Metagenomic Barcoding of Pollen Loads Offers Insights on the Foraging Patterns of Queen Bumble Bees

.  
1) DNA Barcoding has been remarkably successful in nearly all kingdoms of life and has allowed for the rapid analysis of ecological assemblies. Successful DNA barcoding in the plant kingdom has been more difficult than other kingdoms. Due to this understanding plants in ecological contexts and understandings of their synecology in some instances may begin to lag behind other kingdoms.

2) Here we utilize hyb-seq, museum studies, and species distribution modelling, to detect the plant species present in pollen loads collected from Queen Bumble Bees.

3) By utilizing Species distribution modelling we allow for one to process hyb-seq data, create user specified sequence databases which may use MORE ACCURATE alignment algorithms on personal computers over realistic time periods.

4) We show that hyb-seq using the Angiosperms 353 probes, which are currently being used in the largest ever plant systematic endeavor, offers significant promise to metagenomic approaches in real world scenarios.

5) We conclude that these probes offer promise for the identification of plant tissue in both single sample, and metasample contexts.

# 1 | INTRODUCTION

The inability to reliably identify plants to the level of species often leaves our understanding of ecosystem function and interactions wanting. Current methods to ameliorate this situation include: ignoring these ecologically relevant levels of detail, revisiting plots as diagnostic material becomes temporally available, assistance from taxonomic specialists, or the use of barcoding or other molecular techniques. These approaches are untenable in light of the benefits offered by: species in several morphologically difficult genera which serve as bioindicators, preferred partners in ecological interactions, as well as an increasing lack of taxonomic experts (Hebert *et al.* ([2003](#ref-hebert2003biological))). Many genera, especially with the formalized advent of integrative taxonomy, have species which are well defined based upon ecological and behavioral rather than morphological properties, the identification of these taxa in degraded areas or without their mutualistic partners is fraught with difficulty. Hindering an understanding of the breadth of habitat which some species occupy, and the interactions they have with other species.

The identification of many plant species to terminal taxon is an essential component of nearly all land management programs, where many species in the same genus (e.g. Sagebrush - *Artemisia* L., Willows - *Salix* L., and Sedges - *Carex* L.) serve as bioindicators (respectively for ‘rangelands’, streams, and wetlands), as well as in academic research (Gage & Cooper ([2013](#ref-Gage2013HistoricalRO)), AIM). This endeavour is often mired by lack of diagnostic characters (e.g. flowers, fruits, roots or combinations thereof), and increasingly the description of cryptic species (Janzen *et al.* ([2017](#ref-janzen2017nuclear)), Oliver *et al.* ([2009](#ref-oliver2009cryptic))). Solutions to this problem are wanting, certain programmes have relied increasingly upon revisiting field sites to identify material using morphological or chemical approaches, whereas academic research has often used high copy number plastid genes as barcodes (Rosentreter et al. 2021, MORE MORE). However, both approaches have significant downsides, the former resource intensive at the landscape scale - and often does not work, while the latter seldom works due to a lack of variability in the currently available barcodes (Liu *et al.* ([2014](#ref-liu2014identification))).

Recently barcoding (the identification of a sample from a single organism *e.g.* a piece of leaf), 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))). With plants the identification of members of certain clades using barcoding has been quite successful (REF), whereas with other clades results have been elusive (Liu *et al.* ([2014](#ref-liu2014identification)), Group *et al.* ([2011](#ref-china2011comparative)), Coissac *et al.* ([2012](#ref-coissac2012bioinformatic))), however 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 challenges with barcodes include the utilization of high-copy number sequences are associated with their rates of divergence, gene tree conflict, and hybridization (Coissac *et al.* ([2016](#ref-coissac2016barcodes)), Fazekas *et al.* ([2009](#ref-fazekas2009plant))). Particular challenges with the utilization of high-copy number sequences are associated with their rates of divergence, gene tree conflict, and hybridization (Coissac *et al.* ([2016](#ref-coissac2016barcodes))).

Currently the largest plant systematic endeavor ever undertaken, the Kew Plant and Fungal Tree of Life (PAFTOL), is approaching completion (Baker *et al.* ([2021](#ref-baker2021PAFTOL))). This dataset will contain Hyb-Seq data from at least one species representing each genus in the plant kingdom using the popular A353 probes (Baker *et al.* ([2021](#ref-baker2021PAFTOL))), resulting in over 14,000 represented species. These publicly available data serve to provide a taxonomically comprehensive backbone for plant metabarcoding. Data from the 10kP project, which seeks to develop reference genomes from a phylogenetically diverse suite of plants will contribute many more records upon it’s intended completion, now slated to be by 2030, similar projects which seek to sequence high amounts of genomes in regions e.g. the ‘Darwin Tree of Life’ are being undertaken which will contribute data applicable to enormous spatial domains (Cheng *et al.* ([2018](#ref-cheng2018tenkp)), Life Project Consortium *et al.* ([2022](#ref-darwin2022project)), Lewin *et al.* ([2022](#ref-lewin2022biogenome))). These data will promote the ability to apply metabarcoding to resolve a diversity of questions relevant to theoretical and applied ecology (cite). However, the application of metabarcoding still face challenges relating to the enormity of the genomic datasets and the computational power required to process sequence data. Herein we have resolved major components of the problems of identifying plant material without diagnostic morphological character states using the Angiosperms353 (A353) Hyb-Seq probes (Johnson *et al.* ([2019](#ref-johnson2019universal))), and custom species sequence databases derived via species distribution modelling, and temporal filtering.

To increase the quality of metabarcoding results in plants, we suggest reducing the number of possible plant species candidates by generating user selected sequence databases relevant to the the study region and its ecological characteristics (CITE !?). To achieve this goal, we first create a list of candidate species using digital collections gleaned from herbaria, survey work, and citizen science (e.g. iNaturalist), from a region exceeding the study area. To these candidate species, modelling approaches - such as logistic regression, may be used to identify taxa which warrant further exploration e.g. modelling to determine their possibility of presence in metabarcoding samples. We then use species distribution models to create potential distribution maps for the candidate species 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. Species distribution models (SDM’s) examine the ecological conditions associated with known occurrence of a species to identify where else in the study area might suitable habitats be found. This approach has the additional benefit of greatly reducing the size of a sequence database, which allows for the usage of genomic size data on personal computers. This can also significantly reduce processing time, particularly as as most next-generation sequence data is deposited as raw-sequence reads.

Considerable amounts of species interactions vary along time (CaraDonna *et al.* ([2021](#ref-caradonna2021seeing))). For the tropics and subtropics, contrasts in the flowering periods of many plant species, can provide an additional filter for identifying material in many metagenomic samples (Janzen ([1967](#ref-janzen1967synchronization)), Newstrom *et al.* ([1994](#ref-Newstrom1994ANC))). In temperate regions, pollination interactions also vary temporally (CaraDonna *et al.* ([2017](#ref-caradonna2017interaction))), however the overall shorter extent of the active growing season in these systems results in the presence of few to any natural breaks in these systems which reduces the utility of these to operate as filters in the post-processing of sequence matches. Nonetheless, we work through a process which seems applicable to the tropics and subtropics to utilize the temporal dimension for classifying sequencing results.

We apply these metagenomic and informatics approaches to determine whether the foraging record of Queen Bumble Bee’s is consistent across direct observations and the pollen record, an incongruency in several floral visitation networks involving smaller bodied fauna (Barker & Arceo-Gomez ([2021](#ref-barker2021pollen)), Zhao *et al.* ([2019](#ref-zhao2019topology)), Alarcón ([2010](#ref-alarcon2010congruence))). The two foraging phases of the Queen Bumble Bee life cycle is essential to 1) increase their weight before diapause, 2) increase their ovary weights while establishing their recently found nests, both of these time periods represent potential demographic bottlenecks in bumble bee populations (Sarro *et al.* ([2022](#ref-sarro2022bumble))). Bumblebees are one of the only groups of insects with unequivocal quantitative evidence for numerous populations declines, while simultaneously serving as the most effective pollinators in temperate montane ecosystems (Cameron & Sadd ([2020](#ref-cameron2020global)), Goulson *et al.* ([2008](#ref-goulson2008decline)), Williams ([1982](#ref-williams1982distribution)), Colla *et al.* ([2012](#ref-colla2012assessing)), Bergman *et al.* ([1996](#ref-bergman1996micrometeorological)), Bingham & Orthner ([1998](#ref-bingham1998efficient))). *Heterogeneous montane ecosystems simultaneously represent one of the most ecologically resilient areas in the temperate and oftentimes offer the sole potential refugia for multiple dimensions of biodiversity under climate change, whilst experiencing the most abrupt changes in mean annual temperature (CITE). An immediate understanding of how to manage previously overlooked keystone insect species, such as bumble bees, is essential if these refugial ecosystemn services are to be utilized while maintaining their current species richness.*

# 2 | METHODS

## Study System & Field Work

Observations and sample collection was conducted at The Rocky Mountain Biological Laboratory (RMBL; 38°57.5” N, 106°59.3” W (WGS 84), 2900 m.a.s.l.), Colorado, USA (APPENDIX 1 for site information). Pollinator observations of Bombus Latreille spp. (Apidae Latreille) were conducted from June - August of 2015 in six study sites characterized by high-montane/subalpine Parkland vegetation communities. Observations of Bombus foraging took place for one hour at each field site in three 100m transects, where all flowers were also counted and placed into abundance bins. Corbiculae loads were, non-lethally, collected once from all Queen individuals encountered.

### 2.1 | Spatial Analyses

#### 2.1.1 Candidate Species

We downloaded from the Botanical Information and Ecology Network ‘BIEN’ (Maitner ([2022](#ref-bien2022))) all records adjacent to the field sites to develop an ecologically relevant list of vascular plant species, with expected biotic pollination, which may be present at the study area . We then generated Species Distribution Models (SDMs) to predict their distribution throughout the study area. These maps of potential distribution served as a reference to reduce the list of species to include in the genomic sequence databases.

In order to minimise the number of species for which SDM’s were to be generated, BIEN was queried at a distance of up to 100km from our study area and all plant species records were downloaded. ***In order to emulate the stochasticity of botanical collecting, this dataset was bootstrap re-sampled 250 times, with 90% of samples selected, to create a testing dataset***. The median of the logistic regression assessing the probability of occurrence of a species record as a function of distance from the study area was used as a threshold distance, under which, to include species as candidates for distribution modelling.

#### 2.1.2 Distribution Modelling

We used all occurrence records from BIEN (n = 23,919) within a 50km border of the Omernik level 3 ecoregion, which includes the study area *(No. 21 “Southern Rockies”)* to construct the species distribution model (Omernik ([1987](#ref-omernik1987ecoregions))). These records were copied into two, initially identical, sets, one for generating machine learning models (ML; Random Forest, and Boosted Regression Tree’s), and the other for Generalised Linear (GLM) and Generalized Additive Models (GAM) (Barbet-Massin *et al.* ([2012](#ref-barbet2012selecting))). **Ensembled predictions have been shown to outperform their constituent models, on average, and to reduce the ecological signal to the analytical noise of individual runs (Araujo & New (**[**2007**](#ref-araujo2007ensemble)**)). No single method of producing SDMs has been shown to universally outperform others when faced with a large and diverse number of applications, in our case a great number of species with differing biologies and ecologies (Elith\* *et al.* (**[**2006**](#ref-elith2006novel)**), Qiao *et al.* (**[**2015**](#ref-qiao2015no)**)). In the spirit of these findings, multiple families of models, which can be generated together as they have similar requirements regarding the number and ratios of Presence to Absence records were ensembled together (Barbet-Massin *et al.* (**[**2012**](#ref-barbet2012selecting)**)).**

We then generated 4,029 absence points , locations where the focal taxon is anticipated missing, through a random stratification of 19% of the land cover in the area and included them in (BLM CITATION - need appropriate format for journal). To achieve a larger absence dataset, we generated 1,000 pseudo-absence records for each taxon by randomly selecting coordinates located at least 10km away from an occurrence record. For ML 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 the presence records; the required number of absence records were then randomly sampled.

We used 26 environmental variables at 30m resolution to predict the potential distribution of each species, 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 6*). **These publicly available datasets, were selected as they …** . 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))).

Modelling: Random Forest and Boosted Regression Trees, were sub sampled with 30% test and two replicates each before weighted ensemble based on True Skill Statistics (tss) (Naimi & Araujo ([2016](#ref-sdmPackage))). Generalised linear models (GLM) and Generalised additive models (GAM) with 30% sub sampling and three replicates each were also ensembled using the tss (Naimi & Araujo ([2016](#ref-sdmPackage)), @). TSS was chosen as the ensemble criterion as it has been shown to work across a range of species occurrences prevalences (Allouche *et al.* ([2006](#ref-allouche2006assessing))). The results of these models were extracted on a cell-by-cell basis to a polygon feature derived from a minimum-spanning tree which encompasses the study sites, and species from either ensemble with greater than 50% mean habitat suitability across all cells were considered present for further purposes (Prim ([1957](#ref-prim1957minimum))).

535 species were modelled using Generalized Linear Models and Generalized Additive Models. 534 species were modelled using Random Forest and Boosted Regression Trees. 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 (CITE AIM AND Occdownload Gbif.Org ([2021](#ref-gbifDL2021sdms))). Additional novel absence records were generated from the AIM dataset to create a dataset where each species has balanced presence and absences. 11 or more paired presence and absence records were required for this testing, resulting in 334 species being included in the logistic regression (Mdn = 110.0, = 223.1, max = 1568 record pairs used) with a 70% test split (Kuhn ([2022](#ref-caret))).

## 2.2 | Molecular Lab Work

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, U.S.A.

#### 2.2.1 | Reference Plant Library Generation

Using five years of observational data on *Bombus* Queen Bee foraging at these studies sites, we identified the plant taxa most frequently visited by Queens across all years. We sequenced the 12 most commonly visited taxa twice using samples from one site within the Gunnison River Drainage and one individual from another population. In addition, for any of these 12 focal species which did not have a congener pair in this filtered sample, we included a congener - or a species from a closely related genus to serve as an outgroup. We also sequenced another 15 abundant taxa commonly visited by *Bombus* workers, based on the aforementioned data set (*APPENDIX 4*).

#### 2.2.2 | Plant Genomic DNA Extraction

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. DNA was quantified using a Nanodrop 2000 (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and Qubit fluorometer (Thermo Fisher Scientific).

#### 2.2.3 | Pollen Genomic DNA Extraction

Pollen genomic DNA was extracted from corbiculae using a CTAB based protocol modified from Lahlamgiahi et al. and Guertler et al. (2014, 2014). A SDS extraction buffer (350µL , 100mM Tris-HCl, 50 mM EDTA, 50 mM NaCl, 10% SDS v/v., pH 7.5) was added followed by vortexing 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 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.

To precipitate the DNA, chilled Isopropyl alcohol & 3 mM Sodium acetate (5:1) equivalent to 2/3 of the volume of sample were added, with 1 hour of chilling at -20°C, followed by 10 minutes of centrifuging at 13,000 rpm. The supernatant was pipetted to a new 1.5 ml centrifuge tube, and 70% EtOH (400 µL) were added before chilling at -20°C for 20 minutes followed by centrifugation at 13,000 rpm for 10 minutes. Both tubes were then washed with 75% EtOH (400 µL), inverted, centrifuged at 13,000 rpm for 4 minutes, and the solution discarded, then washed with 95% EtOH (400 µL) , inverted, centrifuged at 13,000 rpm for 4 minutes, and the solution discarded. Pellets were dried at room temperature overnight before resuspension in Nuclease free H2O. 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 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.

#### 2.2.4 | Fragmentation, Library Preparation & Target Enrichment

Library preparation was performed using the NEBNext Ultra II FS-DNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, Massachusetts, USA) using slightly modified 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 with ½ volume of 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. Products were analysed on 4% agarose gels, and a Qubit fluorometer. Libraries were pooled and enriched with the Angiosperms 353 probe kit V.4 (Arbor Biosciences myBaits Target Sequence Capture Kit) by following the manufacturer’s protocol and Brewer et al. 2019. Sequencing was performed using an Illumina mi-Seq with 150-bp end reads, (NUSeq Core, Chicago, Illinois).

### 2.2.5 | Computational Processes and Analyses.

#### 2.2.5.1 | Reference Library Data Processing

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 were generated using HybPiper using target files created by M353 (Johnson *et al.* ([2016](#ref-johnson2016hybpiper)), McLay *et al.* ([2021](#ref-mclay2021new))).

#### 2.2.5.2 | Sequence Identification

A custom Kraken2 database was created by downloading representative species of each genus 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 before being placed into the database. The Kraken2 database was built using default parameters. Kraken2 was run on sequences using default parameters (*APPENDIX 5*). Following Kraken2, Bracken was used to classify sequences to terminal taxa (Lu *et al.* ([2017](#ref-lu2017bracken))). Results from both Kraken2 and Bracken, results were reclassified manually to identify terminal taxa. For example, when only a single species of a genus was known in the study area, but our database used a representative of another taxon in the genus, this species was coded as the result. The re-coding of sequences from another representative species for the genus to the sole RMBL representative allowed the identification of XX & % more species.

#### 2.2.5.3 | Identification of Sequence Matching Loci

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

## 2.2.5.4 | Morphological Pollen identification

To develop a reference library of pollen grains which may be present in corbiculae loads, an image reference collection of fuchsin-jelly stained (Beattie ([1971](#ref-beattie1971technique))) slides was assembled from slides previously prepared by the authors (n = 21), and other researchers (n = 38) (Brosi & Briggs ([2013](#ref-brosi2013single))). Using five years of observational data on *Bombus* Queen Bee foraging at these studies sites (Ogilvie unpublished), as well as the Vascular Plant Checklist (Frase & Buck ([2007](#ref-fraser2007vpc))), an additional 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 non accessioned herbarium collections to supplement the number of species and clades covered (Appendix 3).  
We used Divisive Hierarchical Clustering techniques to determine 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 clustering 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 NO.). This key was then used to identify the pollen grains sampled from corbiculae loads to morphotypes in a consistent manner. To prepare the pollen slides from corbiculae, all corbiculae loads were broken apart and rolled using dissection needlepoints to increase heterogeneity of samples. *Cerca* 0.5mm^2 of pollen was placed onto a ~4mm^2 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 nail polish; all samples are noted in *APPENDIX 3*. 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 an unique morphotype than all grains in that transect were also identified and counted. 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 2*.

### 2.3 | Temporal Analyses

To estimate the duration of dates in which plant species were flowering weibull estimates of several phenological parameters all spatially modelled taxa were developed (Belitz *et al.* ([2020](#ref-belitz2020accuracy)), Pearse *et al.* ([2017](#ref-pearse2017statistical))). Only BIEN records which occurred in the Omernik Level 4 Ecoregions within 15km of the study area (n = 5 Level 4 Ecoregions, or conditionally 6 if enough records not be found in the nearest 5), and which were from herbarium records were included. To remove temporally irrelevant herbarium records, i.e. material collected during times which flowering is impossible at the study area due to snow cover, we used the SnowUS dataset (Iler *et al.* ([2021](#ref-iler2021conceptual)), Tran *et al.* ([2019](#ref-tran2019cloud))) from 2000-2017 was analyzed for the first three days of contiguous snow absence, and 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. 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.

### 2.4 | Floral Observations

# 3 | RESULTS

## 3.1 | Spatial Analyses

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 was used as a threshold distance to include species for distribution modelling. 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, where 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.

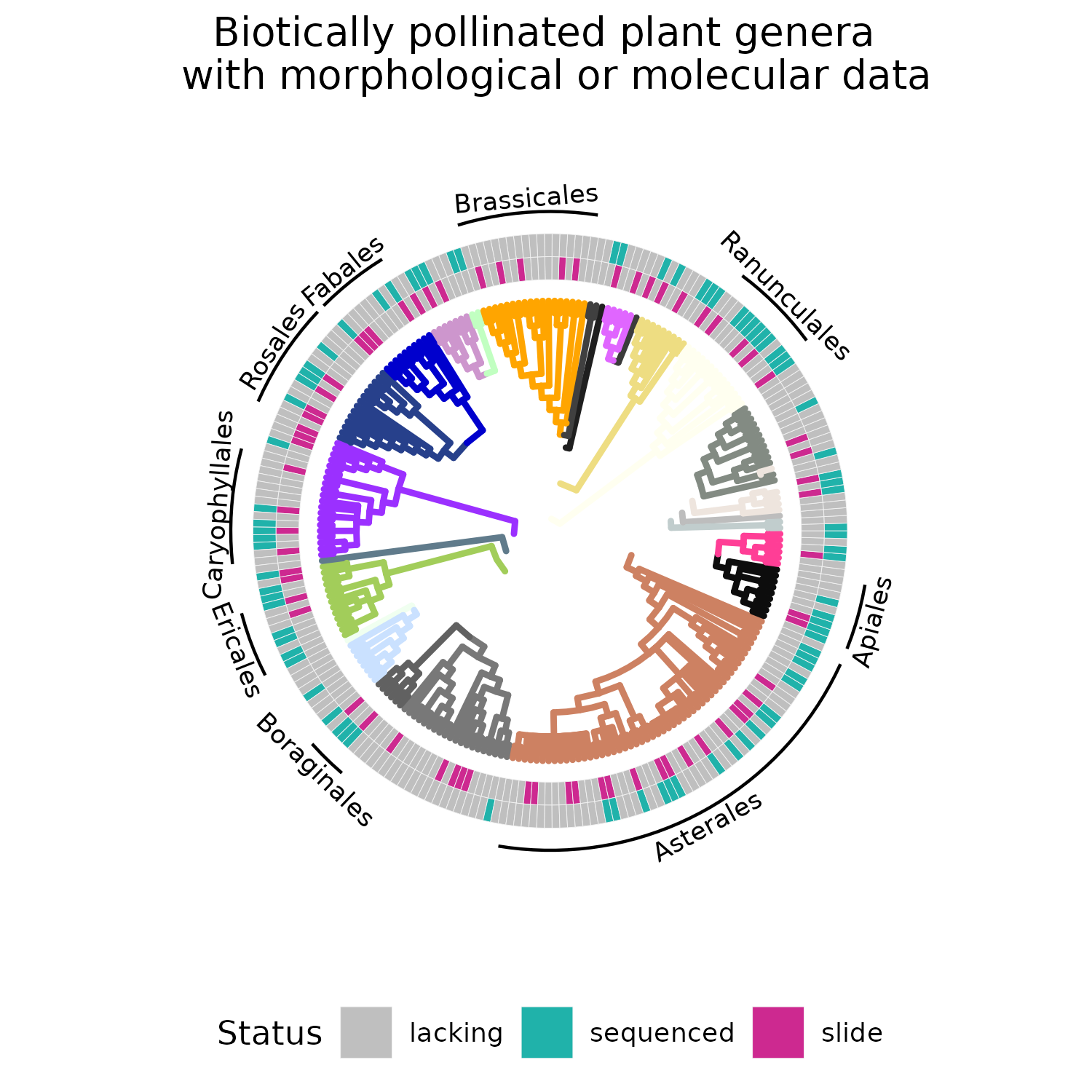
Across the entire spatial domain of modelling all ensembled models (n = 968) had an accuracy of 0.84 (95% CI 0.8356 - 0.8443), kappa 0.68, p-value < 0.001, sensitivity = 0.80, specificity = 0.87, AUC = 0.92.

At the field site, of the 554 vascular plants with biotic pollination syndromes, the 493 ML ensembles accurately predicted the presence of 362 (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%). The balanced accuracy of the ensembled models is 0.627 (Sensitivity = 0.340, Specificity 0.914). Of the 554 vascular plants with biotic pollination syndromes, 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 554 vascular plants with biotic pollination syndromes in the flora 13 (2.3%) were in the Orchid family and 41 (7.4%) are non-natives, both of which are restricted from the database, and can only reduce the number of true predicted presences by roughly 10%.

At the six study plots, 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.

## 3.2 | 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 7*. 60 samples were counted and based on rarefaction **had over % of expected morphotypes found** (morphotype richness, = 4.5, Mdn = 4, min = 1, max = 9), all samples had expected morphotype diversity reach the asymptote *APPENDIX 8*. The number of counted pollen grains in each sample range from (*MIN* - 16,293, = 2788.685, Mdn = 1453).



A caption

*note this figure is draft mode, i reached out to C.H. Cole to get the official APG colors so we are gonna colour edges with that, I have also drawn phylo pics for almost all the labelled order and need to add them in !*

## 3.3 | Metabarcoding Pollen identification

54 corbiculae loads had DNA extracted and underwent various steps towards hyb-seq, in the end a total of 44 corbiculae samples were sequenced, 7,752,353 reads were recovered from sequencing. 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 X**.

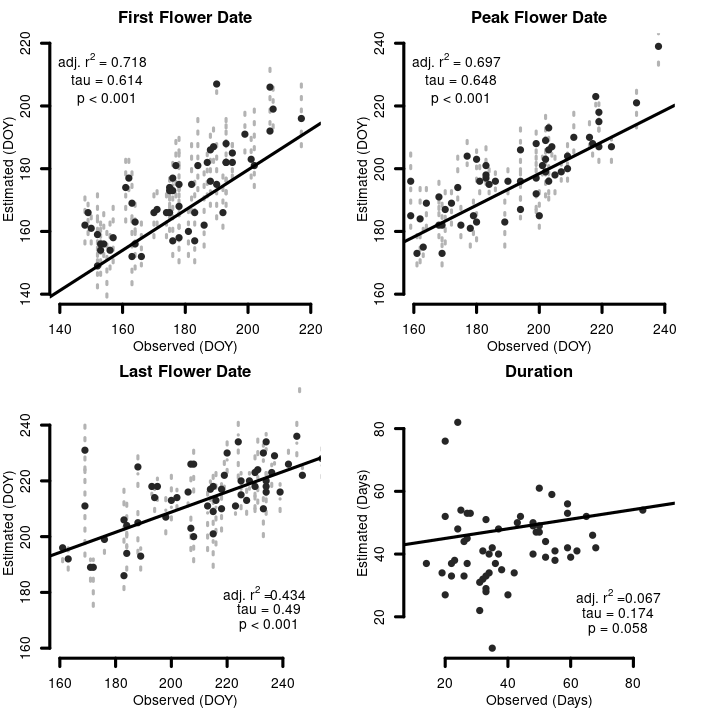
After trimming 7,865,680 sequences remained. 10,682,538 reads were matched using Kraken, of the reads classified by Kraken 10,160,768 reads were matched using Bracken, of the reads classified by Kraken 7,302,876 reads were matched using BLAST.

Based upon subjective review of the three classifiers, BLAST was chosen as the classification method which yielded the most probable results, and it’s values were used for all subsequent analyses.

## 3.4 | Temporal Analyses

The first date of modeled snow melt in the Gothic area (n = 17, = 137.9, Mdn = 135, 3rd quantile = 151), and the first date of a consistent winter snow base (n = 17, = 299.9, Mdn = 300, 1st quantile = 291) from 2000-2017, were used as delimiters for the inclusions of herbarium records in modelling. Of the **500** 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 ( = 35.01657, 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 in general just beneath the elevation bands occupied by the five ecoregions around the study area had weibull estimates also calculated ( = 13.86885, Mdn = 13, max = 24).

Only 58 of these 388 species (n = 34.56897, Mdn = 31) were able to be compared to plot based observational data from the long term (1974–2012) dataset. Of these species relatively high accord was observed between the long-term ground truthed dataset, and the modelled species. There was very strong evidence that the weibull estimates were positively associated with the observed onset (r2 = 0.72, p < 0.0001, tau = 0.61) and peak (r2 = 0.70, p < 0.0001, tau = 0.65) of flowering, and that the number of herbarium samples had a moderate effect on the estimates (p = 0.004 and p = 0.034 respectively). There was very strong evidence that the weibull estimates had a positive association with the observed cessation of flowering (r2 = 0.4339, p < 0.0001, tau = 0.489), however their was no evidence that sample size had an effect (p = 0.349). There was moderate evidence that the weibull estimates, with an effect of sample size, had a weak positive association with the observed duration of flowering (p = 0.0401, r2 = 0.07, tau = 0.17).



A caption

## 3.5 | Floral Observations

The six sites were surveyed for a total of 52 hours from May 27-July 27. A total of 723 queen-pollen foraging interactions were observed (range per bee species by week range = 1 - 18, = 3.46, Mdn = 2), with a range of total observed interactions per bee species across this time period (min = 1, = 59.08, Mdn = 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 min = 1 - 20, = 3.51, Mdn = 2), with a range of total observed interactions per plant species over this time period (min = 1, = 20.26, Mdn = 4, max = 141). The number of plant species which bees were observed interacting with varied more narrowly (range = 1 - 18, = 8, Mdn = 6).

A total of 66 corbiculae loads were collected from Bees, 64 of them from Queens.

# 4 | DISCUSSION

Although we were able to use an actually fine scale flora to determine the species present at the field site, we suspect a similar approach may be accomplished via quick species richness inventories at sites, and then utilizing a bootstrap approach akin to ours, to the taxa returned from databases to derive these estimates.

\*\* Mention in here that the general effect of sample size on predictions means that larger samples sizes are required for this to work!!!! \*\* Although our temporal results were lackluster, we note that our study area has an incredibly brief growing period. and we suspect these temporal results would be useful in sub-tropical and tropical ecosystems. FURTHER, the sites used for ground truthing the temporal flowering periods were not randomly selected across the study area, and cannot be used to make inference to the population across the entirety of the study area as we did here. Regardless both show good agreement on flower onset, peak flowering, and moderate agreement with flowering cessation. The disagreement in flowering cessation is perhaps due to more microclimates which retain water, rather than microclimates which allow the early accumulation of heat.

Fewer modelling runs for SDM’s likely to be effective for determining inclusion, elastic inclusion criteria. The actual dataset which was used for training and testing all of the models incorporated into SDM’s represented only roughly one quarter of the records available for such purposes. We consciously chose to do this in order to showcase the possibility of this approach working in less data rich areas.

Bayesian framework

Future Directions:

While at the time of writing this there are limited A353 sequence data, the Plant and Fungal Trees of Life (PAFTOL) project, which is sequencing at least a species of each genera in the plant Kingdom will produce sequence data from over 14,000 species. Given the extant publicly available genomic data, we conservatively estimate that upon completion of PAFTOL there will be no fewer than 15,500 species (4.4% of all ca. 350,000 plant species) for which sequence data of a majority of these loci exist (Govaerts *et al.* ([2021](#ref-govaerts2021world))). Accordingly, projects in the near future may increase the number of metagenomics samples while decreasing the need to create their own plant sequence reference libraries. As a result of PAFTOL the first ever comprehensive phylogenetic hypotheses of all plant genera will be presented. In tandem with an increased number of digitised and geo-referenced herbarium specimens, and monitoring programs in natural areas, we believe that geo-informatics, and phylogenetic inference will increase the ability of researchers applying this technique to identifying sequence reads. While our approach emphasises the use of this metagenomic technique for the purpose of identifying pollen, I argue the template and resources we provide here make this approach a suitable candidate for many plant metagenomic tasks. While we did not have the resources to explore the possibility of characterising infraspecific characteristics, preliminary results from others (Wenzell *et al.* ([2021](#ref-wenzell2021incomplete)), Loke et al. in prep) indicate a possibility for these probes to also collect data at the level of populations and individuals. \*\*

In regards to better understanding the foraging preferences of *Bombus* feeding in subalpine ecosystems. **JANE AND PAUL SET UP FOR NEAR FUTURE RESULTS?**

**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 all ecological fieldwork, assisted with 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. P.J.C assisted with ecological analyses and writing. J.B.F. conceived, and designed all lab work, analyses, 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. Ian Breckheimer for sharing the SDM predictor variables. Hilary Noble, Zoe Diaz-Martinez, Angela McDonnell, & Elena Loke for assistance with genomic library preparation. 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). We thank the original collectors of these specimens: D. Knoke, L. Brummer, J. Boyd, C. Davidson, I. Gilman, M. Kirkpatrick, S. McCauley, J. Smith, K. Taylor, & C. Williams. David Giblin 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 NEB are gratefully acknowledged for technical support and generously sharing samples.

**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…

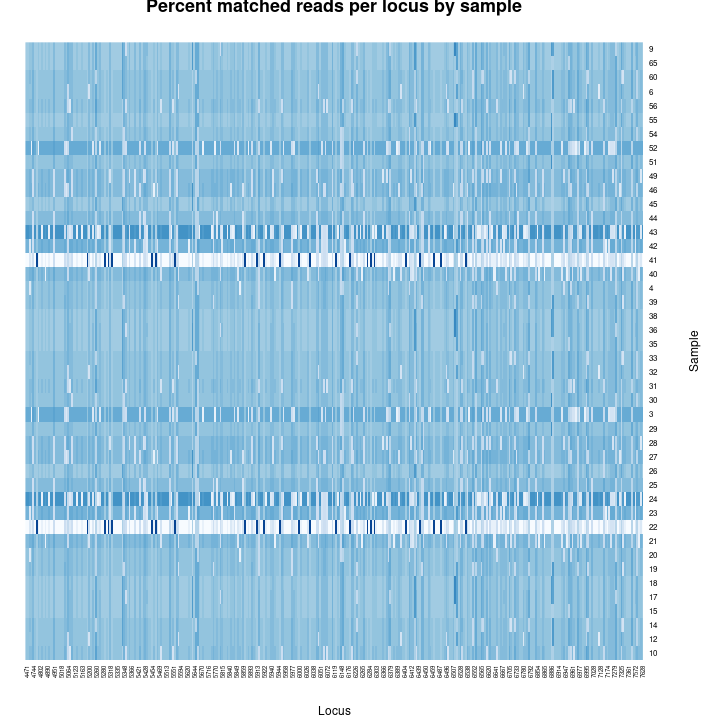
**ORCID**

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>

# References

# Supporting

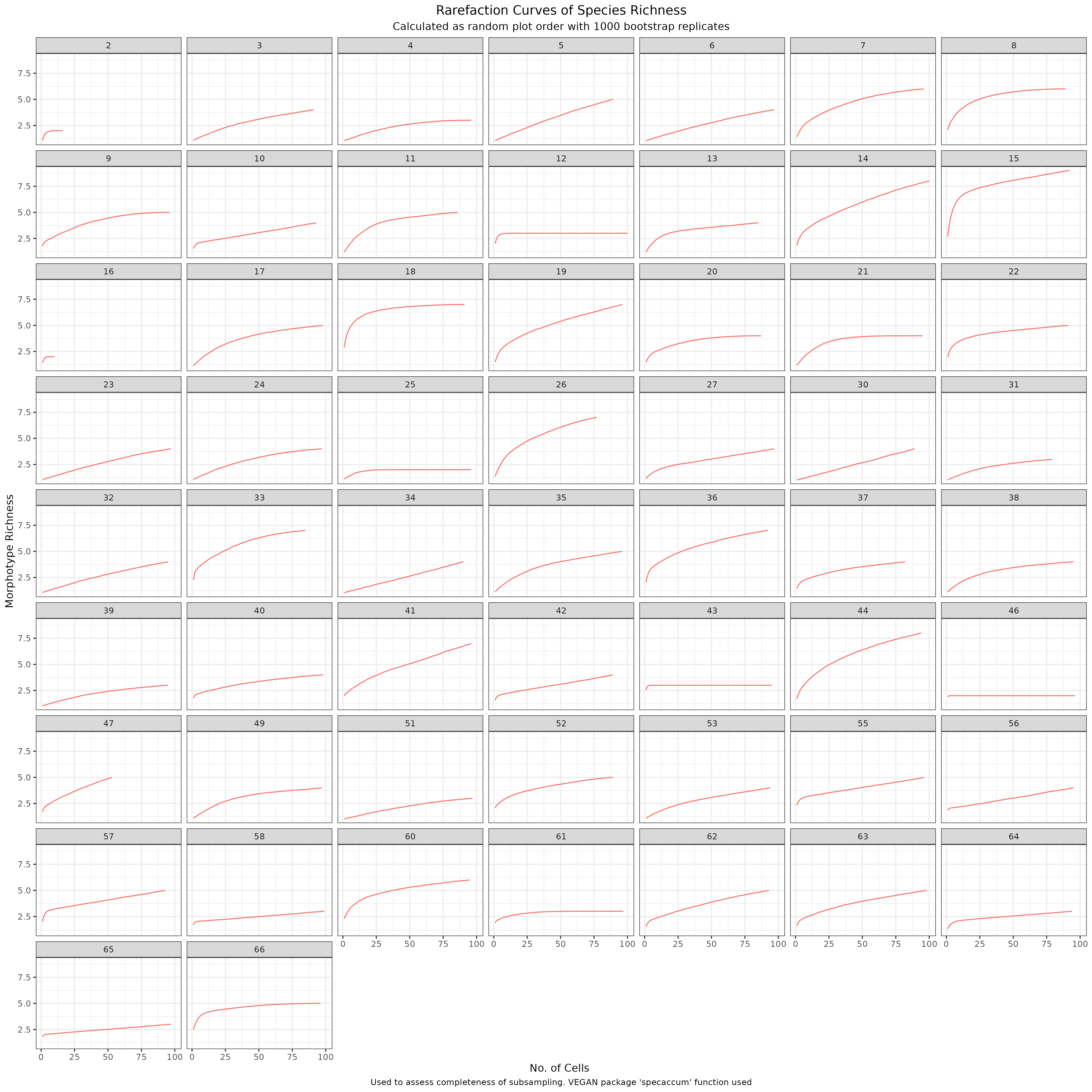
Appendix XX - Reads Per Loci



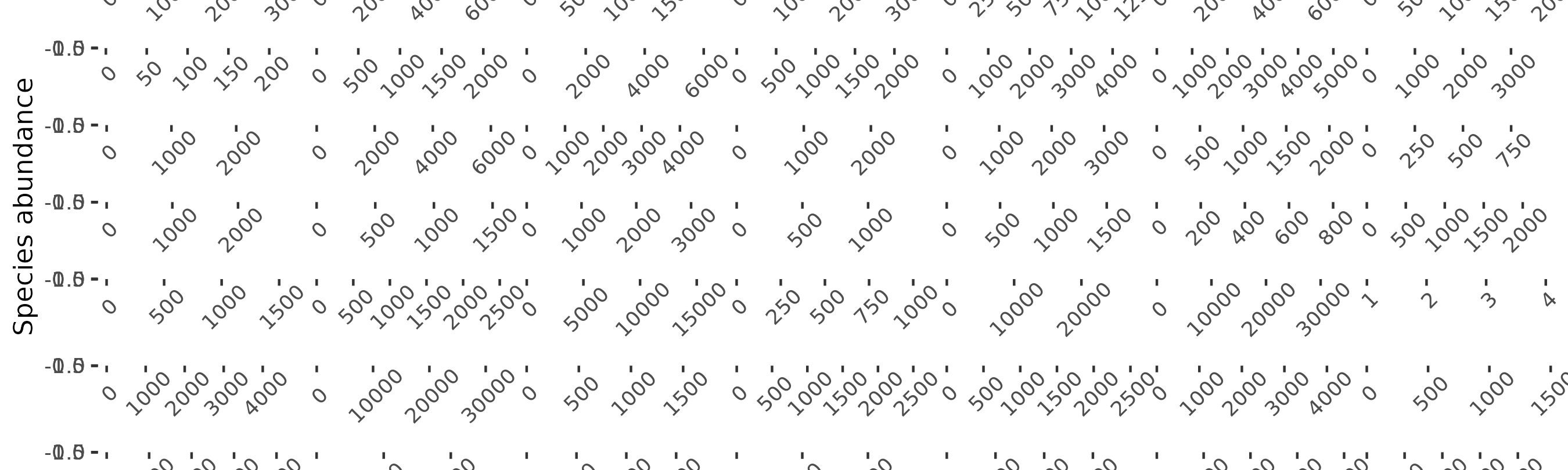
Appendix XX - Species Distribution Models Predictors

| Layer | Description | Source |
| --- | --- | --- |
| 1. | Mean annual cloudiness - MODIS | Wilson et al. 2016 |
| 2. | Cloudiness seasonality 1 - MODIS | Wilson et al. 2016 |
| 3. | Cloudiness seasonality 2 - MODIS | Wilson et al. 2016 |
| 4. | Cloudiness seasonality 3 - MODIS | Wilson et al. 2016 |
| 5. | Beginning of the frost-free period | Wang et al. |
| 6. | Climatic moisture deficit | Wang et al. |
| 7. | Degree-days above 5C from | Wang et al. |
| 8. | Mean annual precipitation | Wang et al. |
| 9. | Mean annual precipitation as snow | Wang et al. |
| 10. | Temperature seasonality | Wang et al. |
| 11. | 2015 Percent Grass/Herbaceous cover - MODIS | (MOD44B) |
| 12. | 2015 Percent Tree cover from Landsat 7/8 | (GLCF) |
| 13. | Soil probability of bedrock (R Horizon) | SoilGrids |
| 14. | Soil organic carbon (Tonnes / ha) | SoilGrids |
| 15. | Surface soil pH in H2O | SoilGrids |
| 16. | Surface soil percent sand | SoilGrids |
| 17. | Soil USDA class | SoilGrids |
| 18. | Topographic elevation | EarthEnv DEM |
| 19. | Topographic elevation, moving window. | EarthEnv DEM |
| 20. | Topographic percent slope | EarthEnv DEM |
| 21. | Topographic wetness index | EarthEnv DEM |
| 22. | Topographic aspect from | EarthEnv DEM |
| 23. | Annual potential solar radiation computed | r.sun |
| 24. | Estimated actual (w/-cloud) solar radiation r | .sun / Wilson et al. 2016 |
| 25. | Log-transformed distance to surface water Gl | obal Surface Water Explorer |
| 26. | Percent surface water Gl | obal Surface Water Explorer |

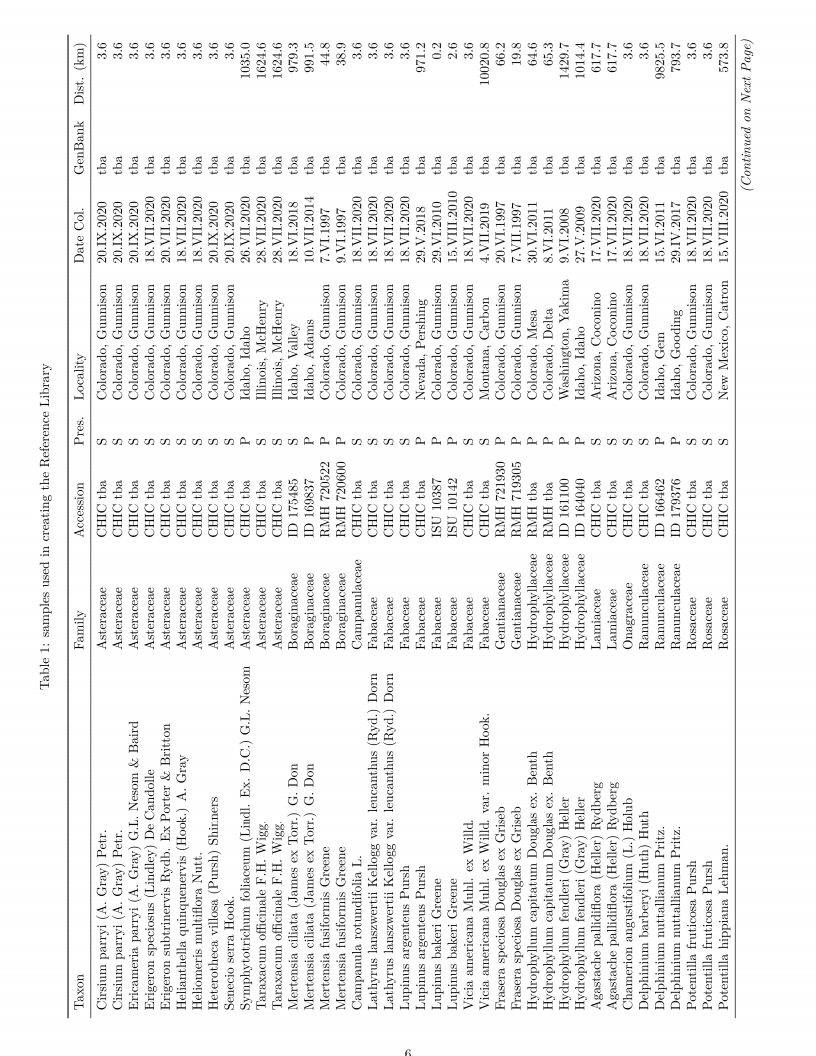
APPENDIX XX - Pollen Morphotype Richness Rarefaction Curves



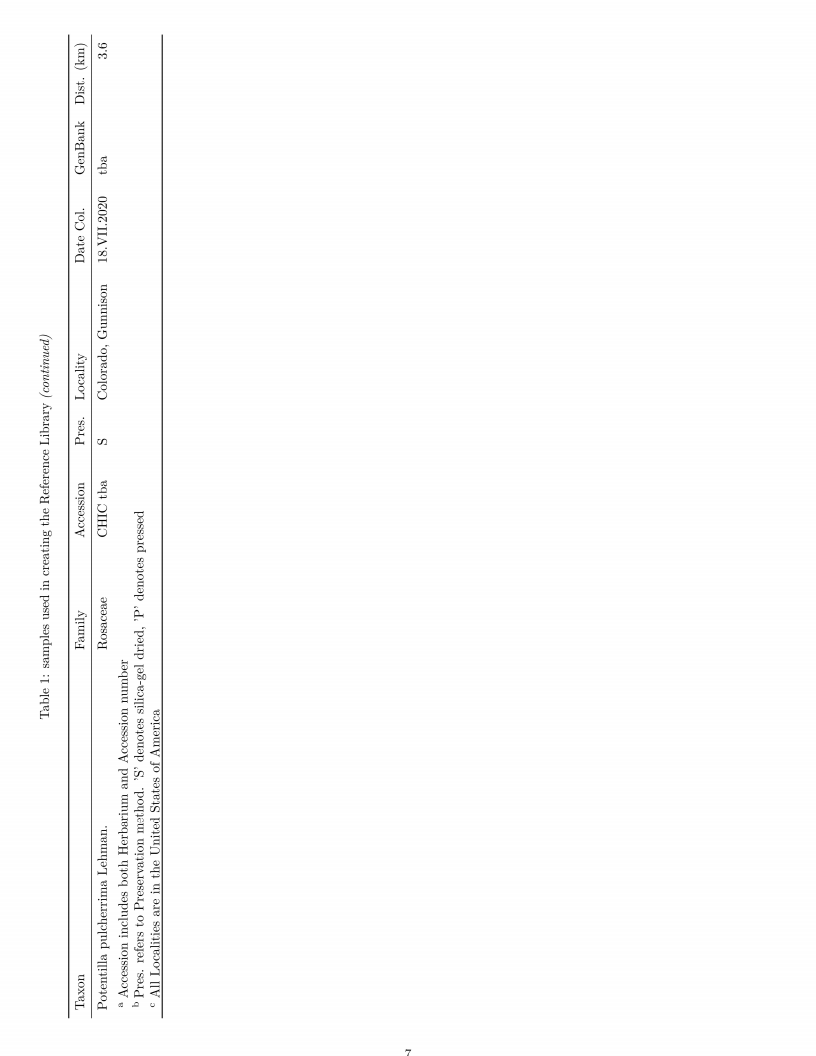
Appendix XX - Pollen Morphotype Abundance Rarefaction Curves



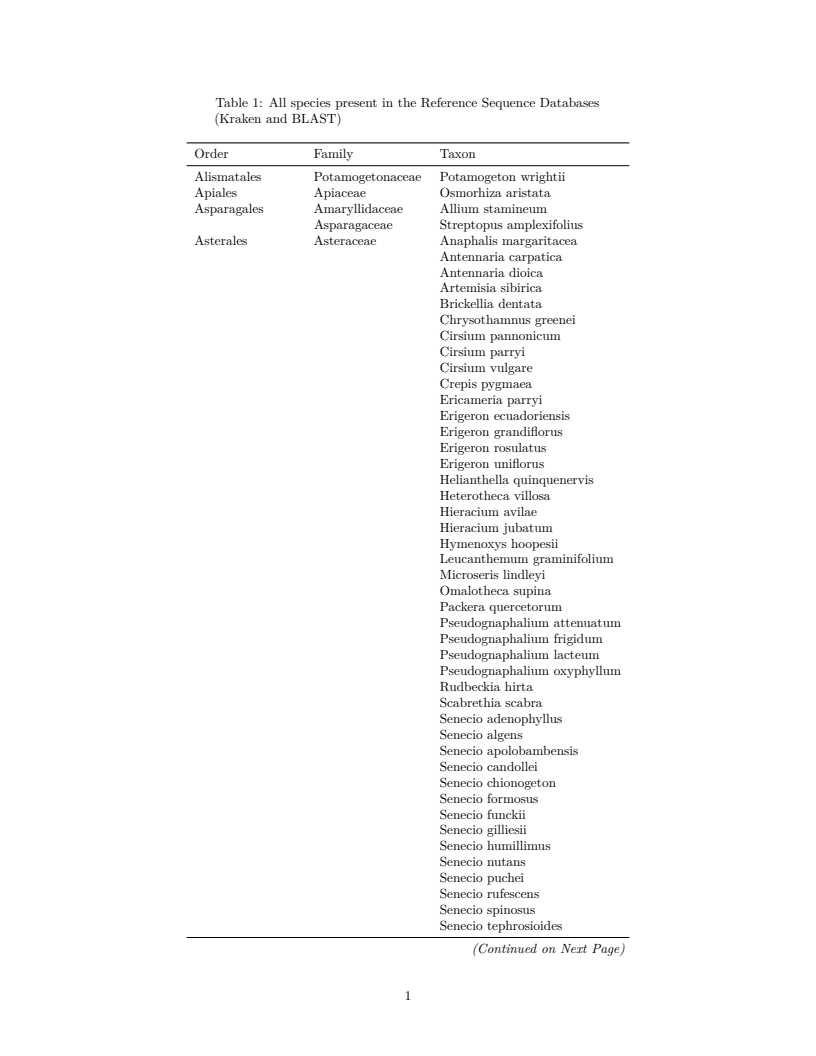
Appendix XX - Molecular Reference Specimen Table



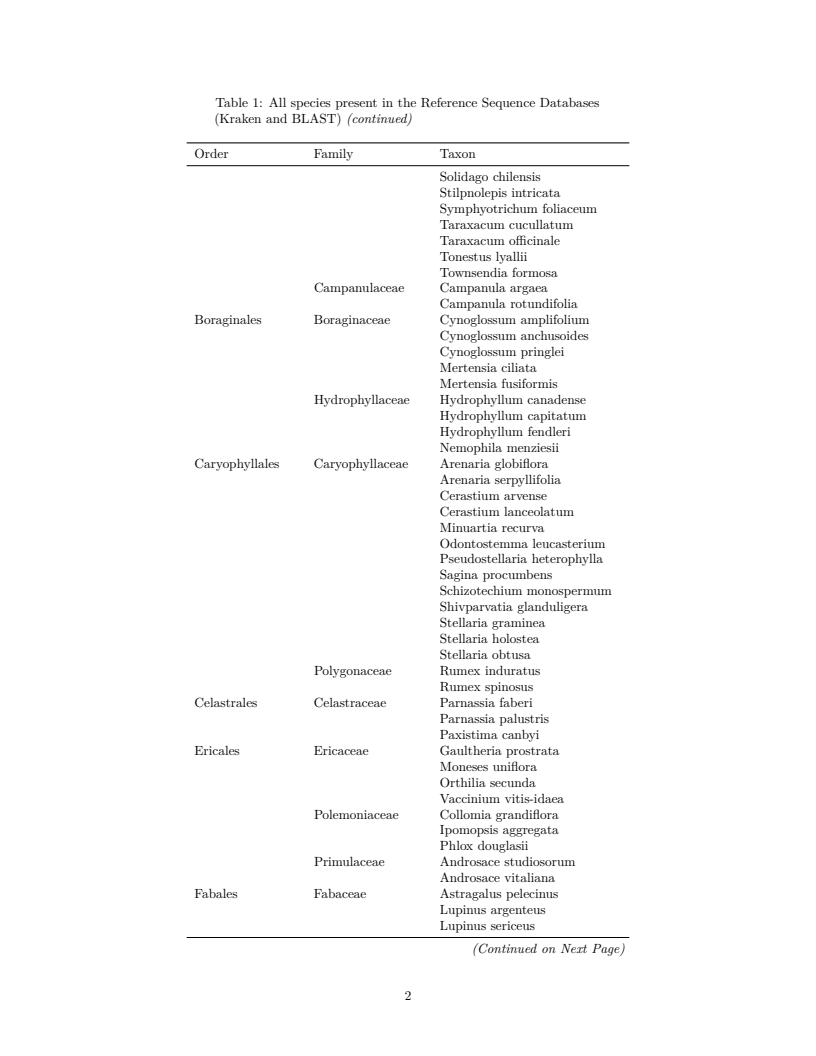
Appendix XX - Molecular Reference Specimen Table (con’t)



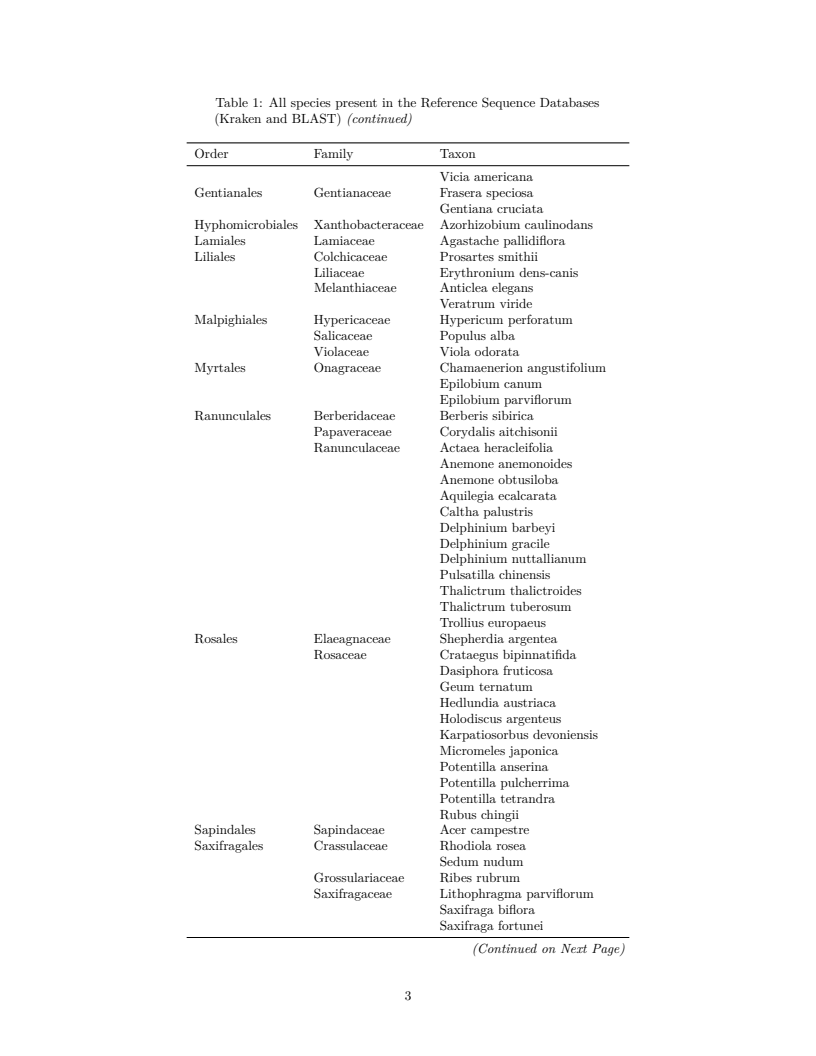
Appendix XX - All Species in the Sequence Databases



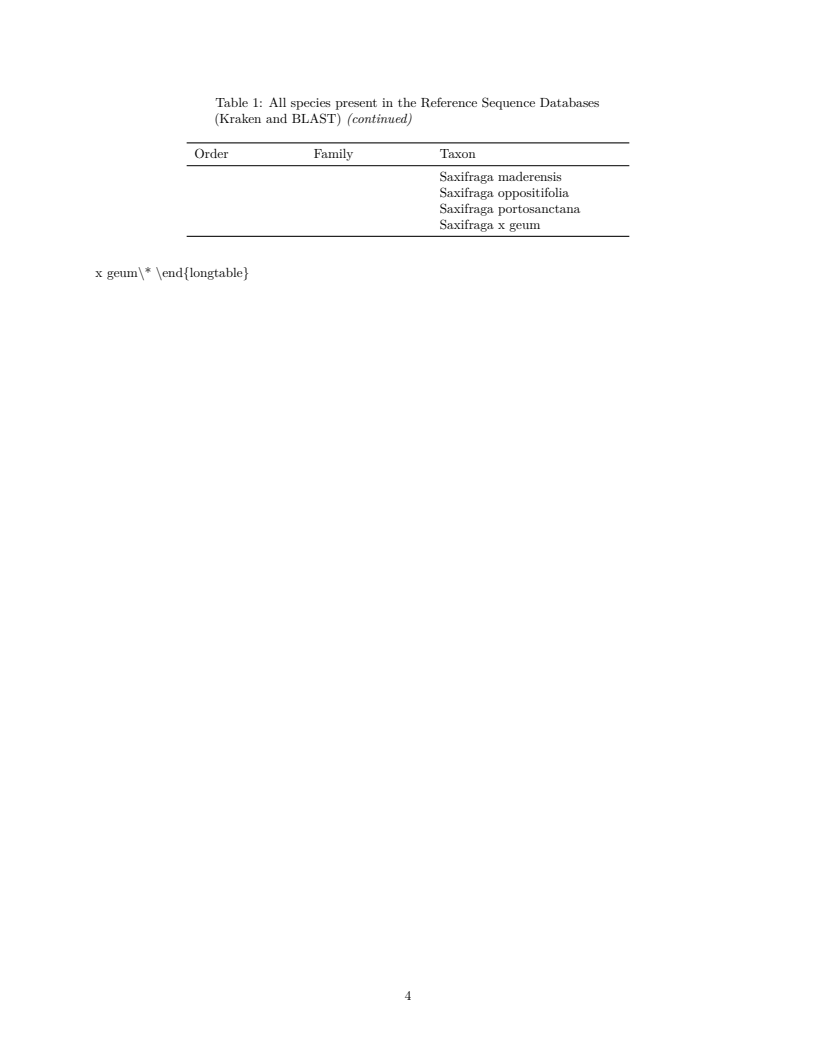
Appendix XX - All Species in the Sequence Databases (con’t)



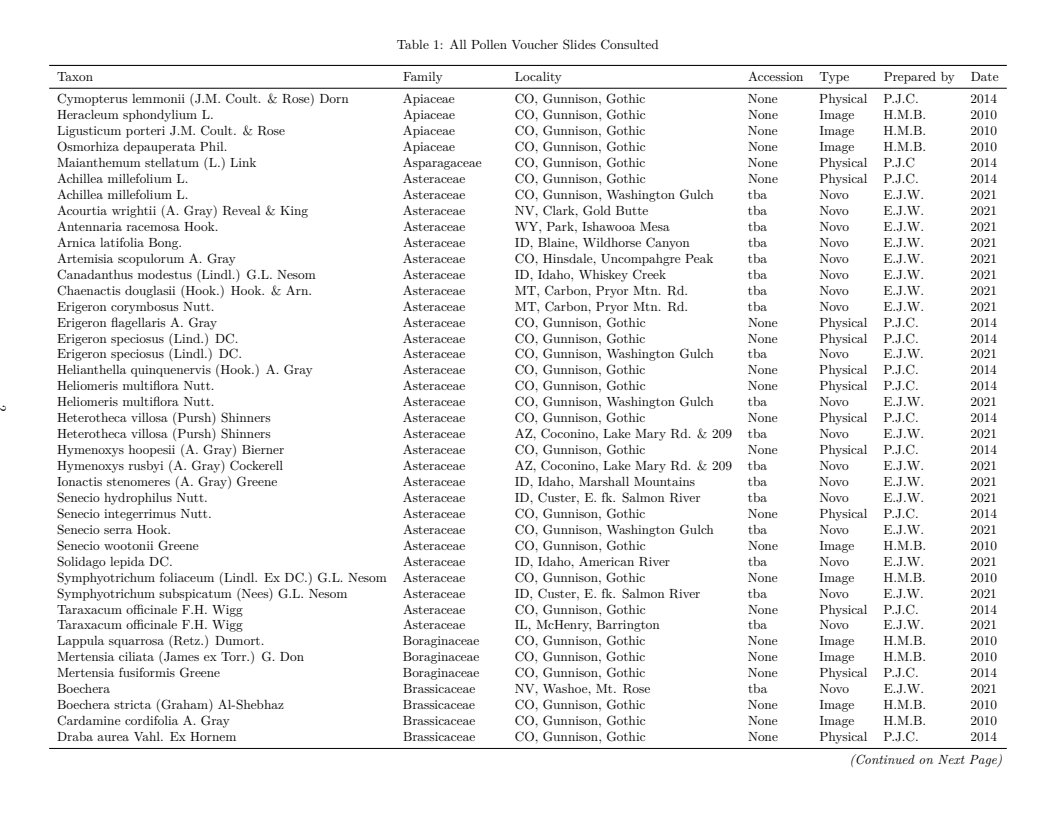
Appendix XX - All Species in the Sequence Databases (con’t)



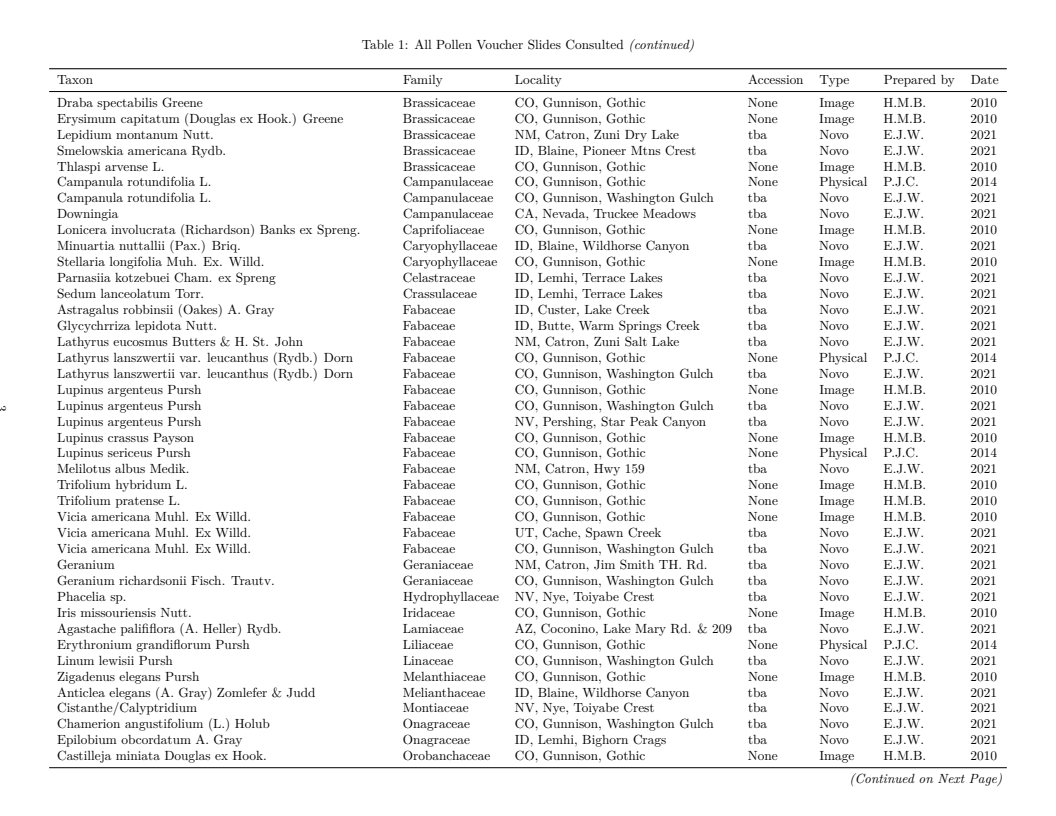
Appendix XX - All Species in the Sequence Databases (con’t)



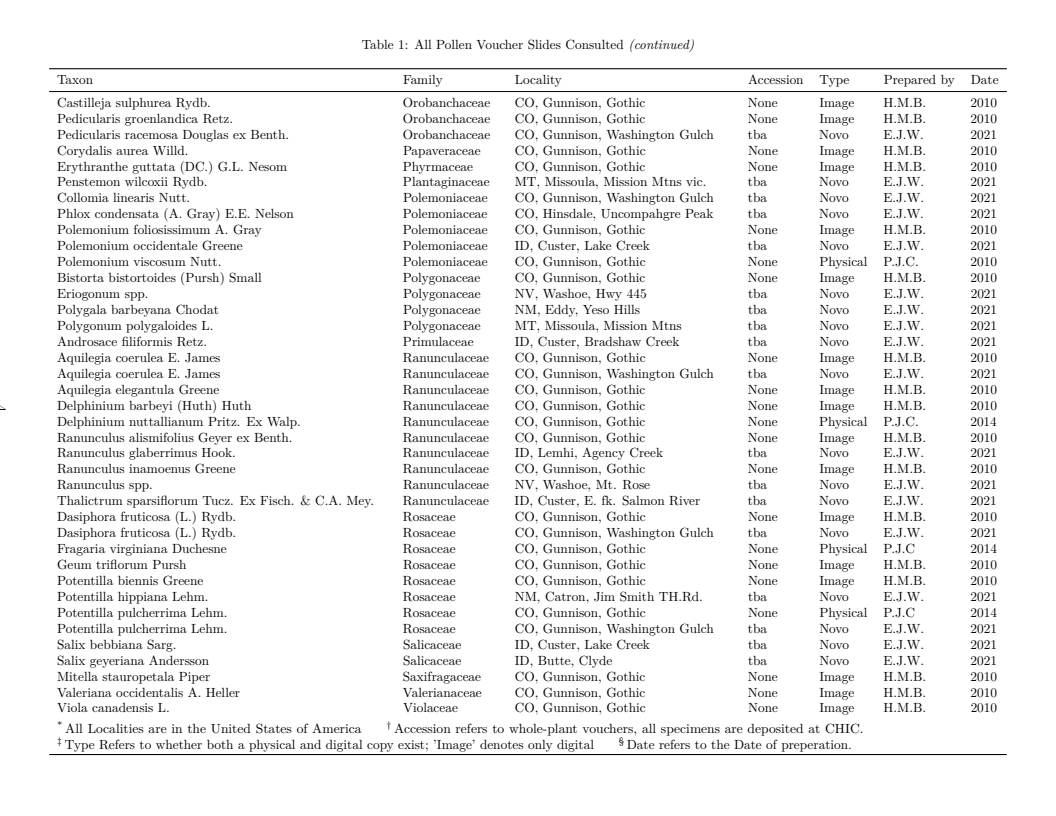
Appendix XX - All Pollen Reference Slides Used to Establish Morphotypes



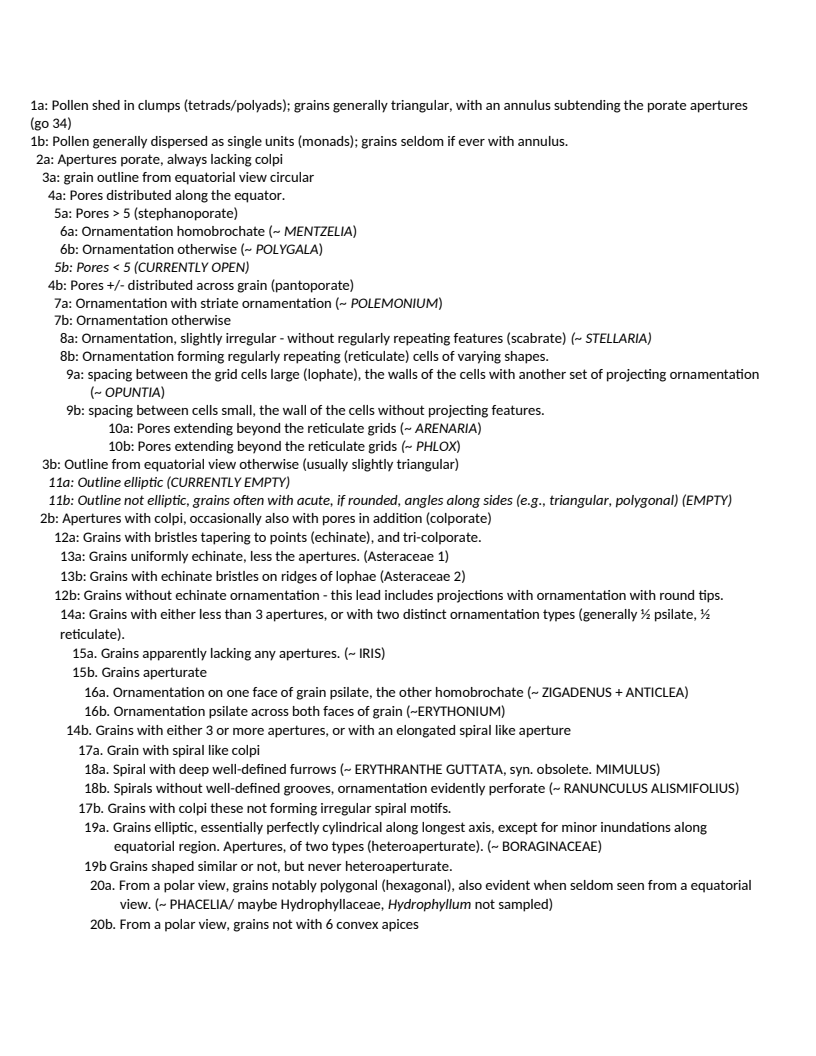
Appendix XX - All Pollen Reference Slides Used to Establish Morphotypes (con’t)

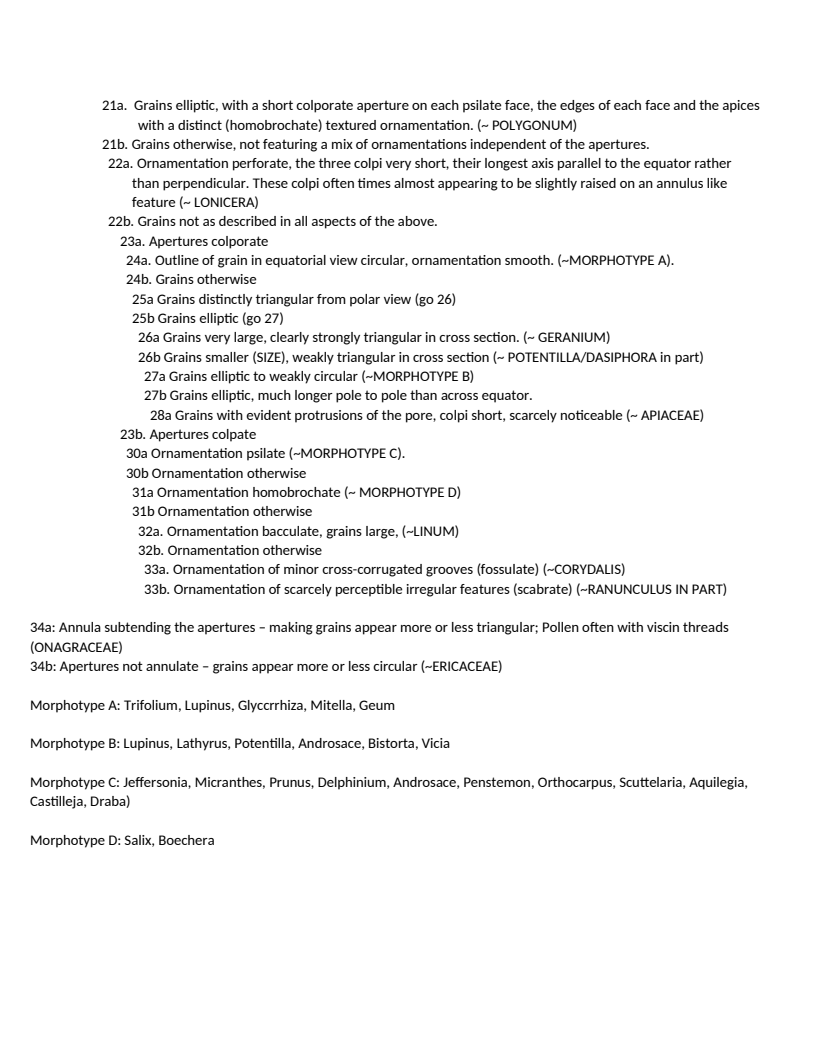


Appendix XX - All Pollen Reference Slides Used to Establish Morphotypes (con’t)



Appendix XX - Pollen Key





Appendix XX - Models used for Species Distribution Model Ensembles

*Generalised Linear Models (GLM)*

*Generalised Additive Models (GAM)*

Ensemble learning utilizes many sets of trees, each composed of many decisions, to create a single model. Each independent variable ( - or *feature*, may) become a node on the tree, a location on the tree where a binary decision will move towards a predicted outcome. Each of the decision tree models which ensemble learning utilizes is a weak models, each of which may suffer due to high variance or bias, but which produce better outcomes than would be expected via random chance. When ensembled these models generate a strong model, a model which should have more appropriately balanced variance and bias and predicts outcomes which are more strongly correlated with the expected values than the individual weak models.

*Random Forest (RF)* the training data are continually bootstrap re-sampled, in combination with random subsets of features, to create nodes which attempt to optimally predict a known outcome. A large number of trees are then aggregated, via the most common predictions, to generate a final classification prediction tree. Each individual prediction tree is generated independently of the others.

*Boosted Regression Tree (BRT)* An initial tree is grown, and all other trees are derived sequentially from it, as each new tree is grown the errors in responses from the last tree are weighed more heavily so that the model focuses on selecting dependent variables which refine predictions. All response data and predictor variables are kept available to all trees.

Alarcón, R. (2010). Congruence between visitation and pollen-transport networks in a california plant–pollinator community. *Oikos*, **119**, 35–44. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0706.2009.17694.x>

Allouche, O., Tsoar, A. & 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.

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. (2021). 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>

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in ecology and evolution*, **3**, 327–338.

Barker, D.A. & Arceo-Gomez, G. (2021). Pollen transport networks reveal highly diverse and temporally stable plant–pollinator interactions in an Appalachian floral community. *AoB PLANTS*, **13**. Retrieved from <https://doi.org/10.1093/aobpla/plab062>

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.

Bergman, P., Molau, U. & Holmgren, B. (1996). Micrometeorological impacts on insect activity and plant reproductive success in an alpine environment, swedish lapland. *Arctic and alpine research*, **28**, 196–202.

Bingham, R.A. & Orthner, A.R. (1998). Efficient pollination of alpine plants. *Nature*, **391**, 238–239.

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.

Cameron, S.A. & Sadd, B.M. (2020). Global trends in bumble bee health. *Annual review of entomology*, **65**, 209–232.

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

CaraDonna, P.J., Petry, W.K., Brennan, R.M., Cunningham, J.L., Bronstein, J.L., Waser, N.M. & Sanders, N.J. (2017). Interaction rewiring and the rapid turnover of plant–pollinator networks. *Ecology letters*, **20**, 385–394.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sande, E.L., Ma, K.H., Colwell, R.K. & Ellison, A.M. (2014). Rarefaction and extrapolation with hill numbers: A framework for sampling and estimation in species diversity studies. *Ecological Monographs*, **84**, 45–67.

Cheng, S., Melkonian, M., Smith, S.A., Brockington, S., Archibald, J.M., Delaux, P.-M., Li, F.-W., Melkonian, B., Mavrodiev, E.V., Sun, W., Fu, Y., Yang, H., Soltis, D.E., Graham, S.W., Soltis, P.S., Liu, X., Xu, X. & Wong, G.K.-S. (2018). 10KP: A phylodiverse genome sequencing plan. *GigaScience*, **7**. Retrieved from <https://doi.org/10.1093/gigascience/giy013>

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.

Colla, S.R., Gadallah, F., Richardson, L., Wagner, D. & Gall, L. (2012). Assessing declines of north american bumble bees (bombus spp.) Using museum specimens. *Biodiversity and Conservation*, **21**, 3585–3595.

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

Elith\*, J., H. Graham\*, C., P. Anderson, R., Dudik, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A. & others. (2006). Novel methods improve prediction of species’ distributions from occurrence data. *Ecography*, **29**, 129–151.

Fazekas, A.J., Kesanakurti, P.R., Burgess, K.S., Percy, D.M., Graham, S.W., Barrett, S.C., Newmaster, S.G., Hajibabaei, M. & Husband, B.C. (2009). Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources*, **9**, 130–139.

Frase, Barbara A. & Buck, P. (2007). Vascular Plants of the Gothic Area. Retrieved from <https://www.digitalrmbl.org/wp-content/uploads/2016/05/vascularplantlist_20071.pdf>

Gage, E. & Cooper, D.J. (2013). Historical range of variation assessment for wetland and riparian ecosystems, u.s. Forest service rocky mountain region

Goulson, D., Lye, G. & Darvill, B. (2008). The decline and conservation of bumblebees. *Annual review of entomology*, **53**, 191–208.

Govaerts, R., Nic Lughadha, E., Black, N., Turner, R. & Paton, A. (2021). The world checklist of vascular plants, a continuously updated resource for exploring global plant diversity. *Scientific Data*, **8**, 1–10.

Group, C.P.W., Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., Bank, M. van der, Chase, M.W., Cowan, R.S., Erickson, D.L. & others. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, **106**, 12794–12797.

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.

Hebert, P.D., Cywinska, A., Ball, S.L. & DeWaard, J.R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **270**, 313–321.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B. & others. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one*, **12**, e0169748.

Hennig, C. (2020). *Fpc: Flexible procedures for clustering*. Retrieved from <https://CRAN.R-project.org/package=fpc>

Hsieh, T.C., Ma, K.H. & Chao, A. (2020). *iNEXT: Interpolation and extrapolation for species diversity*. Retrieved from <http://chao.stat.nthu.edu.tw/wordpress/software_download/>

Iler, A.M., Humphrey, P.T., Ogilvie, J.E. & CaraDonna, P.J. (2021). Conceptual and practical issues limit the utility of statistical estimators of phenological events. *Ecosphere*, **12**, e03828.

Janzen, D.H. (1967). Synchronization of sexual reproduction of trees within the dry season in central america. *Evolution*, **21**, 620–637.

Janzen, D.H., Burns, J.M., Cong, Q., Hallwachs, W., Dapkey, T., Manjunath, R., Hajibabaei, M., Hebert, P.D. & Grishin, N.V. (2017). Nuclear genomes distinguish cryptic species suggested by their DNA barcodes and ecology. *Proceedings of the National Academy of Sciences*, **114**, 8313–8318.

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.

Kress, W.J. & Erickson, D.L. (2007). A two-locus global DNA barcode for land plants: The coding rbcL gene complements the non-coding trnH-psbA spacer region. *PLoS one*, **2**, e508.

Kuhn, M. (2022). *Caret: Classification and regression training*. Retrieved from <https://CRAN.R-project.org/package=caret>

Lewin, H.A., Richards, S., Aiden, E.L., Allende, M.L., Archibald, J.M., Bálint, M., Barker, K.B., Baumgartner, B., Belov, K., Bertorelle, G., Blaxter, M.L., Cai, J., Caperello, N.D., Carlson, K., Castilla-Rubio, J.C., Chaw, S.-M., Chen, L., Childers, A.K., Coddington, J.A., Conde, D.A., Corominas, M., Crandall, K.A., Crawford, A.J., DiPalma, F., Durbin, R., Ebenezer, T.E., Edwards, S.V., Fedrigo, O., Flicek, P., Formenti, G., Gibbs, R.A., Gilbert, M.T.P., Goldstein, M.M., Graves, J.M., Greely, H.T., Grigoriev, I.V., Hackett, K.J., Hall, N., Haussler, D., Helgen, K.M., Hogg, C.J., Isobe, S., Jakobsen, K.S., Janke, A., Jarvis, E.D., Johnson, W.E., Jones, S.J.M., Karlsson, E.K., Kersey, P.J., Kim, J.-H., Kress, W.J., Kuraku, S., Lawniczak, M.K.N., Leebens-Mack, J.H., Li, X., Lindblad-Toh, K., Liu, X., Lopez, J.V., Marques-Bonet, T., Mazard, S., Mazet, J.A.K., Mazzoni, C.J., Myers, E.W., O’Neill, R.J., Paez, S., Park, H., Robinson, G.E., Roquet, C., Ryder, O.A., Sabir, J.S.M., Shaffer, H.B., Shank, T.M., Sherkow, J.S., Soltis, P.S., Tang, B., Tedersoo, L., Uliano-Silva, M., Wang, K., Wei, X., Wetzer, R., Wilson, J.L., Xu, X., Yang, H., Yoder, A.D. & Zhang, G. (2022). The earth BioGenome project 2020: Starting the clock. *Proceedings of the National Academy of Sciences*, **119**, e2115635118. Retrieved from <https://www.pnas.org/doi/abs/10.1073/pnas.2115635118>

Life Project Consortium, D.T. of, Blaxter, M., Mieszkowska, N., Palma, F.D., Holland, P., Durbin, R., Richards, T., Berriman, M., Kersey, P., Hollingsworth, P., Wilson, W., Twyford, A., Gaya, E., Lawniczak, M., Lewis, O., Broad, G., Howe, K., Hart, M., Flicek, P. & Barnes, I. (2022). Sequence locally, think globally: The darwin tree of life project. *Proceedings of the National Academy of Sciences*, **119**, e2115642118. Retrieved from <https://www.pnas.org/doi/abs/10.1073/pnas.2115642118>

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.

Li, X., Yang, Y., Henry, R.J., Rossetto, M., Wang, Y. & Chen, S. (2015). Plant DNA barcoding: From gene to genome. *Biological Reviews*, **90**, 157–166.

Lu, J., Breitwieser, F.P., Thielen, P. & Salzberg, S.L. (2017). Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science*, **3**, e104.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. (2022). *Cluster: Cluster analysis basics and extensions*. Retrieved from <https://CRAN.R-project.org/package=cluster>

Maitner, B. (2022). *BIEN: Tools for accessing the botanical information and ecology network database*. Retrieved from <https://CRAN.R-project.org/package=BIEN>

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**.

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.

Newstrom, L.E., Frankie, G.W. & Baker, H.G. (1994). A new classification for plant phenology based on flowering patterns in lowland tropical rain forest trees at la selva, costa rica. *Biotropica*, **26**, 141–159.

Occdownload Gbif.Org. (2021). Occurrence download. Retrieved from <https://www.gbif.org/occurrence/download/0206948-200613084148143>

Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O’Hara, R.B., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., Evangelista, H.B.A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M.O., Lahti, L., McGlinn, D., Ouellette, M.-H., Ribeiro Cunha, E., Smith, T., Stier, A., Ter Braak, C.J.F. & Weedon, J. (2022). *Vegan: Community ecology package*. Retrieved from <https://CRAN.R-project.org/package=vegan>

Oliver, P.M., Adams, M., Lee, M.S., Hutchinson, M.N. & Doughty, P. (2009). Cryptic diversity in vertebrates: Molecular data double estimates of species diversity in a radiation of australian lizards (diplodactylus, gekkota). *Proceedings of the Royal Society B: Biological Sciences*, **276**, 2001–2007.

Omernik, J.M. (1987). Ecoregions of the conterminous united states. *Annals of the Association of American geographers*, **77**, 118–125.

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.

Prim, R.C. (1957). Shortest connection networks and some generalisations. *Bell System Technical Journal*, **36**, 1389–1401.

Qiao, H., Soberon, J. & Peterson, A.T. (2015). No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, **6**, 1126–1136.

Robinson, N., Regetz, J. & Guralnick, R.P. (2014). EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data. *ISPRS Journal of Photogrammetry and Remote Sensing*, **87**, 57–67.

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.

Sarro, E., Tripodi, A. & Woodard, S.H. (2022). Bumble bee (bombus vosnesenskii) queen nest searching occurs independent of ovary developmental status. *Integrative Organismal Biology*, **4**, obac007.

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

Tran, H., Nguyen, P., Ombadi, M., Hsu, K., Sorooshian, S. & Qing, X. (2019). A cloud-free MODIS snow cover dataset for the contiguous united states from 2000 to 2017. *Scientific data*, **6**, 1–13.

Wang, T., Hamann, A., Spittlehouse, D. & Carroll, C. (2016). Locally downscaled and spatially customizable climate data for historical and future periods for north america. *PloS one*, **11**, e0156720.

Wenzell, K.E., McDonnell, A.J., Wickett, N.J., Fant, J.B. & Skogen, K.A. (2021). Incomplete reproductive isolation and low genetic differentiation despite floral divergence across varying geographic scales in castilleja. *American Journal of Botany*, **108**, 1270–1288.

Williams, P.H. (1982). The distribution and decline of british bumble bees (bombus latr.). *Journal of Apicultural Research*, **21**, 236–245. Retrieved from <https://doi.org/10.1080/00218839.1982.11100549>

Wilson, A.M. & Jetz, W. (2016). Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. *PLoS biology*, **14**, e1002415.

Wood, D.E., Lu, J. & Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome biology*, **20**, 1–13.

Zhao, Y.-H., Lázaro, A., Ren, Z.-X., Zhou, W., Li, H.-D., Tao, Z.-B., Xu, K., Wu, Z.-K., Wolfe, L.M., Li, D.-Z. & Wang, H. (2019). The topological differences between visitation and pollen transport networks: A comparison in species rich communities of the himalaya–hengduan mountains. *Oikos*, **128**, 551–562. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1111/oik.05262>