**Title:** Protocol to increase accuracy and fidelity of Pollen Meta-genomic Barcoding using Angiosperms353: Case study using Corbiculae loads from wild Bumble Bees from RMBL

**Short title:**

**Authors and Affiliations: (Alphabetical right now)**

Reed Benkendorf

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 Wordsworth :

**Abstract**

1. DNA Barcoding has been successful for the rapid identification of species in complex ecological assemblages. However, metabarcoding in the plant kingdom has been limited due to a lack of universal gene regions that works across all taxa, limiting the applications of eDNA and metagenomics in ecology.
2. To circumvent these limitations, we propose using more holistic approach which combines a multiple gene approach that incorporates existing plant occurrence databases, species distribution models, and phenology analyses to generate a list of likely candidate species to increase accuracy. Such an effort can help guide plant sample collection and library construction, particularly for projects with limited resources or local taxonomic knowledge. In addition, we demonstrate that building a custom candidate plant list based on known occurrences, ecological requirements, and phenology can improve computing efficiency, speed, and accuracy of metabarcoding. We tested this approach to identify the plant species contributing to the corbiculae pollen loads of bumble bees. To validate the efficiency of this approach we compare results of the DNA barcoding to expert-led pollen identification and field observations of plant-pollinator interactions that generated these pollen loads.
3. We show that the Angiosperms 353 probes, developed for phylogenomics, and which are currently being used in the largest ever plant systematic endeavor, offer significant promise to metagenomic approaches around the globe. The DNA barcoding of bumble bee corbiculae pollen loads was most accurate when combined with knowledge of what plant species were flowering in the plant community when they were collected.
4. Supplementing DNA barcoding data with ecological context is most accurate and powerful.

**Introduction**

Large scale species loss (Joppa et al. 2011; Pimm et al. 2014) and biotic homogeneization (REF), and the impacts of these processes on ecosystem functions and human well-being (Cardinale et al. 2012), have inspired numerous calls for a more consistent monitoring of ecosystems and their diversity (Pereira and Cooper 2006; Pereira et al. 2013). Monitoring ecological communities structure and interaction will be critical to informing and prioritizing conservation efforts (REF), particularly considering the large proportion of know species currently threaten or likely to be threaten in the near future (Pimm et al. 2014; Pereira et al. 2010). However, ecosystem monitoring remains inconsistent in time and space (Yoccoz et al. 2011) as it requires substantial resources, thereby leaving several regions, ecosystems, and even species under-observed (Pereira et al. 2012; Collen et al. 2008; Meyer et al. 2015; Ruete 2015).

Species interactions are a critical component of ecosystem stability and therefore biodiversity conservation (Soltis *et al.* ([2019](#ref-soltis2019darwin)), Futuyma & Agrawal ([2009](#ref-futuyma2009macroevolution)), Voje *et al.* ([2015](#ref-voje2015role)), Weber *et al.* ([2017](#ref-weber2017evolution)), Hembry & Weber ([2020](#ref-hembry2020ecological))). Species interactions, which span from positive (e.g., symbiosis, mutualistic) to negative (e.g., predation, competition), shape species distribution (Bascompte 2009; Wisz et al. 2013) and tolerance to environmental stressors (REF). They impact ecosystem stability and diversity (Agrawal *et al.* ([2007](#ref-agrawal2007filling)), Valiente-Banuet *et al.* ([2015](#ref-valiente2015beyond)), Bascompte *et al.* ([2006](#ref-bascompte2006asymmetric)), and the relationships between biodiversity and ecosystem functions (Thebault and Loreau 2006). However, species interactions are challenging to observe (REFs). They are typically measured through shorter-term observational studies yet are highly variable in time and space (Liu and Gaines 2022). Studying species interactions also requires extensive taxonomic expertise, especially those from diverse clades (Hebert *et al.* ([2003](#ref-hebert2003biological))), further challenging their frequent assessment. As a result, few ecological networks have been fully mapped (REFs). Improving our capacity to efficiently study species interactions consistently across several ecosystems, landscapes, and management context is thus pivotal to enacting conservation interventions (Tylianakis et al. 2010). This is particularly urgent as increased biological introductions and climate change are impacting species interactions at a rapid rate, thus calling for more frequent and consistent monitoring (REFs).

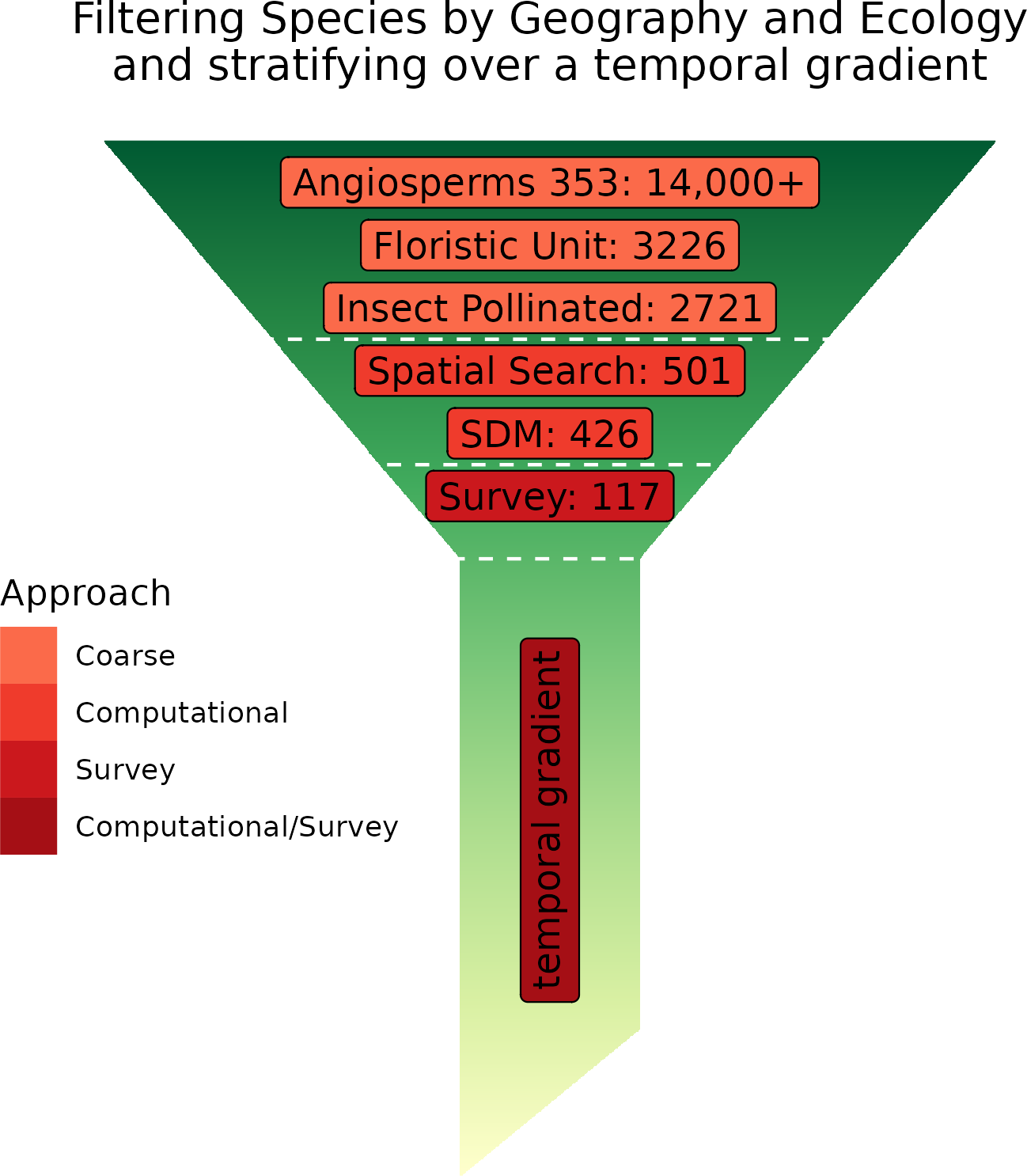
Recently barcoding (the identification of a sample from a single fragment of an organism), and metabarcoding (the identification of a sample containing a mix of organisms *e.g.* soil), have shown considerable promise in all Kingdoms of Life (Ruppert *et al.* ([2019](#ref-ruppert2019past))). The ability to identify species from a fragments of organisms (e.g. hair, scat, leaf tissue) increases our ability to understand the interactions of not only entire ecosystems but also a focal, generally rare and hence difficult to detect, organism with their surrounding; allowing for the most precise allocation of conservation decisions and funds e.g. those for restoration processes (Banerjee *et al.* ([2022](#ref-banerjee2022environmental)), Johnson *et al.* ([2023](#ref-johnson2023environmental))). With plants the identification of members of certain clades to species using barcoding has been successful in some groups (Kress ([2017](#ref-kress2017plant))), whereas many other clades have proven more problematic (Liu *et al.* ([2014](#ref-liu2014identification)), Group *et al.* ([2011](#ref-china2011comparative)), Coissac *et al.* ([2012](#ref-coissac2012bioinformatic))). Metabarcoding incurs additional challenges to those which exist for the currently implemented barcodes (Li *et al.* ([2015](#ref-li2015plant)), Kress & Erickson ([2007](#ref-kress2007two)), Group *et al.* ([2009](#ref-cbol2009dna)), Coissac *et al.* ([2012](#ref-coissac2012bioinformatic))). Particular limitations of the traditional high copy number barcodes (e.g. ITS2, *rbcL*, *matK*, *trnH-psbA*) include the utilization their rates of divergence, gene tree conflict, and hybridization (Coissac *et al.* ([2016](#ref-coissac2016barcodes)), Fazekas *et al.* ([2009](#ref-fazekas2009plant))). Currently, most plant metabarcoding endeavors only allow the identification of material to the level of family or genus.

Currently one of the largest plant systematic endeavor, undertaken by the Royal Botanic Gardens Kew, is the Plant and Fungal Tree of Life (PAFTOL) is approaching completion (Baker *et al.* ([2021a](#ref-baker2021PAFTOL))). This data set will contain hybridization capture (Hyb-Seq) data from at least one species in each genus of the plant kingdom, 14,000+ represented species, using the popular Angiosperms353 (A353) probes (Baker *et al.* ([2021a](#ref-baker2021PAFTOL)), Johnson *et al.* ([2019](#ref-johnson2019universal))). Data from the 10kP project, which seeks to develop reference genomes from a phylogenetically diverse suite of 10,000 plant species, will contribute more species by 2030 (Cheng *et al.* ([2018](#ref-cheng2018tenkp))). These publicly available data serve to provide a phylogenetically comprehensive backbone for plant metabarcoding. Similar projects, focused on whole genome sequencing of all organisms in certain geographic areas, such as the ‘Darwin Tree of Life’, which aims to sequence all described taxa in Britain and Ireland, will provide geographically dense data sets to seed projects globally (Life Project Consortium *et al.* ([2022](#ref-darwin2022project)), Lewin *et al.* ([2022](#ref-lewin2022biogenome))). These novel genomic data sets will promote the ability to apply metabarcoding to resolve a diverse array of questions relevant to theoretical and applied ecology (Kress ([2017](#ref-kress2017plant)), Hollingsworth *et al.* ([2016](#ref-hollingsworth2016telling))). However, the application of metabarcoding still faces challenges relating to the enormity of the genomic data sets and the computational power required to process sequence data.

To address these challenges, we are proposing the generation of a smaller, and more appropriate reference list for identifying candidate taxa from barcode data. By reducing the list of likely species in a study site it can increase the accuracy and efficiency of metabarcoding results in plants, which will better inform data collection for projects with limited access to taxonomic expertise. Specifically, we propose using current distribution data to generate a list of candidate species relevant to the study area and its ecological characteristics (Bell *et al.* ([2022](#ref-bell2022plants))). Initially such a species list can be leveraged to identify a target collections list of plant tissues to be included in genetic studies. It can subsequently help to reduce the size of a reference sequence database, which increases computation speed and efficiency, allowing for the use of genomic data on personal computers. This approach can significantly reduce processing time, increasing project efficiency, particularly as most next-generation sequence data is deposited as raw-sequence reads. We test this metagenomic and informatics approach by examining the plant species composition within corbiculae pollen loads of bumble bees as a case study. We use DNA barcoding to identify the plant species in corbiculae pollen loads collected from overwintered queen bumble bees, and compare this approach with direct observations of bee flower visits and microscopically-examined pollen slides, which has shown some incongruency in several floral visitation networks involving smaller bodied fauna (Olesen *et al.* ([2011](#ref-olesen2011missing)), Barker & Arceo-Gomez ([2021](#ref-barker2021pollen)), Zhao *et al.* ([2019](#ref-zhao2019topology)), Alarcón ([2010](#ref-alarcon2010congruence))). The assessment of the plant species in pollen is a commonly desired, across several potential applications, despite numerous potential complications (Pornon *et al.* ([2017](#ref-pornon2017dna)), Bell *et al.* ([2017](#ref-bell2017applying)), Sickel *et al.* ([2015](#ref-sickel2015increased)), Bell *et al.* ([2019](#ref-bell2019quantitative)), Suchan *et al.* ([2019](#ref-suchan2019pollen)), Johnson *et al.* ([2021](#ref-johnson2021airborne))).

**Methods**

To improve metabarcoding reliability and efficiency, we suggest creating a regional list of candidate species using digital collections gleaned from herbaria, survey work, and community science (Figure 1). This list can further be refined using species distribution models and temporal filtering to limit the impact of spatial and taxonomic biases in the species list and account for spatial variations in niche availability throughout the study area. The final list is then used to inform collection of plant samples to create a library and inform metabarcoding . We apply this methodological framework to the metabarcoding of corbiculae pollen loads of bumble bees and compare the accuracy of our metabarcoding approach both prior and after applying a spatial and temporal filtering to pollen identification conducted by experts and field observations.



**Figure** 1. Simplified Conceptual Diagram of three approaches leading towards classification of sequencing results, and the number of species associated with them in our area. The upper three boxes indicate a common coarse approach, assuming one has a digitized Flora, which is not always the case. The center two boxes indicate the computational approach illustrated here. The final box indicates the use of the expert field data in the case study. The stem of the final applies to both Computational and Expert Survey results, and should be thought of as using time ala chromatography, in the post-classification process.

**Create Species List**

Survey databases

The first step of our methodological approach consists in collating existing species occurrence data to create a preliminary list of candidate species likely to occur in the study area. Such occurrence data can be retrieved from databases including the Botanical Information and Ecology Network ‘BIEN’ (Maitner ([2022](#ref-bien2022))) or the Global Biodiversity Information Facility (GBIF; REF) which aggregate occurrence records from various sources including community science datasets (e.g., iNaturalist) and herbarium records. These databases can be queried to extract a list of ecologically relevant vascular plant species with occurrence records in the administrative (e.g. county, region) or ecological unit (e.g. ecoregion) including the study site or within a search radius surrounding the cite. Focusing on a broader geographic region than the study area alone can account for inconsistent sampling effort and undetected species, which are both common in herbarium records and community science datasets (Pereira and Copper 2006; Geldmann et al. 2016; Girardello et al. 2018).

Distribution modeling to see which taxa are likely to be within his area

While increasingly used in biodiversity assessments (REFs), datasets like BIEN and GBIF can nonetheless include occurrence records with erroneous taxonomic identification, geographic inaccuracies, or historical records (Smith et al. 2016; Freitas et al. 2020). Hence additional filtering might be needed to develop a more robust list of candidate species. Once a regional list of potential species is established, Species Distribution Models can be used to further reduce the species list to only keep species with ecological requirements found within study sites (Figure 1). This step can also help filter for species adapted to the local ecological condition of the study sites, which can be particularly useful in highly heterogeneous landscapes in which a simple survey of databases may overestimate the list of potential species or when limited taxonomic expertise is available to filter the preliminary species list. Species Distribution Models examine relationships between known species occurrences data and a set of environmental variables to identify potentially suitable habitats with similar ecological conditions elsewhere in the landscape (Guisan and Zimmermann 2000; Phillips et al. 2006). They are often used to inform conservation strategies, evaluate invasion risks, or estimate species richness (Thuillier et al. 2005; Pineda and Lobo 2009). Environmental variables used as predictors in SDMs typically describe the fundamental ecological niche of the species (e.g., climate, elevation, soil types) or constraints to its distribution (e.g., land uses/land covers, canopy openness) and commonly rely on publicly available geospatial datasets (see Table 1 for examples). Once relationships between known occurrence records and these environmental variables are established, SDM algorithms then score remaining portions of the study area on a suitability scale ranging from 0 (highly unsuitable) to 1 (highly suitable) on the basis of their similarity to habitats currently sheltering species populations. Once SDMs are generated for the entire pool of species identified through a database search (section 2.3), these suitability maps can be aggregated to refine the list of species likely to occur in study sites.

**Table 1**. Example of environmental variables and data sources commonly used in species distribution models

|  |  |
| --- | --- |
| **Variable** | **Data source** |
| Climate | WorldClim, IPCC Data Distribution Center, GRIMET, PRISM |
| Elevation | National Atlas |
| Land use/Land Cover | MODIS |
| Vegetation cover | Remote sensing products (e.g. Landsat, MODIS) |

Temporal filtering

When studying species interactions, a potentially helpful additional step is to query the candidate species list to retain plant species with phenology that aligns with the study species preferences. For example, when studying plant-pollinator interactions, considering that plant flowering varies over space and time, the project may want to focus only on the species flowering during the pollinators’ active period,. These contrasts in the phenological periods can thus provide an additional filter for identifying material in certain types of metagenomic samples (Janzen ([1967](#ref-janzen1967synchronization)), Newstrom *et al.* ([1994](#ref-Newstrom1994ANC)), Thompson ([1994](#ref-thompson1994coevolutionary)), CaraDonna *et al.* ([2021](#ref-caradonna2021seeing))). Community science efforts focusing on plant phenology can provide helpful information on the phenology of commonly observed species. Examples of such datasets include: the National Phenology Network in the United States (Denny et al. 2014), SeasonWatch in India, PlantWatch in Canada, and the Pan European Phenology Project. Local flora can also provide useful information on the phenological characteristics of species.

**Collecting representative samples for reference library.**

Although several online genomic databases do contain numerous taxa, these often contain only a single representative of each genus, and do not always account for geographic variability. For this reason, we are recommending using the reduced list of potential candidate plant species as a target list to collect leaf samples to generate a reference library. This will increase efficiency and accuracy of analyses down the line. We recognize that this might not be feasible for all study systems, or for extensively long candidate species list, however using current knowledge of the system and key players in an interaction will allow for more accurate hypothesis testing. For example, if interested in pollination, you can focus on known candidate species that represent flower morphologies associated with specific pollinator guilds (Fenster et al 2004). DNA extraction and processing of these should be done alongside the samples collected for this system.

**Using baits to pull out specific regions (barcode) of the genome**

For barcoding the environmental samples and leaf samples collected from candidate taxa, we are proposing using a Next generation sequencing approach which incorporates a target capture probes (CITE). There are several approaches available for preparing genomic libraries that vary in cost, ease of use (CITE), and the quality and quantity of sample DNA required that are all appropriate for target capture approach. The choice of best library preparation approach will depend on the type of samples, the lab and sequencing platform preferred. Once a library has been generated, and size selected to generate appropriate sized fragments, and appropriately barcoded to distinguish individuals, a synthetic probes, often referred to as baits, is used to preferably retain fragments within the library which contain sequences associated with desired loci, while remainder of fragments which do not are washed away, maximizing sequencing efficiency. There are a number of probe sets which have been developed, some are taxon specific, but a handful are more general and appropriate for barcoding multiple taxa, including Angiosperms353 probe set (Johnson et al., 2019), which is appropriate for most flowering species, and goFlag (cite) might be more appropriate for non-flowering species. These approaches capture data from multiple loci across genome, which are both expressed, therefore have lower mutation rates, and are found to occur in single copies in most taxa, making them appropriate for taxonomic studies.

**Compare barcode to reference library**

The data generated is filtered and trimmed using pipleines such as TRIMMOMATIC (Bolger et al., 2014) to identify and remove sequences added during library preparation process and not associated with sample. These sequence fragments (contigs) are then compared back to the reference probe sequences to sort them by loci of origin (HybPiper ; Johnson et al., 2016) . For libraries which contain only a single taxa, the contigs from the same loci then be aligned to generates a consensus sequence (MAFFT (Katoh and Standley, 2013). This consensus sequence can then be used as a reference for which to compare the multiple species samples and identify taxa.

For samples which ontain multiple species, the process iS X.

**Case Study: Measuring bumblebee visitation based on corbiculae load.**

**2.5.1. Bee-Flower Observations and Pollen Load Collection**

To test this metagenomic and informatics approach, we used the identification of plant species within corbiculae pollen loads of bumble bees collected at the Rocky Mountain Biological Laboratory (RMBL; 38°57.5” N, 106°59.3” W, 2900 m.a.s.l.), Colorado, USA (Appendix 1 for site information). Bumblebees are one of the only groups of insects with unequivocal quantitative evidence for numerous population declines. They are also the most effective pollinators in many temperate montane ecosystems (Cameron & Sadd ([2020](#ref-cameron2020global)), Goulson *et al.* ([2008b](#ref-goulson2008decline)), Williams ([1982](#ref-williams1982distribution)), Colla *et al.* ([2012](#ref-colla2012assessing)), Bergman *et al.* ([1996](#ref-bergman1996micrometeorological)), Bingham & Orthner ([1998](#ref-bingham1998efficient)), Grixti *et al.* ([2009](#ref-grixti2009decline)), which are experiencing steep changes in annual temperature under climate change (Brito-Morales *et al.* ([2018](#ref-brito2018climate)), Pepin *et al.* ([2022](#ref-pepin2022climate))). As such, the effective management of these ecosystems relies on the study of keystone species, such as bumble bees, that play a critical role in maintaining their current diversity (Loarie *et al.* ([2009](#ref-loarie2009velocity)), Dobrowski & Parks ([2016](#ref-dobrowski2016climate))). The queen bumblebees emerge early in the season to identify nesting sites. Floral resources are essential at these foraging stages for increasing their weight before diapause and increasing their ovary weights while establishing their found nests, and therefore this period represent potential demographic bottlenecks in bumble bee populations (Sarro *et al.* ([2022](#ref-sarro2022bumble))).

Species list

To identify a list of potential candidate species that a bumblebee might visit within the study area (i.e., Southern Rockies in Colorado), we queried the BIEN database to retain all vascular plants records identified within a distance of up to 100km from our study area. We used the resulting 23,919 occurrences to create species distribution models for the 501 species identified through this database query. We selected 90% of records for each species to train the species distribution models, while the remaining 10% were retained for validation. We then generated 4,029 absence points (i.e., locations where the focal taxon is anticipated missing) through a random stratification of 19% of the land cover (Land Management ([2019](#ref-aim2019database))). To achieve a larger absence data set, we generated an additional 1,000 pseudo-absence records for each taxon by randomly selecting coordinates located at least 10km away from any occurrence record. For Machine Learning models, these pseudo-absences were reduced so that the ratio of presence to absence records were balanced (Barbet-Massin *et al.* ([2012](#ref-barbet2012selecting))). To achieve this, we removed absence records inside of 10% of the mean sample value of any predictor variable the presence records; the required number of absence records were then randomly sampled.

We used 26 environmental variables at a 30m resolution to construct the species distribution models; six related to climate, five soil, four topographic, four related to cloud cover, with the remaining reflecting assorted abiotic parameters (Wilson & Jetz ([2016](#ref-wilson2016remotely)), Wang *et al.* ([2016](#ref-wang2016locally)), Hengl *et al.* ([2017](#ref-hengl2017soilgrids250m)), Robinson *et al.* ([2014](#ref-robinson2014earthenv))) (Appendix 2). These environmental variables were extracted from publicly available datasets and selected to represent variables interacting with plant physiology. For linear regression models these predictors underwent both *vifstep* (theta = 10, max observations = 12,500) and *vifcor* (theta = 0.7, max observations = 12,500) to detect highly correlated variables, and collinear features were removed leaving 16 variables (Naimi *et al.* ([2014](#ref-usdm2014))).

We created species distribution models using four different algorithms commonly used in the literature: Random Forest (def) and Boosted Regression Tree’s (def), Generalized Linear (def; GLM) and Generalized Additive Models (def; GAM) (Barbet-Massin *et al.* ([2012](#ref-barbet2012selecting)) (Elith 2011). We then combined all the previous model outputs into an ensemble forecasting, as ensembled predictions have been shown to outperform their constituent models and minimize the analytical noise of individual runs (Araujo & New ([2007](#ref-araujo2007ensemble))).

To evaluate the accuracy of the species distribution models, additional presence records from GBIF (n = 61,789), and AIM (n = 12,730) were used as test and training sets (n = 74,519) for logistic regression (Occdownload Gbif.Org ([2021](#ref-gbifDL2021sdms)), Land Management ([2019](#ref-aim2019database))). Additional novel absence records were generated from the AIM data set to create a data set where each species has balanced presence and absences. The True Skill Statistics (TSS), was used to measure model accuracy as it has been shown to work across a wide range of species occurrences prevalence (Allouche *et al.* ([2006](#ref-allouche2006assessing))). [Something about whether any models were left out if they did not meet TSS]. Finally, we converted model outputs (i.e., maps of habitat suitability) into presence/absence by classifying as “present” cells with a mean suitability value greater than 0.5 (Prim ([1957](#ref-prim1957minimum))). This resulted in a list of 426 species with ecological requirements that matched out study sites.

We further refined the species list by identifying species with a flowering phenology likely to overlap with either foraging phase of the bumble bee. We used Weibull estimates of phenological parameters developed by (Belitz *et al.* ([2020](#ref-belitz2020accuracy)), Pearse *et al.* ([2017](#ref-pearse2017statistical))). to estimate the flowering periods of all 426 modeled species. We used the SnowUS data set (Iler *et al.* ([2021](#ref-iler2021conceptual)), Tran *et al.* ([2019](#ref-tran2019cloud))) from 2000-2017 to remove herbarium records collected during times in which flowering is impossible at the study area due to snow cover. We used the SnowUS database to identify an annual “snow-free” period starting after the first three days of contiguous snow absence and ending after the first three days of contiguous snow cover in fall. Herbarium records after the 3rd quantile for melt, and the 1st quantile for snow cover of these metrics were removed from the candidate list. Species with > 10 records had their Weibull distributions generated for the date 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 effective start and end of flowering. These estimates were compared to a long-term observational study of flowering phenology 1974-2012 (CaraDonna *et al.* ([2014](#ref-caradonna2014shifts))), and the floral abundance data from 2015, using Kendall’s tau.

Pollen and plant material collection

Observations of visitation and bee corbiculae pollen collection was conducted at the Rocky Mountain Biological Laboratory (RMBL; 38°57.5” N, 106°59.3” W, 2900 m.a.s.l.), Colorado, USA (Appendix 1 for site information). Pollinator observations of *Bombus Latreille spp. (Apidae Latreille*) were conducted from May 29th – July 23rd of 2015 in six study sites as a part of a larger study (described in Ogilvie and CaraDonna 2022). Observations of *Bombus* foraging took place for one hour at each field site. Corbiculae pollens loads were non-lethally collected from queens encountered by capturing them in an insect net and transferring them into a restraining device (Kearns *et al.* ([2001](#ref-kearns2001natural))). We then collected a single pollen load (i.e., from one leg) from the bee and then released it. At weekly intervals at each site, we also recorded the abundances of flowers visited by bumble bees within belt transects spread over three vegetation types (0.5 x 40 m transects in each vegetation type, 60 m2 total area per site). Using five years (2015-2020) of observational data on *Bombus* queen interactions with flowering plants at these studies sites, we identified the plant taxa most frequently visited by queens across all years. In order to capture more variability inherit in the 353 loci we sampled the 12 most visited taxa twice using samples collected from one site within the Gunnison Basin River Drainage and one individual collected from another more distal population. In addition, we included a congener - or a species from a closely related genus to serve as an outgroup for all 12 taxa to confirm the efficiency of 353 baits for species level identification. We also sequenced another 15 taxa of plants commonly visited by *Bombus* workers, based on the abundances, and immediate access to plant tissue, in the aforementioned data set (Appendix 3). Plant collections were identified typically using a combination, of dichotomous keys and primary literature as required (Flora of North America Editorial Committee ([1993+](#ref-flora1993flora)), Hitchcock & Cronquist ([2018](#ref-hitchcock2018flora)), Ackerfield ([2015](#ref-ackerfield2015flora)), Lesica *et al.* ([2012](#ref-lesica2012manual)), Cronquist *et al.* ([1977+](#ref-cronquist1977intermountain)), Allred & Ivey ([2012](#ref-allred2012flora)), *Jepson flora project* ([2020](#ref-jepson2022online)), Mohlenbrock ([2002](#ref-mohlenbrock2002vascular))).

Pollen DNA Extraction

All lab work was carried out at The Daniel F. and Ada L. Rice Plant Conservation Science Center at the Chicago Botanic Garden, Glencoe, Illinois, USA. 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](#ref-doylesCTAB))) that included two chloroform washes. Pollen DNA was also extracted using a modified CTAB method (Lahlamgiahi et al. and Guertler et al. (2014, 2014) which included using a SDS extraction buffer (350µL, 100mM Tris-HCl, 50 mM EDTA, 50 mM NaCl, 10% SDS v/v., pH 7.5) and vortexed to allow dissolution of corbiculae. Pollen grains were then macerated with Kontes Pellet Pestles, and the tip of these washed with 130 µL of the SDS extraction buffer, samples were then incubated for 1 hour at 30°C. This was followed by the addition of 10% CTAB solution (450ul, of 20 mM Tris-Cl pH. 8.0, 1.4 M NaCl, 10 mM EDTA pH 7.5, 10% CTAB, 5% PVP, ~85% Deionized water) and RNAse (10 uL of 10 mg/mL) and samples were incubated for 40 minutes at 37°C, on a heat block (Multi-Blok, Thermo Fisher Scientific, Waltham Massachusetts) set to 40°C. After 20 minutes incubation, Proteinase K (15 µL of 20mg/ml) and DTT (12.5 µL of 1M in water) were added, and the samples were further incubated at 60°C for 1 hour. Samples were then incubated overnight at 40°C. 500 µL of Phenol-Chloroform-Isoamyl alcohol (25:24:1) were added, vortexed, and centrifuged at 10,000 rpm for 10 minutes and the aqueous phase was pipetted to a 1.5 ml centrifuge tube.

Extractions were assessed using a Nanodrop 2000 (Thermo Fisher Scientific) and Qubit fluorometer (Thermo Fisher Scientific). DNA extracts were then cleaned using 2:1 v./v. Sera-Mag beads (Cytiva, Little Chalfont, UK) to solute ratio following the manufacturer’s protocol, eluted in 0.5x TE, and the eluent allowed to reduce by half volume in ambient conditions. DNA was quantified using a Qubit fluorometer.

Using baits to pull out specific regions (barcode) of the genome

Sequence library preparation was performed using the NEBNext Ultra II FS-DNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, Massachusetts, USA) which was slightly modified from manufacturers recommendation. 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, with similar library preparation of all samples. Libraries were pooled and enriched with the Angiosperms 353 probe kitV.4 (Arbor Biosciences myBaits Target Sequence Capture Kit) by following the manufacturer’s protocol (Brewer et al. 2019). Sequencing was performed using an Illumina mi-Seq with 150-bp end reads, (NUSeq Core, Chicago, Illinois).

Processing Sequences,

Sequences were processed using Trimmomatic, which removed sequence adapters, clipped the first 3 bp, discarding reads less than 36 bp, and removing reads if their average PHRED score dropped beneath 20 over a window of 5 bp (Bolger & Giorgi ([2014](#ref-bolger2014trimmomatic)), Tange ([2021](#ref-tange_2022_6377950))). Contigs generated were mapped to a reference with HybPiper with using target files created by M353 (Johnson *et al.* ([2016](#ref-johnson2016hybpiper)), McLay *et al.* ([2021](#ref-mclay2021new))).

Plant samples:

Pollen samples: A custom Kraken2 database was created by downloading representative species indicated as being present in the study area by the spatial analyses from the Sequence Read Archive (SRA) NCBI (Wood *et al.* ([2019](#ref-wood2019improved))). These sequences were processed in the same manner as our novel sequences. The Kraken2 database was built using default parameters. Kraken2 was run on sequences using default parameters (Appendix 9). Following Kraken2, Bracken was used to classify sequences to terminal taxa (Lu *et al.* ([2017](#ref-lu2017bracken))). Finally all reads which could be classified by these databases were passed to a local BLAST database.

A local NCBI database was built using the same processed novel and downloaded sequences as the previous database (Camacho *et al.* ([2009](#ref-camacho2009blast))).

Validation process

To precisely classify the contents of each corbiculae load, the sequences classified by molecular methods were compared with the fieldwork which recorded the presence and absence of species and their duration of flowering at a very fine resolution, . The quantitative counts of grains from microscopy were combined with the semi-quantitative sequencing results to estimate the abundance of each identified species in each corbiculae load.

[Validate using pollen identification].

Five years of observational data on *Bombus* Queen Bee foraging at these studies sites (Ogilvie & CaraDonna ([2022](#ref-ogilvie2022shifting))) was used as baseline of what we should expect for visitation within this system. This was supplemented with the RMBL Vascular Plant Checklist (Frase & Buck ([2007](#ref-fraser2007vpc)). Microscopy was use to qualitatively identify, and quantitatively note, the plant species present in corbiculae loads. A pollen reference library of know taxa visited by Bumbleebees and therefore expected to be present in the corbiculae loads was made using a fuchsin-jelly stained grains from slides previously prepared by the authors (n = 21), and other researchers (n = 38) (Beattie ([1971](#ref-beattie1971technique)), Brosi & Briggs ([2013](#ref-brosi2013single))). A total of 62 voucher slides for species were prepared and imaged at 400x (Leica DMLB, Leica MC170 HD Camera, Leica Application Suite V. 4.13.0) from previously prepared slides by (Ogilvie & CaraDonna ([2022](#ref-ogilvie2022shifting)) and both non-accessioned and RMBL herbarium collections to supplement the number of species and clades covered (Appendix 3). To prepare the pollen slides from corbiculae, all corbiculae loads were broken apart and rolled using dissection needlepoints to increase heterogeneity of samples. *Circa* 0.5mm2 of pollen was placed onto a ~4mm2 fuchsin jelly cube (Beattie ([1971](#ref-beattie1971technique))) atop a graticulated microscope slide, with 20 transects and 20 rows (400 quadrants) (EMS, Hartfield, PA). The jelly was melted, with stirring, until pollen grains were homogeneously spread across the microscope slide. Slides were sealed with Canada Balsam (Rublev Colours, Willits, CA) followed by sealing with clear nail polish to prevent oxidation; all samples are noted in Appendix 4.

To identify the pollen present in corbiculae loads, light microscopy at 400x (Zeiss Axioscope A1) was used. In initial sampling in three transects, each pollen grain was identified to morphotype and counted; an additional two transects were scanned for morphotypes unique to that slide, if either transect contained a unique morphotype then all grains in that transect were also identified and counted. We used clustering techniques to supplement our subjective opinions of which plant taxa were distinguishable via light microscopy, and to develop a dichotomous key to pollen morphotypes. Ten readily discernible categorical traits were collected from each specimen in the image collection. These traits were transformed using Gower distances, and clustered using Divisive Hierarchical clu**st**ering techniques (Maechler *et al.* ([2022](#ref-cluster2022))). Using the cluster dendrogram, elbow plot, and heatmaps (Hennig ([2020](#ref-fpc2022))), of these results morphological groups of pollen which could not be resolved via microscopy were delineated, and a dichotomous key was prepared (Appendix 6). This key was then used to identify the pollen grains sampled from corbiculae loads to morphotypes in a consistent manner.

Subsequent to the first round of sampling, non-parametric species richness rarefaction curves (Oksanen *et al.* ([2022](#ref-vegans2022))), and non-parametric species diversity rarefaction curves were used to assess the completeness of sampling (Chao *et al.* ([2014](#ref-inextArticle)), Hsieh *et al.* ([2020](#ref-inextPackage))). Slides not approaching the asymptote of the rarefaction curve were then re-sampled, and analysed iteratively for up to a total of seven transects (Appendix 7 & 8).

To reclassify the sequence reads, these data were combined with the flora observation data, and mapped by genus. If more than one species in the genus was flowering at that time and site, then the reads were split evenly between the taxa. For sequence data which did not match at the genus level, a user subjectively scored them based on the species composition and phenological activity at each site, the queen interaction data, and pollen assignments. To estimate the abundance of each of these species in the corbiculae loads, these data were combined with the microscopy data. For each morphotype detected in pollen, and each classified sequence read which was not detected via microscopy, they were given a value of 0.5% to indicate their trace presences. When more than a single species belonged to a morphotype group in a single sample, the quantitative values from the morphological work were multiplied by the relative sequence abundance of each species in the load. All final compositions were standardized to a sum of 100%, by adding or subtracting the differences (induced by classifying records as ‘trace’) to all species with abundances > 1%.

**Results**

**Species list**

Spatial Analyses

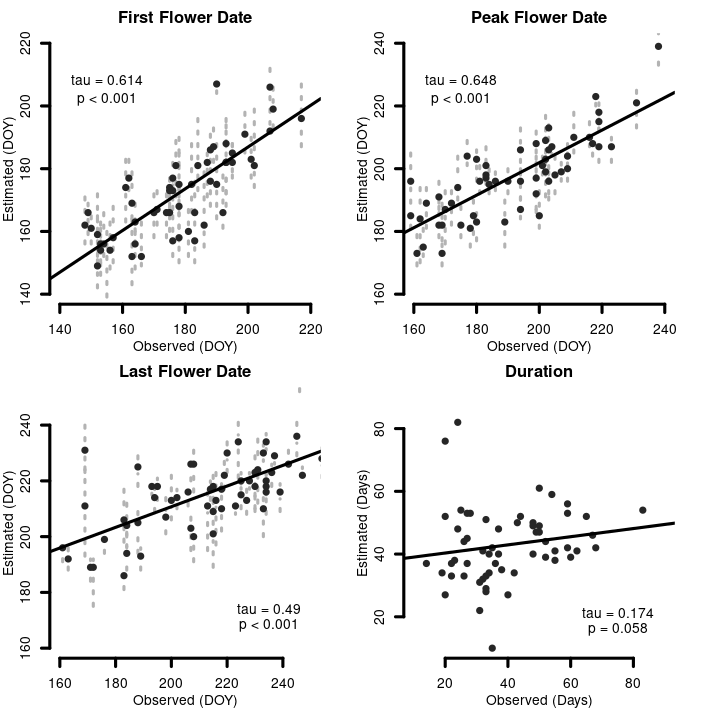
The threshold distance under which a species would undergo species distribution modelling was the median (25.009 km) of the logistic regression assessing the probability of occurrence of a species record as a function of distance from the study area. A 2-sample test for equality of proportions with continuity correction (X-squared = 13.254, df = 1, p-value = 0.000136, 95% CI 0.04-1.00) was used to test whether more of the records located in the broad ecological sites present at the field station, between the distance of the median (25.009 km) to the third quantile (ca 43.830 km) of the regression distance, were true presences at the field station. Including these records would have resulted in modelling an additional 222 species distributions of which 30 are true presences, these taxa were not modelled.

Across the entire spatial domain of modelling, all ensembled models (n = 968) had an accuracy of 0.84 (95% CI 0.8356 - 0.8443), kappa = 0.68, p-value < 0.001, sensitivity = 0.80, specificity = 0.87, AUC = 0.92. The 493 ML ensembles accurately predicted the presence of 362 occurrences (65.3%), incorrectly predicted the presence of 64 (11.6%), incorrectly predicted 34 true presences (6.1%) as being absent, and correctly predicted the true absence of 33 (6.0%). Of the 554 vascular plants with biotic pollination, the 475 LM ensembles accurately predicted the presence of 286 (51.6%), incorrectly predicted the presence of 41 (14.3%), incorrectly predicted 93 true presences (16.8%) as being absent, and correctly predicted the true absence of 55 (9.9%). The balanced accuracy of the ensembled models is 0.664 (Sensitivity = 0.573, Specificity 0.754). Of the 117 plant species identified to the species level across the spatial extents of all plots and duration of queen bee activity, the ML ensembles predicted the presence of 105 (89.7%) of them, and LM ensembles 102 (87.2%). Of the missing species two (1.7%) are Orchids, six (5.1%) are non-native, and one (0.85%) is of contested taxonomic standing, all of which (7.65%) are restricted from the initial query database.

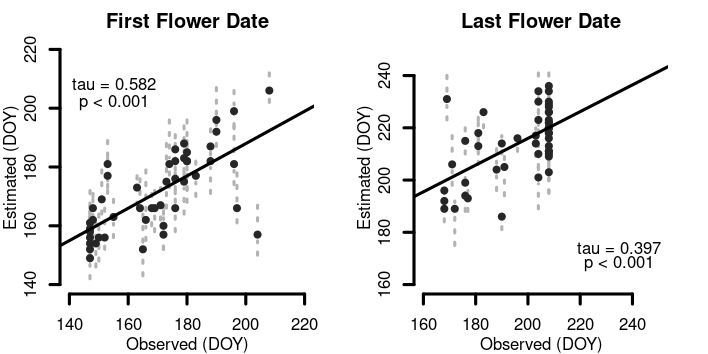
Temporal Analysis

The first date of modeled snow melt in the Gothic area (n = 17, = 137.9, Mdn = 135, 3rd quartile = 151), and the first date of a consistent winter snow base (n = 17, = 299.9, Mdn = 300, 1st quartile = 291) from 2000-2017 were used as delimiters for the inclusions of herbarium records in modelling. Of the 439 species predicted likely present in the area via logistic regression, 332 species (64.4%) with more than 10 records in the focal level 4 ecoregions ( = 35.016, Mdn = 35, max = 96) had Weibull estimates calculated, an additional 56 species (11.2%) with enough contributing records from the “Sedimentary Mid-Elevation Forests”, a large ecoregion generally just beneath the elevation bands occupied by the five ecoregions around the study area had Weibull estimates also calculated ( = 13.868, Mdn = 13, max = 24). We could only compare 58 of these 388 species to plot based observational data from the long term (1974–2012) data set (CaraDonna *et al.* ([2014](#ref-caradonna2014shifts))), but this revealed high accord between the long-term ground truthed data set and the modelled species. 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). There was moderate evidence that the Weibull estimates had a weak positive association with the observed duration of flowering (p = 0.58, tau = 0.17).

Of the previous 58 species compared, 47 of these could be compared to transect based data from the six sites observed in 2015There was very strong evidence that the Weibull estimates were positively associated with the observed onset (p < 0.0001, tau = 0.58), and cessation of flowering (p < 0.0001, tau = 0.40).



*Modelled dates of when major flowering events occurred compared between long term and modelled data*

asddasdasdasdasdasdasdasdasdasdasd

*Modelled dates of when major flowering events occurred compared between 2015 and modelled data*

**Molecular analysis**

**Plant data**

Need some text here… how well did plant capture go?

**Corbiculae loads**

The 54 corbiculae loads had DNA extracted and underwent various steps towards target capture, in the end a total of 44 corbiculae samples were sequenced, with 7,752,353 reads recovered. The number of reads per sequence varied widely (range = 76 - 508,795, = 176,189.8, Mdn = 138,395). Of the possible 353 loci, the number which were recovered from each sample, and informative to BLAST were range = 24 - 353, = 305.5, Mdn = 331. The number of reads per loci from across all samples had a range of 178 - 506,653, = 20,688, Mdn = 12,616 (Appendix 11). After trimming 7,865,680 sequences remained. 10,682,538 reads were matched using Kraken, of the reads classified by Kraken 10,160,768 reads were matched using Bracken, of the reads classified by Kraken 7,549,608 reads were matched using BLAST. Based upon subjective review of the three classifiers (Appendix 12) BLAST was chosen as the classification method which yielded the most probable results by the field ecologist, and its values were used for all subsequent analyses.

The initial classification of sequences which were made by BLAST were reviewed programmatically, using predicted presence of the species (from spatial modelling), modelled flowering time (from temporal modelling), and taxonomy (from existing sources). A sequential process was utilized which reassigned sequences based on binary combinations of the factors above (Appendix 15). 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 DB 2) Reduce false classifications of focal species 3) Identify high confidence sequence matches. Of the top ten taxa which were identified by BLAST for the 680 distinct records, 55.4% of the reads were classified to a species representing 48.3% of all classified reads, 41.9% of the reads were classified to genus representing 48.3% of all classified reads, and 0% of the records were classified to family.

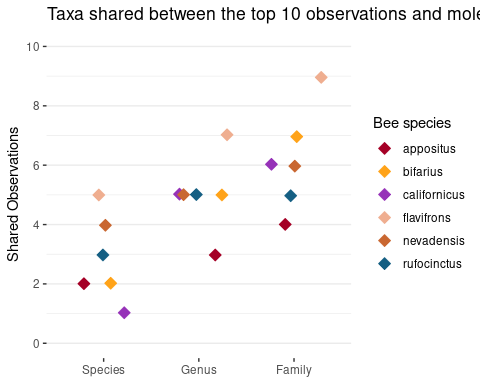
Metabarcoding Pollen Identification

Of the 0 classifications which were assigned to genera without any species predicted by spatial analyses, were investigated by hand after post-processing steps. These were all assigned via post-processing conditions (: , Appendix 15). These were manually assigned to a variety of ranks, occasionally to genus - 0, and species - 0, by consultation of the alpha-taxonomic literature (Sadeghian *et al.* ([2015](#ref-sadeghian2015molecular)), Sennikov & Kurtto ([2017](#ref-sennikov2017phylogenetic)), Rabeler & Wagner ([2016](#ref-rabeler2016new)), Pusalkar & Singh ([2015](#ref-pusalkar2015taxonomic)), Moore & Bohs ([2003](#ref-moore2003its)), Weber ([1998](#ref-weber1998new))).

**Validation**

In 2015 the six sites were surveyed once weekly from May 27-July 27 for a total of 52 hours. A total of 723 overwintered queen-pollen foraging interactions were observed (range per bee species by week range = 1-18, = 3.46, median = 2), with a range of total observed interactions per bee species across this time period (minimum = 1, = 59.08, median = 19, max = 184). Plants varied widely in the number of interactions which they partook in with each species of bee (range per plant species by week minimum = 1 - 20, = 3.51, median = 2), with a range of total observed interactions per plant species over this time period (minimum = 1, = 20.26, median = 4, max = 141). The number of plant species which bees were observed interacting with varied more narrowly (range = 1 - 18, = 8, median = 6), interactions were observed with a total of 36 plant species.

A total of 66 corbiculae loads were collected from bees, 64 of them from queens.

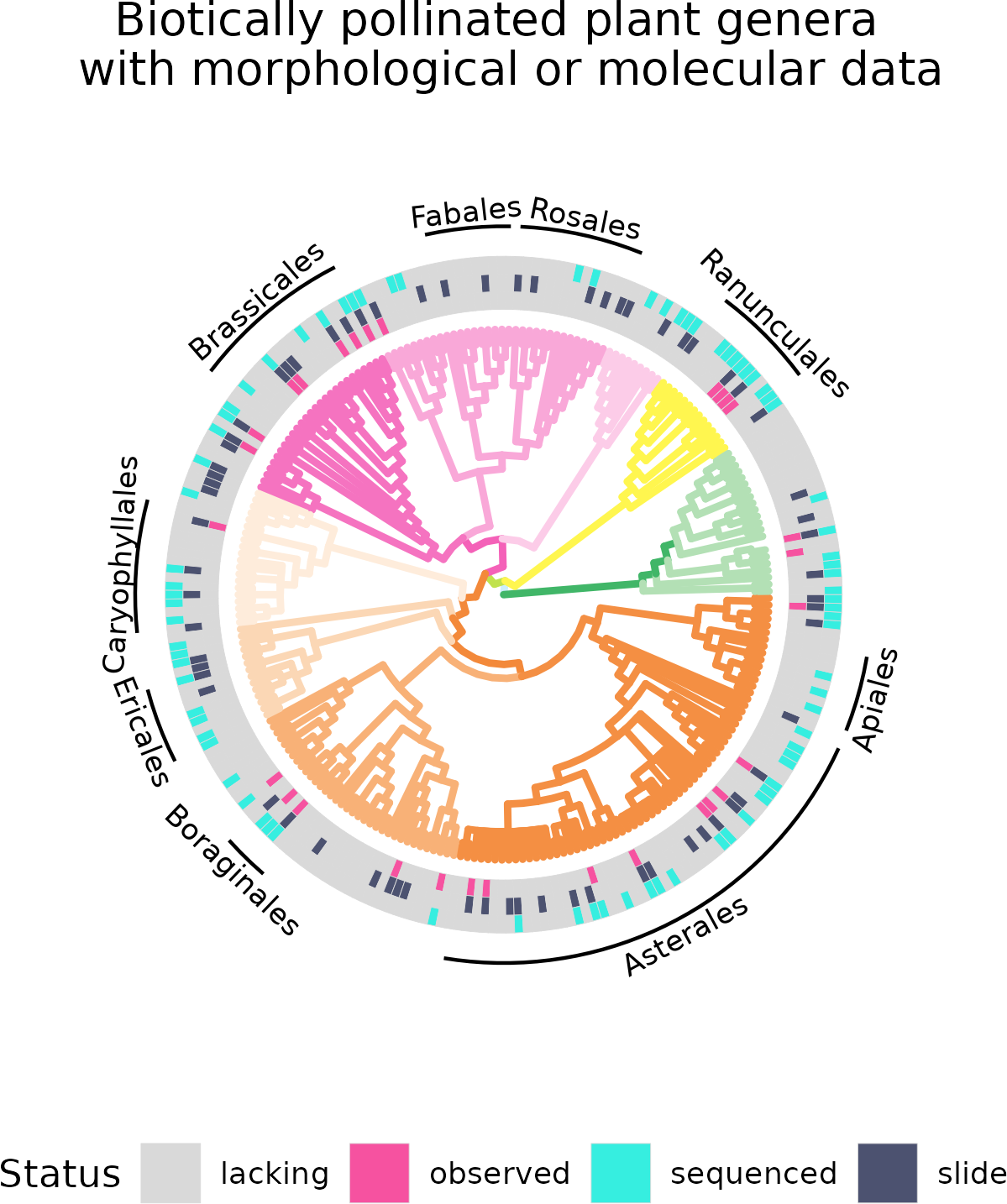


*Number of the ten most commonly visited plants which are also in the top ten most common sequences*

**Microscopic Pollen identification**

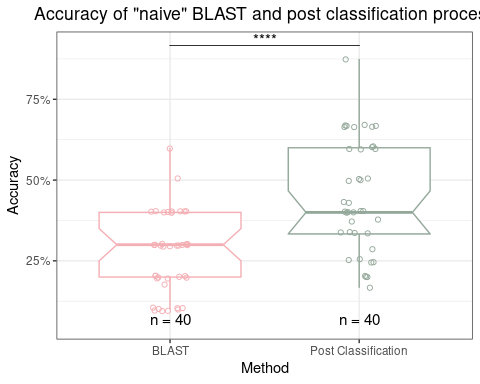
Using the fuchsin jelly preparation and light microscopic analyses of grains and scoring of 10 character states resulted in the establishment of 28 morphotypes which grains could be classified into (Appendix 6). From the 37 samples that were counted and based on rarefaction we identified substantial amounts of the abundance and morphotype richness of the samples (morphotype richness, = 4.5, median = 4, min = 1, max = 9) (Appendix 7 & 8). The number of counted pollen grains in each sample range from (514 - 19924, = 3319, median = 1891).

**Floral Observations**



*Phylogenetic tree of all biotically pollinated plant genera in the study area. The innermost ring indicates every genus which Queen Bee’s were observed to visit. The intermediate ring indicates that at least a single morphological pollen voucher slide was prepared for a member of the genus. The outermost ring indicates that sequence data were available for at least a member of that genus. Branch colors follow APG 4.*

**Integrated Observational, Molecular, and Palynological Results**



*Comparision of Accuracy between the initial output data from BLAST, and these same data subjected to the post-classification process which removes surrogate, and temporally restricted species*

a The mean only calculated across the samples where the species was detected

**Discussion**

* Summarize of case study
* Talk about accuracy and efficiency of approach
* Limitations
* Application on how this would be useful
* Conclusion.

We have demonstrated how the Angiosperms533 hyb-seq probes may be used for plant barcoding in a metagenomic context (Johnson *et al.* ([2019](#ref-johnson2019universal)), Hollingsworth *et al.* ([2016](#ref-hollingsworth2016telling))). This was exemplified in an ecologically relevant scenario, where the results have immediate implications for natural history guided fundamental science and land management. The test pollen loads contained a number of closely related taxa, some in notoriously morphologically difficult clades with rapid rates of diversification (e.g. *Mertensia*, *Lupinus* L.), at naturally occurring proportions (Nevado *et al.* ([2016](#ref-nevado2016widespread)), Nazaire & Hufford ([2014](#ref-nazaire2014phylogenetic))). We incorporated spatial and temporal approaches for creating custom sequence databases an approach which is readily applicable to any lab group with the capacity to perform next-generation sequencing across the entirety of multiple continents, and which we expect to be highly beneficial in many study areas. By combining insights from these novel approaches with an extensive observational field based study we show how these methods may be applied to test a variety of hypotheses related to ecological interactions.

The SDM’s which we generated, with relatively few occurrence records and few modelling iterations, performed beyond expectations, likely due to the utility of the predictor variables and strong alignment of vegetation by orographic precipitation in the study area. However, we had difficulties in evaluating our predictions in an operational context. We utilized the database query approach, to only model species with a high probability of not being dispersal limited to the focal area, and focused on a relevant subset of many of these species ranges to reduce the contributions of range wide adaptions on habitat (Sork ([2018](#ref-sork2018genomic)), Joshi *et al.* ([2001](#ref-joshi2001local))). While the models worked well compared to both test, and validation with external point data, moving from points to polygon features was more difficult. We were able to compare our results to 1) a Flora, 2) lists of plants used by Bumble Bees at plots; the former inappropriate in that it contained a great number of species which we sought to use modelling to reduce *e.g.* all strictly alpine species, and the latter inappropriate in that it contained only species relevant to *Bombus* but had no official ‘absence’ data. Further given the, size of the minimum spanning tree which we extracted points to, a formal floristic inventory would still be a time intensive process. Accordingly, we expect the real results of our data lay somewhere in between these two evaluations; with an excess of species predicted present (Dubuis *et al.* ([2011](#ref-dubuis2011predicting)), Calabrese *et al.* ([2014](#ref-calabrese2014stacking)), Pinto-Ledezma & Cavender-Bares ([2021](#ref-pinto2021predicting))), but few enough that they lend themselves to metabarcoding. We observe that our models seemed very capable of effectively identifying alpine species and removing them in binomial contexts.

In regards to the case study, our results indicate the overall information gleaned from observations of queen Bumble Bee foraging and analysis of pollen records are largely congruent. Relaxing concerns regarding differences between the broad insights gleaned from observational, as compared to data derived from the pollen records (Barker & Arceo-Gomez ([2021](#ref-barker2021pollen)), Zhao *et al.* ([2019](#ref-zhao2019topology)), Alarcón ([2010](#ref-alarcon2010congruence))). In general when interaction networks are considered at coarse levels, such as the duration of a season, our perceptions regarding the generality of interactions at smaller time scales may be inflated relative to the actualized interactions within them, e.g. a week (CaraDonna & Waser ([2020](#ref-caradonna2020temporal))). These results indicate a possibility that at even finer levels *Bombus* display high amounts of floral fidelity within foraging bouts, an observation which implies that part of the reason for the high efficiency of *Bombus* as a pollinator might partially be related to their lack of movement of hetero-specific pollen (Brosi & Briggs ([2013](#ref-brosi2013single)), Ashman & Arceo-Gómez ([2013](#ref-ashman2013toward)), Galloni *et al.* ([2008](#ref-galloni2008visitor)), Brosi ([2016](#ref-brosi2016pollinator))). The mechanisms behind this observed fidelity are likely related to pollen nutritional values, specifically high concentrations of protein, and the absence of particular amino acids required for larval development in other flower more commonly used by workers (Genissel *et al.* ([2002](#ref-genissel2002influence)), Tasei & Aupinel ([2008](#ref-tasei2008nutritive)), Goulson *et al.* ([2005](#ref-goulson2005causes)), Goulson *et al.* ([2008a](#ref-goulson2008diet)), Hanley *et al.* ([2008](#ref-hanley2008breeding))).

Also regarding the case study, some foraging preferences of *Bombus*, at this field site and across several localities globally emerge. These suggest the need for land managers to maintain relatively high amounts of members of the plant families Fabaceae, Boraginaceae, and Ranunculaceae, in Western North American montane landscapes (Goulson *et al.* ([2005](#ref-goulson2005causes)), Goulson ([2010](#ref-goulson2010bumblebees)), Liang *et al.* ([2021](#ref-liang2021evolutionary)), Bontsutsnaja *et al.* ([2021](#ref-bontvsutvsnaja2021bumble))). Numerous historic practices reduce the ability of many landscapes to support stable populations of *Bombus*. Historic livestock grazing involved the removal of many species known to have compounds toxic to cattle. In particular, the removal of locoweeds (*Astragalus* & *Oxytropis*) and larkspurs (*Delphinium*) were common across public lands administered by the U.S. Forest Service (Ralphs & Ueckert ([1988](#ref-ralphs1988herbicide)), Aldous ([1919](#ref-aldous1919eradicating)), Ralphs *et al.* ([2003](#ref-ralphs2003mechanism))). Further actions, generally initiated by early settlers, involved the channelization and incising of streams, culling of beavers, and leaving cattle concentrated on higher order stream banks, processes which lowered water tables and reduced the extent of stream-associated wetlands and the mesic meadows fringes which provide habitat for many species of ‘tall’ *Mertensia* (e.g. *M. ciliata*), to an extent *Delphinium barbeyi* and many species of native *Trifolium* (Dahl ([1990](#ref-dahl1990wetlands)), Naiman *et al.* ([1988](#ref-naiman1988alteration)), Belsky *et al.* ([1999](#ref-belsky1999survey)), Cooke & Reeves ([1976](#ref-cooke1976arroyos))). Fire suppression resulted in the succession of many Aspen (*Populus tremuloides*) groves to Conifer stands, decreasing the mosaic of age structured habitats in many landscapes, adversely effecting habitat for tall *Mertensia* species and several species of *Delphinium* (Brewen *et al.* ([2021](#ref-brewen202176)), Keane ([2002](#ref-keane2002cascading))). Finally the effects of Nitrogen deposition, given the West’s rapidly growing population still pose adverse effects on the abundance of species of Fabaceae at urban-rural interfaces (see Stevens *et al.* ([2018](#ref-stevens2018atmospheric)), Fenn *et al.* ([2003](#ref-fenn2003ecological))). Current solutions to the above issues, involve targeted burns, reintroduction of beavers and beaver habitat analogs, and the possibility of re-seeding a variety of ‘locoweeds’ and ‘larkspurs’ in areas now seldom used, or only used for early, grazing. The highly enthusiastic response of land managers, and homeowners, to plant *Ascelpias*, using genetically appropriate materials, to improve Monarch Butterfly (*Danaus plexippus*) habitat provides an effective framework for the latter (Oberhauser *et al.* ([2015](#ref-oberhauser2015monarchs)), Basey *et al.* ([2015](#ref-basey2015producing))).

We have concerns regarding the number of persons training to become and practice botany, and grave concerns regarding the funding mechanisms for floristic and field based botanical research and for centralized authorities to produce consensus opinions on alpha taxonomy (Prather *et al.* ([2004b](#ref-prather2004decline)), Kramer & Havens ([2015](#ref-kramer2015report)), Prather *et al.* ([2004a](#ref-prather2004implications)), Crisci *et al.* ([2020](#ref-crisci2020end)), Manzano ([2021](#ref-manzano2021flippant)), Stroud *et al.* ([2022](#ref-stroud2022botanical))). To reduce the effects of a low population density of botanists on the maintenance of and production of Flora’s and to foster meta-genomics across landscapes without field stations we utilized Species Distribution Modelling to generate predictive species lists. In this proof-of-concept example we performed several iterations of modelling runs, and several approaches (i.e. the ‘linear models’, and the ‘machine learning’), which took notable amounts of compute power. We suspect the possible deleterious nature of this endeavor may be reduced by: 1) more field surveying by crews will reduce the need to generate as many species 2) fewer runs of models, 3) only running machine learning models which do not require an explicit process to reduce spatial autocorrelation. However, given the time required to perform all aspects of a study, even our amount of computation was negligible. Further, we are very optimistic about the possibility for persons to perform these tasks, as mentioned we utilized roughly only one quarter of the records which were digitally available for presence, and we suspect others will have enough records to perform this process nearly anywhere else in the temperate. In certain scenarios modelling of predicted species via more formally tailored S(tacked)-SDM or J(oint)-SDM approaches may be beneficial (Wilkinson *et al.* ([2021](#ref-wilkinson2021defining)), Pinto-Ledezma & Cavender-Bares ([2021](#ref-pinto2021predicting)), Schmitt *et al.* ([2017](#ref-schmitt2017ssdm))).

Tandem to the lack of continued expertise required to generate and maintain species lists, is the expertise required to continue tracking when major phenological events occur in many plant species at relatively fine scales or under novel climates. Knowledge of these events is currently limited to general time periods of only a handful of phenological events and groups of organisms (e.g. flowering initiation, or trees) (Prather *et al.* ([2004a](#ref-prather2004implications)), Li *et al.* ([2016](#ref-li2016responses))). While many programs and initiatives exist to collect phenological information on subsets of easily identifiable charismatic species to detect major trends in phenology, these capture only a subset of the extent diversity (Betancourt *et al.* ([2005](#ref-betancourt2005implementing)), Havens *et al.* ([2007](#ref-havens2007chicago))). In many instances it appears that while landscapes respond similarly to environmental variables which predict phenological responses, that individual species vary widely in their responses to similar environmental cues, or respond to different cues (Augspurger & Zaya ([2020](#ref-augspurger2020concordance)), Xie *et al.* ([2015](#ref-xie2015deciduous)), Xie *et al.* ([2018](#ref-xie2018predicting)), CaraDonna *et al.* ([2014](#ref-caradonna2014shifts))). As can be seen here, predictions of when a single, major phenological event occurs is already data limited. A more promising approach for the tropics may lay in utilizing circular statistics (Park *et al.* ([2022](#ref-park2022herbarium))).

The nearly complete Plant and Fungal Tree of Life (PAFTOL) will provide a comprehensive phylogenetic backbone of the entire plant kingdom, and the inclusion of A353 probes with lineage specific probe sets is common in producing massive genetic datasets (Baker *et al.* ([2021b](#ref-baker2021exploring))). We predict that the A353 probes which it is utilizing to work nearly immediately for DNA barcoding of whole plant material, and that more elaborate validation studies in controlled metabarcoding settings, utilizing existing experimental designs, will have favorable results (Bell *et al.* ([2017](#ref-bell2017applying)), Bell *et al.* ([2019](#ref-bell2019quantitative)), Bell *et al.* ([2021](#ref-bell2021comparing)), Lamb *et al.* ([2019](#ref-lamb2019quantitative))). In particular the harvesting of loci with more variation in certain lineages, and or with more variable flanking regions, will prove promising for identifying closely related plant material. We suspect that conserved reaches of genes resulted in the high amounts of reads in somewhat obscure species. Given that the A353 loci are nuclear, single copy, and a variety are present the possibility of identifying target loci for quantitative purposes is high, without continual PCR enrichment is possible; this would align with relatively high efficacy of WGS (Lang *et al.* ([2019](#ref-lang2019genome)), Peel *et al.* ([2019](#ref-peel2019semi)), Bell *et al.* ([2021](#ref-bell2021comparing))). Recent evidence indicates that the potential for identifying nearly cryptic taxa and even infra-specific inference, of either whole plant material, and perhaps in metagenomic context are possible (Ottenlips *et al.* ([2021](#ref-ottenlips2021resolving)), Wenzell *et al.* ([2021](#ref-wenzell2021incomplete)), Loke et al. in prep, Slimp *et al.* ([2021](#ref-slimp2021potential)), Beck *et al.* ([2021](#ref-beck2021palmer))). We further believe that in synthetic phylogenetic trees - with incorporation of NGS backbones - will allow in automatic reassignment of reads as a function of phylogenetic distance with measures of uncertainty (Hinchliff *et al.* ([2015](#ref-hinchliff2015synthesis)), Smith & Brown ([2018](#ref-smith2018constructing)), Baker *et al.* ([2021a](#ref-baker2021PAFTOL))).

**Conclusion**

We believe that the combination of spatial and temporal models, united and guided by localized natural history knowledge, provides the essential components of a framework for approaching the coarse elucidation of ecological interactions using DNA Barcoding. Herein we crudely utilized this thinking via binary outcomes, should a species predicted be predicted present or not? Is it unequivocally flowering or not? Myriad data show biological systems and ecological interactions have more variance than can be reasonably discretely parsed. We expect that within a framework developed from our preliminary works studies of pollinator behavior may be enacted via this approach at a landscape level, e.g. the scale of an entire drainage basin such as the Gunnison which is quickly becoming one of the worlds few model ecosystems. We hope that the A353 probes as tools for metabarcoding play a role in these endeavors.

**AUTHOR CONTRIBUTIONS:** R.C.B conducted botanical collections, conducted all molecular lab work, lead all analyses, and writing. J.E.O conceived, designed, and conducted ecological fieldwork, assisted with analyses, and writing. P.J.C conducted ecological fieldwork, assisted with ecological analyses, and writing. E.J.W. prepared, imaged, and collected trait data on pollen reference slides, and assisted with analysis of trait data and writing a dichotomous key. S.T. assisted with spatial analyses and writing. J.B.F. conceived, and designed lab work, analyses, and integration of approaches, assisted with writing, and secured funding for molecular work.

**ACKNOWLEDGMENTS:** Nyree Zerega for assistance obtaining herbaria loans and accessioning our collections at CHIC. Pat Herendeen for assistance with virtually all aspects of preparing pollen vouchers and the identification process. Hilary Noble, Zoe Diaz-Martinez, Angela McDonnell, & Elena Loke for assistance with genomic library preparation. Ian Breckheimer for sharing the SDM predictor variables. We thank the curators at the following herbaria for supplying tissue: Ben Legler at Stillinger (ID), Charles (Rick) Williams at Ray J. Davis (IDS), (B)Ernie Nelson at Rocky Mountain (RM); and the collectors: D. Knoke, L. Brummer, J. Boyd, C. Davidson, I. Gilman, M. Kirkpatrick, S. McCauley, J. Smith, K. Taylor, & C. Williams. David Giblin & Mare Nazaire for sharing relevant sections of an advanced draft of FNA V. 15. The Bureau of Land Management is thanked as many plant specimens were collected by R.C.B as a partner or contractor to the agency; Sarah Burnett and Lauren Price are thanked for sharing AIM data. Sanda and New England Biotech are gratefully acknowledged for technical support and generously sharing samples. T.C.H. Cole for sharing the Angiosperm Phylogeny 4 colour palette. The Program in Plant Biology and Conservation is thanked for funding. The holdings of the following herbaria were essential for this project: AK, ALTA, ASU, BABY, BC, BM, BMO, BOON, BRIT, CANB, CAS, CHSC, CM, CMN, CNS, COLO, CONN, CS, CSU, DAV, DBG, DES, ENCB, F, FR, G, GH, GZU, IAC, K, KR, KSP, KSTC, KU, LD, LOB, LSU, MA, MACF, MEL, MICH, MIL, MIN, MNHN, MO, MT, MW, NCSC, NSW, NY, O, OBI, PI, RBG, RSA, SD, SDSU, SFV, TENN, TRT, UA, UAC, UAM, UAZ, UBC, UBC, UCR, UCS, UCSB, UMO, UNM, UPS, US, USCH, USF, USU, UTEP, UWBM, V, VT, W, WSCO, WU, XAL, YPM, Z.

**CONFLICT OF INTERESTS** The authors declare no conflicts of interest.

**PEER REVIEW** The peer review history for this document is available at …

**DATA AVAILABILITY STATEMENT** The queries required to download all data used in this project are located in… All novel sequencing data are located at NCBI…

**References**