Rank-Based Characterization of Pollen Assemblages Collected by Honey Bees Using a Multi-Locus Metabarcoding Approach

Authors: Richardson, Rodney T., Lin, Chia-Hua, Quijia, Juan O., Riusech, Natalia S., Goodell, Karen, et al.

Source: Applications in Plant Sciences, 3(11)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1500043

Applications in Plant Sciences 2015 3(11): 1500043

Protocol Note

Rank-based characterization of pollen assemblages collected by honey bees using a multi-locus metabarcoding approach[1]

Rodney T. Richardson[2,4], Chia-Hua Lin[2], Juan O. Quijia[2], Natalia S. Riusech[2], Karen Goodell[3], and Reed M. Johnson[2]

[2]Department of Entomology, The Ohio State University—Ohio Agricultural Research and Development Center, 1680 Madison Avenue, Wooster, Ohio 44691 USA; and [3]Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 1179 University Drive, Newark, Ohio 43023 USA

- *Premise of the study:* Difficulties inherent in microscopic pollen identification have resulted in limited implementation for large-scale studies. Metabarcoding, a relatively novel approach, could make pollen analysis less onerous; however, improved understanding of the quantitative capacity of various plant metabarcode regions and primer sets is needed to ensure that such applications are accurate and precise.
- *Methods and Results:* We applied metabarcoding, targeting the ITS2, *matK*, and *rbcL* loci, to characterize six samples of pollen collected by honey bees, *Apis mellifera*. Additionally, samples were analyzed by light microscopy. We found significant rank-based associations between the relative abundance of pollen types within our samples as inferred by the two methods.
- *Conclusions:* Our findings suggest metabarcoding data from plastid loci, as opposed to the ribosomal locus, are more reliable for quantitative characterization of pollen assemblages. Furthermore, multilocus metabarcoding of pollen may be more reliable than single-locus analyses, underscoring the need for discovering novel barcodes and barcode combinations optimized for molecular palynology.

Key words: *Apis mellifera*; *Fraxinus*; *matK*; molecular palynology; pollen plastid biology; *rbcL*.

Quantitative identification of pollen by taxonomic origin is important for applications in pollination biology and conservation (Kearns and Inouye, 1993; Wilson et al., 2010; Forcone et al., 2011; Girard et al., 2012; Cusser and Goodell, 2013), authentication of apicultural products (Louveaux et al., 1978; Jones and Bryant, 1992; Dimou and Thrasyvoulou, 2007), and allergy-related airborne pollen monitoring (Longhi et al., 2009; Kraaijeveld et al., 2015). Traditionally, pollen analysis has been accomplished using microscopic palynology, a technique involving the discrimination of pollen types by morphology (Erdtman, 1943). Due to the expertise required and difficulties associated with accurately distinguishing and identifying pollen from morphologically similar taxa, this technique has been difficult to implement on a large scale. Thus, the development and improvement of novel techniques for pollen analysis is an area of current interest (Keller et al., 2015; Kraaijeveld et al., 2015; Richardson et al., 2015).

[1]Manuscript received 16 April 2015; revision accepted 13 September 2015.

The authors thank the apiary managers for site access, M. E. Hernandez-Gonzalez and the Ohio Agricultural Research and Development Center Molecular and Cellular Imaging Center staff for technical support, and D. B. Sponsler and R. A. Klips for helpful manuscript review and botanical advice, respectively. This study was funded by a Pollinator Partnership Corn Dust Research Consortium grant to R.M.J. and an Ohio State University–Newark Scholarly Activity Grant to K.G. and was supported by an allocation of computing time from the Ohio Supercomputer Center.

[4]Author for correspondence: richardson.827@osu.edu

doi:10.3732/apps.1500043

The application of DNA barcoding to pollen analysis displays promise as an efficient and reliable approach. Similar to comparing morphological features of unknown pollen to those of reference pollen from voucher specimens, DNA sequences of unknown origin can be compared to sequences from voucher specimens. Initial applications of this approach to pollen analysis have used capillary sequencing technology (Longhi et al., 2009; Wilson et al., 2010; Galimberti et al., 2014); however, advances in the accuracy and length capabilities of next-generation sequencing provide researchers a practical, high-throughput alternative (Keller et al., 2015; Kraaijeveld et al., 2015; Richardson et al., 2015).

Recently, next-generation sequencing was used to characterize the botanical origins of bee-collected pollen using the ribosomal intergenic ITS2 locus (Richardson et al., 2015). This target locus was chosen because previous studies suggested that plastids are rarely incorporated into pollen (Reboud and Zeyl, 1994; Mogensen, 1996; Azhagiri and Maliga, 2007). However, evidence from more recent studies suggests that pollen plastids may be common (Tang et al., 2009), enabling pollen metabarcoding of plastid loci (Galimberti et al., 2014; Kraaijeveld et al., 2015). Although the approach using ITS2 was successful in identifying pollen (Richardson et al., 2015), it suffered from two limitations: (1) the method failed to detect certain prominent taxa identified microscopically and (2) while the method generated a useful taxonomic list, the relative abundance of different pollen types could not be inferred from the sequence data. Here, we present an improvement in pollen metabarcoding by targeting the plastid loci *matK* and *rbcL*, in addition to

the ribosomal ITS2 locus, to characterize polyfloral samples of pollen collected by honey bees. In addition, we compare our metabarcoding results with results from microscopic analysis to evaluate the range of taxa detected and the capacity for quantitative inference of rank order pollen type abundance using a multilocus metabarcoding approach.

METHODS AND RESULTS

***Sample collection and homogenization*—**During spring 2014, bee-collected pollen samples were collected at six apiaries, all greater than 15 km apart, in west-central Ohio. The latitude and longitude of each apiary is provided in Appendix 1, and apiaries are herein denoted as A, B, C, D, E, and F. Using Sundance I bottom-mounted pollen traps (Ross Rounds, Albany, New York, USA), we collected four samples from each site from 5–11 May, sampling every other day. After collection, samples were pooled by site before homogenization. A 10% subsample (by weight) was taken from each pooled sample, mixed in 50% ethanol, and stirred for 25 min using a magnetic stir plate. Using Buchner funnel vacuum filtration (Whatman grade 1; Sigma-Aldrich, St. Louis, Missouri, USA), we separated the homogenized pollen from the solvent and transferred it to a flow hood to air dry at room temperature.

***Pollen identification and quantification by microscopy*—**We mixed 100 mg of the dried, homogenized pollen sample from each site in 0.5 mL of water and mounted five separate smears onto microscope slides in basic fuchsin jelly (Kearns and Inouye, 1993). We then counted and identified approximately 1000 pollen grains per slide for each pooled sample under a compound microscope at 400–1000× magnification. The voucher specimens used for pollen identification are listed in Richardson et al. (2015). A total of approximately 5000 grains were analyzed per sample. Due to the difficulty in distinguishing some related plant taxa (e.g., within Rosaceae [Moore et al., 1991]), we chose to limit microscopic identification to the family level. The total number of grains of pollen from each plant family, summed from each of the five slides, is available in Appendix S1.

***Pollen identification by metabarcoding*—**After drying our homogenized samples, we freed DNA from 50 mg of pollen per sample using bead-beater pulverization (Mini-BeadBeater-1; BioSpec Products, Bartlesville, Oklahoma, USA) (Simel et al., 1997). Each sample was placed in a 2.0-mL microcentrifuge tube with 600 µL of lysis buffer from the QIAGEN DNeasy Plant Mini Kit (QIAGEN, Venlo, Limburg, Netherlands). Zirconium/silica beads (0.5 mm diameter) were added until the total contents of each tube reached 1.5 mL, and the sample was pulverized for 2 min. Then, 300 µL of deionized water was transferred to each tube and mixed with the contents and a 300-µL portion of the resulting lysate mix was transferred to a sterile 1.5-mL microcentrifuge tube. DNA was extracted using the QIAGEN DNeasy Plant Mini Kit (QIAGEN), and the ribosomal ITS2 and plastid *matK* and *rbcL* loci were amplified in separate PCR reactions. Amplification was conducted using previously published primer sets (Fay et al., 1997; Cuénoud et al., 2002; Chen et al., 2010) and the Phusion High-Fidelity PCR Kit (New England Biolabs, Ipswich, Massachusetts, USA) in a Mastercycler ep Gradient PCR machine (Eppendorf AG, Hamburg, Germany). Primer sequences, reagents, and PCR conditions for each barcoding locus are presented in Appendix 2. The ITS2, *matK*, and *rbcL* amplicons were subsequently purified using the PureLink PCR Purification kit (Life Technologies, Carlsbad, California, USA). At this point, 500 ng of purified PCR product for each locus was indexed independently using the NEBNext Ultra DNA Library Prep Kit for Illumina and NEBNext Multiplex Oligos for Illumina (New England Biolabs). Multiplexed samples were purified before being pooled (Agencourt AMPure XP; Beckman Coulter, Brea, California, USA). A final nine-cycle library amplification step was performed and samples were analyzed on a Qubit 2.0 fluorometer (Life Technologies) and an Agilent 2100 Bioanalyzer (DNA 1000 kit; Agilent Technologies, Santa Clara, California, USA) to ensure sample quality before sequencing. Paired-end sequencing was performed with the Illumina MiSeq platform using the TruSeq LT assay (600 cycles). Sequence data are available from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (accession code SRP055937).

Sequences were analyzed using an alignment-based approach. All computation was performed at the Ohio Supercomputer Center on a 12-core HP Intel Xeon X5650 machine with 48 GB of RAM. Reads were first trimmed by quality using Trimmomatic (v0.32; Bolger et al., 2014) with Phred scale 33 quality thresholds of 20 for both the 5′ and 3′ ends of each read. Reads less than 50 bp in length were discarded. Reads were then dereplicated to minimize PCR amplification bias and converted to FASTA format using the FASTX-Toolkit (version 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/). Next, reads were aligned against reference ITS2, *matK*, and *rbcL* plant sequences downloaded from NCBI GenBank on 23 September 2014. Reference libraries were constrained to only include plant species known to be present in Ohio and surrounding states based on the USDA Plants Database (http://plants.usda.gov/). Reference libraries are available in FASTA format in Appendices S2, S3, and S4. Venn diagrams showing the completeness of each of the reference libraries, at both the genus and species level, are presented in Appendix S5. Complete lists of the genera and species represented in each reference library are presented in Appendix S6. Alignment was performed using the BLASTN algorithm (version 2.2.29+; Altschul et al., 1997). Alignment quality-control thresholds were set as follows: *E*-value cutoff 1e-150, number of alignments 1, output format 0, number of descriptions 1. An additional setting, percent identity threshold, was used and its value differed between loci. For ITS2, we used a percent identity threshold of 95%, as in Richardson et al. (2015). However, given the relatively low sequence divergence between species at the *matK* and *rbcL* loci, we used a stringent setting of 99% identity. Following BLAST, we used MEGAN 5 (version 5.1.5; Huson et al., 2011) to taxonomically summarize our results with the following settings: min support 1, min score 50.0, max expected 1e-150, top percent 100.0, min complexity 0.00, min support percent 0.0 (off), paired end mode. Complete family-level and genus-level metabarcoding results are summarized in Appendix S7 and Appendix S8, respectively.

***Analysis of results*—**After sequencing and implementing quality control, we obtained from 78,975 to 224,428 forward reads and 134,133 to 557,713 reverse reads across all 18 amplicon libraries. The median number of reads per locus was 258,987, 194,856, and 134,183 for ITS2, *rbcL*, and *matK*, respectively. In total, these reads had best hits to plant species from 49 families. To limit the potential for false identification, we limited our analysis using a consensus-based approach, counting only families found in more than one of the three amplicon libraries for each sample. Consensus lists of the families detected and their relative abundance in each sample are provided in Appendix S9. Using this approach, we confidently detected 25 plant families across the six sites. Using microscopy, 25 plant families were identified, six of which (Asparagaceae, Elaeagnaceae, Hamamelidaceae, Lamiaceae, Magnoliaceae, and Poaceae) were not identified by the metabarcoding consensus analysis. Although these families were detected microscopically, they were present at very low abundance, never constituting more than 0.5% of the 5000 counted grains in any sample.

To test the ability to infer the rank order abundance of different pollen types from the metabarcoding data, we conducted Spearman's rank-based correlation between the number of mate-paired read alignments and the number of pollen grains per plant family for each locus individually as well as for the mean of the *rbcL* and *matK* loci, excluding the ITS2 data. We chose to exclude ITS2 because data from this locus exhibited poor quantitative capacity in a prior study (Richardson et al., 2015). Lastly, we calculated *R* coefficients for families detected across at least five of the six samples to determine which families were over- or under-represented in the metabarcoding analysis relative to microscopic analysis. The *R* coefficient is used in authenticating honey provenance (Bryant and Jones, 2001). In the context of this paper, the *R* coefficient is the quotient, for a particular taxon, of the relative abundance as inferred by metabarcoding and the relative abundance as inferred by microscopy. We conducted this analysis on *rbcL* data because this locus exhibited a broad scope of detection and was the only single locus to produce significant rank-based correlations when compared to the microscopy data.

Pollen from the families Rosaceae (commonly species of *Malus* Mill., *Crataegus* L., *Amelanchier* Medik., *Prunus* L., and other cultivated relatives) and Salicaceae (predominantly *Salix* L. spp.) comprised over 65% of our samples (Fig. 1). Pollen from plants in the Asteraceae (*Taraxacum officinale* F. H. Wigg.) and Oleaceae (*Fraxinus* L. spp.) were also abundant. Using Spearman's rank-based correlation, we found moderate to strong associations between the rank order abundance of pollen types within our samples as inferred by the molecular and microscopic approaches. For the *rbcL* locus, ρ values ranged from 0.536 to 0.939, and the associations were significant for five out of six samples (Table 1). For the mean of *rbcL* and *matK*, the associations were significant across all samples and ρ values ranged from 0.570 to 0.939 (Fig. 2). When *matK* and ITS2 were analyzed separately, associations between the molecular and microscopic relative abundances were not significant for any sample (Table 1). In our analysis of average *R* coefficients, we found that certain families were consistently over- or under-represented in the molecular results relative to the microscopic results (Table 2). In particular, the average

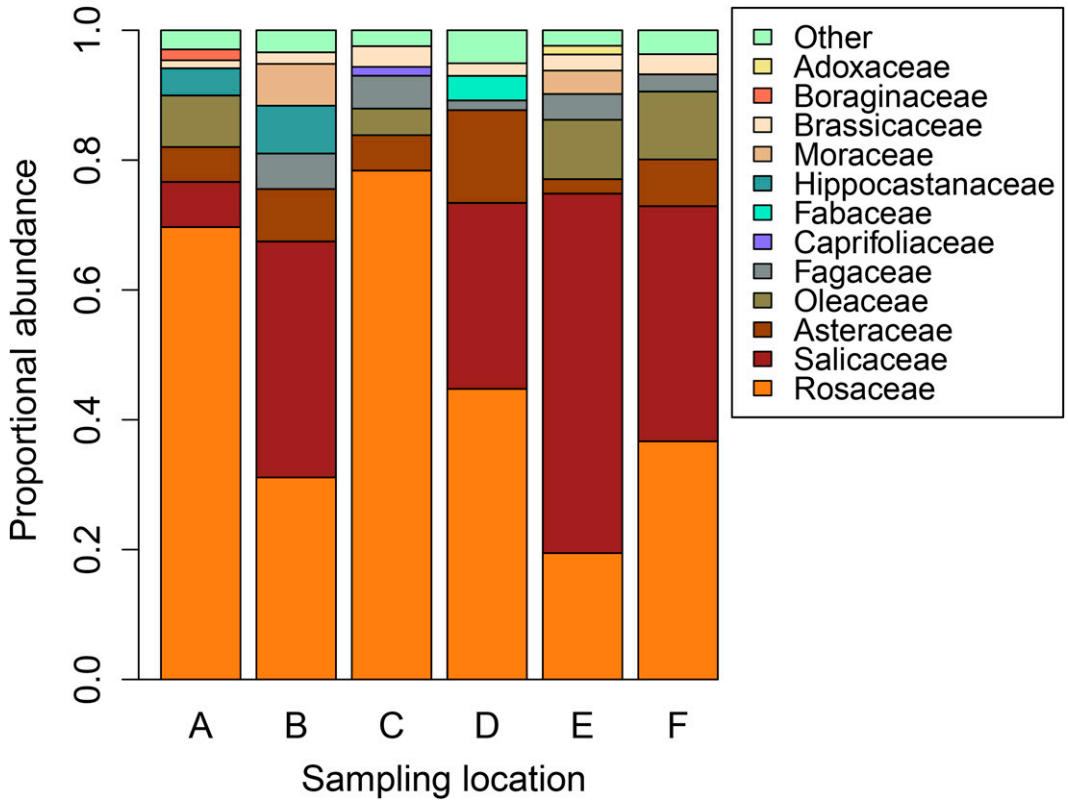

Fig. 1. Proportional taxonomic abundances of each sample as estimated by microscopy. Unidentified pollen grains and taxa present at less than or equal to 1% of the sample are grouped as "Other."

R coefficients for Brassicaceae, Caprifoliaceae, and Salicaceae were under-represented in the molecular data by greater than threefold relative to the microscopy data, while Fabaceae and Fagaceae were over-represented by greater than threefold (Table 2).

CONCLUSIONS

We employed multilocus metabarcoding alongside traditional microscopic pollen identification, with the latter being considered the current standard of practice. Using a consensus-based approach, we found significant rank-based correlations between *rbcL* sequence abundance and microscopically examined pollen grain abundance for five of six samples. However, using the mean of *rbcL* and *matK* sequence abundance, we found significant associations between metabarcoding and microscopic results across all sites. This suggests that while the *rbcL* locus may be quantitatively useful, the simultaneous use of multiple loci may improve quantitative measurement of pollen abundance.

Our multilocus, consensus-based method exhibits promise as a powerful approach to pollen identification using metabarcoding. While no significant associations were found between *matK* sequence abundance and microscopy data, significant associations were found across all samples when *matK* sequence abundance was averaged with *rbcL* abundance. The poor performance of *matK* when used individually may be a result of incomplete universality displayed by the *matK* primer set (Chen et al., 2010). Despite its discriminatory power as a rapidly evolving plastidial coding region (Hilu and Liang, 1997), our data suggest the *matK* primer set used here may not be ideal for characterizing diverse pollen samples and may only be useful for supplementing data from other loci through average- or median-based analyses. Performing such analyses could enable researchers to both broaden the scope of detectable taxa and increase the quantitative capacity of metabarcoding efforts. Using one primer set to coamplify a genetic region across taxonomically diverse samples can be problematic, because priming site sequence divergence may hinder or prevent amplification for some taxa, potentially leading to under-representation or even nondetection in the metabarcoding sequence data. Employing a suite of primers enables researchers to overcome this limitation.

An additional metabarcoding issue involves minimizing the potential for false-positive identifications. Across a diverse sample, it can be expected that some closely related taxa exhibit

Table 1. Spearman's rank-based correlation between the total number of grains per plant taxon as determined by microscopy and the number of mate-paired aligned reads per plant taxon as determined by metabarcoding. Numbers indicate Spearman's ρ, with the *P* value for each test in parentheses.

Sampling location[a]	ITS2	*matK*	*rbcL*
A	0.381 (0.360)	0.469 (0.203)	0.587 (0.049)
B	–0.130 (0.658)	0.515 (0.072)	0.764 (0.001)
C	0.238 (0.582)	0.750 (0.066)	0.939 (<0.001)
D	0.204 (0.504)	0.262 (0.536)	0.575 (0.028)
E	0.067 (0.854)	0.483 (0.194)	0.536 (0.073)
F	–0.005 (0.989)	0.617 (0.086)	0.762 (0.002)

[a] Letters indicate the apiary associated with each sample.

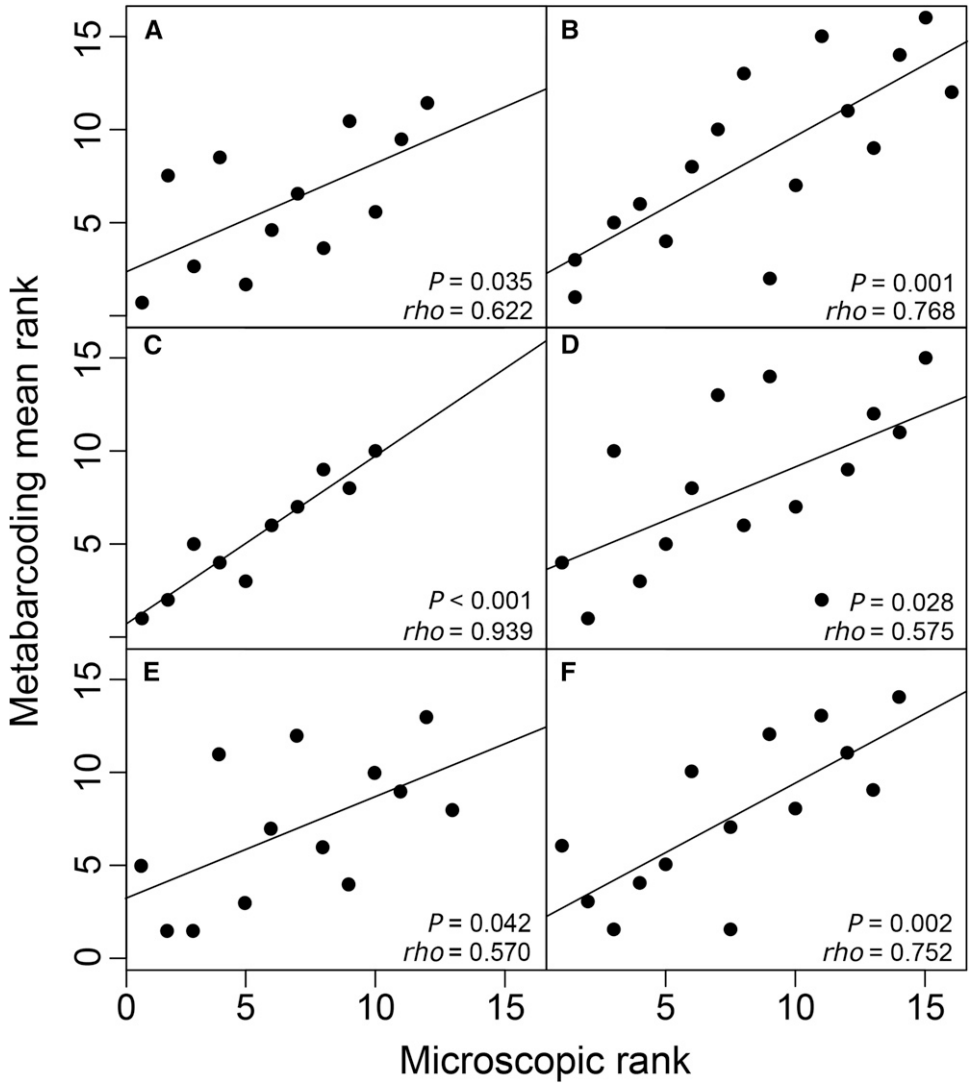

Fig. 2. Rank-transformed taxonomic abundance as estimated by the mean number of *rbcL* and *matK* metabarcoding reads (*y*-axis) and the number of grains estimated microscopically (*x*-axis). Spearman's ρ and *P* values are provided. Part letters indicate the apiary associated with each sample.

little sequence divergence at a particular locus. Employing multiple loci in conjunction with consensus-based analysis limits the potential for false-positive identifications as the probability of the same false-positive identification occurring across multiple independent loci is decreased relative to the probability for a single locus. Lastly, the completeness of the reference database is crucial for the successful application of metabarcoding. Although none of the libraries used here were entirely complete with respect to Ohio taxa, a large majority of the known species were represented (Appendix S5).

Future research into different bioinformatic analyses, such as classifier-based analysis as opposed to alignment-based analysis, is warranted. The current alignment-based approach does not provide confidence estimates for individual sequence to taxon assignments. Classifier-based approaches are commonly used in microbial ecology, where they have been designed for the analysis of ribosomal amplicon libraries (Wang et al., 2007). Keller et al. (2015) successfully applied a classifier-based approach to ribosomal amplicons originating from pollen DNA, but to our knowledge, this approach has never been applied to nonribosomal loci, such as *matK* or *rbcL*. Successful application of this approach may enable researchers to better understand the confidence of taxonomic assignments on a read-by-read basis as well as across taxonomic ranks.

Although significant associations were found between the microscopic and molecular method, the presence of outliers cannot

TABLE 2. Average *R* coefficients for taxa present in at least five of the six samples.

Family	Average *R*	SD
Rosaceae	1.507	0.868
Asteraceae	2.922	2.744
Brassicaceae	0.190	0.116
Fagaceae	5.972	4.734
Oleaceae	2.674	5.186
Caprifoliaceae	0.091	0.017
Fabaceae	7.509	9.631
Salicaceae	0.097	0.083
Aceraceae	0.618	0.893

Note: SD = standard deviation.

be overlooked (Fig. 2). Our analysis of family-specific *R* coefficients shows that some families were consistently over- or under-represented in the molecular results when compared to the microscopic results (Table 2), suggesting that, in addition to stochastic sampling error, some systemic mechanism may bias results. Such systemic biases could be attributable to aspects of pollen plastid biology, such as taxon-specific rates of plastid incorporation or relationships between average pollen grain volume and plastid abundance. To our knowledge, no studies have directly addressed such basic questions of plastid biology within pollen tissue. Alternatively, these biases may be the result of decreased amplification efficiency for certain plant families, resulting in nondetection or underestimation of abundance.

Unless validated, pollen metabarcoding data should be questioned in terms of its capacity for quantitative inference. Such validation requires comparison between the results of novel molecular approaches and the standard method of microscopic palynology. Contemporary studies have applied this approach, using statistical tests including Pearson's product moment correlation, Spearman's rank-based correlation, and generalized linear modeling to determine the quantitative capacity of metabarcoding techniques (Keller et al., 2015; Kraaijeveld et al., 2015; Richardson et al., 2015). One conclusion consistent with the analysis presented here, and that presented by Kraaijeveld et al. (2015), is that low-copy-number plastid loci provide generally quantitative results. However, studies disagree on the quantitative capacity of the repetitive ITS2 locus for metabarcoding (Keller et al., 2015; Richardson et al., 2015).

We employed multilocus metabarcoding to characterize the taxonomic composition of polyfloral honey bee–collected pollen. Requiring only minute quantities of pollen, our approach can easily be applied to studying the foraging habits of individual honey bees, as well as other ecologically and economically important pollinators, such as solitary bees and bumble bees. Our results suggest that sequencing plastid loci produces semi-quantitative results. Furthermore, our results support the use of a multilocus, consensus-based approach over single-locus barcoding. Further research is needed to validate these trends across a larger sample of plant taxa and additional barcode loci.

LITERATURE CITED

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25: 3389–3402.

Azhagiri, A. K., and P. Maliga. 2007. Exceptional paternal inheritance of plastids in *Arabidopsis* suggests that low-frequency leakage of plastids via pollen may be universal in plants. *Plant Journal* 52: 817–823.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Bryant, V. M. Jr., and G. D. Jones. 2001. The r-values of honey: Pollen coefficients. *Palynology* 25: 11–28.

Chen, S., H. Yao, J. Han, C. Liu, J. Song, L. Shi, and Y. Zhu. 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* 5: e8613.

Cuénoud, P., V. Savolainen, L. W. Chatrou, M. Powell, R. J. Grayer, and M. W. Chase. 2002. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences. *American Journal of Botany* 89: 132–144.

Cusser, S., and K. Goodell. 2013. Diversity and distribution of floral resources influence the restoration of plant-pollinator networks on a reclaimed strip mine. *Restoration Ecology* 21: 713–721.

Dimou, M. G., and A. Thrasyvoulou. 2007. A comparison of three methods for assessing the relative abundance of pollen resources collected by honey bee colonies. *Journal of Apicultural Research* 46: 144–148.

Erdtman, G. 1943. An introduction to pollen analysis. Chronica Botanica Company, Waltham, Massachusetts, USA.

Fay, M. F., S. M. Swenson, and M. W. Chase. 1997. Taxonomic affinities of *Medusagyne oppositifolia* (Medusagynaceae). *Kew Bulletin* 52: 111–120.

Forcone, A., P. V. Aloisi, S. Ruppel, and M. Muñoz. 2011. Botanical composition and protein content of pollen collected by *Apis mellifera* L. in the north-west of Santa Cruz (Argentinean Patagonia). *Grana* 50: 30–39.

Galimberti, A., F. De Mattia, I. Bruni, D. Scaccabarozzi, A. Sandionigi, M. Barbuto, M. Casiraghi, and M. Labra. 2014. A DNA barcoding approach to characterize pollen collected by honeybees. *PLoS ONE* 9: e109363.

Girard, M., M. Chagnon, and V. Fournier. 2012. Pollen diversity collected by honey bees in the vicinity of *Vaccinium* spp. crops and its importance for colony development. *Botany* 90: 545–555.

Hilu, K., and H. Liang. 1997. The *matK* gene: Sequence variation and application in plant systematics. *American Journal of Botany* 84: 830–839.

Huson, D. H., S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. 2011. Integrative analysis of environmental sequences using MEGAN 4. *Genome Research* 21: 1552–1560.

Jones, G. D., and V. M. Bryant Jr. 1992. Melissopalynology in the United States: A review and critique. *Palynology* 16: 63–71.

Kearns, C. A., and D. W. Inouye. 1993. Techniques for pollination biologists. University Press of Colorado, Boulder, Colorado, USA.

Keller, A., N. Danner, G. Grimmer, M. Ankenbrand, K. von der Ohe, W. von der Ohe, S. Rost, et al. 2015. Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. *Plant Biology* 17: 558–566.

Kraaijeveld, K., L. A. de Weger, M. V. García, H. Buermans, J. Frank, P. S. Hiemstra, and J. T. den Dunnen. 2015. Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. *Molecular Ecology Resources* 15: 8–16.

Longhi, S., A. Cristofori, P. Gatto, F. Cristofolini, M. S. Grando, and E. Gottardini. 2009. Biomolecular identification of allergenic pollen: A new perspective for aerobiological monitoring? *Annals of Allergy, Asthma & Immunology* 103: 508–514.

Louveaux, J., A. Maurizio, and G. Vorwohl. 1978. Methods of melissopalynology. *Bee World* 59: 139–157.

Mogensen, H. L. 1996. The hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany* 83: 383–404.

Moore, P. D., J. A. Webb, and M. E. Collinson. 1991. Pollen analysis, 2nd ed. Blackwell Scientific Publications, Oxford, United Kingdom.

Reboud, X., and C. Zeyl. 1994. Organelle inheritance in plants. *Heredity* 72: 132–140.

Richardson, R. T., C.-H. Lin, D. B. Sponsler, J. O. Quijia, K. Goodell, and R. M. Johnson. 2015. Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey bees in an agroecosystem. *Applications in Plant Sciences* 3: 1400066.

Simel, E. J., L. R. Saidak, and G. A. Tuskan. 1997. Method of extracting genomic DNA from non-germinated gymnosperm and angiosperm pollen. *BioTechniques* 22: 390–392, 394.

Tang, L. Y., N. Nagata, R. Matsushima, Y. Chen, Y. Yoshioka, and W. Sakamoto. 2009. Visualization of plastids in pollen grains: Involvement of FtsZ1 in pollen plastid division. *Plant & Cell Physiology* 50: 904–908.

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

Wilson, E. E., C. S. Sidhu, K. E. LeVan, and D. A. Holway. 2010. Pollen foraging behaviour of solitary Hawaiian bees revealed through molecular pollen analysis. *Molecular Ecology* 19: 4823–4829.

Appendix 1. GPS coordinates of each sampling location.

Sampling location	Latitude (°N)	Longitude (°W)
A	40.09	83.39
B	40.05	84.15
C	39.96	83.43
D	39.99	83.59
E	39.91	84.00
F	39.86	83.66

Appendix 2. Supply list and protocol sheet.

A. Reagents and kits used

1. QIAGEN DNeasy Plant Mini Kit (QIAGEN, Venlo, The Netherlands)
2. 0.5-mm zirconium beads
3. Phusion High-Fidelity PCR Kit (New England Biolabs, Ipswich, Massachusetts, USA)
4. PureLink PCR Purification Kit (Life Technologies, Carlsbad, California, USA)
5. NEBNext Ultra DNA Library Prep Kit for Illumina and NEBNext Multiplex Oligos for Illumina (New England Biolabs)
6. Agencourt AMPure XP purification kit (Beckman Coulter, Brea, California, USA)

B. Equipment

1. Populated 8-frame Langstroth honey bee hives
2. Sundance I bottom-mounted pollen trap (Ross Rounds, Albany, New York, USA)
3. Mini-BeadBeater-1 (BioSpec Products, Bartlesville, Oklahoma, USA)
4. PCR machine (Mastercycler ep Gradient) (Eppendorf AG, Hamburg, Germany)

C. Pollen preparation

1. Weigh out 10% of sample into appropriately sized beaker with magnetic stir bar.
2. Add approximately four volumes of 50% EtOH to the pollen (approximate the pollen volume from the beaker).
3. Stir on high for 25 min (if pollen particles remain, add an additional volume of water and stir 5 min, repeat if necessary).
4. Vacuum-filter pollen from EtOH in Büchner funnel, transfer pollen to dry weigh boat, and allow to air dry.
5. Use spatula to break pollen up until powdery and place pollen in 50-mL conical tube for storage at −20°C.

D. DNA extraction (QIAGEN DNeasy Plant Mini Kit)

1. Weigh 50 mg of prepared pollen into BeadBeater-safe microcentrifuge tube.
2. Add 600 µL of QIAGEN lysis buffer (AP1).
3. Add 0.5-mm zirconium beads (fill tube 3/4 full).
4. Place tube in BeadBeater, pulverize for 1 min, hold in ice for 1 min, and pulverize for an additional minute.
5. Add 300 µL of deionized water to BeadBeater tube and mix thoroughly.
6. Transfer 300 µL of solution from BeadBeater tube to sterile microcentrifuge tube.
7. Add 130 µL of buffer P3 and incubate on ice for 5 min.
8. Centrifuge 5 min at 20,000 *g*.
9. Transfer supernatant into QIAShredder Mini Spin Column (purple column) with a 2-mL collection tube attached.
10. Centrifuge 2 min at 20,000 *g*.
11. Transfer flowthrough to a new microcentrifuge tube.
12. Add 1.5 volumes of buffer AW1 and mix by pipetting.
13. Pipette 675 µL of above mixture into DNeasy Mini Spin Column (white column).
14. Centrifuge 1 min at 6000 *g* and discard flowthrough.
15. Repeat, adding 675 µL more.

Note: The tube can only hold 675 µL at a time, but you will likely have 1000 µL in each sample, so you have to do this multiple times, centrifuging and discarding the flowthrough each time. Be sure to use the same spin column with the same sample each time.

16. Add 500 µL of buffer AW2, centrifuge 1 min at 6000 *g*, and discard flowthrough.
17. Repeat above step. After discarding flowthrough, transfer spin column to clean microcentrifuge tube.
18. Pipette 30 µL of elution buffer AE onto middle of spin column membrane, incubate 1 min before centrifuging at 6000 *g* for 1 min.
19. Repeat the above step. After incubation the tube should contain 60 µL of DNA solution ready for PCR.

E. PCR amplification (Phusion High-Fidelity PCR Kit)

Assemble all reaction components on ice and quickly transfer the reactions to a thermocycler preheated to the denaturation temperature (98°C). All components should be mixed and lightly centrifuged prior to use. It is best to add Millipore water first and important to **add Phusion DNA Polymerase last**.

1. Assemble reagents from Table A1 in a microcentrifuge tube on ice.

Table A1. PCR mastermix reagents.

Reagent	Volume (µL)	Final concentration
Nuclease-free water	30.5	
5× Phusion HF	10	1×
10 mM dNTPs	1	200 µM
10 µM Forward primer	2.5	0.5 µM
10 µM Reverse primer	2.5	0.5 µM
2 U/µL Phusion DNA polymerase	0.5	0.4 U/20 µL PCR
Template DNA	3	150 ng (total amount)
Total	50	

2. Mix and spin down in centrifuge.
3. Perform PCR using conditions appropriate for the primers being used. (Primer sequences are shown in Table A2, conditions are shown in Tables A3, A4, and A5 for ITS2, *matK*, and *rbcL* primer sets, respectively.)

TABLE A2. Primer sequences and PCR conditions used for the amplification of plant ITS2, *matK*, and *rbcL* loci.

Loci	Primer sequences (5′–3′)
ITS2	F: ATGCGATACTTGGTGTGAAT
	R: GACGCTTCTCCAGACTACAAT
matK	F: CGATCTATTCATTCAATATTTC
	R: TCTAGCACACGAAAGTCGAAGT
rbcL	F: ATGTCACCACAAACAGAAAC
	R: TCGCATGTACCTGCAGTAGC

TABLE A3. PCR conditions for ITS2 locus.

Step	Cycles	Temperature (°C)	Time
Initial denaturation	1	98	30 s
Denaturation	30	98	10 s
Annealing	30	59	30 s
Extension	30	72	30 s
Final extension	1	72	10 min
Hold	1	4	∞

TABLE A4. PCR conditions for *matK* locus.

Step	Cycles	Temperature (°C)	Time
Initial denaturation	1	98	30 s
Denaturation	30	98	10 s
Annealing	30	58	30 s
Extension	30	72	30 s
Final extension	1	72	10 min
Hold	1	4	∞

TABLE A5. PCR conditions for *rbcL* locus.

Step	Cycles	Temperature (°C)	Time
Initial denaturation	1	98	30 s
Denaturation	30	98	10 s
Annealing	30	57	30 s
Extension	30	72	30 s
Final extension	1	72	10 min
Hold	1	4	∞

F. Purification of PCR products (PureLink PCR Purification Kit)

1. Add 200 µL of PureLink Binding Buffer (B2) with isopropanol to 50 µL of the PCR product (50 µL). Mix by pipetting up and down slowly.
2. Put the sample into a PureLink Spin Column with a collection tube.
3. Centrifuge the column at 12,000 *g* for 1 min. Discard the flowthrough.
4. Reinsert the column into the collection tube and add 650 µL PureLink Wash Buffer (W1) with ethanol. Centrifuge the column at 12,000 *g* for 1 min. Discard the flowthrough and place the column in the same collection tube.
5. Centrifuge the column at 21,000 *g* for 2 min, discard the flowthrough.
6. Place the column into a clean 1.7-mL elution tube. Add 50 µL of PureLink Genomic Elution Buffer to the center of the column. Incubate the column at room temperature for 1 min. Centrifuge the column at 21,000 *g* for 2 min.
7. The elution tube contains the purified PCR product. Store the purified DNA at 4°C for immediate use or at −20°C for long-term storage.

G. Library preparation (NEBNext Ultra DNA Library Prep Kit for Illumina and NEBNext Multiplex Oligos for Illumina)

i. NEBNext End Prep

Keep reagents and tubes on ice unless otherwise specified.

1. Mix the following components in a sterile nuclease-free tube:

End Prep Enzyme Mix (green)	3.0 µL
End Repair Reaction Buffer (10×) (green)	6.5 µL
Fragmented DNA	55.5 µL
Total volume	65 µL

2. Place 500 ng of sample (determine volume using Nanodrop spectroscopy) in designated tube and add Millipore water to a total volume of 55.5 µL.
3. Mix by pipetting slowly, then centrifuge briefly to collect all liquid from the sides of the tube.
4. Place in thermocycler and run the following program: 30 min at 20°C, 30 min at 65°C, then hold at 4°C.

ii. Adapter ligation

1. Add the following components directly to the End Prep reaction mixture and mix well:

Blunt/TA Ligase Master Mix (red)	15 µL
NEBNext Adaptor for Illumina* (red)	2.5 µL
Ligation Enhancer (red)	1 µL
Total volume	83.5 µL

*The NEBNext Adaptor is provided in the Multiplex Oligos for Illumina kit.

2. Mix by pipetting slowly, then centrifuge briefly to collect all liquid from the sides of the tube.
3. Incubate at 20°C for 15 min in a thermal cycler.
4. Add 3 µL of USER enzyme.
5. Mix well and incubate at 37°C for 15 min.

iii. Size selection of adapter-ligated DNA

The size-selection protocol is based on a starting volume of 100 µL.

1. Vortex AMPure XP beads to resuspend.
2. Add 13.5 µL of distilled water to the adapter ligation reaction for a 100-µL total volume.
3. Add 35 µL of resuspended AMPure XP beads to the 100-µL ligation reaction.
4. Mix well by pipetting up and down slowly at least 10 times.
5. Incubate for 5 min at room temperature.
6. Quickly spin the tube and place the tube on an appropriate magnetic stand to separate the beads from the supernatant.
7. After the solution is clear (about 5 min), carefully transfer the supernatant containing your DNA to a new tube. (**Caution: do not discard the supernatant.**) Discard the beads that contain the unwanted large DNA fragments.
8. Add 15 µL of resuspended AMPure XP beads to the supernatant, mix well, and incubate for 5 min at room temperature.
9. Briefly centrifuge the tube and place it on an appropriate magnetic stand to separate the beads from the supernatant. After the solution is clear (about 5 min), carefully remove and discard the supernatant that contains unwanted small fragments of DNA. Be careful not to disturb the beads that contain the desired DNA targets. (**Caution: do not discard beads.**)
10. Add 200 µL of 80% ethanol to the tube while in the magnetic stand. Incubate at room temperature for 30 s, and then carefully remove and discard the supernatant.
11. Repeat Step 8 twice for a total of three washes.
12. Air dry the beads for 10 min while the tube is on the magnetic stand with the lid open.
13. Elute the DNA target from the beads into 28 µL of 10 mM Tris-HCl (pH 8.0). Mix well on a vortex mixer or by pipetting up and down slowly. Briefly centrifuge the tube and place it on a magnetic stand. After the solution is clear (about 5 min), transfer 23 µL to a new PCR tube for amplification.

iv. PCR amplification

1. Mix the following components in sterile strip tubes:

Adapter-ligated DNA fragments	23 µL
NEBNext High Fidelity 2× PCR Master Mix (blue)	25 µL
Index primer* (blue)	1 µL
Universal PCR primer* (blue)	1 µL
Total volume	50 µL

*The primers are provided in NEBNext Multiplex Oligos for Illumina.

2. Perform PCR using the following conditions:

Step	Cycles	Temperature (°C)	Time
Initial denaturation	1	98	30 s
Denaturation	9	98	10 s
Annealing	9	65	30 s
Extension	9	72	30 s
Final extension	1	72	5 min
Hold	1	4	∞

v. Cleanup of PCR amplification

1. Vortex AMPure XP beads to resuspend.
2. Add 50 µL of resuspended AMPure XP beads to the PCR reactions (~50 µL). Mix well by pipetting up and down slowly at least 10 times.
3. Incubate for 5 min at room temperature.
4. Briefly centrifuge the tube and place it on an appropriate magnetic stand to separate beads from supernatant. After the solution is clear (about 5 min), carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets. (**Caution: do not discard beads.**)
5. Add 200 µL of 80% ethanol to the PCR plate while the tube is in the magnetic stand. Incubate at room temperature for 30 s, and then carefully remove and discard the supernatant.
6. Repeat Step 5 once.
7. Air dry the beads for 10 min while the PCR plate is on the magnetic stand with the lid open.

8. Elute DNA target from beads into 33 µL of 10 mM Tris-HCl (pH 8.0). Mix well by pipetting up and down slowly at least 10 times. Briefly centrifuge the tube and place it on an appropriate magnetic stand to separate beads from supernatant. After the solution is clear (about 5 min), carefully transfer 28 µL of supernatant to a new PCR tube.
9. Store libraries at −20°C.

Note: Libraries must have at least a concentration of 7 ng/µL for Qubit analysis. However, a concentration range of 5–50 ng/µL is sufficient for the bioanalyzer analysis.

H. Bioinformatics workflow

All analyses were performed on a 12-core HP Intel Xeon X5650 Unix machine with 48 GB of RAM.

i. Quality control

File names are designated by the following parenthetical symbols: { }

Trim reads by quality

Unix-command-line-$ java -jar /path/to/Trimmomatic/software/Trimmomatic-0.32/trimmomatic-0.32.jar SE -threads 12 -phred33 -trimlog logfile {DATA_FILE.fastq.gz} {DATA_trimmed.fastq} TRAILING:20 LEADING:20 MINLEN:50

Dereplicate reads and convert to FASTA

Unix-command-line-$ /path/to/fastx/toolkit/fastx_collapser -i {DATA_trimmed.fastq} -o {DATA_trimmed_collapsed.fasta} -Q 33

ii. BLAST alignment

Unix-command-line-$ /path/to/blast/software/ncbi-blast-2.2.29+/bin/blastn -db /path/to/reference/database/database.fasta -query {DATA_trimmed_collapsed.fasta} -out {OUTPUT_FILE_NAME} -evalue 1e-150 -num_descriptions 1 -num_alignments 1 -num_threads 12 -outfmt 0 -perc_identity 99 (95% identity was used for ITS2)