Plant Metagenomic Barcoding of Pollen Loads Offers Insights on the Foraging Patterns of Queen Bumble Bees in the Southern Rocky Mountains, U.S.A.

Reed Clark Benkendorf*[1,2], Jane E. Ogilvie [3], Emily J. Woodworth [1,2], Sophie Taddeo [1,2], Paul J. Caral

Abstract

an abstract will be written to fill this space.

INTRODUCTION

The inability to reliably identify plants to the level of species often leaves our understanding of ecosystem function and interactions nebulous. Current methods to remediate this situation include: ignoring these ecologically relevant levels of detail, revisiting plots for diagnostic material, 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 complex genera which serve as bioindicators, and preferred partners in ecological interactions, as well as an increasing lack of taxonomic experts. Many genera have species which are well defined based upon ecological 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, as well as in academic research (Gage & Cooper (2013), AIM). This endeavour is often mired by lack of diagnostic characters (e.g. flowers, fruits, or roots), and increasingly the description of cryptic species (Janzen *et al.* (2017), Oliver *et al.* (2009)). Solutions to this problem are wanting, certain programmes have relied increasingly upon revisiting field sites to identify

*Correspondence: reedbenkendorf2021@u.northwestern.edu 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)). Recently barcoding, and metabarcoding, have shown much promise in all Kingdoms of life. For example . . . . . With plants the identification of members of certain clades has been quite successful, whereas with others results have been elusive (Liu *et al.* (2014)), most studies seem to be in between this spectrum (CITE). Particular challenges with the utilization of high-copy number sequences are associated with their rates of divergence. . . . and. . . 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)), and custom species sequence databases derived via species distribution modelling.

Our foundation for increasing the quality of metabarcoding results in plants is reducing the number of possible plant species candidates by generating user selected sequence databases. While there are numerous possible approaches for this process, we achieve the selection of possible plant candidate species using digital collections gleaned from Herbaria, (typically Governmental) survey work, and citizen science from a domain exceeding the study area. To these candidate species an approach, such as logistic regression, may be used to identify distances under which taxa are worth further exploring. To these candidate species, we generate species distribution models, which indicate the probability of suitable habitat in a domain, and base the inclusion of these taxa, or representative congeners, upon these results. 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. Moreover, as most next-generation sequence data is deposited as raw-sequence reads, from a processing perspective, it is essential to reduce the candidate species via an approach as such. Concomitant to next-generation sequencing has been high-throughput sequencing (cite). Currently the largest plant systematic endeavor ever undertaken, the Kew Plant and Fungal Tree of Life (PAFTOL), is approaching completion (Baker *et al.* (2021)). The dataset it creates will contain Hyb-Seq data from at least one species representing each genus in the Plant Kingdom using the popular A353 probes (). These widely available data serve to provide a taxonomically comprehensive backbone for plant metabarcoding. *A brief consideration and example on the inclusion of the temporal dimension of plant flowering, which we suspect will be of importance in sub-tropical, and tropical ecosystems - and others with high seasonality in flowering, is also discussed.* We show that by combining both candidate plant species, and representative sequence data, with only a small amount of new reference sequencing, one is capable of generating ecologically realistic, and biologically relevant results. To test these assertions we utilize a long-term observational study of

Bumble Bees, the most important genus of pollinators in the temperate, and one with consistently concerning conservation assessments (Cameron & Sadd (2020), Goulson *et al.* (2008)). By generating sequence data from corbiculae, pollen baskets, we seek to determine whether this approach is viable, and if the foraging record of pollen matches the record derived from observations. **bee stuff here**

METHODS

Study System

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.), Gunnison County, Colorado, USA (see... *APPENDIX 1* for site information). Pollinator observations of *Bombus* Latreille spp. (Apidae Latreille) were conducted from June - August of 2015. The six study sites are in areas characterised by high-montane/subalpine Parkland vegetation communities.

Spatial Analyses

To develop an ecologically relevant list of vascular plant species, with expected biotic pollination, which may be present at the study sites all records adjacent to the field site were downloaded from BIEN (Maitner (2022)), and these taxa had Species Distribution Models generated to infer their suitability. This list of species served as a reference for which 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 field site and all plant species records were downloaded. In order to estimate the stochasticity of collections, 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 to include species for distribution modelling.

Species had all records from BIEN within a 50km border of the Omernik level 3 ecoregion which the site is located in (No. 21' Southern Rockies'), downloaded (n = 23,919), (Omernik (1987)). These records were split into two, initially identical, sets, one for generating machine learning models, and the other for Generalised Linear (GLM) and Generalized Additive Models (GAM). The set for generating GLM and GAM records was thinned to reduce spatial autocorrelation in the dataset, as measured by Morans Index (Moran (1950), Bivand & Wong (2018)). To both datasets an additional 4029 plots collected from a random stratification of 19% of the land cover in the area of analysis were searched to create true absences (BLM CITATION-need appropriate format for journal). To achieve a larger absence dataset 1000 pseudo-absence records were generated for each taxon, each of which was greater than 10km from an occurrence record. For ML models, these pseudo-absences were reduced so that the number of presence to absence records were 1:1; to achieve this, absence records inside of 10% of the mean sample value, of the presence records, for any predictor were removed.

Species abiotic niche predictors were 26 variables at 30m resolution, six related to climate, five soil, four topographic, four related to cloud cover, with the remaining … (Wilson & Jetz (2016), Wang *et al.* (2016), Hengl *et al.* (2017), Robinson *et al.* (2014)) (*APPENDIX 6*). For linear regression models these predictors underwent both vifstep (theta = 10, max observations = 12,500) and vifcor (theta = 0.7, max observations = 12,500), and collinear features were removed leaving 16 variables (Naimi *et al.* (2014)).

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)). 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)). The results of these models were extracted to a polygon feature derived from a minimum-spanning tree which encompasses the study area, and species from either ensemble with greater than 50% habitat suitability were considered present for further purposes (Prim (1957)). To evaluate these results GBIF (n = 61789) records which had not been assimilated to the BIEN database yet were evaluated using logistic regression (GBIF citation).

535 species were modelled using Generalized Linear Models and Generalized Additive Models. 534 species were modelled using Random Forest and Boosted Regression Trees. *Of the species modelled % were accurately classified as being present at the field station, while % were accurately classified as being absent from the field station* TABLE XX* for more results.*

To evaluate the accuracy of the Species distribution models, additional presence records from GBIF (n = 61789), and AIM (n = 12730) were used as test and training sets (n = 74,519) (CITE AIM AND GBIF). 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 (median = 110.0, $\bar{x}$ = 223.1, max = 1568 record pairs used) with a 70% test split (Kuhn (2022)).

Molecular Lab Work

Reference Plant Library Generation

Using 5 years of observational data on *Bombus* Queen Bee foraging at these studies sites (Ogilvie unpublished), 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*).

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

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.

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).

Computational Processes and Analyses.

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)). Contigs were generated using HybPiper using target files created by M353 (Johnson *et al.* (2016), McLay *et al.* (2021)).

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)). 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)). 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.

Identification of Sequence Matching Loci

A local NCBI database was built using the same processed novel and downloaded sequences (Camacho *et al.* (2009)).

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)) slides was assembled from slides previously prepared by the authors (n = 21), and other researchers (n = 38) (Brosi & Briggs (2013)). Using 5 years of observational data on *Bombus* Queen Bee foraging at these studies sites (Ogilvie unpublished), as well as the Vascular Plant Checklist (FRASER BUCK 2007), an additional 62 voucher slides 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).

In order to determine which plant taxa were distinguishable via light microscopy, and to develop a dichotomous key to pollen morphotypes, Divisive Hierarchical Clustering techniques were used. 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)). Using the cluster dendrogram, elbow plot, and heatmaps (Hennig (2020)), 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 0.5mm2 of pollen was placed onto a ~4mm2 fuchsin jelly cube (Beattie (1971)) atop a graticulated microscope slide, with 20 transects and 20 rows (400 quadrats) (EMS, Hartfield, PA). The jelly was melted until pollen grains were homogeneously spread across the microscope slide. Slides were sealed with Canada Balsam (Rublev Colours, Willits, CA); 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)), and non-parametric Species Diversity rarefaction curves were used to assess the completeness of sampling (Chao *et al.* (2014), Hsieh *et al.* (2020)). 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.*

Temporal Analyses

To estimate the duration of dates in which plant species were flowering Weibull estimates of all spatially modelled taxa were developed (Belitz *et al.* (2020), Pearse *et al.* (2017)). Only BIEN records which occurred in the Omernik Level 4 Ecoregions within 15km of the study area (n = 5, 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, the SnowUS dataset (Iler *et al.* (2021), Tran *et al.* (2019)) 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 estimated for the date when 10% of individuals had begun flowering, when 50% were flowering, and when 90% of individuals had flowered.

Results

Microscopic Pollen identification

Using the fuchsin jelly preparation and light microscopic analyses of grains and scoring of 12 character states resulted in the establishment of XX morphotypes which grains could be reliably classified into. ***APPENDIX 7*** . XX Samples were counted and based on rarefaction had over % of expected morphotypes found. The relative abundance of pollen grains in each sample (max % of any species, mean % of all species, min % trace amounts detected).

Metabarcoding Pollen identification

Kraken2 was able to identify the species richness of pollen samples ($\bar{x}$ = , min = , max =). Bracken was able to estimate the relative abundance of pollen grains in each sample (max % of any species, $\bar{x}$ % of all species, min % trace amounts detected).

Spatial Analyses

[Table 1 about here.]

[Table 2 about here.]

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, we declined to perform this computational task.

overall quality of SDMS across entire range of study

Taken together all models throughout the entire domain of modelling 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.

The quality of the models themselves in predicting a large component of the alpha richness of a region

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), a P VALUE IS NOT REPORTED AS THE VALUES WERE MANUALLY PARSED INTO CLASSES BASED ON SUITABILITY PER UNIT AREA. Of the Of the 554 vascular plants with biotic pollination syndromes, the 475 ML 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), a P VALUE IS NOT REPORTED AS THE VALUES WERE MANUALLY PARSED INTO CLASSES BASED ON SUITABILITY PER UNIT AREA. 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.

The quality of the models in predicting ecologically dominant plant taxa in a pollination network

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 are restricted from the initial query database,

Temporal Analyses

The first date of modelled snow melt in the Gothic area (n = 17, mean = 137.9, median = 135, 3rd quantile = 151), and the first date of a consistent winter snow base (n = 17, mean = 299.9, median = 300, 1st quantile = 291)... 332 species with more than 10 records in the focal level 4 ecoregions ($\bar{x}$ = 35.01657, median = 35, max = 96) had weibull estimates calculated, an additional 56 species 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 ($\bar{x}$ = 13.86885, median = 13, max = 24).

Only 58 of these 388 species ($\bar{x}$ n = 34.56897, median n = 31) were able to be compared to ground truth data.

[Figure 1 about here.]

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 floristic inventories at sites, and then utilizing a bootstrap approach akin to ours.

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.

Fewer modelling runs for SDM's likely to be effective for determining inclusion, elastic inclusion criteria.

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. 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. in prep., 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.

Acknowledgements: Nyree Zerega for assistance obtaining herbaria loans and accessioning our collections at CHIC. We thank the curators at the following herbaria for supplying tissue: Ben Legler at Stillinger (ID), Charles Williams at Ray J. Davis (IDS), 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. Zoe Diaz-Martinez, Angela McDonnell, & Elena Loke for assistance with genomic library preparation. 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.

References

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.

Layer	Description	Source
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)	
15.	Surface soil pH in H2O	
16.	Surface soil percent sand	
17.	Soil USDA class	SoilGrids
18.	Topographic elevation	EarthEnv DEM.
19.	Topographic elevation, moving window.	
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	
25.	Log-transformed distance to surface water	Global Surface Water Explorer
26.	Percent surface water from	Global Surface Water Explorer

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

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.

Bivand, R. & Wong, D.W.S. (2018). Comparing implementations of global and local indicators of spatial association. *TEST*, **27**, 716–748.

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.

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.

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

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.

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., 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.

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

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

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

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., Baker, W.J. & others. (2021). New targets acquired: Improving locus recovery from the Angiosperms353 probe set. *Applications in plant sciences*, **9**.

Moran, P.A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

Naimi, B. & Araujo, M.B. (2016). Sdm: A reproducible and extensible r platform for species distribution modelling. *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. *Ecography*, **37**, 191–203.

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.

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

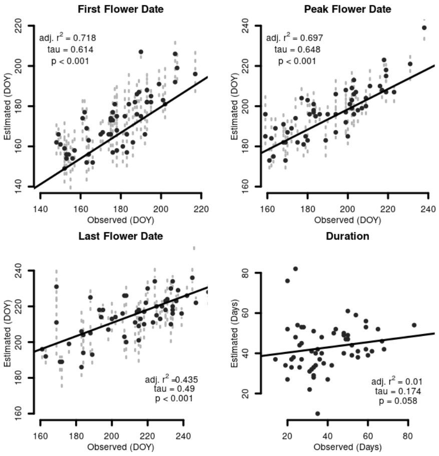
Tran, H., Nguyen, P., Ombadi, M., Hsu, K., Sorooshian, S. & Qing, X. (2019). A cloud-free MODIS snow cover dataset for the contiguous united states from 2000 to 2017. *Scientific data*, **6**, 1–13.

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.

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.

List of Figures



First Flower Date

Peak Flower Date

Last Flower Date

Duration

Figure 1: A caption

List of Tables

Table 2: Logistic regression assessing accuracy of SDMs

Metric	Value	Metric	Value
Accuracy (Training)	83.75	F-Score	0.84
Accuracy (Test)	84.00	AUC	0.92
Recall	81.03	Concordance	0.92
True Neg. Rate	86.97	Discordance	0.08
Precision	88.04	Tied	0.00

Table 3: SDM evaluation contingency table

	Training		Testing	
	Absence	Presence	Absence	Presence
Absence	25620	3838	11130	1653
Presence	6614	28248	2758	12024