Overview

Many bee species have dietary restrictions that cause them to consume pollen from plants in a single genus or family – just a portion of the floral resources available to them in their environment. It's not immediately obvious why any species would adopt this life history strategy, called pollen specialization. From a mathematical perspective, a pollen-specialist bee adopts more risk than a pollen generalist: when a bee's diet is composed of fewer food species, the food's availability varies more over time as result of statistical averaging (the same concept explains why it is riskier to invest in a few stocks than a diverse portfolio) (Batstone et al., 2018; Schindler et al., 2015). Yet despite these risks, North America is home to at least 1,079 species of pollen-specialist bees (Fowler, 2020a, 2020b; Fowler and Droege, 2020), an estimated quarter of the continent's total number of bee species. Understanding why bees have adopted this dietary strategy is important for their conservation, for a foundational understanding of their biology, and for a more general understanding of why specialization is found across the tree of life.

One hypothesis for why bees maintain specialized diets is that generalizing is prohibitive: it requires the animal to evolve or maintain the ability to use multiple types of resources, which may be difficult due to evolutionary constraints on diet (Hardy et al., 2020). However, because specializing leads to a greater variability in the animal's food supply, specialists might minimize risk by specializing on a single resource that provides a **predictable plethora** (Wcislo and Cane, 1996)—that it is, a resource that is consistently abundant through time. While specializing on a single *randomly selected* resource would be a poor strategy, specializing specifically on a resource that provides a predictable plethora could be a viable long-term option. Despite strong theoretical (Sexton et al., 2017) and anecdotal (González-Varo et al., 2016; Minckley and Roulston, 2006; Wcislo and Cane, 1996) evidence for this hypothesis, it has not been formally assessed with empirical data.

To test the predictable plethora hypothesis, we will investigate whether pollen specialist bee species in the eastern United States specialize on pollen from the region's most abundant angiosperms. We propose using two different approaches. First, we will compare the regional abundance of plant genera hosting specialist bees in our study region with the regional abundance of close relatives that don't host any specialists, but that share similar habitat associations. Second, we will conduct a phylogenetic regression on all angiosperms in the eastern US to test whether a plant taxon's abundance predicts whether it hosts a specialist bee. Both approaches are described in more detail below.

Research questions

1) Are plants that host specialist bees more abundant than close relatives with the same habitat association that do not?
2) Are more abundant angiosperm genera more likely to support pollen-specialist bees?

Methods

Data

All analyses were conducted in R (R Core Team, 2020).

Regional abundance dataset: We used citizen science data from the website iNaturalist (www.inaturalist.org) to measure the regional abundance of angiosperm genera in the eastern United States. Users of iNaturalist upload observations of organisms to the website, often with photo(s), the date, and spatial coordinates. Because the website is quite popular, it provides a large amount of opportunistically collected data; however, like most citizen science data, it has various biases (Dickinson et al., 2010). When possible, we have accounted for these biases analytically (see section 'Measuring regional abundance,' below).

We used the R package *spocc* (Chamberlain, 2020) to pull observations from iNaturalist. We pulled all observations of angiosperm genera with native species in our study region (20 states total: Alabama, Connecticut, Delaware, Florida, Georgia, Louisiana, Maine, Maryland, Maine, Maryland Massachusetts, Mississippi, New Hampshire, New Jersey, New York, North Carolina, Pennsylvania, Rhode Island, South Carolina, Vermont, Virginia, West Virginia). We obtained the list of native angiosperm plants for this region from the USDA PLANTS database (www.plants.usda.gov). We pulled only research-grade observations, which have a photo, date and coordinates, and their identification has been verified by at least two iNaturalist users. When pulling observations, we used the search terms "genus *" (e.g., "Claytonia *"), which returns all observations for a genus, including synonyms.

After pulling the observations from iNaturalist, we updated and filtered the dataset using the following steps. First, we removed duplicate observations of species at the same latitude and longitude. Second, we updated all synonyms to accepted species names according to 'The Plant List' (www.theplantlist.org), using the R package *Taxonstand* (Cayuela et al., 2019). Third, we excluded hybrids, introduced plants, and all genera that included crop species (based on Table 1 of Meyer and Purugganan, 2013). We excluded crop genera because we suspect that crop plants are more widespread in the wild now than they were during most of bees' evolutionary histories, making their abundance more difficult to estimate accurately. We excluded introduced plants because these plants were not present for the large part of bees' evolutionary histories in the region. In addition, we excluded observations that were not identified to the species level, as we could not determine whether these observations were of native or introduced species.

Lastly, we excluded observations in anthropogenic habitat using QGIS and landcover data from the National Landcover Database (USGS 2016). We considered anthropogenic habitat to be land classified as 'Developed' and 'Planted/Cultivated' from the database. We also excluded the 0.5% of observations that were in the ocean (most likely there because of inaccurate spatial coordinates). After these steps the iNaturalist dataset had 805,414 observations from 3,584 species in 978 genera and 176 families.

Measuring regional abundance

For each taxon, we used the observations from iNaturalist to measure regional abundance across the entire eastern United States. Our proxy for regional abundance is the total number of unique locations (latitude-longitude combinations) at which a plant taxon was recorded. We corrected our measures of regional abundance for sampling effort, as use of iNaturalist is not randomly distributed in space – there are a greater number of users and observations near metropolitan areas and in the northern half of our study region (e.g., see Figure S1). As a result, use of raw abundance from the dataset (without correcting for sampling effort) might cause us to overestimate the abundance of plants in better sampled areas – for example, the abundance of plant taxa with more northern ranges relative to those with more southern ones.

To measure sampling effort, we calculated the total number of observations in hexagonal grid cells spaced approximately 285-km apart over the entire eastern US (see hexagonal grid cells in Figure S1 for an example). We chose this grid size by testing the explanatory power of models of plant abundance measured at various grid sizes, and picking the grid size of the model with the greatest explanatory power (see Supporting Methods for more details). We used hexagonal grid cells rather than rectangular ones, because hexagons have substantially less spatial distortion when overlayed on top of a sphere (Sahr et al., 2003). We created them using the R package *dggridR* (Barnes and Sahr, 2020).

For each plant genus, we then measured the number of observations in each grid cell, and corrected for sampling effort within the grid cell by dividing by the cell's total number of observations across genera. This gave us each genus's relative abundance in each grid cell. We then summed a genus's relative abundance across all the grid cells, to obtain the effort-corrected measure of regional abundance. Effort-corrected estimates of abundance were positively correlated with raw abundance values (Figure S2). Note that when measuring effort-corrected abundance, we binned records for all species within the same genus, because bees that specialize on one plant genus probably don't differentiate strongly between species within the genus (though there are exceptions to this, see e.g., Simpson and Neff [1983]).

Constructing the phylogeny

Because closely related plants are not independent data points, we accounted for phylogenetic relationships between plant genera in our analyses (see the analysis methods section) by constructing a genus-level phylogeny of the plants in our data. We generated the phylogeny using the R package *V.phylomaker* (Jin and Qian, 2019). The package uses the angiosperm megatree from Smith and Brown (2018), and adds new taxa that are not on the megatree by placing them near their closest relatives. We used Scenario 3 from Jin and Qian (2019) to add new taxa to the megatree (n=167 genera). Because a genus-level phylogeny was sufficient for our analyses, we pruned the species-level megatree by randomly picking one species from each genus to represent that genus. For genera with no species on the megatree, we randomly

picked a species of that genus from our study region to be added to the megatree (i.e., using Scenario 3).

Categorizing specialist host plants and picking close relatives

To determine which plant genera in our study region host pollen specialist bee species, we used a list of pollen specialist bee species and their host plants for the eastern United States (Fowler and Droege, 2020). The list has 108 bee species, which collectively specialize on 59 plant genera. We excluded twelve plant genera from the list that included major crop plants (Meyer and Purugganan, 2013). For the "close relatives" analysis (question 1), we also excluded plant genera on this list that were part of families that host pollen-specialist bee species in our study region (e.g., the plant genus *Symphyotrichum* hosts specialist bees; so does the family it's a part of, Asteraceae). We did this because the closest relatives of these plants also host specialist bee species, though at the family level.

For each plant on our list, we picked one close relative, searching specifically for plants that shared the same broad habitat association. We chose close relatives to approximately control for phylogenetic relatedness, and we matched plants by habitat association because this variable can strongly affect how plants change in abundance in response to anthropogenic change (Clavel et al., 2011). Specifically, two plants with the same habitat association should respond more similarly than two plants associated with different habitats; thus, the difference between them in abundance may be more likely to reflect what it was in the past (e.g., a forest specialist would be a poor match with a habitat generalist, because it is probably less abundant than it once was whereas the habitat generalist might be more abundant). Our detailed methods for matching plant genera by their habitat associations are described in the Supporting Methods.

To pick close relatives we calculated the phylogenetic distances between all genera in our phylogeny using the function 'fastDist' from the R package *phytools* (Revell, 2012). Then, for each plant genus used by specialists on our list, we picked the closest relative with a similar habitat association (see Supporting Methods). When a plant had more than one closest relative with a similar habitat association, we randomly picked among them. In cases where plants used by specialists were each other's closest relative (e.g., *Hydrophyllum* and *Nemophila*; *Triodanis* and *Campanula*) they also shared the same next-closest relative. In these cases, we randomly assigned one plant to be paired with next closest relative and the other plant to be paired with the next closest relative after that (assuming they shared the same habitat association). We excluded as options any plant genera that we knew to host specialist bees in other parts of the world. Figure S3 shows the phylogeny of the plant genera hosting specialists we used in the paired analysis and their close relatives.

Analyses: blinding the data

We conducted blind data analyses for all analyses (MacCoun and Perlmutter, 2015). In this type of analysis, a researcher blinds the data by randomizing it or by adding random noise; they then

conduct fake runs of the analysis on this blinded data, making the major analysis decisions before knowing what the results will be. Because researchers must wait to see the outcome of the analysis decisions, this method is a way of reducing cognitive biases that can affect the outcome of a data analysis. It prevents the researcher from subconsciously making analysis decisions that favor the results they believe a priori are more likely or more publishable. In the manuscript for this paper, we will clearly identify any analysis decisions made on unblinded data as 'post-blind.' We blinded the data in different ways for the different analyses, and describe how in each section below. The results reported in this document are from analyses conducted on the blinded data.

A script of the blinded analysis will be posted to github (https://github.com/cmsmith91/specialist_bees), as will this document, before the data are unblinded.

Analysis: Are plants that host specialist bees more abundant than close relatives with the same habitat association that do not?

To determine whether plant genera that host specialist bees have higher regional abundance than close relatives that do not, we used a one-sided paired t-test. We log-transformed the abundance data to meet assumptions of normality. (*Note*: if after unblinding the data, the assumption of normality is no longer met, we plan to analyze the data using a Wilcoxin signed rank test, which is non-parametric).

In blinding the data for this analysis, we kept pairs of plants together, and randomized within the pair which plant hosted specialist bees and which did not.

Analysis: Are more abundant angiosperm genera more likely to support pollen-specialist bees?

Prior to doing the analysis we excluded abiotically pollinated plants from the dataset in order to simplify our analysis and because we know a priori that no wind- or water-pollinated plants in our study region host specialist bees. We categorized a plant family or genus as abiotically pollinated if it was included on a list of wind- and water-pollinated plant genera or families, which we compiled from four different references (Ackerman, 2000; Eriksson and Bremer, 1992; Regal, 1982; Renner, 2014). (Note, we considered recently added members of the Plantaginaceae to be animal-pollinated rather than wind-pollinated, following [Saunders, 2018]). If a plant was not included in one of these sources, we considered it to be pollinated by animals.

To determine how a plant's abundance affects the probability it hosts specialists, we a used phylogenetic logistic regression model (Ives and Garland, 2010). More specifically, we modeled the log-odds a plant hosts specialist bees as a function of its relative abundance (corrected for sampling effort and scaled with mean=0 and s.d.=1 to allow for easier interpretation of the model coefficients) and used a log link function. The phylogenetic logistic regression model

accounts for the shared evolutionary ancestry of plant genera through the variance-covariance matrix, and the parameter α, which estimates the strength of phylogenetic signal in the model residuals. We conducted the phylogenetic logistic regression using the 'phyloglm' function of the R package *phylolm* (Tung Ho and Ané, 2014).

We used permutation tests to assess the significance of model coefficients. For these tests, we randomly reordered the response variable *n*=999 times, refitting the model to the null data each time, and recording the estimate of the null slope coefficient. (For these, we generated more permutations than needed, and discarded model runs that didn't converge, selecting the first *n*=999 that did). Our *P*-value was (*r*+1)/(*n*+1), where *r* is the number of null slope coefficients greater than the observed value (North et al., 2002). This analysis method had reasonable type I error rates (5 and 4% assuming a weak and moderate phylogenetic signal, respectively), and 80% power (assuming a weak phylogenetic signal; 97% assuming a moderate one), assuming that a one standard deviation increase in a plant's abundance increases the log-odds it hosts a specialist bee by 30% (for more details, see *Error analysis* in the Supporting Methods). Although this analysis method had an acceptable type I error rate, our simulations showed that it had little power to detect phylogenetic signal in our simulated data, and as a result produced biased estimates of the slope coefficient for data generated with the stronger phylogenetic signal (see *Error analysis* in the Supporting Methods and Figures S4). Thus, all regression coefficients from this analysis should be interpreted with caution.

In blinding the data for this analysis, we simulated data with the same phylogenetic logistic regression model that was used to analyze the data (Ives and Garland, 2010), using the function 'rbinTrait' from the R package *phylolm* (Tung Ho and Ané, 2014). For the simulation, we set the value of the slope coefficient equal to 0.3, the value of the model intercept equal to -2.725, and the value of α equal to 0.1, and used the observed angiosperm phylogeny for the variance-covariance matrix.

Results of the blind analysis

Are plants that host specialist bees more abundant than close relatives with the same habitat association that do not?

There was no difference in the abundance of plants hosting pollen-specialist bees and close relatives not hosting pollen-specialist bees (paired t-test, mean difference in log-abundance = -0.40, *t*=-1.5, *df*=30, *P* = 0.93; Figure 1). (*Note: we will only report the following if the difference is significant upon unblinding the data*). Plants hosting specialist bees were on average 4.7% less abundant than close relatives not hosting specialist bees (median percent difference = -2.7%).

Are more abundant angiosperm genera more likely to support pollen-specialist bees?

We found that more abundant angiosperm genera were significantly more likely to support pollen specialist bee species (b_0=-2.70; b_1=0.34; α =0.39; *P*=0.006). The least abundant

angiosperm in our data had an 5% chance of hosting pollen specialist bee species, whereas the most abundant angiosperm in our data had an 46% chance (Figure 2).

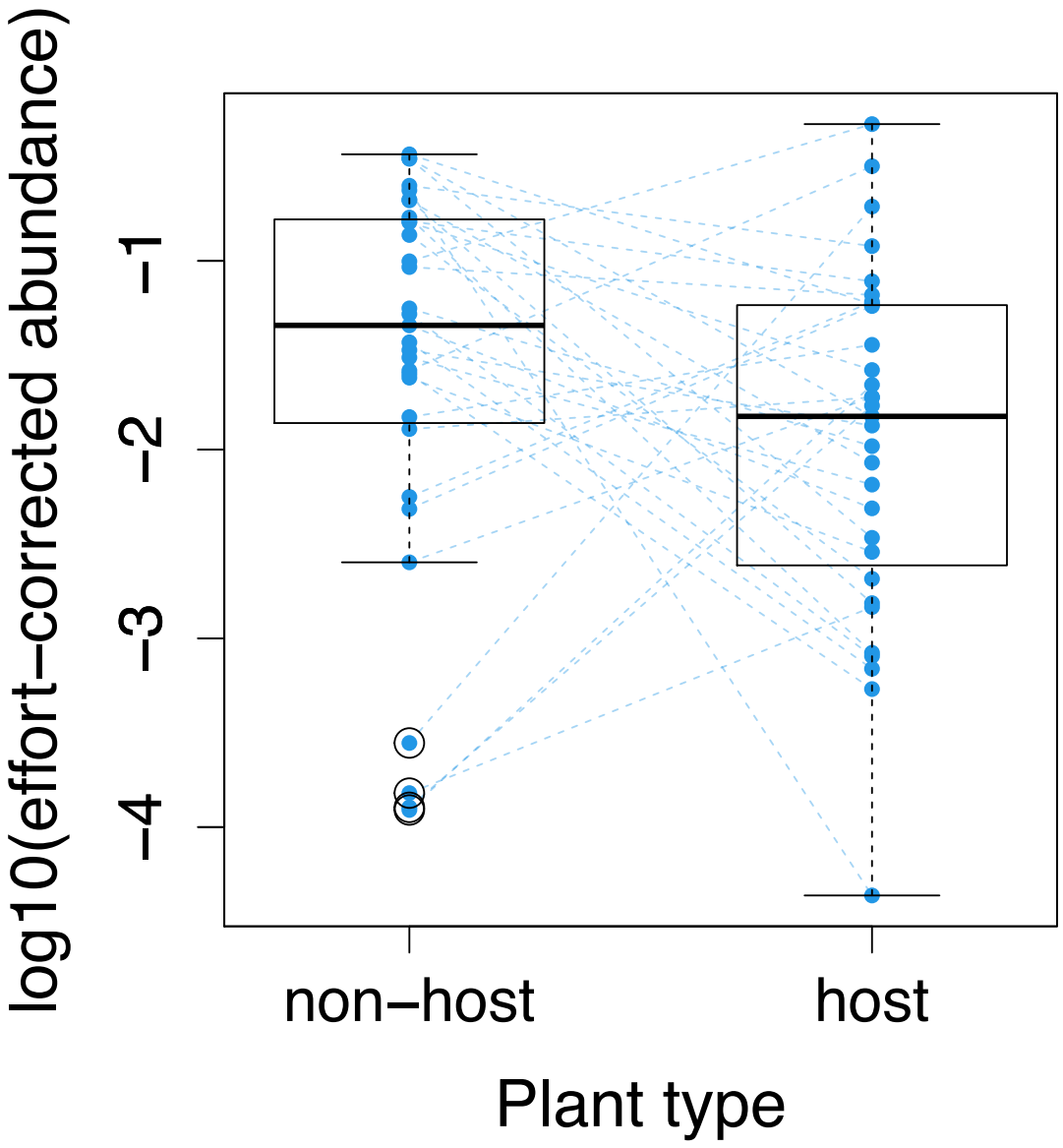

Figure 1. Abundance (log-transformed) by plant type. Each plant genus is represented by a blue circle, and dotted blue lines connect plants hosting specialist bees by their non-host close relatives. The boxes encompass the first and third quartiles of the data and the thick black line is the median. The plot whiskers extend to 1.5 times the interquartile range and larger black circles represent outliers.

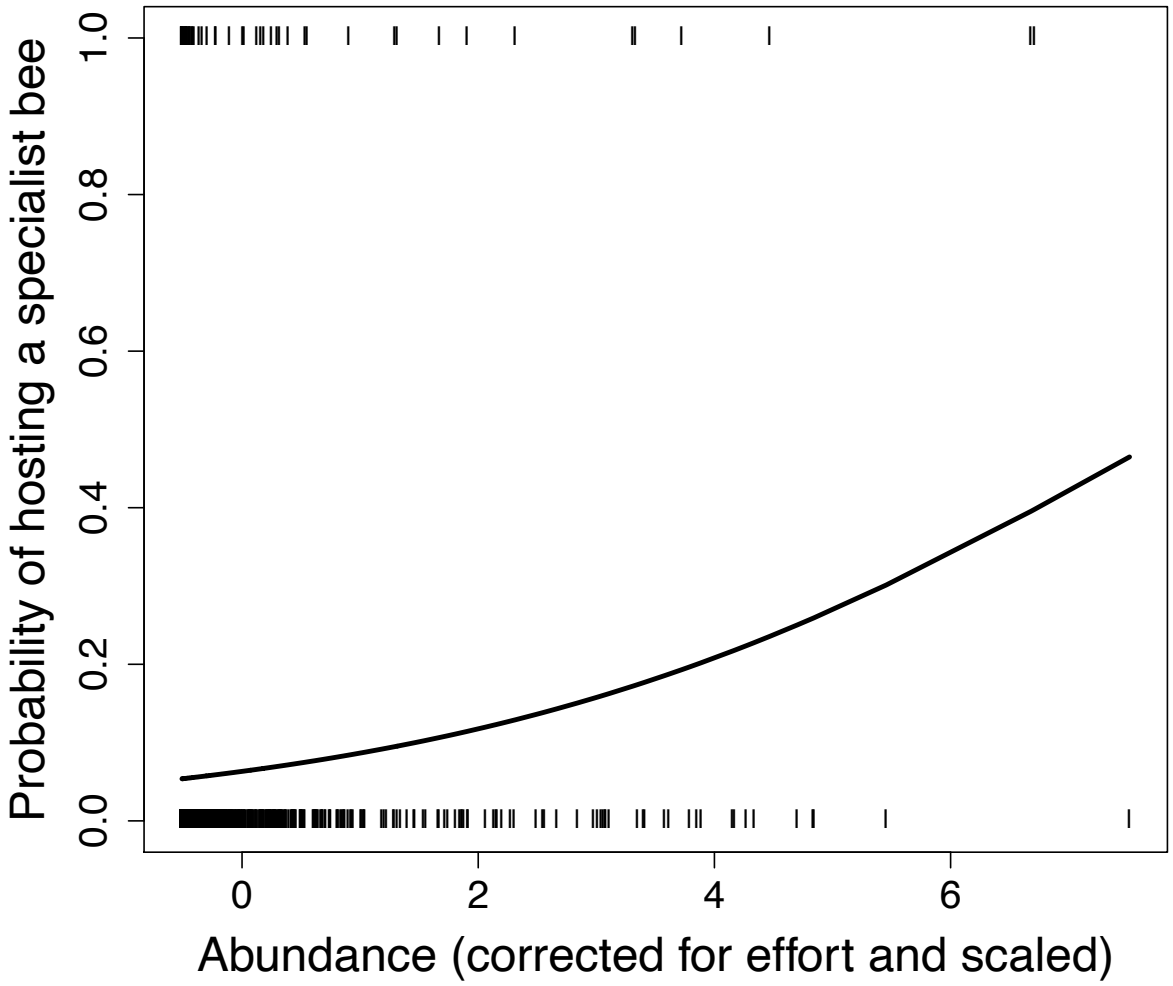

Figure 2. The relationship between a plant's abundance and the probability it hosts specialist bee species. The black line shows the model prediction, and the tick points represent the data. Tick points located at one on the y-axis indicate that the plant hosts at least one specialist bee species, and tick points located at zero indicate that it does not.

Literature cited

Ackerman, J.D., 2000. Abiotic pollen and pollination: ecological, functional, and evolutionary perspectives. Plant Syst. Evol. 222, 167–185. https://doi.org/10.1007/BF00984101

Barnes, R., Sahr, K., 2020. dggridR: Discrete Global Grids. R package version 2.0.8.

Batstone, R.T., Carscadden, K.A., Afkhami, M.E., Frederickson, M.E., 2018. Using niche breadth theory to explain generalization in mutualisms. Ecology 99, 1039–1050. https://doi.org/10.1002/ecy.2188

Cayuela, L., Macarro, I., Stein, A., Oksanen, J., 2019. Taxonstand: taxonomic standardization of plant species names.

Chamberlain, S., 2020. spocc: Interface to Species Occurrence Data Sources. R package version 1.1.0.

Clavel, J., Julliard, R., Devictor, V., 2011. Worldwide decline of specialist species: toward a global functional homogenization? Front. Ecol. Environ. 9, 222–228. https://doi.org/10.1890/080216

Cooper, N., Thomas, G.H., Venditti, C., Meade, A., Freckleton, R.P., 2016. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. Biol. J. Linn. Soc. 118, 64–77. https://doi.org/10.1111/bij.12701

Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen science as an ecological research tool: Challenges and benefits. Annu. Rev. Ecol. Evol. Syst. 41, 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Eriksson, O., Bremer, B., 1992. Pollination systems, dispersal modes, life forms, and diversification rates in angiosperm families. Evolution (N. Y). 46, 258–266.

Fowler, J., 2020a. Pollen specialist bees of the central United States [WWW Document]. URL https://jarrodfowler.com/bees_pollen.html

Fowler, J., 2020b. Pollen specialist bees of the western United States [WWW Document]. URL https://jarrodfowler.com/pollen_specialist.html

Fowler, J., Droege, S., 2020. Pollen specialist bees of the eastern United States [WWW Document]. URL https://jarrodfowler.com/specialist_bees.html

González-Varo, J.P., Ortiz-Sanchez, F.J., Vilïa, M., 2016. Total bee dependence on one flower species despite available congeners of similar floral shape. PLoS One 11, 1–17. https://doi.org/10.1371/journal.pone.0163122

Hardy, N.B., Kaczvinsky, C., Bird, G., Normark, B.B., 2020. What we don't know about diet-breadth evolution in herbivorous insects. Annu. Rev. Ecol. Evol. Syst. 51, 103–122. https://doi.org/10.1146/annurev-ecolsys-011720-023322

Ives, A.R., Garland, T., 2010. Phylogenetic logistic regression for binary dependent variables. Syst. Biol. 59, 9–26. https://doi.org/10.1093/sysbio/syp074

Jin, Y., Qian, H., 2019. V.PhyloMaker: an R package that can generate very large phylogenies for vascular plants. Ecography (Cop.). 42, 1353–1359. https://doi.org/10.1111/ecog.04434

MacCoun, R., Perlmutter, S., 2015. Blind analysis: hide results to seek the truth. Nature 526, 187–189. https://doi.org/10.1038/526187a

Meyer, R.S., Purugganan, M.D., 2013. Evolution of crop species: genetics of domestication and diversification. Nat. Rev. Genet. 14, 840–852. https://doi.org/10.1038/nrg3605

Minckley, R.L., Roulston, T.H., 2006. Incidental mutualisms and pollen specialization among bees, in: Waser, N., Ollerton, J. (Eds.), Plant-Pollinator Interactions: From Specialization to Generalization. University of Chicago Press, Chicago, pp. 69–98.

North, B., Curtis, D., Sham, P., 2002. A note on the calculation of empirical P values from Monte Carlo procedures. Am. J. Hum. Genet. 71, 439–441.

R Core Team, 2020. R: a language and environment for statistical computing.

Regal, P.J., 1982. Pollination by wind and animals: ecology of geographic patterns. Annu. Rev. Ecol. Syst. 13, 497–524. https://doi.org/10.1146/annurev.es.13.110182.002433

Renner, S.S., 2014. The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. Am. J. Bot. 101, 1588–1596. https://doi.org/10.3732/ajb.1400196

Revell, L.J., 2012. phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Sahr, K., White, D., Kimerling, A.J., 2003. Geodesic discrete global grid systems. Cartogr. Geogr. Inf. Sci. 30, 121–134. https://doi.org/10.1559/152304003100011090

Schindler, D.E., Armstrong, J.B., Reed, T.E., 2015. The portfolio concept in ecology and evolution. Front. Ecol. Environ. 13, 257–263. https://doi.org/10.1890/140275

Sexton, J.P., Montiel, J., Shay, J.E., Stephens, M.R., Slatyer, R.A., 2017. Evolution of ecological niche breadth. Annu. Rev. Ecol. Evol. Syst. 48, 183–206. https://doi.org/10.1146/annurev-

ecolsys-110316-023003

Simpson, B.B., Neff, J.L., 1983. Floral biology and floral rewards of *Lysimachia* (Primulaceae). Am. Midl. Nat. 110, 249–256.

Smith, S.A., Brown, J.W., 2018. Constructing a broadly inclusive seed plant phylogeny. Am. J. Bot. 105, 302–314. https://doi.org/10.1002/ajb2.1019

Tung Ho, L.S., Ané, C., 2014. A linear-time algorithm for gaussian and non-gaussian trait evolution models. Syst. Biol. 63, 397–408. https://doi.org/10.1093/sysbio/syu005

Wcislo, W.T., Cane, J.H., 1996. Floral resource utilization by solitary bees and exploitation of their stored foods by natural enemies. Annu. Rev. Entomol. 41, 257–286. https://doi.org/10.1146/annurev.ento.41.1.257

Supporting figures

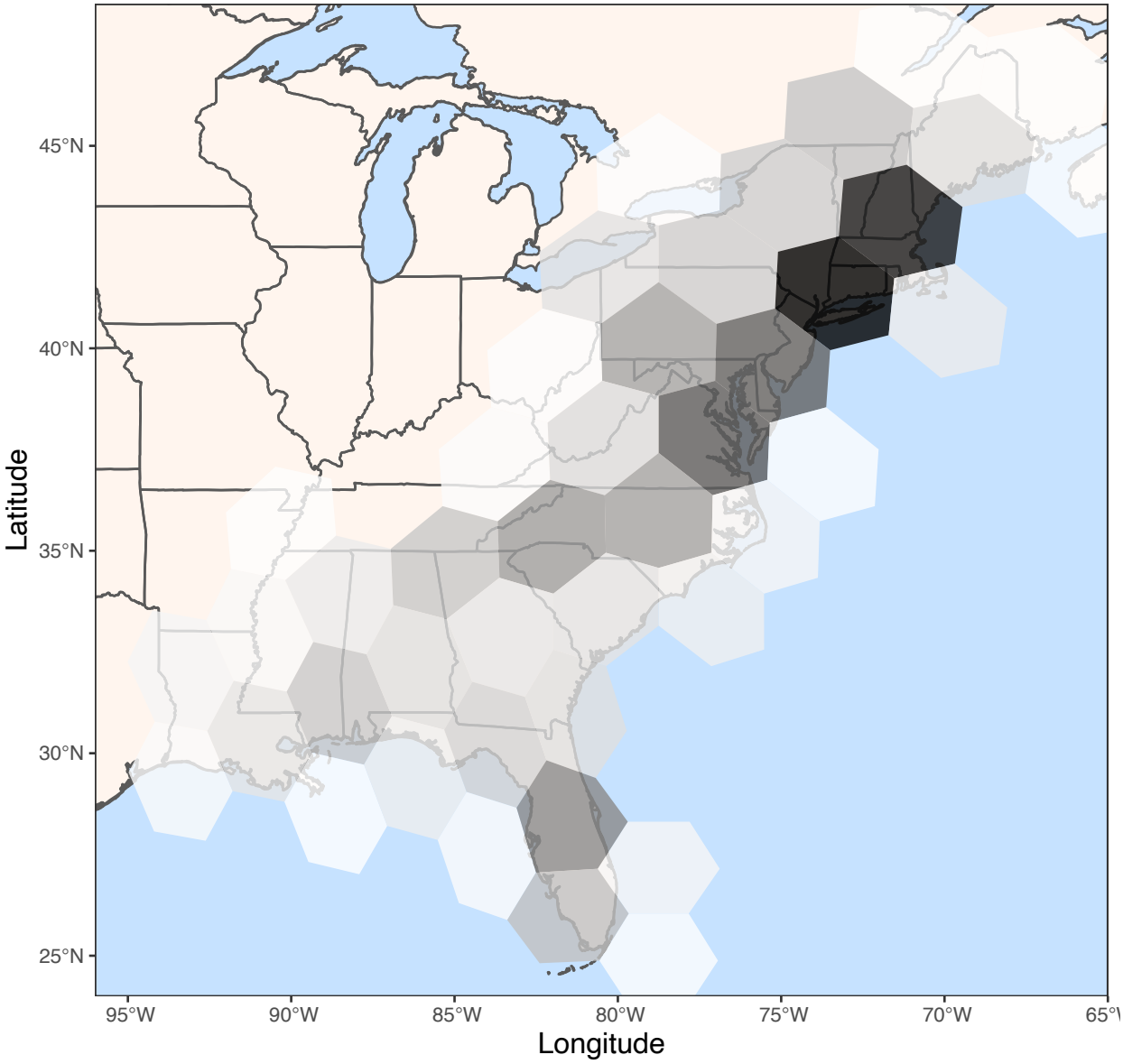

Figure S1. Map of the eastern United States, demonstrating how sampling effort in the iNaturalist dataset is greater in the north than in the south. The shading of the hexagons represents the number of observations in the hexagon, with darker shades corresponding to a greater number of observations.

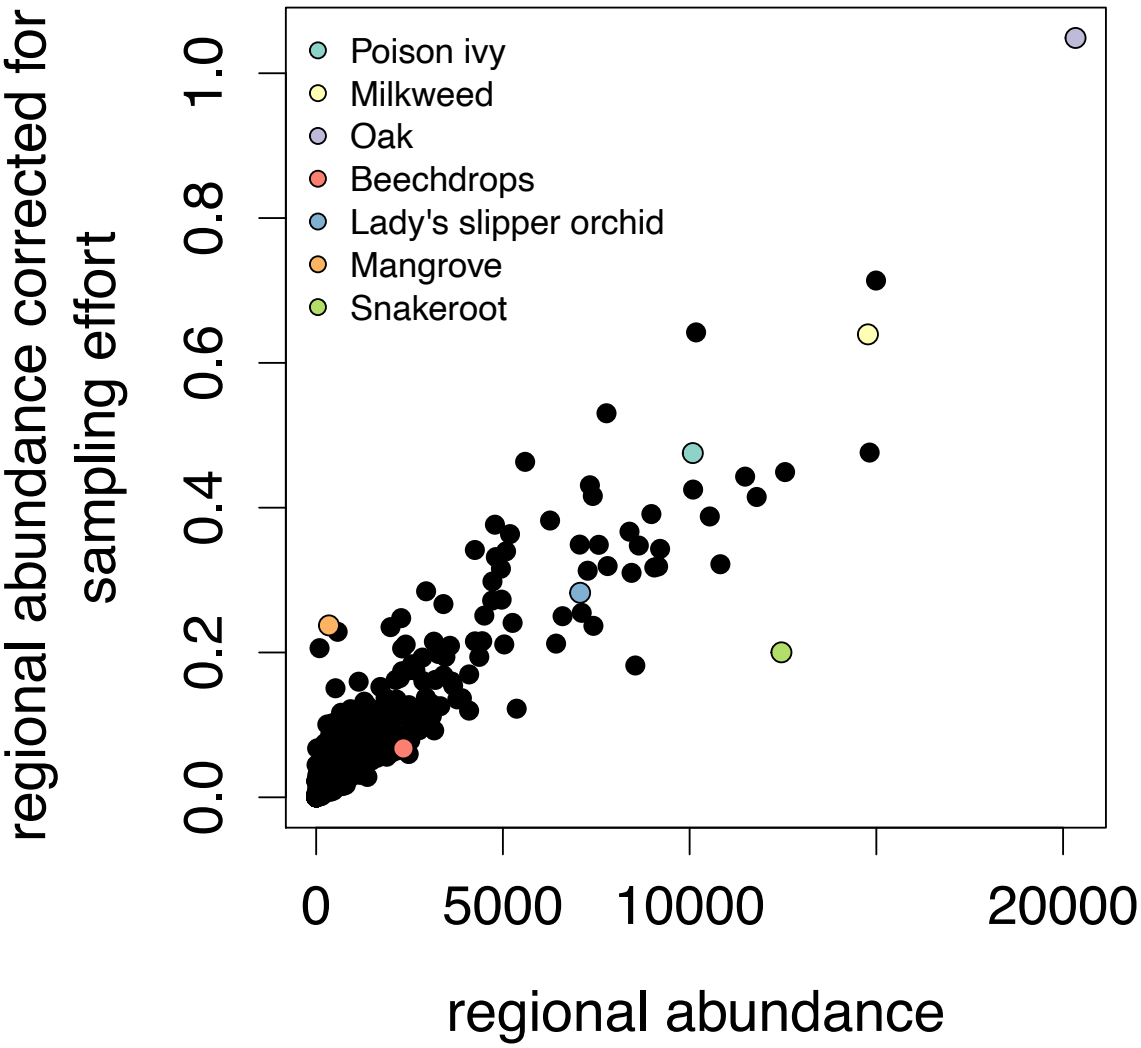

Figure S2. The relationship between regional abundance (number of unique records on iNaturalist) and our measure of effort-corrected regional abundance for each genus. The points that are labelled in color are notable genera or stand-out datapoints, and are presented to allow the reader to calibrate the data against their own impressions of taxon abundance. Common names are used in the legend are for the genera: *Toxicodendron* (Poison ivy), *Asclepias* (Milkweed), *Quercus* (Oak), *Epifagus* (Beechdrop), *Cypripedium* (Lady's slipper orchid), *Conocarpus* (Mangrove), *Ageratina* (Snakeroot).

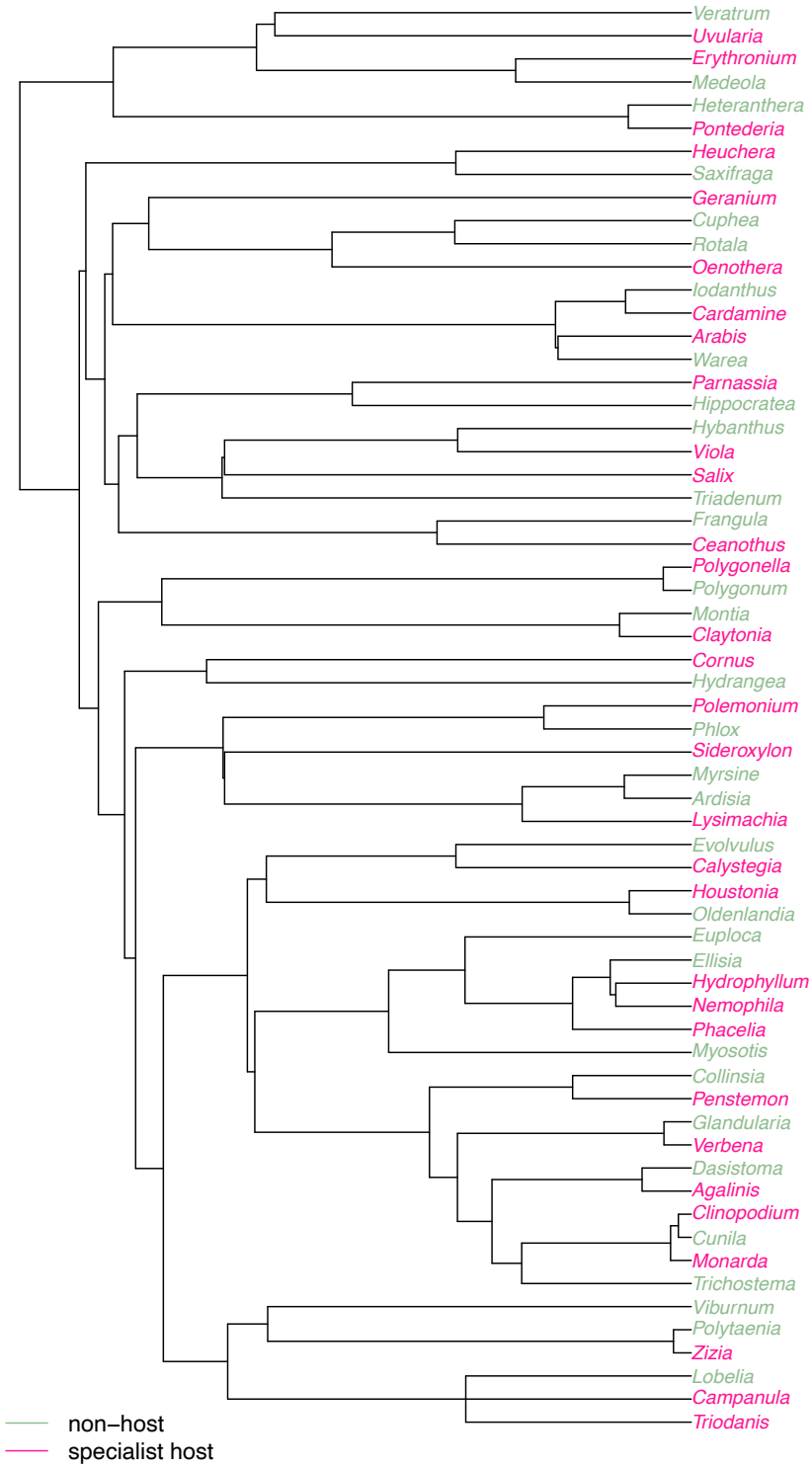

Figure S3. Phylogeny of the plants used by specialist and their close relatives chosen for the paired analysis.

Supporting methods

Comparing the habitat associations of specialist host plants and their close relatives

Our goal was to determine whether the plant species in two closely related genera have similar habitat associations. Because there are many different types of habitats in our study region, and because a plant species can be associated with anywhere from one habitat type to many, we used a multivariate approach to decide whether plant species from different genera have different habitat associations. We used habitat descriptions for each plant species from keys, online sources, and the scientific literature, and categorized each species as either being associated or not associated with thirteen different types of habitat (Table S1), assigning a value of one if the plant species was associated with the habitat type and a zero otherwise. This gave us a binary matrix, with plant species as rows and the thirteen habitat types as columns for each of the pairs of close relatives. For each pair of plant species in this matrix, we calculated the distance between their habitat associations using the Sørensen index. We then used a permutational multivariate analysis of variance (hereafter, a 'PERMANOVA') to determine whether the plant species from the two genera differed substantially in their habitat associations in multivariate space.

We used the F-statistic from the PERMANOVA to decide whether plant genera differ in their habitat associations. We considered a pair of genera to have different habitat associations if the F-statistic was greater than three (we decided on this threshold based on visual examination of NMDS plots). For plant genera that did not share the same habitat association we picked a new, next-closest relative and repeated the process again. We kept repeating this process iteratively until all specialist host plants were paired with a close relative sharing the same habitat association.

Table S1. The habitat types we used and their indicator words. When we saw an indicator word in a plant species' habitat description, we categorized the plant species as being associated with the relevant habitat type. We found the plant species' habitat descriptions from plant keys, various online sources, NatureServe, and the scientific literature.

Habitat type	**Indicator words**
Early-successional/open	fields, roadsides, grasslands, barrens, pastures, open areas, openings, prairies, open rights-of-way, clearings, trail edges, meadows, along streams, stream banks, stream sides, grassy slopes, margins of crop fields, open areas in forests, edges of forests, mesas, shrubland, brushland, scrub forest, stream margin, forest margin
Forest	woodland-conifer, oakwoods, pinelands, woodland, woods, mixed forest, pine-oak woodland, alluvial woods, pine barrens, scrub woodland flat, pineland swamps, forested seep

Moist forest/forested wetland	moist pinelands, wet pinelands, floodplain woods, moist woods, shady streams, moist forest, floodplain forests, wet-mesic forest, pine bogs
Savannah	woodland-savannah, savanna, open woodlands, flatwoods, open pineland, open woods, longleaf pine-wiregrass, hammock, sandhill
Wetland	bog, fen, saltmarsh, tidal marshes, shorelines, edges of temporary pools, streambed, swamp, wet areas, springs, wet meadows, herbaceous wetland, wet soil, damp places, pond edge, gully, seepage, seepage swamp, pond margin, pocosin, around limesinks
Coastal	dunes, near-coastal areas, near the coast, sandy beaches, mudflat, sandy seashores, coastal dunes, sandy flat, sand/dune, mangroves, brackish, coastal swales, maritime forest, dune woodlands and scrub, coastal strand
Rocky	crevices of rocks, rocks, bare rock, limestone cliffs and outcrops, limestone slope, bedrock, rocky outcrop, rocky slope, outcrop, talus, shale barren
Mountaintop	ridge, bald, ledge, cliff, bluff, mountains
Others (each coded as an additional column)	tundra: tundra, alpine areas desert: desert, desert scrub, xeric canyons: canyons chaparral: chaparral lake/pond: lake, pond

Picking the spatial scale for measuring sampling effort

To pick the most appropriate hexagon size for the grid cells used to measure sampling effort we compared the explanatory power of models of plant abundance measured within eight different sizes of hexagon grids. Specifically, we tested the explanatory power of the following eight sizes: 855, 495, 286, 165, 95, 55, 32, or 18 km mean distance between hexagons (corresponding to resolutions 4-11 in *dggridR*, respectively; Barnes and Sahr, 2020).

For each hexagon size, we modeled the mean number of observations of each genus present within the grid cell (excluding absences) as a function of the sum of all observations in the grid cell. We calculated fit using Pearson correlation coefficients. Because the grids created with smaller hexagons had a greater number of grid cells and because Pearson correlation coefficients are affected by sample size, we used subsampling when calculating the correlation coefficients for the larger datasets. For these datasets, we drew random subsamples the size of the smallest dataset and with each subsample we calculated the correlation coefficient. We repeated this process one thousand times, and use the average correlation coefficient across these one thousand iterations. For the smallest dataset (which we did not subsample) we used the Pearson correlation coefficient fit to the entire dataset. We compared the correlation

coefficients across the eight different models and picked the grid size producing the model with the greatest correlation coefficient. This grid size was 285 km.

Error analysis

To calculate type I error and power, we simulated datasets using the phylogenetic logistic regression model that was also used to analyze the data (Ives and Garland, 2010). In the model, the log-odds that a plant genus hosts pollen-specialist bees is a function of that plant's abundance. The specific model we used took the form:

$$
\begin{aligned}
&Y = \text{binom}(n, p) \\
&\log(p/(1-p)) = b_0 + b_1 * \text{abundance} \qquad \text{(Equation 1)} \\
&\text{cov}(Y) \sim \mathbf{V}(p, \alpha)
\end{aligned}
$$

where Y is a vector of 1s and 0s of length n, where 1 signifies that a plant hosts pollen-specialists bees and 0 signifies that it does not; p is the probability a plant hosts pollen-specialist bees; abundance is a vector of plants' abundances, corrected for sampling effort and standardized with mean = 0 and s.d.=1; b_0 is the coefficient of the model intercept; and b_1 is the slope coefficient, and describes the strength of the effect of abundance on the log-odds a plant hosts a specialist; $\mathbf{V}(p, \alpha)$ is the covariance matrix, in which the parameter α is equal to -log(α), and where α is a measure of phylogenetic signal; smaller values of α indicate greater phylogenetic signal (Ives and Garland, 2010).

We simulated data under this model using the function 'rbinTrait' from the R package *phylolm* (Tung Ho and Ané, 2014). To calculate type I error, we set the slope coefficient equal to zero (no effect of plant abundance on the probability the plant hosts specialists). To calculate power, we set it equal to 0.3 (a moderate effect of plant abundance on the probability the plant hosts specialists). We used two values of phylogenetic signal: a moderate signal (α = 0.01) and a weaker one (α = 0.10). In our data, these phylogenetic signals correspond to phylogenetic half-lives in Y of 69.5-my, or half the height of the angiosperm phylogeny, when α = 0.01; and 6-my, or 5% of the height of the angiosperm phylogeny, when α = 0.1 (Cooper et al., 2016).

We set the other parameter values of the model to match our observed data. We used the observed values of effort-corrected abundance from the data (standardized with mean=0 and s.d.=1); we used the observed phylogeny in modeling the phylogenetic structure. We used the same sample size as for the observed data by simulating random draws of size N=811. We made the ratio of 0s to 1s in the outcome variable approximately consistent with the observed data, by varying the value of the model intercept so that this ratio matched between the observed and simulated data. This involved a parameter exploration where we varied the model intercept for each combination of parameters (two values of phylogenetic signal by two values of the slope coefficient), simulated N=5000 replicate datasets for each parameter combination and calculated the ratio of 0s to 1s. We picked the value of the model intercept where the ratio of 0s to 1s in the data most closely matched the observed value.

To calculate the type I error rate and power, we simulated 130 replicate datasets for each parameter combination (two values of the phylogenetic signal by two values of the slope coefficient). We analyzed the 130 replicate datasets using the analysis method described in the main text, saving the p-values and estimated model coefficients from each run of the analysis. Because the model did not converge on every dataset, we discarded datasets where model failed to converge and, from 130 initial replicates, selected the first *n*=100 replicate datasets where the model converged. We calculated the error rates on these final 100 datasets.

Note:
Our simulations showed that the phylogenetic logistic regression model was not able to accurately estimate phylogenetic signal (α) from our simulated data. For both of the observed values of α (α = 0.01 and α =0.10), the function 'phyloglm' nearly always provided estimates of α near the parameter's upper bound of 0.389, with the median estimates of α ranging between 0.378 and 0.386 for the four parameter combinations. Further simulations suggested that this was typical for a range of values of α tested (n=100 replicates were generated for 10 values of α spaced evenly between 0.005-0.388; median estimates of α from the fitted models ranged from 0.385 to 0.389). Others have shown previously that the model estimator for α performs poorly when there is a high ratio to zeros and ones in a dataset, because there is less information for model fitting (Ives and Garland, 2010). Our data has a high ratio of zeros to ones (only 6.3% of the plant genera on our dataset host pollen specialist bees) and we suspect this is the reason for the poor performance of the estimator. Because the 'phyloglm' function was not able to accurately estimate α, it also tended to overestimate the value of the slope coefficient for the data simulated with the stronger value of phylogenetic signal (see bottom left graph of Figure S4). Although our simulations suggest that the type I error was acceptable (see main text), the slope coefficients should be interpreted with caution.

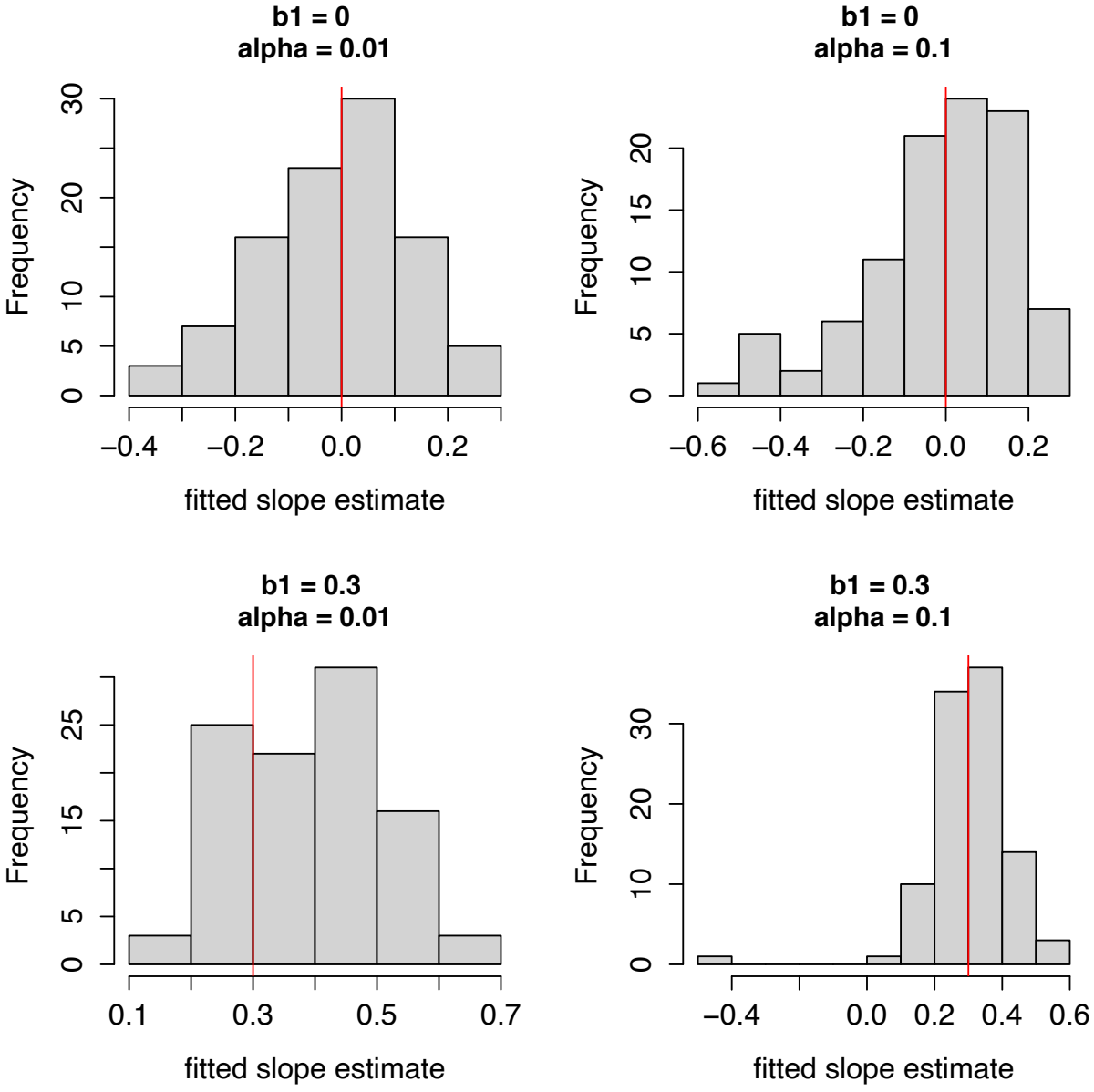

Figure S4. Distribution of the fitted slope estimates from the error analysis simulations, with each graph representing a different parameter combination. The red line shows the true value of the slope coefficient.