SEXUALLY DIMORPHIC FORAGING IN BEES AND ROBUST MEASUREMENT OF BIODIVERSITY

By

MICHAEL ROSWELL

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Ecology and Evolution

Written under the direction of

RACHAEL WINFREE

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

JANUARY, 2020

ABSTRACT OF THE DISSERTATION

Sexually dimorphic foraging in bees and robust measurement of biodiversity

By MICHAEL ROSWELL

Dissertation Director:

Rachael Winfree

The field of ecology relies on the concept of *species diversity* to describe the structure of biological communities and ecosystems. But do we measure and interpret species diversity correctly? In my dissertation, I address two problems with the current species diversity paradigm. First, in measuring species diversity, we imagine that the roles that individuals play in ecosystems are cast by their species identities. However, if variation in the traits and behaviors of individuals of the same species is large, relative to variation between different species, this paradigm may fail. Second, we imagine that we can estimate species diversity in comparable ways in different systems. Yet our tools to measure diversity pertain to samples, which, even if collected the same way, may not be equally representative. Finally, diversity estimates contain uncertainty, but existing tools to describe that uncertainty make risky statistical assumptions.

In my dissertation research, I tested whether variation within bee species might be as large as variation between them. Then, I addressed methodological challenges of comparing biodiversity. To support this research, I collected an unusual bee-plant interaction dataset. I designed my study to reveal aspects of community diversity and variation within species that

could be masked by undersampling in more typical datasets. My dataset is particularly large: ~20,000 records of bees visiting flowers, collected at only six meadows in a single summer. Each community is relatively well-sampled; the number of individuals and interactions per community is several times what pollination ecologists usually collect.

In my first chapter, on the degree of variation within species, I show how different sexes of the same bee species interact with different groups of plants. Pollination ecologists tend to lump all bees, regardless of sex, into groups by species. When lumping the sexes, ecologists assume that male bees are, from an ecological standpoint, the same as females, which, in fact, collect food for their offspring and thus forage at higher rates. I show that differences between sexes, within species, are comparable to differences between bee species. These differences arise from distinct activity periods for male and female bees, and also from flower choices made by the two sexes when they co-occur. My findings challenge lumping organisms simply by species, and may help land managers decide which mix of plants will best support bee populations.

In my second chapter, I use my dataset to illustrate best practices for measuring diversity. Rather than simply reveal patterns emerging from my own data, this chapter provides guidance on measuring the biodiversity of ecological communities more generally. Currently, conservation decisions and peer-reviewed papers rest on estimates of biodiversity that undershoot more in some places than others, often to the point of incorrectly identifying which communities are more diverse. Ecologists also use many metrics to compare biodiversity, and may lack conceptual justification for choosing one versus the other. This

chapter clarifies the biases arising from traditional tools of standardizing samples to make them comparable, reviews newer methods to reduce these biases, and provides a novel conceptual overview of diversity metrics themselves.

In my third chapter, I showed that tools to quantify the uncertainty associated with different diversity estimates are not all reliable. In this chapter I define criteria for valid confidence intervals, and introduce two tools asses those criteria, “slugplots” and “checkplots.” These plots show that popular asymptotic biodiversity estimates lack robust uncertainty estimates, which makes using them even harder. I also show an under-reported bias/variance tradeoff within a popular family of diversity metrics.

This dissertation provides new insights into the natural history of bees, clarifies best practices for measuring diversity, and rigorously assesses the statistical tools that ecologists rely upon to compare biodiversity.

ACKNOWLEDGEMENTS

First and foremost, I am thankful to Rachael Winfree for her guidance, encouragement, and high expectations. I am also extremely grateful for her support, wisdom, generosity, and insight in developing the ideas, the data, and the manuscripts that comprise my dissertation. I am very appreciative of collaborative intellectual environment Rachael maintains in our lab group, from which I have learned tremendously. I thank Jonathan Dushoff, who not only joined my dissertation committee, but also worked with me as coauthor on all three chapters, and initiated many of the intellectual insights contained therein. I would also like to thank my other committee members, Olaf Jensen and Malin Pinsky, who have been excellent teachers and provided invaluable professional mentorship. I owe thanks to Rob Muldowney for teaching me to use bigger computers and helping me navigate Rutgers's computing infrastructure. I especially acknowledge Marsha Morin, whose caring guidance not only enabled this dissertation, but underlies the community I've called home for the past several years.

I thank members of the Winfree lab past and present for their friendship, savvy editing skills, rich intellects, and piercing questions. I am so grateful to have had wonderful friendships with coworkers here, and I am grateful for Colleen Smith's company through the dissertation process, and to Bethanne Bruninga-Socolar for bringing me to the lab, adding layers of goodness to our friendship. I owe special thanks to Tina Harrison and Daniel Cariveau for teaching me about bees, but more importantly, for their friendship, counsel, mentorship, and ongoing scientific collaborations. I am grateful to Neal Williams and Nacho Bartomeus for their generous support of my research from afar. I am also grateful to the students of the

Ecology and Evolution graduate program.

Numerous Rutgers University undergraduates and technicians have helped me in the field and in the lab, and I would not have been successful in data collection without them; I owe special thanks to Joe Zientek, Kurtis Himmler, Elena Suglia, Cameron Kanterman, Riva Letchinger, Nick Cieslick, Tiffany Bennet, Shermila Villanueva, Kiara Londono, and Julie Criscione.

I feel very fortunate to have been supported by my family and friends throughout my time in graduate school. I am grateful to Chris Free for his instant trust and friendship, and the continuing discussions about research, our careers, and so much else that followed. My grandmother, Edith Sherr, instilled in me a curiosity about the natural world that she continues to fuel with loving questions and observations; my parents Barbara and Bob Roswell raised me to be curious, critical, ambitious, and caring, and they support my research with gentle, high expectations, and my siblings push me to be my best self. I am extremely grateful to my partner Jackie Specht, whose patience, support, interest, love, friendship, and insistence on work/ life balance sustained this dissertation, our relationship, and me.

I was supported by a Rutgers Excellence Fellowship and the National Science Foundation's Graduate Research Fellowship. I received research support from a small grant through the Ecology & Evolution graduate program and the USDA's Federal Conservation Innovation Grant program, though an addendum to an existing grant, *Next steps in pollinator conservation…* awarded to The Xerces Society for Invertebrate Conservation, with Rachael Winfree as one of the Co-PIs .

At the time this dissertation was submitted, the following chapters were published or under review:

Roswell M, Dushoff J, Winfree R. “Male and female bees show large differences in floral preference.” *PLoS One.* 2019;14: e0214909. doi:10.1371/journal.pone.0214909 (Chapter 1)

Roswell, M., J. Dushoff, and R. Winfree. “A conceptual guide to measuring species diversity” *Under reivew; Oikos* (Chapter 2)

Chapter 1 is reproduced here under the Creative Commons licenses. Chapter 2 is included here with permission from *Oikos*, provided that the thesis document be embargoed for 2 years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Chapter 1

Chapter 2

INTRODUCTION

I began my work in the Winfree lab on a project to assess federally-funded pollinator habitat enhancements. One of the project goals was to determine whether enhanced meadow habitats (i.e. those that had bee-attractive flowers planted in them) could support a greater diversity of wild bees. We found that much of the information about bee biodiversity that we collected over the course of the whole summer was, in fact, available to us within the first ten minutes of sampling at each site (Ward et al. 2014). At least as surprising, even when trying to estimate bee richness, it was as informative to look at bee abundance in those 10 minutes as it was to look at bee richness.

During this initial work in the lab, I was also struck by apparent patterns of bee behavior in the field that I could not find documented in the literature. It seemed possible that these patterns, for example, that male and female bees seemed to have different flower preferences, were mere extrapolations of stochastic variation in the field. I was curious, however, how these patterns, if they were real, might shape how we asked and answered questions about bee ecology.

I was pulled in two directions. On the one hand, my initial research suggested that we could avoid nuance: simply count bees for 10 minutes, and we could compare bee biodiversity, specifically, species richness, between sites. On the other, it seemed like the bee ecology literature the lab group discussed neglected the cool biology that I thought I witnessed in the field. The papers I read at the time largely assumed that differences between bee species

really mattered for bee and pollination ecology, but that the finer distinctions like time of day, bee sex, and even seasonal activity patterns would come out in the wash.

My thesis research has been motivated by a desire to reconcile my observations, my intuitions, and the data and metrics my colleagues use to answer ecological questions. In my first chapter I explored whether the nuance that stood out to me in the field could be ecology meaningful . Indeed, I found that in most bee species, males and females tended to visit different species of flowers, and that the differences between the sexes could be as great as typical differences between species.

In my second chapter, I wrestled with the strikingly strong relationship between abundance and richness in our lab datasets, and generated a conceptual review and synthesis of practices for measuring biodiversity that minimize sampling artefacts. In close collaboration with my coauthor and committee member Jonathan Dushoff, I developed a new way to visualize a mathematical insight that diversity quantifies the average species rarity of a community. This chapter is currently in review, and I am eager to see how it can impact the field of ecology broadly.

During this work we uncovered some patterns about the statistical uncertainty in biodiversity estimates that seemed at odds with the literature. My third chapter addresses these inconsistencies. Again, I collaborated closely with Jonathan to highlight statistical challenges in biodiversity measurement. Jonathan provided structured ideas on how to approach the inconsistencies we were seeing, and oversaw the simulations, analysis, and toolkit

development while I initiated our dialog with the literature and helped manifest the ideas in code, figures, and writing for an in-prep manuscript. This work contributes two key findings about Hill diversity estimation. First, the reciprocal of Simpson's concentration, often touted as a stable and reliable diversity estimate, is subject to large statistical uncertainty. Second, tools to describe statistical uncertainty in Hill diversity estimates are often unreliable, but they are much more reasonable approximations for observed sample diversity than for estimates of the true diversity of a full community. This chapter also contains methods contributions beyond critiquing Hill diversity estimation procedures. It explicates the use of two new tools for assessing frequentist statistics, both confidence intervals and p-values. It also contains a simple but to the best of my knowledge, novel approach to simulating species abundance distributions that takes true diversity as an input, rather than the true diversity of the full distribution manifesting only as an epiphenomenon. These tools will hopefully prove immediately useful to other researchers upon publication.

In the preparation of my dissertation, I have also received substantial training for a career as an ecologist. In my first chapter, I designed and executed an ambitious and successful field study that rendered a dataset that has been valuable for teaching and for intra- and extramural scientific collaborations beyond my dissertation work. I expanded my quantitative toolkit, both through developing my null model analyses, and also by learning to use random effects models. I trained and mentored several undergraduates and technicians in the process of collecting and preparing the data, and learned a great deal of bee identification.

In my second chapter, I became an expert in the literature on biodiversity assessment. I invested in conceptual and theoretical work that was neither driven by a particular hypothesis nor patterns in a particular dataset. I took on an overly ambitious project and at several points wrestled with letting go of a paper I did not believe was complete. I also wrestled with reviewer feedback that challenged the basic premises of my research, and was (with lots and lots of support and work from my two co-authors, Jonathan and Rachael) able to preserve the best parts of an earlier draft while preparing a truly major revision. I worked very hard to find confidence in the ideas and suggestions in this manuscript, and then to write them clearly and coherently.

In my third chapter, I learned foundational ideas in classical statistics. I expanded my computational toolkit, using high-performance computing resources through the university, and analyzing simulated data. I had the experience of largely drafting a manuscript without lots of step-by-step handholding or feedback from coauthors.

While some theses build strength through focus on a single guiding question, I am proud to submit this diverse dissertation, which testifies not only to my modest scholarly contributions to date, but also to the breadth and depth of both training and expertise that I acquired during my doctoral work.

CHAPTER 1

Male and female bees show large differences in floral preference

Manuscript Authors: Michael Roswell, Jonathan Dushoff, Rachael Winfree

Post-publication notes on Chapter 1

This chapter consists of a manuscript published by *PLOS One* in 2019. Since publication, I revisited the random effects model, to better understand both the median odds ratio estimates and also to determine how much of a problem the singular fits posed for model interpretation.

Figure 4 in the manuscript displays the estimated median odds ratio for each random effect in the model. This is a summary of the model fit, and describes the importance of each effect in terms of how much the odds of a bee being male are expected to change between levels of that effect. There are two features of this figure that gave me some pause upon recent re-inspection. First, in the more complex model with more interaction terms, the median odds ratios for the interactions with sampling round are often estimated with extremely high precision. This is puzzling, my intuition is that estimating complex interaction terms should be subject to greater, not less, uncertainty than estimating the main effects. Second, several median odds ratios are at exactly 1 with no error, corresponding to estimates of 0 variance for some effects from the model fit.

The small error terms on the interaction effects make some degree of sense. Imagine for a moment that each level for these interaction effects has been exactly and correctly captured by the BLUP in the model (that is, that the predicted effect of each combination of factors is always right). In that case, the precision with which the median odds ratio can be estimated would be entirely governed by the number of BLUPs, i.e. the number of levels of that random effect. In this case, median odds ratio should be estimated with highest precision for those effects with the greatest number of realized levels in the data, and with the least precision for the effects represented by the fewest levels in the data. Therefore, uncertainty for the interaction terms would be smaller than for the main effect terms because there are more levels realized in the data for the interaction effects (i.e. more BLUPs).

One complicating issue here is that the number of observations at each level may be very small. While a single observation is an unbiased estimate of the mean, it is a terribly unreliable estimate. Because a small number of observations at each level occur more commonly for the interaction terms (same number of observations, just split among more levels), I could imagine that the model confounds variability within a level and variability between levels, and I expect this should contribute to the real uncertainty in the median odds ratio for interaction terms, though the way we computed this uncertainty it would not. I have not explored this issue further.

The singular fit is a potentially more serious issue. It is literally not conceivable that there is truly 0 variability in sex ratio between sites, for example. The singular fit, then, could arise for several reasons, two of which I outline below. First, a singular fit can occur when there is not

enough data to specify a model, leading to overfitting and essentially meaningless estimates. Second, a singular fit can occur when fitting some models when an effect is "close to the boundary." This type of problem is expected in binomial models, where correctly estimating probabilities that are truly close to 0 or 1 is technically challenging. I did not attempt to distinguish between these possibilities directly, opting instead to determine if the singular fit indicated that our conclusions were suspect.

Two approaches I took to troubleshoot and examine the singular fits are 1) to fit the same model with different optimizers, and 2) to fit the same model in a Bayesian framework. We did the first of these in preparing the manuscript for publication, and it gave us some confidence that we were not making really foolish claims. Since then, packages for fitting random effects models in STAN through R have become readily available (Ali et al. 2019). These tools allow users to specify models using *glm*-type linear model syntax in R, and the software pre-selects sensible priors. I imagine that singular fits are not technically possible when fitting a binomial random effects model with STAN, assuming that the prior does not include probabilities of 0 or 1. If the singular fits in our model resulted from boundary issues, then I expect that fitting the same models in STAN will give very similar estimates, with the effects given a variance of 0 in *lme4* (Bates et al. 2016) having a small variance term in the fit from STAN. If the singular fits in our model are a symptom of an overly complex model for the available data, then I expect that the STAN fit could differ substantially.

Whereas *lme4* and *rstanarm* (Ali et al. 2019) use tools from somewhat different schools of statistical philosophy, if we fit the same models in both, we can leverage the fact that they

have quite different fitting machinery to assess the gravity of our singular fits for model interpretation.

Effects that did not have a 0-variance estimate in *lme4* had very similar variance estimates in *rstanarm*, and those that had a variance estimate of 0 from *lme4* had very low variance estimates when fit with *rstanarm*. This gives me confidence that our published results are approximately correct.

Abstract

Background

Intraspecific variation in foraging niche can drive food web dynamics and ecosystem processes. In particular, male and female animals can exhibit different, often cascading, impacts on their interaction partners. Despite this, studies of plant-pollinator interaction networks have focused on the partitioning of the floral community between pollinator species, with little attention paid to intraspecific variation in plant preference between male and female bees. We designed a field study to evaluate the strength and prevalence of sexually dimorphic foraging, and particularly resource preferences, in bees.

Study design

We observed bees visiting flowers in semi-natural meadows in New Jersey, USA. To detect differences in flower use against a shared background of resource (flower) availability, we maximized the number of interactions observed within narrow spatio-temporal windows. To distinguish observed differences in bee use of flower species, which can reflect abundance patterns and sampling effects, from underlying differences in bee preferences, we analyzed our data with both a permutation-based null model and random effects models.

Findings

We found that the diets of male and female bees of the same species were often dissimilar as the diets of different species of bees. Furthermore, we demonstrate differences in preference between male and female bees. We show that intraspecific differences in preference can be robustly identified among hundreds of unique species-species interactions, without precisely quantifying resource availability, and despite high phenological turnover of both bees and plant bloom. Given the large differences in both flower use and preferences between male and

female bees, ecological sex differences should be integrated into studies of bee demography, plant pollination, and coevolutionary relationships between flowers and insects.

Key words

dimorphism, dissimilarity, Morisita-Horn, phenology, plant-pollinator interaction, pollination, pollinator habitat, preference

Introduction

Intraspecific variation in traits and behavior, including foraging niche, has important consequences for species interactions and conservation (Durell 2000, Bolnick et al. 2011). Sexual dimorphism is a large source of individual niche variation, and an important factor in plant-animal interactions, such as seed dispersal (Zwolak 2018). Sexual dimorphism underlies adaptation, speciation, and the way in which animals exploit their ecological niche (Butler et al. 2007, Temeles et al. 2010). Morphological, behavioral, and life-history dimorphisms can also drive the form and function of ecosystems, for example when predator sex ratio drives the community composition of lower trophic levels, affecting the physical and chemical properties of the environment (Fryxell et al. 2015, Start and De Lisle 2018).

Though ecological dimorphisms were first studied in vertebrates (Selander 1966), they are common across taxa, including insects (Shine 1989). Surprisingly, in bees (Hymenoptera, Apoidea) for which both foraging (Willmer 2011) and sexual dimorphism (Alcock et al. 1978) have been well studied, sexually dimorphic foraging has rarely been documented. Intraspecific variation in floral preference is known for social (Heinrich 1979) and to a lesser

extent, solitary bees (Tur et al. 2014, Bruninga-Socolar et al. 2016), yet most community-level studies focus on species-level interactions, and specifically on how female bees forage.

Male bees differ from their better-studied female counterparts in their life history and ecology. Female bees construct, maintain, provision, and defend nests, whereas male bees primarily seek mates (Willmer and Stone 2004). Both sexes drink floral nectar for their own caloric needs, but only females collect pollen to provision young, and thus forage at greater rates. Pollens from different flower species (the term we use throughout for the flowers from a species of plant) tend to be distinct not only in morphology but also in terms of nutritional content, and both of these (factors drive plant-specific foraging by bees (Cane and Sipes 2006). While recent work shows variation in nectar is more important than previously acknowledged (Parachnowitsch et al. 2018), even species of bees that specialize narrowly on pollen hosts (oligolects) tend to nectar from many flower species. It is unclear what factors drive male bee preferences, though the criteria males use to select floral partners probably differ from those used by females.

Although female bees are more prolific pollinators due to the greater time they spend foraging at flowers, when male bees have been studied, they prove to be important pollinators as well. This is true not only in specialized oil- or scent-collecting pollination systems, where males would be predicted to be important (Janzen 1971, Eltz et al. 2007, Etl et al. 2017), but also for males simply foraging for nectar (Cane 2002, Cane et al. 2011, Ogilvie and Thomson 2015). Male bees may also be particularly relevant for bee conservation. Males may be limiting in declining populations, either because genetic diversity is necessary for the development of

female offspring as a result of complementary sex determination, or because mate or sperm limitation results from poor male condition (Elias et al. 2010, Straub et al. 2016, Gloag et al. 2019). As the dispersing sex in most bee species, males may be crucial for gene flow and metapopulation persistence even when they are not locally limiting (Ulrich et al. 2009, López-Uribe et al. 2015).

Observed differences in resource use, which reflects the overlap of consumers and resources (availability) as well as consumer preference, may fail to reveal more essential differences in foraging niche. Preferences may be more important than use alone in the context of species conservation, and may mediate the strength of selection imposed by interaction partners. Preference—the use of a resource in excess of its relative availability—is challenging to measure, because both resource use and availability must be known. Floral resource availability for pollinators is particularly hard to quantify outside an experimental context because the appropriate scale and units of floral resource availability are unclear. The composition, amount, and supply rate of pollen and nectar per flower, the number of flowers per inflorescence, of inflorescences per individual, and the number and distribution of individual plants over the square kilometers of a bee's foraging range are all important components of availability (Hicks et al. 2016). Furthermore, floral availability can change rapidly over time. However, differences in flower use between bees foraging at the same place and time indicate differences in preference, which may occur between species, or between individuals of the same species.

In this study, we assess differences between floral preferences of male and female bees in the

field. We collected bees foraging on flowers in meadows in New Jersey, USA. In order to observe preference differences, we collected as many individuals as possible during replicated, short (3-day) windows, during which we assumed floral availability and bee abundance were constant at each site. We compare the species composition of flowers visited by males and females of the most common bee species across the entire study as a naïve measure of differences in preference between the sexes. Then, using random effects models, we assess when differential flower species *use* by male and female bees likely arises from sex-specific floral *preference*, as opposed to shifting overlap between foragers and floral resources (i.e. changes in *availability* without differences in preference). Specifically, we ask

1) How much do male and female bee diets overlap?
2) To what degree are particular flower species disproportionately visited by bees of one sex?
3) To what extent are differences in floral use driven by preference, rather than phenological differences between male and female bees?

Materials and methods

Study design and data collection

Because absolute preference is nearly impossible to observe outside of an experiment, we designed our study to reveal differences in preference between groups of bees. In order to collect a large number of males and females from

many native bee species, we selected six meadows (sites) in New Jersey, USA with a high abundance and diversity of flowers. These semi-natural meadows were managed for pollinator-attractive, summer-blooming forbs through seed addition, and a combination of mowing, burning, and weed removal. Most flower species present in the meadows are native to the eastern United States. We collected our data during peak bloom and maximum day length (6 June to 20 August 2016), and during good weather (sunny enough for observers to see their own shadow, no precipitation). We visited each site for three consecutive good weather days over five evenly spaced sampling rounds in the 11-week period of our study. In all analyses, we assume that bees and flowers detected at a site within one 3-day sampling round co-occurred. In contrast, we assume that turnover of both plant species in bloom and bee species activity can occur in the ~10 days between sampling rounds.

During each 3-day sampling round, an observer walked parallel transects through the meadow (which ranged in size from 0.8–2.2 ha; mean=1.4 ha), observing every open flower within a moving 1-m semicircle, and net-collecting any bee seen actively foraging, which we defined as contacting anthers or collecting nectar from a flower (Figure A in S1 File). We collected all bee species except *Apis mellifera* L.*,* the domesticated western honey bee, because *Apis* males do not forage. Observations began as soon as pollinator activity picked up in the morning (7–9 am) and continued into the late afternoon or evening until pollinator activity slowed substantially. Observers sampled nearly continuously, in 30-minute timed collection bouts with short breaks in between. If inclement weather precluded a minimum of six 30-minute sampling bouts in a day, we added an additional day to the sampling round as soon as weather permitted.

Flower species were identified in the field by the data collector. Bee species were identified using a dissecting microscope and published keys; Jason Gibbs (University of Manitoba), Joel Gardner (University of Manitoba), and Sam Droege (USGS) assisted with identification for bees in the genera *Andrena, Anthophora, Coelioxys, Halictus, Heriades, Hoplitis, Hylaeus, Lasioglossum, Megachile, Melissodes, Nomada, Osmia, Pseudoanthidium, Ptilothrix, Sphecodes, Stelis,* and *Triepeolus,* and at least one of them confirmed voucher specimens for every species. We determined every specimen to species except for the following four complexes: Most bees in the genus *Nomada* with bidentate mandibles (*ruficornis* group) were treated as one species. All specimens from the *Hylaeus* species complex that includes *Hylaeus affinis*, *H. modestus*, and at least one additional species, informally dubbed "species A*,"* were treated as a single species, denoted *Hylaeus affinis-modestus*, because females cannot be reliably distinguished. There is a cryptic species in the genus *Halictus* unlikely to occur in our area, *Halictus poeyi*, which is not morphologically distinct from *H. ligatus*; we treat all specimens in this complex as *Halictus ligatus*. We could not confidently separate all specimens of the two closely related *Lasioglossum* species *Lasioglossum hitchensi* and *L. weemsi*. Thus, we treat all specimens from either species as one, denoted *Lasioglossum hitchensi-weemsi.*

All bee specimens are curated in the Winfree lab collection at Rutgers University, and the data used in this paper, along with R scripts used in data analysis and figure preparation, are available from the Dryad Digital Repository (doi:10.5061/dryad.c3rr6q1). No specific permits were required to collect these data, however, we obtained permission to access meadow

habitats and sample insects from Mercer County, the Institute for Advanced Study, Somerset County, and the Raritan Headwaters Association

Analytical methods

We performed all statistical analyses and simulations using R 3.5.1 (R Core Team 2018).

1) How much do male and female bee diets overlap?

To compare the diets of male and female bees, we used the Morisita-Horn index of resource overlap (Morisita 1959, Horn 1966). This dissimilarity index compares the proportion of all female bees found on each flower species to the proportion of all male bees found on each flower species. In other words, it compares the contribution of each flower species to female diets (where this term includes the food that females collect for themselves and also to feed to young) to the contribution of the same flower species to male diets. The Morisita-Horn index ranges from zero (completely similar) to one (maximally dissimilar), and has several good properties for our purposes. First, it uses proportions, placing visits from male and female bees on the same scale, even though most visits come from females. Second, it is much more sensitive to large proportions than to small ones, thereby down-weighting the contribution of flower species for which we have little information. Third, the Morisita-Horn estimates are resilient to undersampling and uneven sample size between groups (Barwell et al. 2015).

To determine whether the male-female differences we observed exceeded those expected by chance, we compared the observed compositional dissimilarity between flower visits from male and from female bees to dissimilarity measures from a null model that randomly permuted the bee sex associated with each flower-visit record. This permutation holds constant the total number of male and of female visits, and the total number of visits to each flower species from both sexes combined (Figure B in S1 File). The range of dissimilarity values from this simulation is the difference we would observe in our sample, if there were no true difference in flower species use between males and females of the same bee species. We evaluated the hypothesis that male and female diets overlap less than would be expected by chance; thus, we use a one-sided alpha of $p<0.05$. We iterated this null simulation 9999 times, which was sufficient to stabilize p-values near our chosen alpha (North, Curtis & Sham 2002). When the observed dissimilarity was greater than 9500 of the 9999 simulated dissimilarities, we concluded that we had detected a difference in the pattern of floral visitation between conspecific male and female bees, given the observed diet breadth and abundance of each sex. We also computed the mean null model value, and a 95% confidence interval for this mean using the 0.025 and 0.975 quantiles of the dissimilarity values generated for each null model.

To compare the diet overlap we observed between sexes to a meaningful benchmark, interspecific diet overlap, we repeated the same null model analysis, this time comparing females of the focal species to females of other species. To compare the results, we present the difference between the observed dissimilarities and the null dissimilarities for each female-male and species-species comparison. We performed one analysis for each bee

species for which we collected at least 20 visitation records for each sex (19 species). This sample size threshold is arbitrary, but null model variance shrinks with sample size, such that apparent patterns for species with smaller sample sizes are rarely interpretable (Figure C in S1 File).

Because we analyze 19 bee species, females of each species are compared to 18 others. We then compared the female-male difference (observed minus mean null dissimilarity in flower communities visited) to the analogous species-species difference (observed minus null dissimilarity).

For this analysis, which evaluates holistic differences between male and female bees of the same species, we combined observations across the full season and all sites. This allows us to observe foraging niche differences that are driven by flower and/ or bee phenology, in addition to any sex-specific floral preference.

2) To what degree are particular flower species disproportionately visited by bees of one sex?

To answer this question, we fit a random intercepts model to our entire data set of 153 bee species to determine whether particular flower species are disproportionately visited by male or female bees, and whether the answer varies by bee species. In our model, bee sex is the response, and flower species, bee species, site, and their interactions are all random effects; we included no fixed effects. The random effects provide partial pooling, which is especially useful when there are many levels, few data associated with some or all levels, and/ or

inconsistent amounts of data across levels (Gelman and Hill 2007). We can infer disproportionate visitation by male vs. female bees for a flower species when predicted odds of visitors to that flower species being male are especially high or low.

We statistically control for variation in the overall sex ratio across bee species through a random intercept of bee species, and variation in sex ratios across sites, through random intercepts for site, and the site-bee species and site-flower species interactions. Although we deemed it unlikely that, within bee species, sex ratios at birth vary greatly across space, any variability attributed to site terms could result from differences in bee sex ratios, or from differential overlap of bee foraging activity and flower bloom across space.

We call this model the “summed model” because we sum interactions observed across the entire season (all five sampling rounds) at each site. In the summed model, the relationship between phenological overlap and the odds of flower-visiting bees being male would be incorporated into the species effects. This perspective is helpful for considering flower species’ contributions to the overall diets of male versus female bees. We fit the model with the R package lme4 (Bates et al. 2016) with the following call:

Summed model

```
lme4::glmer(bee_sex ~ (1|site)+ (1|flower_species)+ (1|bee_species)+
        (1|flower_species:bee_species)+ (1|site:bee_species)+ (1|site:flower_species),
        family=”binomial”, data=data)
```

We included bee species and site as random, rather than fixed, effects to directly compare the variability in bee sex associated with each of these predictors to the variability associated with flower species (preference). Comparing the overall variability across these groups was more important to us than assessing predictions on a per-site or per-bee-species basis. We fit flower species, the primary covariate of interest, as a random effect to facilitate model fitting (fewer degrees of freedom) as well as interpretation. In our summed model, we included all two-way interactions, but omitted the three-way interaction, bee species by flower species by site. Although the sort of context-dependent preference this term could represent (e.g. males from bee species *1* prefer flower *A* at one site (relative to females), but shun it at another) may exist in nature, it is unlikely we would estimate it accurately in our model.

We confirmed model convergence by comparing several fitting methods using the allFit function in lme4 (Bates et al. 2016), which all showed similar parameter estimates (Table A in S1 File). We tested whether residuals from our model were over dispersed using Bolker's function "overdisp" (Bolker 2017), and visually assessed our additivity assumptions with binned residual plots (Gelman and Hill 2007) (Figure D in S1 File).

3) To what extent are differences in floral use driven by preference, rather than phenological differences between male and female bees?

Over the 11 weeks of our study, we observed turnover in bee species, in flower bloom, and within-bee species changes in sex ratio. Therefore, phenological overlap between male versus female bees and the bloom period of particular flower species, rather than preference of those

bees for those flowers, may explain much of the variation in sex ratio we observed across visitors. In question 3, we are explicitly interested in distinguishing sex-specific diet *preferences* from variable *use* resulting from seasonal resource availability and male vs. female abundance. We do this in the "seasonal model" by incorporating sampling round (our measure of phenology) as an additional random intercept effect, along with random intercepts for the interactions between sampling round and the other covariates. We chose to include sampling round as a random effect because this enables direct comparison to all other terms in both models. We ignored the three- and four-way interactions between bee species, flower species, and other covariates. We fit this model with the following call in the R package lme4, with new terms in bold:

Seasonal model

glmer(bee_sex ~ (1|site)+ (1|flower_species)+ (1|bee_species)+

(1|flower_species:bee_species)+ (1|site:bee_species)+ (1|site:flower_species)**+**

(1|sampling_round)+ (1|site:sampling_round)+

(1|flower_species:sampling_round)+

(1| bee_species:sampling_round)+ (1|site:bee_species:sampling_round)+

(1|site:flower_species:sampling_round), family="binomial", data=data)

Our index of preference for both the *summed model* and the *seasonal model* is the change in predicted odds that a bee is male when the flower species it visits is given. To describe the importance of model terms, we calculated a bootstrapped median odds ratio using code from Seth (Seth 2017), which gives the expected difference in odds that a flower-visiting bee is male

between levels of a predictor (Merlo et al. 2006). For example, a median odds ratio of five for the main effect of sampling round would indicate that the odds of a flower-visiting bee being male differ by about a factor of five between sampling rounds, while a median odds ratio of one would indicate that the odds of a flower-visiting bee being male do not change across rounds. If the median odds ratio is large for flower species in both models (and the random effects predictions for each species are consistent across both models), we could say that there are intrinsic (i.e. not simply phenological) properties of flower species identity that male or female bees prefer. If flower species is a strong predictor of bee sex in the summed model but not in the seasonal one, we would still conclude that flower species often contribute more strongly to the diet of one sex than the other, though these differences may not arise due to differing preferences. If the sampling round terms have large median odds ratios, then accounting for phenology is critical for identifying differences in preference in addition to differences in use.

Because male bees require less pollen for their own diets and do not collect pollen to provision offspring, we predict that they would be less likely to visit flower species that do not produce nectar than females would. Post-hoc, we examined whether the predicted odds that visitors to nectarless flower species would be male were lower than for flower species known to produce both pollen and nectar in the seasonal model (Appendix A in S1 File).

Results

In total we collected 18,698 bee specimens belonging to 152 bee species (Table B in S1 File) from a total of 109 flower species (Table C in S1 File), which together comprised 1417 unique species-species interactions. Although the ratio of male to female bees was highly variable across bee species (Table B in S1 File), roughly 18% of specimens were male (n=3372). Thus, the overall ratio of male to female bees we collected was 0.22, although this ratio varied markedly between flower species (Fig 1).

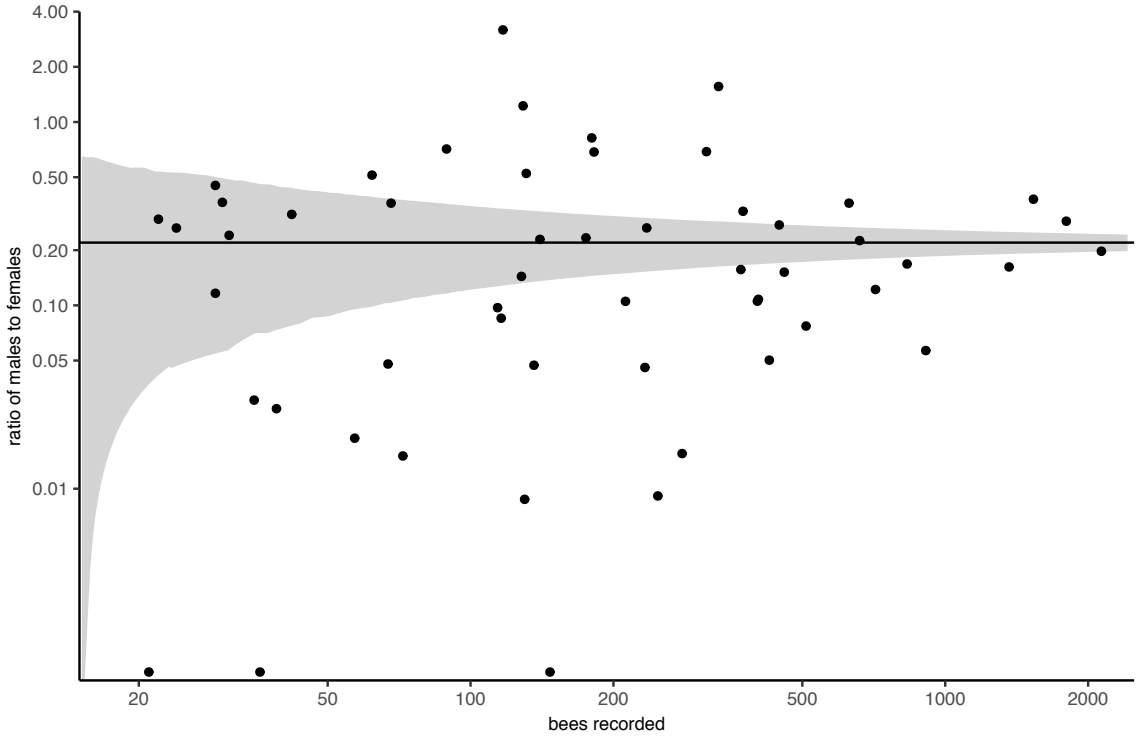


Fig 1. The sex ratio (M:F) of flower-visiting bees varies across flower species.

Each point represents a flower species; the x-axis is the number of bees collected from that species, the y-axis is the ratio of male to female bees collected from the flower. Flower species that received >19 visits are plotted (n=54). The shaded region is bounded by a smoothed fit to the 97.5th and 2.5th percentiles of the binomial distribution given by the observed ratio of males to females in our overall dataset (M/F=0.22; i.e. M/(M+F)=0.18). This distribution represents our expectation for random variation in sex ratio across flower species, if the sex

ratio of flower-visiting bees is independent of flower species identity (male and female bees exhibit the same floral preferences), and remains nearly constant across time and space.

How much do male and female bee diets overlap?

We found that male and female bee diets overlap significantly less than would be expected given random sampling of the flowers visited by both sexes (Fig 2), and that the differences in diet composition between male and female bees of several species were of similar magnitude to the differences in diet between species of bee (Fig 3). The patterns we observed did not result from a single tendency across all bee species, such as males always visiting a nested subset of flower species visited by females (Figure E in S1 File).

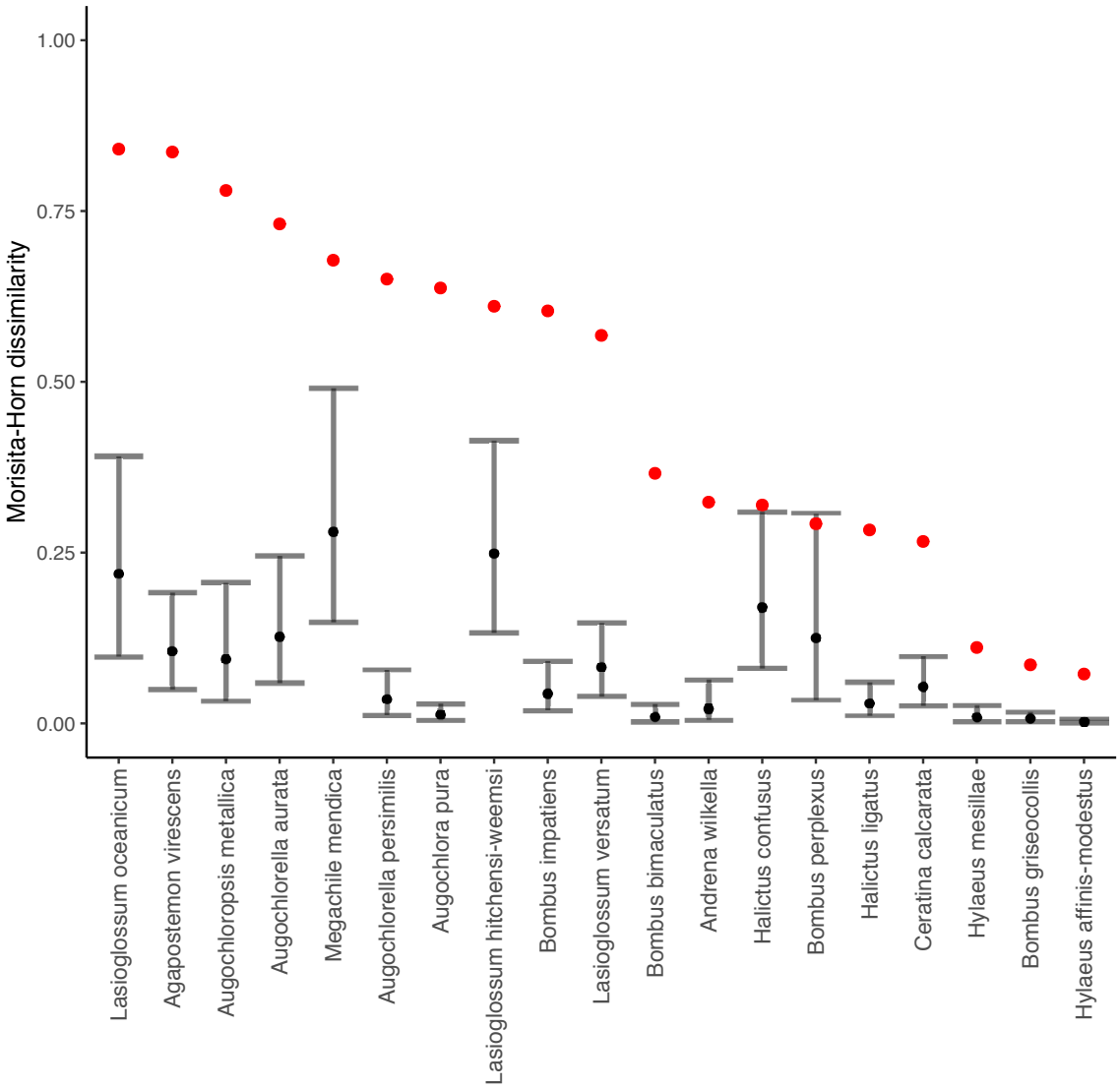


Fig 2. Flower visit patterns of male and female bees of the same species differed significantly.

Red points are observed Morisita-Horn dissimilarities between flower communities visited by all male and all female bees of a particular species across all sites and sampling rounds. Black points are the mean dissimilarity (gray bars, 95% CI) from a permutation-based null model that randomly shuffles the sex associated with each visit record, maintaining the total number of males, females, and overall combined visits to each floral species.

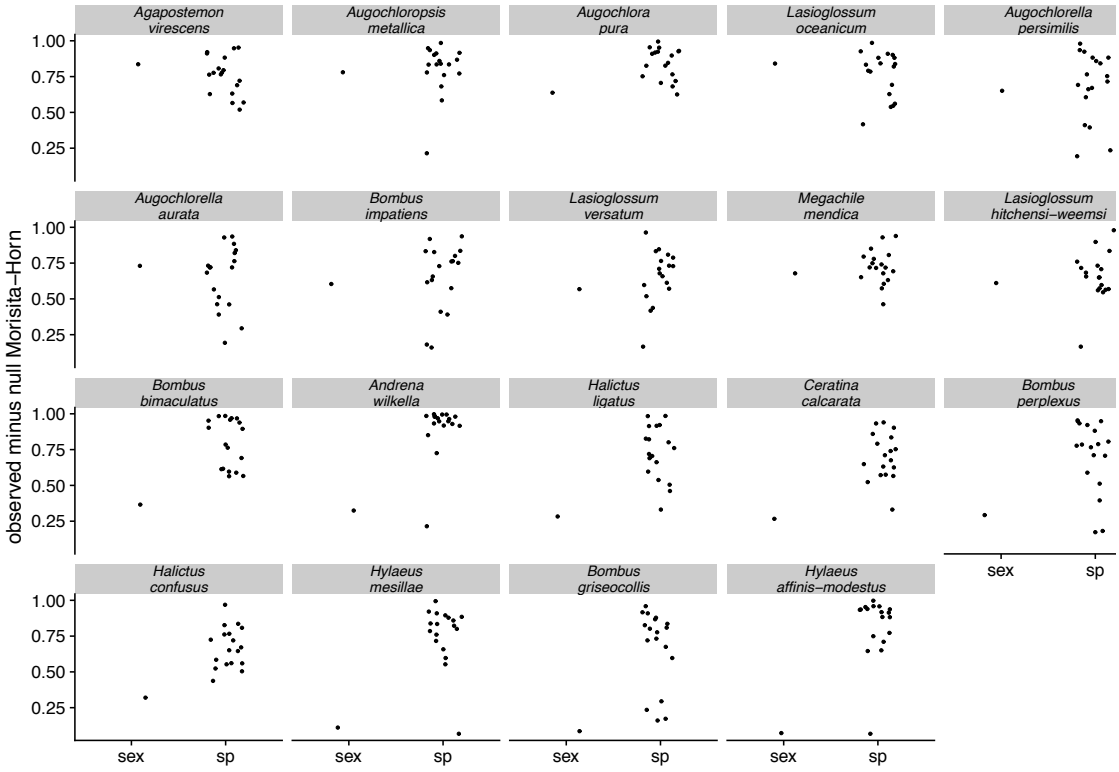


Fig 3. The diets of male and female bees of the same species can be as dissimilar as the diets of females of two different bee species.

Dissimilarities in this figure are the observed statistic minus, for each pairwise comparison, the mean dissimilarity in the null model. Each panel focuses on a bee species (panel name) and shows: above the label “sex”, observed diet dissimilarity between male and female bees of the focal species, minus the average null dissimilarity resulting from randomly permuting the sex identity of each visit record; above the label “sp”, observed diet dissimilarity between female bees of the focal species and each other bee species, minus the average null dissimilarity resulting from randomly permuting the species identity of each visit record.

To what degree are particular flower species disproportionately visited by bees of one sex?

The sex ratio of flower-visiting bees varied across species of flower (Fig 2). After controlling for bee species identity (the strongest predictor of sex in our models, Fig 4), and site, we still found that some flower species received a disproportionate number of male bee visitors (Figs 4 and 5). The median odds ratio for the main effect of flower species was 3.6 (bootstrapped CI 3.0–4.2) in our summed model, indicating that, typically, the visitor sex ratio differs between two flower species by more than a factor of 3. Furthermore, we observed sex-based differences in flower use specific to particular bee species: the median odds ratio for the flower species by bee species interaction in our summed model was nearly as large (median=3.1, bootstrapped CI 3.0–3.3) as the main effect of flower species. By contrast, sex ratios did not differ between sites (median odds ratio for main effect of site=1).

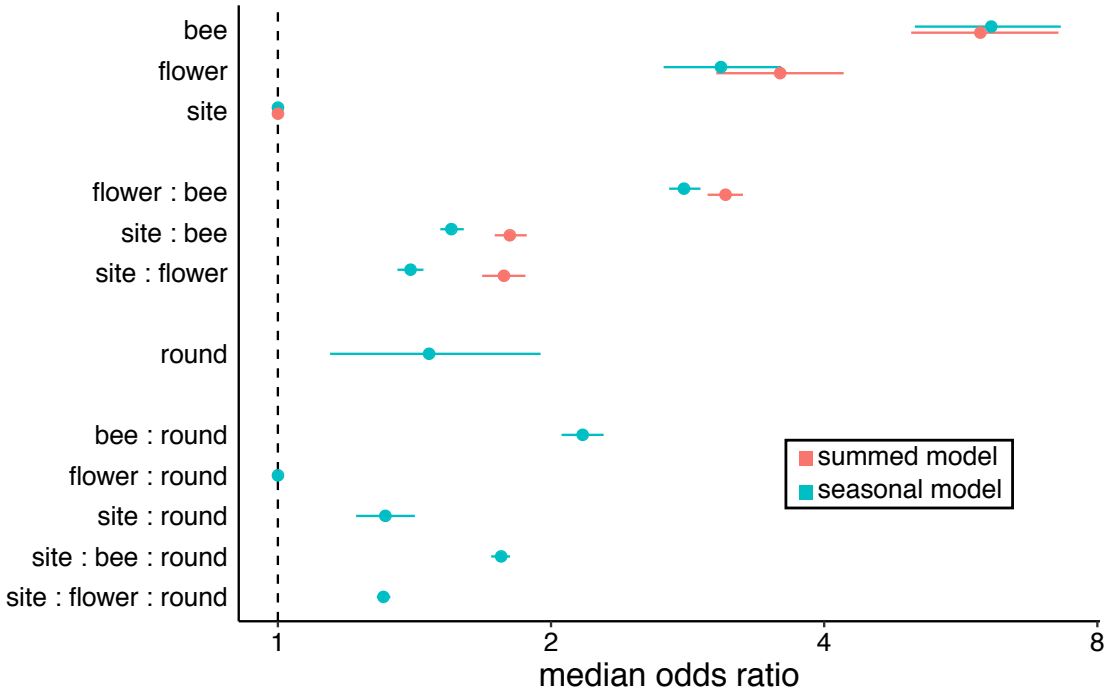


Fig 4. Flower species, along with bee species, predicts the sex of visiting bees, indicating that floral preferences differ between male and female bees.

Flower species is an important predictor of bee sex even after accounting for phenology (seasonal model). For each term ("bee"= bee species, "flower"=flower species, "round"=sampling round) in each model, the median odds ratio (+/- 95% bootstrapped credible interval) indicates the expected difference in odds that a flower-visiting bee is male between two levels. For example, a median odds ratio of 3.7 for the flower species term means the odds of a visitor being male are expected to differ by a factor of 3.7 between two randomly selected species of flower.

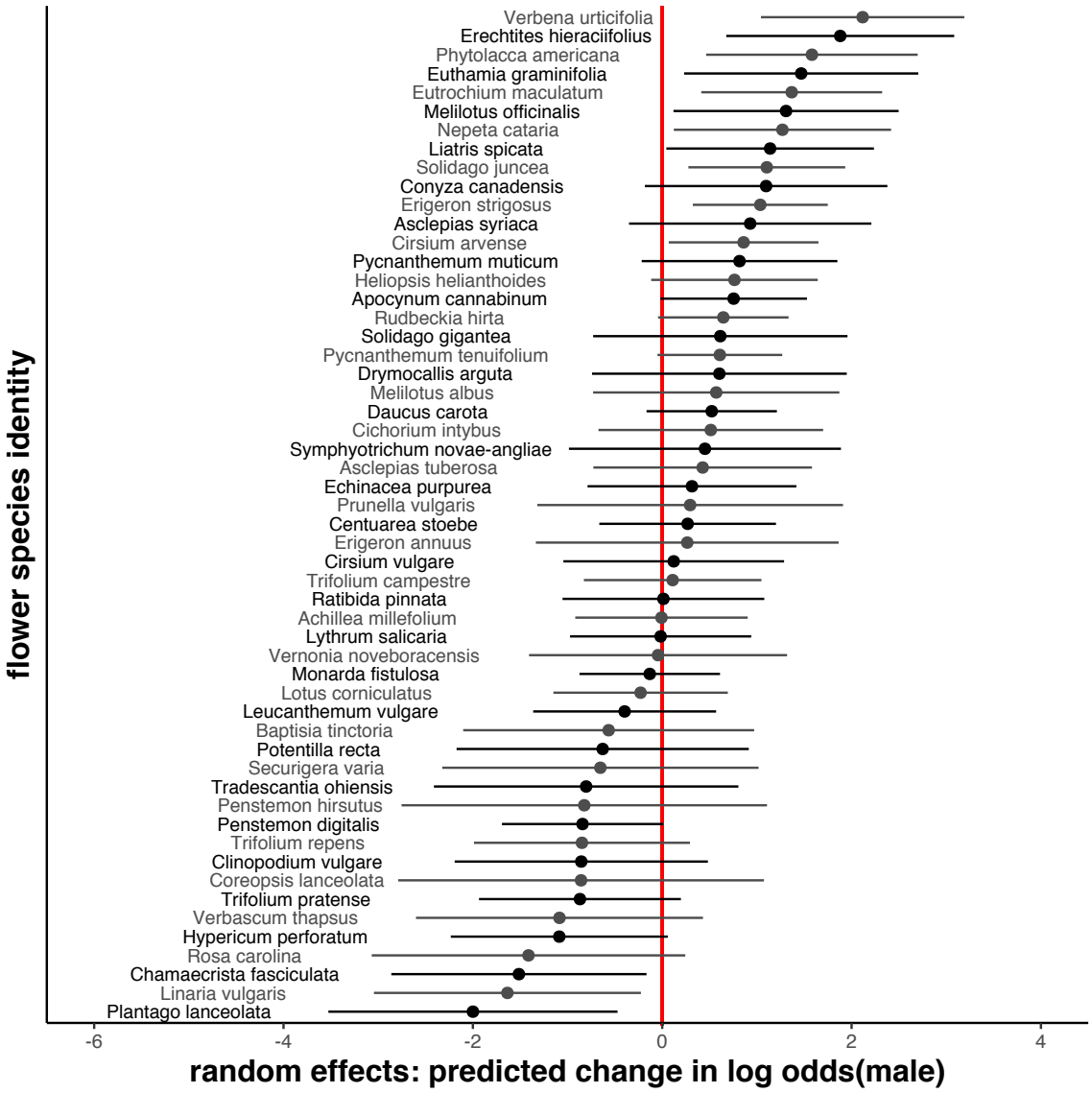


Fig 5. Male bee preferences for and against flower species vary across flower species. Each point is the conditional mode of the random effects prediction (the random-effects analog to an estimate), for a flower species that received at least 20 visits, on the logit scale. Zero represents the odds of a visitor being male on a random flower, and -2 or 2 indicates a ~7 fold decrease or increase in those odds, given flower species identity. Error bars are the square root of the conditional variances on the conditional mode * 1.96, and can be interpreted as the expected range in which the random effect for a particular flower truly lies, analogous to 1.96

times the standard error of the mean for a fixed effect.

To what extent are differences in floral use driven by preference, rather than phenological differences between male and female bees?

The flower species blooming in our system turned over throughout our 11-week sampling period, with several highly visited species blooming for only one of the three months during which we sampled. This turnover, along with potential sex-specific bee flight seasons (e.g. males emerge first in many solitary species, but are not produced until the end of the colony cycle for many social species), means that differences in diet between male and female bees could reflect seasonal availability and use, without also indicating preference differences between the sexes. Indeed, phenology predicts bee sex somewhat, with the odds of a flower-visiting bee being male expected to change by a factor of 1.5 (bootstrapped CI 1.1–1.9) between sampling rounds (Fig 4). Phenological patterns of male vs. female flight seasons vary across bee species; the median odds ratio for the bee species by sampling round interaction is 2.2 (bootstrapped CI 2.1–2.3) (Fig 4). Even after accounting for these effects, however, there remains a strong association between the species of flower a bee visits and its sex (Figs 4–5).

The relative effects of each flower species on the sex of its visitors were changed very little by accounting for phenology; Pearson and Spearman correlations between the random effect of flower species in the seasonal model and the same random effect in the simpler summed model were both 0.98. Each flower species recorded in our study appears in Table C in S1 File, along with the number of male and female bees collected from it and the conditional mode of

the random effect prediction from the seasonal model. In addition to finding overall preference difference between males and females, we found evidence for bee-species-specific difference in floral preferences between the sexes (median odds ratios in both models for the bee species by flower species interaction > 2.8). A post-hoc examination of random effects predictions was consistent with our prediction that male bees avoid flower species that do not produce nectar, with the odds of visiting bees being male nearly twice as high, on average, for flower species that do produce nectar (Figure E in S1 File).

Discussion

We found strong evidence for sexually dimorphic foraging preferences in bees. At the level of individual flower species, we found that commonly, a disproportionate number of bee visitors were male, and that sexually dimorphic preferences drove these patterns. We found that the difference between the flower species visited by male and female bees of the same species was similar in magnitude to differences between females of different species. The partitioning of the floral community among bee species is a primary focus of pollination ecology and ecological network analysis (Bascompte and Jordano 2014), but male bees are typically disregarded or lumped together with their female counterparts. Our study suggests this may represent an important oversight. Further, our study provides strong confirmation of the few studies that investigate the foraging behavior of male bees, which found that males play a unique role in plant pollination (Cane 2002, Pascarella 2010, Cane et al. 2011, Ogilvie and Thomson 2015). Lastly, our result implies that male bees contribute substantially to the complexity of plant-pollinator networks in nature, and that network analyses might benefit

from separating males and females into different nodes (Bolnick et al. 2011, Zwolak 2018).

Phenology, a previously reported mechanism for distinct use of floral resources by male and female bees (Robertson 1925, Ogilvie and Thomson 2015), explained some variation in the sex ratio of flower-visiting bees, but was less important than flower species identity over the period of our study. We expected to find an effect of phenology because both the identity of the flower species blooming within sites, and also the sex ratio of foragers within bee species, vary across the season. Males emerge first in most solitary bees. In contrast, for social species, initial broods usually consist primarily of female workers, then males and reproductive females emerge at the end of the colony cycle (Willmer and Stone 2004). To account for the possibility that phenology explained the disproportionate use of many flower species by one sex of bee, we extended our summed model by adding phenology terms. Surprisingly, phenology only weakly predicted the sex of flower-visiting bees. This is despite the fact that, as predicted by natural history, the sampling round(s) in which males were relatively more prevalent depended on bee species (the bee species by sampling round interaction was much bigger than the sampling round main effect; Fig 4). This indicates that our evidence for floral preference differences between male and female bees was robust to accounting for seasonal turnover in flower species bloom, bee species flight seasons, and the sex ratios within bee species.

Patterns in bee-flower interaction data can arise from the sampling process itself (Blüthgen 2010, Fründ et al. 2016). Our analyses control for these patterns. To evaluate diet overlap, we used a dissimilarity index that downweights rarely used resources, and implemented a null

model that accounts for differences that could arise from sampling effects or the fact that females outnumber males in our dataset by nearly a factor of 5. To evaluate preference, we used random effects models that incorporated all (nearly 19,000) observations, and shrank extreme values for rarely observed species-species interactions towards the global mean for each effect. Thus, our estimates for sex-specific preferences should be robust to the inevitable under-sampling of rarer taxa. Establishing differences in preference between categories of bees such as males and females, even when resource availability is seasonal and difficult to quantify, is possible using methods such as these, though absolute preference remains elusive.

Some studies show, and conservation practice assumes that floral diversity is associated with more bee individuals and diversity, although this pattern could arise from many processes (Roulston and Goodell 2011, De Palma et al. 2015, Senapathi et al. 2016, Spiesman et al. 2017, Sutter et al. 2017). Complementary flower species use between the sexes implies one mechanism by which a bee species could benefit from a diversity of flower choices: a resource used in small proportions at the species level may be crucial for fitness in one sex. Such a dependency would likely be overlooked when individuals of both sexes are pooled before analysis.

Within pollinating insects, sexually dimorphic preferences and contribution to plant reproduction have been reported before (Rusterholtz and Erhardt 2000, Alarcón et al. 2010, Broadhead et al. 2017). Though studies examining foraging differences between male and female bees (Ne'eman et al. 2006, Rundlöf et al. 2014, Ritchie et al. 2016, Ogilvie and Forrest

2017) and pollination by male bees (Cane 2002, Ostevik et al. 2010, Pascarella 2010, Cane et al. 2011, Fliszkiewicz et al. 2011, Wolf and Moritz 2014) have been few outside sexual mimicry and scent collection pollination systems, they found that male and female bees visit different flowers, and that male bees could be important pollinators. Male bees may be especially implicated in long-distance pollen transfer (Janzen 1971, Roubik 1993, Kraus et al. 2009, Wolf and Moritz 2014), although outside the tropics there is little direct evidence this is true.

Mating behaviors of male bees (Barrows 1976a, b, Alcock et al. 1978, Alcock 1983, 2013, Stone 1995, Willmer and Stone 2004, Pinheiro et al. 2017) are better known than foraging behaviors. The two, however, are likely closely linked. Mating-related selection may drive differences in the sensory systems of male and female bees (Streinzer et al. 2013, Somanathan et al. 2017, Brand et al. 2018), or even their approaches to learning (Dötterl et al. 2011, Robert et al. 2016). The mate seeking behaviors of male bees, such as patrolling routes (Barrows 1976a) or seeking flowers visited by conspecific females (Rossi et al. 2010) could generate differences from females via complementarity (males visiting flower species not visited by females), or nestedness (one sex primarily visiting a subset of species visited by the other). We found evidence for both (Figure E in S1 File). Divergent floral preferences between sexes may reflect nutritional needs or mating behavior, but could also reflect visual or olfactory sensitivities that differ between the sexes (Streinzer et al. 2013, Somanathan et al. 2017).

Differences in body size and thermal ecology between male and female bees may also determine foraging behavior (Willmer and Stone 2004). In animals, size dimorphisms often mediate trophic relationships (Selander 1966, Givens 1978, Beck et al. 2007, Temeles et al.

2010, Alonso et al. 2016, Start and De Lisle 2018). Body size often manifests ecologically through thermal constraints, which also drive bee-plant interactions (Heinrich and Heinrich 1983, Chappell 1984, Stone et al. 1999, Rader et al. 2013). In fact, male bees may have preferences relative to females for some flower species based solely on their thermal rewards (Sapir et al. 2006).

Lastly, nutritional rewards likely drive differences in flower species preference between male and female bees. Whereas most female bees collect both nectar and pollen, male bees forage primarily for nectar to fuel flight (Willmer and Stone 2004). Thus, we predicted that male bees would avoid flowers that produce no nectar. Indeed, in both our models, the predicted odds of a bee visiting a nectar-less flower species being male were approximately half that of a bee visiting flower species that produce nectar (Figure E in S1 File). Within flower species that produce pollen and nectar, we found large variation in the relative preferences of male and female bees. Further investigation could reveal which floral traits mediate these sex-specific flower preferences and visitation rates in bees.

Scaling up, it is currently unknown how the distinct foraging niches of male bees mediate either the robustness of pollinator communities to species loss and environmental perturbations (Ramos-Jiliberto et al. 2012, Brosi and Briggs 2013, Tur et al. 2014), or the effectiveness of different habitat ameliorations (Rusterholtz and Erhardt 2000, Rundlöf et al. 2014, Williams and Lonsdorf 2018). This study suggests that both questions warrant further investigation.

Acknowledgements

We thank Daniel Cariveau and Neal Williams for study design suggestions, and Jason Gibbs, Tina Harrison, Dylan Simpson, and three anonymous reviewers for comments on an earlier draft. Cameron Kanterman and Riva Letchinger helped collect field data; Kurtis Himmler, Tiffany Bennet, and professional taxonomists Jason Gibbs, Joel Gardiner, and Sam Droege helped with bee species determinations. Mercer County, the Institute for Advanced Study, Somerset County, and the Raritan Headwaters Association provided site permission.

Supporting information

S1 File. Sampling scheme. (a) The six study sites in central New Jersey, USA. (b) Schematic sampling diagram (not to scale). One observer walked parallel 2m transects covering the entire sampling area. Each 30-minute sampling bout resumed where the previous one left off; observers typically covered the entire meadow once over a 3-day sampling round. (c) The southwestern-most site in peak bloom **(Figure A)**. Schematic cartoon of our simulation for the dissimilarity values associated with our null hypothesis that diets of male and female bees do not differ. (a) Each collection record for each bee species associates the sex of an individual bee to the flower species from which it was collected. (b) To compute the dissimilarity between males and females, we compare all visits to each flower species from males (purple vector) to all visits to each flower species from females (green vector). (c) The Morisita-Horn index summarizes the differences between the two vectors as a value between 0 (identical) and 1 (maximally dissimilar). (d) For our null model, we shuffle the sex column from our observation table. (e) This produces two null vectors. The row and column sums for the matrices in (b) and (c) are identical, but the elements can differ. (f) For our null model, we compute the dissimilarity between the null vectors. We repeated steps d-f 9999 times to generate confidence intervals for the null hypothesis that the sex of a visiting bee is unrelated to the flower species it is collected from. When comparing the flower species visited by different species of bee, we conducted an analysis identical except that rather than comparing two sexes of the same species, we compared two species of the same sex (i.e. exchanging "sex" and "species" throughout Figure A in S1 File) **(Figure B)**. Effect size for diet dissimilarity is independent of sample size, while standardized effect is strongly driven by the number of individuals of the sex with the fewest records. a) Observed Morisita-Horn dissimilarity in

flower communities visited by male and female bees of a single species, minus average null dissimilarity vs. the number of records for the less frequently observed sex. b) Observed minus null dissimilarity in composition of flowers visited by male and female bees of a single species, scaled by the variation in the null model, versus the number of records for the less frequently observed sex **(Figure C)**. Binned residual plots for each model show minor violation of the additivity assumption. Residuals and predicted values on the probability scale **(Figure D)**. Seasonal model predictions are consistent with the hypothesis that male bees avoid flower species that do not produce nectar, relative to females. Each point is the random effect prediction (change in odds that a bee visiting that flower is male) for a flower species. Boxplots show the 25th, 50th, and 75th percentiles, with whiskers extending to more extreme values within 1.5x the interquartile range **(Figure E)**. Methods for post-hoc analysis of male avoidance of nectar-free flowers **(Appendix A)**. Model convergence confirmed based on similar parameter estimates across fitting routines. For each model, the estimate for each term is given for each of 6 fitting algorithms in the R package lme4. Subsequent analyses used parameter estimates in yellow, in both cases tied for the highest estimated likelihood with other very similar fits **(Table A)**. Bee species with number of female and male specimens collected **(Table B)**. Number of male and female visitors to each plant species, and bias towards attracting male bee visitors. This bias is the random effect prediction from the seasonal model, which indicates the change in log(odds) that a visiting bee is male when the species of flower it visits is given; greater values indicate male bias **(Table C)**.

Supporting Information S1

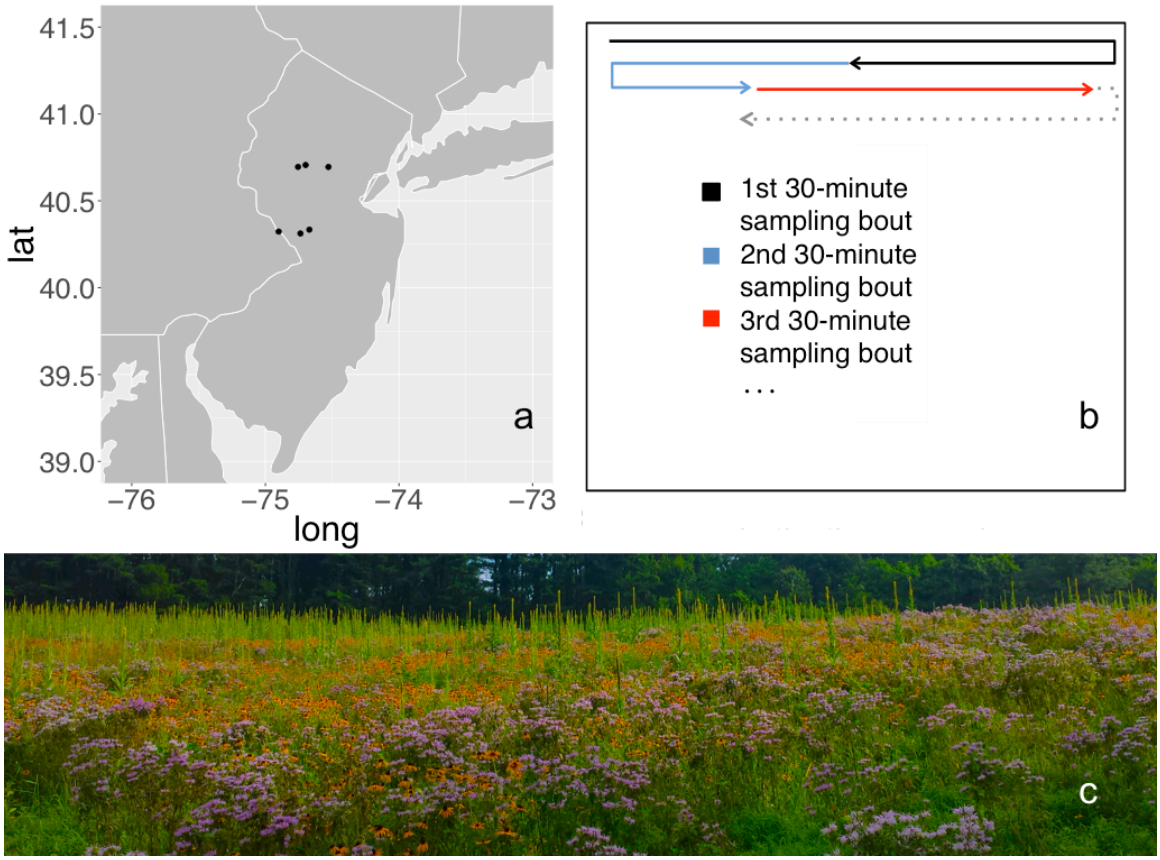


Figure A: Sampling scheme. (a) The six study sites in central New Jersey, USA. (b) Schematic sampling diagram (not to scale). One observer walked parallel 2m transects covering the entire sampling area. Each 30-minute sampling bout resumed where the previous one left off; observers typically covered the entire meadow once over a 3-day sampling round. (c) The southwestern-most site in peak bloom.

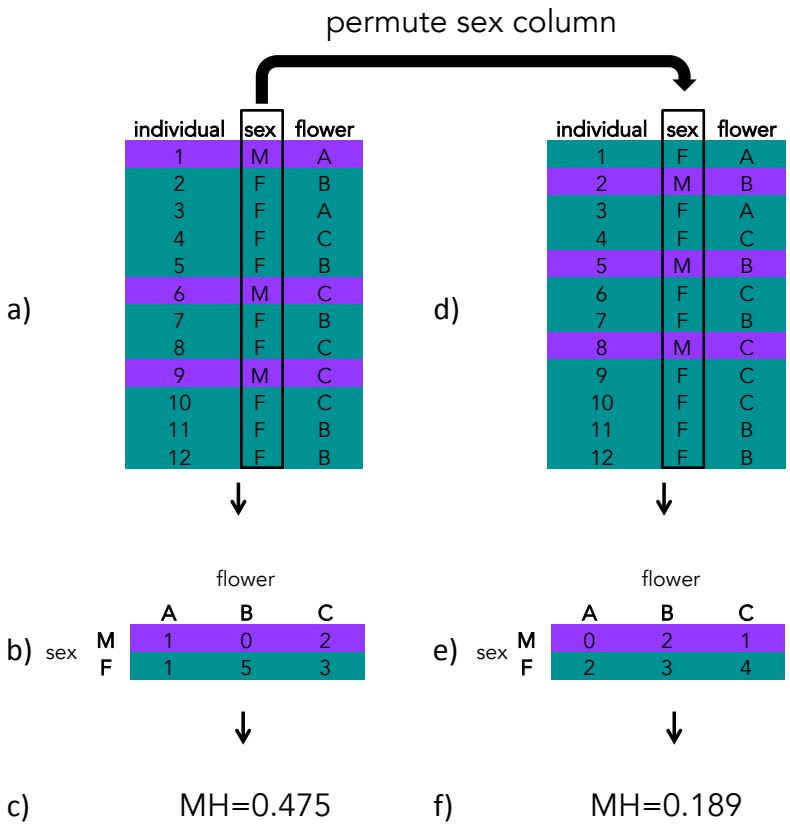


Figure B. Schematic cartoon of our simulation for the dissimilarity values associated with our null hypothesis that diets of male and female bees do not differ. (a) Each collection record for each bee species associates the sex of an individual bee to the flower species from which it was collected. (b) To compute the dissimilarity between males and females, we compare all visits to each flower species from males (purple vector) to all visits to each flower species from females (green vector). (c) The Morisita-Horn index summarizes the differences between the two vectors as a value between 0 (identical) and 1 (maximally dissimilar). (d) For our null model, we shuffle the sex column from our observation table. (e) This produces two null vectors. The row and column sums for the matrices in (b) and (c) are identical, but the

elements can differ. (f) For our null model, we compute the dissimilarity between the null vectors. We repeated steps d-f 9999 times to generate confidence intervals for the null hypothesis that the sex of a visiting bee is unrelated to the flower species it is collected from. When comparing the flower species visited by different species of bee, we conducted an analysis identical except that rather than comparing two sexes of the same species, we compared two species of the same sex (i.e. exchanging “sex” and “species” throughout figure A).

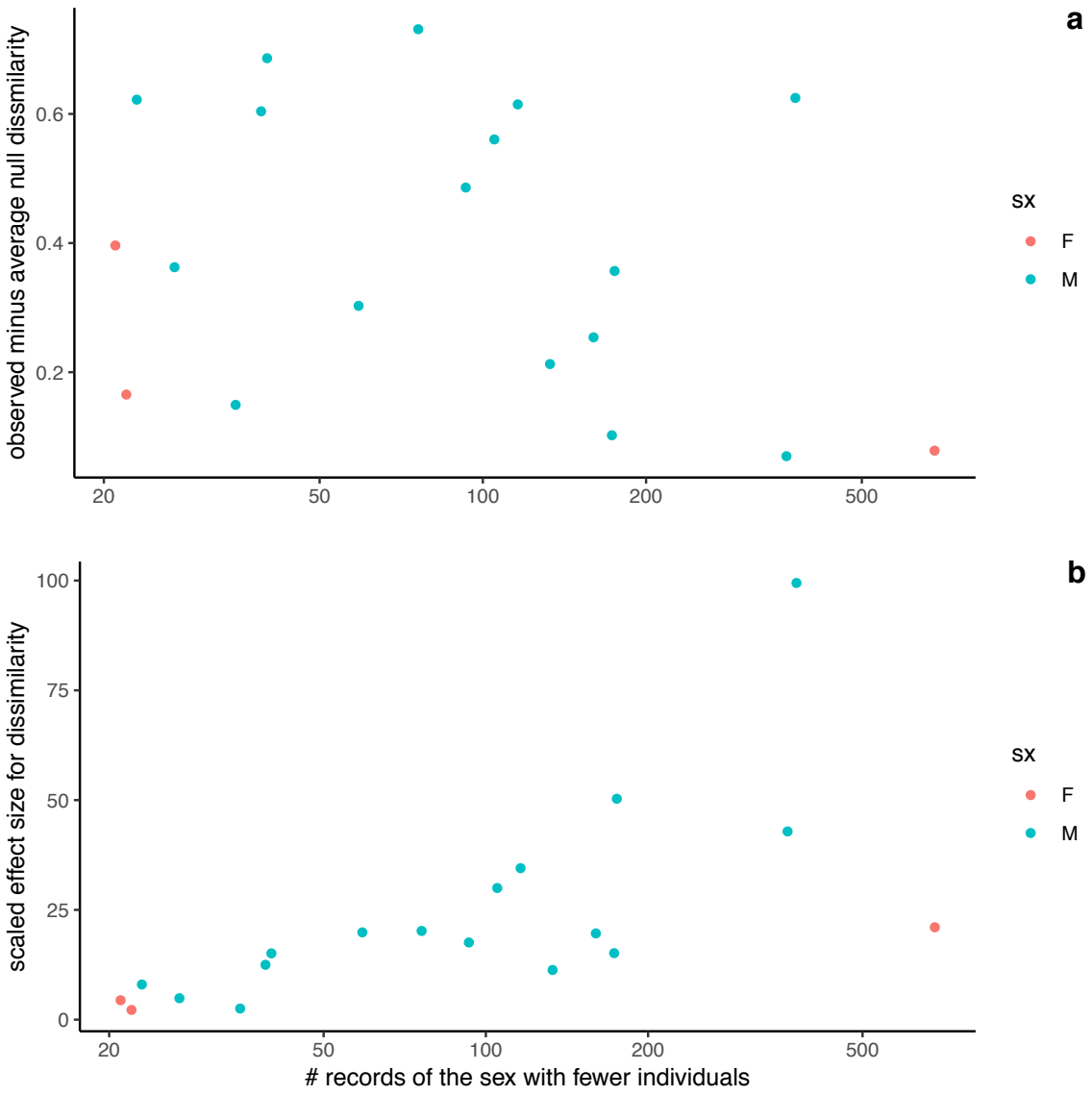


Figure C. Effect size for diet dissimilarity is independent of sample size, while standardized effect is strongly driven by the number of individuals of the sex with the fewest records. a) Observed Morisita-Horn dissimilarity in flower communities visited by male and female bees of a single species, minus average null dissimilarity vs. the number of records for the less frequently observed sex. b) Observed minus null dissimilarity in composition of flowers visited by male and female bees of a single species, scaled by the variation in the null model, versus the number of records for the less frequently observed sex.

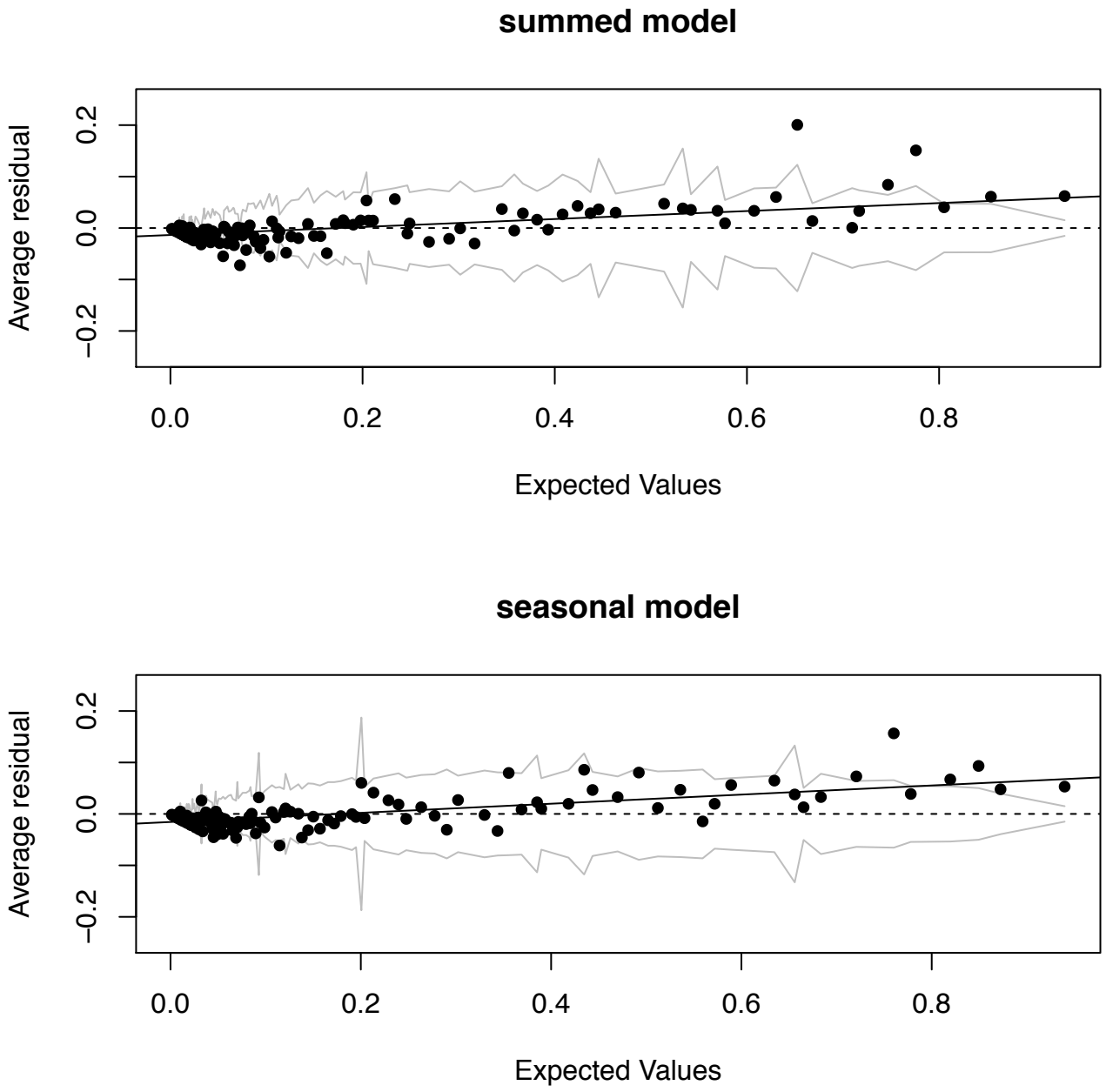


Figure D. Binned residual plots for each model show minor violation of the additivity assumption. Residuals and predicted values on the probability scale.

Appendix A. Methods for post-hoc analysis of male avoidance of nectar-free flowers**.**

Based on previously published work we determined that 11 flower species in our dataset do not produce floral nectar: *Chamaecrista fasciculata* flowers (Rutter and Rausher 2004), *Senna hebecarpa* (Vaudo et al. 2014), *Desmodium paniculatum* (Robertson 1890), *Solanum carolinianum* (Bernardello 2007), *Securigera varia* (Bernardello 2007), *Plantago lanceolata* (Sharma et al. 1993), *Hypericum perforatum* (Willmer 2011), *Hypericum punctatum* (Willmer 2011), *Tradescantia ohioensis* (Vaudo et al. 2016), *Sisyrinchium angustifolium* (Silvério et al. 2012), *Glyceria grandis* (Bernardello 2007), *Sorgahastrum nutans* (Bernardello 2007).

We compared the mean random effects predictions for each of these species from our seasonal model (main text figure 6) with the random effects predictions for nectar-producing species. We compared the mean value for each set of random effects predictions with a Welch's t-test. The difference was nearly significant according to this test ($p=0.055$), although the assumption of independence between observations was certainly invalidated by the random effects structure of our model. We report the difference in means as an odds ratio in the text and present boxplots below (Figure S5)

1. Rutter MT, Rausher MD. Natural selection on extrafloral nectar production in Chamaecrista fasciculata: the costs and benefits of a mutualism trait. Evolution (N Y). 2004;58: 2657–2668.
2. Vaudo AD, Patch HM, Mortensen DA, Grozinger CM, Tooker JF. Bumble bees exhibit daily behavioral patterns in pollen foraging. Arthropod Plant Interact. 2014;8: 273–283. doi:10.1007/s11829-014-9312-5
3. Robertson C. Flowers and Insects IV. Bot Gaz. 1890;15: 79–84.
4. Bernardello G. A systematic survey of floral nectaries. In: Nicolson SW, Nepi M,

Pacini E, editors. Nectaries and Nectar. Springer; 2007. pp. 19–128.

5. Sharma N, Koul P, Koul AK. Pollination biology of some species of genus Plantago L. Bot J Linn Soc. 1993;111: 129–138.
6. Willmer P. Pollination and floral ecology. Princeton: Princeton University Press; 2011.
7. Vaudo AD, Patch HM, Mortensen DA, Tooker JF, Grozinger CM. Macronutrient ratios in pollen shape bumble bee (Bombus impatiens) foraging strategies and floral preferences. Proc Natl Acad Sci. 2016;113: E4035–E4042. doi:10.1073/pnas.1606101113
8. Silvério A, Nadot S, Souza-Chies TT, Chauveau O. Floral rewards in the tribe Sisyrinchieae (Iridaceae): Oil as an alternative to pollen and nectar? Sex Plant Reprod. 2012;25: 267–279. doi:10.1007/s00497-0d-dispersing animals matters for plants. - Biol. Rev. 93: 897–913.

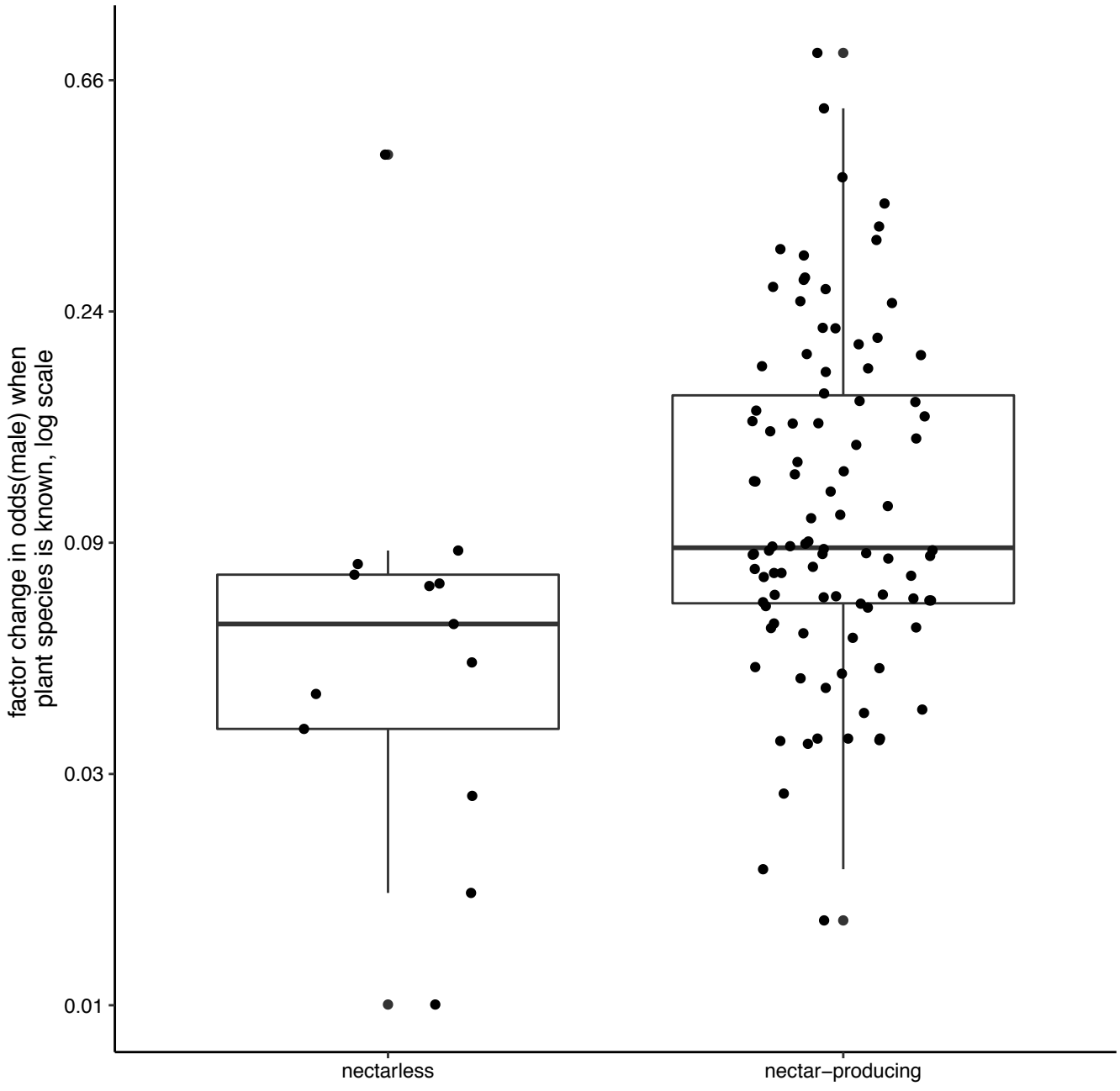


Figure E. Seasonal model predictions are consistent with the hypothesis that male bees avoid flower species that do not produce nectar, relative to females. Each point is the random effect prediction (change in odds that a bee visiting that flower is male) for a flower species. Boxplots show the 25th, 50th, and 75th percentiles, with whiskers extending to more extreme values within 1.5x the interquartile range.

Table A. Model convergence confirmed based on similar parameter estimates across fitting routines. For each model, the estimate for each term is given for each of 6 fitting algorithms in the R package lme4. Subsequent analyses used parameter estimates in yellow, in both cases tied for the highest estimated likelihood with other very similar fits.

term	**model**	bobyqa	Nelder_Mead	nlminbw	optimx.L-BFGS-B	nloptwrap.NLOPT_LN_NELDERMEAD	nloptwrap.NLOPT_LN_BOBYQA
intercept	summed	-2.43	-2.43	-2.43	-2.43	-2.43	-2.43
bee species	summed	2.04	2.04	2.04	2.04	2.04	2.04
flower species	summed	1.40	1.40	1.40	1.40	1.40	1.40
site	summed	0.00	0.00	0.00	0.00	0.00	0.00
bee species:flower species	summed	1.21	1.21	1.21	1.21	1.21	1.21
site:bee species	summed	0.62	0.62	0.62	0.62	0.62	0.62
site:flower species	summed	0.61	0.61	0.61	0.61	0.61	0.61
intercept	seasonal	-2.38	-2.45	-2.38	-2.38	-2.45	-2.45
bee species	seasonal	2.09	2.14	2.09	2.09	2.13	2.13
flower species	seasonal	1.25	1.27	1.25	1.25	1.27	1.27
site	seasonal	0.00	0.00	0.00	0.00	0.00	0.00

bee species:flower species	seasonal	1.09	1.10	1.09	1.09	1.10	1.10
site:bee species	seasonal	0.46	0.47	0.46	0.46	0.47	0.47
site:flower species	seasonal	0.35	0.35	0.35	0.35	0.35	0.35
sampling round	seasonal	0.38	0.36	0.38	0.38	0.36	0.36
sampling round:bee species	seasonal	0.83	0.84	0.83	0.83	0.84	0.84
sampling round:flower species	seasonal	0.00	0.00	0.00	0.00	0.00	0.00
sampling round:site	seasonal	0.29	0.29	0.29	0.29	0.29	0.29
sampling round:site:bee species	seasonal	0.60	0.60	0.60	0.60	0.60	0.60
sampling round:site:flower species	seasonal	0.28	0.29	0.28	0.28	0.28	0.28

Table B. Bee species with number of female and male specimens collected.

family	genus	species	females	males
Andrenidae	*Andrena*	*brevipalpis*	1	0
Andrenidae	*Andrena*	*carlini*	3	0
Andrenidae	*Andrena*	*commoda*	3	0
Andrenidae	*Andrena*	*cressonii*	16	0
Andrenidae	*Andrena*	*fragilis*	2	0
Andrenidae	*Andrena*	*hippotes*	4	0
Andrenidae	*Andrena*	*imitatrix*	6	0
Andrenidae	*Andrena*	*krigiana*	14	0
Andrenidae	*Andrena*	*nasonii*	13	0
Andrenidae	*Andrena*	*nuda*	2	0
Andrenidae	*Andrena*	*pruni*	6	0
Andrenidae	*Andrena*	*robertsonii*	8	0
Andrenidae	*Andrena*	*rudbeckiae*	8	11
Andrenidae	*Andrena*	*rugosa*	1	0
Andrenidae	*Andrena*	*spiraeana*	1	0
Andrenidae	*Andrena*	*vicina*	6	0
Andrenidae	*Andrena*	*wilkella*	277	59
Andrenidae	*Andrena*	*wilmattae*	2	0
Andrenidae	*Calliopsis*	*andreniformis*	4	1
Apidae	*Anthophora*	*abrupta*	4	0
Apidae	*Anthophora*	*terminalis*	3	2
Apidae	*Bombus*	*auricomus*	1	0
Apidae	*Bombus*	*bimaculatus*	577	175
Apidae	*Bombus*	*citrinus*	0	5
Apidae	*Bombus*	*fervidus*	18	0
Apidae	*Bombus*	*griseocollis*	681	815
Apidae	*Bombus*	*impatiens*	2358	105
Apidae	*Bombus*	*perplexus*	22	36
Apidae	*Bombus*	*vagans*	14	2
Apidae	*Ceratina*	*calcarata*	1417	133
Apidae	*Ceratina*	*dupla*	151	19
Apidae	*Ceratina*	*mikmaqi*	130	5
Apidae	*Ceratina*	*strenua*	285	13

Apidae	*Melissodes*	*agilis*	0	7
Apidae	*Melissodes*	*bimaculatus*	9	1
Apidae	*Melissodes*	*denticulatus*	7	73
Apidae	*Melissodes*	*desponsus*	1	7
Apidae	*Melissodes*	*subillatus*	31	6
Apidae	*Melissodes*	*trinodis*	1	7
Apidae	*Nomada*	*articulata*	4	0
Apidae	*Nomada*	*bidentate_gr*	8	0
Apidae	*Nomada*	*erigeronis*	1	0
Apidae	*Nomada*	*lehighensis*	1	0
Apidae	*Nomada*	*maculata*	2	0
Apidae	*Nomada*	*pygmaea*	15	0
Apidae	*Ptilothrix*	*bombiformis*	0	1
Apidae	*Triepeolus*	*cressonii*	0	1
Apidae	*Triepeolus*	*eliseae*	1	0
Apidae	*Triepeolus*	*remigatus*	1	0
Apidae	*Xylocopa*	*virginica*	137	13
Colletidae	*Hylaeus*	*affinis_modestus*	1376	363
Colletidae	*Hylaeus*	*fedorica*	1	0
Colletidae	*Hylaeus*	*leptocephalus*	1	3
Colletidae	*Hylaeus*	*mesillae*	575	173
Halictidae	*Agapostemon*	*sericeus*	5	5
Halictidae	*Agapostemon*	*virescens*	203	76
Halictidae	*Augochlora*	*pura*	1036	377
Halictidae	*Augochlorella*	*aurata*	397	39
Halictidae	*Augochlorella*	*persimilis*	434	116
Halictidae	*Augochloropsis*	*metallica*	121	40
Halictidae	*Dufourea*	*novaeangliae*	0	1
Halictidae	*Halictus*	*confusus*	174	35
Halictidae	*Halictus*	*ligatus*	2432	160
Halictidae	*Halictus*	*parallelus*	6	18
Halictidae	*Halictus*	*rubicundus*	31	19
Halictidae	*Lasioglossum*	*abanci*	6	0
Halictidae	*Lasioglossum*	*admirandum*	15	0
Halictidae	*Lasioglossum*	*anomalum*	17	0
Halictidae	*Lasioglossum*	*atwoodi*	7	1
Halictidae	*Lasioglossum*	*birkmanni*	1	0
Halictidae	*Lasioglossum*	*bruneri*	6	4
Halictidae	*Lasioglossum*	*callidum*	54	0
Halictidae	*Lasioglossum*	*cattellae*	14	4
Halictidae	*Lasioglossum*	*coeruleum*	2	0

Halictidae	*Lasioglossum*	*coreopsis*	1	0
Halictidae	*Lasioglossum*	*coriaceum*	14	0
Halictidae	*Lasioglossum*	*cressonii*	16	5
Halictidae	*Lasioglossum*	*ellisiae*	0	3
Halictidae	*Lasioglossum*	*ephialtum*	1	0
Halictidae	*Lasioglossum*	*foxii*	2	2
Halictidae	*Lasioglossum*	*fuscipenne*	9	0
Halictidae	*Lasioglossum*	*gotham*	74	2
Halictidae	*Lasioglossum*	*hitchensi_weemsi*	152	27
Halictidae	*Lasioglossum*	*illinoense*	70	7
Halictidae	*Lasioglossum*	*imitatum*	462	15
Halictidae	*Lasioglossum*	*leucocomum*	2	0
Halictidae	*Lasioglossum*	*leucozonium*	2	0
Halictidae	*Lasioglossum*	*nigroviride*	2	0
Halictidae	*Lasioglossum*	*oblongum*	4	2
Halictidae	*Lasioglossum*	*obscurum*	7	1
Halictidae	*Lasioglossum*	*oceanicum*	104	23
Halictidae	*Lasioglossum*	*oenotherae*	1	0
Halictidae	*Lasioglossum*	*paradmirandum*	50	0
Halictidae	*Lasioglossum*	*pectorale*	3	0
Halictidae	*Lasioglossum*	*pilosum*	2	0
Halictidae	*Lasioglossum*	*platyparium*	2	3
Halictidae	*Lasioglossum*	*rozeni*	15	11
Halictidae	*Lasioglossum*	*smilacinae*	4	0
Halictidae	*Lasioglossum*	*subviridatum*	5	1
Halictidae	*Lasioglossum*	*tegulare*	31	2
Halictidae	*Lasioglossum*	*trigeminum*	44	0
Halictidae	*Lasioglossum*	*truncatum*	2	0
Halictidae	*Lasioglossum*	*versatum*	681	93
Halictidae	*Lasioglossum*	*viridatum*	11	2
Halictidae	*Lasioglossum*	*zephyrum*	12	1
Halictidae	*Sphecodes*	*atlantis*	0	1
Halictidae	*Sphecodes*	*dichrous*	3	5
Halictidae	*Sphecodes*	*heraclei*	10	5
Megachilidae	*Anthidiellum*	*notatum*	4	1
Megachilidae	*Anthidium*	*manicatum*	7	8
Megachilidae	*Anthidium*	*oblongatum*	18	19
Megachilidae	*Coelioxys*	*alternatus*	1	2
Megachilidae	*Coelioxys*	*banksi*	1	0
Megachilidae	*Coelioxys*	*germanus*	0	1
Megachilidae	*Coelioxys*	*hunteri*	0	1

Megachilidae	*Coelioxys*	*modestus*	0	1
Megachilidae	*Coelioxys*	*obtusiventris*	1	0
Megachilidae	*Coelioxys*	*octodentatus*	1	1
Megachilidae	*Coelioxys*	*porterae*	0	1
Megachilidae	*Coelioxys*	*sayi*	2	6
Megachilidae	*Heriades*	*carinatus*	31	2
Megachilidae	*Heriades*	*leavitti*	1	6
Megachilidae	*Heriades*	*variolosus*	10	0
Megachilidae	*Hoplitis*	*pilosifrons*	46	1
Megachilidae	*Hoplitis*	*producta*	8	0
Megachilidae	*Hoplitis*	*spoliata*	2	1
Megachilidae	*Lithurgus*	*chrysurus*	0	6
Megachilidae	*Megachile*	*brevis*	25	3
Megachilidae	*Megachile*	*campanulae*	6	18
Megachilidae	*Megachile*	*exilis*	11	29
Megachilidae	*Megachile*	*frugalis*	26	6
Megachilidae	*Megachile*	*gemula*	4	2
Megachilidae	*Megachile*	*georgica*	1	0
Megachilidae	*Megachile*	*inimica*	4	0
Megachilidae	*Megachile*	*integra*	1	0
Megachilidae	*Megachile*	*melanophaea*	0	1
Megachilidae	*Megachile*	*mendica*	22	56
Megachilidae	*Megachile*	*montivaga*	15	9
Megachilidae	*Megachile*	*petulans*	0	2
Megachilidae	*Megachile*	*pugnata*	2	3
Megachilidae	*Megachile*	*rotundata*	11	8
Megachilidae	*Megachile*	*sculpturalis*	17	32
Megachilidae	*Megachile*	*xylocopoides*	2	1
Megachilidae	*Osmia*	*albiventris*	3	0
Megachilidae	*Osmia*	*atriventris*	9	0
Megachilidae	*Osmia*	*bucephala*	21	0
Megachilidae	*Osmia*	*distincta*	7	0
Megachilidae	*Osmia*	*georgica*	5	0
Megachilidae	*Osmia*	*pumila*	30	0
Megachilidae	*Pseudoanthidium*	*nanum*	0	1
Megachilidae	*Stelis*	*lateralis*	1	0
Megachilidae	*Stelis*	*louisae*	1	2

Table C. Number of male and female visitors to each plant species, and bias towards attracting male bee visitors. This bias is the random effect prediction from the seasonal

model, which indicates the change in log(odds) that a visiting bee is male when the species of flower it visits is given; greater values indicate male bias.

family	genus	species	female visits	male visits	random effect
Verbenaceae	*Verbena*	*urticifolia*	58	71	2.117
Asteraceae	*Erechtites*	*hieraciifolius*	130	203	1.880
Fabaceae	*Senna*	*hebecarpa*	5	5	1.680
Phytolaccaceae	*Phytolacca*	*americana*	108	74	1.581
Asteraceae	*Euthamia*	*graminifolia*	50	18	1.468
Asteraceae	*Eutrochium*	*maculatum*	461	166	1.367
Fabaceae	*Melilotus*	*officinalis*	41	21	1.308
Lamiaceae	*Nepeta*	*cataria*	99	81	1.270
Lamiaceae	*Monarda*	*punctata*	0	1	1.243
Campanulaceae	*Lobelia*	*inflata*	12	3	1.146
Asteraceae	*Liatris*	*spicata*	186	128	1.140
Asteraceae	*Solidago*	*juncea*	636	77	1.104
Asteraceae	*Conyza*	*canadensis*	32	10	1.097
Polygonaceae	*Fallopia*	*convolvulus*	3	4	1.046
Asteraceae	*Erigeron*	*strigosus*	712	119	1.036
Asclepidaceae	*Asclepias*	*syriaca*	28	89	0.929
Verbenaceae	*Verbena*	*hastata*	8	3	0.925
Lamiaceae	*Pycnanthemum*	*verticillatum*	5	8	0.886
Asteraceae	*Cirsium*	*arvense*	351	96	0.859
Lamiaceae	*Pycnanthemum*	*muticum*	398	60	0.817
Asteraceae	*Solidago*	*canadensis*	8	2	0.816
Asteraceae	*Heliopsis*	*helianthoides*	186	49	0.763
Apocynaceae	*Apocynum*	*cannabinum*	283	92	0.754
Fabaceae	*Trifolium*	*hybridum*	11	7	0.742
Asteraceae	*Rudbeckia*	*hirta*	1174	189	0.645
Asteraceae	*Solidago*	*gigantea*	107	9	0.613
Lamiaceae	*Pycnanthemum*	*tenuifolium*	1113	421	0.608
Rosaceae	*Drymocallis*	*arguta*	22	8	0.604
Fabaceae	*Melilotus*	*albus*	20	9	0.570
Cornaceae	*Swida*	*racemosa*	12	1	0.544
Apiaceae	*Daucus*	*carota*	1783	350	0.523
Asteraceae	*Helianthus*	*strumosus*	1	1	0.518
Asteraceae	*Cichorium*	*intybus*	104	10	0.513
Verbenaceae	*Verbena*	*simplex*	11	3	0.482
Asteraceae	*Symphyotrichum*	*novae-angliae*	25	6	0.451
Asclepidaceae	*Asclepias*	*tuberosa*	114	26	0.427

Asteraceae	*Cirsium*	*discolor*	5	2	0.350
Asteraceae	*Echinacea*	*purpurea*	86	45	0.314
Lamiaceae	*Prunella*	*vulgaris*	17	5	0.295
Asteraceae	*Centuarea*	*stoebe*	321	50	0.268
Asteraceae	*Erigeron*	*annuus*	26	3	0.264
Onagraceae	*Oenothera*	*fruticosa*	2	1	0.220
Polygonaceae	*Persicaria*	*setacea*	3	1	0.158
Asteraceae	*Cirsium*	*vulgare*	112	16	0.121
Fabaceae	*Trifolium*	*campestre*	365	39	0.110
Asteraceae	*Ratibida*	*pinnata*	539	121	0.012
Asteraceae	*Achillea*	*millefolium*	473	36	-0.008
Asteraceae	*Bidens*	*trichosperma*	1	0	-0.015
Asteraceae	*Solidago*	*rugosa*	1	0	-0.015
Lythraceae	*Lythrum*	*salicaria*	364	38	-0.017
Rosaceae	*Rubus*	*flagellaris*	1	0	-0.028
Fabaceae	*Trifolium*	*aureum*	1	0	-0.033
Campanulaceae	*Lobelia*	*siphilitica*	2	0	-0.034
Gentianaceae	*Sabatia*	*angularis*	1	0	-0.034
Asteraceae	*Vernonia*	*noveboracensis*	52	37	-0.044
Fabaceae	*Vicia*	*tetrasperma*	1	0	-0.048
Apiaceae	*Sanicula*	*canadensis*	1	0	-0.049
Asteraceae	*Doellingeria*	*umbellata*	1	0	-0.052
Ranunculaceae	*Ranunculus*	*hispidus*	1	0	-0.057
Asteraceae	*Coreopsis*	*tinctoria*	1	0	-0.068
Poaceae	*Sorghastrum*	*nutans*	1	0	-0.092
Lamiaceae	*Teucrium*	*canadense*	3	0	-0.104
Brassicaceae	*Barbarea*	*vulgaris*	3	0	-0.115
Loniceraceae	*Lonicera*	*japonica*	1	0	-0.129
Lamiaceae	*Monarda*	*fistulosa*	1398	401	-0.132
Fabaceae	*Desmodium*	*paniculatum*	6	1	-0.137
Onagraceae	*Oenothera*	*biennis*	2	0	-0.143
Asteraceae	*Hieracium*	*pilosella*	2	0	-0.149
Hypericaceae	*Hypericum*	*punctatum*	1	0	-0.177
Poaceae	*Glyceria*	*grandis*	1	0	-0.187
Alliaceae	*Allium*	*vineale*	3	0	-0.225
Apiaceae	*Eryngium*	*yuccifolium*	2	0	-0.226
Fabaceae	*Lotus*	*corniculatus*	142	33	-0.228
Asteraceae	*Crepis*	*capillaris*	6	0	-0.236
Rosaceae	*Rubus*	*pensilvanicus*	7	0	-0.242
Cornaceae	*Swida*	*amomum*	4	0	-0.250
Oxalidaceae	*Oxalis*	*stricta*	8	0	-0.250

Asteraceae	*Lactuca*	*serriola*	4	0	-0.257
Rosaceae	*Rosa*	*multiflora*	5	0	-0.264
Asteraceae	*Gaillardia*	*aristata*	4	0	-0.275
Caryophyllaceae	*Dianthus*	*armeria*	3	0	-0.281
Iridaceae	*Sisyrinchium*	*angustifolium*	9	0	-0.353
Asteraceae	*Carduus*	*nutans*	1	0	-0.367
Rubiaceae	*Galium*	*mollugo*	4	0	-0.369
Polygonaceae	*Persicaria*	*pensylvanica*	4	0	-0.392
Asteraceae	*Leucanthemum*	*vulgare*	406	20	-0.397
Asteraceae	*Helianthus*	*angustifolius*	11	0	-0.412
Solanaceae	*Solanum*	*carolinense*	14	0	-0.517
Convulvulaceae	*Calystegia*	*silvatica*	5	0	-0.539
Asteraceae	*Krigia*	*biflora*	19	0	-0.544
Fabaceae	*Baptisia*	*tinctoria*	19	5	-0.566
Asteraceae	*Solidago*	*altissima*	8	0	-0.587
Rosaceae	*Potentilla*	*recta*	56	1	-0.629
Fabaceae	*Securigera*	*varia*	38	1	-0.653
Scrophulariaceae	*Verbascum*	*blattaria*	15	0	-0.722
Asclepidaceae	*Asclepias*	*incarnata*	7	0	-0.737
Commelinaceae	*Tradescantia*	*ohiensis*	34	1	-0.804
Scrophulariaceae	*Penstemon*	*hirsutus*	36	0	-0.823
Scrophulariaceae	*Penstemon*	*digitalis*	862	48	-0.842
Fabaceae	*Trifolium*	*repens*	130	6	-0.848
Lamiaceae	*Clinopodium*	*vulgare*	64	3	-0.855
Asteraceae	*Coreopsis*	*lanceolata*	21	0	-0.857
Fabaceae	*Trifolium*	*pratense*	192	20	-0.869
Scrophulariaceae	*Verbascum*	*thapsus*	129	1	-1.085
Hypericaceae	*Hypericum*	*perforatum*	223	10	-1.087
Rosaceae	*Rosa*	*carolina*	71	1	-1.413
Fabaceae	*Chamaecrista*	*fasciculata*	246	2	-1.514
Scrophulariaceae	*Linaria*	*vulgaris*	275	4	-1.635
Plantaginaceae	*Plantago*	*lanceolata*	147	0	-1.999

CHAPTER 2

A conceptual guide to measuring species diversity

Manuscript authors: Michael Roswell, Jonathan Dushoff, Rachael Winfree

Key words: *Hill numbers, rarefaction, coverage, rarity*

ABSTRACT

Three metrics of species diversity — species richness, the Shannon index, and the Simpson index — are still widely used in ecology, despite decades of valid critiques leveled against them. Developing a robust diversity metric has been challenging because, unlike many variables ecologists measure, the diversity of a community often cannot be estimated in an unbiased way based on a random sample from that community. Over the past decade, ecologists have begun to incorporate two important tools for estimating diversity, coverage and Hill diversity. Coverage is a method for equalizing samples that is, on theoretical grounds, preferable to other commonly used methods such as equal-effort sampling, or rarefying datasets to equal sample size. Hill diversity comprises a spectrum of diversity metrics and is based on three key insights. First, species richness and variants of the Shannon and Simpson indices are all special cases of one general equation. Second, richness, Shannon, and Simpson can be expressed on the same scale and in units of species. Third, there is no way to eliminate the effect of relative abundance from estimates of any of these diversity metrics, including species richness. Rather, a researcher must choose the relative sensitivity of the metric towards rare and common species, a concept which we describe as “leverage.” In this paper we explain coverage and Hill diversity, provide guidelines for how to use them together to measure species diversity, and demonstrate their use with examples from our own data. We show why researchers will obtain more robust results when they estimate the Hill diversity of

equal-coverage samples, rather than using other methods such as equal-effort sampling or traditional sample rarefaction.

INTRODUCTION

Species diversity is one of the more frequently measured quantities in ecology, yet *how* to measure it is complex, and sometimes contentious. The past decade has seen great advances in comparing and unifying various diversity metrics, and also in developing ways to standardize samples prior to measuring diversity (Jost 2006, Ellison 2010, Chiarucci et al. 2011, Chao and Jost 2012, Colwell et al. 2012, Chase and Knight 2013, Chao and Chiu 2016, Cox et al. 2017). This latter step is necessary because — in contrast to many variables ecologists measure, for which a random sample from a community provides a reasonably unbiased estimate of the community itself — most species diversity values estimated from samples are a biased measure of the diversity of the larger community. This is mainly because the true relative abundance of rare species is poorly captured in samples, in which those species tend to appear only once or not at all. Here, we provide a conceptual guide to best practices for comparing the level of biodiversity of two or more communities, based on samples from those communities (Fig. 1). We begin by reviewing methods for standardizing samples, which is an important but often overlooked step in measuring diversity. In this section we review "coverage," a conceptually elegant, but under-used, method for standardizing samples. We then provide a guide to using Hill diversity. We try to make this concept more intuitive to ecologists by showing how the different Hill diversities are all calculating the mean rarity of the species in the community, but doing so using different types of means (arithmetic, geometric, and harmonic). We also draw parallels between Hill diversity and the workings of calculations familiar to many ecologists: the link functions of generalized linear models.

Throughout the paper, for simplicity, we assume researchers sample discrete individuals, and discuss statistical matters in terms of individual- but not incidence-based diversity estimation (Colwell et al. 2012, Chao et al. 2014a, Chao and Jost 2015). Throughout, the references we cite typically discuss both approaches. Similarly, when we use the term "relative abundance," we refer to the proportion of individuals belonging to a given species, but in most cases other measures like proportional biomass or percent cover could be used instead. We assume that ecologists wish to determine which communities are more and less diverse, and by how much; in other words, that they aim to measure an "effect size" (Chao and Jost 2012, Chase and Knight 2013). Thus, we advocate for methods that will accurately reflect relative (but not necessarily absolute) differences in diversity. To demonstrate the preferred tools for standardizing samples and quantifying diversity — coverage and Hill diversity — we analyze a small data set on wild bees we collected from four meadows.

EQUALIZING SAMPLES

Diversity can only be meaningfully compared across communities that have been sampled equivalently in some way. Unfortunately, there are multiple ways to standardize samples, and the choice of sample standardization method can strongly influence results. In this section, we consider three main ways ecologists standardize their samples: by equalizing effort, equalizing sample size, or equalizing coverage.

Conceptual problems with traditional methods of equalizing samples

Many ecologists build *equal-effort sampling* into their study designs. Effort can be measured as the amount of time spent sampling, the area sampled, the number of traps set out, or the

like. This seems like the right way to compare communities: sample the same way and the same amount in each, and any differences should reflect only the diversity of each community, and not how the communities were sampled. But this is not true. In reality, two factors determine how well the sample represents the **true diversity** of the community: how many species there are and in what relative abundances, but also how hard one looks for those species.

Equal-effort sampling only deals with the second factor. The problem with equal-effort sampling is that sample size generally varies across communities given equal effort, and sample size partly determines how well the observed abundance distribution matches the true species abundance distribution of the community. For instance, a small sample is likely to contain only a few species, all of them common. As samples contain more individuals, the number of species rises, and **sample diversity** grows (Preston 1948). In sum, diversity estimates (and especially species richness estimates) based on equal-effort sampling underestimate community diversity from samples that contain fewer individuals, because these samples often include fewer species by chance alone, regardless of the community from which they are drawn (Gotelli and Colwell 2001).

A second way ecologists standardize samples is by *sample size*; for example, by removal of individuals from larger samples until all samples have the same number of individuals (**rarefaction**). However, rarefaction doesn't provide unbiased samples either, because it still doesn't account for the distribution of relative abundances in the true, larger community (Willis 2019). Standardizing by sample size doesn't work because more diverse communities

usually have both more and also rarer species, and thus require more effort to characterize. Even for two communities with very different numbers of rare species (and thus different diversities), the number of rare species that 'fit' into a fixed sample size, e.g. 100 individuals, could be similar. Furthermore, the differences in diversity between two samples of the same size changes idiosyncratically as sample size increases. It is not always possible to predict from smaller samples which of two communities would appear more diverse with much larger samples. In sum, sample-size standardization leads to larger underestimates of diversity for more diverse communities (Chao & Jost 2012).

Coverage: a solution

Sample-size and effort-based standardization do not fairly represent community diversity because they do not account for the underlying species abundance distribution of the community being sampled (Brose et al. 2003, Cao et al. 2007, Beck and Schwanghart 2010, Willis 2019). In contrast, a newer method, ***coverage*** (Box 1), accounts for both the amount of sampling and, to a much greater extent than the other methods, the true diversity of the community. Coverage thereby recognizes that *more diverse communities require more sampling in order to be equally well-characterized.* Coverage was discovered in the 1940s by the founder of computer science, Alan Turing, but was only recently introduced as a tool for standardizing samples in ecology (Alroy 2010, 2017, Jost et al. 2010, Chao and Jost 2012). Coverage is a theoretically elegant way to standardize samples, and is increasingly used in the ecological literature.

Coverage describes how well a sample captures the true diversity of the whole community, including species that have not yet been detected. More precisely, coverage estimates the proportion of individuals in the (true, larger) community that are represented by species present in the sample. As this proportion increases, the share of individuals in the true community that belong to undetected species falls. For example, a coverage of 0.98 means that 2% of the individuals in the community being sampled belong to species the researcher has missed. For a sample to contain enough species to represent 98% of a more diverse community, it usually must be larger than a sample with 98% coverage from a less diverse community. Thus, when ecologists standardize samples by coverage, they compare samples that have more individuals (and/ or that required more effort to collect) from some communities than others. This results in more balanced information from each community.

Sampling with equal coverage isn't quite what we might want: to sample each community until the same *proportion of its diversity* had been recorded. For example, if one used species richness as the diversity metric, one could imagine sampling each community until 90% of the species in each community had been detected. In this case, the comparison would be fair. Unfortunately, this method is not possible, because it is not usually possible to know how many species are truly there, nor in what proportions – if it were, we wouldn't need to estimate diversity. Given that this ideal cannot be implemented, coverage is a practical approach to achieving more comparable samples, using information available to researchers.

The key insight behind coverage is that the proportion of *individuals* in the community belonging to undetected species can be estimated reliably, based only on the frequencies of

species already in the sample (Good and Toulmin 1956, Chao and Jost 2012, Zhang 2016). This concept is best illustrated with a species accumulation curve (Box 1, Fig B1(a)). Imagine being at the endpoint of the curve, about to sample one more individual. The pool that individual will be sampled from contains all the as-yet-unsampled individuals in the community, most of which belong to species already detected, but some of which do not. If the next individual you obtain is a new species, the species accumulation curve goes up one step for a slope of 1. If it is not a new species, the curve moves horizontally one step for a slope of 0 (Fig B1(a), arrows). Thus, the slope of the species accumulation curve represents the probability that the next individual sampled will belong to a new species. This slope is (1-coverage). As coverage approaches 1, the species accumulation curve asymptotes.

While ecologists have long used the slope of the species accumulation curve to measure sampling completeness, the advantage of the more recent formalization is that even when sampling is incomplete, samples can be *compared at equal coverage* (Fig. B1(b)). This comparison is 'fair' in the sense that the same proportion of individuals from each community is represented by the species in each sample. In sum, while it cannot remove sample diversity's dependence on sample completeness (Willis 2019), coverage is the fairest available way to standardize samples because it standardizes what is known (the sample) relative to what is there (the true community).

What coverage clarifies about species richness

Ecologists may be attracted to species richness as a diversity metric because true richness depends only the number of species, but not on their relative abundances. However,

estimates of richness are, in fact, highly sensitive to the relative abundances of species in the community being sampled. The concept of coverage offers a nice demonstration of how this is so.

Although sample coverage increases with sampling, the rate of this increase slows as sampling proceeds (Fig. B1(c)). This is because after initial sampling, the vast majority of individuals in a community *do* belong to species represented in the sample, and it takes a lot of work to find those comparatively few individuals belonging to the new, rare species. This means that sample richness depends on *how rare* the rare species are. For example, imagine two communities, one in which all species have the same abundance, and the second in which a few species are very common, but most are very, very rare. In the first community, at low and medium sample sizes, finding a new species with additional sampling remains quite likely. In the second, once samples are large enough that the common species have been detected, the chance of detecting a new (and very rare) species with additional sampling is low. This means that even if both communities had the same richness, samples from the first community would usually contain more species. In sum, species richness estimates are not only sensitive to the size of the sample and the true number of species in the community, but also to *species relative abundances in the community*, just like other diversity indices.

A note on extrapolation

To this point, we have discussed standardizing via rarefaction, but not **extrapolation** –that is, extending the pattern of species detection to a greater sample size, effort, or coverage than already obtained. Inferring what one might have seen with additional sampling is obviously

appealing, and compatible, at least in principle, with sample standardization. The past decade saw the introduction of unified methods for rarefaction and extrapolation for diversity estimation, based not only on sample size or effort, but also on sample coverage (Chao and Jost 2012, Colwell et al. 2012, Chao et al. 2014a).

Standardizing to a level that involves extrapolation for at least some samples could be preferable to analyzing only very incomplete samples for every community. The caveat is, of course, that the farther from the sample an extrapolation extends, the more sensitive it is to the extrapolation method's assumptions. A second issue is that neither empirical nor theoretical work yet guides what level of sample completeness is "good enough" to serve as a target level, though Chao and Jost (2012) suggest that extrapolating to double the observed sample size entails little risk. Because of these complexities and dangers, guidance on extrapolation is beyond the scope of this guide. However, we discuss using asymptotic diversity estimators, which could be considered an extreme form of extrapolation, below (See Asymptotic estimators).

DIVERSITY METRICS

In this section, we briefly review problems with species richness, and the traditional Shannon and Simpson indices, which are the ways ecologists most often measure the diversity of a community (Magurran and McGill 2011). We then explain **Hill diversity**, a general approach that includes, as special cases, species richness and modified versions of the traditional Shannon and Simpson indices. There is an increasing consensus that Hill diversity is the preferred way to measure not only the species diversity of a community, which is the subject

of this paper, but also differentiation among communities (Ellison 2010, Chao and Chiu 2016, Botta-Dukát 2018, Chao et al. 2019, Ohlmann et al. 2019), functional and phylogenetic diversity (Chao et al. 2014b, Kang et al. 2016), and genetic diversity (Sherwin et al. 2017, Alberdi and Gilbert 2019).

To illustrate the main points we make here and elsewhere in the text, we present some analyses of a small data set extracted from a larger study (Roswell et al. 2019a, b). This data set includes wild bees we collected with hand nets from four meadows, using equal effort: 7 person-hours over three consecutive days in each meadow (Fig. 2). In the first meadow (green squares in Fig. 2), we collected 578 individual bees that we identified to 40 bee species. In the second meadow (purple triangles), we collected 442 individuals of 40 species. In the third meadow (orange circles), we collected 745 individuals of 32 species. In the fourth (pink diamonds), we collected 225 individuals of 29 species. The question we seek to answer is, "which bee communities are more and less diverse, and by how much?" We use the data to show how the answer can vary with the amount of data collected, the method chosen to standardize samples, and the diversity metric used.

Conceptual problems with traditional diversity metrics

The number of species in a sample (species richness) is a very flawed measure of diversity. Richness is strongly associated with the number of individuals in the sample, especially at the earlier stages of sampling (Fig. 3). Furthermore, as sampling proceeds, the accumulation curves representing different communities often cross (Lande *et al.* 2000; colored lines in Fig. 3). This means that the relative richness of two communities, as measured at a smaller sample

size, does not predict their relative richness at a larger sample size well (Cao et al. 2007, Coddington et al. 2009, Haegeman et al. 2013). This is often true even when estimators such as Chao1 are used to predict true diversity (colored clouds of points in Fig. 3; see 'Asymptotic estimators' below).

Because richness is so sensitive to sampling effort and relative abundance, its estimation can hinge on how samples are standardized. Even the best asymptotic richness estimators, such as Chao1 (Gotelli and Colwell 2011), cannot reliably predict the true community diversity (Fig. 3). Both sample richness and sample-based richness estimators are strongly influenced by the rarest species, which are precisely the species that we know least about. This is another way of saying that richness has high uncertainty. In fact, in the context of estimating and comparing community diversity from samples, this uncertainty is often insurmountable (Haegeman et al. 2013).

The traditional diversity indices that explicitly include relative abundance (Magurran and McGill 2011), such as the Shannon (Shannon and Weaver 1963) and Simpson (Simpson 1949) indices, are more robust than richness to the sampling problems outlined above. However, their use creates a new set of problems: these indices do not scale intuitively, or similarly, with species gain and loss (Box 2; Jost 2009, Tuomisto 2010). These problems have led to the suggestion that diversity lacks any conceptual grounding (Hurlbert 1971).

Hill diversity: a solution

A unified method for measuring diversity was developed by Hill (1973), and re-introduced to

ecologists by Jost (2006). This method takes as its starting point that both the number and the relative abundance of species are components of diversity, and that these components cannot be fully separated. A community with more species is more diverse. And a community with a given number of species is more diverse when all species are equally abundant (extreme evenness) than when, for example, it has one dominant species and all other species are very rare (extreme dominance). Diversity metrics vary in the degree to which they reflect these two components. The diversity metric developed by Hill (1973) consists of a single equation that, depending on the value taken by its sole parameter, the exponent that we call "ℓ," can vary from counting all species equally, even if they are vanishingly rare, to heavily emphasizing the species that are most common (Box 3).

Hill diversity has several important advantages. First, Hill diversities behave in ways that are logically reasonable for a measure of diversity (Hurlbert 1971, Jost 2009, Tuomisto 2010, Chao et al. 2014b). For example, if some proportion of a community's species were randomly removed, all Hill diversities decrease by the same proportion. Traditional diversity indices fail this and other common-sense expectations.

But how do Hill diversities do this? One interpretation is that Hill diversities express the diversity of a community in terms of an imaginary community with that same diversity, but in which all species are equally abundant (Jost 2006). For example, imagine comparing two communities using a given Hill diversity (i.e., with a given exponent in eqn. B2). Imagine that community A has a diversity of 5 and community B has a diversity of 25. This means that community A has the same diversity as a perfectly even community with 5 species, and

community B has the same diversity as a perfectly even community with 25 species. Thus, there is a concrete sense in which community B is 5 times more diverse than community A. All Hill diversities can be interpreted in this same way.

A second advantage is that the calculation of Hill diversity is simple and already familiar to ecologists. Like the traditional diversity indices, Hill diversity summarizes relative (but not absolute) abundances, and the only data required to compute the sample Hill diversity are the relative abundances of species in a sample. The three forms of Hill diversity most commonly used by ecologists are species richness, and modifications of the traditional Shannon and Simpson indices. The key insight of Hill (1973) was that these three measures are special cases of the same general equation (Box 4). These three forms of Hill diversity — which we will refer to as **species richness**, **Hill-Shannon** diversity, and **Hill-Simpson** diversity — differ only in how they scale rarity (Box 5). Richness uses an arithmetic rarity scale, which gives high **leverage** to, and therefore remains very sensitive to, rare species; Hill-Simpson diversity uses a reciprocal scale, which shifts leverage towards, and is thus dominated by common species; Hill-Shannon uses a logarithmic scale, and falls between the two.

A third, elegant aspect of Hill diversity is that each of its forms can be thought of as taking the mean of the sample. Specifically, here we develop the idea that rarity can be defined as the reciprocal of relative abundance, and that Hill diversities calculate the mean of the rarities of the species in the sample (Patil and Taillie 1982). If a community includes many species, all rare, that community has high mean rarity. In contrast, a community with only a few species, none of which is rare, has low diversity and low mean rarity. This way of understanding what

Hill diversities “really are” may be intuitive for many ecologists, who are accustomed to thinking about rarity in the context of diversity.

The difference between richness, Hill-Shannon diversity, and Hill-Simpson diversity is that they calculate mean rarity using different types of means: the arithmetic, geometric, and harmonic means, respectively. An important point, which has generally been overlooked in the literature, is that these means differ not in how they *weight* the values they average (each value is always weighted by its frequency) but instead by how they *scale* these values. Each type of mean locates a balance point among a set of items. But the different means spread these items apart and squish them together differently. Thus, they provide greater leverage to either higher or lower values, i.e., to either common or rare species. Many ecologists are already familiar with this scaling process as it is directly analogous to the use of link functions in generalized linear models. We explore this new way of visualizing Hill diversities, and the different forms of means generally, in Box 5.

WHICH HILL DIVERSITY TO USE?

Which variant of Hill diversity to use, then? There is no one answer to this question. As Southwood quipped, about diversity indices in general, “There can be no universal ‘best-buy,’ although there are rich opportunities for inappropriate usage” (Southwood 1978). Hill diversity diminishes these opportunities, because Hill diversities require researchers to consciously chose how much leverage they want to afford to rare species. This decision is reflected in the value of the exponent ℓ. We discuss some advantages and disadvantages of using different values of ℓ below.

Species richness ($\ell = 1$) is not recommended by any of the authors who have systematically tested diversity metrics (Hurlbert 1971, Kempton 1979, Magurran and McGill 2011, Chase and Knight 2013, Haegeman et al. 2013), because it is difficult to estimate accurately outside of an experimental setting. Sample richness varies drastically with sample size and sample equalization method. This is because it is very sensitive to the rarest species: species with the largest rarity values have especially high leverage over the arithmetic mean. The same problem affects asymptotic richness estimators (Melo 2004, Chao and Jost 2015). Species richness is best reserved for special cases, such as when the community is completely known, or if there is enough information to parameterize an occupancy model (Iknayan et al. 2014, Guillera-Arroita et al. 2019).

Hill-Simpson diversity ($\ell = -1$) may be a good choice for a research question that mainly concerns the patterns in the relative abundances of common species, requires confidence that the expected diversity would not change substantially with additional sampling, or relates to the probability that two randomly selected individuals are the same species (Simpson 1949, Hurlbert 1971). The reciprocal scale used to calculate Hill-Simpson diversity spreads low rarity values apart and squishes high ones together (Box 5). Therefore, Hill-Simpson diversity is most sensitive to the differences in low rarity values (i.e., the relative abundance of common species). The expected value of sample Hill-Simpson diversity tends to be robust to sample standardization and to change little as sample sizes increases. Furthermore, true Hill-Simpson diversity may be estimated with little bias (Simpson 1949, Chao and Jost 2015, Grabchak et al. 2017), although the uncertainty in these estimates shrinks

slowly with additional sampling, and *precise* estimates remain difficult when there are many rare species.

Hill-Shannon diversity ($\ell = 0$) lies between richness and Hill-Simpson diversity, and may be the "just right" measure in many applications (Peet 1974, Kempton 1979). The geometric mean affords leverage to extreme values according to their proportional, not absolute, difference from the mean. Thus, it can respond strongly to both very high and to very low rarity values. Another argument in favor of Hill-Shannon is that many species abundance distributions are approximately log-normal (Williamson and Gaston 2005, McGill et al. 2007), and thus their central tendency might be well-described by the geometric mean. Observed Hill-Shannon diversity begins to stabilize at achievable sample sizes, and furthermore, asymptotic estimators for Hill-Shannon diversity perform reasonably well (Beck and Schwanghart 2010). The Hill-Shannon diversity retains some of the sensitivity of Hill diversities with higher exponents (such as richness), and also the robustness to sampling and sample standardization of Hill diversities with lower exponents (such as Hill-Simpson diversity). As a result, Hill-Shannon may be a good choice for characterizing gradients in biodiversity in an ecologically meaningful way.

For research questions about diversity in a more general sense, researchers should consider using all three metrics, as well as intermediate values for the exponent ℓ (Fig. 4). Although Hill diversities with different scaling exponents tend to be highly correlated within communities (Magurran and McGill 2011), they emphasize different aspects of the community, and are not fully exchangeable (Hurlbert 1971, Patil and Taillie 1982). Using more than one diversity metric

portrays the diversities of the communities most fully because, for example, one community can be the most diverse when its many rare species are given great leverage (when ℓ is large), but a different community most diverse when its more even distribution of more common species is emphasized (when ℓ is small) (Patil and Taillie 1982).

A diversity profile is constructed by estimating Hill diversity over the range of ℓ values. Researchers can do this in R with the functions “Diversity” in the package *SpadeR* (for asymptotic and sample diversities with estimated uncertainty; plotting features built in), “iNEXT” in the package *iNEXT* (for asymptotic and coverage-rarefied diversity estimates, with estimated uncertainty), and “renyi” in the package *vegan* (raw sample diversity) (Chao et al. 2016, Hsieh et al. 2016, Oksanen 2016). Each of these packages parameterizes Hill diversity with the exponent $q = 1 - \ell$.

ASYMPTOTIC ESTIMATORS

Are asymptotic estimators the solution?

This guide has focused on standardizing samples and then calculating sample diversity as a means to compare true diversities. This approach is imperfect, because sample diversity is not expected to equal true diversity (Hurlbert 1971, Dauby and Hardy 2012, Chao et al. 2014a). An alternative method is using asymptotic estimators to predict what diversity would be, if each community were sampled until the species accumulation curve reached its asymptote. Asymptotic estimators that do this (Chao and Jost 2015) are quite popular. However, we

believe that asymptotic estimators have two important limitations that ecologists sometimes overlook.

First, at feasible levels of sampling, asymptotic Hill diversity estimators frequently do not reach their own asymptote (Melo 2004, Beck and Schwanghart 2010, Chiu and Chao 2016, Close et al. 2018). In other words, the estimate of the asymptotic, "true" diversity value can increase with sampling (e.g. Fig. 3). This means that the diversity estimates given by asymptotic estimators are usually conditional on sample completeness, which hinders comparisons between communities and between studies.

Second, the uncertainty associated with asymptotic estimators can be large and difficult to quantify, particularly for richness (Haegeman et al. 2013). When sample coverage is low, the approximated confidence intervals (CI) around asymptotic diversity estimates for all Hill diversities are wide. Even so, they are not reliably wide enough. In fact, the CI for asymptotic Hill diversity estimators frequently do not overlap the true community diversity at their stated level (e.g. 95%) (Mao et al. 2017). For example, for a simulated community with a richness of 200, a Hill-Simpson diversity of 50, and a log-normal distribution of species relative abundances, the "95% CI" around Chao 1 asymptotic estimates of richness (Chao and Jost 2015) include the true richness value less than 50% of the time for a random sample of a few hundred individuals (Fig. 5). In contrast, the CI for sample diversity include the expected sample diversity at a rate closer to their stated confidence level (Fig. 5).

This stark difference in accuracy between the CI for sample diversity and the CI for asymptotic diversity estimators ties back to the fact that asymptotic estimators can be biased by incomplete sampling. While variation in species' sample frequencies drives uncertainty in both sample diversity and asymptotic diversity estimators (Chao and Jost 2015), CI for sample diversities are less ambitious, as they aim simply to contain the expected diversity of a sample, conditioned on size or effort (Smith and Grassle 1977). The CI for asymptotic estimators, by contrast, aspire to contain the true diversity of the full community, but often they do not. In particular, when the estimator is too low because it has not reached its own asymptote, the CI are likely to be too low also.

Is lacking valid confidence intervals a fatal flaw for a method to estimate diversity? We believe it depends on the application. Ecologists studying biodiversity will likely estimate biodiversity across many communities, and then use a statistical model to understand how biodiversity responds to predictors, such as forest cover or temperature. In the model, the uncertainty in the diversity estimates gets conflated with unmodeled but true variation between communities, and both contribute to the regression's error term. This problem can be remedied by increasing sample sizes (i.e., diversity estimates from more communities). For example, imagine sampling logged and unlogged forests to determine how logging affects species diversity (Chao and Jost 2012). Using a method such as standardizing by coverage or computing asymptotic diversity estimates may fail to provide a reliable estimate for any given site, but if enough sites are sampled, could reliably identify a group-level pattern. These methods would be preferable to a method that gave misleading estimates with better-known sampling uncertainty, such as sample diversity estimates under traditional rarefaction (Chao

and Jost 2012)

Whether to use coverage or asymptotic estimators

Thus far, we have presented two options for comparing community diversity based on sample data: either equalizing samples, preferably by coverage, and then calculating sample diversity, or using asymptotic estimators. We believe that comparing sample diversity after standardizing by coverage is the better method, but we do not mean to suggest that coverage is without its problems. Valid CI do not yet exist for expected sample diversity given coverage, as discussed in box 1. Yet, as mentioned above, a more accurate estimate with unknown uncertainty may be preferable to a more biased estimate with better known sampling variance. Standardizing by coverage and using asymptotic estimators are two different approaches to reducing biases that arise from comparing incomplete samples (Chao et al. 2014a, Chao and Jost 2015, Willis 2019), yet both lack valid CI, and both are sensitive to sample completeness. Overall, we identify two advantages to using sample diversity after standardizing by coverage, rather than using asymptotic diversity estimators.

First, the coverage method obeys the "replication principle," which states that when combining completely distinct communities of equal size and diversity (on the chosen scale), the diversity of the combination will be roughly the sum of the components. The replication principle was proven to hold for sample diversity after standardizing by coverage (Chao et al. 2014a), but to our knowledge, not for asymptotic diversity estimators. The advantage of using a metric that obeys the replication principle is admittedly less clear when comparing real-world communities for which the replication principle is unlikely to apply (i.e., the

communities being compared are unlikely to be composed of modular, equal-size and equal-diversity subcommunities). Nevertheless, it is reasonable to assume that asymptotic estimators that do not work in simple, unambiguous thought experiments are also likely to be problematic in

messy, real-world applications.

Second, and more practically, while both types of estimates depend on sampling completeness, when researchers standardize by coverage, they are at least explicit about what their sampling completeness is. By contrast, asymptotic estimators attempt to estimate the true diversity of the full community, a quantity that is not conditioned on sampling. Because neither method is robust to sampling completeness, we advocate using the one (i.e. coverage) that both accounts for sample completeness and describes it in ecologically meaningful terms. Conditioning comparisons on sample completeness can help ecologists guard against interpreting patterns that reflect researcher decisions rather than ecological processes.

Why not use both coverage and asymptotic estimators together?

In practice, ecologists typically choose a method of standardization (effort, sample size, or increasingly, coverage), *or* use asymptotic diversity estimators with unstandardized samples. However, it is tempting to combine the two methods, because asymptotic estimators tend to be sensitive to sampling completeness (Close et al. 2018).

While this sounds promising, in our view there are important issues to resolve before

coverage-based sample standardization should be combined with asymptotic estimators. First, both coverage and asymptotic diversity estimators usually rely on similar features of the data (such as the number of singletons), so combining them could introduce circularity. Second, it is not yet clear that more accurate comparisons emerge from any given combination of standardization method and asymptotic estimator, compared to using only coverage-based standardization or asymptotic estimators alone. Future theoretical and simulation-based work could build the case for a combined approach.

STANDARDIZING SAMPLES, THEN CALCULATING HILL DIVERSITY: A WORKED EXAMPLE WITH OUR BEE DATA

In this section we provide a demonstration analysis, using some of our own data, to show how the researcher's choice of how to standardize samples and calculate diversity affects interpretation of diversity patterns. We use three data standardization methods (effort, size, and coverage), as well as all three Hill diversities (richness, Hill-Shannon, and Hill-Simpson). We also compare asymptotic Hill diversity estimates to the standardized sample diversities. Our purpose is not to determine the accuracy of these methods, which we cannot do: we do not know the true diversities of our bee communities. Rather, our goal is to show how our choice of standardization method and diversity metric, as well as our level of sampling, can determine the results.

First, our answer depends on *how we standardize* our data. This can be seen by focusing on one row at a time in Fig. 6. When we standardize by size, we could conclude that species richness is fairly similar across the four bee communities, but when we standardize by effort

or coverage, strong differences among communities emerge. These findings reinforce our argument that sample standardization is an important choice that researchers need to make carefully when measuring diversity.

Second, the *choice of Hill diversity* (richness, Hill-Shannon, or Hill-Simpson) drives our understanding of the relative diversity of these four communities. We can see this by focusing on one column at a time in Fig. 6. For example, consider the column for which the data are standardized by coverage (Fig. 6, third column from left). Using richness as our metric indicates that there are large differences in diversity between the four communities, and that the purple community is the most diverse. Using Hill-Shannon or Hill-Simpson, however, leads us to the conclusion that the pink community is most diverse. Furthermore, when Hill-Shannon diversity is used, the pink diamond community appears around 25% more diverse than the purple triangle community, but when Hill-Simpson diversity is used, this difference increases to about 80%. We expected this result, because communities with many rare species need not also be “most diverse” when emphasizing the more common ones. This underscores the importance of researchers explicitly stating what aspects of diversity matter most for a particular question, and then choosing the appropriate Hill diversity to reflect those aspects.

Third, there are interactions among the choice of sample standardization method and the choice of diversity metric. As expected, although relative Hill-Shannon diversities depend on standardization, Hill-Shannon is far more robust than richness to standardization method. Hill-Simpson diversity is even more robust than Hill-Shannon diversity. While the absolute values of asymptotic diversity estimates are higher than the sample diversities, we see similar

relative diversity patterns using asymptotic estimators and coverage-based rarefaction. The robustness of Hill-Shannon and Hill-Simpson to sample standardization method is a strong argument for using these Hill diversities, rather than richness.

Finally, we note that the relative diversities for our four samples are sensitive to *sampling completeness*, even after standardization (Fig. 7). If any Hill diversities were robust to sampling completeness, we would observe a constant distance between the lines for each community within a panel of Figure 7 (i.e., the colored lines would increase in parallel). Clearly this is not the case for any of the Hill diversities shown, regardless of how we standardize samples. It is also not the case for asymptotic Hill diversity estimators, which exhibit different – but not less – sensitivity to sampling completeness up to the point we sampled each community (Appendix C). This should be concerning to field ecologists, who rarely have the luxury of comparing complete samples. Even the sample diversity *rankings*, not to mention the relative differences in diversity, vary with sample completeness for all Hill diversities. In sum, sampling completeness almost always affects diversity estimates.

CONCLUSIONS

The unavoidable truth is that when ecologists compare local diversity, they must choose how sensitive their comparison will be to the rarest species, which are always inadequately represented in samples. There is no robust way to simply “count” the species in most natural communities; richness estimated from samples depends on species’ relative abundances and sampling completeness.

Whereas ecologists usually cannot compare true species richness, we have shown how ecologists can compare communities using sample richness, Hill-Shannon, and Hill-Simpson diversity, after rarefying samples to equal coverage. Using Hill diversities requires only minor modifications to the diversity metrics that ecologists already use. These small modifications make a big difference, as Hill diversities scale intuitively, are always expressed as rarities, and require that ecologists explicitly choose how sensitive their diversity metric is to rare species.

Standardizing samples by coverage improves upon simply acknowledging the fact that both sample size and the true distribution of species abundances drive diversity estimates. It takes bigger samples to capture the diversity of more diverse communities. Coverage measures how representative samples are of the communities from which they are drawn, so equalizing coverage before measuring diversity can reduce bias in biodiversity comparisons.

Even though the tools contained in this guide are the best available at present, they are still under development. Ecologists still lack heuristics for identifying sufficient sample coverage levels, for choosing the appropriate Hill diversity scaling exponent for a given question or dataset, and for robustly accounting for uncertainty in diversity estimates. Nonetheless, when researchers have a strong argument for comparing the species diversity of communities, the tools in this guide should facilitate doing so in a principled manner. Using coverage with Hill diversity, ecologists can assess relative differences in diversity between communities based on sample data, while clearly expressing the sample completeness upon which their inferences depend.

Glossary

Sample diversity	The (Hill) diversity of a sample. This quantity can be calculated directly, as the number of species and their relative abundances are known.
True diversity	The true (Hill) diversity of an entire community.
Asymptotic diversity	An estimate of the true (Hill) diversity of the community. This is known as the "asymptotic" diversity because as the sample size increases, sample diversity and other diversity estimates converge on their true values, which are seldom known *a priori*.
Coverage	The proportion of individuals in the community belonging to species represented in a sample.
Hill diversity	Also called Hill numbers; the generalized mean species rarity.
Hill-Richness	The Hill diversity when $\ell = 1$, the arithmetic mean rarity, or the total number of species. Referred to simply as "richness" throughout.
Hill-Shannon	The Hill diversity when $\ell = 0$, the geometric mean rarity, or the exponential of Shannon's entropy.
Hill-Simpson	The Hill diversity when $\ell = -1$, the harmonic mean rarity, or the inverse of Simpson's concentration index.
Leverage	The influence of a value on the mean depends on the frequency of that value ("weight"), but also its displacement from other values in the set ("leverage"). The farther a given value from the others, the more leverage that value has. Rescaling shifts leverage from low to high values, or vice-versa.

Rarefaction	A process of randomly subsampling by removing individuals or subsamples.
Extrapolation	An approach to estimating the diversity of an augmented sample that may resemble rarefaction “in reverse.”
Rarity	1/relative abundance (Patil and Taillie 1982).

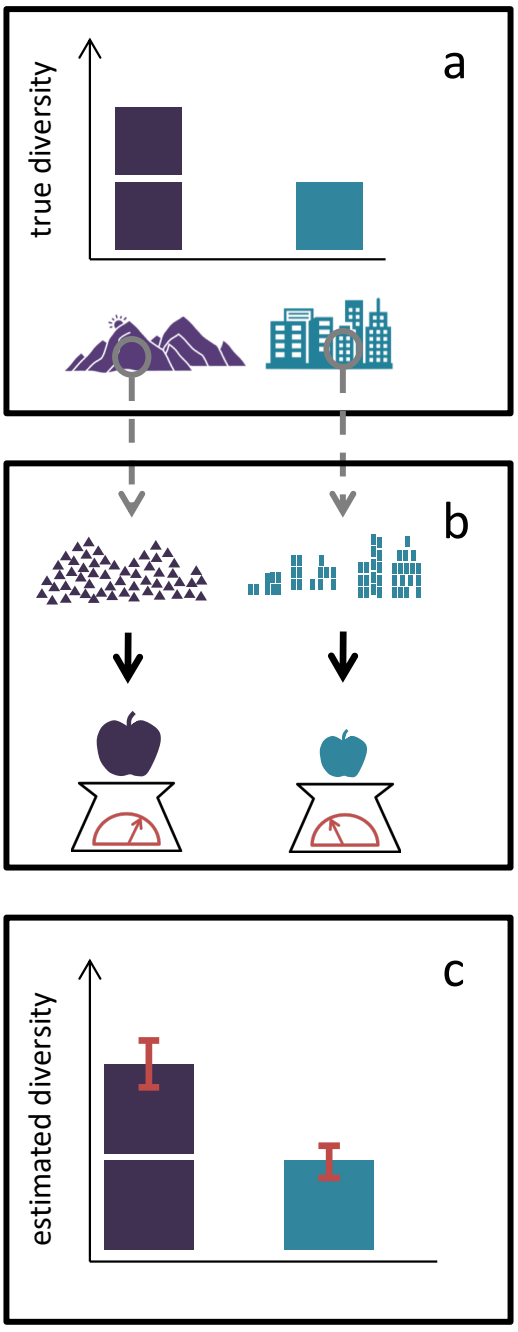


Figure 1. Comparing biodiversity between communities using sample data requires both sample standardization and appropriate metrics of biodiversity. (a) The true species diversity of an ecological community in the mountains may be double the true diversity of a community in the city. Nevertheless, samples collected from the mountains and city are unlikely to be directly comparable in terms of their diversity, even if the samples were collected using equal

effort. (b) This guide describes methods for standardizing samples, and diversity metrics for making equitable comparisons. (c) With appropriate sample standardization and metrics, ecologists can correctly estimate that the mountain community is twice as diverse as the city community.

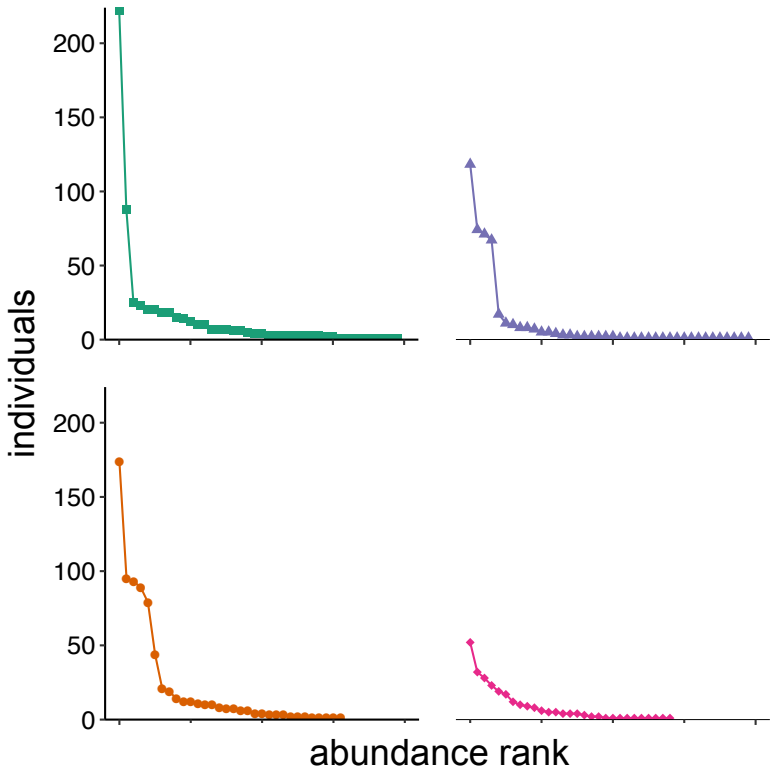


Figure 2. Observed rank-abundance distributions for the bee samples from our four meadows. The sample from the green square community has strong dominance by a small number of species with a long 'tail' of rare species. The sample from the purple triangle community also has strong dominance and a long a tail of rare species, although it has fewer species of intermediate rarity. The sample from the orange circle community has a much shorter tail of rare species. The pink diamond community sample has the least variation in rarity, and the fewest species. Diversity metrics summarize these distributions to enable quantitative comparisons.

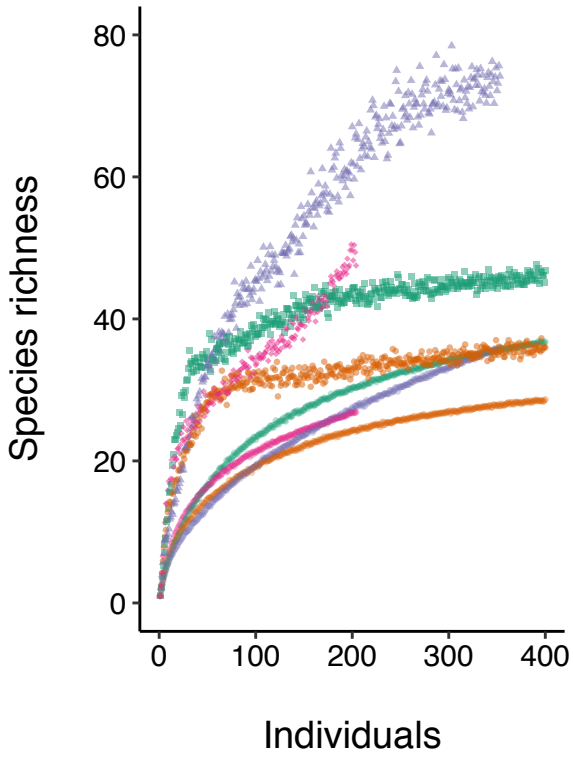


Figure 3. Species accumulation curves for number of species observed (y) versus the number of individuals sampled (x) for the bee communities in four meadows. The clouds of points represent the Chao1 estimates for the meadow of the same color. Chao1 predictions seem likely to continue increasing in most examples.

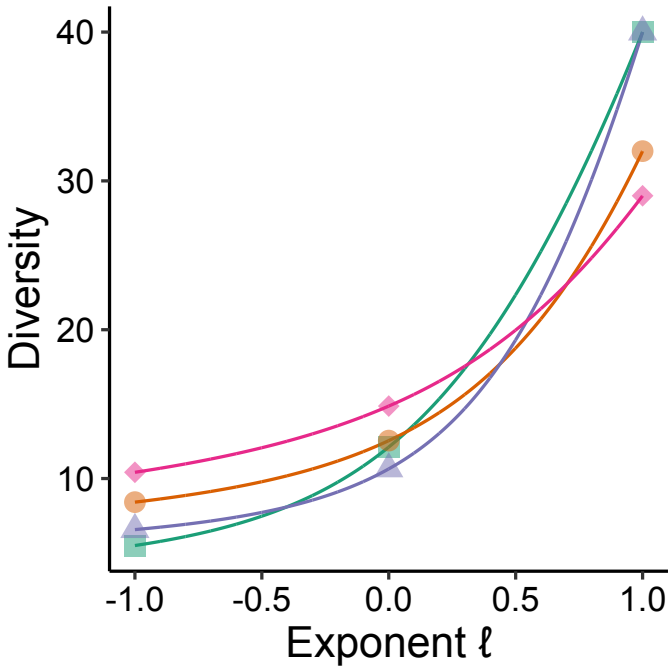


Figure 4. It can be useful to visualize a “diversity profile” across values of the exponent ℓ. Here we show the sample diversities of our four bee communities plotted as a function of the exponent ℓ in Equation B2. The y-axis is the value of the diversity metric, as calculated from the raw sample. The lines can cross because a sample can have, for example, a large number of rare species (high richness, rightmost points) but a small number of common species (low Hill-Simpson, leftmost points), as compared with another sample (middle points are Hill-Shannon).

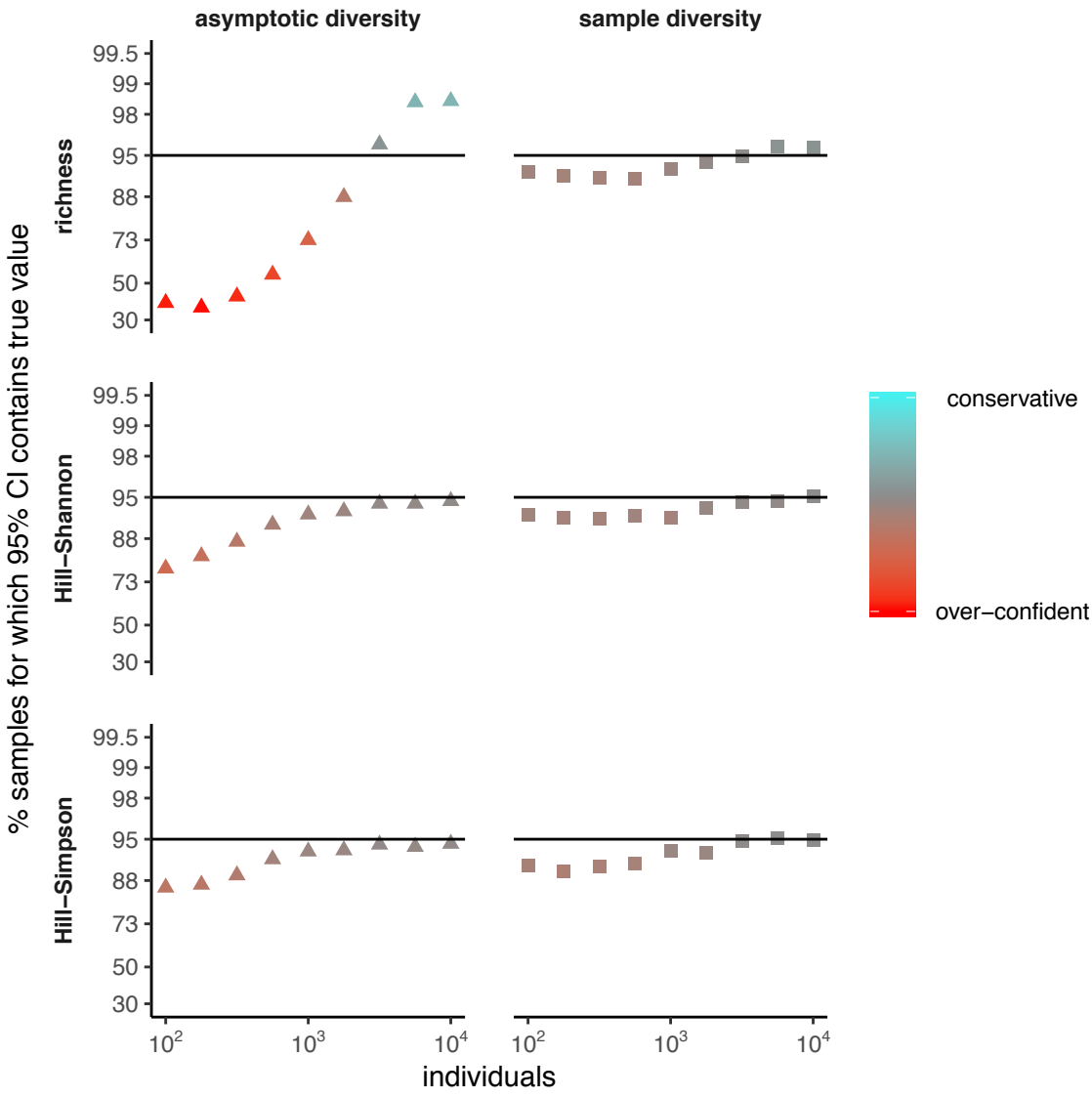


Figure 5**.** Nominal 95% CI do not consistently include their target value 95% of the time for either sample diversity or asymptotic diversity estimates, but are much closer for sample diversity. To generate this figure, fixed numbers of individuals were randomly sampled 5000 times for each sample size, and for each sample, both the asymptotic diversity estimate and associated confidence intervals (left) and the sample diversity and associated confidence intervals (right) were computed for species richness, Hill-Shannon, and Hill-Simpson diversity. The y-axis in each panel

represents the percentage of samples for which the estimated 95% confidence intervals contained (left) the true diversity of the simulated community or (right) the average diversity of samples with a given number of individuals; in each case this should be 95%. We plotted the Y-axis on the log-odds scale to be able to show both strong conservativism and strong overconfidence. Simulation methods described in appendix B.

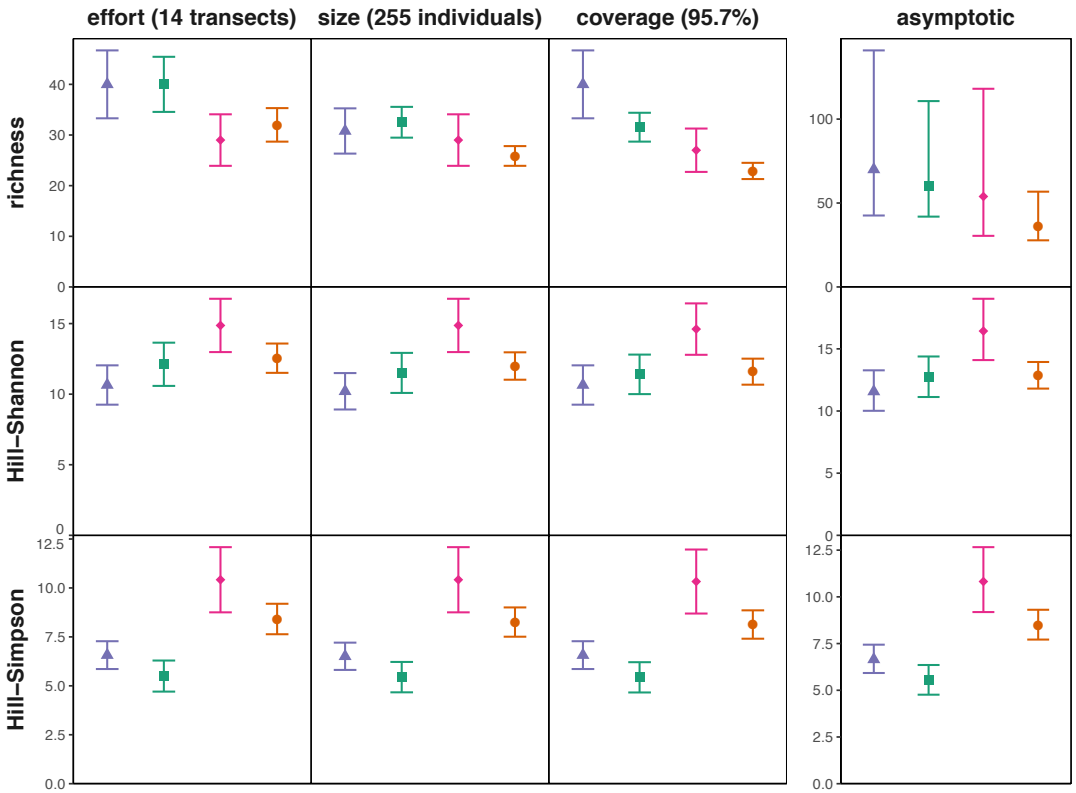


Figure 6. The answer to, “which communities are more and less diverse, and by how much?” depends on both how the samples are standardized (columns), and which diversity metric is used (rows). Standardization method matters most when Hill diversity is strongly driven by the rarity of the rarest species (species richness, top row) and matters least when rare species have little leverage (Hill-Simpson, bottom row). Error bars are “95% CI” that assume uncertainty arises from the process of randomly sampling a fixed number of individuals (i.e. the number of individuals in the sample after standardization) from each community; raw (i.e. equal effort) samples used for asymptotic estimates. We plotted the asymptotic Hill diversity estimators with separate y-axes, to facilitate comparing relative differences in estimated diversity.

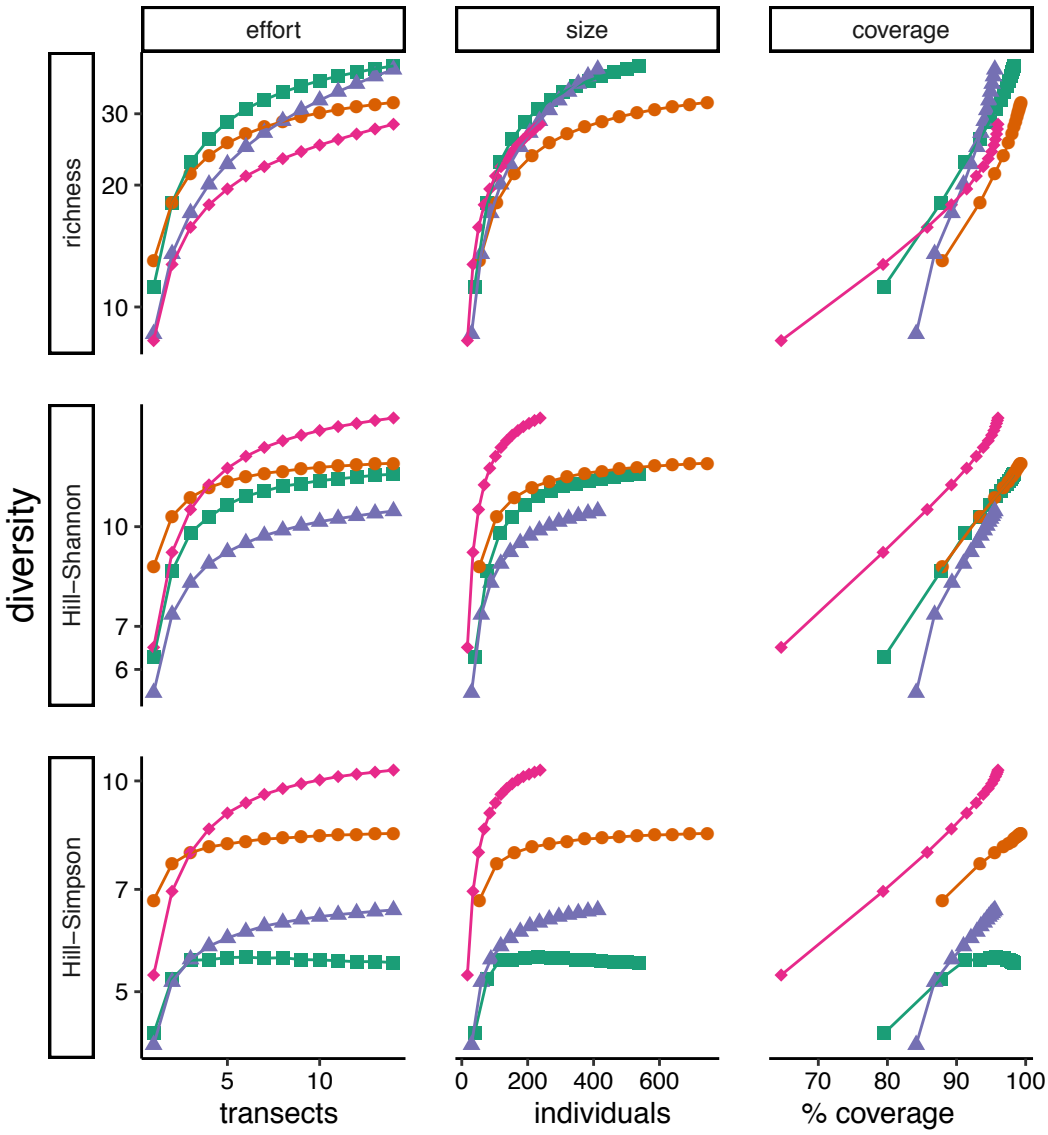


Figure 7. In addition to standardization method (columns), Hill diversities (rows) are sensitive to the amount of sampling (x-axis). To generate this figure, one to fourteen 30-minute data collection events per community were resampled without replacement 9999 times. For each group of 9999 random subsamples, average effort (number of 30-minute transects resampled), average size (number of individuals), or average coverage is plotted on the x-axis, and average sample diversity is plotted on the y (with tick marks places at log scaled intervals, but actual diversity values shown).

The logarithmic y-axis reveals a constant relative difference in diversity as a constant distance between lines. Uncertainty estimates omitted for clarity. Diversity often increases rapidly as coverage gets very close to one, because in our communities (and in natural communities in general) there are many rare species, each of which makes up a small share of the total abundance.

BOX 1: What is coverage?

Coverage is a measure of how completely a community has been sampled. Specifically, it estimates the total true relative abundance in the community of all the species represented in the sample. Coverage can be visualized as the complement of the slope of a species accumulation curve (Fig. B1). Coverage increases more slowly as sample size increases and more and more species are detected. In ecological communities, this slowing is often quite dramatic because, while most species in an ecological community are likely rare, most individuals in the community belong to common species.

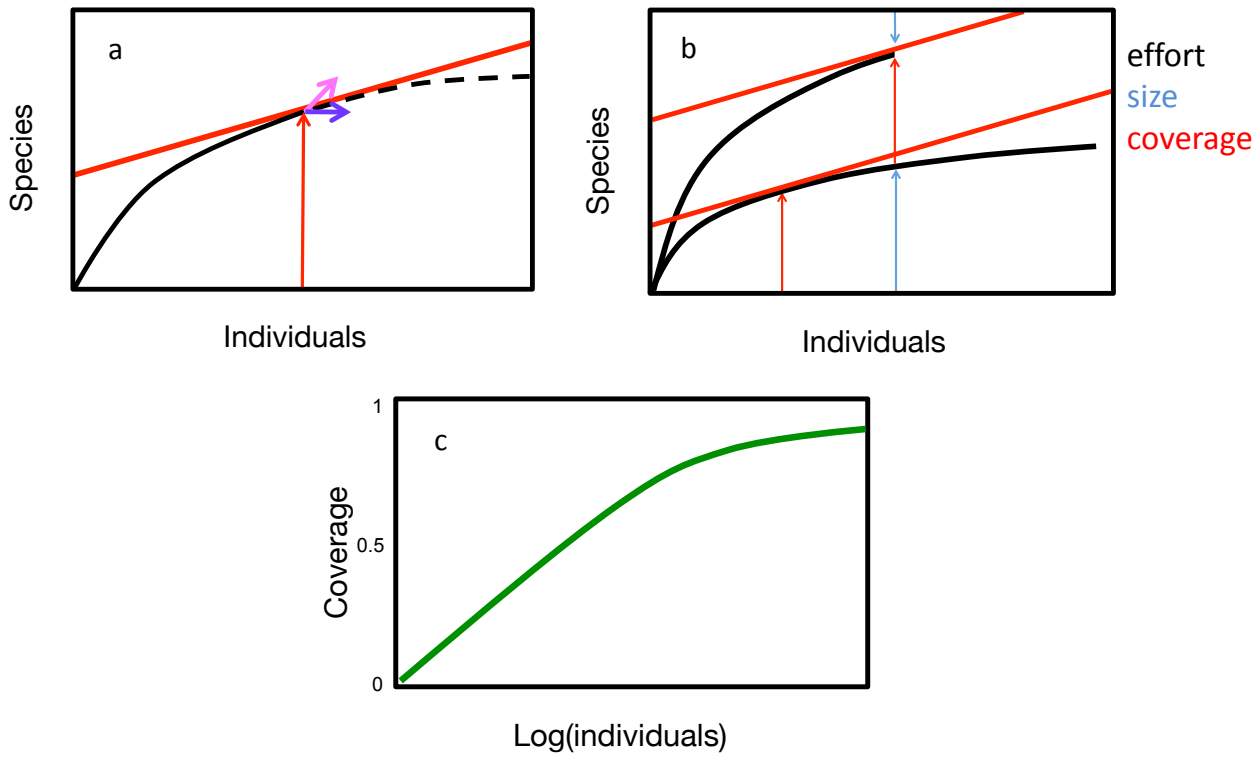


Figure B1. (a) As an ecologist collects individuals, there is some probability that the next individual would be from an already detected species (short horizontal arrow), or

from a new one (short diagonal arrow). The chance that it is from a new species is the slope of the species accumulation curve (red line) at that point. The probability of *not* picking a new species (1-the slope) is the coverage, which approaches 1 as the curve flattens out. (b) Two species accumulation curves at equal effort (ends of black curves), equal size (light blue arrows), and equal coverage (red arrows). These three data standardization methods often result in different diversity estimates. (c) At higher values of coverage, to obtain even modest gains in coverage, sample sizes may need to increase by orders of magnitude.

To estimate coverage, only 3 parameters are needed (Chao & Jost 2012):

f_1, the number of singletons (species represented by only 1 individual) in the sample

f_2, the number of doubletons (species represented by only 2 individuals) in the sample

n, the total number of individuals in the sample

Chao and Jost (Chao and Jost 2012) provide the following equation for coverage (C):

$$C = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right] \qquad \text{Equation B1}$$

R code that will estimate coverage from sample abundances is available with the functions "iNEXT" and "estimateD" in the R package *iNEXT* (Hsieh et al. 2016).

At present, standardizing samples based on coverage faces two unsolved issues. First, the methodology for rarefying to equal coverage is a bit clunky. The R package *iNEXT* uses rarefaction procedures for size-based rarefaction, but describes the sample size for each community in terms of the expected coverage for a sample of that size. The package authors justify this because both expected coverage and expected diversity are non-decreasing functions of sample size (Chao and Jost 2012, Chao et al. 2014a, Hsieh et al. 2016). This is "clunky" because the relationship derived to predict coverage given sample size is run backwards, effectively assuming that there is a perfectly predictable relationship between them. However, these approximations can work well
(Chao and Jost 2012, Chao et al. 2014a).

Second, as a result of the indirect rarefaction procedure, the confidence intervals (CI) provided for coverage-standardized samples are too narrow (i.e., anti-conservative). This is because the number of individuals required to achieve a target coverage level is uncertain, but the CI only reflect variability in sample diversity given a particular, fixed sample size (Chao and Jost 2015).

BOX 2: **Problems with the traditional Shannon and Simpson indices**

The first problem with traditional diversity indices is that they measure very different things (Tuomisto 2010). Species richness, of course, measures the number of species. The Shannon index measures uncertainty about the identity of species in the sample, and its units quantify information (bits; Hurlbert 1971), while the Gini-Simpson (1- Simpson's original index) measures a probability, specifically, the probability that two individuals, drawn randomly from the sample, will be of different species (Simpson 1949, Hurlbert 1971). Because species richness, the Shannon index, and the Gini-Simpson index do not measure the same quantities, justifying the choice of one of them to represent diversity is particularly difficult.

A second problem is that the Shannon and Gini-Simpson indices behave in ways that do not make sense for a metric of diversity. For example, if a diverse community (Fig. B2(a)) loses 1/3 of its species, the traditional Shannon and Gini-Simpson indices shown only small proportional changes (Fig. B2(b)). Even a loss of 2/3 of species does not result in dramatic changes in index values (Fig. B2(c)). In contrast, all of the Hill diversity measures presented in this guide would give values of 30, 20, and 10 for the three communities. This property of Hill numbers is called the "replication principle" (Hill 1973, Chao et al. 2014a). Note that although in the illustrations, individuals are lost along with their species, the values of all diversity metrics would be the same if total abundance were held constant even as species were lost. That is because all the diversity metrics discussed in this guide consider only relative, not absolute, abundance.

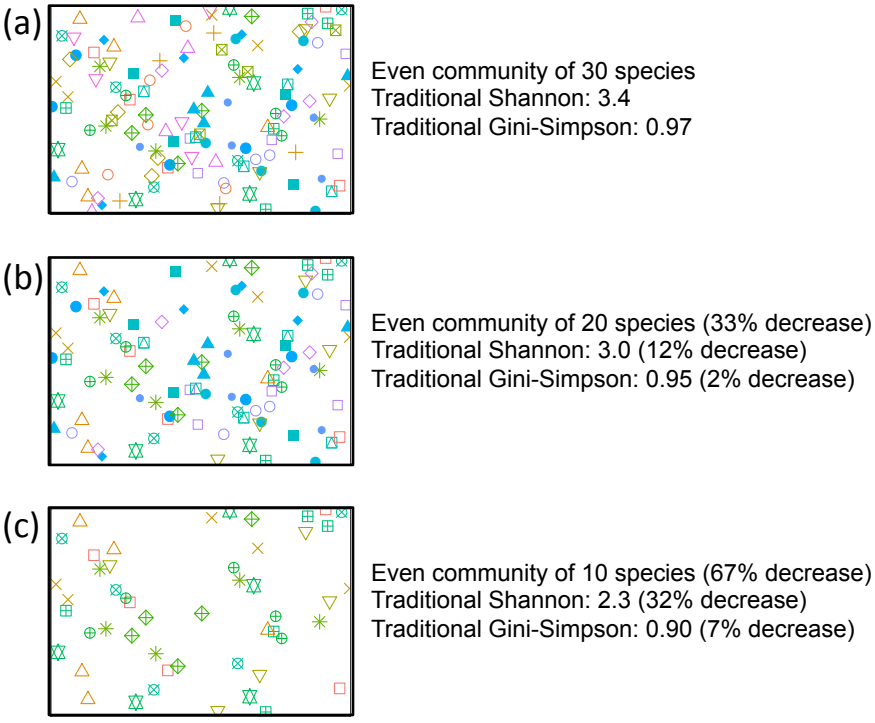


Figure B2. Values for the traditional Shannon and Gini-Simpson indices, calculated for communities that have decreasing numbers of species.

BOX 3: **Defining Hill diversity**

We define Hill diversity by the equation

$$D = \left(\sum_{i=1}^{S} p_i (r_i)^{\ell} \right)^{1/\ell}$$

Equation B2

where *D* is diversity, *S* is the number of species, p_i is the proportion of all individuals that belong to species *i*, r_I is the rarity of species *i,* defined as $1/p_i$, and ℓ is the exponent that determines the rarity scale on which the mean is taken.

An elegant aspect of Hill diversities is that Equation 2 is the equation for the generalized weighted mean, or Hölder mean (Bullen 2003). We intentionally use the exponent ℓ rather than "q" (Jost 2006) to highlight this insight; it is easily shown that our equation is algebraically equivalent to Jost's, with ℓ= 1-q (appendix A).

Hill diversities take the mean of the sample. Specifically, they measure the mean rarity of the species in the sample, where the rarity of a species is the reciprocal of its relative abundance (Patil and Taillie 1982). When computing this average, the rarity of each species is first scaled by the exponent ℓ, and then weighted by the relative abundance of that species. This average is then back-transformed onto the diversity scale because of the outer exponent, the power of $1/\ell$. It may be helpful to think of the exponent ℓ as determining the leverage provided to rare species, and to recognize that for all values of the exponent, each species is weighted by its

relative abundance. We discuss this idea further in Box 5.

Hill diversity formalizes a simple truism: a community consisting of species that are, on average, more rare has higher diversity (Patil and Taillie 1982, Tuomisto 2010, Botta-Dukát 2018, Kondratyeva et al. 2019).

BOX 4: **Three particularly useful Hill diversity metrics**

While Hill diversities are a continuous function of the exponent ℓ in equation B2, three particular integer values of ℓ produce versions of metrics that are already familiar to ecologists: species richness, Hill-Shannon, and Hill-Simpson.

The only data required to calculate the Hill diversity of a sample are the number of individuals of each species found in each sample. The equations below have only two types of parameters:

S = number of species in the sample

p_i = (number of individuals of species i) / (total number of individuals in the sample)

Species richness emphasizes (provides higher leverage to) rare species, and can be simply calculated as:

$$S$$

This is equivalent to Equation B2 when ℓ = 1.

Hill-Shannon diversity emphasizes neither rare nor common species. It is defined as the limit of Equation 2 as ℓ approaches 0, and is calculated with the base of the natural logarithm, e, raised to the power of the traditional Shannon entropy index:

$$e^{-\sum_{i=1}^{S} p_i \ln(p_i)}$$

Equation B3

Hill-Simpson diversity emphasizes (provides higher leverage to) the common species. It is equivalent to Equation 2 when $\ell = -1$. It is also the inverse of the traditional Simpson index:

$$\frac{1}{\sum_{i=1}^{S} (p_i)^2}$$

Equation B4

Sample Hill diversities can be computed using the function “renyi” in the R package *vegan* (Oksanen 2016), and Hill diversities of equal-sized or equal-coverage samples can be approximately compared using the functions “iNEXT” and “estimateD” in the R package *iNEXT* (Hsieh et al. 2016). Estimates for asymptotic values of Hill diversity are available in *SpadeR* (Chao and Jost 2015, Chao et al. 2016).

Box 5: A new way to visualize mean rarity

When ecologists calculate Hill diversity, though they may not realize it, they calculate the arithmetic, geometric, or harmonic mean species rarity. The exponent ℓ in equation B2, which scales the rarities and determines what type of mean is calculated, could also be thought of as a link function. Every ecologist has used a link function to transform values onto a scale at which they will be additive, calculated the mean, and then back-transformed the mean onto the original scale. In fact, we do this just to calculate the standard deviation of a sample.

To calculate the standard deviation, we raise each difference from the mean to the power 2, add these new, squared values together, divide by the sample size, and then back-transform to the original scale of the data by raising the computed mean to the power of 1/2. In other words, we use the quadratic link function. The root mean square error of a model is computed the same way.

A generalized linear model with an identity link estimates the arithmetic mean of the data, and could be thought of as raising each value to the power of 1, taking the mean, and back transforming the mean by raising to the power of 1. Of course, this is the same as not transforming at all. When a log link is used in a generalized linear model, the data are transformed by taking the logarithm, and then typically the mean is back transformed to the original scale by exponentiating. Thus, the mean that is calculated with the log link is the geometric mean (which is the limit of the generalized mean when the scaling exponent ℓ approaches 0). The harmonic mean uses the reciprocal function as the link (to transform, raise

to the power of -1; to back transform, raise to the power of -1). A similar link function is used with gamma error structures in generalized linear models.

But what do these transformed scales, these link functions, look like (figure B3)?

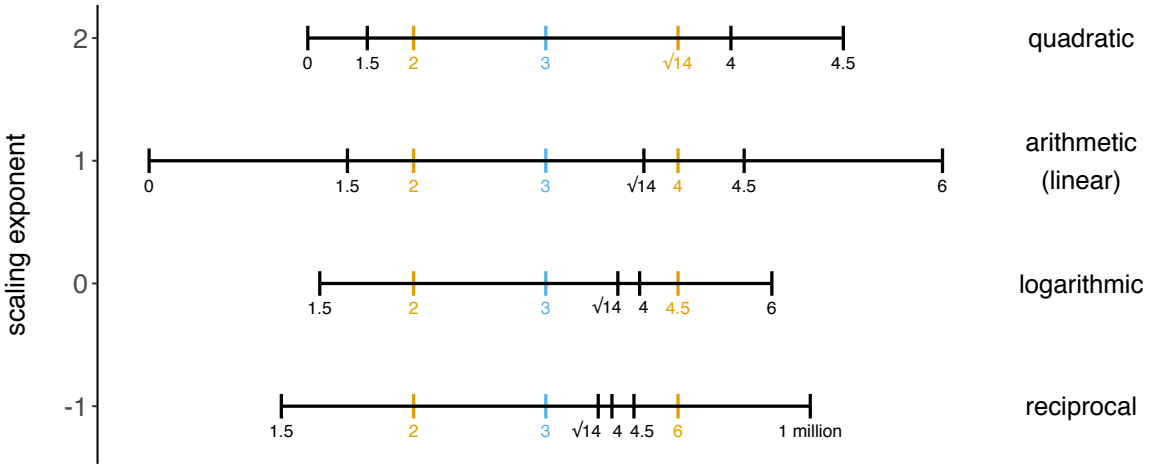


Figure B3. A link function can be visualized in terms of how it scales values. On each scale, a power transformation with the scaling exponent shown on the left, the two yellow points are equidistant from the blue one. It is not possible to align more than 2 points on two different scales; here, each scale is aligned to show that 2 and the higher yellow value are equidistant from 3. On the quadratic scale, the distance between two values is the difference between their squares. Thus, the distance between 2 and 3 is equal to the distance between 3 and √14 (~3.74) because 2^2 (4) and (√14)^2 (14) both differ from 3^2 (9) by 5. On the arithmetic scale, distances between pairs of values are their arithmetic differences. Thus, the distance between 3 and 2 is equal to the distance between 3 and 4; both differ from 3 by 1. On the logarithmic scale, the distance between two values is the factor (proportion) by which the two values differ.

> Thus, the distance between 3 and 2 is equal to the distance between 3 and 4.5, because both differ from 3 by a factor of 1.5. On the reciprocal scale, the distance between two values is equal to the difference in their reciprocals. Thus, the distance between 3 and 2 is equal to the distance between 3 and 6.

The three "link functions" used in computing the arithmetic, geometric, and harmonic means correspond to scaling exponents of 1, 0, and -1 in equation B2, respectively. The mean of a set of values, when put on the appropriate scale, is the balance point between them. This could be visualized as the fulcrum on a balanced lever. The scales differ in which values are spaced farthest apart (Fig. B3), and thus which extreme values will be most displaced from the center, or given the highest leverage. As the scaling exponent decreases, the leverage afforded to high values shrinks, and the leverage afforded the lowest values grows. For example, relative to the arithmetic scale (exponent = 1), a log-transformation (exponent = 0) spreads the small values out but compresses the largest values together.

When thinking about the different Hill diversities, it may be useful to consider this leverage metaphor. Historically, the differences between Hill diversities with different exponents (for example, the difference between species richness, Hill-Shannon, and Hill-Simpson) have been discussed in terms of how heavily the exponents "weight" rare or abundant species (Jost 2006, Magurran and McGill 2011). From equation B2, it is clear that this is not the simplest interpretation. Regardless of the exponent, each species is always weighted by its relative abundance, and every individual "counts" towards the average by the same amount. What

changes with the exponent is the *scaling* of the species' rarities, or how far apart rarity values fall (Fig B4).

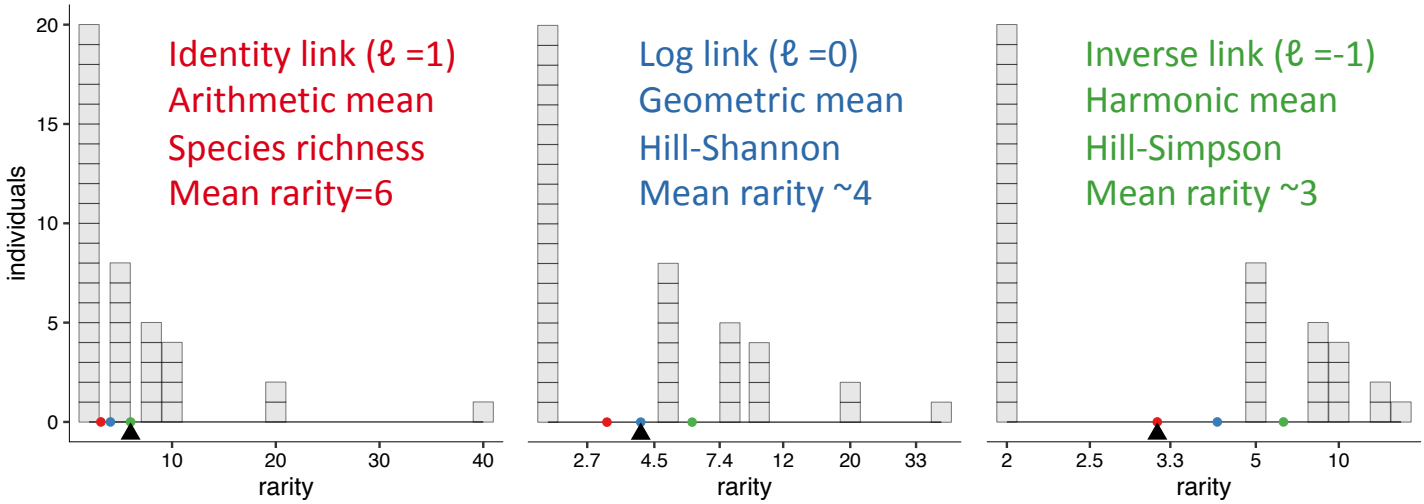


Figure B4. Diversity, or mean rarity, is the balance point for the community along the rarity scale. Each panel shows the same ecological community of 40 individuals and 6 species with abundances 20, 8, 5, 4, 2, and 1, and rarities 2, 5, 8,10, 20, and 40, respectively. Each block represents an individual; in this metaphor, the "mass" of each "block" is the same. Each individual's x-axis value is its species' rarity, which is the reciprocal of its relative abundance. All three panels display the same community: the same individuals, the same species, and the same rarities; only the scaling of the rarities changes between panels. The community's balance point along the rarity scale, pictured in this figure as the fulcrum in each panel, is the mean rarity, or *diversity*, of the community. To ease comparison across scales, in each panel, we marked the arithmetic mean with a rose dot, the geometric mean with a blue dot, and the harmonic mean with a green dot.

Panels differ in which exponent is used to transform the rarity scale. The arithmetic scale provides high leverage to very rare species; although they carry little weight (few individuals), these species influence the mean a great deal because they sit far to the right of the rarity scale. The arithmetic mean rarity of the community is the Hill diversity when $\ell = 1$, and is equal to species richness (value = 6). The logarithmic scale provides less leverage to very rare species. Thus, the geometric mean rarity of the community is lower (value ≈ 4). The geometric mean rarity is also known as the Hill-Shannon diversity, or the Hill diversity when $\ell = 0$. The reciprocal scale accords more leverage to low rarity values. Thus, the harmonic mean rarity, also known as the Hill-Simpson diversity, or Hill diversity when $\ell = -1$, is much lower still (value ≈ 3). An interactive online application that enables users to specify species abundances and the scaling parameter is available at https://mean-rarity.shinyapps.o/rshiny_app1/

Appendix A: Linking equations for Hill diversity

It has previously been observed that Hill diversity is a measure of mean rarity (Patil and Taillie 1982). However, may not be immediately obvious how Hill diversities are in fact calculating means. Here we link our intuition that Hill diversity computes mean rarity and notational conventions for Hill diversity in the literature.

Suppose we have a simple community of 10 individuals of 4 species with abundances 5, 2, 2, and 1. Let us use the relative abundance of each species as a measure of its commonness, such that the commonness of species *i* is

$$c_i = \frac{abundance_i}{\sum_{i=1}^{S} abundance_i}$$

Equation 1

where the sum is taken over the total number of species in the community, *S*. Rarity will simply be the reciprocal of this quantity,

$$r_i = {}^{1}/_{c_i} = \frac{\sum_{i=1}^{S} abundance_i}{abundance_i}$$

Equation 2

Then, we could ask what the average rarity of species in this community is. One approach would be to sum the rarity of each of the four species, and then divide this sum by 4:

$$\frac{(\frac{10}{5}+\frac{10}{2}+\frac{10}{2}+\frac{10}{1})}{4}$$

Equation 3

This would give us an average rarity of 22/4, or 5.5. But perhaps this is an unintuitive way to average; there is only one individual with rarity 10, after all, yet the way we computed the average community rarity, it contributes as much to the average as the 5 individuals with rarity 2. Perhaps this seems strange.

A popular solution to this kind of problem is to weight each unique quantity by the number of times it is observed, thus producing a weighted mean:

$$\sum_i^n w_i x_i,$$

Equation 4

which describes the mean of a set of *n* items (in our case rarities) indexed by *i*, x_i, each of which is weighted by something, w_i (In our case that weight is the relative abundance of species *i*). To return to our previous notation, we now represent each species by its rarity value, and weight each species by its relative abundance in the sample, or its commonness:

$$\sum_i^S c_i r_i$$

Equation 5

If we re-compute the mean using equation 5, we find each species *i* contributing exactly a

quarter of the rarity, i.e. 1, to a community mean rarity of 4. This is somewhat odd: we just determined that in order to count the rarities fairly, we needed to account for the relative abundance of each species, yet all species appear to contribute equally to the mean rarity regardless of relative abundance. Nonetheless, this is a widely used diversity metric: species richness. In other words, richness is the arithmetic mean community rarity.

From this example, it should be clear how this quantity is at once "independent" of the distribution of relative abundances of the species, and also an abundance-weighted average of species rarity (a "Hill diversity"). It should also be clear why this diversity metric is so sensitive to the number of rare (and poorly sampled) species.

As discussed in the main text, richness, the arithmetic mean, is a special instance of the *generalized mean*:

$$(\sum_i^n w_i {x_i}^l)^{1/l}$$

Equation 6

This looks similar to equation 5 but there's a new parameter, ℓ. The *generalized mean* of a set (Equation 6) always lies between the smallest and greatest element (inclusive), and the exponent ℓ determines the leverage accorded to large versus small elements in defining the location of center of the set. Our original (arithmetic) mean, where $\ell = 1$, is very sensitive to large (outlier) values of rarity (i.e. especially rare species). In fact, the arithmetic mean rarity increases just as much when you include one individual of a new, rarest species as if you

included hundreds or even millions of individuals of a new, but very common one!

The *geometric mean* (the limit of Equation 5 as ℓ -> 0) or the *harmonic mean* (ℓ = -1) are also sensible ways to compute the mean community rarity. Rather than giving extreme leverage to the rarest species, these means emphasize more common ones, but to different extents.

Hill-Simpson is the harmonic mean community rarity. The harmonic mean is, in sharp contrast to the arithmetic mean, insensitive to outliers (in our case, very rare species). It increases greatly with more common species, but hardly at all with rarer ones.

Perhaps when we discuss the average rarity of our community, we seek a goldilocks answer that is driven by the rarity of typical species, rather than the rarest or most common? There is a special mean for this, too: the geometric mean (Hill-Shannon). The geometric mean lies between the arithmetic and harmonic means, and might be just right (Jost 2006).

In closing, let us return to Hill numbers, which ecologists agree are the best way to describe the diversity of ecological communities (Ellison 2010, Haegeman et al. 2013), and show that they measure the mean community rarity. The equation for Hill numbers is in fact an equation for the generalized mean (Equation 5), but traditionally it is expressed in a somewhat different form. We will start with the generalized mean. The mean rarity in the community, in which the rarity of each species i (r_i) is weighted by its commonness (c_i, relative abundance, also the inverse of rarity) is given by:

$$\left(\sum_i^n c_i \, {r_i}^l\right)^{1/l}$$

Equation 6

Replacing rarity, r_i, with $1/c_i$, we have

$$\left(\sum_i^n {c_i}^{1-l}\right)^{1/l}$$

Equation 7

Substituting q for 1-ℓ we have

$$\left(\sum_i^n {c_i}^{q}\right)^{1/1-q}$$

Equation 8

which has been the traditional equation for the Hill diversity of order q (Jost 2006).

Appendix B: Simulation methods for assessing confidence interval performance

We simulated a log-normal species abundance distribution with 200 species and a Hill-Simpson diversity of 50. We took 5000 samples each for 9 sample sizes ranging from 100 to 10,000 (with replacement). For each sample, we generated the nominal 95% confidence intervals for both the sample diversity and the asymptotic diversity following (Hsieh et al. 2016). Briefly, this procedure generates an extrapolated species abundance distribution by augmenting the sample. The augmented sample is then bootstrapped to simulate sampling uncertainty, and sample or asymptotic diversity is then computed for each bootstrapped sample. The resulting distribution of bootstrapped diversity values is then centered on the sample diversity or asymptotic estimate for the original sample; 0.025 and 0.975 quantiles of this new distribution are then chosen to provide a confidence interval. Here, we asked whether these confidence intervals contained, in the case of asymptotic diversity, the true diversity of the original simulated community, or in the case of sample diversity, the average sample from 5000 random samples pf the same size. R code for the protocol described above will be publicly available on FigShare or GitHub at time of publication.

Appendix C

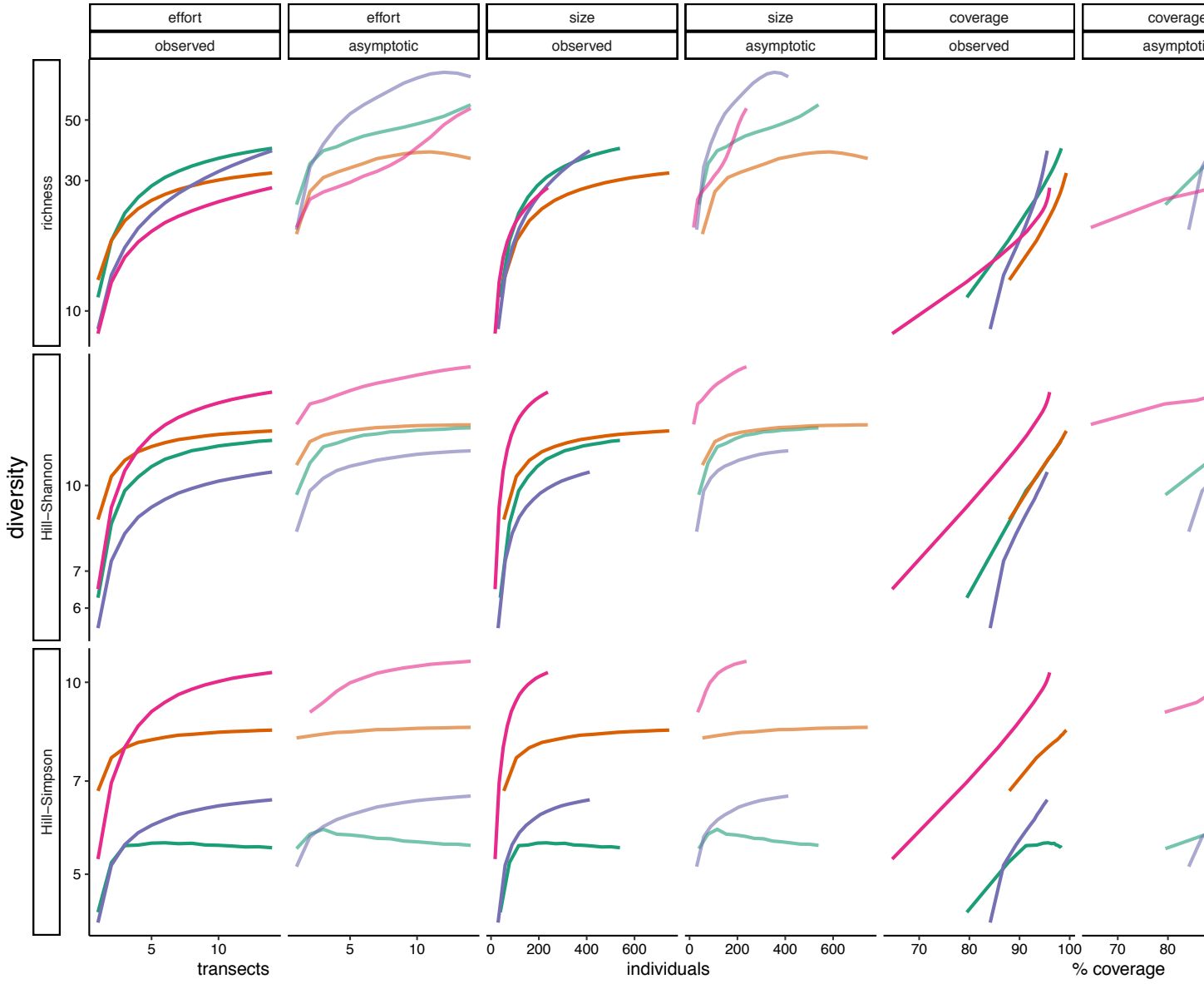


Figure 1. Asymptotic estimators and sample diversity are both sensitive to sample completeness, but not in identical ways. 9999 random samples of one to fourteen 30-minute transects were randomly sampled without replacement. For each sample, the number of individuals, estimated coverage, sample Hill diversity, and asymptotic Hill diversity estimate were computed. For each sample size, the mean value for effort (i.e. number of transects), number of individuals, sample coverage, and diversity estimate is plotted. For each group of

9999 random subsamples, average effort (number of 30-minute transects resampled), average size (number of individuals), or average coverage is plotted on the x-axis, and average sample diversity is plotted on the y (with tick marks placed at log scaled intervals, but actual diversity values shown). The logarithmic y-axis reveals a constant relative difference in diversity as a constant distance between lines.

CHAPTER 3

Assessing diversity estimators and their uncertainty with checkplots

Manuscript authors: Michael Roswell, Rachael Winfree, Jonathan Dushoff

ABSTRACT

1. Ecologists frequently measure and compare quantities, such as Hill diversity, for which parametric uncertainty estimates are unknown or invalid. Approximate uncertainty estimates may be generated through Monte-Carlo resampling schemes such as bootstrapping. However, the bootstrapping simulation may not realize the variability in ecological metrics that occurs under random sampling of natural communities.

2. We simulated species abundance distributions with known diversity, and then sampled from them to evaluate the sampling uncertainty in Hill diversity. We then assessed whether recently proposed confidence intervals for Hill diversity estimates were valid, using two new visual diagnostic tools, the “slugplot” and the “checkplot.”

3. We found that observed Hill-Simpson diversity had high sampling uncertainty even with large samples, in comparison to Hill-Shannon diversity and richness. Similarly, asymptotic Hill-Simpson diversity estimators often had higher variability than asymptotic Hill-Shannon diversity and asymptotic richness estimators, though both of these are more sensitive to rare species. The proposed confidence intervals often performed well for sample diversity, especially for more even communities. Only with very large samples could proposed confidence intervals obtain nominal coverage for

asymptotic Hill-Shannon and Hill-Simpson estimators; the proposed confidence intervals performed poorly for the Chao1 richness estimator.

4. Rough estimates of relative sampling uncertainty can help guide ecologists choice of which Hill diversity to use. We found that Hill-Simpson diversity, which can be estimated with little bias, often has greater sampling uncertainty than Hill-Shannon and in some cases, richness. Slugplots and checkplots are flexible tools for assessing confidence intervals and related statistics. We showed that proposed confidence intervals may work well for sample Hill diversities, but not for asymptotic Hill diversity estimates.

Introduction

Ecologists lack precise, robust, unbiased tools to estimate the species diversity of a community based on the diversity of samples drawn from it (Haegeman et al. 2013). While ecologists know this, do they know how accurately they can estimate diversity indices? The literature conveys that richness is very hard to estimate robustly, that Simpson's index and its Hill-number version can be estimated with little bias from samples of nearly any size, and that Shannon's entropy and its Hill number equivalent fall somewhere in between (Beck and Schwanghart 2010, Chase and Knight 2013, Haegeman et al. 2013, Chao and Jost 2015). In spite of these observations, we believe that ecologists often lack intuition for how accurate and precise commonly used diversity estimates are. We suspect that ecologists tend to overestimate their ability to precisely estimate true species richness and true Hill-Simpson

diversity, but underestimate their ability to describe the statistical uncertainty in the expected richness of a finite sample. As the field coalesces around the consensus that Hill diversities (eq. 1) are the best family of metrics for measuring and comparing species diversity, identifying methods to quantify the uncertainty in Hill diversity estimates becomes more important (Willis 2019).

We define Hill diversity D as the mean species rarity in the community

$$D = \left(\sum_{i=1}^{S} p_i (r_i)^{\ell} \right)^{1/\ell} \quad 1)$$

where D is diversity or mean rarity, p_i is the relative abundance of species i, r_i is the rarity of species i (defined as the reciprocal of p_i), S is the total species richness, and ℓ is the scaling exponent that determines the type of mean computed (Roswell et al. ND)

Quantitative measures of the statistical uncertainty in estimates of Hill diversity (D) are technically difficult to produce, and furthermore, the existing tools rely on a variety of statistical approaches with incompatible interpretations. Thus, while a number of recent publications provide quantitative interval estimates for community diversity (Dauby and Hardy 2012, Zhang 2012, Haegeman et al. 2013, Chao and Jost 2015, Willis and Bunge 2015, Zhang and Grabchak 2016, Mao et al. 2017), the intervals they provide are not all equivalent. To the best of our knowledge, only one of these approaches is widely used in practice, namely variations of the “confidence intervals” described by Chao and Jost 2015 (Chao and Shen

2003, Colwell et al. 2012, Chao et al. 2014, 2019, Chao and Jost 2015), and included in the popular R packages SpadeR (Chao et al. 2016) and iNEXT (Hsieh et al. 2016).

The Chao and Jost 2015 confidence intervals were designed to indicate the statistical uncertainty for both sample Hill diversity and also for asymptotic Hill diversity estimates. The sampling distributions of Hill diversity estimates are poorly known, which makes defining confidence intervals for these quantities very difficult.

In this study, we describe a novel approach to assessing confidence intervals, and test the confidence intervals proposed by Chao and Jost (2015) using simulated species abundance distributions. First, we sample from simulated species abundance distributions to empirically describe the sampling distributions of Hill diversity estimates. Next, we ask whether the proposed intervals for both sample Hill diversities and for asymptotic Hill diversity estimates for the three most commonly used values for the scaling exponent ℓ (1, 0 and -1, corresponding to richness, Hill-Shannon diversity, and Hill-Simpson diversity) should be considered "confidence intervals," using tools we introduce as "slugplots" and "checkplots."

Specifically, we ask

1) How much sampling variability is there in sample Hill diversities and in Chao and Jost's (2015) asymptotic Hill diversity estimators?

2) Does the method proposed by Chao and Jost (2015) generate valid confidence intervals that reflect this variability? If not, are these intervals conservative or overconfident?

3) Are the proposed intervals from this method biased high or low?

4) Does confidence interval performance depend on the species abundance distribution or choice of Hill diversity?

STATISTICAL BACKGROUND

p-values

We define a p-value to be the probability of a particular observation (or one that is more extreme), given that a particular statistical hypothesis is true.

Formally, we define the one-tailed p-value for an observation x and the statistical hypothesis Θ as $p_{\Theta)(x|\Theta)}$ as the probability of observing a parameter as or more extreme than x, and define 1-tailed p-values as

1) $p_{\Theta}^{-} = P(X \leq x|\Theta)$ and
2) $p_{\Theta}^{+} = P(X \geq x|\Theta)$

where X is a random variable, and Θ describes the statistical hypothesis.

p-values describe an observation x in terms of the quantiles of a random variable. If Θ is true, and X is a continuous random variable, then the probability density function of $p_{\Theta}(X)$ is uniform on $[0, 1]$, by the probability integral transform (Casella and Berger 2002).

If X is a discrete random variable, however, p_{Θ^-} is not always equal to $1 - p_{\Theta^+}$. When for $x \in X$, $p^+_\Theta(x) \neq 1 - p^-_\Theta(x)$, $p_\Theta(x)$ is associated, in theory, with a range of "platonic" p-values between p^+_Θ and p^-_Θ. When using p-values to construct confidence intervals or evaluate statistical hypotheses, it is prudent to be conservative, i.e. select only p^+_Θ or p^-_Θ, in accordance with the tail appropriate for a given hypothesis. When evaluating a p-value, it can be informative break the tie not by selecting the most conservative p-value associated with x, but instead by randomly sampling the range of p-values consistent with x. p-values estimated this way, even for discrete random variables, always have a discrete uniform distribution on $[0, 1]$.

What are confidence intervals for?

A frequentist confidence interval (CI) with confidence level *L* is an interval estimate obtained by a method that is expected to result in the interval containing the true value a proportion *L* of the time no matter what the true value is (Cox and Hinkley 1974).

CIs are commonly constructed as a collection of values that would not be rejected (at an aggregate p-value of $\alpha = 1 - L$) if they were treated as null hypotheses (Neyman 1937). In other words, we construct CIs using a counterfactual null hypothesis: a value falls within the CI if, imagining that that value were the true parameter value under the null hypothesis, the null hypothesis would not be rejected for the observed statistic.

Frequentist confidence intervals are commonly used to express the range of population parameter values that are consistent with the data. Thus, when valid confidence intervals can be found, they provide practical guidance about uncertainty. Because a valid confidence interval depends on a sound method to estimate p-values (in order to conduct the counterfactual null hypothesis test), to test the validity of a confidence interval, it is sufficient to test the p-value upon which it is based.

Defining confidence intervals

In this manuscript, we only explore equal-tailed confidence intervals, or those that have equal chances of being too high and too low. Thus, we set $a = b = \alpha/2$. Furthermore, 2-sided $(1-\alpha) * 100\%$ confidence intervals are bounded by l and u, where l and u are given by the 1-sided $(1-\alpha/2) * 100\%$ confidence intervals $[l, \inf)$ and $(-\inf, u]$. However, the process of imagining tests where every possible value of a parameter is the true value under the null, and finding the parameter value (l or u) that, if it were the true parameter value under the null, would result in $p_\Theta(x) = \alpha/2$ may be computationally or conceptually infeasible. A simpler approach is to assume a kind of symmetry such that $p_{\Theta_y}(x) = p_{\Theta_x}(y)$, where Θ_x indicates that x is the true value of the parameter of interest under the statistical hypothesis.

Here, we introduce two diagnostic tools to assess confidence intervals and the p-values upon which they are based, which we call “slugplots” and “checkplots,” respectively. A slugplot is an ordered plot of estimates and their confidence intervals, computed from Monte-Carlo random samples of $X|\Theta$. A checkplot is a histogram of estimated p-values for the simulated, true parameter value, but where p-values are computed based on a null hypothesis

parameterized by a Monte-Carlo random sample of $X|\Theta$ rather than the full knowledge of Θ itself.

Slugplots

In a slugplot, confidence intervals for 1000 random samples are ordered to ease visual comparison with the true value. A slugplot makes it is easy to both verify if $\alpha/2 * 100\%$ of random samples have confidence intervals above the target value and $\alpha/2 * 100\%$ have confidence intervals below the target value, and also to examine bias when nominal coverage is not achieved. A slugplot (Fig 1) is thus more nuanced than a simple measure of statistical coverage, and is useful for examining CI performance for a given combination of parameter values and sample size. Slugplots are informative for continuous statistics but may be difficult to interpret for discrete statistics.

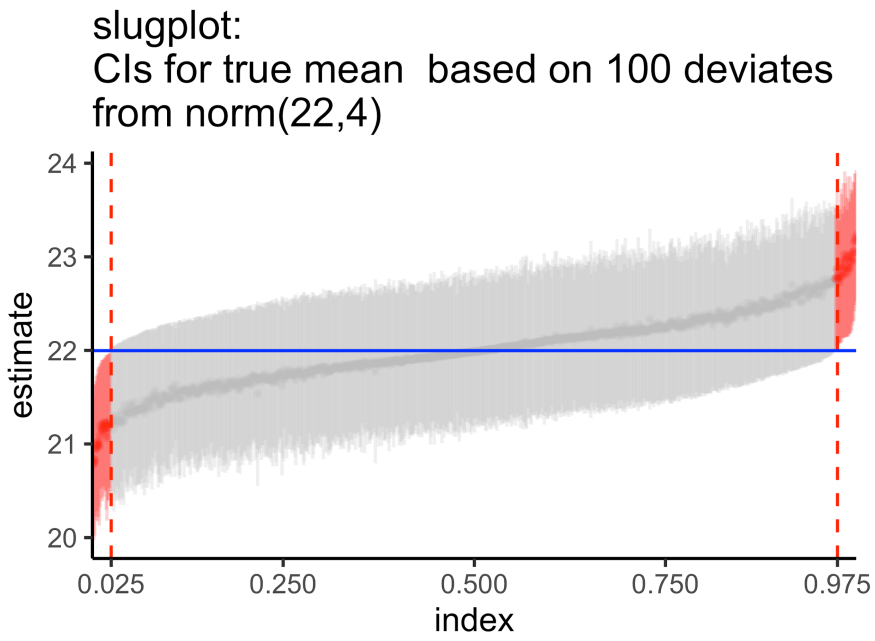


Figure 1. A slugplot for 95% CIs based on Student's t- type p-values, constructed for 1000 samples of 100 deviates from a normal distribution with mean= 22, standard deviation = 4. The slugplot exhibits near ideal behavior of these CIs, with 2.5% of CIs too low (red, at left) and 2.5% too high (red, at right). The point estimate for the mean for each sample is a darker shade for each CI.

Checkplots

Whereas a slugplot displays the empirical performance of a confidence interval (on the scale of the parameter of interest), it may be difficult to see from a given slugplot whether a confidence interval is theoretically sound. Partly, this is because it is possible to achieve nominal coverage for a particular parameter value and sample size by coincidence, despite having invalid confidence intervals. Furthermore, slugplots may be of little use for discrete statistics, for which valid confidence intervals will be conservative for most parameter values at modest sample sizes. A more general tool to assess p-values (and thus CI; a valid method to

compute p-values is both necessary and sufficient to construct valid CI) is a “checkplot,” a type of rank histogram.

To create a checkplot, one must first generate samples via Monte-Carlo simulation based on known parameters. From each simulated sample, the researcher then estimates the probability that the population parameter value is less than or equal to the simulated parameter (whereas the population parameter *is* the simulated parameter; the probabilities are computed without considering this information). This is of course slightly different from the idealized test inversion invoked in ideal frequentist CI, and relies on a symmetry assumption that the 2-sided p-value of the upper or lower confidence limit, given the estimate, is equal to the 2-sided p-value of the sample estimate, given the confidence limit. When all parameters for the true distribution (These probabilities (p-values) are then binned into a histogram (a checkplot, Fig. 2). Because we break ties for discrete statistics randomly, a valid p-value always has a flat checkplot.

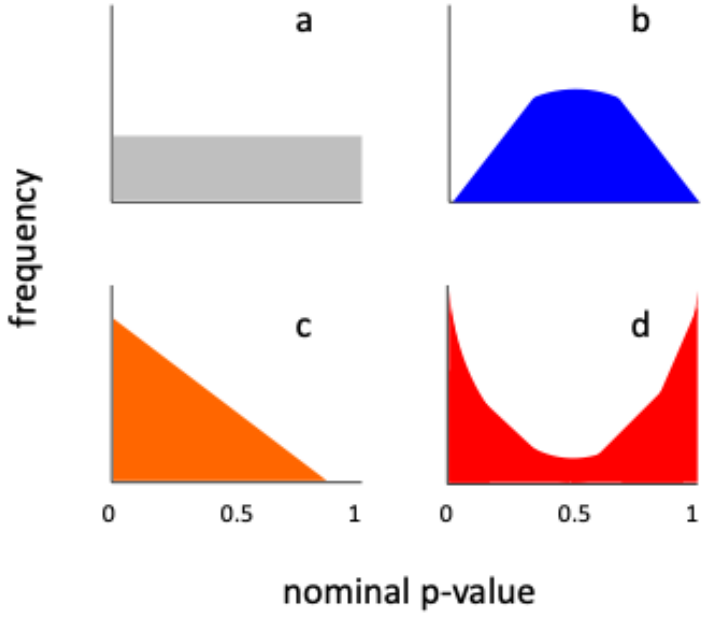
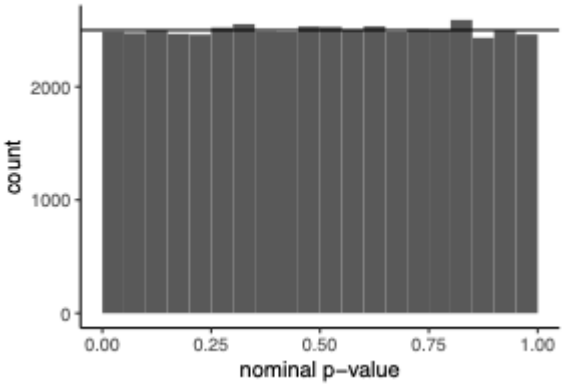


Fig 2 Checkplots can reveal accuracy, bias, and conservatism of a proposed p-value or confidence interval. a) an ideal checkplot has a uniform, unbroken distribution across all nominal p-values. b) checkplot with higher density around p=0.5, implying conservative confidence intervals c) Checkplot with low nominal p-values occurring more frequently than higher nominal p-values, implying biased, anti-conservative confidence intervals d) Checkplot with high density towards extreme p-values implies anti-conservative confidence intervals.

Checkplots can reveal how approximated confidence intervals or p-values may provide misleading statistical guidance ("How to interpret a p-value histogram" in press, Hamill 2001, Talts et al. 2018). Humped, centered checkplots imply conservative confidence intervals and p-values. These are obviously preferable to U-shaped checkplots, which imply anti-

conservative confidence intervals and p-values. When checkplots are skewed to one side, they imply biased statistics that consistently over- or under-predict population parameters.

Because small deviations from the uniform distribution may be difficult to diagnose visually, we suggest generating up to 50,000 p-values per checkplot. Alternatively, the flatness of a checkplot might be assessed with a Chi-squared test (Wilks 2019). When developing or comparing approximated p-values, these nuanced diagnostics enable assessing CI performance beyond the simple statistical coverage probability (Fig 3).



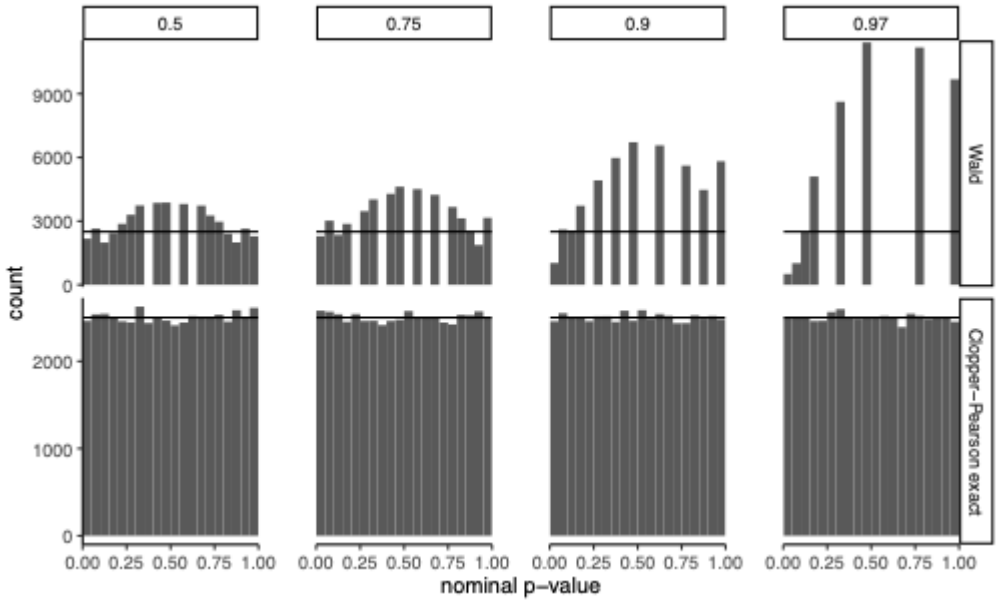


Fig 3 Checkplots for p-values and their associated CIs for samples from a univariate normal distribution (based on Student's t-statistics) and for samples from a binomial distribution (based on Wald and exact statistics) reveal the validity of each statistic and their associated CI. (a) Confidence intervals for the true mean of sample from a normal distribution may be constructed by inverting a t-test, taking the sample mean as the "null hypothesis." Therefore $p(\mu < \hat{\mu})$, where $\hat{\mu}$ is the sample mean, from 50,000 samples of 100 random deviates from a normal distribution with $\mu = 22$ and $\sigma = 4$, falls uniformly across $[0, 1]$, generating a flat checkplot. (b) The Wald binomial test and associated confidence interval depend on the discrete binomial distribution converging on a continuous chi-squared distribution, which occurs asymptotically. However, at smaller sample sizes, the approximation can be poor (Brown et al. 2001). When the binomial proportion is close to 0.5, the Wald confidence intervals are conservative. The p-values for the true proportion, taking the sample proportion as the null, from 50,000 samples of 100 random Bernoulli trials with frequency of success = 0.5 is not uniform, and the checkplot shows a higher density of nominal p-values near 0.5. With more extreme binomial proportions, the Wald intervals become biased towards 0.5, and the checkplots show a high density p-values >0.95 (c-e). In addition to providing a misleading interval, the Wald statistic depends on which condition a researcher determines to be "success." The inverse β distribution is used to estimate p-values for a binomial distribution in the exact binomial test and Clopper-Pearson interval (f-i). Due to the discreteness of the binomial distribution, $p_{\Theta} \neq 1 - p_{-\Theta}$ for

every value of Θ. Thus, when generating a checkplot, we randomly sample from the interval $[p_{\Theta}, 1 - p_{-\Theta}]$ to simulate a "platonic" p-value for each sample. The exact binomial test delivers valid p-values, and therefore a flat checkplot can be made for any binomial proportion. The Clopper-Pearson interval is constructed from inverting the exact binomial test (using the observed frequency as the null), and is therefore also valid.

METHODS

All simulations and analyses were conducted in R 3.4.1 (R Development Core Team 2015)

Simulating species abundance distributions

We simulated species abundance distributions (fig 4) based on their true diversity and a parametric shape assumption. Rather than selecting the parameter values for these distributions based on a community assembly process, or fitting curves to field data, we chose parameter values by fitting curves, constrained by the true diversity of the full species abundance distribution. The distributions we simulated have known, finite diversity but infinite species abundances. There is effectively no difference between sampling with and without replacement from these distributions. Simulation details may be found in appendix 1.

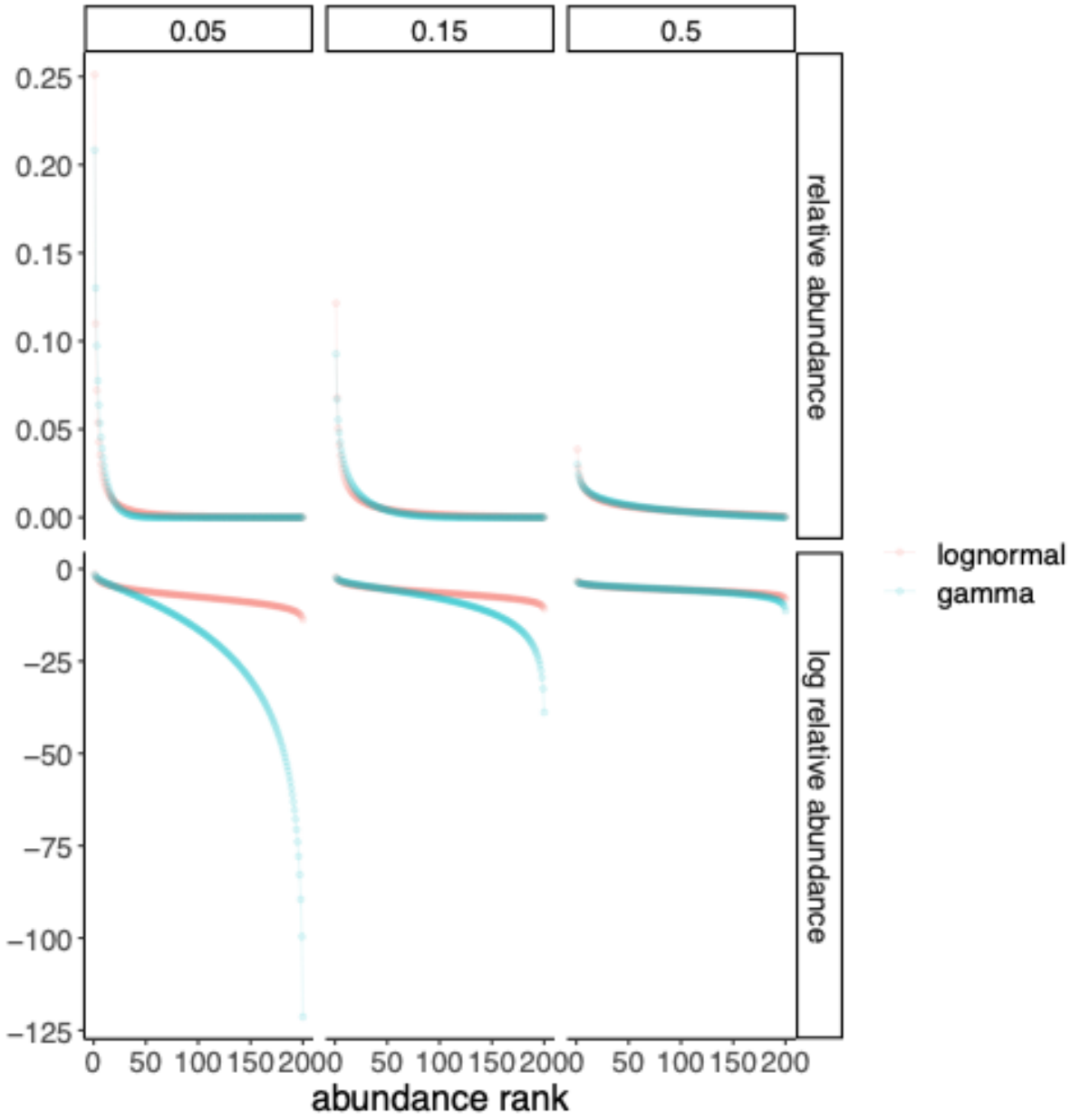


Fig 4 While the log-normal and gamma distributions were similar for even communities, for uneven communities, the rarest species were much rarer for the gamma than for the log-normal. The top row shows rank-abundance distributions, the bottom shows a “Whittaker plot” where relative abundances are log-transformed (Whittaker, 1965). Columns labelled with their normalized evenness (Chao and Ricotta 2019), which is approximately the ratio of Hill-Simpson diversity to richness. We also

simulated distributions with evennesses = {0.25, 0.75, 0.85}, and for each evenness and distributional assumption, distributions with 100 instead of 200 species.

Community sampling and diversity estimation

From each species abundance distribution, we sampled 10^2 to 10^5 individuals, 50,000 times each. For each sample, we computed the sample Hill diversity and estimated asymptotic Hill diversity (Chao and Jost 2015). We also generated "augmented" bootstrapped assemblages. We resampled from the bootstrapped assemblages 2000 times per sample, and computed diversity estimates for each bootstrap sample. We did this for sample and estimated asymptotic Hill diversities with the scaling exponent $\ell = 1$, 0, and -1 (richness, Hill-Shannon diversity, and Hill-Simpson diversity).

Assessing the sampling distribution of Hill diversities

To describe the empirical sampling uncertainty for each sample and asymptotic Hill diversity across sample sizes, we computed the standard deviation of the log sample diversity or log estimated asymptotic diversity from the 50,000 random samples described above (McArdle et al. 1990). The standard deviation of the log diversity is a unitless measure that describes the expected proportional variability of the data, and is numerically and conceptually similar to the coefficient of variation. Because diversities are always positive, sd(log(diversity)) is an appropriate variability measure(McArdle et al. 1990, Gaston and McArdle 1994).

Assessing confidence intervals for Hill diversity estimates

For each combination of sample size and simulated abundance distribution, we took 50,000 empirical samples. For each sample, we then generated an “augmented” sample, and resampled this 2000 times. The p-value we computed was the rank the true value (mean sample diversity in the case of sample diversity, or true diversity in the case of asymptotic diversity) would have, were it among the 2000 bootstrap samples. In other words, we determined how many of the bootstrapped samples had sample- or estimated true diversity less than or equal to the true value. We also recorded the 0.025 and 0.975 quantiles of the bootstrap distributions, as these define the 95% CI (Chao and Jost 2015). (fig 5).

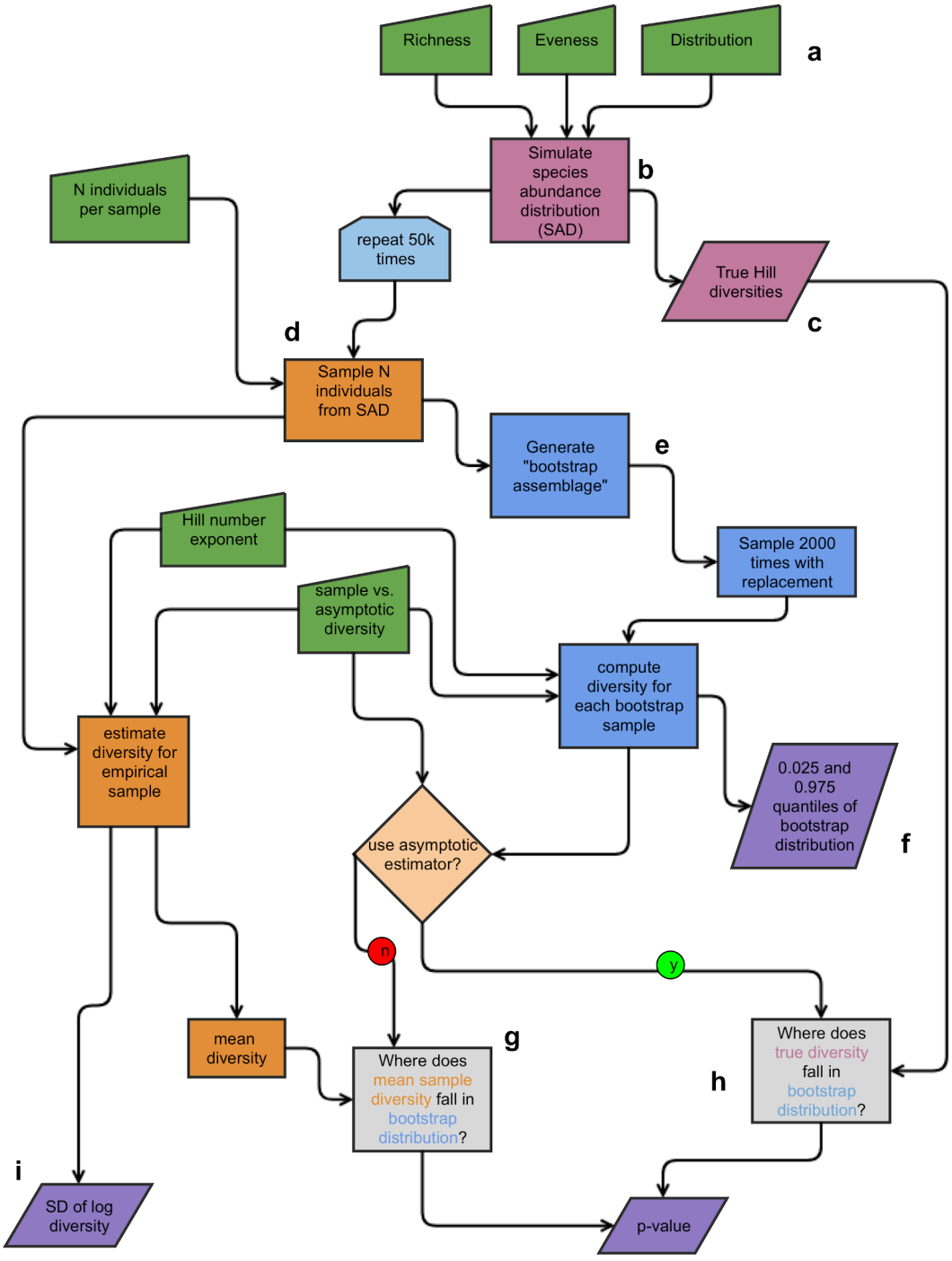

Richness
Eveness
Distribution
a
Simulate species abundance distribution (SAD)
b
N individuals per sample
repeat 50k times
True Hill diversities
c
d
Sample N individuals from SAD
Generate "bootstrap assemblage"
e
Hill number exponent
Sample 2000 times with replacement
sample vs. asymptotic diversity
compute diversity for each bootstrap sample
estimate diversity for empirical sample
0.025 and 0.975 quantiles of bootstrap distribution
f
use asymptotic estimator?
n
y
mean diversity
g
Where does mean sample diversity fall in bootstrap distribution?
h
Where does true diversity fall in bootstrap distribution?
i
SD of log diversity
p-value

Figure 5. Flowchart for simulation and analysis methods. (a) simulations took as inputs the true richness, the true evenness (and thus the true Hill-Simpson diversity, Chao and Ricotta 2019), and a distributional shape. (b) We wrote an R function that selected the optimal shape parameter for the continuous distribution and diversity constraints given by (a). (c) The true Hill diversities of the complete species abundance distribution were therefore known at the time of simulation. (d) We repeatedly took random samples of fixed size from each simulated species abundance distribution, and computed the observed diversity for the sample as well as the asymptotic diversity estimate based on the sample. (e) we used the methods proposed by Chao and Jost 2015 to generate an augmented species frequency distribution based on each sample. CI are generated by randomly resampling (bootstrapping) this distribution. We compute diversities for each bootstrap sample as well. Thus, for each of the 50,000 samples for a given species abundance distribution and sample size, we generated 1 augmented distribution and bootstrapped it 2000 times. (f) The bootstrap samples are ranked by their diversity, and the 0.025 and 0.975 quantiles of these ordered samples are equivalent to the upper and lower bounds of the 95% confidence interval. (g) We find the rank of the expected sample diversity within the bootstrap diversities, which is the p-value for the true parameter value. In the presence of ties, which are especially likely for richness, we sample all ranks compatible with that diversity. These p-values are then used to generate checkplots. (h) We use a similar procedure to estimate p-values for asymptotic diversity estimates. However, the population parameter of interest, in this case, is the true diversity, which may not be the expected value of the estimator under random sampling. (i) for each combination

of species abundance distribution, sample size, Hill number, and estimate type (i.e. sample or asymptotic), we compute the standard deviation of the log diversity of each of the 50,000 samples drawn from the true SAD.

To assess the validity of the Chao and Jost 2015 confidence intervals, we created a slugplot and a checkplot for each combination of sample size, species abundance distribution, and Hill diversity, for both sample diversities and asymptotic diversity estimates.

RESULTS

How much sampling variability is there in sample Hill diversity and asymptotic Hill diversity estimators?

The relative sampling variability of the Hill diversity estimates we examined depended on evenness and estimation method. For sample diversity, and often for asymptotic diversity estimates, Hill-Simpson diversity had the greatest sampling variability, and richness the least (fig 6). While this result may be counterintuitive, we note that the sampling variability of Hill diversities is not typically compared in a unitless manner. For a given community, richness is always greater than Hill-Shannon and Hill-Simpson, and so if all had the same proportional variability, the variance would be greatest for richness. Hill-Shannon diversity almost always has less proportional variability than Hill-Simpson diversity. As evenness increases, the sampling variability of all Hill diversity decreases, but it decreases most slowly for Hill-

Simpson diversity. Thus, for all but the most skewed possible communities, sample Hill-Simpson diversity has the greatest sampling uncertainty at all sample sizes.

Whereas differences in evenness and distributional shape affected results, doubling richness from 100 to 200 species did not qualitatively affect results, and thus we display results for only a representative set of the 200-species communities. Full results can be found in Appendix 2.

For the asymptotic diversity estimators, sampling uncertainty was initially very high, and always exceeded the variability we saw with sample diversity. Note that because our measure of variability was unitless, this is not simply because the asymptotic diversity estimates had greater mean values. Unlike for sample diversity, the ranking of sampling variability for asymptotic Hill diversity estimates was more consistent, with asymptotic richness estimators always more variable than asymptotic Hill-Shannon or Hill-Simpson estimators at small for all but the most even of communities (fig 7).

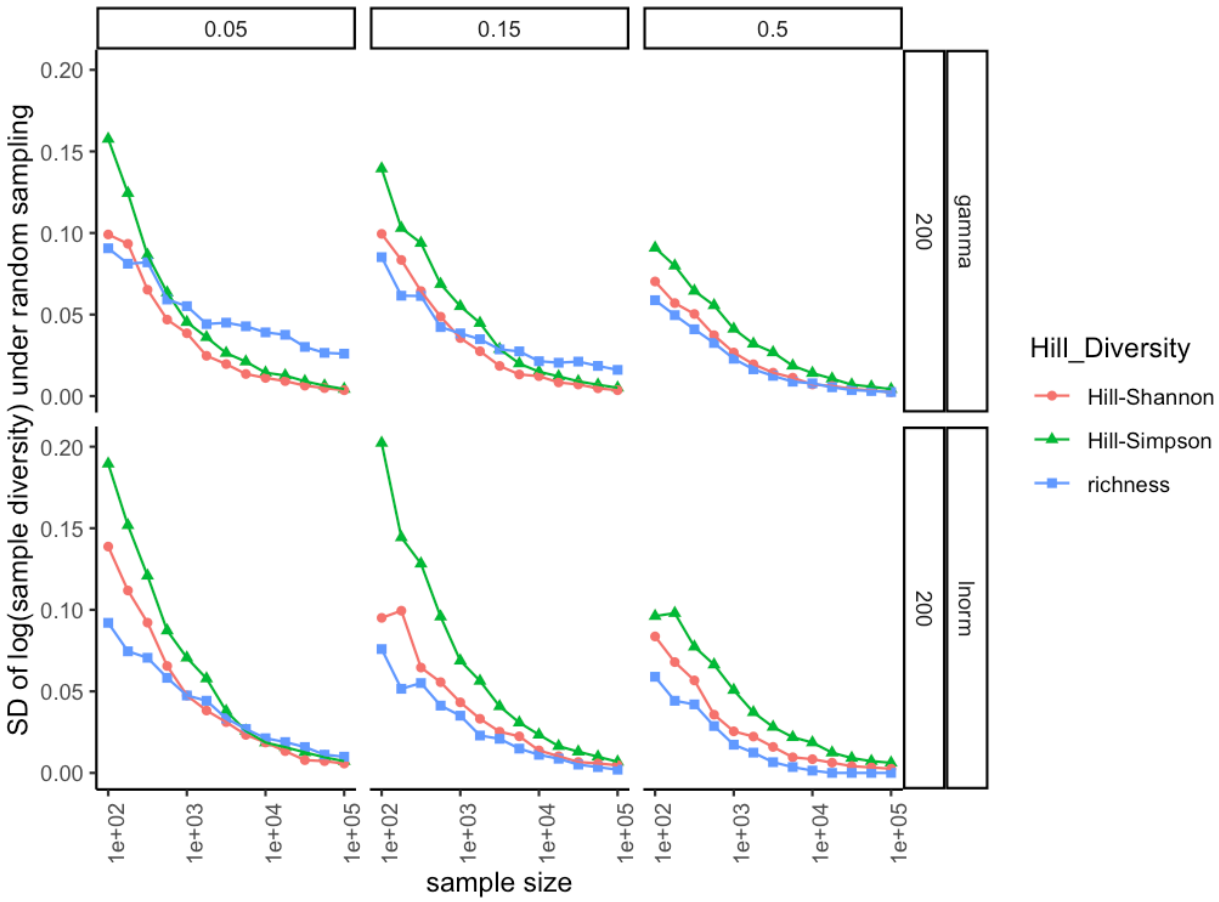


Figure 6. As community evenness increases, the sampling uncertainty in sample diversity decreases. However, it decreases fastest for richness, and slowest for Hill-Simpson diversity, which had the greatest sampling uncertainty for the species abundance distributions we simulated. Sample Hill-Shannon diversity typically had less sampling variability than Hill-Simpson diversity, but if communities were sufficiently uneven, this pattern could be reversed. Columns indicate the evenness of the simulated species abundance distribution, and with one row for the gamma shaped SADs and one row for the log-normally shaped SADs.

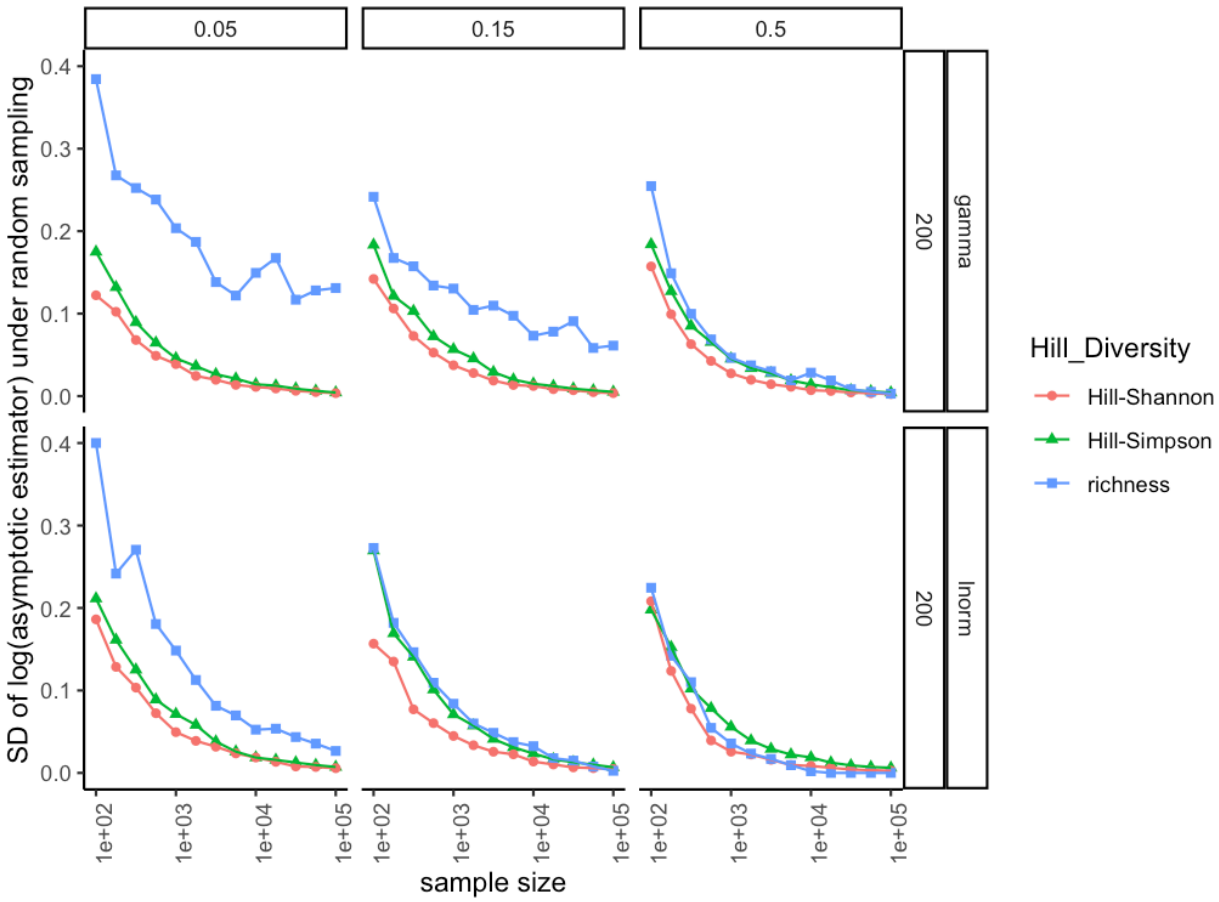
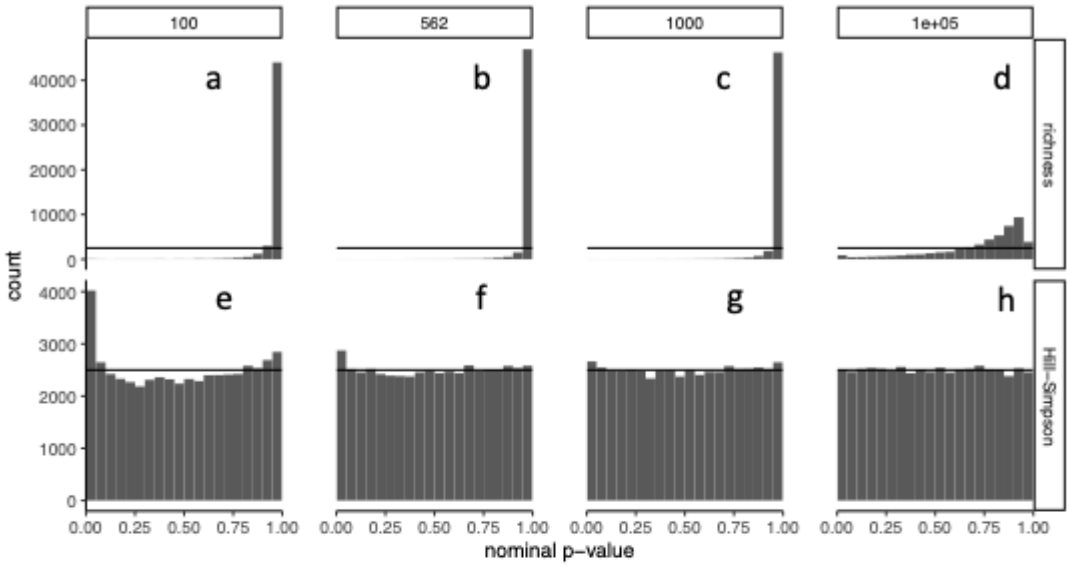


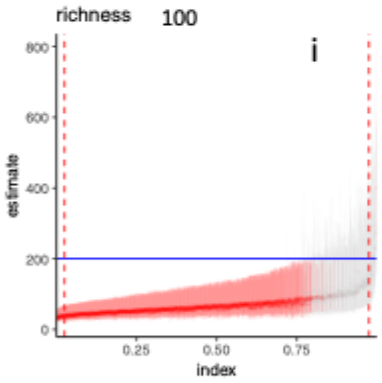
Figure 7. Sampling variability in asymptotic estimators is high for all Hill diversities at low sample sizes. For all the species abundance distributions we examined, it was lower asymptotic for Hill-Shannon diversity estimates than asymptotic Hill-Simpson diversity estimates. For more even species abundance distributions, all asymptotic Hill number estimates had similar sampling uncertainty, uncertainty rapidly grew for asymptotic richness estimates as evenness declined.

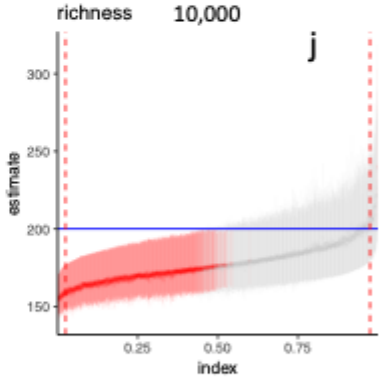
Do proposed uncertainty estimators capture observed sampling variability for sample- and asymptotic Hill diversities?

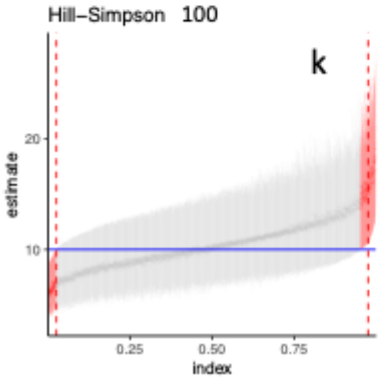
For small and moderate levels of sampling, the Chao and Jost method for constructing confidence intervals achieved nearly nominal coverage of expected sample diversity for all sample Hill diversities (Appendix 3). Nevertheless, nominal p-values derived from the bootstrap approximation were not exact, with deviations depending on both the species abundance distribution and Hill diversity exponent. (Appendix 4). However, as sample sizes increase, sample richness can equal true richness, and therefore the true sample uncertainty declines. The Chao and Jost method for estimating this uncertainty was unstable. The uncertainty estimates for sample Hill-Simpson and Hill-Shannon diversity generally improved with sample sizes and nominal p-values appeared valid, with flat checkplots and perfect-looking slugplots.

For small sample sizes, the Chao and Jost method for constructing confidence intervals for asymptotic diversity estimates performed poorly (fig 8, Appendices 5,6). For large sample sizes, the Chao and Jost method for estimating the uncertainty in asymptotic diversity improved for Hill-Shannon and Hill-Simpson, but not for richness, where the downward bias of the estimator precluded good confidence interval coverage.


100
562
1000
1e+05
a
b
c
d
e
f
g
h
richness
Hill–Simpson
count
nominal p–value


richness 100
i
estimate
index


richness 10,000
j
estimate
index


Hill–Simpson 100
k
estimate
index

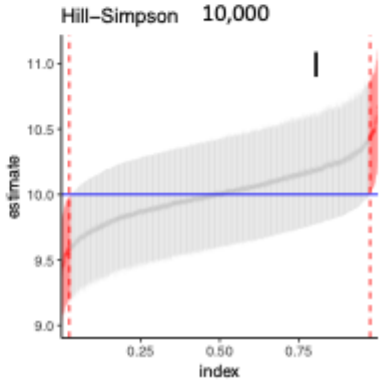

Hill–Simpson 10,000
l
estimate
index

Fig 8 Checkplots for asymptotic richness and asymptotic Hill-Simpson estimates for increasing sample sizes. For richness estimates (a-d), the estimator and its nominal 95% CI are too low for almost every sample. The downward bias (resulting in a high p-value associated with the true richness) hardly shrinks even as sample sizes increase by several orders of magnitude. For asymptotic Hill-Simpson estimates (e-h), the nominal 95% CI are anti-conservative for lower sample sizes. By 10,000 individuals for this community, the checkplot (h)is essentially flat, which means that at this sample size confidence intervals are valid. Slugplots for richness at 100 (i) and 10,000 (j) individuals show many random samples for which the entire confidence interval sits below true richness (blue line, 200 species), and none where the entire interval sits above true richness. Slugplots for asymptotic Hill-Simpson estimates at 100 individuals (k) shows the tendency to overestimate true Hill-Simpson diversity, which is entirely resolved by 10,0000 individuals (l), where nearly exactly 2.5% of the samples have CI that are too high and 2.5% are too low, consistent with the nominal 95% confidence.

DISCUSSION

We found that sampling uncertainty for sample Hill-Simpson diversity was higher than sampling uncertainty for sample Hill-Shannon diversity and richness. This was not expected based on statements prevalent in the diversity estimation literature that Hill-Simpson diversity can be estimated accurately from modest samples (Magurran and McGill 2011, Chase and Knight 2013). While the estimation *bias* (that is, a tendency to be consistently higher or

lower than the true value) for asymptotic or even observed Hill-Simpson diversity can be small, even with small sample sizes (Simpson 1949, Chao and Jost 2015), both observed Hill-Simpson diversity and asymptotic Hill-Simpson estimators remain sensitive to fluctuations in species frequencies, even in large samples. This sensitivity means that there is a bias-variance tradeoff in diversity estimation, where Hill-Shannon diversity can be estimated with more bias but less variability between samples, while Hill-Simpson diversity is estimated with less bias but often high variability between samples.

Proposed confidence intervals for sample Hill diversity capture the uncertainty arising from stochastic sampling variation in species frequencies. While the correspondence between sample diversity and the true diversity of a community is often weak (Willis 2019), Roswell et al. *in review*), the statistical tools for comparing rarefied diversity are impressively robust. Not only does statistical theory suggest such estimates should be possible (Smith and Grassle 1977), recent estimators for true species frequencies enable simulating a realistic enough species pool for bootstrap approximations to work well in many cases (Chao and Jost 2015). However, the confidence intervals for sample richness become unstable, ironically, as sample richness approaches true richness. We note that field ecologists seldom have the luxury of worrying about that scenario.

Despite progress on estimating true Hill diversity based on samples (i.e. asymptotic estimators), current tools have large uncertainty that cannot be robustly estimated. The estimators exhibit bias, and these biases are recapitulated when estimating uncertainty via the modified bootstrap proposed by Chao and Jost 2015. As sample sizes increase, the

performance of the confidence intervals for asymptotic Hill-Simpson and Hill-Shannon diversity improves, but the intervals for asymptotic richness estimates are not directly interpretable. One reason for this is that the best richness estimators are formally “lower bounds” (Chiu et al. 2014), for which a 95% confidence interval does not have a clear interpretation. The intervals suggested by (Chao and Jost 2015) do not achieve nominal coverage of true richness, even at very large sample sizes and under idealized sampling conditions, such as those we simulated here.

We introduced three new tools in this paper: a species abundance distribution simulator, slugplots, and checkplots. Our community simulation algorithm allows users to simulate a species abundance distribution based on *a priori* determination of true diversity, in contrast to selecting the parameters of an infinite distribution, and then determining true diversity based on finite samples from it (Chao and Jost 2015, May et al. 2018).

Slugplots provide a transparent visual representation of confidence interval performance for continuous statistics. These summaries clarify the direction and nature of bias or conservatism associated with nominal confidence intervals. Slugplots provide an empirical complement to checkplots, visualizing confidence interval performance on the scale of parameters of interest.

Checkplots, a type of p-value histogram, address whether nominal p-values are valid probability measures. Because valid p-values are fundamental to equal-tail confidence intervals, they are useful to researchers who are interested in expressing effect sizes with

confidence intervals, even researchers who are skeptical of using p-values for statistical inference. Any researcher able to simulate data can also generate and read checkplots to evaluate whether a statistical tool is sound, even if analytical proofs of pivotality are not possible or accessible.

In many realms of statistics, normal approximations are valid with very large sample sizes, but the approximations break down with smaller samples. Checkplots and slugplots can help researchers assess whether violating particular assumptions renders approximate p-values and confidence intervals uninterpretable, or if the approximations are acceptable for statistical practice. However, slugplots and checkplots are both information rich, and future work is needed to improve their capacity for throughput, to explore and communicate how a statistical approximation fares across a range of parameter values.

In the real world, where species have finite abundances, and species abundance distributions are not stationary, sampling variability is both greater and more difficult to estimate than in our simulations. We intentionally generated simple, replicable species abundance distributions and sampling procedures to test Chao and Jost's uncertainty estimation tools. We assume that if the estimates are inaccurate for our simulations, they cannot work well in the real world. If they did work well on our simulated data, it would still be important to determine how the predictions from Chao and Jost's methods would be affected by sampling without replacement, by changing species frequencies during sampling, or by the other real-world complexities that we assumed away.

To robustly compare true diversity, ecologists need good diversity estimators, as well as tools to accurately quantify sampling uncertainty (Willis 2019). The approach to constructing confidence intervals proposed by Chao and Jost (2015) approximates statistical uncertainty in the diversity of a finite sample. The approach is less useful for asymptotic Hill diversity estimators, because at smaller samples, when uncertainty is greatest and most important to quantify, the CI can be both biased and anti-conservative.

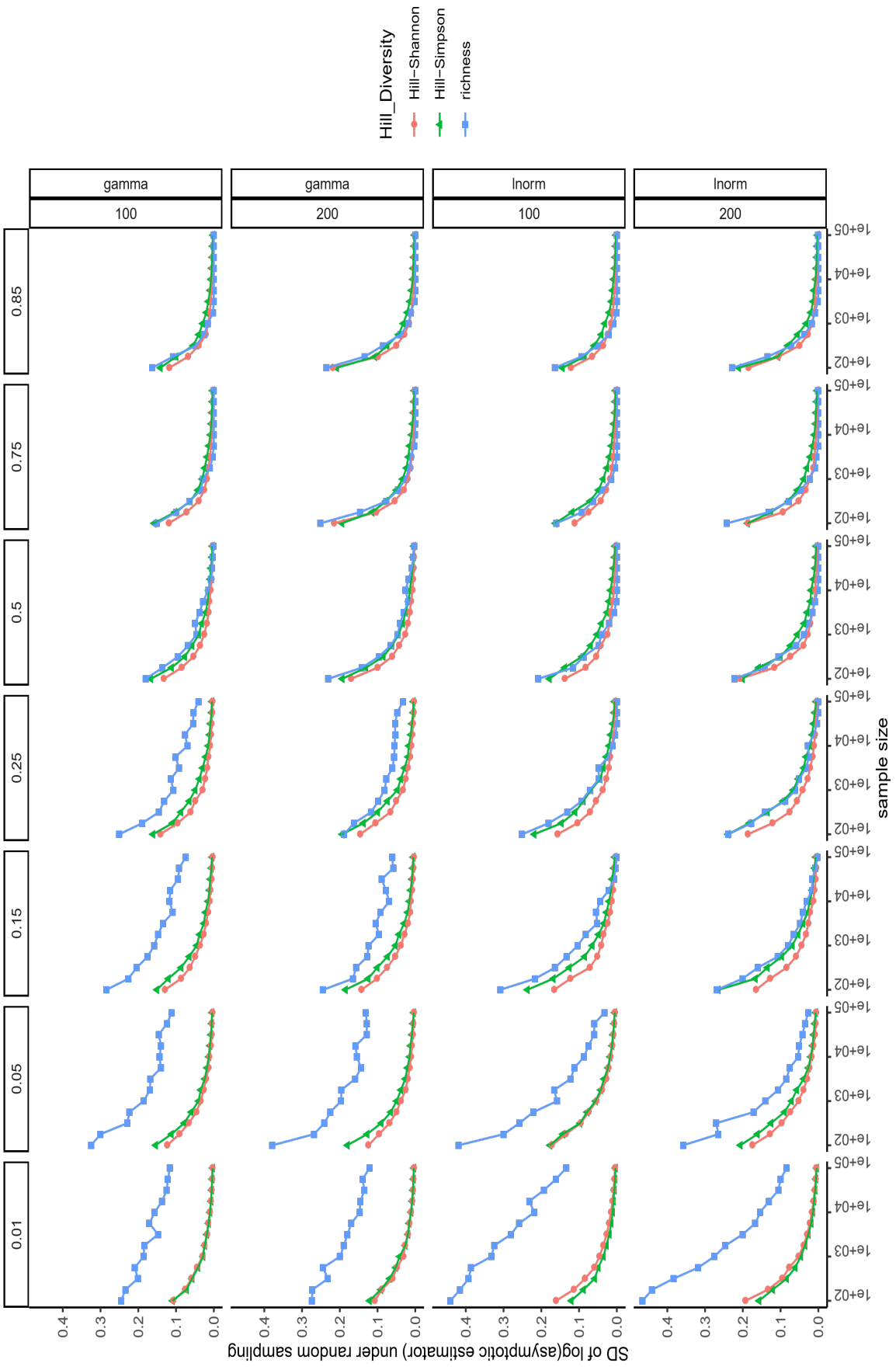
ACKNOWLEDGEMENTS

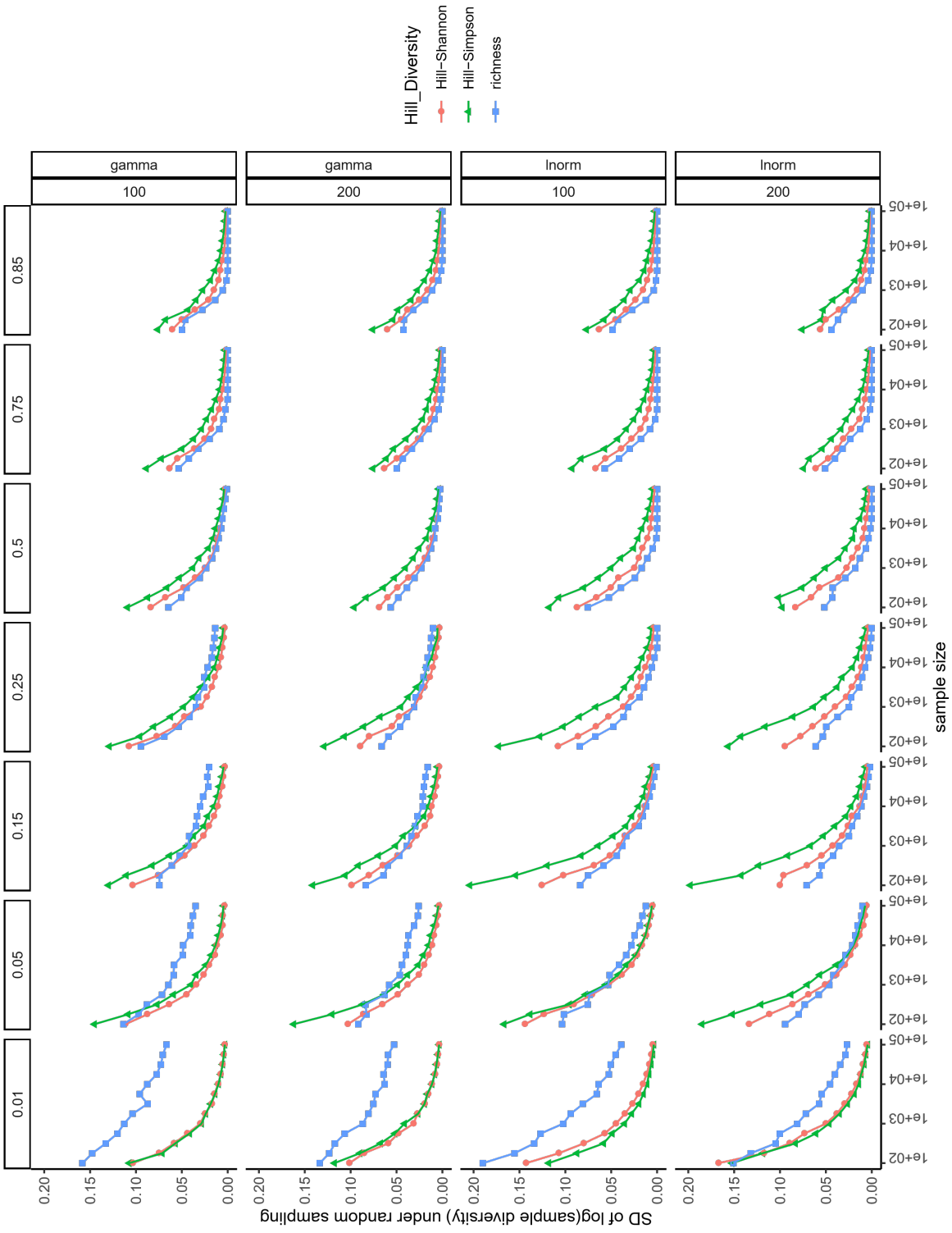
We thank Rob Muldowney for computational support, and the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here http://oarc.rutgers.edu. Thanks also to Michael Li, who invented slugplots.

Appendix 1

To simulate species abundance distributions, we selected a true richness and evenness value, and 1 of 2 distributional shape assumptions (log-normal and gamma). To parameterize evenness, we used (Hill-Simpson diversity -1)/(richness-1), which is equal to 1 with a perfectly even species-abundance distribution, and 0 when Hill-Simpson diversity reaches its minimum, 1, for any richness value (Chao and Ricotta 2019). We used the log-normal distribution because it fits observed sample species abundance distributions well (Enquist et al., n.d.; Baldridge, Harris, Xiao, & White, 2016; Matthews, Sadler, Kubota, Woodall, & Pugh, 2019). We used the gamma distribution because it fits observed species abundance distributions as well (Matthews & Whittaker, 2014; Matthews et al., 2019), but makes different assumptions about rare species, providing scope to test proposed confidence intervals (Béguinot, 2018). To fit each parametric distribution, we wrote a function in R that identifies the optimal shape parameter, given richness, Hill-Simpson diversity, and a distributional assumption (code will be on FigShare or GitHub at time of publication). The basic logic of this simulation was to use the target Simpson diversity and distributional family to set the relative abundance for a fixed number of species, where their relative abundances were given by evenly-spaced quantiles of the continuous gamma or lognormal distribution.

Appendix 2: Sampling uncertainty in sample and asymptotic Hill diversities


Hill_Diversity
Hill−Shannon
Hill−Simpson
richness
gamma
lnorm
100
200
0.01
0.05
0.15
0.25
0.5
0.75
0.85
SD of log(asymptotic estimator) under random sampling
sample size
0.0
0.1
0.2
0.3
0.4
1e+02
1e+03
1e+04
1e+05


Hill_Diversity
Hill−Shannon
Hill−Simpson
richness
gamma
100
gamma
200
lnorm
100
lnorm
200
0.85
0.75
0.5
0.25
0.15
0.05
0.01
1e+02
1e+03
1e+04
1e+05
sample size
0.20
0.15
0.10
0.05
0.00
SD of log(sample diversity) under random sampling

Appendices 3-6 available upon request from the author:

Appendix 3: Slugplots for sample Hill diversity

Appendix 4: Checkplots for sample Hill diversity

Appendix 5: Slugplots for asymptotic Hill diversity

Appendix 6: Checkplots for asymptotic Hill diversity

BIBLIOGRAPHY

Alarcón, R. et al. 2010. Sex-dependent variation in the floral preferences of the hawkmoth Manduca sexta. - Anim. Behav. 80: 289–296.

Alberdi, A. and Gilbert, M. T. P. 2019. A guide to the application of Hill numbers to DNA-based diversity analyses. - Mol. Ecol. Resour. 19: 804–817.

Alcock, J. 1983. Male behvaiour in two bumblebees, *Bombus nevadensis auricomus* and *B. griseicollis* (Hymenoptera: Apidae). - J. Zool. 200: 561–570.

Alcock, J. 2013. Sexual selection and the mating behavior of solitary bees. - Elsevier Inc.

Alcock, J. et al. 1978. The ecology and evolution of male reproductive behaviour in the bees and wasps. - Zool. J. Linn. Soc. 64: 293–326.

Ali, I. et al. 2019. Package rstanarm. in press.

Alonso, J. C. et al. 2016. Thermal tolerance may cause sexual segregation in sexually dimorphic species living in hot environments. - Behav. Ecol. 27: 717–724.

Alroy, J. 2010. The shifting balance of diversity among major marine animal groups. - Science (80-.). 329: 1191–1194.

Alroy, J. 2017. Effects of habitat disturbance on tropical forest biodiversity. - Proc. Natl. Acad. Sci. 114: 6056–6061.

Barrows, E. M. 1976a. Mating behavior in halictine bees (Hymenoptera: Halictidae): II. microterritorial and patrolling behavior in males of Lasioglossum rohweri. - Zeitschrift Fur Tierpsychologie-Journal Comp. Ethol. 40: 377–389.

Barrows, E. M. 1976b. Mating behavior in halictine bees (Hymenoptera: Halictidae): I. patrolling and age-specific behavior in males. - J. Kansas Entomol. Soc. 49: 105–119.

Barwell, L. J. et al. 2015. Measuring beta-diversity with species abundance data. - J. Anim. Ecol.: 1112–1122.

Bascompte, J. and Jordano, P. 2014. Mutualistic networks in time and space. - In: Bascompte, J. and Jordano, P. (eds), Mutualistic networks. Princeton University Press, pp. 87–106.

Bates, D. et al. 2016. Package “lme4.” - CRAN Repos.: 113.

Beck, J. and Schwanghart, W. 2010. Comparing measures of species diversity from incomplete inventories: an update. - Methods Ecol. Evol. 1: 38–44.

Beck, C. A. et al. 2007. Sex differences in grey seal diet reflect seasonal variation in foraging behaviour and reproductive expenditure: Evidence from quantitative fatty acid signature analysis. - J. Anim. Ecol. 76: 490–502.

Bernardello, G. 2007. A systematic survey of floral nectaries. - In: Nicolson, S. W. et al. (eds), Nectaries and Nectar. Springer, pp. 19–128.

Blüthgen, N. 2010. Why network analysis is often disconnected from community ecology: A critique and an ecologist’s guide. - Basic Appl. Ecol. 11: 185–195.

Bolker, B. M. 2017. GLMM FAQ.

Bolnick, D. I. et al. 2011. Why intraspecific trait variation matters in community ecology. - Trends Ecol. Evol. 26: 183–192.

Botta-Dukát, Z. 2018. The generalized replication principle and the partitioning of functional diversity into independent alpha and beta components. - Ecography (Cop.). 41: 40–50.

Brand, P. et al. 2018. Sexual dimorphism in visual and olfactory brain centers in the perfume-collecting orchid bee Euglossa dilemma (Hymenoptera, Apidae). - J. Comp. Neurol. 526: 2068–2077.

Broadhead, G. T. et al. 2017. Diel rhythms and sex differences in the locomotor activity of

hawkmoths. - J. Exp. Biol. 220: 1472–1480.
Brose, U. et al. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. - Ecology 84: 2364–2377.
Brosi, B. J. and Briggs, H. M. 2013. Single pollinator species losses reduce floral fidelity and plant reproductive function. - Proc. Natl. Acad. Sci. 110: 13044–8.
Bruninga-Socolar, B. et al. 2016. The role of floral density in determining bee foraging behavior: a natural experiment. - Nat. Areas J. 36: 392–399.
Bullen, P. S. 2003. Handbook of means and their inequalities. - Kluwer Academic Publishers.
Butler, M. A. et al. 2007. Sexual dimorphism and adaptive radiation in Anolis lizards. - Nature 447: 202–5.
Cane, J. H. 2002. Pollinating bees (Hymenoptera: Apiformes) of U.S. alfalfa compared for rates of pod and seed set. - J. Econ. Entomol. 95: 22–27.
Cane, J. H. and Sipes, S. S. 2006. Characterizing floral specialization by bees: Analytical methods and a revised lexicon for oligolecty. - In: Plant–pollinator interactions: From specialization to generalization. in press.
Cane, J. H. et al. 2011. Pollination value of male bees: The specialist bee Peponapis pruinosa (Apidae) at Summer Squash (Cucurbita pepo). - Environ. Entomol. 40: 614–620.
Cao, Y. et al. 2007. Effects of sample standardization on mean species detectabilities and estimates of relative differences in species richness among assemblages. - Am. Nat. 170: 381–395.
Chao, A. and Jost, L. 2012. Coverage-based rarefaction and extrapolation: Standardizing samples by completeness rather than size. - Ecology 93: 2533–2547.
Chao, A. and Jost, L. 2015. Estimating diversity and entropy profiles via discovery rates of new species. - Methods Ecol. Evol. 6: 873–882.
Chao, A. and Chiu, C. H. 2016. Bridging the variance and diversity decomposition approaches to beta diversity via similarity and differentiation measures. - Methods Ecol. Evol. 7: 919–928.
Chao, A. et al. 2014a. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. - Ecol. Monogr. 84: 45–67.
Chao, A. et al. 2014b. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. - Annu. Rev. Ecol. Evol. Syst. 45: 297–324.
Chao, A. et al. 2016. R: SpadeR. in press.
Chao, A. et al. 2019. Proportional mixture of two rarefaction/extrapolation curves to forecast biodiversity changes under landscape transformation. - Ecol. Lett.: ele.13322.
Chappell, M. A. 1984. Temperature regulation and energetics of the solitary bee Centris pallida during foraging and intermale mate competition. - Physiol. Zool. 57: 215–225.
Chase, J. M. and Knight, T. M. 2013. Scale-dependent effect sizes of ecological drivers on biodiversity: Why standardised sampling is not enough. - Ecol. Lett. 16: 17–26.
Chiarucci, A. et al. 2011. Old and new challenges in using species diversity for assessing biodiversity. - Philos. Trans. R. Soc. Lond. B. Biol. Sci. 366: 2426–2437.
Chiu, C.-H. and Chao, A. 2016. Estimating and comparing microbial diversity in the presence of sequencing errors. - PeerJ 4: e1634.
Close, R. A. et al. 2018. How should we estimate diversity in the fossil record? Testing richness estimators using sampling-standardised discovery curves. - Methods Ecol. Evol. 9: 1386–1400.

Coddington, J. A. et al. 2009. Undersampling bias: The null hypothesis for singleton species in tropical arthropod surveys. - J. Anim. Ecol. 78: 573–584.
Colwell, R. K. et al. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. - J. Plant Ecol. 5: 3–21.
Cox, K. D. et al. 2017. Community assessment techniques and the implications for rarefaction and extrapolation with Hill numbers. - Ecol. Evol.: 1–14.
Dauby, G. and Hardy, O. J. 2012. Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species. - Ecography (Cop.). 35: 661–672.
De Palma, A. et al. 2015. Ecological traits affect the sensitivity of bees to land-use pressures in European agricultural landscapes. - J. Appl. Ecol. 52: 1567–1577.
Dötterl, S. et al. 2011. Behavioural plasticity and sex differences in host finding of a specialized bee species. - J. Comp. Physiol. A Neuroethol. Sensory, Neural, Behav. Physiol. 197: 1119–1126.
Durell, S. E. A. L. V. dit 2000. Individual feeding specialization in shorebirds: population consequences and conservation implications. - Biol. Rev. Camb. Philos. Soc. 75: 503–518.
Elias, J. et al. 2010. No evidence for increased extinction proneness with decreasing effective population size in a parasitoid with complementary sex determination and fertile diploid males. - BMC Evol. Biol. 10: 366.
Ellison, A. M. 2010. Partitioning diversity. - Ecology 91: 1962–1963.
Eltz, T. et al. 2007. Enfleurage, lipid recycling and the origin of perfume collection in orchid bees. - Proc. R. Soc. - B 274: 2843–2848.
Etl, F. et al. 2017. A perfume-collecting male oil bee? Evidences of a novel pollination system involving Anthurium acutifolium (Araceae) and Paratetrapedia chocoensis (Apidae, Tapinotaspidini). - Flora Morphol. Distrib. Funct. Ecol. Plants 232: 7–15.
Fliszkiewicz, M. et al. 2011. The importance of male red mason bee (Osmia rufa L.) and male bufftailed bumblebee (Bombus terrestris L.) pollination in blackcurrant (Ribes nigrum L.). - J. Hortic. Sci. Biotechnol. 86: 457–460.
Fründ, J. et al. 2016. Sampling bias is a challenge for quantifying specialization and network structure: Lessons from a quantitative niche model. - Oikos 125: 502–513.
Fryxell, D. C. et al. 2015. Sex ratio variation shapes the ecological effects of a globally introduced freshwater fish. - Proc. R. Soc. B 282: 20151970.
Gelman, A. and Hill, J. 2007. Data analysis using regression and multilevel/ hierarchical models. - Cambridge University Press.
Givens, R. P. 1978. Dimorphic foraging strategies of a salticid spider (Phidippus audax). - Ecology 59: 309–321.
Gloag, R. S. et al. 2019. Workers' sons rescue genetic diversity at the sex locus in an invasive honey bee population. - Mol. Ecol.: 0–2.
Good, I. J. and Toulmin, G. H. 1956. The number of new species and the increase in population coverage when a sample is increased. - Biometrika 77: 45–63.
Gotelli, N. J. and Colwell, R. K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurment and comparison of species richness. - Ecol. Lett. 4: 379–391.
Gotelli, N. J. and Colwell, R. K. 2011. Estimating species richness. - Biol. Divers. Front. Meas. Assess.: 39–54.
Grabchak, M. et al. 2017. The generalized Simpson's entropy is a measure of biodiversity. - PLoS One 12: 1–11.

Guillera-Arroita, G. et al. 2019. Inferring species richness using multispecies occupancy modeling: estimation performance and interpretation. - Ecol. Evol. 9: 780–792.
Haegeman, B. et al. 2013. Robust estimation of microbial diversity in theory and in practice. - ISME J. 7: 1092–1101.
Heinrich, B. 1979. “Majoring” and “Minoring” by foraging bumblebees, Bombus vagans: an experimental analysis. - Ecology 60: 245–255.
Heinrich, B. and Heinrich, M. J. E. 1983. Size and caste in temperature regulation by bumblebees. - Physiol. Zool. 56: 552–562.
Hicks, D. M. et al. 2016. Food for pollinators: quantifying the nectar and pollen resources of urban flower meadows. - PLoS One 11: 1–37.
Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. - Ecology 54: 427–432.
Horn, H. S. 1966. Measurement of “Overlap” in comparative ecological studies. - Am. Nat. 100: 419–424.
Hsieh, T. C. et al. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). - Methods Ecol. Evol. 7: 1451–1456.
Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. - Ecology 52: 577–586.
Iknayan, K. J. et al. 2014. Detecting diversity: emerging methods to estimate species diversity. - Trends Ecol. Evol. 29: 97–106.
Janzen, D. H. 1971. Euglossine bees as long-distance pollinators of tropical plants. - Science (80-.). 171: 203–205.
Jost, L. 2006. Entropy and diversity. - Oikos 113: 363–375.
Jost, L. 2009. Mismeasuring biological diversity: Response to Hoffmann and Hoffmann (2008). - Ecol. Econ. 68: 925–928.
Jost, L. et al. 2010. Partitioning diversity for conservation analyses. - Divers. Distrib. 16: 65–76.
Kang, S. et al. 2016. Hill number as a bacterial diversity measure framework with high-throughput sequence data. - Sci. Rep. 6: 38263.
Kempton, R. A. 1979. The structure of species abundance and measurement of diversity. - Biometrics 35: 307–321.
Kondratyeva, A. et al. 2019. Reconciling the concepts and measures of diversity, rarity and originality in ecology and evolution. - Biol. Rev. 94: 1317–1337.
Kraus, F. B. et al. 2009. Male flight distance and population substructure in the bumblebee Bombus terrestris. - J. Anim. Ecol. 78: 247–252.
Lande, R. et al. 2000. When species accumulation curves intersect : implications for ranking diversity using small samples. - Oikos 89: 601–605.
López-Uribe, M. M. et al. 2015. Nest suitability, fine-scale population structure and male-mediated dispersal of a solitary ground nesting bee in an urban landscape. - PLoS One 10: 1–20.
Magurran, A. E. and McGill, B. J. 2011. Biological diversity: frontiers in measurement and assessment. - Oxford University Press.
Mao, C. X. et al. 2017. On the asymptotic variance of the Chao estimator for species richness estimation. - Stat. Sin. 27: 1193–1203.
McGill, B. J. et al. 2007. Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. - Ecol. Lett. 10: 995–1015.
Melo, A. S. 2004. A critique of the use of jackknife and related non-parametric techniques to

estimate species richness. - Community Ecol. 5: 149–157.
Merlo, J. et al. 2006. A brief conceptual tutorial of multilevel analysis in social epidemiology: Using measures of clustering in multilevel logistic regression to investigate contextual phenomena. - J. Epidemiol. Community Health 60: 290–297.
Morisita, M. 1959. Measuring of interspecific association and similarity between communities. - Mem. Fac. Sci. Kyushu Univ. Ser. E 3: 65–80.
Ne'eman, G. et al. 2006. Foraging by male and female solitary bees with implications for pollination. - J. Insect Behav. 19: 383–401.
Ogilvie, J. E. and Thomson, J. D. 2015. Male bumble bees are important pollinators of a late-blooming plant. - Arthropod. Plant. Interact. 9: 205–213.
Ogilvie, J. E. and Forrest, J. R. 2017. Interactions between bee foraging and floral resource phenology shape bee populations and communities. - Curr. Opin. Insect Sci. 21: 75–82.
Ohlmann, M. et al. 2019. Diversity indices for ecological networks: a unifying framework using Hill numbers. - Ecol. Lett. 22: 737–747.
Oksanen, J. 2016. Vegan: ecological diversity. - R Doc.: 12.
Ostevik, K. et al. 2010. Pollination potential of male bumble bees (Bombus impatiens): movement patterns and pollen-transfer efficiency. - J. Pollinat. Ecol. 2: 21–26.
Parachnowitsch, A. L. et al. 2018. Evolutionary ecology of nectar. - Ann. Bot.: 247–261.
Pascarella, J. B. 2010. Pollination biology of Gelsemium sempervirens L. (Ait.) (Gelsemiaceae): do male and female Habropoda laboriosa F. (Hymenoptera, Apidae) differ in pollination efficiency? - J. Apic. Res. 49: 170–176.
Patil, G. P. and Taillie, C. 1982. Diversity as a concept and its measurement. - J. Am. Stat. Assoc. 77: 548–561.
Peet, R. K. 1974. The measurement of species diversity. - Annu. Rev. Ecol. Syst. 5: 285–307.
Pinheiro, M. et al. 2017. Flowers as sleeping places for male bees: somehow the males know which flowers their females prefer. - Arthropod. Plant. Interact. 11: 329–337.
Preston, F. W. 1948. The commonness, and rarity, of species. - Ecology 29: 254–283.
R Core Team 2018. R: a language and environment for statistical computing. in press.
Rader, R. et al. 2013. Native bees buffer the negative impact of climate warming on honey bee pollination of watermelon crops. - Glob. Chang. Biol. 19: 3103–3110.
Ramos-Jiliberto, R. et al. 2012. Topological plasticity increases robustness of mutualistic networks. - J. Anim. Ecol. 81: 896–904.
Ritchie, A. D. et al. 2016. Generalist behavior describes pollen foraging for perceived oligolectic and polylectic bees. - Environ. Entomol. 45: 909–919.
Robert, T. et al. 2016. Male bumblebees perform learning flights on leaving a flower but not when leaving their nest. - J. Exp. Biol. 220: 930–937.
Robertson, C. 1890. Flowers and Insects IV. - Bot. Gaz. 15: 79–84.
Robertson, C. 1925. Heterotropic bees. - Ecology 6: 412–436.
Rossi, B. H. et al. 2010. Sexual harassment by males reduces female fecundity in the alfalfa leafcutting bee, Megachile rotundata. - Anim. Behav. 79: 165–171.
Roswell, M. et al. 2019a. Data from: Male and female bees show large differences in floral preference. - PLoS One 14: e0214909.
Roswell, M. et al. 2019b. Male and female bees show large differences in floral preference (ME Saunders, Ed.). - PLoS One 14: e0214909.
Roubik, D. W. 1993. Tropical pollinators in the canopy and understory: Field data and theory for stratum "preferences." - J. Insect Behav. 6: 659–673.

Roulston, T. H. and Goodell, K. 2011. The role of resources and risks in regulating wild bee populations. - Annu. Rev. Entomol. 56: 293–312.
Rundlöf, M. et al. 2014. Late-season mass-flowering red clover increases bumble bee queen and male densities. - Biol. Conserv. 172: 138–145.
Rusterholtz, H.-P. and Erhardt, A. 2000. Can nectar properties explain sex-specific flower preferences in the Adonis Blue butterfly Lysandra bellargus? - Ecol. Entomol. 25: 81–90.
Rutter, M. T. and Rausher, M. D. 2004. Natural selection on extrafloral nectar production in Chamaecrista fasciculata: the costs and benefits of a mutualism trait. - Evolution (N. Y). 58: 2657–2668.
Sapir, Y. et al. 2006. Morning floral heat as a reward to the pollinators of the Oncocyclus irises. - Oecologia 147: 53–59.
Selander, R. K. 1966. Sexual dimorphism and differential niche utilization in birds. - Condor 68: 113–151.
Senapathi, D. et al. 2016. Landscape impacts on pollinator communities in temperate systems: evidence and knowledge gaps. - Funct. Ecol. in press.
Seth, M. 2017. How to implement credible 95% interval for median odds ratio using JAGS?
Shannon, C. E. and Weaver, W. 1963. The mathematical theory of communication. - Univ. Illinois Press 5: 1–131.
Sharma, N. et al. 1993. Pollination biology of some species of genus Plantago L. - Bot. J. Linn. Soc. 111: 129–138.
Sherwin, W. B. et al. 2017. Information theory broadens the spectrum of molecular ecology and evolution. - Trends Ecol. Evol. 32: 948–963.
Shine, R. 1989. Ecological causes for the evolution of sexual dimorphism: a review of the evidence. - Q. Rev. Biol. 64: 419–461.
Silvério, A. et al. 2012. Floral rewards in the tribe Sisyrinchieae (Iridaceae): Oil as an alternative to pollen and nectar? - Sex. Plant Reprod. 25: 267–279.
Simpson, E. H. 1949. Measurement of diversity. - Nature 163: 688–688.
Smith, W. and Grassle, J. F. 1977. Sampling properties of a family of diversity measures. - Biometrics 33: 283–292.
Somanathan, H. et al. 2017. Visual adaptations for mate detection in the male carpenter bee Xylocopa tenuiscapa. - PLoS One in press.
Southwood, R. 1978. Ecological methods: with particular reference to the study of insect populations. - Chapman and Hall.
Spiesman, B. J. et al. 2017. Bumble bee colony growth and reproduction depend on local flower dominance and natural habitat area in the surrounding landscape. - Biol. Conserv. 206: 217–223.
Start, D. and De Lisle, S. 2018. Sexual dimorphism in a top predator (Notophthalmus viridescens) drives aquatic prey community assembly. - Proc. R. Soc. B in press.
Stone, G. N. 1995. Female foraging responses to sexual harassment in the solitary bee *Anthophora plumipes*. - Anim. Behav. 50: 405–412.
Stone, G. N. et al. 1999. Windows of opportunity and the temporal structuring of foraging activity in a desert solitary bee. - Ecol. Entomol. 24: 208–221.
Straub, L. et al. 2016. Neonicotinoid insecticides can serve as inadvertent insect contraceptives. - Proc. R. Soc. - B in press.
Streinzer, M. et al. 2013. Sexual dimorphism in the olfactory system of a solitary and a eusocial bee species. - J. Comp. Neurol. 521: 2742–2755.

Sutter, L. et al. 2017. Enhancing plant diversity in agricultural landscapes promotes both rare bees and dominant crop-pollinating bees through complementary increase in key floral resources. - J. Appl. Ecol. 54: 1856–1864.
Temeles, E. J. et al. 2010. Evolution of sexual dimorphism in bill size and shape of hermit hummingbirds (Phaethornithinae): a role for ecological causation. - Philos. Trans. R. Soc. B 365: 1053–1063.
Tuomisto, H. 2010. A consistent terminology for quantifying species diversity? Yes, it does exist. - Oecologia 164: 853–860.
Tur, C. et al. 2014. Downscaling pollen-transport networks to the level of individuals. - J. Anim. Ecol. 83: 306–317.
Ulrich, Y. et al. 2009. Flexible social organization and high incidence of drifting in the sweat bee, Halictus scabiosae. - Mol. Ecol. 18: 1791–1800.
Vaudo, A. D. et al. 2014. Bumble bees exhibit daily behavioral patterns in pollen foraging. - Arthropod. Plant. Interact. 8: 273–283.
Vaudo, A. D. et al. 2016. Macronutrient ratios in pollen shape bumble bee (Bombus impatiens) foraging strategies and floral preferences. - Proc. Natl. Acad. Sci. 113: E4035–E4042.
Ward, K. et al. 2014. Streamlined bee monitoring protocol for assessing pollinator habitat. - Xerces Soc. Invertebr. Conserv.
Williams, N. M. and Lonsdorf, E. V. 2018. Selecting cost-effective plant mixes to support pollinators. - Biol. Conserv. 217: 195–202.
Williamson, M. and Gaston, K. J. 2005. The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. - J. Anim. Ecol. 74: 409–422.
Willis, A. D. 2019. Rarefaction, alpha diversity, and statistics. - Front. Microbiol. in press.
Willmer, P. 2011. Pollination and floral ecology. - Princeton University Press.
Willmer, P. G. and Stone, G. N. 2004. Behavioral, ecological, and physiological determinants of the activity patterns of bees. - Adv. Study Behav. 34: 347–466.
Wolf, S. and Moritz, R. F. A. 2014. The pollination potential of free-foraging bumblebee (Bombus spp.) males (Hymenoptera: Apidae). - Apidologie 45: 440–450.
Zhang, Z. 2016. Statistical implications of Turing's formula. - John Wiley and Sons, Inc.
Zwolak, R. 2018. How intraspecific variation in seed-dispersing animals matters for plants. - Biol. Rev. 93: 897–913.