

Improving plant DNA metabarcoding accuracy with ecological filters and Angiosperms353: Field and pollen microscopy validation

Reed Clark Benkendorf^{1,2}  | Emily J. Woodworth^{1,2} | Paul J. CaraDonna^{1,2,3}  |
Jane E. Ogilvie³  | Sophie Taddeo^{1,2,4}  | Jeremie B. Fant^{1,2} 

¹Chicago Botanic Garden, Glencoe, Illinois 60022, USA

²Plant Biology and Conservation, Northwestern University, Evanston, Illinois 60208, USA

³Rocky Mountain Biological Laboratory, P.O. Box 519, Crested Butte, Colorado 81224, USA

⁴Department of Environmental and Ocean Sciences, University of San Diego, San Diego, California 92110, USA

Correspondence

Jeremie B. Fant, Chicago Botanic Garden, Glencoe, Illinois 60022, USA.
Email: jfant@chicagobotanic.org

Abstract

Premise: Metabarcoding has become a successful tool for the identification of species in ecological assemblages. However, the usefulness of metabarcoding for identifying plant species has been hampered due to a lack of universal gene regions that work across all taxa, limiting the applications of metabarcoding in ecology.

Methods: Here, we outline a spatiotemporal approach that combines Angiosperms353 baits with species distribution models and phenological analyses to generate a list of candidate species to increase metabarcoding accuracy. To evaluate the ecological realism of our framework, we compared the results of DNA metabarcoding pollen loads of wild bumble bees to long-term field observations of bee–plant interactions and visual pollen identification.

Results: We show that metabarcoding bumble bee pollen loads was most accurate when combined with a candidate taxa list of plants flowering when the bumble bees were foraging, which improved the accuracy and taxonomic precision of 77.5% of samples relative to non-filtered matches.

Discussion: With the proliferation of species occurrence and phenology data and advances in computing and software, spatiotemporal filtering provides an improved approach for interpreting metabarcoding results. Additionally, we demonstrate that Angiosperms353 offers significant promise for metabarcoding projects to reveal species interactions.

KEYWORDS

Angiosperms353, bee–plant interactions, *Bombus*, herbarium, metabarcoding

Large-scale species loss, biotic homogenization, and the impacts of these processes on ecosystem stability and functions have inspired calls for more consistent monitoring of species and their diversity (Cardinale et al., 2012; Pimm et al., 2014; Socolar et al., 2016). In addition to which species are present, the interspecific interactions that occur among them are critical components of ecosystem stability, quality, and function (Futuyma and Agrawal, 2009). Therefore, biodiversity monitoring must consider both species and their interactions (Lindenmeyer and Likens, 2009; Westgate et al., 2013). However, monitoring species *and* their interactions remains logistically challenging (Yoccoz et al., 2001;

Lindenmeyer and Likens, 2009), thereby leaving many regions, ecosystems, and taxa under-observed (Collen et al., 2008; Meyer et al., 2015; Ruete, 2015). These resulting data gaps impact the capacity to identify and implement robust conservation interventions (Tylianakis et al., 2010). DNA barcoding and metabarcoding—the identification of a sample from a single fragment of an organism or a mix of organisms, respectively—have shown considerable promise in all kingdoms of life (Ruppert et al., 2019). The ability to identify species from fragments of organisms (e.g., hair, scat, soil, pollen) dramatically increases our ability to identify not only the presence of species, but also interactions among

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2025 The Author(s). *Applications in Plant Sciences* published by Wiley Periodicals LLC on behalf of Botanical Society of America.

species (Bell et al., 2023). DNA barcoding and metabarcoding approaches have immense potential because they can help increase our ability to monitor more ecosystems in greater depth, but a major challenge is that the results of these studies can be spurious and can lead to misleading ecological inferences (Zinger et al., 2019).

New and emerging genomic datasets can enable the use of metabarcoding to resolve a diverse array of questions relevant to our applied and conceptual understanding of ecology and evolution (Hollingsworth et al., 2016; Kress, 2017). Such datasets include the Plant and Fungal Tree of Life (PAFTOL)—the largest-ever plant systematic endeavour (Baker et al., 2021a)—which will contain hybridization capture (Hyb-Seq) data from at least one species in each of the 14,000 genera of the plant kingdom with the Angiosperms353 probes (Johnson et al., 2019; Baker et al., 2021a, 2021b; Zuntini et al., 2024). These publicly available data provide a phylogenetically comprehensive backbone for plant barcoding, but they still only contain a fraction of all known plant species. To date, for some groups of plants, DNA barcoding has successfully identified organisms to species (Kress, 2017), but this has proven more challenging for many other clades (Spooner, 2009; China Plant CBOL Group et al., 2011; Coissac et al., 2012; Liu et al., 2014; Caetano Wyler and Naciri, 2016; Hollingsworth et al., 2016).

Although metabarcoding is often proposed as a rapid and easy method for monitoring ecological interactions, making generalizations about its utility is difficult as researchers rarely implement the same methodology or analyses (Zinger et al., 2019). An issue for plants is the lack of reliable universal

barcodes and the poor resolution of existing barcodes, leading to spurious results (Spooner, 2009; Caetano Wyler and Naciri, 2016; Hollingsworth et al., 2016). To address current metabarcoding challenges, we present a more efficient approach that generates a shorter and more appropriate reference list for identifying candidate taxa from barcode data (Figure 1). This spatiotemporal approach uses current species distribution data, species associations with key environmental predictors (i.e., spatial filtering), and phenology data (i.e., temporal filtering) to generate a list of species relevant to the study area and sampling period. The study area list of plants can then be leveraged to identify target plant tissue collection and sequencing, reduce the size of a reference sequence database, increase computation speed and efficiency, and filter metabarcoding results for accuracy. In combination, this approach can help increase the accuracy and efficiency of metabarcoding results in plants. To demonstrate its utility, we apply our methodological approach to the metabarcoding of the pollen loads of several wild bumble bee species, which we validate by applying spatial and temporal filtering with molecular approaches, using both morphological pollen identifications and multi-year field observations.

METHODS

Overview of the methodological approach

Our approach has three main parts (Appendices S1, S2). Given that there are over 350,000 known vascular plant

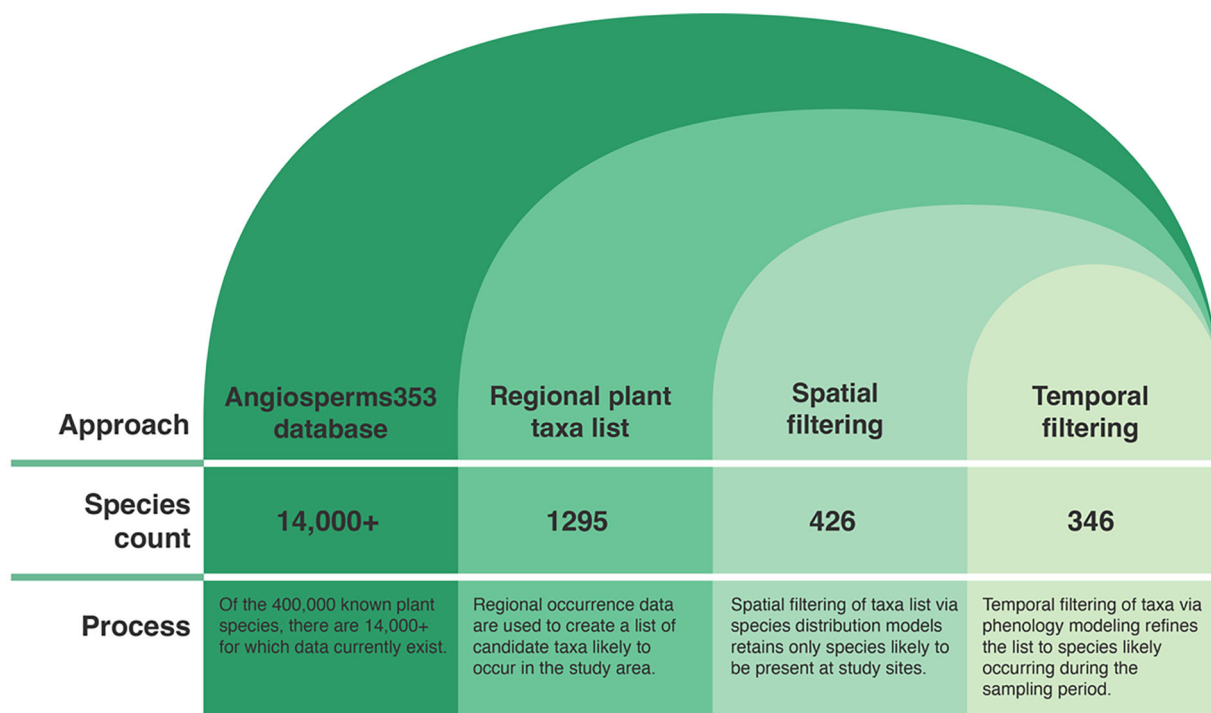


FIGURE 1 Key methodological steps in our approach to improve metabarcoding species identification, with corresponding species filtering from our example.

species (Antonelli et al., 2023), we begin by (1) generating a regional plant list of potential candidate taxa from occurrence data, which is necessary to improve the reliability and efficiency of metabarcoding (Figure 1). Because occurrence data are coarse and their effort is distributed unevenly, any plant list based purely on geography is likely to overstate the species composition at any one site. Therefore, the regional list is further refined using (2) spatial filtering via species distribution modeling to identify species likely to occur within a site based on their environmental requirements. After spatial filtering, we continue to refine the candidate taxa list via (3) temporal filtering to identify species that are likely in a phenophase of interest at the time of sampling. This final list serves to facilitate all other steps in the metabarcoding process, including the sequencing of additional focal species likely to be found in environmental samples used for metabarcoding to improve the odds that interacting species are detected. We test this metagenomic and spatiotemporal approach in the context of plant–pollinator interactions by identifying the plant species found within pollen loads collected by several species of wild bumble bees. Finally, we validate the accuracy of plant species identification from our approach by comparing molecular results to direct, long-term field observations of bumble bee–flower interactions and direct identification of pollen via microscopy.

Potential candidate taxa list

The first step of our approach is to collate existing species occurrence data to create a regional list of candidate species likely to occur in a study area (e.g., county-level plant occurrence). Such occurrence data can be retrieved and aggregated from many databases, such as the Global Biodiversity Information Facility (GBIF [<https://www.gbif.org/>]), herbaria consortia, and community science datasets (e.g., iNaturalist [<https://www.inaturalist.org/>]). These databases can be queried to extract a list of plant species with occurrence records in the ecological or administrative unit that includes the study site or within a search radius.

Given the coarseness of regional occurrence data, especially data sources that aggregate occurrences at county levels such as the USDA PLANTS database (<https://plants.usda.gov/>) in the United States where counties often have several thousand species, additional filtering of taxa is needed to further refine the list of candidate species. This is in part because regional occurrence datasets often include occurrence records with taxonomic and geographic inaccuracies or extirpated historical records (Smith et al., 2016; Freitas et al., 2020). Spatial filtering of likely taxa can be achieved using species distribution models (SDMs) to retain only species that are likely to be present at study sites (Figure 1). SDMs allow filtering of species occurrences based on environmental variables that are associated with occurrences of a species (e.g., climate, elevation, soil types) or that constrain its distribution (e.g., land uses/land covers, habitat type). Using widely available, public geospatial

datasets as covariates (see Appendix S3 for examples), SDM algorithms can score remaining portions of the study area based on a probability of suitability scale ranging from 0 (highly unsuitable) to 1 (highly suitable).

Filtering for plant species that are likely to be in a particular phenophase at the time of sampling collection further refines the species list and improves the accuracy of metabarcoding. This is especially true when studying species interactions, where temporal co-occurrence can rule out huge numbers of interactions that are unlikely (Olesen et al., 2010; CaraDonna et al., 2021). Recent increases in the digitization of herbarium collections and various community science projects (e.g., National Phenology Network [<https://www.usanpn.org/>], Budburst [<https://budburst.org/>]) provide data that can elucidate coarse phenological patterns to retain only plant species with phenophases that align with the sampling period.

Using spatiotemporal filtering in metabarcoding

The reduced candidate taxa list can then be used to guide the collection of plant tissue samples to generate a reference library. As online genomic databases often contain only a single representative of each genus and do not always account for geographic variability, it can be worthwhile supplementing with additional samples of likely plant candidates to increase the efficiency and accuracy of future analyses. However, given the cost of sequencing, it is best to minimize reference sequence generation.

While there are numerous next-generation sequencing approaches available for metabarcoding, we advocate for a target capture approach that uses select regions of the genome (loci) known to be useful for systematic inference at multiple taxonomic resolutions (Weitemier et al., 2014). Regardless of the approach used, the process begins with preparing a genomic library (a pool of DNA extract composed of short fragments [100–300 bp] from across the genome) from each sample. Once a library has been generated, a synthetic probe, often referred to as a “bait,” is used to preferentially retain sequences from desired loci.

The sequence fragments generated are compared back to the original reference sequences used for the bait set using a sequence alignment algorithm. This allows all sequencing contigs to be sorted by loci and then assembled into a consensus sequence for each locus (e.g., HybPiper v.1.3, Johnson et al., 2016; MAFFT, Katoh and Standley, 2013). These consensus sequences of each potential taxon can then be added to the existing reference sequence library for screening metabarcoding data and used to compare metagenomics sequences to ecologically relevant results.

Metabarcoding

Isolation of DNA from environmental samples will be similar to the process described above, with the target

“barcode” loci being isolated and amplified before sequencing. The short sequencing fragments (contigs) generated during the next-generation sequencing process will then be compared to the known sequences within the custom-made sequence reference library, using any one of a variety of alignment methods.

Application of the approach

System background

To test the effectiveness of our methodological approach, we applied it to identify the plant species found in the pollen loads (corbiculae) of queen bumble bees (*Bombus latreillei* spp.) collected from the Rocky Mountain Biological Laboratory in Colorado, USA (Appendix S4). We collected pollen loads from wild foraging queen bees between May and July of 2015 at six permanent study sites, along a 16-km stretch with each site separated by more than 2 km, where we monitored bumble bees and floral resources (details in Ogilvie and CaraDonna, 2022). To harvest the pollen loads, we captured queens in an insect net, transferred them into a restraining device (Kearns and Thomson, 2001), collected a pollen load from one leg, and then released them. We collected 64 corbiculae pollen loads from queens of several common wild bumble bee species: *Bombus appositus* Cresson, *B. bifarius* Cresson, *B. californicus* Smith, *B. flavifrons* Cresson, *B. nevadensis* Cresson, and *B. rufocinctus* Cresson (Pyke, 1982; Ogilvie and CaraDonna, 2022). Details of vegetation surveys are also described in the literature above.

We used eight years (2015–2022) of observational data on *Bombus* flower visits to identify the plant taxa most frequently visited by queens across all years and compared these observations with metagenomic data. Bumble bee abundance and interactions with flowering plants were monitored for one hour per week at the six study sites (Ogilvie and CaraDonna, 2022). We focused on queen bumble bees as they represent an important component of the colony life cycle, and the amount and quality of pollen they collect for nest provisioning can directly influence offspring production and colony success (Goulson, 2010; Woodard and Jha, 2017; Ogilvie and CaraDonna, 2022).

Survey databases to generate a regional taxa list

To create a preliminary list of subalpine plant species likely to occur in the study area (Gunnison County, Colorado, USA), we queried the Botanical Information and Ecology Network (BIEN) database (<https://bien.nceas.ucsb.edu/bien/>) and identified plant species observed within the Southern Rockies in Colorado. For this candidate taxa list, we then gathered all occurrences recorded within a bounding box around the Omernik Level 3 ecoregion in which the study site is located.

Spatial filtering via species distribution modeling

We conducted all statistical tests in R 3.6+ (R Core Team, 2019). To refine the resulting species list based on their abiotic requirements, we created SDMs with the sdm R package using four commonly implemented algorithms: random forest, boosted regression trees, generalized linear models, and generalized additive models (see Appendix S1 for a complete description) (Naimi and Araujo, 2016). We used 26 environmental variables at a 30-m resolution to construct the SDMs (see Appendix S3) and removed highly collinear predictors before using linear regression models (Naimi et al., 2014). We then followed standard protocols to evaluate the accuracy of all four SDM outputs (Allouche et al., 2006; Araujo and New, 2007). This resulted in a list of 426 species with environmental requirements that matched our study sites (Figure 1).

Temporal filtering of the species list

Using the spatially filtered species list (426 species; Figure 1), we identified species with flowering phenologies that overlapped with the pollen load collection dates. We used Weibull estimates of phenological parameters developed by Belitz et al. (2020) and Pearse et al. (2017) to estimate the flowering periods of 383 species from herbarium records (other species had insufficient herbarium records). We generated Weibull distributions for species with more than 10 herbarium records, including the dates when 10% of individuals had begun flowering (initiation), when 50% were flowering, and when 90% of individuals had flowered (cessation). We used the initiation and cessation dates, respectively, as the effective start and end of flowering and assessed the number of species likely to flower in each week of the 11-week period of queen bee activity. For validation, these estimates were compared to a long-term observational study (1974–2012) of flowering phenology (CaraDonna et al., 2014) and the floral abundance data from 2015 observations (see “System background” in Ogilvie and CaraDonna, 2022), using Kendall's tau.

Microscopic pollen identification

To validate the metagenomic identification of species in pollen loads, we also visually identified pollen in the bee corbiculae loads to species (or to the lowest taxonomic level possible) under a microscope. To do so, we first gathered a pollen reference library of taxa known to be visited by bumble bees from fuchsin jelly-stained grains previously prepared by the authors and other researchers (Beattie, 1971; Brosi and Briggs, 2013) and augmented the reference library with herbarium collections (121 slides in total). We identified pollen to distinguishable morphotypes using microscopy. Details of our sampling and identification protocol are included in the Supporting Information (Appendices S1, S5, S6).

Metabarcoding: Additional plant tissue collection and extraction

We supplemented the available reference sequences by sequencing 15 additional species collected around our bumble bee study sites (Appendix S7). Plant genomic DNA was isolated from ~1 cm² of leaf tissue from silica gel-dried or herbarium material using a modified cetyltrimethylammonium bromide (CTAB) protocol (Doyle and Doyle, 1987) that included two chloroform washes.

Pollen DNA extraction

We extracted DNA from 54 corbiculae pollen load samples using a modified CTAB method (Guertler et al., 2014; Lalhmangaihi et al., 2014), which included using a sodium dodecyl sulfate (SDS) extraction buffer (350 µL, 100 mM Tris-HCl, 50 mM EDTA, 50 mM NaCl, 10% SDS v/v, pH 7.5). DNA extracts were then cleaned using 2:1 v/v Sera-Mag beads (Cytiva, Little Chalfont, United Kingdom) to solute ratio, eluted in 0.5× Tris-EDTA, and the eluent allowed to reduce by half volume in ambient conditions. DNA was quantified using a Qubit fluorometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA).

Library preparation and bait capture (barcoding)

Sequence library preparation was performed using the NEBNext Ultra II FS-DNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, Massachusetts, USA). Fragmentation was performed at one-half volume of reagents and one-quarter enzyme mix for 40 min at 37°C, with an input of 500 ng of cleaned DNA. Adapter ligation and PCR enrichment were performed with one-half volumes, while cleanup of products was performed using SPRI beads (Beckman Coulter, Indianapolis, Indiana, USA) and recommended volumes of 80% v/v ethanol washes. The exception was the herbarium specimens, which were not fragmented and only end-repaired. Libraries were pooled and enriched with the Angiosperms353 probe kit V.4 (Arbor Biosciences myBaits Target Sequence Capture Kit; Arbor Biosciences, Ann Arbor, Michigan, USA) following the manufacturer's protocol. Sequencing was performed using an Illumina MiSeq with 150-bp paired-end reads (NUSeq Core, Chicago, Illinois, USA).

Bioinformatics

Sequences were processed using Trimmomatic 0.39, which removed sequence adapters, clipped the first 3 bp, discarded read lengths less than 36 bp, and removed reads if their average PHRED score dropped beneath 20 over a 5-bp window (Bolger and Giorgi, 2014; Tange, 2021). The only exception was that we discarded pollen samples with reads

less than 30 bp. We mapped the generated contigs to a reference genome with HybPiper using target files created by M353 (Johnson et al., 2016; McLay et al., 2021) and separated contigs according to gene regions. For the plant samples, contigs were further processed to generate a consensus sequence for each of the 353 gene regions (Appendix S8). To process the pollen sequence of contigs, we created a custom Kraken2 v2.1.2 (Wood et al., 2019) database by downloading representative species or genera from our spatiotemporal filtered taxa list (Appendix S9). This database was built and run using default parameters. Following Kraken2, Bracken v2.6.2 was used to classify sequences to terminal taxa (Lu et al., 2017). Finally, all reads that could be classified by these databases were passed to a local BLAST 2.13.0 database composed of the same sequences as the former databases (Camacho et al., 2009). We manually reviewed the initial sequence classifications made by BLAST using species presence predicted by spatial modeling, modeled flowering time from temporal modeling, and taxonomy from existing sources. We used a sequential process that reassigned sequences based on binary combinations of the factors above (Appendix S10). Given the relative sparsity of the number and relatedness of species represented in the sequence database, this was performed to (1) identify locally present species represented by surrogates in the database, (2) reduce false classifications of focal species, and (3) identify high confidence sequence matches.

In addition, we manually investigated all the reads that were classified to genera without any species predicted by spatial analyses. These reads were assigned to a variety of ranks, occasionally to genus, by consulting the alpha-taxonomic literature (Moore and Bohs, 2003; Pusalkar and Singh, 2015; Sadeghian et al., 2015; Sennikov and Kurtto, 2017). These values represented a synthesis of molecular, morphological, and observational data to compare the accuracy of BLAST and the reassignment algorithm, utilizing the potential candidate taxa list and phenological estimates.

Validating the metabarcoding approach

We compared the sequences classified by molecular methods with the results of direct bumble bee floral visitation data (hereafter “field”; Ogilvie and CaraDonna, 2022) and pollen microscopic identification (hereafter “microscopy”). We first compared the capacity of the three methods (i.e., molecular, field, microscopy) to identify species visited by bumble bees by comparing the identity of species identified at three taxonomic levels (order, genus, family). To determine how reliable each technique was in predicting the identification and abundance of pollen in the corbicula load, we compared how often the plant species identified matched across at least two of the methods used. To measure the relative frequencies of each pollen type or visit, we calculated the percentage representation of each species by dividing the count for a specific species by the sum of all data points produced by that

technique. Although the data for the three methods represent slightly different processes—number of visits (field), number of pollen grains harvested (microscopy), and gene frequencies (molecular)—we compared percentages across methodologies to determine how well the abundance predicted the importance of each plant species. In cases where a plant species was only identified by one methodology, we created a threshold to distinguish between meaningful differences and potential biological noise. We expect all approaches to produce some biological noise from either unintentional visits, contamination, or low power of identification; therefore, plant species recorded by one method that represented less than one or two observed visits, or less than 1% of the total number of pollen grains or molecular species confirmation, were designated as “false positives.”

Comparison of sequence classifications pre- and post-processing

The sequences assigned by the BLAST database (of 163 species), which included surrogate taxa (i.e., congeners to represent) and locally occurring taxa, were subjected to an automated post-processing step to replace the sequence matches for surrogates with a suitable local species (Appendix S10). After this, all sequences were manually classified collectively by researchers, using all sources of data, to serve as a reference dataset for comparison with each of the three methods.

RESULTS

Field observations and microscopic pollen identification

During field observations conducted in 2015–2022, we observed 1424 overwintered queen–pollen foraging interactions. Bumble bee queens visited 40 plant species from 35 genera and 16 families, with each bumble bee species visiting 8–20 plant species ($\bar{x} = 14.86$, median = 17). Our microscopic analyses of bumble bee corbiculae pollen loads identified 28 pollen grain morphotypes based on 10 characters (number of grains: $\bar{x} = 3319$, median = 1891, range = 514–19,924; morphotype richness: $\bar{x} = 4.5$, median = 4, range = 1–9, SD = 1.6; $n = 37$ samples) (Appendices S11, S12).

Spatial filtering

Our spatial filtering using species distribution models and logistic regression reduced the potential plant species list from 1295 to 426 species (Figure 1). All ensemble models (i.e., a final model averaging characteristics of submodels; $n = 968$) had an accuracy of 0.84 (95% confidence interval [CI] = 0.836–0.844, kappa = 0.68, $P < 0.001$, sensitivity = 0.80, specificity = 0.87, area

under the curve [AUC] = 0.92). The 493 machine learning ensembles accurately predicted the presence of 362 species (65.3% of all species modeled) and the absence of 33 species (6.0%) in the study area. Ensemble models incorrectly predicted the presence of 64 species (i.e., species predicted as present but in fact absent from the study area; 11.6% of all species modeled) and failed to meet habitat suitability thresholds for 34 species (6.1% of species modeled) that are in fact present in the study area. The linear model ensembles accurately predicted the presence of 286 species (51.6%) and the absence of 55 species (9.9%) in the study area. They incorrectly predicted the presence of 41 species (14.3%) and the absence of 93 species (16.8%) that are in fact present in the study area. The balanced accuracy of the ensemble models is 0.66 (sensitivity = 0.57, specificity = 0.75). Of the 117 plant species identified to the species level during field observations (see “System background,” above) across all plots and duration of queen bee activity, the machine learning ensembles predicted the presence of 105 (89.7% of all species observed) of them, and linear model ensembles predicted the presence of 102 (87.2%). Of the missing species, two are orchids (1.7%), six are non-native (5.1%), and one is of contested taxonomic standing (0.85%), all of which (7.65%) were not included in the initial regional plant list.

Temporal filtering

Our temporal filtering, which was applied to 383 species, further reduced our plant species list from 426 at the spatial filtering stage to at most 346 species in the week with the most species flowering (Figure 2). We compared the Weibull estimates for 58 species (15% of 383 species estimated) with direct phenological observations conducted in our study sites (CaraDonna et al., 2014), which revealed high accord with our estimates. There was very strong evidence that the Weibull estimates were positively associated with the observed

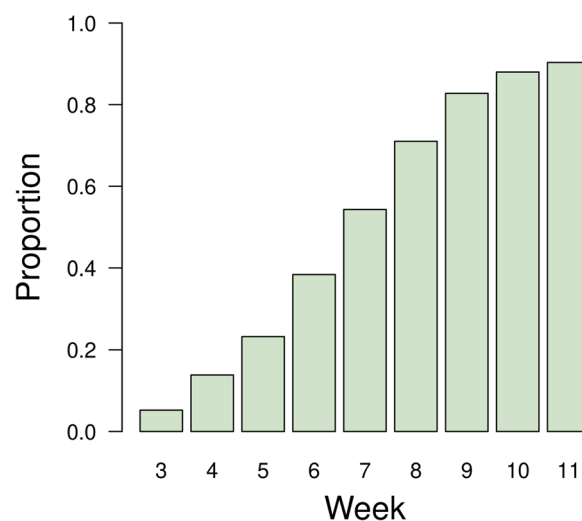


FIGURE 2 Proportion of the 383 plant species expected to flower during each week of the active period of queen bumble bees.

onset ($P < 0.0001$, $\tau = 0.61$), peak ($P < 0.0001$, $\tau = 0.65$), and cessation of flowering ($P < 0.0001$, $\tau = 0.49$).

Molecular analysis

Corbiculae loads

Of the 54 corbiculae loads from which we extracted DNA, a total of 44 could be sequenced, with 7,752,353 reads recovered. The number of reads per corbicula load varied widely ($\bar{x} = 176,190$, median = 138,395, range = 76–508,795). Of the possible 353 loci, the number that was recovered from each sample and informative to BLAST ranged from 24–353 ($\bar{x} = 305.5$, median = 331). The number of reads per loci from across all samples had a range of 178–506,653 ($\bar{x} = 20,688$, median = 12,616) (Appendix S13). A total of 10,682,538 reads were matched using Kraken, of which 10,160,768 reads were matched using Bracken and 7,549,608 reads were matched using BLAST. Based on a subjective review of the three classifiers (Appendix S14), we chose BLAST as the classification method that yielded the most probable results, as a number of the species returned by the other classifiers had flowers too small to be handled by the species of *Bombus* in this study, and therefore BLAST values were used for all subsequent analyses. Of the sequences passed into BLAST, 55.4% of the reads were classified to a species and 41.9% of the reads were classified to genus.

Comparison of sequence classifications pre- and post-processing

The improvement of accuracy between samples associated with the naive BLAST results versus temporal filtering was compared using a paired Wilcoxon signed-rank test. A Wilcoxon effect size, with a one-sided hypothesis of greater, shows strong evidence ($P < 1e-04$) that the filtered results had higher accuracy (effect size of 0.732 [95% CI = 0.57–0.84, $n = 40$, bootstrap replicates = 1000]) (Figure 3) (Hothorn et al., 2006; Kassambara, 2023).

Validating molecular methods using field and microscopy data

Given that pollen under the microscope could rarely be identified to species, we focused our comparisons of species identified by the different methodological approaches (i.e., field, molecular, microscopy) at the genus level. Of the 91 bumble bee and plant genera combinations identified across all three methodological approaches ($n = 182$ observations and samples analyzed), there were 28 bee species–plant genera combinations that were exclusively observed in the field (i.e., bee $\times$ plant species) with no supporting data from either of the lab-based methods (i.e., molecular and microscopy; Figure 4). For six of these bee–plant genera

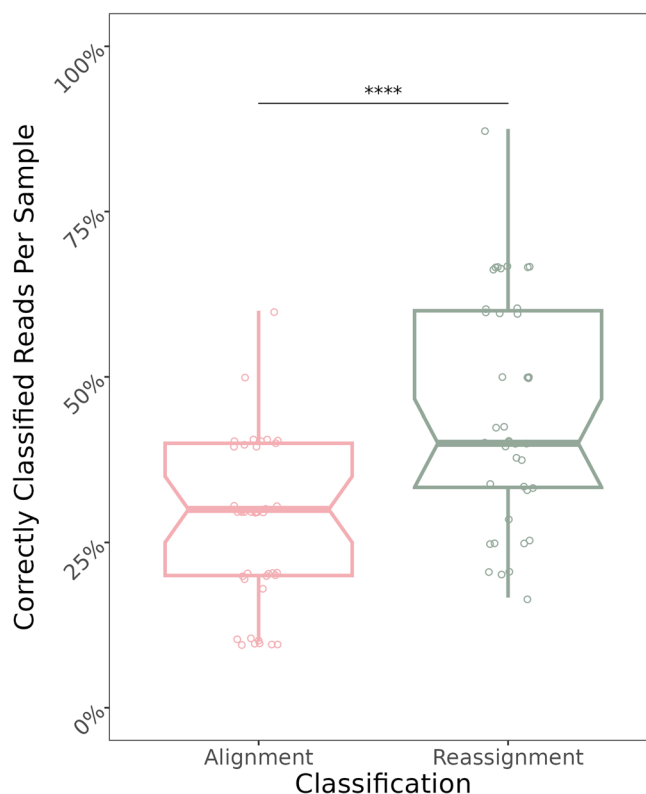


FIGURE 3 Comparison of accuracy between the initial output data from BLAST (alignment) and these same data subjected to the post-classification process, which removes surrogate and temporally restricted species (reassignment), for the top 10 most abundant reads per sample. **** denotes $P \leq 0.0001$.

interactions (*Calochortus* Pursh–*B. appositus*, *Dodecatheon* L.–*B. bifarius*, *Lonicera* L.–*B. flavifrons*, *Linum* L.–*B. appositus*, *Claytonia* L.–*B. bifarius*, *Valeriana* L.–*B. rufocinctus*), the observed bumble bee species was the only bumble bee species seen visiting this genus, and in all cases the frequency of visits was low (<2% of total visits observed) (Figure 4). By contrast, there were five specific bumble bee–plant combinations that did not have corresponding lab-based data (*Cirsium* Mill.–*B. flavifrons*, *Hydrophyllum* L.–*B. nevadensis*, *Lathyrus* L.–*B. appositus* and *B. californicus*, *Vicia* L.–*B. rufocinctus*, and *Pedicularis* L.–*B. californicus*), despite those plant genera being visited by other bumble bees, as shown by either microscopy or molecular data (Figure 4). Hence, it is likely that in these situations, pollen was collected during a visit but only in low quantities. The remaining five plant genera (*Hymenoxys* Cass., *Senecio* L., *Frasera* Walter, *Aconitum* L., and *Erythronium* L.) were visited by multiple bumble bee species in our dataset but pollen was not detected via any of the other laboratory methods, suggesting bees were primarily collecting nectar resources from these taxa.

Plant genera found in pollen microscopy samples with no corresponding field observations likely represent rare or uncommon interactions. Of the 16 genera observed 76 times in the raw BLAST classification results for the molecular

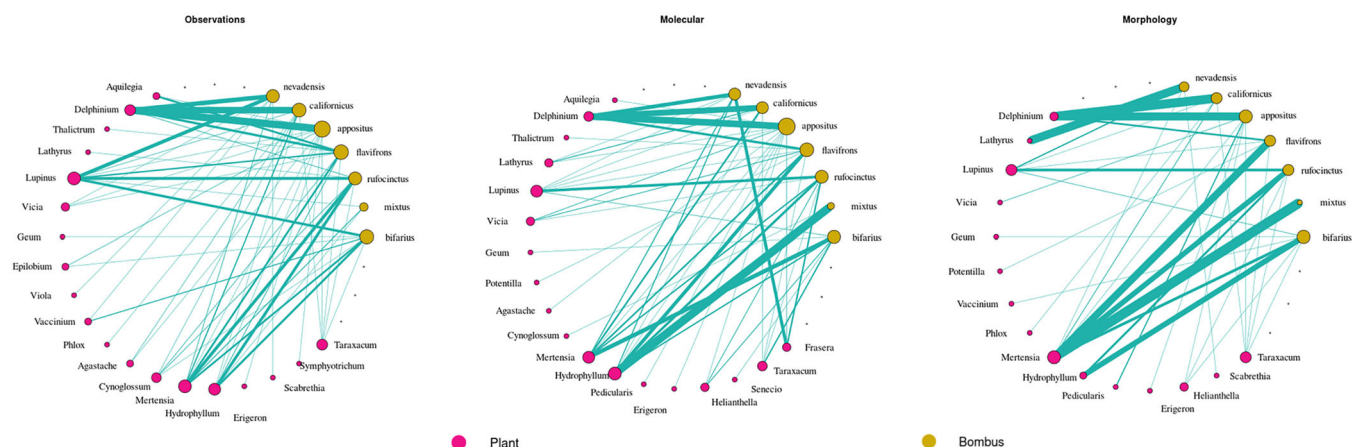


FIGURE 4 Network graph diagrams illustrating bee–plant relationships generated from each methodology: (A) field observations of bee–plant interactions (limited to species visited for pollen), (B) pollen metabarcoding from bumble bee pollen loads, and (C) pollen identification via microscopy (from bumble bee pollen loads).

data but not observed in the field, 25 of these instances were in very low abundances (representing <1% of the counts or sequences for that sample) and might represent false positives or low-level contamination. There were 66 instances that constituted less than 5% of the reads per sample, suggesting few visits or potential contamination (Figure 4). Of the bumble bee–plant genera combinations that were confirmed in two of the three methodologies, the Asteraceae genera were most likely to be recorded in pollen microscopy and field visitation but not molecular data (Figure 4). All of these were recorded in very low frequencies, suggesting bees might be picking up small amounts when foraging on Asteraceae flowers for nectar, or that there was difficulty obtaining sufficient reads to identify them. The molecular technique produced the most unique bee–plant genera combinations, detecting 41 genera relative to 31 detected by field observations (Figure 4). For the most part, these were often in very low frequency, suggesting they might represent false positives or low-level contamination.

DISCUSSION

Using wild bumble bee pollen loads, we demonstrate the benefits of creating custom regional plant species lists—filtered based on their environmental habitats (spatial filtering) and phenology (temporal filtering)—which can enhance the accuracy and precision of metabarcoding results. In combination with a bait capture approach, we show moderate to high concordance of metabarcoding to in-depth field observations of bumble bee interactions with flowering plants and identification of pollen in corbiculae loads via microscopy. The use of the Angiosperms353 baits allows for higher taxonomic resolution, often to species, although creating a regionally relevant candidate taxa list was critical for assessing whether databases have appropriate representation. An important insight gained from our

metabarcoding validation with field observation and microscopy is that the primary limitation of our approach was not insufficient data, but instead an overabundance of sequence data. In particular, the enormous amount of available global data within these databases greatly increases processing time and the potential for spurious matches. We demonstrate that our custom sequence database approach can substantially improve the accuracy of metabarcoding, with taxonomic identification accuracy improving for 77.5% of our samples after applying our spatiotemporal filtering approach.

The accuracy of the metabarcoding, especially when compared to in-depth field and microscopy data, confirms that our spatiotemporal filtering approach can be a valuable tool not only for identifying species, but also for testing hypotheses related to ecological interactions. In all cases, the most common plant species present in the pollen loads were correctly identified by all three techniques. In fact, even though each of the three methodological approaches measures slightly different biological processes (e.g., flower visitation versus pollen collected), the relative frequency of the most common plants was consistent across samples. This suggests that all three techniques can accurately identify the species and frequencies for most of the common interactions in this dataset, although field observations were essential for validating the potential plant species in the other two methods.

The application of our spatiotemporal filtering approach also revealed a few interesting natural history insights into bumble bee ecology. First, while bumble bees are often considered generalist and flexible foragers across space and time at a population level, the molecular and microscopy pollen load data revealed that individual queens are often more specialized, focusing on collecting pollen from a single or a few species during a foraging trip (consistent with Macior, 1994). Next, we observed several cases of exclusively field-observed visits to plant species, sometimes in

high frequency, to *Frasera speciosa* Douglas ex Griseb. and many Asteraceae genera. This pattern suggests that such floral visits represent nectar foraging without pollen collection (Pornon et al., 2017; Milla et al., 2022; Encinas-Viso, 2023). In contrast to this pattern, we also observed several cases in which there was pollen identified in bumble bee corbicular loads but the interaction was not observed with the field data, in particular plant–pollinator interactions with willow family members (Salicaceae). Given natural history knowledge of the study system, this observation likely represents pollen loads containing *Salix* L. spp. (willow)—yet molecular evidence instead suggests that *Populus tremuloides* Michx. (trembling aspen) is the more likely species. This generates two hypotheses to be explored. If the high number of these DNA reads in the pollen loads represent willow, then this suggests more frequent visits to willow for pollen than captured with the field data. Indeed, willows are an important spring resource for bumble bee queens, they flower in abundance around our field sites, and we have observed bumble bees on them, but individual plants are patchy and infrequent. If instead the pollen belongs to *Populus tremuloides*, as the molecular evidence suggests, then this would indicate a novel interaction that we have not observed in the field and may reflect opportunistic pollen foraging from abundant wind-pollinated aspen trees, indicating the approach can be used to identify interactions in the canopies.

The comparison of the plant species detected across the three methodological approaches, using a filtered species list, highlights that using any of the three approaches independently could lead to the identification of unlikely biological interactions. All three techniques identified unique bumble bee–plant species combinations. In most cases, these unique combinations were at low relative frequencies, suggesting that—even if they represent real interactions—they are likely rare and of limited biological importance for provisioning nest cells. Alternatively, these interactions may be spurious, representing background noise inherent to that dataset. Here, the distinction between the two sides of biological probability was fairly clear, given the multiple datasets, and for this reason we recommend a combination of two approaches to winnow artifacts from reality.

A key finding from our approach, consistent with other metabarcoding studies, is that without a candidate species list, molecular methods generate many spurious results that can be misleading (Drake et al., 2022). Given the growing size of molecular datasets, the use of short reads in next-generation sequencing, and that traditional barcoding approaches use markers of relatively conserved regions, it is increasingly likely that multiple matches will be retrieved, some of which are inevitably incorrect (Bell et al., 2017; Calderón-Sanou et al., 2020; Furlan et al., 2020). This highlights the value of having an environmentally realistic species list to cross-validate the taxa matched, which for urban studies may include popular horticultural species (Arstingstall et al., 2021; Milla et al., 2022). Although metabarcoding produced accurate and reliable identification of

the common taxa, it also identified by far the most false-positive bumble bee–plant interactions (up to one-third of the total interactions identified). Although these false positives made up only a fraction of the total dataset (0–6%), their effect is outsized. Most of these spurious matches were associated with species from families or genera that also contained true interacting partners, but which we believe arose from the misalignment of highly conserved gene regions. Accordingly, research programs that seek to use metagenomic barcoding methods must ensure that the appropriate resources are dedicated to fieldwork to anchor the datasets in biological reality. Furthermore, researchers must consider how their hypothesis will be affected by having either more false absences (missed interactions) or more false positives (non-interactions). Here, false positives were minimized in an effort to characterize the core plant species on which *Bombus* queens rely to provision nest cells. Other studies may instead want to characterize all insects that pollinate a plant species and will consequently use less stringent filters to minimize false absences.

Conclusions

Our spatiotemporal filtering approach can improve both the efficiency and accuracy of metabarcoding, supporting both basic and applied science by improving biodiversity monitoring. There are many contexts in which our approach can be useful, such as facilitating the monitoring of plant–animal interactions (Banerjee et al., 2022), particularly those that are challenging to visually observe (Bell et al., 2023). Combining metabarcoding with species distribution models (as we do here) can be used to improve the detection of species that are otherwise hard to identify in the field. This includes graminoids, mosses, lichens, and ferns—clades that are difficult to distinguish without reproductive organs—or individuals at multiple reproductive phases.

AUTHOR CONTRIBUTIONS

R.C.B., J.E.O., J.B.F., and S.T. conceived the ideas and designed methodology; J.E.O., P.J.C., R.C.B., and E.J.W. collected the data; R.C.B. analyzed the data. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGMENTS

The authors would like to sincerely thank N. Zerega, P. Herendeen, H. Noble, A. McDonnell, and E. Loke for their assistance with acquiring herbarium loans, pollen identification, and preparing the genomic libraries; I. Breckheimer for sharing spatial data; herbarium curators B. Legler, R. Williams, and E. Nelson for tissue loans; and J. Reithel and Rocky Mountain Biological Laboratory staff for access to field sites and logistical support. We are also grateful for the comments made by two reviewers and *Applications in Plant Sciences* Associate Editor Ed McAssey, which greatly improved the manuscript.

DATA AVAILABILITY STATEMENT

All novel sequencing data are available in the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) under PRJNA1093153 (metagenomic) and PRJNA1083211 (reference). Further data and code are located on Dryad (<https://doi.org/10.5061/dryad.8sf7m0d1c>). Plant geolocation data from the Bureau of Land Management terrestrial Assess, Inventory, and Monitor program cannot be shared without permission from the BLM itself. Interested parties can apply at <https://gbp-blm-egis.hub.arcgis.com/pages/aim> for access to this portion of the plant geolocation data.

ORCID

Reed Clark Benkendorf  <https://orcid.org/0000-0003-3110-6687>

Paul J. CaraDonna  <https://orcid.org/0000-0003-3517-9090>

Jane E. Ogilvie  <https://orcid.org/0000-0001-8546-0417>

Sophie Taddeo  <https://orcid.org/0000-0002-7789-1417>

Jeremie B. Fant  <https://orcid.org/0000-0001-9276-1111>

REFERENCES

- Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43: 1223–1232. <https://doi.org/10.1111/j.1365-2664.2006.01214.x>
- Antonelli, A., R. Govaerts, E. Nic Lughadha, R. E. Onstein, R. J. Smith, and A. Zizka. 2023. Why plant diversity and distribution matter. *New Phytologist* 240: 1331–1336. <https://doi.org/10.1111/nph.19282>
- Araujo, M. B., and M. New. 2007. Ensemble forecasting of species distributions. *Trends in Ecology & Evolution* 22: 42–47.
- Arstingstall, K. A., S. J. DeBano, X. Li, D. E. Wooster, M. M. Rowland, S. Burrows, and K. Frost. 2021. Capabilities and limitations of using DNA metabarcoding to study plant–pollinator interactions. *Molecular Ecology* 30(20): 5266–5297. <https://doi.org/10.1111/mec.16112>
- Baker, W. J., P. Bailey, V. Barber, A. Barker, S. Bellot, D. Bishop, L. R. Botigué, et al. 2021a. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Systematic Biology* 71: 301–319. <https://doi.org/10.1093/sysbio/syab035>
- Baker, W., S. Dodsworth, F. Forest, S. Graham, M. Johnson, A. McDonnell, L. Pokorny, et al. 2021b. Exploring Angiosperms353: An open, community toolkit for collaborative phylogenomic research on flowering plants. *American Journal of Botany* 108(7): 1059–1065.
- Banerjee, P., K. A. Stewart, C. M. Antognazza, I. V. Bunholi, K. Deiner, M. A. Barnes, S. Saha, et al. 2022. Plant–animal interactions in the era of environmental DNA (eDNA)—A review. *Environmental DNA* 4: 987–999. <https://doi.org/10.1002/edn3.308>
- Beattie, A. 1971. A technique for the study of insect-borne pollen. *The Pan-Pacific Entomologist* 47: 82.
- Belitz, M. W., E. A. Larsen, L. Ries, and R. P. Guralnick. 2020. The accuracy of phenology estimators for use with sparsely sampled presence-only observations. *Methods in Ecology and Evolution* 11: 1273–1285.
- Bell, K. L., J. Fowler, K. S. Burgess, E. K. Dobbs, D. Gruenewald, B. Lawley, C. Morozumi, and B. J. Brosi. 2017. Applying pollen DNA metabarcoding to the study of plant–pollinator interactions. *Applications in Plant Sciences* 5(6): e1600124.
- Bell, K. L., K. J. Turo, A. Lowe, K. Nota, A. Keller, F. Encinas-Viso, L. Parducci, et al. 2023. Plants, pollinators and their interactions under global ecological change: The role of pollen DNA metabarcoding. *Molecular Ecology* 32(23): 6345–6362.
- Bolger, A., and F. Giorgi. 2014. Trimmomatic: A flexible read trimming tool for Illumina NGS data. *Bioinformatics* 30: 2114–2120.
- Brosi, B. J., and H. M. Briggs. 2013. Single pollinator species losses reduce floral fidelity and plant reproductive function. *Proceedings of the National Academy of Sciences, USA* 110: 13044–13048.
- Caetano Wyler, S., and Y. Naciri. 2016. Evolutionary histories determine DNA barcoding success in vascular plants: Seven case studies using intraspecific broad sampling of closely related species. *BMC Evolutionary Biology* 16: 103.
- Calderón-Sanou, I., T. Münkemüller, F. Boyer, L. Zinger, and W. Thuiller. 2020. From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices? *Journal of Biogeography* 47(1): 193–206.
- Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10: e421.
- CaraDonna, P. J., A. M. Iler, and D. W. Inouye. 2014. Shifts in flowering phenology reshape a subalpine plant community. *Proceedings of the National Academy of Sciences, USA* 111: 4916–4921.
- CaraDonna, P. J., L. A. Burkle, B. Schwarz, J. Resasco, T. M. Knight, G. Benadi, N. Bluthgen, et al. 2021. Seeing through the static: The temporal dimension of plant–animal mutualistic interactions. *Ecology Letters* 24: 149–161.
- Cardinale, B. J., J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A. Narwani, et al. 2012. Biodiversity loss and its impact on humanity. *Nature* 486: 59–67.
- China Plant CBO Group, D.-Z. Li, L.-M. Gao, H.-T. Li, H. Wang, X.-J. Ge, J.-Q. Liu, et al. 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences, USA* 108: 19641–19646.
- Coissac, E., T. Riaz, and N. Puillandre. 2012. Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology* 21: 1834–1847.
- Collen, B., M. Ram, T. Zamin, and L. McRae. 2008. The tropical biodiversity data gap: Addressing disparity in global monitoring. *Tropical Conservation Science* 1: 75–88.
- Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.
- Drake, L. E., J. P. Cuff, R. E. Young, A. Marchbank, E. A. Chadwick, and W. O. C. Symondson. 2022. An assessment of minimum sequence copy thresholds for identifying and reducing the prevalence of artefacts in dietary metabarcoding data. *Methods in Ecology and Evolution* 13: 694–710. <https://doi.org/10.1111/2041-210X.13780>
- Encinas-Viso, F., J. Bovill, D. E. Albrecht, J. Florez-Fernandez, B. Lessard, J. Lumbers, J. Rodriguez, et al. 2023. Pollen DNA metabarcoding reveals cryptic diversity and high spatial turnover in alpine plant–pollinator networks. *Molecular Ecology* 32(23): 6377–6393. <https://doi.org/10.1111/mec.16682>
- Freitas, T. M., L. F. Montag, P. De Marco Jr., and J. Hortal. 2020. How reliable are species identifications in biodiversity big data? Evaluating the records of a neotropical fish family in online repositories. *Systematics and Biodiversity* 18: 181–191.
- Furlan, E. M., J. Davis, and R. P. Duncan. 2020. Identifying error and accurately interpreting environmental DNA metabarcoding results: A case study to detect vertebrates at arid zone waterholes. *Molecular Ecology Resources* 20(5): 1259–1276.
- Futuyma, D. J., and A. A. Agrawal. 2009. Macroevolution and the biological diversity of plants and herbivores. *Proceedings of the National Academy of Sciences, USA* 106: 18054–18061. <https://doi.org/10.1073/pnas.0904106106>
- Goulson, D. 2010. Bumblebees: Behaviour, ecology, and conservation. Oxford University Press, New York, New York, USA.
- Guertler, P., A. Eicheldinger, P. Muschler, O. Goerlich, and U. Busch. 2014. Automated DNA extraction from pollen in honey. *Food Chemistry* 149: 302–306.
- Hollingsworth, P. M., D.-Z. Li, M. van der Bank, and A. D. Twyford. 2016. Telling plant species apart with DNA: From barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371: e20150338.

- Hothorn, T., K. Hornik, M. A. van de Wiel, and A. Zeileis. 2006. A lego system for conditional inference. *American Statistician* 60: 257–263.
- Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and N. J. Wickett. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4: e1600016.
- Johnson, M. G., L. Pokorny, S. Dodsworth, L. R. Botigue, R. S. Cowan, A. Devault, W. L. Eiserhardt, et al. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 68: 594–606.
- Kassambara, A. 2023. Rstatix: Pipe-friendly framework for basic statistical tests. Website <https://CRAN.R-project.org/package=rstatix> [accessed 24 September 2025].
- Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.
- Kearns, C. A., and J. D. Thomson. 2001. The natural history of bumblebees. University Press of Colorado, Denver, Colorado, USA.
- Kress, W. J. 2017. Plant DNA barcodes: Applications today and in the future. *Journal of Systematics and Evolution* 55: 291–307.
- Lalmangaihi, R., S. Ghatak, R. Laha, G. Gurusubramanian, and N. S. Kumar. 2014. Protocol for optimal quality and quantity pollen DNA isolation from honey samples. *Journal of Biomolecular Techniques* 25(4): 92.
- Lindenmayer, D. B., and G. E. Likens. 2009. Adaptive monitoring: A new paradigm for long-term research and monitoring. *Trends in Ecology and Evolution* 24: 482–486.
- Liu, J., L. Shi, J. Han, G. Li, H. Lu, J. Hou, X. Zhou, et al. 2014. Identification of species in the angiosperm family Apiaceae using DNA barcodes. *Molecular Ecology Resources* 14: 1231–1238.
- Lu, J., F. P. Breitwieser, P. Thielen, and S. L. Salzberg. 2017. Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science* 3: e104.
- Macior, L. W. 1994. Pollen-foraging dynamics of subalpine bumblebees (*Bombus* Latr.). *Plant Species Biology* 9(2): 99–106.
- McLay, T. G., J. L. Birch, B. F. Gunn, W. Ning, J. A. Tate, L. Nauheimer, E. M. Joyce, et al. 2021. New targets acquired: Improving locus recovery from the Angiosperms353 probe set. *Applications in Plant Sciences* 9(7): e11420.
- Meyer, C., H. Kreft, R. Guralnick, and W. Jetz. 2015. Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* 6: 8221. <https://doi.org/10.1038/ncomms9221>
- Milla, L., A. Schmidt-Lebuhn, J. Bovill, and F. Encinas-Viso. 2022. Monitoring of honey bee floral resources with pollen DNA metabarcoding as a complementary tool to vegetation surveys. *Ecological Solutions and Evidence* 3: e12120. <https://doi.org/10.1002/2688-8319.12120>
- Moore, A. J., and L. Bohs. 2003. An ITS phylogeny of *Balsamorhiza* and *Wyethia* (Asteraceae: Heliantheae). *American Journal of Botany* 90: 1653–1660.
- Naimi, B., and M. B. Araujo. 2016. Sdm: A reproducible and extensible R platform for species distribution modelling. *Ecography* 39: 368–375.
- Naimi, B., N. A. S. Hamm, T. A. Groen, A. K. Skidmore, and A. G. Toxopeus. 2014. Where is positional uncertainty a problem for species distribution modelling? *Ecography* 37: 191–203.
- Ogilvie, J. E., and P. J. CaraDonna. 2022. The shifting importance of abiotic and biotic factors across the life cycles of wild pollinators. *Journal of Animal Ecology* 91: 2412–2423.
- Olesen, J. M., J. Bascompte, Y. L. Dupont, H. Elberling, C. Rasmussen, and P. Jordano. 2010. Missing and forbidden links in mutualistic networks. *Proceedings of the Royal Society B: Biological Sciences* 278(1706): 725–732.
- Pearse, W. D., C. C. Davis, D. W. Inouye, R. B. Primack, and T. J. Davies. 2017. A statistical estimator for determining the limits of contemporary and historic phenology. *Nature Ecology and Evolution* 1: 1876–1882.
- Pimm, S. L., C. N. Jenkins, R. Abell, T. M. Brooks, J. L. Gittleman, L. N. Joppa, P. H. Raven, et al. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. *Science* 344(6187): e1246752.
- Pornon, A., C. Andalo, M. Burrus, and N. Escaravage. 2017. DNA metabarcoding data unveils invisible pollination networks. *Scientific Reports* 7: e16828. <https://doi.org/10.1038/s41598-017-16785-5>
- Pusalkar, P. K., and D. K. Singh. 2015. Taxonomic rearrangement of *Arenaria* (Caryophyllaceae) in Indian Western Himalaya. *Journal of Japanese Botany* 90: 77–91.
- Pyke, G. H. 1982. Local geographic distributions of bumblebees near Crested Butte, Colorado: Competition and community structure. *Ecology* 63: 555–573.
- R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website <http://www.R-project.org/> [accessed 24 September 2025].
- Ruete, A. 2015. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. *Biodiversity Data Journal* 3: e5361. <https://doi.org/10.3897/BDJ.3.e5361>
- Ruppert, K. M., R. J. Kline, and M. S. Rahman. 2019. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation* 17: e00547.
- Sadeghian, S., S. Zarre, R. K. Rabeler, and G. Heubl. 2015. Molecular phylogenetic analysis of *Arenaria* (Caryophyllaceae: Tribe Arenarieae) and its allies inferred from nuclear DNA internal transcribed spacer and plastid DNA rps16 sequences. *Botanical Journal of the Linnean Society* 178: 648–669.
- Sennikov, A. N., and A. Kurtto. 2017. A phylogenetic checklist of *Sorbus* s.l. (Rosaceae) in Europe. *Memoranda Societatis Pro Fauna et Flora Fennica* 93: 1–78.
- Smith, B. E., M. K. Johnston, and R. Lücking. 2016. From GenBank to GBIF: Phylogeny-based predictive niche modeling tests accuracy of taxonomic identifications in large occurrence data repositories. *PLoS ONE* 11(3): e0151232. <https://doi.org/10.1371/journal.pone.0151232>
- Socolar, J. B., J. J. Gilroy, W. E. Kunin, and D. P. Edwards. 2016. How should beta-diversity inform biodiversity conservation? *Trends in Ecology and Evolution* 31: 67–80.
- Spooner, D. M. 2009. DNA barcoding will frequently fail in complicated groups: An example in wild potatoes. *American Journal of Botany* 96(6): 1177–1189. <https://doi.org/10.3732/ajb.0800246>
- Tange, O. 2021. GNU parallel 20220322 (savannah). Available at Zenodo repository <https://doi.org/10.5281/zenodo.6377950> [accessed 24 September 2025].
- Tylianakis, J. M., E. Laliberté, A. Nielsen, and J. Bascompte. 2010. Conservation of species interaction networks. *Biological Conservation* 143: 2270–2279.
- Weitemier, K., S. C. Straub, R. C. Cronn, M. Fishbein, R. Schmickl, A. McDonnell, and A. Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2: e1400042.
- Westgate, M. J., G. E. Likens, and D. B. Lindenmayer. 2013. Adaptive management of biological systems: a review. *Biological Conservation* 158: 128–139.
- Wood, D. E., J. Lu, and B. Langmead. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biology* 20: 257.
- Woodard, S. H., and S. Jha. 2017. Wild bee nutritional ecology: Predicting pollinator population dynamics, movement, and services from floral resources. *Current Opinion in Insect Science* 21: 83–90.
- Yoccoz, N. G., J. D. Nichols, and T. Boulinier. 2001. Monitoring of biological diversity in space and time. *Trends in Ecology and Evolution* 16: 446–453.
- Zinger, L., A. Bonin, I. G. Alsos, M. Bálint, H. Bik, F. Boyer, A. A. Chariton, et al. 2019. DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology* 28: 1857–1862. <https://doi.org/10.1111/mec.15060>
- Zuntini, A. R., T. Carruthers, O. Maurin, P. C. Bailey, K. Leempoel, G. E. Brewer, N. Epiawalage, et al. 2024. Phylogenomics and the rise of the angiosperms. *Nature* 629: 843–850. <https://doi.org/10.1038/s41586-024-07324-0>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Supplemental methods.

Appendix S2. Methods workflow.

Appendix S3. Independent variables used in the species distribution models.

Appendix S4. Site map.

Appendix S5. All pollen reference slides used to establish morphotypes.

Appendix S6. Pollen key developed and used to identify morphotypes with microscopy.

Appendix S7. Molecular reference specimen table.

Appendix S8. Reads per locus for additional reference samples.

Appendix S9. All species in the sequence databases.

Appendix S10. Post-BLAST sequence alignment process.

Appendix S11. Pollen morphotype richness rarefaction curves.

Appendix S12. Pollen morphotype abundance rarefaction curves.

Appendix S13. Reads per locus for corbicula samples.

Appendix S14. Comparison of Kraken2, Bracken, and BLAST.

How to cite this article: Benkendorf, R. C., E. J. Woodworth, P. J. CaraDonna, J. E. Ogilvie, S. Taddeo, and J. B. Fant. 2025. Improving plant DNA metabarcoding accuracy with ecological filters and Angiosperms353: Field and pollen microscopy validation. *Applications in Plant Sciences* 13(5): e70026. <https://doi.org/10.1002/aps3.70026>