**Title**: Protocol to increase accuracy and fidelity of metagenomic barcoding using Angiosperms353: a case study using pollen loads from wild bumble bees

**Running headline**: Protocol to increase metabarcoding accuracy and fidelity

**Authors and Affiliations**: Reed Clark Benkendorf1,2\*, Emily J. Woodworth1,2 , Paul J. CaraDonna1,2,3, Jane E. Ogilvie3, Sophie Taddeo1,2,4 , Jeremie B. Fant1,2

1 Chicago Botanic Garden, Glencoe, Illinois 60022, USA

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

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

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

\*Corresponding author

Reed Benkendorf *https://orcid.org/0000-0003-3110-6687*

Paul CaraDonna *https://orcid.org/0000-0003-3517-9090*

Jeremie Fant *https://orcid.org/0000-0001-9276-1111*

Jane Ogilvie *https://orcid.org/0000-0001-8546-0417*

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

Emily Woodworth

**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, E. Nelson for tissue loans; J. Reithel and RMBL staff for access to field sites and logistical support.

**AUTHOR CONTRIBUTIONS**: R.C.B., J.E.O., J.B.F., 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.

**DATA AVAILABILITY STATEMENT**: The queries required to download all data used in this project are located in the lead authors Orcid. All novel sequencing data are located at NCBI Short Read Archive (SRA) under PRJNA1093153 (metagenomic) and PRJNA1083211 (reference).

# Abstract

1) DNA metabarcoding has been successful for the rapid identification of species in complex ecological assemblages, including identifying interspecific interactions among species. However, advances in metabarcoding within the plant kingdom have been hampered due to a lack of universal gene regions that work across all taxa, which limit the applications of eDNA and metagenomics in ecology more broadly.

2) To circumvent these limitations, we propose a holistic spatio-temporal approach that combines multi-gene barcoding with existing plant occurrence databases, species distribution models, and phenological analyses to generate a shortened list of candidate species to increase metabarcoding accuracy and computing efficiency. To validate the accuracy and efficiency of our methodological framework, we compared the results of the DNA metabarcoding from pollen loads of several species of wild bumble bees to in-depth, long-term field observations of bee-plant interactions, along with expert-led pollen identification.

3) We show that DNA metabarcoding of the plant species included in bumble bee pollen loads was most accurate when combined with a candidate taxa list of plant species flowering in the community when the bumble bees were foraging, which improved the accuracy and taxonomic precision of 77.5% of samples.

4) With the recent proliferation of species occurrence and phenology data in tandem with advances in computing and software development, we believe that spatio-temporal filtering provides a simple approach for interpreting metagenomic studies globally. Additionally, we demonstrate that the Angiosperms 353 probes (developed for phylogenomics) offer significant promise for metagenomics projects globally, including metabarcoding to reveal species interactions within complex communities. Further, our approach demonstrates that integrating DNA metabarcoding is most accurate and powerful when combined with local ecological data.

**Keywords**: barcoding, plant-pollinator interactions, phenology, species distribution models, angiosperms353, metagenomics

# Introduction

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 (Pimm et al. 2014, Socolar et al. 2016, Cardinale et al. 2012). The interactions that occur among species are critical components of ecosystem stability, quality, and function (Futuyma & Agrawal 2009), therefore biodiversity monitoring must consider both species and their interactions (Lindemeyer & Likens 2009, Westgate et al. 2013). However, monitoring species interactions is challenging and remains inconsistent in time and space (Yoccoz et al. 2001, Lindemeyer & Likens 2009), thereby leaving many regions, ecosystems, and taxa under-observed (Collen et al. 2008; Meyer et al. 2015; Ruete 2015). The resulting data gaps impact the capacity to identify and implement robust conservation interventions (Tylianakis et al. 2010). DNA barcoding and metabarcoding (i.e., 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 the presence of species as well as their interactions. These approaches have immense potential because they can help increase our ability to monitor more ecosystems in greater depth. However, the results of these studies may often be spurious and can lead to misleading ecological inferences.

New and emerging genomic data sets can enable the use of metabarcoding to resolve a diverse array of questions relevant to our applied and conceptual understanding of ecology and evolution (Kress 2017, Hollingsworth et al. 2016). Such datasets include the Plant and Fungal Tree of Life (PAFTOL)—the largest 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 (A353) probes (Baker et al. 2021a, Baker et al. 2021b, Johnson et al. 2019). These publicly available data provide a phylogenetically comprehensive backbone for plant metagenomic barcoding, but they still only contain a fraction of all known plant species. Despite their potential, the enormity of these genomic datasets requires considerable computational power, which might limit their widespread use in metabarcoding as it requires a large amount of sequencing to generate a sufficiently detailed reference database. 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 (Liu et al. 2014, CBOL Group et al. 2011, Coissac et al. 2012).

To address current metabarcoding challenges, we present a holistic approach that generates a shorter and more appropriate reference list for identifying candidate taxa from barcode data. This can help increase the accuracy and efficiency of metabarcoding results in plants. This spatio-temporal 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. This list is then leveraged to identify target plant tissue collection, reduce the size of a reference sequence database, increase computation speed and efficiency, and filter metabarcoding results for accuracy.

Although metabarcoding is often proposed as a rapid and easy method for monitoring biodiversity in environmental samples, the accuracy and consistency of this approach are rarely tested. Hence, to demonstrate the usefulness of this process, we apply our methodological approach to the metabarcoding of corbiculae pollen loads of several wild bumble bee species. We then validate the accuracy of our approach to identify plant species before and after applying spatial and temporal filtering by comparing the molecular data with both morphological pollen identifications and multi-year field observations.

# Methods

## Overview of methodological approach

Our approach has three main parts. Given that there are over 350,000 known vascular plant 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 modelling 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 (see “Collecting representative samples”) to improve the odds that interacting species are detected. We test this metagenomic and spatio-temporal 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.

A diagram of a plant

Description automatically generated with medium confidence

**Figure 1**. Key methodological steps in our approach to improve metabarcoding species identification, with corresponding species filtering from our case study.

## Potential candidate taxa list

### Survey databases to generate a regional 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.org), herbaria consortia, and community science datasets (e.g., iNaturalist). 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.

### Spatial filtering for likely taxa via species distribution models

Given the coarseness of regional occurrence data, 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 (see SI Table S1 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).

### Temporal filtering for likely taxa using phenology data

Filtering for plant species 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, Budburst) provide data that can elucidate coarse phenological patterns to retain only plant species with phenophases that align with the sampling period.

## Using spatio-temporal filtering in metagenomics

### Collecting representative samples

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 analyses down the line. However, given the cost of sequencing, it is best to minimize reference sequence generation.

While there are numerous next-generation sequencing approaches available for metagenomics, 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 (100-300bp) 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 loci (HybPiper; Johnson et al., 2016; MAFFT, Katoh and Standley, 2013). These consensus sequences of each potential taxon can then added to the existing reference sequence library used to screen metabarcoding data

## Metabarcoding

Isolation of DNA from environmental samples will be be similar to 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 compared to the known sequences within the custom-made sequence reference library, using any one of a variety of alignment methods.

# Case Study

### 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* spp.) collected from the Rocky Mountain Biological Laboratory in Colorado, United States (SI Figure S1). We collected pollen loads from wild foraging queen bees between May and July of 2015 at six permanent study sites where we monitor bumble bees and floral resources (details in *Anonymized reference*). To harvest the pollen loads, we captured queens in an insect net, transferred them into a restraining device (Kearns et al. 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*, *B. bifarius*, *B. californicus*, *B. flavifrons*, *B. nevadensis*, and *B. rufocinctus* (Pyke 1982; Ogilvie & CaraDonna 2022).

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 compare 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.

### 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 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 modelling

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 (GLMs), and generalized additive models (GAMs) (see SI Table S2 for complete description) (Naimi & Araujo 2016). We used 26 environmental variables at a 30 m resolution to construct the SDMs (SI Table 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 & 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 phenological records including the dates when 10% of individuals had begun flowering, when 50% were flowering, and when 90% of individuals had flowered. We used the initiation and cessation dates, respectively, as the effective start and end of flowering and consequently assessed the number of species likely to flower in each week of the 11-week period of queen be 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 & 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 & 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 (SI Figure 2).

### Metagenomics: additional plant tissue collection and extraction

We supplemented the available reference sequences by successfully sequencing 15 additional species collected around our bumble bee study sites (see SI Table 4). Plant genomic DNA was isolated from ~ 1 cm2 of leaf tissue from silica-gel dried or herbarium material using a modified cetyltrimethylammonium (CTAB) protocol (Doyle & Doyle 1987) that included two chloroform washes.

### Pollen DNA extraction

We extracted DNA from 54 corbiculae pollen load samples using a modified Cetyltrimethylammonium bromide (CTAB) method (Lalhmangaihi et al. 2014; Guertler et al. 2014) which included using a 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, UK) to solute ratio, eluted in 0.5x TE, and the eluent allowed to reduce by half volume in ambient conditions. DNA was quantified using a Qubit fluorometer.

### 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 ½ volume of reagents and ¼ enzyme mix for 40 minutes at 37°C, with an input of 500 ng cleaned DNA. Adapter Ligation and PCR enrichment were performed with ½ 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 Angiosperms 353 probe kit VkitV.4 (Arbor Biosciences myBaits Target Sequence Capture Kit) following the manufacturer’s protocol. Sequencing was performed using an Illumina mi-Seq with 150-bp paired-end reads (NUSeq Core, Chicago, Illinois).

### Bioinformatics

Sequences were processed using Trimmomatic, which removed sequence adapters, clipped the first 3 bp, discarded reads less than 36 bp, and removed reads if their average PHRED score dropped beneath 20 over a 5 bp window (Bolger & 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 by gene-regions. For the plant samples, contigs could be further processed to generate a consensus sequence for each of the 353 gene-regions. To process the pollen sequence of contigs, we created a custom Kraken2 database by downloading representative species or genera from our spatio-temporal filtered taxa list. This database was built and run using default parameters (SI Table S5). Following Kraken2, Bracken 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 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 predicted species presence from spatial modelling, modelled flowering time from temporal modelling, and taxonomy from existing sources. We used a sequential process that reassigned sequences based on binary combinations of the factors above (SI, Table 6). 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 which 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 (Sadeghian et al. 2015, Sennikov & Kurtto 2017, Pusalkar & Singh 2015, Moore & Bohs 2003). 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 metagenomic approach

We compared the sequences classified by molecular methods with the results of direct bumble bee floral visitation data (hereafter “field”; Ogilvie & 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 in 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. Hence any plant species recorded by one method that represented less than 1-2 observed visits, or less than 1% of the total number of pollen grains or molecular species confirmation, we designated these interactions as potential “false positives”.

# Results

## Field observations and microscopic pollen identification

During field observations conducted in 2015-2022, we observed 1,424 overwintered queen-pollen foraging interactions. Bumble bee queens visited 40 plant species from 35 genera and 16 families, with each bumble bee species visiting between 8-20 plant species (x̄ = 14.86, median = 17). Our microscopic analyses of bumble bee corbiculae pollen loads identified 28 pollen grain morphotypes based on 10 characters (SI Figure 2) (number of grains: x̄ = 3,319, median = 1,891, range = 514-19,924; morphotype richness, x̄ = 4.5, median = 4, range = 1-9; n = 37 samples) (SI Figures 7 & 8).

## Spatial filtering

Our spatial filtering using species distribution models and logistic regression reduced the plant species list from 1,295 to 426 species (Figure 1). All ensembled models (i.e., combination of models using different algorithms; n = 968) had an accuracy of 0.84 (95% CI 0.836 - 0.844; kappa = 0.68, p < 0.001, sensitivity = 0.80, specificity = 0.87, AUC = 0.92). The 493 machine learning ensembles accurately predicted the presence of 362 species (65.3% of all species modelled) 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 modelled) and failed to meet habitat suitability thresholds for 34 species (6.1 % species) 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 ensembled 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 section) 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 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 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 onset (p < 0.0001, tau = 0.61), peak (p < 0.0001, tau = 0.65), and cessation of flowering (p < 0.0001, tau = 0.49).

A graph of a number of species

Description automatically generated

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

## 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 sequence varied widely (x̄ = 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 (x̄ = 305.5, median = 331). The number of reads per loci from across all samples had a range of 178 - 506,653 (x̄ = 20,688, median = 12,616) (SI Figure 4). A total of 10,682,538 reads were matched using Kraken and 10,160,768 of these reads were matched using Bracken, and 7,549,608 reads were matched using BLAST. Based on a subjective review of the three classifiers (SI Figure 5), we chose BLAST as the classification method which yielded the most probable results, and its values were used for all subsequent analyses. Of the top ten taxa that were identified by BLAST across all pollen samples, 55.4% of the reads were classified to a species, and 41.9% of the reads were classified to genus.

A diagram of a bar graph

Description automatically generated with medium confidence

***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 abundance reads per sample.*

## Validating molecular methods with 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 81 bumble bee and plant genera combinations identified across all three methodological approaches (n = 182), there were 28 bee species-plant genera combinations which were exclusively observed in the field (i.e., bee x 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 interactions (*Calochortus* (*B. appositus*), *Dodecatheon* (*B. bifarius*), *Lonicera* (*B. flavifrons*), *Linum* (*B. appositus*), *Claytonia* (*B. bifarius*), *Valeriana* (*B. rufocinctus*), this was only bumble bee species seen visiting this genus and in all cases the frequency of visits were low (>2% of total visits observed) (Figure 5). By contrast, there were five of these bumble bee-plant combinations that did not have corresponding lab-based data (*Cirsium* (*B. flavifrons*), *Hydrophyllum* (*B. nevadensis*), *Lathyrus* (*B. appositus* and *B. californicus*), *Vicia* (*B. rufocinctus*), and *Pedicularis* (*B. californicus*)), those plant genera were visited by other bumble bee species, and for those bee species, the plant pollen was verified using the microscope or using molecular approach (Figure 5). 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*, *Senecio*, *Frasera*, *Aconitum*, and *Erythronium*), were visited by multiple bumble bee species in our dataset but pollen was never 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 uncommon interactions, or perhaps heterospecific pollen deposited by other floral visitors. Regardless, as we pooled all data across individuals, dates, and sites, we assumed that these still represent infrequent visits. Of the genera observed in the pollen microscopy and molecular data but not in field observations, most were in very low numbers (<1% of data), suggesting few visits (Figure 4). Of the bumble bee-plant genera combinations that were confirmed in two out 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), although all were in very low frequencies, suggesting bees might be picking up small amounts when foraging on *Asteraceae* flowers for nectar. The molecular technique produced by far the most unique bee-plant genera combinations, accounting for over one-third of all data points for the three methods combined (Figure 5). For the most part, these were often in very low frequency and suggest they might represent false positives or low-level contamination.

A diagram of a plant

Description automatically generated

***Figure 4****. Comparison of the interactions detected between bumble bee queens and plant genera among the different approaches (bee-plant field observations, pollen microscopy of pollen loads, and molecular metabarcoding of pollen loads). The rings indicate different lines of evidence for binary interactions between Bombus and that genus.*

A diagram of a molecule

Description automatically generated

***Figure 5****: 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).*

# Discussion

We demonstrate that our spatio-temporal filtering approach increased the accuracy and efficiency of species identification using traditional metagenomic and barcoding approaches. 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)— can enhance the accuracy and precision of metabarcoding results. In combination with a bait capture approach, we show moderate to high concordance of the 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 Angiosperm 353 baits allows for higher taxonomic resolution, often to species, although creating a regionally relevant candidate taxa list was critical for assessing if 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 rather, the quantity of sequence data available. In particular, the enormous amount of global data within these databases increases processing time and the potential for spurious matches. We demonstrate that our custom sequence database approach can substantially improve the processing time and accuracy of metabarcoding, with taxonomic identification accuracy improving for 77.5% of our samples after applying our spatio-temporal filtering approach.

The accuracy of the metabarcoding, when compared to in-depth field and microscopy data, confirms that our spatio-temporal filtering approach can be a valuable tool for testing hypotheses related to ecological interactions across various taxonomic groups and study areas. In all cases, the most common plant species present in the corbicula loads were correctly identified by all three techniques. However, the other methods (i.e., field, molecular) were critical in enabling inferences about the identity of species identified through microscopy, as those could only be identified at the genus level. What was somewhat surprising was that even the relative frequency of these plants remained consistent across the three methodological approaches, despite the fact they measure slightly different biological processes (e.g., flower visitation versus pollen collected). This suggests that all three techniques can accurately identify the species and frequencies for most of the common interactions in this dataset, although the field observations were essential for validating the potential plant species in the other two methods.

Validating our approach with a case study revealed some insights into bumble bee foraging ecology. Most obviously, 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 often focus on collecting pollen from a single or a few species during a foraging trip, supporting earlier work (e.g., Macior 1994). This foraging tendency may be related to a combination of flower constancy and learning constraints (e.g., Raine & Chittka 2007), and nutritional preferences in pollen (e.g., Hanley et al. 2008). Our results reiterate the need to maintain plants within *Fabaceae*, *Boraginaceae*, and *Ranunculaceae* in Western North American landscapes as pollen resources to support *Bombus* (Goulson 2010).

Other ecological insights came from discrepancies among the three methods. For example, there were exclusively field-observed bumble bee visits to flowers, sometimes in high frequency, such as to *Frasera speciosa* and many *Asteraceae* genera. Such a pattern is indicative of flower visits for nectar without pollen collection—a phenomenon that can otherwise be overlooked without examining pollen loads. By contrast, there were several situations in which there was pollen identified in corbicula loads but the interaction was not observed in the field, such as the willow family (*Salicaceae*). Given the available evidence, this observation could represent pollen loads containing *Salix* spp.(willow) or *Populus tremuloides* (trembling aspen), although molecular evidence suggests that *Populus* may be more likely. In either case, this observation generates two hypotheses. First, we know that willows (*Salix* spp., *Salicaceae*) are an important spring resource for bumble bee queens: many willow species flower profusely around our field sites and we observe bumble bees on them, but the plants are patchy and often tall, and the bees can be difficult to observe and identify from afar. 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 in the field data. Second, if the pollen belongs to *Populus tremuloides*, this would indicate a novel interaction that we have not observed in the field and that might reflect an opportunistic pollen foraging from wind-pollinated aspen trees.

Overall, our data elucidated a single aspect of bumble bee interactions. The comparison across three methodological approaches, along with the use of a filtered species list, highlights limitations to each of the three approaches (field, microscope, metabarcoding) that could lead to the identification of unlikely biological interactions if used on their own. First, all three techniques identified unique bumble bee-plant species combinations. In most cases, these unique combinations were at very low relative frequencies, suggesting that even if they represent real interactions, they are likely rare and of limited biological importance. Because they were unique to only one of three approaches, we assumed these interactions to be spurious, representing background noise inherent to that dataset. Here the distinction between the two sides of biological probability was somewhat clear, given the enormity of this, and other observational interaction data sets. However, in many systems, we suspect that the combination of two approaches may be required to winnow artifacts from reality.

A key finding both here, and in previous metabarcoding studies, is that molecular methods generate many spurious results which can be misleading (Drake et al. 2021). Given the growing size of molecular datasets, the use of short reads in next-generation sequencing, and that barcoding approaches use markers of highly conserved regions, it is increasingly likely that multiple matches will be retrieved, some of which are inevitably incorrect. This highlights the value of an environmentally realistic species list to cross-validate the taxa matched. Although metagenomic barcoding 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-positive interactions made up only a fraction of the total dataset (0-6%), their effect is outsized. Most of these spurious matches were for species in families that also contained true interacting partners, but which we believe arose from the alignment of highly conserved gene regions. This phenomenon was even observed within genera, as the species identified occasionally included taxa that were geographically unlikely (e.g., *Delphinium gracile*, a taxon endemic to the Mediterranean basin), even though the geographically correct species was still the most abundant in the dataset. Accordingly, research programs that seek to use metagenomic barcoding methods must ensure that the appropriate resources are dedicated to fieldwork to anchor the data sets in biological reality.

# Conclusion

Our spatio-temporal 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 that can be useful, such as facilitating the monitoring of plant-animal interactions (Banerjee et al. 2022), particularly those that are challenging to visually observe. It may also be used in a variety of other pollen sample identification applications (Bell et al. 2017, Bell et al. 2019). Further afield it can be used in the identification of plants in animal diet components (e.g., pollen, frass, scat), detecting noxious weeds in water samples, and assessing soil seed banks. Finally combining metabarcoding with SDMs could 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 which are difficult to distinguish without reproductive organs, or individuals at multiple reproductive phases.

**References**

Allouche, O., Tsoar, A. and Kadmon, R. (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., Govaerts, R., Nic Lughadha, E., Onstein, R.E., Smith, R.J. and Zizka, A. (2023), Why plant diversity and distribution matter. New Phytol, 240: 1331-1336.<https://doi.org/10.1111/nph.19282>

Araujo, M.B. & New, M. (2007). Ensemble forecasting of species distributions. *Trends in ecology & evolution*, 22, 42–47.

Baker, W.J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., Botigué, L.R., Brewer, G., Carruthers, T., Clarkson, J.J., Cook, J., Cowan, R.S., Dodsworth, S., Epitawalage, N., Françoso, E., Gallego, B., Johnson, M.G., Kim, J.T., Leempoel, K., Maurin, O., Mcginnie, C., Pokorny, L., Roy, S., Stone, M., Toledo, E., Wickett, N.J., Zuntini, A.R., Eiserhardt, W.L., Kersey, P.J., Leitch, I.J. & Forest, F. (2021a). A Comprehensive Phylogenomic Platform for Exploring the Angiosperm Tree of Life. *Systematic Biology*, 71, 301–319. Retrieved from<https://doi.org/10.1093/sysbio/syab035>

Baker, W., Dodsworth, S., Forest, F., Graham, S., Johnson, M., McDonnell, A., Pokorny, L., Tate, J., Wicke, S. & Wickett, N. (2021b). [Exploring Angiosperms353: An open, community toolkit for collaborative phylogenomic research on flowering plants](https://doi.org/10.1002/ajb2.1703). *American Journal of Botany*, 108.

Banerjee, P., Stewart, K. A., Antognazza, C. M., Bunholi, I. V., Deiner, K., Barnes, M. A., Saha, S., Verdier, H., Doi, H., Maity, J. P., Chan, M. W., Chen, C. Y. (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., Larsen, E.A., Ries, L. & Guralnick, R.P. (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., Burgess, K.S., Botsch, J.C., Dobbs, E.K., Read, T.D. & Brosi, B.J. (2019). Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. *Molecular Ecology*, 28, 431–455.

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

Bolger, A. & Giorgi, F. (2014). Trimmomatic: A flexible read trimming tool for illumina NGS data. *Bioinformatics*, 30, 2114–2120.

Brosi, B.J. & Briggs, H.M. (2013). Single pollinator species losses reduce floral fidelity and plant reproductive function. *Proceedings of the National Academy of Sciences*, 110, 13044–13048.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. (2009). BLAST+: Architecture and applications. *BMC bioinformatics*, 10, 1–9.

CaraDonna, P.J., Burkle, L.A., Schwarz, B., Resasco, J., Knight, T.M., Benadi, G., Bluthgen, N., Dormann, C.F., Fang, Q., Frund, J. & others. (2021). Seeing through the static: The temporal dimension of plant–animal mutualistic interactions. *Ecology Letters*, 24, 149–161.

CaraDonna, P.J., Iler, A.M. & Inouye, D.W. (2014). Shifts in flowering phenology reshape a subalpine plant community. *Proceedings of the National Academy of Sciences*, 111, 4916–4921.

Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P., Narwani, A., Mace, G.M., Tilman, D., Wardle, D.A. & others. (2012). Biodiversity loss and its impact on humanity. *Nature*, 486, 59–67.

CBOL Group, C.P.B., Li, D.-Z., Gao, L.-M., Li, H.-T., Wang, H., Ge, X.-J., Liu, J.-Q., Chen, Z.-D., Zhou, S.-L., Chen, S.-L. & others. (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*, 108, 19641–19646.

Coissac, E., Hollingsworth, P.M., Lavergne, S. & Taberlet, P. (2016). From barcodes to genomes: Extending the concept of DNA barcoding.

Coissac, E., Riaz, T. & Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular ecology*, 21, 1834–1847.

Collen, B., Ram, M., Zamin, T. & McRae, L. (2008). The tropical biodiversity data gap: Addressing disparity in global monitoring. *Tropical Conservation Science*, 1, 75–88.

Drake, L. E., Cuff, J. P., Young, R. E., Marchbank, A., Chadwick, E. A., & Symondson, W. O. C. (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>

Doyle, J.J. & Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, 19, 11–15.

Freitas, T.M., Montag, L.F., De Marco Jr, P. & Hortal, J. (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.

Futuyma DJ, Agrawal AA. (2009) Macroevolution and the biological diversity of plants and herbivores. *Proceedings of the National Academy of Sciences* 106, 18054-18061, [https://doi.org/10.1073/pnas.090410610](https://doi.org/10.1073/pnas.0904106106)

Goulson, D. *Bumblebees: behaviour, ecology, and conservation*. Oxford University Press, 2010.

Hanley, M.E., Franco, M., Pichon, S., Darvill, B. & Goulson, D. (2008). Breeding system, pollinator choice and variation in pollen quality in British herbaceous plants. *Functional Ecology*, 592–598.

Hollingsworth, P.M., Li, D.-Z., Bank, M. van der & Twyford, A.D. (2016). Telling plant species apart with DNA: From barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371, 20150338.

Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N.J. & Wickett, N.J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in plant sciences*, 4, 1600016.

Johnson, M.G., Pokorny, L., Dodsworth, S., Botigue, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T. & others. (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.

Katoh, K. & Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular biology and evolution*, 30, 772–780.

Kearns, C.A., Thomson, J.D. & others. (2001). *Natural history of bumblebees*. University Press of Colorado.

Kress, W.J. (2017). Plant DNA barcodes: Applications today and in the future. *Journal of systematics and evolution*, 55, 291–307.

Lindenmayer, D.B. & Likens, G.E. (2009). Adaptive monitoring: A new paradigm for long-term research and monitoring. *Trends in ecology & evolution*, 24, 482–486.

Liu, J., Shi, L., Han, J., Li, G., Lu, H., Hou, J., Zhou, X., Meng, F. & Downie, S.R. (2014). Identification of species in the angiosperm family Apiaceae using DNA barcodes. *Molecular ecology resources*, 14, 1231–1238.

Lu, J., Breitwieser, F.P., Thielen, P. & Salzberg, S.L. (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., Birch, J.L., Gunn, B.F., Ning, W., Tate, J.A., Nauheimer, L., Joyce, E.M., Simpson, L., Schmidt-Lebuhn, A.N., William J & others. (2021). New targets acquired: Improving locus recovery from the Angiosperms353 probe set. *Applications in plant sciences*, 9.

Meyer, C., Kreft, H., Guralnick, R. *et al.* Global priorities for an effective information basis of biodiversity distributions. *Nat Commun* **6**, 8221 (2015). https://doi.org/10.1038/ncomms9221

Moore, A.J. & Bohs, L. (2003). An ITS phylogeny of Balsamorhiza and Wyethia (Asteraceae: Heliantheae). *American Journal of Botany*, 90, 1653–1660.

Naimi, B. & Araujo, M.B. (2016). [Sdm: A reproducible and extensible r platform for species distribution modelling](https://doi.org/10.1111/ecog.01881). *Ecography*, 39, 368–375.

Naimi, B., Hamm, N. a.s., Groen, T.A., Skidmore, A.K. & Toxopeus, A.G. (2014). [Where is positional uncertainty a problem for species distribution modelling?](https://doi.org/10.1111/j.1600-0587.2013.00205.x) *Ecography*, 37, 191–203.

Olesen, J.M.,. Bascompte, J., Dupont, Y.L.., Elberling, H., Rasmussen, C., Jordano, P. (2010). Missing and forbidden links in mutualistic networks. Proceedings of the Royal Society B: Biological Sciences

Ogilvie, J.E. & CaraDonna, P.J. (2022). The shifting importance of abiotic and biotic factors across the life cycles of wild pollinators. *Journal of Animal Ecology*.

Pearse, W.D., Davis, C.C., Inouye, D.W., Primack, R.B. & Davies, T.J. (2017). A statistical estimator for determining the limits of contemporary and historic phenology. *Nature Ecology & Evolution*, 1, 1876–1882.

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., ... & Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *science*, *344*(6187), 1246752.

Pusalkar, P.K. & Singh, D.K. (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.

Raine NE, Chittka L (2007) The Adaptive Significance of Sensory Bias in a Foraging Context: Floral Colour Preferences in the Bumblebee *Bombus terrestris*. PLOS ONE 2(6): e556.<https://doi.org/10.1371/journal.pone.0000556>

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., Kline, R.J. & Rahman, M.S. (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., Zarre, S., Rabeler, R.K. & Heubl, G. (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. & Kurtto, A. (2017). A phylogenetic checklist of Sorbus s.l. (Rosaceae) in Europe.

Smith BE, Johnston MK, Lücking R (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., Gilroy, J.J., Kunin, W.E. & Edwards, D.P. (2016). How should beta-diversity inform biodiversity conservation? *Trends in Ecology & Evolution*, 31, 67–80.

Tange, O. (2021). GNU parallel 20220322 (savannah). Retrieved from <https://doi.org/10.5281/zenodo.6377950>

Tylianakis, J.M., Laliberté, E., Nielsen, A. & Bascompte, J. (2010). Conservation of species interaction networks. *Biological conservation*, 143, 2270–2279.

Weitemier, K., Straub, S.C., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A. & Liston, A. (2014). Hyb-seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in plant sciences*, 2, 1400042.

Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001). Monitoring of biological diversity in space and time. *Trends in ecology & evolution*, 16, 446–453.