A novel approach for pollen identification and quantification using hybrid capture-based DNA metabarcoding

Dona Kireta[1], Kor-jent Dijk[2], Stephen Crotty[3], Arif Malik[3], Karen Bell[4], Katja Hogendoorn[5], and Andrew Lowe[1]

[1]The University of Adelaide - North Terrace Campus
[2]University of Adelaide
[3]Affiliation not available
[4]New South Wales Department of Primary Industries
[5]The University of Adelaide

June 2, 2023

Abstract

Efforts to explore optimal molecular methods for identifying plant mixtures, particularly pollen, are increasing. Pollen identification (ID) and quantification is important in many fields, including pollination ecology and agricultural sciences, but quantifying mixture proportions remains challenging. Traditional pollen ID using microscopy is time-consuming, requires expertise, and has limited accuracy and throughput. Molecular barcoding approaches being explored offer improved accuracy and throughput. The common approach, amplicon sequencing, employs PCR amplification to isolate DNA barcodes, but introduces significant bias, impairing downstream quantification. We apply a novel molecular hybridisation capture approach to artificial pollen mixtures, to improve upon current taxon ID and quantification methods. The method randomly fragments DNA, and uses RNA baits to capture DNA barcodes, which allows for PCR duplicate removal, reducing downstream quantification bias. Metabarcoding was tested using two reference libraries constructed from publicly available sequences; the matK plastid barcode, and RefSeq complete chloroplast references. Single barcode-based taxon ID did not consistently resolve to species or genus level. The RefSeq chloroplast database performed better qualitatively but had limited taxon coverage (relative to species used here) and introduced ID issues. At family level, both databases yielded comparable qualitative results, but the RefSeq database performed better quantitatively. A restricted matK database containing only mixture species yielded sequence proportions highly correlated with input pollen proportions, demonstrating that hybridization capture usefulness for metabarcoding and quantifying pollen mixtures. The choice of reference database remains one of the most important factors affecting qualitative and quantitative accuracy.

A novel approach for pollen identification and quantification using hybrid capture-based DNA metabarcoding

Running head

Pollen ID via hybrid capture metabarcoding

Kireta D.[1], Dijk K. v.[1], Crotty S.[1], Malik A.[1], Bell K.[3,4], Hogendoorn K.[2], Lowe A. J.[1]

(1) School of Biological Sciences, The University of Adelaide, North Terrace, Adelaide SA 5005; (2) School of Agriculture, Food and Wine, The University of Adelaide, Urrbrae, Adelaide SA 5005; (3)New South Wales Department of Primary Industries, Wagga Wagga, NSW 2650, Australia;(4)School of Biological Sciences, University of Western Australia, Perth, WA 6009, Australia

Corresponding author

Dona Kireta, dona.kireta@adelaide.edu.au, The Braggs building, level 2, The University of Adelaide, North Terrace, Adelaide, SA, 5005

Ph: +61 404 282 348

Abstract

Efforts to explore optimal molecular methods for identifying plant mixtures, particularly pollen, are increasing. Pollen identification (ID) and quantification is important in many fields, including pollination ecology and agricultural sciences, but quantifying mixture proportions remains challenging. Traditional pollen ID using microscopy is time-consuming, requires expertise, and has limited accuracy and throughput. Molecular barcoding approaches being explored offer improved accuracy and throughput. The common approach, amplicon sequencing, employs PCR amplification to isolate DNA barcodes, but introduces significant bias, impairing downstream quantification. We apply a novel molecular hybridisation capture approach to artificial pollen mixtures, to improve upon current taxon ID and quantification methods. The method randomly fragments DNA, and uses RNA baits to capture DNA barcodes, which allows for PCR duplicate removal, reducing downstream quantification bias. Metabarcoding was tested using two reference libraries constructed from publicly available sequences; the *matK* plastid barcode, and RefSeq complete chloroplast references. Single barcode-based taxon ID did not consistently resolve to species or genus level. The RefSeq chloroplast database performed better qualitatively but had limited taxon coverage (relative to species used here) and introduced ID issues. At family level, both databases yielded comparable qualitative results, but the RefSeq database performed better quantitatively. A restricted *matK* database containing only mixture species yielded sequence proportions highly correlated with input pollen proportions, demonstrating that hybridization capture usefulness for metabarcoding and quantifying pollen mixtures. The choice of reference database remains one of the most important factors affecting qualitative and quantitative accuracy.

Key words

Pollen metabarcoding, Pollen quantification, Hybridization capture, Target enrichment

Introduction

Pollen identification (ID) is important for many scientific fields. Key areas are pollination ecology and agricultural sciences, but accurate pollen ID also supports the study of ancient plant communities (Clarke et al., 2020), human health (e.g. allergy research (Weber, 1998)), and forensics (Alotaibi et al., 2020). Traditional methods of pollen ID rely on microscopy to observe diagnostic characters on the pollen exine. This method is time consuming and requires a high level of expertise, while being limited in accuracy and throughput, and potentially constrains many projects. The limitations of microscopy-based pollen ID are well established. In most cases, taxa can only be identified to family, or in some cases genus (Kraaijeveld et al., 2015; Richardson, Lin, Sponsler, et al., 2015; Smart et al., 2017). The time-consuming nature of microscopy-based ID limits the throughput, and usually only a subsample of each sample can be examined, meaning that rare taxa are often missed (Bell et al., 2016; Smart et al., 2017).

Due to these limitations, alternative methods for pollen ID have been sought. DNA barcoding, or metabarcoding (mixed samples) has advanced taxon ID in many research fields, has been explored extensively for pollen ID, and has been shown to provide accurate identifications at high taxonomic resolution and with high sample throughput (Bell et al., 2019; Bell et al., 2017; de Vere et al., 2017; Keller et al., 2015; Kraaijeveld et al., 2015; Richardson, Lin, Quijia, et al., 2015; Richardson, Lin, Sponsler, et al., 2015; Suchan, Talavera, Saez, Ronikier, & Vila, 2019; Wilson, Sidhu, LeVan, & Holway, 2010). In particular,

metabarcoding is able to recover a taxonomic ID from as few as five pollen grains (Pornon et al., 2016), and the method has the ability to ID many more genera than microscopy-based methods (Keller et al., 2015; Richardson, Lin, Sponsler, et al., 2015).

The accuracy of metabarcoding is limited, however, by the choice of barcode and comprehensiveness of reference databases, since only taxa with reference sequences can be detected. Database collections have been established where references can be stored and accessed, and these are growing. The cytochrome *c* oxidase subunit 1 (CO1) barcode is able to differentiate most animal taxa, and can be accessed through the Barcode Of Life Data system (Ratnasingham & Hebert, 2007). However, the selection of effective plant ID barcodes has presented a much greater challenge, since CO1 is not variable enough in plants to provide taxonomic resolution (CBOL Plant Working Group, 2009). The Consortium for the Barcode of Life (CBOL) Plant Working group recommends the chloroplast genome encoded maturase K (*matK*) and ribulose 1,5-biphosphate carboxylase (*rbcL*) as standard barcodes which can ID approximately 70% of all plant taxa, provided they are present in the reference database (CBOL Plant Working Group, 2009). Other barcodes have also been recommended for specific groups of plants, or as supplementary barcodes, such as the *psbA-trnH* spacer (Kress & Erickson, 2012). The success of standard barcodes relies on sequence variability to allow resolution of taxa, and conserved primer binding sites to allow for sequence analysis across a broad range of taxa. The common barcoding approach uses PCR to amplify the barcode using primer sites, followed by sequencing and comparison to a reference database. When reference sequences for target species are absent, the similarity to the closest sequence(s) in the database can be used to generate a genus or family ID (Liu, Clarke, Baker, Jordan, & Burridge, 2019).

Despite the demonstrated strengths of metabarcoding, the inability to answer quantitative questions regarding sample composition remains problematic. In pollination research, it is often desirable to know the relative proportions of taxa in a pollen sample. This information can shed light on the preference of pollinators or abundance of resources, and can improve understanding of pollination networks and ecosystem robustness, which in turn can help restore pollination services in natural and agricultural settings (Dormontt et al., 2018). Currently, there is mixed success in comparisons of relative proportions of DNA sequencing reads to starting pollen proportions for mixed samples (Bell et al., 2017). Positive correlations have been found between proportions of sequence reads and DNA mixes using *trnL* and ITS1 barcodes (Pornon et al., 2016), sequence proportions and starting pollen proportions using ITS2 (Keller et al., 2015), and between averaged *rbcL* and *matK* sequence abundance (Richardson, Lin, Quijia, et al., 2015). However, the latter study also found poor quantification with ITS2, and others found similarly less conclusive results, with weak correlations between sequence and starting sample proportions using ITS2 (Bell et al., 2019), and no conclusive results using ITS (Smart et al., 2017). A meta-analysis on metabarcoding used in 22 ecological studies found only a weak positive association between starting biomass and sequences recovered, with large uncertainty (Lamb et al., 2019). The weak or poor results arise from bias at several steps in the sample to sequence pipeline. Biases occur which can affect both the qualitative (whether the correct taxa are identified), and quantitative (proportion within mixture) aspect of metabarcoding. Any bias affecting qualitative accuracy can affect quantitative accuracy, by potentially lowering some taxa below the detection limit.

Factors including poor resolution of barcodes and biased representation within reference databases affect ID leading to inaccurate quantitative estimates. Additional factors include: differences in DNA isolation method (Pornon et al., 2016); amplification differences between taxa due to differences in primer binding affinity (Krehenwinkel et al., 2017) - which can lead to false negatives (when a present taxon is not identified) (Pawluczyk et al., 2015; Zinger et al., 2019) and downstream quantification biases; different barcode copy numbers (Krehenwinkel et al., 2017); DNA degradation bias (Krehenwinkel et al., 2018); and database quality issues (Richardson, Bengtsson-Palme, & Johnson, 2017). Sequencing bias can also occur between both barcodes and taxa (Pawluczyk et al., 2015). Unequal PCR replication (mostly affecting related taxa) and variable barcode copy number (particularly affecting chloroplast loci (Golczyk et al., 2014) which contain the standard plant barcodes) likely play the greatest roles in introducing bias (Krehenwinkel et al., 2017). In fact, Pawluczyk et al. (2015) found up to a 2000 fold difference in DNA quantity between taxa and loci after PCR. PCR-free methods are being explored as a means to overcome these quantitative challenges, and they show improvement in quantification over PCR-based metabarcoding, for example genome skimming and chloroplast assembly (Lang, Tang, Hu, & Zhou, 2019), Whole Genome Shotgun sequencing (Bell et al., 2021), and MinION Reverse Metagenomics (Peel et al., 2019). However, these methods have other drawbacks. Genome skimming and Whole Genome Sequencing (WGS) for example require a larger amount of DNA, which can be difficult to obtain from small solitary pollinators (Bell et al., 2021; Lang et al., 2019), and MinION Reverse Metagenomics requires the user to curate their own reference databases (Peel et al., 2019).

One method that could overcome these shortcomings and improve accuracy and quantification compared to existing methods of pollen metabarcoding ID is hybridisation (hereafter hybrid) capture. Hybrid capture is a target enrichment technique that has recently been applied to environmental/ecological studies. It can be used for degraded DNA, and has been used to create a reference database from herbarium specimens (Dormontt et al., 2018), explore historic ecological communities through sediment cores (Foster et al., 2021; Schulte et al., 2021), and phylogenetic studies (Nge, Biffin, Thiele, & Waycott, 2021). The method uses a probe, or bait, which is an RNA molecule complementary to the gene region of interest. Since the method does not rely on PCR to isolate the genomic regions of interest, it has the potential to remove PCR bias from the quantification analyses, which has been found to generate large quantitative bias in amplification-based metabarcoding approaches, and can cause taxon-specific amplification bias (Krehenwinkel et al., 2017; Pawluczyk et al., 2015).

For taxonomic ID, the bait is complementary to the barcode of interest (Waycott, van Dijk, & Biffin, 2021). The baits used in this study were designed to target 19 chloroplast genes (see Waycott et al., 2021), applicable to all angiosperm linages. To make them useful for such broad ranges of taxa, the baits do not need to match 100% to the barcode, 80-90% similarity will retrieve the target, and affinity can be controlled with the hybridisation temperature. The sequence overhang generated with hybrid capture baits can often recover complete or near complete chloroplast genomes. In traditional PCR amplification methods, primers are bound to conserved barcode primer sites to amplify the barcodes. This creates exact copies of the barcodes that cannot easily be distinguished from the PCR duplicates. Our approach uses sonication to randomly fragment the DNA after DNA extraction, creating a random DNA

fragment soup. Chloroplast loci (genes) for which baits were designed are then 'fished out' of the soup using the complementary baits (Waycott et al., 2021). Given that each DNA fragment has in theory a unique length, PCR duplicates (amplicons having same sequence and length) can be eliminated bioinformatically and only one copy of every captured sequenced read or read pair is retained. This enables downstream quantification of relative taxon abundances based on the number of reads mapping to references.

The aim of this study was to demonstrate the effectiveness of hybrid capture DNA metabarcoding for identifying taxa in a pollen mix, and determining the accuracy of estimations of relative taxonomic abundances. We used two different reference databases, a *matK* database which is commonly used in amplicon metabarcoding, and a RefSeq whole chloroplast database. We expected that the RefSeq database would produce more accurate qualitative and quantitative results, since many more potentially informative gene regions were recovered using the chloroplast bait set used for hybrid capture, and PCR bias was controlled for. We explored whether, and how closely, the sequence composition of mixed pollen samples reflected starting proportions, to test the potential for broader application of hybrid capture metabarcoding as a useful tool in pollination research.

Materials and Methods

Sample collection

A comprehensive experimental setup was made using pollen of three species from different families. The pollen from these taxa was visually distinct for easy morphological identification by non-experts (Fig. 1). This ensured that the taxa comprising each pollen pellet could be verified through morphology. Pollen was obtained from honey bee hives

fitted with pollen traps. Honey bees forage on one species per foraging trip, so pollen pellets are usually comprised of a single species (Grüter & Ratnieks, 2011; Synge, 1947; Visscher & Seeley, 1982). The hives had been placed in almond orchards (*Prunus dulcis*), brown stringybark plantations (*Eucalyptus baxteri*), and a field with flowering capeweed (*Arctotheca calendula*). *A. calendula* pollen is a distinctive orange colour which was easily separated from pollen pellets of other species that were present at the time of collection.

Pollen mixtures

We constructed 14 different pollen mixtures, with three replicates of each mixture. We used four negative controls (blanks), one for each extraction batch, totalling 48 samples/libraries.

The pollen mixture proportions were weight based. Each taxon varied in quantity from high to low abundance (Table 1, Fig. 2). The mixtures were suspended in ethanol and divided into three replicates for DNA extraction. Ethanol was used for suspension because it evaporated without leaving any residuals that may have affected subsequent DNA extraction and library preparation. Care was taken to strongly agitate the mixture before aliquoting.

DNA extraction and Library preparation

DNA was extracted from the pollen mixtures (9 mg) using the NucleoSpin® Food kit (Macherey-Nagel, Düren, Germany), with the "isolation of genomic DNA from honey or pollen" supplementary protocol. We modified the homogenisation and elution steps. We homogenised the dry pollen mixture aliquots using ceramic beads in 2 mL screw cap tubes on a Bead Ruptor 24 (OMNI International Inc.) at 6 m/s for 20 s cycles (3-4 minutes total) until a powder was formed. Sample tubes were submerged in liquid nitrogen between mill cycles to prevent DNA degradation caused by heat during bead beating, and to allow easier

194 homogenisation by making the pollen brittle. The final elution step was done by passing the
195 60 µL of elution buffer through the spin column membrane twice instead of once, followed
196 by spinning, to maximise DNA yield. Following extraction, DNA was quantified using a
197 Quantus™ Fluorometer and QuantiFluor® dsDNA System (Promega, Madison, WI, USA),
198 normalised to 2 ng/µL (samples with concentration lower than 2 ng/µL were used neat), and
199 sonicated using a Bioruptor® Pico (Diagenode, USA) to create random length fragments
200 (eight cycles of 15 s on, 90 s off).

201 Library preparation was done using an Eppendorf epMotion® 5075t - Liquid Handling
202 Workstation. The DNA libraries were prepared using the NEBNext® Ultra™ II DNA Library
203 Prep kit as described in the protocol by Waycott et al. (2021). In brief, custom made y-stubby
204 adaptors were ligated to the DNA fragments. Each adaptor contained one of 48 unique 8
205 nucleotide in-line barcodes, which were combined in unique combinations (i.e. each sample
206 received a unique combination of two barcodes ligated at each end) allowing downstream
207 sample pooling. The libraries were amplified using PCR (30 s at 94°C, followed by 17 cycles of
208 98°C for 10 s, 65°C for 30 s, and 72°C for 30 s, a final extension at 72°C for 2 mins and held at
209 4°C). To reduce cost, libraries were pooled into groups of 16 according to estimates of library
210 concentration. Pools were purified using a 1:1 volume concentration of MagNA Beads
211 (Rohland & Reich, 2012).

212 ***Hybridization capture***

213 This study used the OZBaits_CP V1.0 universal plastid bait set for hybrid capture developed
214 for targeted capture of angiosperm sequences (Waycott et al., 2021), following the myBaits®
215 Targeted NGS Manual Version 4.01 hybridization protocol.

The baits were added to the pooled libraries and hybridized at 65°C for 48 hours. To avoid evaporation, chill-out™ red liquid wax (Bio-Rad Laboratories, Inc.) was added. Hybridised libraries were then amplified (2 min at 98°C, followed by 20 cycles of 98°C for 20 s, 60°C for 30 s, and 72°C for 45 s, a final extension at 72°C for 5 min and held at 8°C) with custom P7 and P5 Illumina adaptors. Following bait hybridization, target regions were bound to magnetic beads, samples were placed on a magnet and non-target regions were washed out of the product. Resulting libraries were visualised using the high sensitivity DNA assay of a 2100 Bioanalyzer (Agilent), and pooled in equimolar concentrations. Final purification used 1:1 MagNA, and final size selection at 350-600 bp was done using a 2 % agarose Pippin Prep gel cassette (Sage Science).

The unique combination of dual in-line molecular identifiers (adapter barcodes), and unique combination of dual-index primers were only used once for any library preparation in our lab to reduce contamination. The final library was sequenced at the Garvan Institute of Medical Research (Sydney, Australia) on one lane of an Illumina HiSeq X Ten with 2 × 150 cycle chemistry.

Bioinformatics pipeline: Sequence data processing and cleaning

Analyses were done using the Phoenix high performance computing cluster at the University of Adelaide, Australia. Samples were first demultiplexed via the indexes using Bcl2fastq, then demultiplexed via their internal barcodes using Sabre (Sabre-barcode-demultiplexing.). The barcodes had at least 2 degrees of separation, so one base pair mismatch was allowed.

We explored several analysis methods, including the pipelines developed by Sickel et al. (2015) and Bell et al. (2021) which were developed for metabarcoding and WGS

238 respectively. However, we were unsuccessful in implementing methods using qiime2, which
239 appeared incompatible with our non-amplicon data (we also attempted to use the q2-
240 shogun and q2-metaphlan2 plugins for shotgun data, but were unable to overcome the
241 errors encountered). We ultimately used a custom pipeline, which was similar to that of Bell
242 et al. (2021), but used modified pre-processing steps, and additionally used Bracken (Lu,
243 Breitwieser, Thielen, & Salzberg, 2017)(see below) for improved quantification. We removed
244 PCR duplicates using clumpify from BBtools (Bushnell, 2021). Removing PCR duplicates also
245 made subsequent analyses faster and less memory intensive, since the dataset had been
246 reduced by more than half. Sequence filtering and trimming was done using
247 AdapterRemoval (Schubert, Lindgreen, & Orlando, 2016). The 9th base following the 8 nt
248 barcode, reads shorter than 30 nt, reads with a phred quality score < 20, and N tails were
249 removed. Following this, Kraken2 was used to assign taxonomy to reads. Kraken2 is a k-mer
250 based method, so it does not require pre-assembly of the sequences (Wood, Lu, &
251 Langmead, 2019). It was used to classify reads at both species and genus classification levels.
252 Bracken, which is a sister program to Kraken, was then used to estimate read abundance
253 using the Kraken classifications (Lu et al., 2017). A minimum hit group threshold of 5 was set
254 in Kraken (which is useful for custom databases), and a threshold of 5 set in Bracken.
255 Bracken output was analysed using R (RStudio Team, 2020).

256 We explored different reference database approaches for taxonomic identification, the first
257 using a *matK* single barcode database, and the second using a complete chloroplast RefSeq
258 database. The databases were downloaded (January 2022) and built using Kraken and
259 Bracken. A list of all angiosperm species occurring in South Australia was obtained from the
260 Atlas of Living Australia (https://www.ala.org.au/). The publicly available sequences for *matK*

261 references were downloaded using this list. The RefSeq database consisted of all angiosperm
262 chloroplast records available from the NCBI RefSeq database. *A. calendula* did not have a
263 RefSeq chloroplast reference, so the chloroplast sequences available on NCBI were manually
264 added to the database to ensure all taxa used in the mixtures were represented. At the time,
265 15 chloroplast sequences from 8 gene regions were available (Supplementary Table 1), and
266 of the 8 regions, 6 matched barcodes targeted by the chloroplast bait set used (Waycott et
267 al., 2021). For both databases, a modified version was created each including only the three
268 taxa present in the pollen mixtures, to test the quantification independently of taxonomic
269 ID. Then, to simulate a more realistic scenario where pollen identity is unknown, we
270 repeated the analysis with the comprehensive database. The databases are referred to as
271 wide (many taxa) and restricted (mixture taxa only).

272 ***Analysis***

273 Linear regression was used to assess the correlation between the proportions of input pollen
274 weight and resulting sequences. To determine if taxon rarity in the sample had an effect on
275 taxon detection, we used binomial mixed effect models at each taxonomic level, with
276 starting pollen weight proportion as the predictor variable, and a binomial response for
277 detection success or failure. Mix ID was set as a random fixed effect. All modelling was done
278 in RStudio (RStudio Team, 2020) using the lme4 package (Bates, Maechler, Bolker, & Walker,
279 2015).

Results

After sequencing, we retrieved a total of 38,165,440 raw sequencing reads, with an average of 397,557 reads per sample. After filtering, 11,155,855 sequences were retained, an average of 116,207 reads per sample, and 27,009,585 reads were discarded of which an average of 234,035 sequences per sample were PCR duplicates.

Sample M2a had less than 600 reads sequenced after filtering, and was excluded from interpretation as this was likely the result of a technical error and thus unreliable. Of the four blanks, only one retained any reads after the quality filtering steps were carried out.

MatK database

At the species level and using the wide *matK* database, *E. baxteri* was not detected in any sample. *Eucalyptus* was detected in all samples at genus level, apart from within the blank. *A. calendula* was detected in the same 5 samples at species and genus level. The five samples (plus a sixth with failed detection) were from mixes M13 and M14, which had starting proportions of pollen > 0.799, and no samples from mixes with lower starting proportions had positive IDs. *P. dulcis* had the best detection success, and was detected in all samples except the blank. At genus level, *Prunus* was detected in every sample, including the blank. At family level, all three taxa (*Myrtaceae*, *Asteraceae* and *Rosaceae*) were detected in every sample, except for the blank. In the blank no *Myrtaceae* was detected (Fig. 2; Supplementary Table 2).

False positives occurred when taxa which were not present in the sample were detected, or the opposite for false negatives, when taxa present in a sample were not detected. The percentage of false positive sequencing reads was 64.1% using the wide *matK* database at

species level (Fig. 3C), 52.3% at genus level (Fig. 3E), and at family level there was a 10.7% false positive rate (Fig. 3G).

The relationship between input pollen proportion and proportion of reads was generally highly correlated (R^2 = 0.62 - 0.99). *E. baxteri* was undetected at species level, so a correlation could not be calculated. At genus level, R^2 = 0.96, but the proportion of reads fell far below the desired 1:1 input to output ratio. At family level, R^2 = 0.99, and the proportion of reads detected trended closer to the 1:1 ratio, although they remained below the desired level (Fig. 3H). *A. calendula* had the same relationship between input pollen and output reads at species and genus level, which was below the plot threshold, and had the lowest R^2 value (0.62) for both taxonomic levels. At family level, *A. calendula* was similarly correlated as *E. baxteri,* with R^2 = 0.97, and a trend along but consistently below the 1:1 ratio of input pollen to output sequences (Fig. 3H). *P. dulcis* had a very similar relationship between input pollen to output sequences at each taxonomic level (Fig. 3D, F, H), with high R^2 values (species R^2 = 0.75, genus and family R^2 = 0.98). However, the ratio of sequences to starting pollen proportions was positively biased in comparison to the desired 1:1 ratio in each scenario, and the deviation increased with decreasing taxonomic resolution (Fig. 3D, F, H).

The restricted *matK* database (containing only the three taxa used to make mixtures) naturally did not result in any false positives (Fig. 3A). The proportion of sequences versus input pollen was linear and highly correlated for all taxa (R^2 = 0.97 – 0.99; Fig. 3B). The same higher than expected proportion of sequences for *P. dulcis* was seen, but *E. baxteri* and particularly *A. calendula* sequence proportions were much closer to the expected 1:1 ratio (Fig. 3B).

RefSeq database

Using the comprehensive RefSeq database, and at species level, *E. baxteri* and *A. calendula* (although detected in some samples) were found in such low quantities that they were not plottable (Fig. 4C - D). *P. dulcis* sequence proportions were strongly correlated with input pollen proportions (R^2 = 0.97), and closely tracked the 1:1 ratio until the input pollen proportions reached 0.5, beyond which sequences occurred below the expected level (Fig. 4D). At genus level, *Arctotheca* was found at equally low abundances as *A. calendula* at species level. *Eucalyptus* was found at approximately half the expected proportion (Fig. 4E), but was strongly correlated with input pollen proportion (R^2 = 0.98). *Prunus* had slightly higher sequence proportions than expected (Fig. 4F), and was less linear (R^2 = 0.9) with a similar flattening of the curve above 0.5 starting pollen proportion, similar to *P. dulcis* at species level. At family level, all three taxa showed strong correlations between input pollen and sequence proportions (R^2 = 0.81 – 1) and plotted along the 1:1 ratio, although *Rosaceae* (*P. dulcis*) had the least linearity, as previous (R^2 = 0.81; Fig. 4H). *Myrtaceae* (*E. baxteri*) sequence proportions were at expected levels overall, and *Asteraceae* (*A. calendula)* and *Rosaceae* were below and above expected levels respectively (Fig. 4G). Only *P. dulcis* was detected in the blank at all three taxonomic levels, *A. calendula* was detected only at family level, and *E. baxteri* was not detected at all. This was the same as for *matK* except for *P. dulcis* detection at species level.

The percentage of false positive sequencing reads was 72.5% using the wide RefSeq database at species level (Fig. 4C), 47.4% at genus level (Fig. 4E), and a 9.6 % false positive ID rate at family level (Fig. 4G).

346 The restricted RefSeq database (containing only the three taxa used in the mixtures) also
347 naturally did not result in any false positives. The proportion of output sequences versus
348 input pollen was strongly linear for all taxa (R^2 = 0.96 and 0.97). *E. baxteri* and *P. dulcis* points
349 showed more scatter on the plot than for *matK* for samples with less than 0.25 starting
350 pollen proportion. *A. calendula* was close to zero and the other two taxa had higher than
351 expected proportions (Fig. 4B). *E. baxteri* overall had approximately expected read
352 quantities, but *A. calendula* had much lower, and *P. dulcis* much higher than expected read
353 proportions (Fig. 4A).

354 ***Sample rarity***

355 The detection of taxa was successful regardless of the amount of starting pollen in the mix.
356 Starting pollen quantities did not have a significant effect on the detection, using either
357 barcode database for assignment, at any taxonomic level (species, genus or family). Taxon
358 detection versus input pollen proportion was tested in 24 combinations using the four
359 reference databases. In nine cases, the taxon was detected at every pollen input level (every
360 sample), so it was not possible to model (Table 1).

361 **Discussion**

362 We used hybrid capture to metabarcode artificial pollen mixtures and evaluated the efficacy
363 of taxon ID, and quantification of sequence proportions relative to the original pollen
364 mixture. We constructed reference databases using Kraken2 and publicly available
365 references from NCBI. We found that the ID of taxa within the pollen mixture provided by a
366 single barcode did not always have resolution to species or genus level. The RefSeq

chloroplast database yielded better qualitative results at these taxonomic levels, but the database was limited in taxon coverage (relative to the species used here) and read assignment issues likely occurred due to this. At family level, both databases yielded equally good qualitative results, but the RefSeq database performed better quantitatively. This result was not mirrored with restricted databases that only contained the mixture species, probably because *A. calendula* did not have a RefSeq chloroplast genome, and hence it performed better in the wide database which had other *Asteraceae* at Family level. We found overall that this hybrid capture method and bioinformatic pipeline performed well in identifying taxa at higher taxonomic levels, and found close to a 1:1 ratio of input pollen to output sequences depending on the database used. Database quality and choice had a large effect on result accuracy, since our molecular approach seemed to account for potential PCR bias. We discuss these results and limitations to this method as it stands.

Taxon identification

MatK database

At species level the *matK* database resulted in high levels of false negatives. This was unsurprising as the two standard plant barcodes recommended by CBOL for plant ID can discriminate only approximately 70% of plant species, plus there could have been additional reductions in the resolution since this figure relates to longer barcode sequences, rather than the short fragments generated here. Additionally, species within the *Myrtaceae* and *Asteraceae* families (two of the three taxa used here) can be difficult to ID (Arstingstall et al., 2021; Gao et al., 2010). One of the reasons can be high chloroplast similarity in not so closely related *Eucalyptus* species (Bayly et al., 2013), which can make barcoding difficult. In this

study, *Eucalyptus* may have been difficult to identify at species level because it had the most related taxa present in the database.

Prunus dulcis was readily identified at every taxonomic level, while *Eucalyptus baxteri* was more readily detected at genus level *(Eucalyptus)*, and *Arctotheca calendula* was only readily detected at family level *(Asteraceae)*. In the last case, however, there were no other species of *Arctotheca* in the database (there are only 4-5 accepted species in total), which meant that when the reads did not match the *matK* barcode, the closest matches were more distantly related species, contributing to the high false positive rate at genus level. Since there were many other *Prunus* and *Eucalyptus* species present in the database, *P. dulcis* and *E. baxteri* reads had many more closely related options to match to if the sequence did not match correctly, resulting in more accurate genus level IDs. In early analysis exploration with a database containing only one species per genus, the results yielded were poorer, with more false negatives at genus and family levels. This could occur because the hybrid capture method does not extract the entire barcode, so potentially important parts are missing, and the read matches to a different reference. This indicates that it could be important to have closely related species and some 'redundancy' in databases to achieve more accurate genus (if not species) level ID.

Refseq database

Except for *P. dulcis*, which was identified in every sample using the RefSeq database, we had less difficulty identifying the other taxa in the samples compared with the *matK* results. Unlike with *matK, E. baxteri* was identified in some samples at species level, and *Eucalyptus* was readily identified at genus level. At species level, the RefSeq database resulted in more

false positives than the *matK* database results, but there were fewer false negatives as well. For results from both databases, the high false positive rate could be attributed to the Illumina sequencing, which is very sensitive and can easily pick up contamination. Although, most are likely explained by misidentification of sequences that came from the true positive species, since the false positive rate drops off at the higher taxonomic levels (although still not zero at family level).

A. calendula had a poorer representation in the RefSeq database. It did not have a publicly available chloroplast reference at the time of database curation, and the database also did not contain other *Arctotheca* species. Instead, the 15 chloroplast sequences available at the time of this study were added to the database (see methods). This most likely led to the much lower than expected abundance of *A. calendula* using the restricted database. With only the 15 gene regions *A. calendula* reads could possibly hit, versus the entire chloroplast genome for the other two taxa, many of the *A. calendula* sequences which did not match the 15 reference regions well, could have matched to regions of the complete chloroplast references for the other taxa, and increased the quantity of reads to those. However, at family level, and with the wide *RefSeq* database, the proportion of *A. calendula* was closer to expected levels, since with other *Asteraceae* in the database there was more redundancy, and *A. calendula* could match to other more closely related taxa. Again, this suggests that in cases where databases are missing necessary taxa, it is useful to have references of closely related taxa which can provide genus level IDs.

Sample rarity

There was no relationship between pollen input proportion and detection rate. This result was also found by Bell et al. (2019), who additionally tested the influence of other taxa on identification. In both this study and ours, there appears to be a greater influence of taxon identity than rarity on detection.

Comparison of single barcode vs whole chloroplast database

The nature of the hybrid capture baits made the RefSeq database more appropriate for qualitative assessment for a couple of reasons. The first is that more sequences/reads were utilised (*matK* is only one of 19 loci targeted by baits). The *matK* database assigned approximately 1.5% to 3% of reads per sample to a reference, which was unsurprising given the other loci sequenced, but between 85% and 96% of reads assigned to the RefSeq database, resulting in more data being utilised. The second benefit is that the overhang that can occur as a result of randomly sized fragments matching to baits can be made use of. Unlike a single barcode database such as the *matK* database used here, were if the overhang falls outside of the barcode limits, it may prevent sequences from being assigned if the number of nucleotide mismatches exceeds the threshold set.

Quantification

A restricted database only containing the mixture taxa led to linear and highly correlated quantifications of taxon proportions for the *matK* database results, although there appeared to be taxon specific biases (these were present in all instances for both databases used). The RefSeq results, which closely followed the expected 1:1 ratio at family level, were less accurate using the restricted database. The factors discussed above affecting qualitative

success also affected the quantification of relative proportions of the taxa. The greatest deviation from the expected ratio was *A. calendula* using the RefSeq database, likely because a whole chloroplast reference was not available for *A. calendula*, thus the sequences were less readily identified and were underestimated. It is evident from this that it is important wherever possible to have equivalent reference sequences for quantitative accuracy, even though the taxon was identified in many of the samples. The most readily identified species (*P. dulcis)* was overabundant in sequence reads. We expected that there would be a systematic bias arising from the different weights of the pollen taxa. *P. dulcis* was at least twice as large as the other two species meaning that fewer pollen grains would be present in the same weight, and since angiosperm pollen grains have the same number of cells, if each taxon also had the same number of plastids per cell, then we would have expected it to have a lower proportion of sequences than the other two taxa. However, this assumption was not met, and *P. dulcis* was overabundant in all samples, rather than the reverse. This most likely occurred due to two reasons: the assumption about relatively equal numbers of plastids was not met, or the readiness of identification lead it to be overestimated. The number of plastids, and genome copy number of chloroplasts can vary greatly, from few to hundreds, between different species and tissue types, and tissue age (Morley & Nielsen, 2016). While the tissue types were the same in this study, it is likely the species had different numbers of chloroplasts and chloroplast copy number accounting for some quantitative biases. There may also have been biases stemming from the laboratory, in the DNA extraction or sequencing steps, which favoured this taxon over the others.

Comparison with other studies

Compared to other studies, the hybrid capture method of our study, provides weaker qualitative results, whereas our quantitative results are equal or better. All studies considered had highly accurate qualitative results, although the reference databases used, and their breadth, varied.

Our study had accurate identifications at family level, but at species level, we only identified all species correctly in some samples using the RefSeq database. We had high levels of false positives for all species. This is similar to the study by Bell et al. (2021), who used a whole nuclear genome RefSeq database containing publicly available angiosperm species, and found their WGS method to be almost 100% accurate in identifying the species within their pollen mixtures, but they found high levels of false positives. In contrast to this study, we had more highly correlated DNA sequencing and pollen input proportions (R^2 = 0.72 – 1 for all taxa at all taxonomic levels), while they found an increasing correlation of R^2 = 0.60 and R^2 = 0.62 for species and genus levels. The amplicon metabarcoding used by Bell et al. (2019) found largely accurate taxonomic identifications, but only weakly correlated read proportions with *rbcL* and ITS2 barcodes. The study also found that some taxa were more readily detected, as we found with *P. dulcis*. Similar to our comparison between a *matK* and RefSeq database and the results, Bell et al. (2021) found more accurately identified taxa at both species and genus level using a RefSeq database compared to *rbcL* and ITS2 amplicon sequencing (from Bell et al. (2019)).

The study using RevMet by Peel et al. (2019) reliably identified plants in mixed-species samples using their custom database containing 54 species at proportions of ≥ 1%, with ‘few’

false positives and negatives. However, the method was only able to quantify high and low abundance levels of taxa. Lang et al. (2019) also found accurate qualitative results, with a 100% accurate identification rate in all samples, at levels as low as 0.2% of the total mixture. However, their database contained only the species used in their mixtures. Comparatively, our study (although using far fewer species) also had a 100% accurate identification rate of taxa in the samples using the database only containing those samples. The study found significantly and highly correlated sequencing reads with pollen count proportions (R^2 = 86.7%), on par with our quantitative results.

Database selection and limitations

A comprehensive discussion detailing the current limitations of database availability exists in Bell et al. (2021) under the section "4.3 Present feasibility of WGS and future research direction". The main points are that the availability of whole genome or plastid references required for the WGS method used in their paper (and for the RefSeq database used here) are far below that of the number of ITS2 and *rbcL* sequences available. Further, without many upgrades to currently available sequences, this method will remain limited, and researchers may be forced to create their own references which is time consuming and costly. A workaround may be a bioinformatical method for combining data from multiple barcodes into a single analysis, which could utilize the vast quantity of single barcode references already available.

Applications and Conclusion

We have demonstrated that a hybrid capture approach with high throughput sequencing is an appropriate method for metabarcoding pollen mixes. The strength of using hybrid capture lies in the ability to target multiple genomic regions, potentially utilising more informative loci without prior knowledge about the target taxa. Yet, it remains that there is still no applicable method to combine multiple barcodes in a single analysis, so using a RefSeq chloroplast library generated better results than a single *matK* barcode library. However, there are far fewer plastid sequences available compared with barcode sequences, and missing taxa in the database could lead to issues with downstream quantification. Conversely, when the taxa present were known and the database restricted to just those present, the *matK* barcode library resulted in relatively accurate and highly correlated sequence proportions compared with input pollen proportions. This method could be applied to pollinator-collected pollen samples, but care should be taken with reference choice and database curation, particularly when extracting quantitative information.

Aknowledgments

This project was supported by AgriFutures Australia, through funding from the Australian Government Department of Agriculture, Water and the Environment as part of its Rural R&D for Profit program. The funding body was not involved in the undertaking of the study or it's publication. Thank you to Joshua Grist and Leif Tenzin for their assistance with lab work. Thanks to Scott Groom and the Gentle R group at the University of Adelaide for their advice with analyses. The authors declare no conflicts of interest.

References

Alotaibi, S. S., Sayed, S. M., Alosaimi, M., Alharthi, R., Banjar, A., Abdulqader, N., & Alhamed, R. (2020). Pollen molecular biology: applications in the forensic palynology and future prospects: a review. *Saudi journal of biological sciences, 27*(5), 1185-1190. doi:10.1016/j.sjbs.2020.02.019

Arstingstall, K. A., DeBano, S. J., Li, X., Wooster, D. E., Rowland, M. M., Burrows, S., & Frost, K. (2021). Capabilities and limitations of using DNA metabarcoding to study plant-pollinator interactions. *Molecular Ecology, 30*(20), 5266-5297. doi:10.1111/mec.16112

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Bayly, M. J., Rigault, P., Spokevicius, A., Ladiges, P. Y., Ades, P. K., Anderson, C., Bossinger, G., Merchant, A., Udovicic, F., Woodrow, I. E., & Tibbits, J. (2013). Chloroplast genome analysis of Australian eucalypts - *Eucalyptus*, *Corymbia*, *Angophora*, *Allosyncarpia* and *Stockwellia* (Myrtaceae). *Molecular Phylogenetics and Evolution, 69*(3), 704-716. doi:10.1016/j.ympev.2013.07.006

Bell, K. L., Burgess, K. S., Botsch, J. C., Dobbs, E. K., Read, T. D., & Brosi, B. J. (2019). Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. *Molecular Ecology, 28*(2), 431-455. doi:10.1111/mec.14840

Bell, K. L., de Vere, N., Keller, A., Richardson, R. T., Gous, A., Burgess, K. S., & Brosi, B. J. (2016). Pollen DNA barcoding: current applications and future prospects. *Genome, 59*(9), 629-640. doi:10.1139/gen-2015-0200

Bell, K. L., Fowler, J., Burgess, K. S., Dobbs, E. K., Gruenewald, D., Lawley, B., Morozumi, C., & Brosi, B. J. (2017). Applying pollen DNA metabarcoding to the study of plant-pollinator interactions. *Applications in Plant Sciences, 5*(6). doi:10.3732/apps.1600124

Bell, K. L., Petit, R. A., 3rd, Cutler, A., Dobbs, E. K., Macpherson, J. M., Read, T. D., Burgess, K. S., & Brosi, B. J. (2021). Comparing whole-genome shotgun sequencing and DNA metabarcoding approaches for species identification and quantification of pollen species mixtures. *Ecology and Evolution, 11*(22), 16082-16098. doi:10.1002/ece3.8281

Bushnell, B. (2021). BBMap short read aligner, and other bioinformatic tools. Retrieved from sourceforge.net/projects/bbmap/

CBOL Plant Working Group. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences USA, 106*(31), 12794-12797. doi:10.1073/pnas.0905845106

Clarke, C. L., Alsos, I. G., Edwards, M. E., Paus, A., Gielly, L., Haflidason, H., Mangerud, J., Regnéll, C., Hughes, P. D., & Svendsen, J. I. (2020). A 24,000-year ancient DNA and pollen record from the Polar Urals reveals temporal dynamics of arctic and boreal plant communities. *Quaternary Science Reviews, 247*, 106564.

de Vere, N., Jones, L. E., Gilmore, T., Moscrop, J., Lowe, A., Smith, D., Hegarty, M. J., Creer, S., & Ford, C. R. (2017). Using DNA metabarcoding to investigate honey bee foraging reveals

limited flower use despite high floral availability. *Scientific Reports, 7*(1), 42838. doi:10.1038/srep42838

Dormontt, E. E., van Dijk, K.-j., Bell, K. L., Biffin, E., Breed, M. F., Byrne, M., Caddy-Retalic, S., Encinas-Viso, F., Nevill, P. G., Shapcott, A., Young, J. M., Waycott, M., & Lowe, A. J. (2018). Advancing DNA barcoding and metabarcoding applications for plants requires systematic analysis of herbarium collections—an Australian perspective. *Frontiers in Ecology and Evolution, 6*. doi:10.3389/fevo.2018.00134

Foster, N. R., van Dijk, K.-j., Biffin, E., Young, J. M., Thomson, V. A., Gillanders, B. M., Jones, A. R., & Waycott, M. (2021). A multi-gene region targeted capture approach to detect plant DNA in environmental samples: a case study from coastal environments. *Frontiers in Ecology and Evolution, 9*. doi:10.3389/fevo.2021.735744

Gao, T., Yao, H., Song, J., Zhu, Y., Liu, C., & Chen, S. (2010). Evaluating the feasibility of using candidate DNA barcodes in discriminating species of the large Asteraceae family. *BMC Evolutionary Biology, 10*, 324. doi:10.1186/1471-2148-10-324

Golczyk, H., Greiner, S., Wanner, G., Weihe, A., Bock, R., Borner, T., & Herrmann, R. G. (2014). Chloroplast DNA in mature and senescing leaves: a reappraisal. *Plant Cell, 26*(3), 847-854. doi:10.1105/tpc.113.117465

Grüter, C., & Ratnieks, F. L. W. (2011). Honeybee foragers increase the use of waggle dance information when private information becomes unrewarding. *Animal Behaviour, 81*(5), 949-954. doi:10.1016/j.anbehav.2011.01.014

Keller, A., Danner, N., Grimmer, G., Ankenbrand, M., von der Ohe, K., von der Ohe, W., Rost, S., Hartel, S., & Steffan-Dewenter, I. (2015). Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. *Plant Biology (Stuttgart), 17*(2), 558-566. doi:10.1111/plb.12251

Kraaijeveld, K., de Weger, L. A., Ventayol Garcia, 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. *Molecular Ecology Resources, 15*(1), 8-16. doi:10.1111/1755-0998.12288

Krehenwinkel, H., Fong, M., Kennedy, S., Huang, E. G., Noriyuki, S., Cayetano, L., & Gillespie, R. (2018). The effect of DNA degradation bias in passive sampling devices on metabarcoding studies of arthropod communities and their associated microbiota. *PLoS ONE, 13*(1), e0189188. doi:10.1371/journal.pone.0189188

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports, 7*(1), 17668. doi:10.1038/s41598-017-17333-x

Kress, W. J. e., & Erickson, D. L. e. (2012). DNA barcodes: methods and protocols. In W. J. Kress & D. L. Erickson (Eds.), *Methods in Molecular Biology*: Humana Press.

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How quantitative is metabarcoding: a meta-analytical approach. *Molecular Ecology, 28*(2), 420-430. doi:10.1111/mec.14920

615 Lang, D., Tang, M., Hu, J., & Zhou, X. (2019). Genome-skimming provides accurate quantification for
616 pollen mixtures. *Molecular Ecology Resources, 19*(6), 1433-1446. doi:10.1111/1755-
617 0998.13061

618 Liu, M., Clarke, L. J., Baker, S. C., Jordan, G. J., & Burridge, C. P. (2019). A practical guide to DNA
619 metabarcoding for entomological ecologists. *Ecological Entomology, 45*(3), 373-385.
620 doi:10.1111/een.12831

621 Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: estimating species abundance
622 in metagenomics data. *PeerJ Computer Science, 3*(e104). doi:10.7717/peerj-cs.104

623 Morley, S. A., & Nielsen, B. L. (2016). Chloroplast DNA copy number changes during plant
624 development in organelle DNA polymerase mutants. *Frontiers in Plant Science, 7*, 57.
625 doi:10.3389/fpls.2016.00057

626 Nge, F. J., Biffin, E., Thiele, K. R., & Waycott, M. (2021). Reticulate evolution, ancient chloroplast
627 haplotypes, and rapid radiation of the Australian plant genus *Adenanthos* (Proteaceae).
628 *Frontiers in Ecology and Evolution, 8*(492). doi:10.3389/fevo.2020.616741

629 Pawluczyk, M., Weiss, J., Links, M. G., Egana Aranguren, M., Wilkinson, M. D., & Egea-Cortines, M.
630 (2015). Quantitative evaluation of bias in PCR amplification and next-generation sequencing
631 derived from metabarcoding samples. *Analytical and Bioanalytical Chemistry, 407*(7), 1841-
632 1848. doi:10.1007/s00216-014-8435-y

633 Peel, N., Dicks, L. V., Clark, M. D., Heavens, D., Percival-Alwyn, L., Cooper, C., Davies, R. G., Leggett,
634 R. M., Yu, D. W., & Freckleton, R. (2019). Semi-quantitative characterisation of mixed pollen
635 samples using MinION sequencing and Reverse Metagenomics (RevMet). *Methods in Ecology*
636 *and Evolution, 10*(10), 1690-1701. doi:10.1111/2041-210x.13265

637 Pornon, A., Escaravage, N., Burrus, M., Holota, H., Khimoun, A., Mariette, J., Pellizzari, C., Iribar, A.,
638 Etienne, R., Taberlet, P., Vidal, M., Winterton, P., Zinger, L., & Andalo, C. (2016). Using
639 metabarcoding to reveal and quantify plant-pollinator interactions. *Scientific Reports, 6*,
640 27282. doi:10.1038/srep27282

641 Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The Barcode of Life Data System
642 (http://www.barcodinglife.org). *Molecular Ecology Notes, 7*(3), 355-364. doi:10.1111/j.1471-
643 8286.2007.01678.x

644 Richardson, R. T., Bengtsson-Palme, J., & Johnson, R. M. (2017). Evaluating and optimizing the
645 performance of software commonly used for the taxonomic classification of DNA
646 metabarcoding sequence data. *Molecular Ecology Resources, 17*(4), 760-769.
647 doi:10.1111/1755-0998.12628

648 Richardson, R. T., Lin, C. H., Quijia, J. O., Riusech, N. S., Goodell, K., & Johnson, R. M. (2015). Rank-
649 based characterization of pollen assemblages collected by honey bees using a multi-locus
650 metabarcoding approach. *Applications in Plant Sciences, 3*(11). doi:10.3732/apps.1500043

651 Richardson, R. T., Lin, C. H., Sponsler, D. B., Quijia, J. O., Goodell, K., & Johnson, R. M. (2015).
652 Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey
653 bees in an agroecosystem. *Applications in Plant Sciences, 3*(1). doi:10.3732/apps.1400066

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research, 22*(5), 939-946. doi:10.1101/gr.128124.111

RStudio Team. (2020). RStudio: integrated development for R. RStudio. Boston, MA: PBC. Retrieved from http://www.rstudio.com/

Sabre-barcode-demultiplexing. Retrieved from https://github.com/najoshi/sabre.

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes, 9*, 88. doi:10.1186/s13104-016-1900-2

Schulte, L., Bernhardt, N., Stoof-Leichsenring, K., Zimmermann, H. H., Pestryakova, L. A., Epp, L. S., & Herzschuh, U. (2021). Hybridization capture of larch (Larix Mill.) chloroplast genomes from sedimentary ancient DNA reveals past changes of Siberian forest. *Molecu.ar Ecology Resources, 21*(3), 801-815. doi:10.1111/1755-0998.13311

Sickel, W., Ankenbrand, M. J., Grimmer, G., Holzschuh, A., Hartel, S., Lanzen, J., Steffan-Dewenter, I., & Keller, A. (2015). Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecology, 15*, 20. doi:10.1186/s12898-015-0051-y

Smart, M., Cornman, R. S., Iwanowicz, D. D., McDermott-Kubeczko, M., Pettis, J. S., Spivak, M. S., & Otto, C. R. (2017). A comparison of honey bee-collected pollen from working agricultural lands using light microscopy and ITS metabarcoding. *Environmental Entomology, 46*(1), 38-49.

Suchan, T., Talavera, G., Saez, L., Ronikier, M., & Vila, R. (2019). Pollen metabarcoding as a tool for tracking long-distance insect migrations. *Molecular Ecology Resources, 19*(1), 149-162. doi:10.1111/1755-0998.12948

Synge, A. D. (1947). Pollen Collection by Honeybees (*Apis mellifera*). *Journal of Animal Ecology, 16*(2), 122-138.

Visscher, P. K., & Seeley, T. D. (1982). Foraging strategy of honeybee colonies in a temperate deciduous forest. *Ecology, 63*(6). doi:10.2307/1940121

Waycott, M., van Dijk, K.-j., & Biffin, E. (2021). A hybrid capture RNA bait set for resolving genetic and evolutionary relationships in angiosperms from deep phylogeny to intraspecific lineage hybridization. doi:10.1101/2021.09.06.456727

Weber, R. W. (1998). Pollen identification. *Annals of Allergy, Asthma & Immunology, 80*(2), 141-145; quiz 146-147. doi:10.1016/S1081-1206(10)62947-X

Wilson, E. E., Sidhu, C. S., LeVan, K. E., & Holway, D. A. (2010). Pollen foraging behaviour of solitary Hawaiian bees revealed through molecular pollen analysis. *Molecular Ecology, 19*(21), 4823-4829. doi:10.1111/j.1365-294X.2010.04849.x

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology, 20*(1), 257. doi:10.1186/s13059-019-1891-0

690 Zinger, L., Bonin, A., Alsos, I. G., Balint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E.,
691 Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli,
692 L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., Kauserud, H., Orlando, L., Pansu, J.,
693 Pawlowski, J., Tedersoo, L., Thomsen, P. F., Willerslev, E., & Taberlet, P. (2019). DNA
694 metabarcoding-need for robust experimental designs to draw sound ecological conclusions.
695 *Molecular Ecology, 28*(8), 1857-1862. doi:10.1111/mec.15060

Data Accessibility Statement

All sequence data will be made publicly available on the SRA upon manuscript acceptance, before publication.

Benefit-Sharing

Benefits from this research accrue from the sharing of our data and results on public databases as described above.

Author contributions

DK, KvD, AJL, KH, KB designed the experiments; DK, AM undertook the laboratory work; DK, SC analysed the data; AJL, KvD, KH supervised the project, and acquired funding; DK wrote the first draft of the manuscript; all authors contributed substantially to revisions.

Tables and Figures

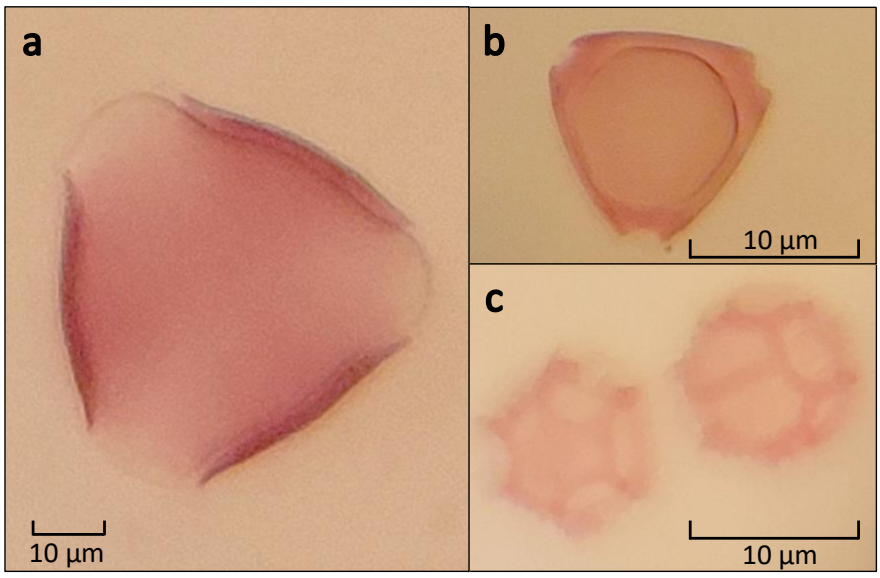


Figure 1. Images of the taxa of pollen used in artificial mixtures. **a)** *Prunus dulcis*, **b)** *Eucalyptus baxteri*, **c)** *Arctotheca calendula.* Photographs were taken from slides under a compound microscope by Leif Currie.

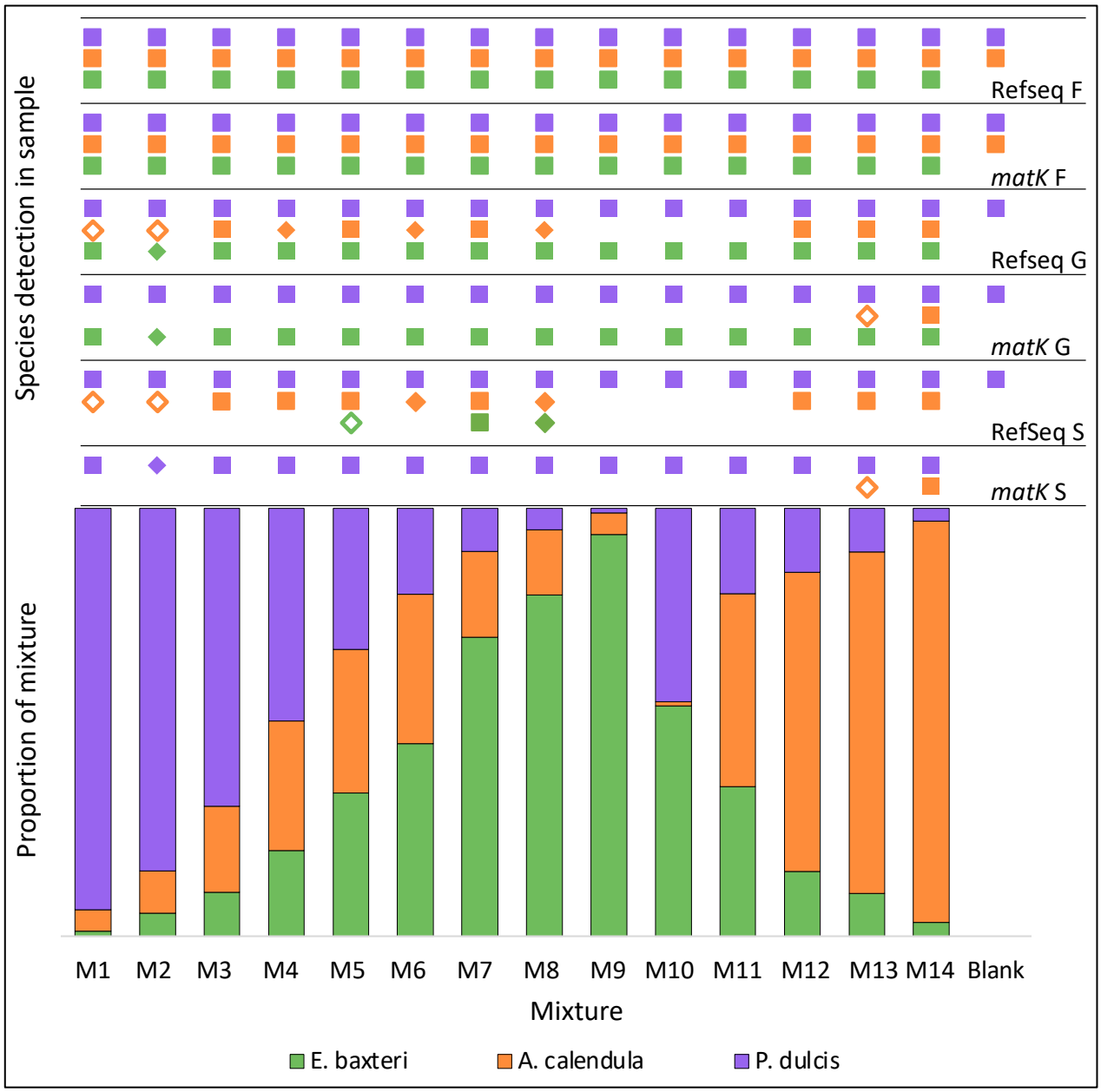


Figure 2. Stacked bar plot of the relative input proportions by weight of three pollen taxa (*Prunus dulcis*, *Arctotheca calendula* and *Eucalyptus baxteri*) in artificially constructed mixtures (M1 - M14), and negative control (Blank). Symbols above each bar indicate whether each taxon was detected in the mixture using metabarcoding with either *matK* or RefSeq databases, identified to family (F), genus (G) and species (S) levels. Solid squares indicate the taxon was detected in three mixture replicates, solid diamonds indicate detection in two of the three replicates, and hollow diamonds indicate detection in only one replicate.

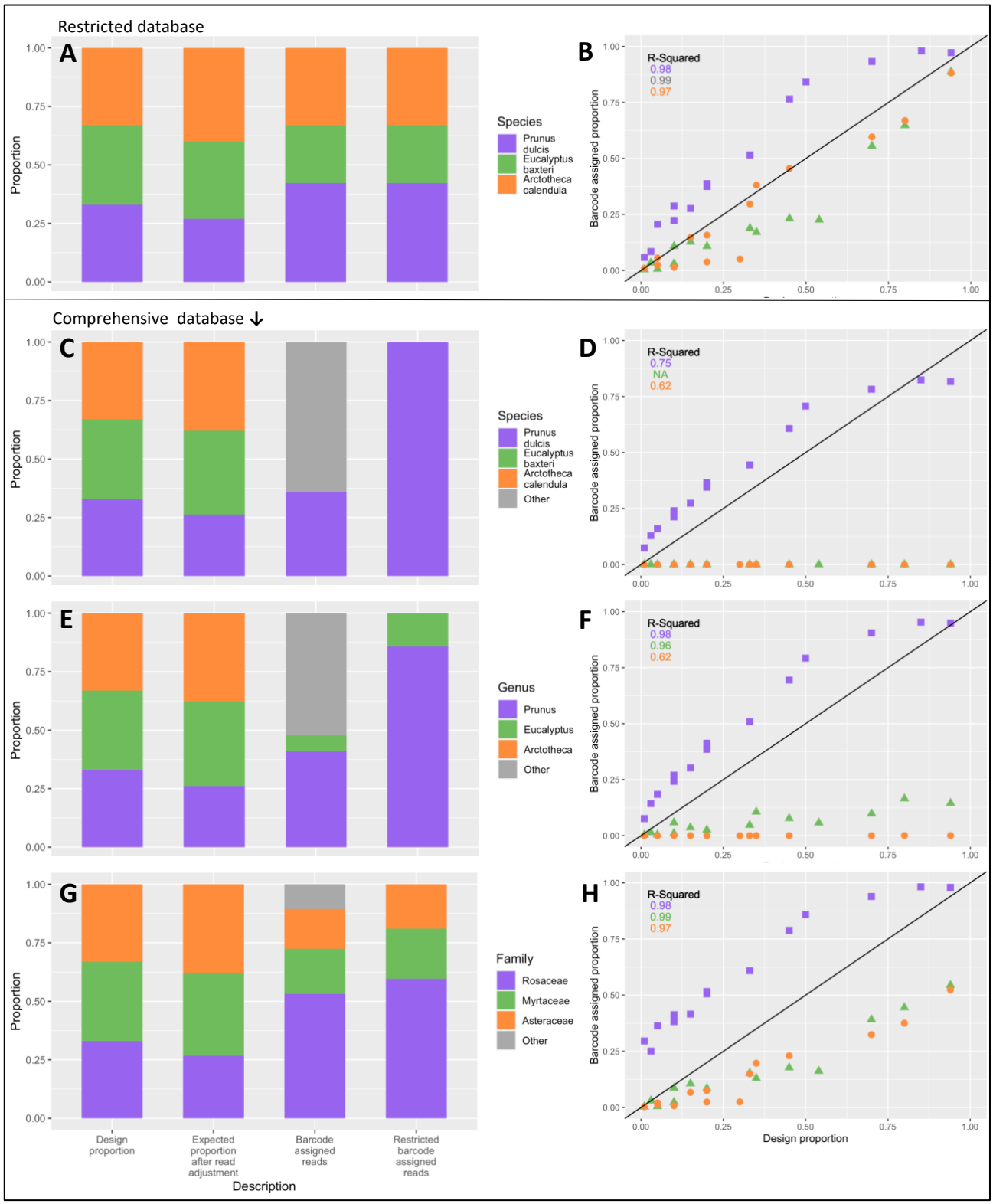
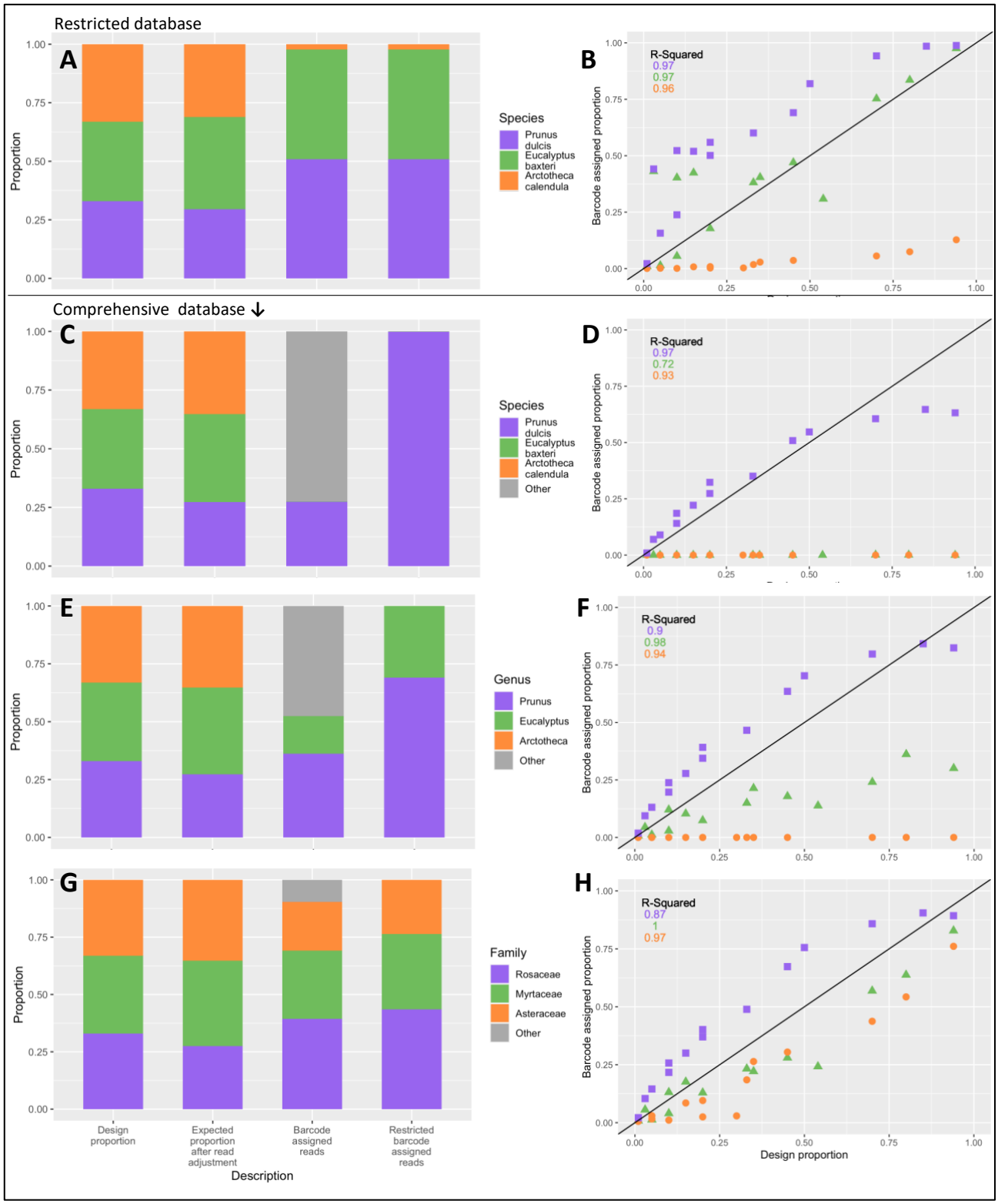


Figure 3. Plots depicting taxonomic assignment in pollen mixtures using a ***matK*** reference database. **A-B:** taxonomic assignment to **species** level made using a **restricted** *matK* database (only containing three taxa used in mixtures); **C-D:** assignment to **species** level using a comprehensive *matK* database; **E-F:** assignment to **genus** level using a comprehensive *matK* database; **G-H:** assignment to **family** level using a comprehensive *matK* database. ***Left side:*** Summary of taxon proportions averaged

across samples. Columns from left to right are: 1) original design proportion according to pollen weight, 2) expected proportion after read correction (given the 14 mixtures had different numbers of reads per taxon), 3) total barcode assigned reads, 4) barcode assigned reads with 'other' (non-target) taxa excluded. ***Right side*:** Sequence proportions versus input (design) proportions.



730 **Figure 4.** Plots depicting taxonomic assignment in pollen mixtures using a **RefSeq** reference
731 database. **A-B:** assignment to **species** level made using a **restricted** database (only containing three
732 taxa used in mixtures); **C-D:** assignment to **species** level using a **comprehensive** database; **E-F:**
733 assignment to **genus** level using a comprehensive RefSeq database; **G-H:** assignment to **family** level
734 using a comprehensive database. ***Left side:*** Summary of taxon proportions averaged across samples.
735 Columns from left to right are: 1) original design proportion according to pollen weight, 2) expected
736 proportion after read correction (given the 14 mixtures had different numbers of reads per taxon), 3)
737 total barcode assigned reads, 4) barcode assigned reads with 'other' (non-target) taxa excluded.
738 ***Right side*:** Sequence proportions versus input (design) proportions of pollen.

739 **Table 1.** Mixed model with binomial distribution to determine if starting pollen proportion affected
740 the success or failure of taxonomic identification to three taxonomic levels in pollen mixtures.

Barcode db	Taxonomic level	Mix taxa	Est.	S.E.	Z-val	P-val
matK restricted	Species	*E. baxteri*	14.42	19.72	0.73	0.46
		A. calendula	-7.20	6.81	-1.06	0.29
		P. dulcis	Response is constant			
matK wide	Species	*E. baxteri*	14.43	19.72	0.73	0.46
		A. calendula	-11.20	35.72	-0.31	0.75
		P. dulcis	-7.13	6.83	-1.04	0.30
	Genus	*E. baxteri*	14.30	19.91	0.79	0.47
		A. calendula	-11.20	35.72	-0.31	0.75
		P. dulcis	-7.13	6.83	-1.04	0.30
	Family	*E. baxteri*	14.43	19.72	0.73	0.46
		A. calendula	-7.20	6.82	-1.06	0.29
		P. dulcis	-7.13	6.83	-1.04	0.30
RefSeq restricted	Species	*E. baxteri*	Response is constant			
		A. calendula				
		P. dulcis				
RefSeq wide	Species	*E. baxteri*	9.03	6.85	1.32	0.19
		A. calendula	-1.98	2.51	-0.79	0.43
		P. dulcis	Response is constant			
	Genus	*E. baxteri*	14.43	19.72	0.73	0.46
		A. calendula	-2.03	2.58	-0.78	0.43
		P. dulcis	Response is constant			
	Family	*E. baxteri*	Response is constant			
		A. calendula				
		P. dulcis				

741
742