study focused on assessing the lysis of morphologically and taxonomically diverse pollen from one of the most common bulk environmental sample types for DNA metabarcoding, surface soil. To achieve this, an artificial pollen mixture was spiked into surface soils from North Carolina, Colorado and Pennsylvania and the DNA subsequently isolated using two commercially available soil extraction kits widely used by the scientific community. The PowerSoil Pro Kit statistically outperformed the PowerSoil Kit based on total DNA yields. For either kit, incubation temperature (65 °C or 90 °C) used had no impact on the recovery of DNA, ASVs, or total reads. A statistically significant increase in the total number of reads for the spiked pollen species was observed with both kits, which confirmed five key findings of this study: 1) pollen was successfully spiked into soil samples, 2) grain lysis releasing high-quality DNA was achieved using both kits and methods (*i.e.*, different incubation temperatures), 3) the DNA contained within dry and corbiculae pollen was of sufficient quality and quantity to permit amplification and sequencing of *ITS2*, 4) the primer pair used permit the recovery of *ITS2* from broad taxonomic groups, and 5) the components and chemicals associated with soil samples did not negatively impact the isolation of DNA from pollen grains using either kit or method. Future studies should assess whether the PowerSoil Pro Kit is appropriate for lysing pollen from other bulk environmental sample types, such as dust and feces for downstream DNA metabarcoding.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

We thank Traci Carlson and Emma Timpano for collecting surface soils, along with Dr. Rebecca Irwin (NC State University) for providing corbiculae pollen. We also thank Dr. Christopher Bernhardt (USGS) for feedback on the design of this study, Teresa Tiedge for assistance with sequence data analysis, and Khushi Patel for assistance with statistical analyses. Seed funding was provided by the NC State University College of Veterinary Medicine.

References

Baksay S, Pornon A, Burrus M, Mariette J, Andalo C, Escaravage N (2020) Experimental quantification of pollen with DNA metabarcoding using ITS1 and trnL. Scientific Reports 10(1): e4202. https://doi.org/10.1038/s41598-020-61198-6

Bell KL, de Vere N, Keller A, Richardson RT, Gous A, Burgess KS, Brosi BJ (2016) Pollen DNA barcoding: Current applications and future prospects. Genome 59(9): 629–640. https://doi.org/10.1139/gen-2015-0200

Bell KL, Fowler J, Burgess KS, Dobbs EK, Gruenewald D, Lawley B, Morozumi C, Brosi BJ (2017) Applying pollen DNA metabarcoding to the study of plant-pollinator interactions. Applications in Plant Sciences 5(6): e1600124. https://doi.org/10.3732/apps.1600124

Bell KL, Burgess KS, Botsch JC, Dobbs EK, Read TD, Brosi BJ (2019) Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. Molecular Ecology 28(2): 431–455. https://doi.org/10.1111/mec.14840

Bryant Jr VM, Jones JG, Mildenhall DC (1990) Forensic palynology in the United States of America. Palynology 14(1): 193–208. https://doi.org/10.1080/01916122.1990.9989380

Bulut H (2019) An R Package for Multivariate Hypothesis Tests: Mvtests. NWSA Academic Journals 14(4): 132–138. https://doi.org/10.12739/NWSA.2019.14.4.2A0175

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13(7): 581–583. https://doi.org/10.1038/nmeth.3869

Chamberlain SA, Szöcs E (2013) taxize: Taxonomic search and retrieval in R. F1000 Research 2: 191. https://doi.org/10.12688/f1000research.2-191.v1

Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C (2010) Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. PLoS ONE 5(1): e8613. https://doi.org/10.1371/journal.pone.0008613

Cheng T, Xu C, Lei L, Li C, Zhang Y, Zhou S (2015) Barcoding the kingdom Plantae: New PCR primers for ITS regions of plants with improved universality and specificity. Molecular Ecology Resources 16(1): 138–149. https://doi.org/10.1111/1755-0998.12438

Delcourt HR, Delcourt PA, Davidson JL (1983) Mapping and calibration of modern pollen-vegetation relationships in the southeastern United States. Review of Palaeobotany and Palynology 39(1–2): 1–45. https://doi.org/10.1016/0034-6667(83)90009-X

Dunker S, Motivans E, Rakosy D, Boho D, Mäder P, Hornick T, Knight TM (2020) Pollen analysis using multispectral imaging flow cytometry and deep learning. The New Phytologist 229(1): 593–606. https://doi.org/10.1111/nph.16882

Fahner NA, Shokralla S, Baird DJ, Hajibabaei M (2016) Large-Scale Monitoring of Plants through Environmental DNA Metabarcoding of Soil: Recovery, Resolution, and Annotation of Four DNA Markers. PLoS ONE 11(6): e0157505. https://doi.org/10.1371/journal.pone.0157505

Hawkins J, de Vere N, Griffith A, Ford CR, Allainguillaume J, Hegarty MJ, Baillie L, Adams-Groom B (2015) Using DNA Metabarcoding to Identify the Floral Composition of Honey: A New Tool for Investigating Honey Bee Foraging Preferences. PLoS ONE 10(8): e0134735. https://doi.org/10.1371/journal.pone.0134735

Holt K, Allen G, Hodgson R, Marsland S, Flenley J (2011) Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. Review of Palaeobotany and Palynology 167(3–4): 175–183. https://doi.org/10.1016/j.revpalbo.2011.08.006

Horikoshi M, Tang Y (2018) *ggfortify*: Data Visualization Tools for Statistical Analysis Results. https://CRAN.R-project.org/package=ggfortify

Jackson JF (1987) DNA repair in pollen. A review. Mutation Research. Fundamental and Molecular Mechanisms of Mutagenesis 181: 17–29. https://doi.org/10.1016/0027-5107(87)90283-1

Jørgensen T, Kjær Kh, Haile J, Rasmussen M, Boessenkool S, Andersen K, Coissac E, Taberlet P, Brochmann C, Orlando L, Gilbert MTP, Willerslev E (2012a) Islands in the ice: Detecting past vegetation on Greenlandic nunataks using historical records and sedimentary ancient DNA Meta-barcoding. Molecular Ecology 21(8): 1980–1988. https://doi.org/10.1111/j.1365-294X.2011.05278.x

Jørgensen T, Haile J, Möller P, Andreev A, Boessenkool S, Rasmussen M, Kienast F, Coissac E, Taberlet P, Brochmann C, Bigelow NH, Andersen K, Orlando L, Gilbert MTP, Willerslev E (2012b) A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. Molecular Ecology 21(8): 1989–2003. https://doi.org/10.1111/j.1365-294X.2011.05287.x

Keller A, Danner N, Grimmer G, Ankenbrand M, Ohe K, Ohe W, Rost S, Härtel S, Steffan-Dewenter I, Mock H-P (2015) Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. Plant Biology 17(2): 558–566. https://doi.org/10.1111/plb.12251

Kelso GK (1994) Pollen Percolation Rates in Euroamerican-Era Cultural Deposits in the Northeastern United States. Journal of Archaeological Science 21(4): 481–488. https://doi.org/10.1006/jasc.1994.1048

Kenjerić D, Mandić ML, Primorac L, Čačić F (2008) Flavonoid pattern of sage (*Salvia officinalis* L.) unifloral honey. Food Chemistry 110(1): 187–192. https://doi.org/10.1016/j.foodchem.2008.01.031

Kraaijeveld K, Weger LA, Ventayol García M, Buermans H, Frank J, Hiemstra PS, Dunnen JT (2015) Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. Molecular Ecology Resources 15(1): 8–16. https://doi.org/10.1111/1755-0998.12288

Lang D, Tang M, Hu J, Zhou X (2019) Genome-skimming provides accurate quantification for pollen mixtures. Molecular Ecology Resources 19(6): 1433–1446. https://doi.org/10.1111/1755-0998.13061

Lê S, Josse J, Husson F (2008) FactoMineR: A Package for Multivariate Analysis. Journal of Statistical Software 25(1): 1–18. https://doi.org/10.18637/jss.v025.i01

Leontidou K, Leontidou K, Vernesi C, Vernesi C, De Groeve J, De Groeve J, Cristofolini F, Cristofolini F, Vokou D, Vokou D, Cristofori A, Cristofori A (2018) DNA metabarcoding of airborne pollen: New protocols for improved taxonomic identification of environmental samples. Aerobiologia 34(1): 63–74. https://doi.org/10.1007/s10453-017-9497-z

Macgregor CJ, Kitson JJN, Fox R, Hahn C, Lunt DH, Pocock MJO, Evans DM (2019) Construction, validation, and application of nocturnal pollen transport networks in an agro-ecosystem: A comparison using light microscopy and DNA metabarcoding. Ecological Entomology 44(1): 17–29. https://doi.org/10.1111/een.12674

Maher Jr LJ (1972) Absolute pollen diagram of Redrock Lake, Boulder County, Colorado. Quaternary Research 2(4): 531–553. https://doi.org/10.1016/0033-5894(72)90090-7

Manivanan P, Rajagopalan SM, Subbarayalu M (2018) Studies on authentication of true source of honey using pollen DNA barcoding. Journal of Entomology and Zoology Studies 6: 255–261.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal 17(1): 10–12. https://doi.org/10.14806/ej.17.1.200

Meiklejohn KA, Jackson ML, Stern LA, Robertson JM (2018) A protocol for obtaining DNA barcodes from plant and insect fragments isolated from forensic-type soils. International Journal of Legal Medicine 132(6): 1515–1526. https://doi.org/10.1007/s00414-018-1772-1

Mildenhall DC (2006) Hypericum pollen determines the presence of burglars at the scene of a crime: An example of forensic palynology. Forensic Science International 163(3): 231–235. https://doi.org/10.1016/j.forsciint.2005.11.028

Mildenhall DC, Wiltshire PEJ, Bryant VM (2006) Forensic palynology: Why do it and how it works. Forensic Science International 163(3): 163–172. https://doi.org/10.1016/j.forsciint.2006.07.012

Moorhouse-Gann RJ, Dunn JC, Symondson WOC (2018) New universal ITS2 primers for high-resolution herbivory analyses using DNA metabarcoding in both tropical and temperate zones. Scientific Reports 8(1): e8542. https://doi.org/10.1038/s41598-018-26648-2

Niemeyer B, Epp LS, Stoof-Leichsenring KR, Pestryakova LA, Herzschuh U (2017) A comparison of sedimentary DNA and pollen from lake sediments in recording vegetation composition at the Siberian treeline. Molecular Ecology Resources 17(6): e46–e62. https://doi.org/10.1111/1755-0998.12689

Peel N, Dicks LV, Clark MD, Heavens D, Percival-Alwyn L, Cooper C, Davies RG, Leggett RM, Yu DW (2019) Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics (RevMet). Methods in Ecology and Evolution 10(10): 1690–1701. https://doi.org/10.1111/2041-210X.13265

Pornon A, Andalo C, Burrus M, Escaravage N (2017) DNA metabarcoding data unveils invisible pollination networks. Scientific Reports 7(1): 1–11. https://doi.org/10.1038/s41598-017-16785-5

R Core Team (2022) R: A language and environment for statistical computing. Vienna, Austria.

Richardson RT, Lin C, Sponsler DB, Quijia JO, Goodell K, Johnson RM (2015a) Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey bees in an agroecosystem. Applications in Plant Sciences 3(1): e1400066. https://doi.org/10.3732/apps.1400066

Richardson RT, Lin C, Quijia JO, Riusech NS, Goodell K, Johnson RM (2015b) Rank-based characterization of pollen assemblages collected by honey bees using a multi-locus metabarcoding approach. Applications in Plant Sciences 3(11): e1500043. https://doi.org/10.3732/apps.1500043

Rojo J, Núñez A, Lara B, Sánchez-Parra B, Moreno DA, Pérez-Badia R (2019) Comprehensive analysis of different adhesives in aerobiological sampling using optical microscopy and high-throughput DNA sequencing. Journal of Environmental Management 240: 441–450. https://doi.org/10.1016/j.jenvman.2019.03.116

Russell EWB (1993) Early Stages of Secondary Succession Recorded in Soil Pollen on the North Carolina Piedmont. American Midland Naturalist 129(2): 384–396. https://doi.org/10.2307/2426519

Sassen M (1964) Fine structure of petunia pollen grain and pollen tube. Acta Botanica Neerlandica 13(2): 175–181. https://doi.org/10.1111/j.1438-8677.1964.tb00150.x

Shivanna KR, Tandon R (2014) Reproductive Ecology of Flowering Plants: A Manual. Springer, India. https://doi.org/10.1007/978-81-322-2003-9

Simel EJ, Saidak LR, Tuskan GA (1997) Method of Extracting Genomic DNA from Non-Germinated Gymnosperm and Angiosperm Pollen. BioTechniques 22(3): 390–394. https://doi.org/10.2144/97223bm02

Sønstebø J, Gielly L, Brysting AK, Elven R, Edwards M, Haile J, Willerslev E, Coissac E, Rioux D, Sannier J, Taberlet P, Brochmann C (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. Molecular Ecology Resources 10(6): 1009–1018. https://doi.org/10.1111/j.1755-0998.2010.02855.x

Stanley RG, Linskens HF (1974) Pollen. Springer, Berlin, 310 pp. https://doi.org/10.1007/978-3-642-65905-8

Sugita S (2007) Theory of quantitative reconstruction of vegetation I: Pollen from large sites REVEALS regional vegetation composition. The Holocene 17(2): 229–241. https://doi.org/10.1177/0959683607075837

Swenson SJ, Gemeinholzer B (2021) Testing the effect of pollen exine rupture on metabarcoding with Illumina sequencing. PLoS ONE 16: e0245611. https://doi.org/10.1371/journal.pone.0245611

Tang Y, Horikoshi M, Li W (2016) ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages. The R Journal 8(2): 474–485. https://doi.org/10.32614/RJ-2016-060

Taylor PE, Jonsson H (2004) Thunderstorm asthma. Current Allergy and Asthma Reports 4(5): 409–413. https://doi.org/10.1007/s11882-004-0092-3

Timpano EK, Scheible MKR, Meiklejohn KA (2020) Optimization of the second internal transcribed spacer (ITS2) for characterizing land plants from soil. PLoS ONE 15(4): e0231436. https://doi.org/10.1371/journal.pone.0231436

van den Boogaart KG, Tolosana-Delgado R, Bren M (2022) compositions: Compositional Data Analysis. http://www.stat.boogaart.de/compositions/

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wiltshire PEJ, Hawksworth DL, Edwards KJ (2015) A Rapid and Efficient Method for Evaluation of Suspect Testimony: Palynological Scanning. Journal of Forensic Sciences 60(6): 1441–1450. https://doi.org/10.1111/1556-4029.12835

Young JM, Weyrich LS, Cooper A (2014) Forensic soil DNA analysis using high-throughput sequencing: A comparison of four molecular markers. Forensic Science International. Genetics 13: 176–184. https://doi.org/10.1016/j.fsigen.2014.07.014

Zhou L-J, Pei K-Q, Zhou B, Ma K-P (2007) A molecular approach to species identification of Chenopodiaceae pollen grains in surface soil. American Journal of Botany 94(3): 477–481. https://doi.org/10.3732/ajb.94.3.477

Supplementary material 1
Table S1

Author: Madison A. Moore, Melissa K.R. Scheible, James B. Robertson, Kelly A. Meiklejohn

Data type: PDF file

Explanation note: **Table S1.** Raw data on the total number of reads and ASV across all samples for each of the 54 plant families identified. * denotes families in which both the % of total reads and % of total ASV were above 1% and subsequently included in downstream statistical analyses (data from remaining 45 families were combined as ‘Other’). Abbreviations are as follows” ASV, amplicon sequence variant.

Link: https://doi.org/10.3897/mbmg.6.89753.suppl1

Supplementary material 2
Table S2

Author: Madison A. Moore, Melissa K.R. Scheible, James B. Robertson, Kelly A. Meiklejohn

Data type: PDF file

Explanation note: **Table S2.** Raw sample data for key metrics in the DNA metabarcoding wet laboratory and bioinformatics processing. Abbreviations are as follows: PS, DNeasy® PowerSoil® kit; PSP, DNeasy® PowerSoil® Pro kit; ASV, amplicon sequence variant.

Link: https://doi.org/10.3897/mbmg.6.89753.suppl2