Supplemental Methods

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

System background

To test the effectiveness of our methodological approach we applied it to identify the plant species found in the corbiculae (pollen loads) of queen bumble bees (*Bombus* Latreille) collected from the Rocky Mountain Biological Laboratory in Colorado, USA (RMBL; 38°57.5" N, 106°59.3" W (WGS 84), 2900 m.a.s.l.). We collected pollen loads from wild foraging queen bees between May and July of 2015 at six permanent study sites (Ogilvie & CaraDonna 2022). To harvest the pollen loads, we captured queens in an insect net, transferred them into a restraining device (Kearns et al. (2001)), collected a pollen load from one leg, and then released them. We collected 64 corbiculae pollen loads from queens of several common wild bumble bee species: *Bombus appositus*, *B. bifarius*, *B. californicus*, *B. flavifrons*, *B. nevadensis*, and *B. rufocinctus* (Pyke 1982; Ogilvie & CaraDonna 2022). At the six study sites, bumble bee abundance and interactions with flowering plants were monitored for one hour at weekly intervals. The abundance of flowers visited by bumble bees within belt transects spread over the three vegetation types (0.5 x 40 m transects in each vegetation type, 60 m^2 total area per site). We used six years (2015-2020) of observational data on *Bombus* flower visits to identify the plant taxa most frequently visited by queens across all years and to compare with metagenomic data.

Survey database to generate a regional taxa list

We first generated a short list of potential candidate species. We downloaded, from the Botanical Information and Ecology Network 'BIEN' (Maitner (2022)), 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. To reduce the number of species to include in the reference sequence databases, we then generated Species Distribution Models (SDMs) for these taxa to predict their distribution throughout the study area.

To minimize 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. To account for the stochasticity of botanical collecting and offset the number of records associated with the research station, this data set was bootstrap re-sampled 250 times, with 90% of samples selected, to create a testing data set. 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.

Spatial filtering via species distribution modelling

To determine which clades to include in the reference sequence database we used Species Distribution Modelling. We used all occurrence records from BIEN (n = 23,919) within a 50km border of the ecoregion, Omernik level 3, which includes the study area *(No. 21 "Southern Rockies")* to construct the species distribution model (Omernik (1987)). 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)). 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)). 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 different biology and ecology (Elith* *et al.* (2006), Qiao *et al.* (2015)). 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)).

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 (Land Management (2019)). To achieve a larger absence data set, 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)). To achieve this, we removed absence records inside of 10% of the mean sample value of any predictor variable the presence records; the required number of absence records were then randomly sampled.

To predict the potential distribution of each species we used 26 environmental variables at 30m resolution, six related to climate, five soil, four topographic, four related to cloud cover, with the remaining reflecting assorted abiotic parameters (Wilson & Jetz (2016), Wang *et al.* (2016), Hengl *et al.* (2017), Robinson *et al.* (2014)). These publicly available data sets, were selected as they pertain to a wide range of variables interacting with plant physiology. For linear regression models these predictors underwent both *vifstep* (theta = 10, max observations = 12,500) and *vifcor* (theta = 0.7, max observations = 12,500) to detect highly correlated variables, and collinear features were removed leaving 16 variables (Naimi *et al.* (2014)).

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)). TSS was chosen as the ensemble criterion as it has been shown to work across a wide range of species occurrences prevalence (Allouche *et al.* (2006)). 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)).

A total of 535 species were modelled using Generalized Linear Models and Generalized Additive Models and 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 (Occdownload Gbif.Org (2021), Land Management (2019)). Additional novel absence records were generated from the AIM data set to create a data set where each species has balanced presence and absences. Eleven 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, $\bar{x}$ = 223.1, max = 1568 record pairs used) with a 70% test split (Kuhn (2022)).

Temporal Filtering of the species list For assignment of reads to ecologically probabilistic species subsequent to BLAST, flowering time was used as a filter. 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), Pearse *et al.* (2017)). 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 ecoregions if enough records were not 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 data set (Iler *et al.* (2021), Tran *et al.* (2019)) from 2000-2017 were 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, we used the initiation and cessation dates, respectively, as effective start and ends of flowering. These estimates were compared to a long-term observational study of flowering phenology 1974-2012 (CaraDonna *et al.* (2014)), and the floral abundance data from 2015, using Kendall's tau.

Microscopic pollen identification

To qualitatively identify, and quantitatively note, the plant species present in corbiculae loads microscopy was used. A pollen reference library of fuchsin-jelly stained grains which may be present in corbiculae loads of slides was assembled from slides previously prepared by the authors ($n = 21$), and other researchers ($n = 38$) (Beattie (1971), Brosi & Briggs (2013)). Using five years of observational data on *Bombus* Queen Bee foraging at these studies sites (Ogilvie & CaraDonna (2022)), as well as the RMBL Vascular Plant Checklist (Frase & Buck (2007)), 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.

We used clustering techniques to supplement our subjective opinions of which plant taxa were distinguishable via light microscopy, and to develop a dichotomous key to pollen morphotypes. Ten readily discernible categorical traits were collected from each specimen in the image collection. These traits were transformed using Gower distances, and clustered using Divisive Hierarchical 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. 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. *Circa* 0.5mm^2 of pollen was placed onto a ~4mm^2 fuchsin jelly cube (Beattie (1971)) atop a graticulated microscope slide, with 20 transects and 20 rows (400 quadrants) (EMS, Hartfield, PA). The jelly was melted, with stirring, until pollen grains were homogeneously spread across the microscope slide. Slides were sealed with Canada Balsam (Rublev Colours, Willits, CA) followed by sealing with clear nail polish to prevent oxidation; all samples are noted in.

To identify the pollen present in corbiculae loads, light microscopy at 400x (Zeiss Axioscope A1) was used. In initial sampling in three transects, each pollen grain was identified to morphotype and counted; an additional two transects were scanned for morphotypes unique to that slide, if either transect contained a unique morphotype 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.

Metagenomics: additional plant tissue collection and extraction Using five years (2015-2020) of observational data on *Bombus* queen interactions with flowering plants at these studies sites, we identified the plant taxa most frequently visited by queens across all years. In order to capture more variability inherit in the 353 loci we sequenced the 12 most visited taxa twice using samples collected from one site within the Gunnison Basin River Drainage and one individual collected from another more distal population. In addition we included a congener - or a species from a closely related genus to serve as an outgroup for all 12 taxa. We also sequenced another 15 taxa of plants commonly visited by *Bombus* workers, based on the abundances, and immediate access to plant tissue, in the aforementioned data set. Plant collections were identified typically using a combination, of dichotomous keys and primary literature as required (Flora of North America Editorial Committee (1993+), Hitchcock & Cronquist (2018), Ackerfield (2015), Lesica *et al.* (2012), Cronquist *et al.* (1977+), Allred & Ivey (2012), *Jepson flora project* (2020), Mohlenbrock (2002)).

Plant genomic DNA was isolated from ~ 1 cm^2 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 DNA extraction

To extract genomic DNA from pollen a CTAB based protocol was 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 to the sampled 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 a heat block (Multi-Blok, Thermo Fisher Scientific, Waltham Massachusetts) set to 40°C. After 20 minutes incubation, Proteinase K (15 µL of 20mg/ml) and DTT (12.5 µL of 1M in water) were added, and the samples were further incubated at 60°C for 1 hour. Samples were then incubated overnight at 40°C. 500 µL of Phenol-Chloroform-Isoamyl alcohol (25:24:1) were added, vortexed, and centrifuged at 10,000 rpm for 10 minutes and the aqueous phase was pipetted to a 1.5 ml centrifuge tube.

To precipitate the DNA, chilled Isopropyl alcohol & 3 mM Sodium acetate (5:1) equivalent to $\frac{2}{3}$ 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 H_2O. Extractions were assessed using a Nanodrop 2000 (Thermo Fisher Scientific) and Qubit fluorometer (Thermo Fisher Scientific). DNA extracts were then cleaned using 2:1 v./v. Sera-Mag beads (Cytiva, Little Chalfont, UK) to solute ratio following the manufacturer's protocol, eluted in 0.5x TE, and the eluent allowed to reduce by half volume in ambient conditions. DNA was quantified using a Qubit fluorometer.

Library preparation & Bait capture (Barcoding)

Sequence library preparation was performed using the NEBNext Ultra II FS-DNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, Massachusetts, USA) 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 using SPRI beads (Beckman Coulter, Indianapolis, Indiana, USA) and recommended volumes of 80% v./v. ethanol washes. The exception was the herbarium specimens which were not fragmented and only end repaired, with similar library preparation of all samples. 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).

Bioinformatics

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), Tange (2021)). Contigs generated were mapped to a reference with HybPiper with using target files created by M353 (Johnson *et al.* (2016), McLay *et al.* (2021)). A custom Kraken2 database was created by downloading representative species indicated as being present in the study area by the spatial analyses from the Sequence Read Archive (SRA) NCBI (Wood *et al.* (2019)). These sequences were processed in the same manner as our novel sequences. The Kraken2 database was built using default parameters. Kraken2 was run on sequences using default parameters. Following Kraken2, Bracken was used to classify sequences to terminal taxa (Lu *et al.* (2017)). Finally all reads which could be classified by these databases were passed to a local BLAST database. A local NCBI database was built using the same processed novel and downloaded sequences as the previous database (Camacho *et al.* (2009)).

Comparision of Sequence classifications pre and post processing

Using wilcox_effsize, with a one-sided hypothesis of greater, we have strong evidence of ($p < 1e\text{-}04$, wilcoxsign_test) an effect size of 0.732 95% CI [0.57, 0.84, $n = 40$, bootstrap replicates = 1000] (Kassambara (2023), Hothorn *et al.* (2006)).

Ackerfield, J. (2015). *Flora of colorado.* BRIT Press Fort Worth.

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.

Allred, K.W. & Ivey, R. (2012). Flora neomexicana III: An illustrated identification manual. *Lulu. com.*

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

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

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.

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

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

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

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

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.

Cronquist, A., Holmgren, A.H., Holmgren, N.H., Reveal, J.L., Holmgren, P.K., Barneby, R & others. (1977+). *Intermountain flora. Vascular plants of the intermountain west, USA volume six. The monocotyledons.* Columbia University.

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.

Flora of North America Editorial Committee, eds. (1993+). *Flora of north america north of mexico [online].* Oxford University Press on Demand.

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

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

Hitchcock, C.L. & Cronquist, A. (2018). *Flora of the pacific northwest: An illustrated manual.* University of Washington Press.

Hothorn, T., Hornik, K., van de Wiel, M.A. & Zeileis, A. (2006). A lego system for conditional inference. *The American Statistician*, **60**, 257–263.

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.

Jepson flora project. (2020).

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.

Kassambara, A. (2023). *Rstatix: Pipe-friendly framework for basic statistical tests.* Retrieved from https://CRAN.R-project.org/package=rstatix

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

org/package=caret

Land Management, B. of. (2019). U.s. Department of interior bureau of land management, BLM - assessment, inventory, and monitoring (AIM) terrestrial indicators raw dataset. Retrieved from https://gbp-blm-egis.hub.arcgis.com/pages/aim

Lesica, P., Lavin, M. & Stickney, P.F. (2012). *Manual of montana vascular plants*. Brit Press.

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**.

Mohlenbrock, R.H. (2002). *Vascular flora of illinois*. SIU Press.

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.

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

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

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., Caceres, M.D., Durand, S., Evangelista, H.B.A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M.O., Lahti, L., McGlinn, D., Ouellette, M.-H., Cunha, E.R., Smith, T., Stier, A., Braak, C.J.F.T. & Weedon, J. (2022). *Vegan: Community ecology package*. Retrieved from https://CRAN.R-project.org/package=vegan

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, multiscale smoothed, 90m digital elevation model from fused ASTER and SRTM data. *ISPRS Journal of Photogrammetry and Remote Sensing*, **87**, 57–67.

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.

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.