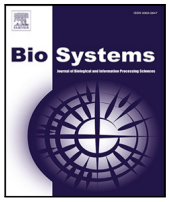
BioSystems 237 (2024) 105153

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems



Mathematical constraints on a family of biodiversity measures via connections with Rényi entropy

Theodore D. Gress, Noah A. Rosenberg *

Department of Biology, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Keywords:
Biodiversity
Hill numbers
Population genetics
Rényi entropy
Shannon entropy

ABSTRACT

The Hill numbers are statistics for biodiversity measurement in ecological studies, closely related to the Rényi and Shannon entropies from information theory. Recent developments in the mathematics of diversity in the setting of population genetics have produced mathematical constraints that characterize how standard measures depend on the highest-frequency class in a discrete probability distribution. Here, we apply these constraints to diversity statistics in ecology, focusing on the Hill numbers and the Rényi and Shannon entropies. The mathematical bounds can shift perspectives on the diversities of communities, in that when upper and lower bounds on Hill numbers are evaluated in a classic butterfly example, Hill numbers that are initially larger in one community switch positions—so that associated normalized Hill numbers are instead smaller than those of the other community. The new bounds hence add to the tools available for interpreting a commonly used family of statistics for ecological data.

1. Introduction

Consider an ecological community containing up to n distinct species, in which the relative abundance of species i is p_i, for $i = 1, 2, \ldots, n$, with $\sum_{i=1}^{n} p_i = 1$. The measurement of the diversity of such a community is a basic task in ecology; a diversity measure has high values if the community contains many species with nontrivial and comparable abundances and low values if it contains few species, one of which predominates.

Many statistics have been proposed for use in diversity measurements (Magurran, 2004; Jost, 2006; Gotelli and Ellison, 2013). Focusing on α-diversity, a diversity concept for a single community (as opposed to concepts involving multiple communities), *species richness* is simply the number of species in the community, or $R = \sum_{i=1}^{n} p_i^0$. *Simpson's index*, $S = \sum_{i=1}^{n} p_i^2$, gives the probability that two individuals randomly drawn from the community, with replacement, are from the same species; S is most naturally viewed as a similarity measure, and $1 - S$ and $1/S$ are diversity measures. *Shannon entropy* is $H = -\sum_{i=1}^{n} p_i \log p_i$, where the natural logarithm is most frequently used (Magurran, 2004, p. 107); an information-theoretic perspective is natural, as a high-diversity probability distribution of species abundances can be regarded as analogous to a complex message, requiring "more bits" to describe than a low-diversity distribution.

In work that has since become influential (Jost, 2006; Chao and Ricotta, 2019; Roswell et al., 2021), Hill (1973) described how standard indices can be placed in a single family of diversity indices related through Rényi's definition of generalized entropy (Rényi, 1961). Again using the natural logarithm, for $q \geqslant 0$ and $q \neq 1$, *Rényi entropy* can be written

$$H_q = \frac{1}{1-q} \log \sum_{i=1}^{n} p_i^q. \tag{1}$$

For $q \geqslant 0$ and $q \neq 1$, the *Hill number of order* q is a diversity index defined by

$$D_q = \left(\sum_{i=1}^{n} p_i^q \right)^{\frac{1}{1-q}}. \tag{2}$$

In other words, $D_q = e^{H_q}$.

For $q = 0$, D_q is simply the species richness n. For $q = 2$, D_q is $1/S$, the reciprocal of Simpson's index. In the limit as $q \to 1$, D_q approaches e^H for Shannon entropy H. The value of q determines the relative weight that common and rare species have in the diversity measurement. An increase in q increases the weight that the most common species has in the calculation, with $q < 1$ emphasizing rare species and $q > 1$ emphasizing common species—and the limits $q = 0$ weighting all species equally, irrespective of their abundances, and $q \to \infty$ weighting only the most abundant species.

Hill numbers have a number of features as diversity measures (Chao et al., 2014a; Leinster, 2021; Roswell et al., 2021). For example, as can be seen by supposing that each $p_i = \frac{1}{n}$ to obtain $D_q = n$ in Eq. (2),

* Corresponding author.
E-mail address: noahr@stanford.edu (N.A. Rosenberg).

https://doi.org/10.1016/j.biosystems.2024.105153
Received 14 September 2023; Received in revised form 24 January 2024; Accepted 9 February 2024
Available online 28 February 2024

the Hill numbers behave sensibly in that when all species have the same abundance, their values increase with the number of species, irrespective of the value of q. Second, for fixed numbers of species, Hill numbers increase toward a maximum achieved when all species abundances are equal. Third, the ability to modulate the value of q in a framework based on the Hill numbers enables a researcher to change the emphasis on common and rare species while maintaining the same conceptual structure for diversity measurement. Finally, that the Hill numbers subsume long-used standard measures of diversity facilitates their interpretation.

With recent attention to the mathematical and statistical properties of diversity measures (Jost, 2006; Ellison, 2010; Chao et al., 2014a), the Hill numbers have become prominent. Studies have proposed them in specific contexts, such as for tracking diversity in communities over time (Maturo and Di Battista, 2018) and for measuring diversities of species interactions (Ohlmann et al., 2019). The Hill numbers have also been employed with newer forms of data beyond counts of individuals—such as acoustic measurements (Luypaert et al., 2022), and with increasing frequency, the DNA sequence data that appear in metagenomics (Kang et al., 2016; Ma and Li, 2018; Alberdi and Gilbert, 2019). Alongside these extensions focused on applications, further mathematical analysis has contributed to advancing the understanding of the behavior of the Hill numbers (Jost, 2007; Chao et al., 2014b; Chiu et al., 2014; Leinster, 2021).

Throughout the long history of diversity measurement studies in ecology—with the co-option of Shannon and Rényi entropies from information theory serving as prominent examples—the mathematical study of ecological diversity measures has benefited from parallels with investigations of analogous problems in other types of data. One such data type is allelic data in population genetics, in which, instead of distinct species, discrete allelic categories are tabulated. Considering counts of observations of alleles, each allele has a frequency that is analogous to a species relative abundance. Mathematically, the two contexts are the same in evaluating diversity from a vector of nonnegative entries whose sum is 1. The analogy of species diversity and allelic diversity enables mathematical properties of diversity measures in one context to inform the use of equivalent measures in the other.

Because of the way they are formulated, the Hill numbers derive a number of their mathematical properties from properties of Rényi entropies. In a mathematical analysis of Rényi entropy in the context of population genetics, Aw and Rosenberg (2018) discerned the upper and lower bounds on Rényi entropy as functions of the frequency of the most frequent allele in an allele frequency distribution. This analysis relied on the study of a quantity termed by Aw and Rosenberg (2018) the *α-homozygosity*, or $J^{(\alpha)} = \sum_{i=1}^{n} p_i^{\alpha}$, and it obtained a variety of results useful for interpreting α-homozygosity. If we replace the notation α with q to avoid confusion with the α in α-diversity, then the formula for q-homozygosity lies within the Hill number formula in Eq. (2). $D_q = (J^{(q)})^{\frac{1}{1-q}}$.

Here, we use results obtained for q-homozygosity and Rényi entropy in the context of population genetics to understand mathematical bounds on Hill numbers. We examine the dependence of the Hill numbers on the frequency of the most abundant species, considering the effect on this dependence of the value of q. We illustrate our mathematical results in an example taken from rainforest butterfly communities, suggesting how the bounds can be used to enhance the interpretation of Hill numbers in ecology.

2. Results

2.1. Homozygosity, Rényi entropy, and Shannon entropy

We make use of connections between the Hill numbers and a family of measures in genetics termed α-homozygosities and α-heterozygosities by Aw and Rosenberg (2018). For parallelism with the Hill numbers, we henceforth call the quantities of Aw and Rosenberg (2018) q-homozygosities and q-heterozygosities.

In population genetics, a standard genetic diversity measure for a genetic locus in a population is *expected heterozygosity*, equal to $1 - \sum_{i=1}^{n} p_i^2$ for a set of nonnegative allele frequencies p_i with $\sum_{i=1}^{n} p_i = 1$; the "expected" in "expected heterozygosity" refers to its computation from the allele frequency distribution as the probability that two draws from the distribution are distinct, as opposed to a separate computation termed "observed heterozygosity," obtained from empirical measurement of the fraction of heterozygotes seen in actual individuals. A complementary homogeneity measure is *expected homozygosity*, $\sum_{i=1}^{n} p_i^2$.

In the same way that Hill (1973) suggested a family of diversity measures with different exponents for use in diversity computations in ecology, Aw and Rosenberg (2018) examined a corresponding family of diversity measures in population genetics. For an allele frequency vector $\mathbf{p}$ and an exponent $q > 1$, generalized diversity and similarity measures can be written

$$H^{(q)}(\mathbf{p}) = 1 - \sum_{i=1}^{n} p_i^q \tag{3}$$

$$J^{(q)}(\mathbf{p}) = \sum_{i=1}^{n} p_i^q, \tag{4}$$

where we show the argument $\mathbf{p}$ here to emphasize that these quantities are computed from allele frequency vectors. Dropping the argument (and also dropping the term "expected"), standard heterozygosity and homozygosity are $H^{(2)}$ and $J^{(2)}$, respectively.

Aw and Rosenberg (2018) showed that if the frequency of the most frequent allele at a locus is equal to $M < 1$, then the q-homozygosities and q-heterozygosities are constrained within the open unit interval. Specifically, each q-homozygosity and each q-heterozygosity has an upper bound that is less than 1 and a lower bound that is greater than 0; these bounds depend on M. Aw and Rosenberg (2018) found that the upper and lower bounds on q-homozygosity decrease as q increases. Furthermore, they proved that the area between the upper and lower bounds of q-homozygosity on the unit square, representing the set of possible ordered pairs $(M, J^{(q)})$, decreases as q increases. Initially, Aw and Rosenberg (2018) did not suppose that the number of distinct alleles was known; this initial assumption amounts to an assumption that the number of distinct alleles is arbitrarily large. Aw and Rosenberg (2018) then obtained a stricter result, assuming that the number of distinct alleles is fixed. That result, their Corollary 3.13, appears here as Result 1 in Box I.

Result 1 assumes $q > 1$, so that q-homozygosity and q-heterozygosity will lie within $[0, 1]$. Aw and Rosenberg (2018) also obtained entirely analogous bounds on the Rényi entropy for $0 < q < 1$ and $1 < q$. Their Corollary 3.19 states the bounds of the Rényi entropy for a fixed frequency of the most frequent allele and fixed number of alleles, and it is restated in Box I as Result 2.

The Shannon entropy and its bounds can be interpreted as a limit of the Rényi entropy as $q \to 1$, or H_1. Aw and Rosenberg (2018) directly derived bounds for the case of the Shannon entropy. Result 3 in Box I provides the bounds on Shannon entropy for a fixed frequency of the most frequent allele and fixed number of alleles, restated from Corollary 3.16 of Aw and Rosenberg (2018).

2.2. Connections to Hill numbers

At this point, we exploit the connections among q-homozygosities, Rényi entropies, and Hill numbers. We can convert among q-homozygosities $J^{(q)}$, Rényi entropies H_q, and Hill numbers D_q by using transformations

$$H_q = \frac{1}{1-q}\log J^{(q)} \tag{5}$$

$$J^{(q)} = e^{(1-q)H_q} \tag{6}$$

$$D_q = e^{H_q} \tag{7}$$

$$H_q = \log D_q \tag{8}$$

$$D_q = (J^{(q)})^{\frac{1}{1-q}} \tag{9}$$

$$J^{(q)} = (D_q)^{1-q}. \tag{10}$$

In particular, for fixed q, the transformation in Eq. (7) is monotonically increasing in Rényi entropy H_q and the transformation in Eq. (9) is monotonically decreasing in q-homozygosity $J^{(q)}$.

Because the Hill number for q is obtained from the corresponding Rényi entropy by a monotonic transformation, we can conclude that the Hill number D_q obtains its minimum and maximum at the same places as the Rényi entropy H_q. Changing alleles to species, we obtain a new result, labeled Result 4 in Box I, which describes the upper and lower bounds of the Hill number D_q as functions of the frequency of the most abundant species in a community with a fixed number of species.

With our mathematical results describing bounds on the statistics now established, we explore the behavior of the statistics in relation to the bounds, viewing the results in terms of species abundances. For parallelism, although Result 2 for Rényi entropy and Result 4 for Hill numbers apply on larger domains, we restrict attention to $q > 1$, the domain that produces Result 1 for q-homozygosity.

2.3. Analysis of the bounds

Fig. 1 graphs the q-homozygosity as a function of the frequency of the most abundant species M, following Result 1. In all panels, both the upper bound and the lower bound increase as M increases. The maximal q-homozygosity is 1, and it occurs when a single species has frequency 1. Within a panel, as the maximal permissible number of distinct species increases, the lower bound becomes less strict; for a fixed number of distinct species I, the lower bound is defined only on the interval $[\frac{1}{I}, 1]$. The upper bound does not depend on the maximal number of distinct species.

In Fig. 1, as the exponent q in q-homozygosity increases, the upper and lower bound curves decrease. Comparing across values of q, we see that for low values of q, a visible difference exists between the upper and lower bounds, indicating an influence of species other than the most abundant species on q-homozygosity. As q increases to a large value, however, the upper and lower bounds nearly coincide, so that q-homozygosity is largely determined by the frequency of the most abundant species.

Fig. 2 displays the bounds on Rényi entropy as a function of M, following Result 2. As Rényi entropy is, unlike q-homozygosity, a diversity measure, a number of patterns reverse those for q-homozygosity. The minimal Rényi entropy is zero and occurs when a single species has frequency 1. As the maximal permissible number of distinct species increases, the lower bound of Rényi entropy remains unchanged, while the upper bound increases. Comparing panels in Fig. 2, as the exponent q increases, the upper and lower bounds decrease. For large q, as was seen with q-homozygosity, the upper and lower bounds are close together, so that Rényi entropy is closely predicted by the frequency of the most abundant species.

The patterns for Hill numbers, shown in Fig. 3 according to Result 4, closely follow those seen for Rényi entropy, but with larger numbers and more separation between upper and lower bounds. The upper and lower bounds decrease with an increasing frequency of the most abundant species. The minimal value of the Hill numbers is 1, occurring at $M = 1$. The upper bound on Hill numbers depends on the number of permissible species, whereas the lower bound does not. For increasing q, comparing panels of Fig. 3, the upper and lower bounds on the Hill numbers decrease, and they approach one another. As was seen for q-homozygosity and for Rényi entropy, for large q, the Hill number with exponent q depends primarily on the frequency of the most abundant species.

2.4. Example data set

To illustrate the application of the bounds to empirical data, we reexamined data analyzed by Jost (2006) based on samples reported by Devries and Walla (2001). Devries and Walla (2001) collected five years of butterfly abundance data in a rainforest region of eastern Ecuador, sampling butterflies in both the canopy and the understory. Within the data of Devries and Walla (2001), Jost (2006) restricted attention to species with at least 8 captures, considering 11,696 observations across 74 species. The total number of canopy observations is 5774 in $I_c = 56$ species, and the total number of understory observations is 5922 in $I_u = 65$ species. Jost (2006) studied the diversity in the two communities, calculating a variety of quantities to compare them. We make use of these data to illustrate the utility of bounds on diversity statistics.

One feature of the two communities is that they differ substantially in the frequency of the most abundant species. The canopy community has 1882 observations of its most abundant species, *Historis acheronta*, for a frequency $M_c \approx 0.3259$. In the understory, the most abundant species is *Nessaea hewitsoni*, with 984 observations and frequency $M_u \approx 0.1662$.

We calculated Hill numbers for values of q from 1.01 to 10 for the canopy and understory butterfly communities, considering upper and lower bounds from Result 4 with the values of I and M associated with the two communities. Fig. 4A shows Hill numbers plotted as a function of q for the canopy butterfly community. The upper and lower bounds use the $I_c = 56$ butterfly species found in that habitat and the frequency $M_c \approx 0.3259$ of the most abundant species. Fig. 4B is an analogous plot for the understory butterfly community ($I_u = 65$, $M_u \approx 0.1662$).

Comparing Figs. 4A and 4B, it appears that the understory has a more diverse community. The understory has 9 more butterfly species than the canopy, $I_u - I_c = 9$. The Hill number at $q = 2$, the inverse of the Simpson index, is 12.48 for the understory, but only 6.63 for the canopy. The limit of the Hill number as $q \to 1$, the exponential of the Shannon entropy, is 19.74 for the understory and 14.02 for the canopy.

If the Hill numbers for the canopy and understory are considered with their upper and lower bounds, however, then the comparison changes. Fig. 4C graphs the position of the Hill numbers in relation to their upper and lower bounds, or

$$\frac{D_q - L_q(M,I)}{U_q(M,I) - L_q(M,I)}, \tag{11}$$

where $U_q(M,I)$ and $L_q(M,I)$ represent the upper and lower bounds on the Hill number D_q as functions of the frequency M of the most abundant species and the number of species I, as obtained in Result 4. For this normalized quantity, a value of 1 indicates that the Hill number lies at its maximum, and a value of 0 gives the minimum. Note that even with monotonicity in q for the upper and lower bounds and for the Hill numbers of a dataset itself, the normalized value need not be monotonic in q.

In Fig. 4C, the canopy has a larger value than the understory for the normalized Hill number. This result suggests that given both the frequency of the most abundant species and the species richness in the two habitats, the canopy appears to be more diverse. Thus, a computation that takes into account bounds on Hill numbers provides an alternative view of the comparison of the two communities.

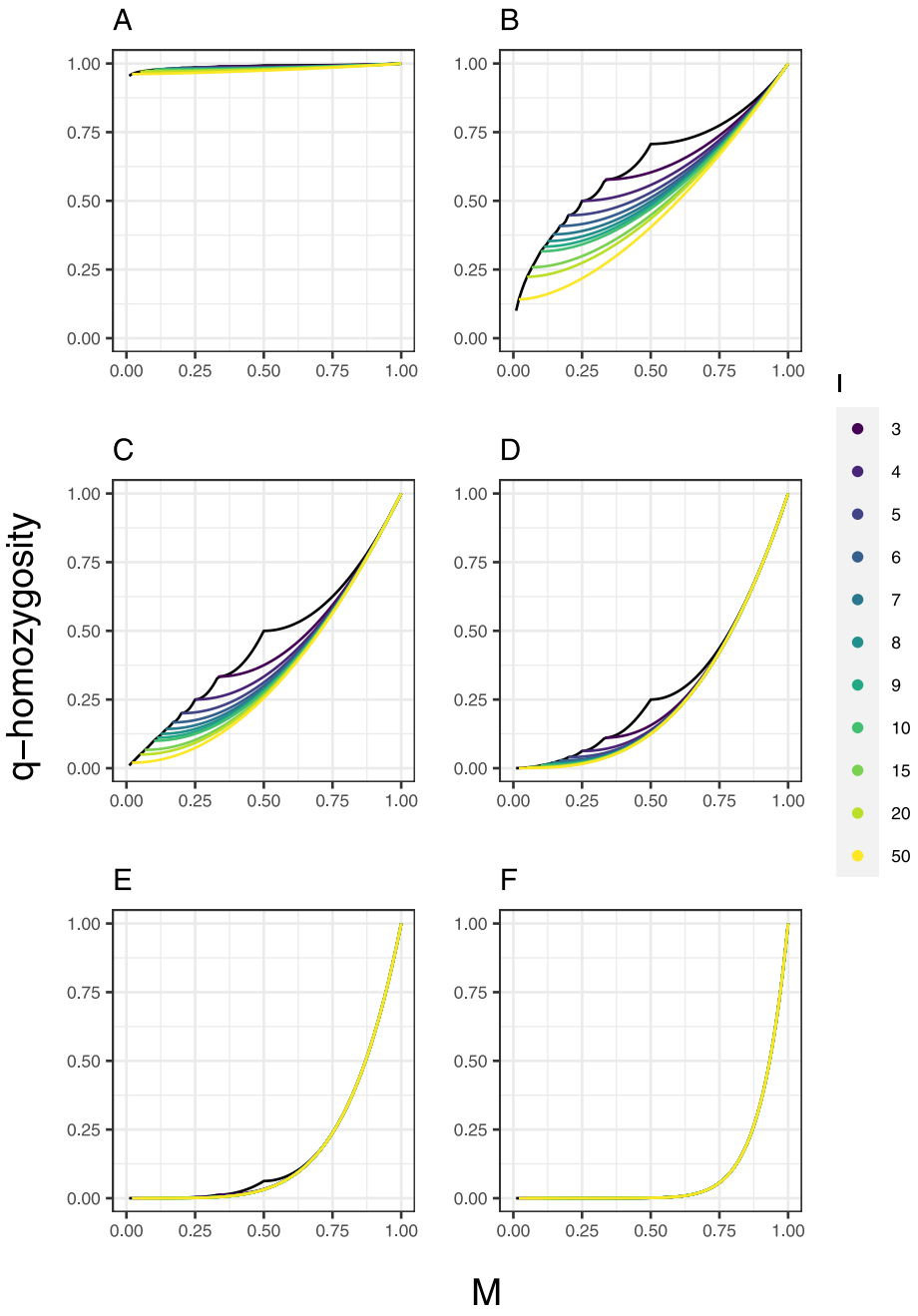


Fig. 1. Lower and upper bounds of q-homozygosity for different choices of q and the maximal number of distinct alleles I. Relying on Result 1 in Box I, each panel represents a different value of q. (A) $q = 1.01$. (B) $q = 1.5$. (C) $q = 2$. (D) $q = 3$. (E) $q = 5$. (F) $q = 10$. Lower bounds are shown for different values of the maximal number of alleles I (from top to bottom, $I = 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 50$). The upper bound is the same irrespective of the number of alleles (black), except that it is defined only for $M \geqslant \frac{1}{I}$. For large q, the upper and lower bounds are very close together.

3. Discussion

Making use of analytical bounds derived previously for q-homozygosity and Rényi entropy, we have obtained bounds on Hill numbers as functions of the exponent q, the number of species, and the frequency of the most abundant species. These bounds limit the possible values of the Hill numbers in relation to both the frequency of this most abundant species in a sample from a community and the number of species present. In an example with data from rainforest butterfly communities, we have shown that viewing the diversity statistics with the bounds can influence a comparison of two communities. The bounds provide context for interpretation of Hill numbers, enabling a normalization that allows a researcher to understand the diversity of a community relative to minimum and maximum values given the frequency of the most abundant species (Eq. (11)). The understanding that such bounds exist can add to a researcher's toolbox of principles for interpreting the values of Hill numbers seen in empirical studies—and the associated normalized statistics can add to the repertoire of statistics that can be usefully computed.

Statistics for measuring diversity from data on species abundances in communities are assessed for a variety of properties, including

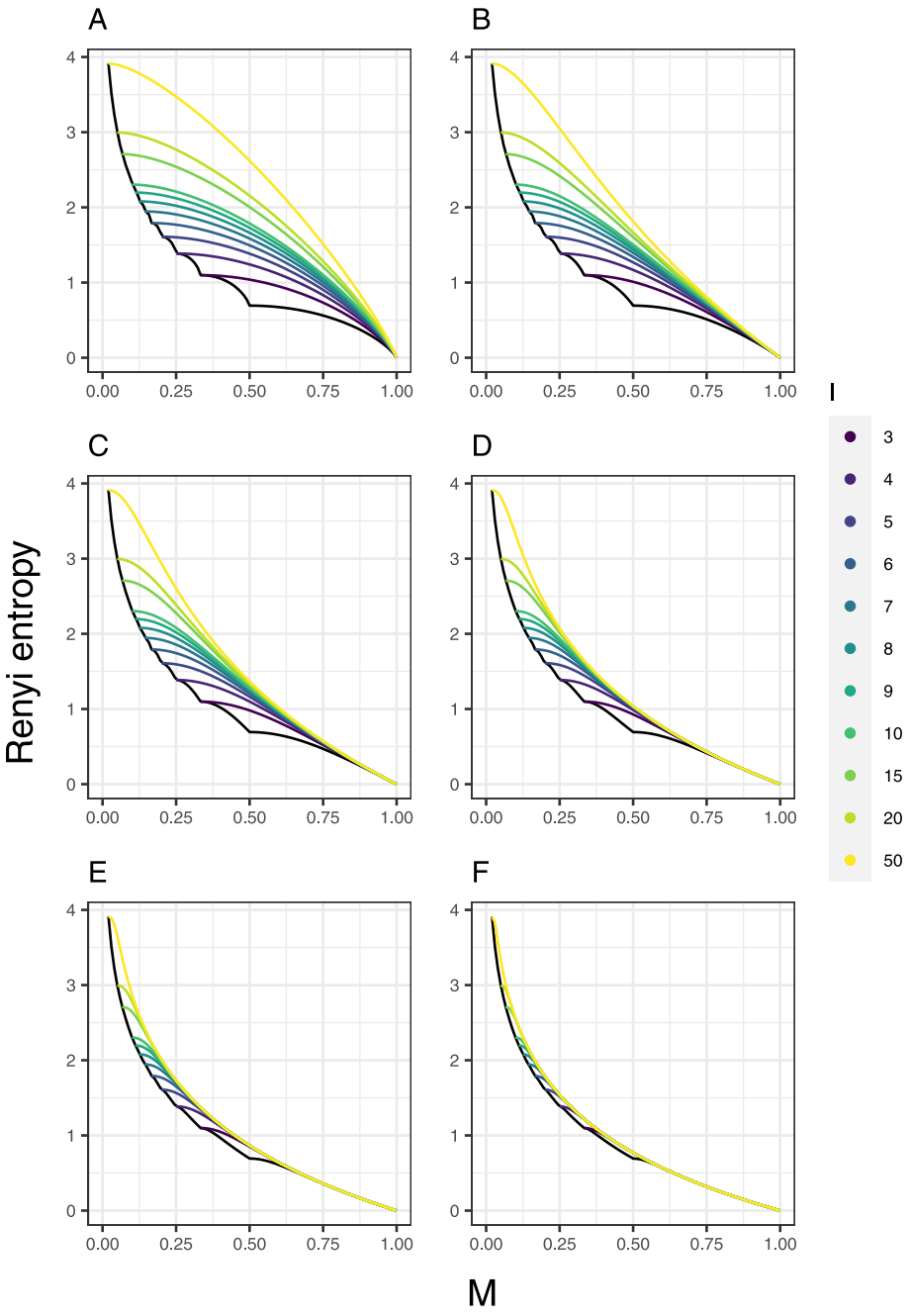


Fig. 2. Lower and upper bounds of Rényi entropy for different choices of q and the maximal number of distinct species I. Relying on Result 2 in Box I, each panel represents a different value of q. (A) $q = 1.01$. (B) $q = 1.5$. (C) $q = 2$. (D) $q = 3$. (E) $q = 5$. (F) $q = 10$. Upper bounds are shown for different values of the maximal number of species I (from bottom to top, $I = 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 50$). The lower bound is the same irrespective of the number of species (black), except that it is defined only for $M \geqslant \frac{1}{I}$. For large q, the upper and lower bounds are very close together.

mathematical simplicity, relationship to historical data sets, and intuitive alignment with researchers' understanding of factors affecting communities (e.g. Pielou, 1975; Magurran, 2004; Jost, 2006; Liu et al., 2007; Leinster and Cobbold, 2012; Gotelli and Ellison, 2013; Chao et al., 2014a; Leinster, 2021; Roswell et al., 2021). Hill numbers have recently been regarded as some of the most useful diversity statistics, as they provide a relatively simple and mathematically desirable family of values that both encompasses the Simpson and Shannon statistics and extends beyond them. Indeed, use of a trajectory of the Hill numbers as a function of q—a Hill number "diversity profile"—is a way to more fully explain the diversity of a community than use of single statistics (Leinster and Cobbold, 2012; Chao and Jost, 2015; Roswell et al., 2021). Our bounds further enable assessments of Hill numbers in relation to a q-dependent maximum. The normalization in Eq. (11) follows a similar form to other normalizations that consider global extrema of diversity statistics over all possible abundance vectors (Jost, 2010), but instead with extrema considered only over vectors with a fixed value for the frequency of the most abundant species.

How do the new bounds augment the long-available potential to modulate q in application of the Hill numbers? Modulation of q already

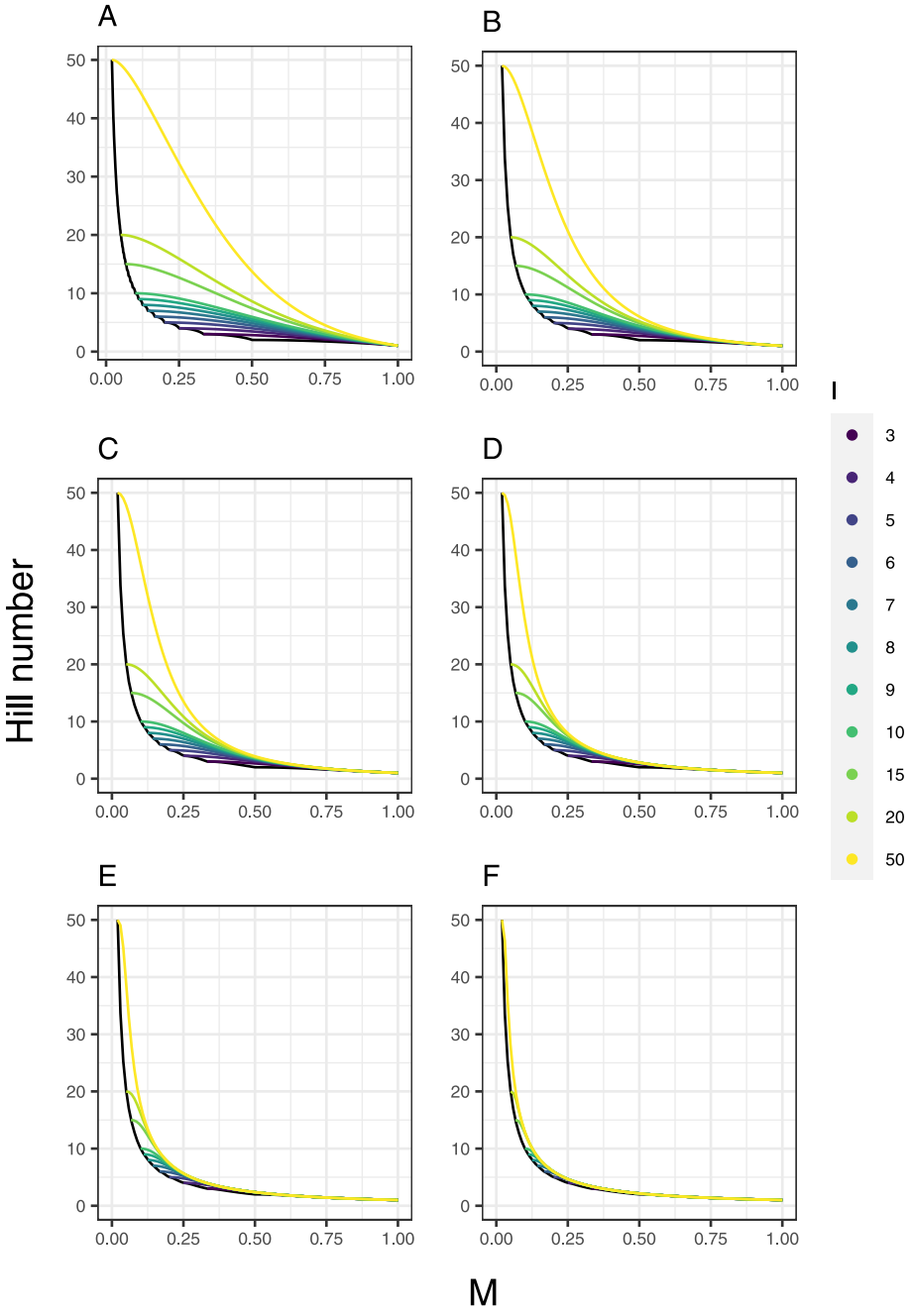


Fig. 3. Lower and upper bounds of Hill numbers for different choices of q and the maximal number of distinct species I. Relying on Result 4 in Box I, each panel represents a different value of q. (A) $q = 1.01$. (B) $q = 1.5$. (C) $q = 2$. (D) $q = 3$. (E) $q = 5$. (F) $q = 10$. Upper bounds are shown for different values of the maximal number of species I (from bottom to top, $I = 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 50$). The lower bound is the same irrespective of the number of species (black), except that it is defined only for $M \geqslant \frac{1}{I}$. For large q, the upper and lower bounds are very close together.

tunes the effect of the largest abundance M on the value of a diversity statistic: with larger q, the Hill number increasingly depends primarily on M. The normalization in Eq. (11) achieves something different: it captures effects of the remaining species, and it does so in the interval $[0, 1]$. For any choice of q, the normalization evaluates a sense of diversity if the remainder of the community is considered without its most abundant species. Consider two communities with the same dominant species, say, Community 1 with abundances $(0.8, 0.2, 0, 0, 0)$ and Community 2 with $(0.8, 0.05, 0.05, 0.05, 0.05)$. For large q, the Hill numbers are nearly identical—$D_3 = 1.387$ for Community 1 and 1.397 for Community 2. A difference in communities is detectable mainly by noting that the pattern of change in Hill numbers with q differs for the two communities: for example, the difference between communities is larger at $q = 2$, with $D_2 = 1.470$ for Community 1 and 1.538 for Community 2. The relative change in Hill numbers between communities is modest, so that their quite different features—excluding the dominant species—are not easily seen. Using the bounds, however, the difference widens: across the full range of values of q, Community 2

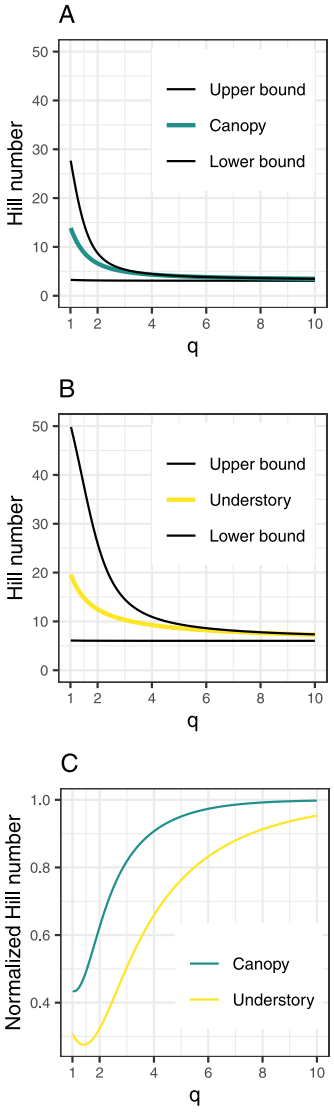


Fig. 4. Hill numbers for two communities. Panels (A) and (B) plot Hill numbers for two communities as a function of q, alongside the lower and upper bounds on the Hill number as a function of the frequency M of the most abundant species and the number of species I. The value of q ranges from 1.01 to 10. The Hill number for the data is plotted alongside bounds taken from Result 4. (A) Canopy community. (B) Understory community. (C) The proximity of the Hill number to its upper bound, calculated for two communities using $[D_q - L_q(M,I)]/[U_q(M,I) - L_q(M,I)]$, as a function of q (Eq. (11)). $L_q(M,I)$ and $U_q(M,I)$ represent the lower and upper bounds on the Hill number D_q as functions of the frequency M of the most abundant species and the number of species I, as obtained in Result 4 in Box I.

always has maximal diversity given $M = 0.8$ and number of species $I = 5$, with Eq. (11) equal to 1, whereas Community 1 always lies at the minimum, with Eq. (11) equal to 0. The normalization thus clarifies relative diversities in sets of rarer species, as obtained by discarding the species that is most abundant.

Indeed, the butterfly data that we have considered, reanalyzing data from Jost (2006), generates an interesting finding. In particular, the understory has higher values for the Hill numbers—but when normalizing Hill numbers based on their maximal values given the frequency of the most abundant species, instead the canopy has higher values. Like Jost (2006), Leinster and Cobbold (2012) had used Hill numbers to reanalyze rainforest butterfly data—in this case from DeVries et al. (1997). They noted the higher values in the understory for values of $q \lesssim 3.1$, writing *"if one is principally concerned with dominance, the population in the canopy appears to be fractionally more diverse, but from any other point of view, there is more diversity in the understory."* In other words, if the diversity index emphasizes the most frequent species by using a high value of q, then the canopy is more diverse—but otherwise, the understory is more diverse. We offer a view in which, in a similar data set, when the frequency of the most abundant species is taken into consideration, the canopy appears to be more diverse than the understory across a range of values of q (Fig. 4C). The point of this observation is not to establish which community is in fact more diverse, but to note that our bounds enable a researcher to control for the contribution to diversity of the most frequent species, and that when accounting for its abundance, the relative order of the diversity values might be transposed. In the analysis of Leinster and Cobbold (2012), modulation of q, which alters the influence of the highest abundance on the Hill numbers, does not transpose the two communities for $q \lesssim 3.1$; it is the normalization, which in effect compares the sets of remaining species as communities in themselves, that changes the relative order of the diversity profiles.

A contribution of our study is its further development of a connection between results concerning diversity indices in population genetics and mathematically related diversity indices in ecology. As both areas rely on measurement of abundances in categorical data for the purpose of understanding features of biological diversity, the measures that they can employ are similar and often mathematically identical. Insights obtained in one domain can be imported into the other, such as in the influence of the ecological use of the information-theoretic Shannon entropy on the choice of Lewontin (1972) to employ this measure for characterizing human population-genetic diversity at different geographical scales (Winther, 2021; Novembre, 2022).

Many studies of population-genetic statistics have been providing mathematical bounds that constrain the values of the statistics in relation to allele frequencies (e.g. Rosenberg and Jakobsson, 2008; Maruki et al., 2012; Alcala and Rosenberg, 2019, 2022). Results for such statistics can be used to obtain corresponding results for related statistics in the ecological context. Indeed, more general results than those we have used can be employed for additional statistics that satisfy the necessary mathematical requirements (Aw and Rosenberg, 2018); results fixing frequencies other than the most abundant can also be obtained (Morrison and Rosenberg, 2023). Further exploration of mathematical bounds can potentially contribute to decisions on researchers' choices of ecological statistics and to improving the interpretation of statistics that are chosen.

CRediT authorship contribution statement

Theodore D. Gress: Conceptualization, Investigation, Writing – original draft, Writing – review & editing. **Noah A. Rosenberg:** Conceptualization, Investigation, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Maike Morrison for helpful comments. We acknowledge National Institutes of Health grant R01 HG005855 for support.

Suppose **p** is a vector of length I of nonnegative numbers that sum to 1, arranged in decreasing order with $p_i \geqslant p_j$ if $i < j$. Suppose the largest element in **p** is fixed at $p_1 = M$, with M in $[\frac{1}{I}, 1]$. In this setting, the following four results hold. Note that the vector **p** can be taken to represent a vector of species relative abundances or a vector of allele frequencies.

Result 1 (*Bounds on q-homozygosity*). *For $q > 1$, q-homozygosity satisfies*

$$M^q + \frac{(1-M)^q}{(I-1)^{q-1}} \leqslant J^{(q)} \leqslant \lfloor M^{-1} \rfloor M^q + (1 - \lfloor M^{-1} \rfloor M)^q.$$

Equality with the upper bound occurs if and only if $p_i = M$ for $1 \leqslant i \leqslant K-1$, $p_K = 1-(K-1)M$, and $p_i = 0$ for $i > K$, where $K = \lceil M^{-1} \rceil$. Equality with the lower bound occurs if and only if $p_i = \frac{1-M}{I-1}$ for $2 \leqslant i \leqslant I$.

Result 2 (*Bounds on Rényi entropy*). *For $q > 0$, $q \neq 1$, Rényi entropy satisfies*

$$H_q \geqslant \frac{1}{1-q} \log \left[\lfloor M^{-1} \rfloor M^q + (1 - \lfloor M^{-1} \rfloor M)^q \right]$$

$$H_q \leqslant \frac{1}{1-q} \log \left[M^q + (I-1) \left(\frac{1-M}{I-1} \right)^q \right].$$

Equality with the upper bound occurs if and only if $p_i = \frac{1-M}{I-1}$ for $2 \leqslant i \leqslant I$. Equality with the lower bound occurs if and only if $p_i = M$ for $1 \leqslant i \leqslant K-1$, $p_K = 1-(K-1)M$, and $p_i = 0$ for $i > K$, where $K = \lceil M^{-1} \rceil$.

Result 3 (*Bounds on Shannon entropy*). *Shannon entropy, obtained as the $q \to 1$ limit of Rényi entropy, satisfies*

$$H_1 \geqslant \lfloor M^{-1} \rfloor M \log \frac{1}{M} + (1 - \lfloor M^{-1} \rfloor M) \log \left(\frac{1}{1 - \lfloor M^{-1} \rfloor M} \right)$$

$$H_1 \leqslant M \log \frac{1}{M} + (1-M) \log \left(\frac{1-M}{I-1} \right).$$

Equality with the upper bound occurs if and only if $p_i = \frac{1-M}{I-1}$ for $2 \leqslant i \leqslant I$. Equality with the lower bound occurs if and only if $p_i = M$ for $1 \leqslant i \leqslant K-1$, $p_K = 1-(K-1)M$, and $p_i = 0$ for $i > K$, where $K = \lceil M^{-1} \rceil$.

Result 4 (*Bounds on Hill numbers*). *The Hill numbers satisfy the following results.*
(i) For $q > 0$, $q \neq 1$,

$$D_q \geqslant \left[\lfloor M^{-1} \rfloor M^q + \left(1 - \lfloor M^{-1} \rfloor M \right)^q \right]^{\frac{1}{1-q}}$$

$$D_q \leqslant \left[M^q + (I-1) \left(\frac{1-M}{I-1} \right)^q \right]^{\frac{1}{1-q}}.$$

(ii) For $q = 1$,

$$D_1 \geqslant e^{\lfloor M^{-1} \rfloor M \log \frac{1}{M} + (1 - \lfloor M^{-1} \rfloor M) \log \left(\frac{1}{1 - \lfloor M^{-1} \rfloor M} \right)}$$

$$D_1 \leqslant e^{M \log \frac{1}{M} + (1-M) \log \left(\frac{1-M}{I-1} \right)}.$$

(iii) Equality with the upper bound occurs if and only if $p_i = \frac{1-M}{I-1}$ for $2 \leqslant i \leqslant I$. Equality with the lower bound occurs if and only if $p_i = M$ for $1 \leqslant i \leqslant K-1$, $p_K = 1-(K-1)M$, and $p_i = 0$ for $i > K$, where $K = \lceil M^{-1} \rceil$.

Box I. Mathematical bounds on diversity statistics for vectors of species relative abundances.

References

Alberdi, A., Gilbert, M.T.P., 2019. A guide to the application of Hill numbers to DNA-based diversity analyses. Mol. Ecol. Resour. 19, 804–817. http://dx.doi.org/10.1111/1755-0998.13014.

Alcala, N., Rosenberg, N.A., 2019. G'_{ST}, Jost's D, and F_{ST} are similarly constrained by allele frequencies: a mathematical, simulation, and empirical study. Mol. Ecol. 28, 1624–1636. http://dx.doi.org/10.1111/mec.15000.

Alcala, N., Rosenberg, N.A., 2022. Mathematical constraints on F_{ST}: multiallelic markers in arbitrarily many populations. Phil. Trans. R. Soc. B 377, 20200414. http://dx.doi.org/10.1098/rstb.2020.0414.

Aw, A.J., Rosenberg, Noah A., 2018. Bounding measures of genetic similarity and diversity using majorization. J. Math. Biol. 77, 711–737. http://dx.doi.org/10.1007/s00285-018-1226-x.

Chao, A., Chiu, C.-H., Jost, L., 2014a. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. Annu. Rev. Ecol. Evol. Syst. 45, 297–324. http://dx.doi.org/10.1146/annurev-ecolsys-120213-091540.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K., Ellison, A.M., 2014b. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol. Monographs 84, 45–67. http://dx.doi.org/10.1890/13-0133.1.

Chao, A., Jost, L., 2015. Estimating diversity and entropy profiles via discovery rates of new species. Methods Ecol. Evol. 6, 873–882. http://dx.doi.org/10.1111/2041-210X.12349.

Chao, A., Ricotta, C., 2019. Quantifying evenness and linking it to diversity, beta diversity, and similarity. Ecology 100, e02852. http://dx.doi.org/10.1002/ecy.2852.

Chiu, C.-H., Jost, L., Chao, A., 2014. Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. Ecol. Monographs 84, 21–44. http://dx.doi.org/10.1890/12-0960.1.

DeVries, P.J., Murray, D., Lande, R., 1997. Species diversity in vertical, horizontal, and temporal dimensions of a fruit-feeding butterfly community in an Ecuadorian rainforest. Biol. J. Linnean Soc. 62, 343–364. http://dx.doi.org/10.1006/bijl.1997.0155.

Devries, P.J., Walla, T.R., 2001. Species diversity and community structure in neotropical fruit-feeding butterflies. Biol. J. Linnean Soc. 74, 1–15. http://dx.doi.org/10.1111/j.1095-8312.2001.tb01372.x.

Ellison, A.M., 2010. Partitioning diversity. Ecology 91, 1962–1963. http://dx.doi.org/10.1890/09-1692.1.

Gotelli, N.J., Ellison, A.M., 2013. A Primer of Ecological Statistics, second ed. Sinauer, Sunderland, MA.

Hill, M.O., 1973. Diversity and evenness: A unifying notation and its consequences. Ecology 54, 427–432. http://dx.doi.org/10.2307/1934352.

Jost, L., 2006. Entropy and diversity. Oikos 113, 363–375. http://dx.doi.org/10.1111/j.2006.0030-1299.14714.x.

Jost, L., 2007. Partitioning diversity into independent alpha and beta components. Ecology 88, 2427–2439. http://dx.doi.org/10.1890/06-1736.1.

Jost, L., 2010. The relation between evenness and diversity. Diversity 2, 207–232. http://dx.doi.org/10.3390/d2020207.

Kang, S., Rodrigues, J.L.M., Ng, J.P., Gentry, T.J., 2016. Hill number as a bacterial diversity measure framework with high-throughput sequence data. Sci. Rep. 6, 38263. http://dx.doi.org/10.1038/srep38263.

Leinster, T., 2021. Entropy and Diversity: The Axiomatic Approach. Cambridge University Press, Cambridge.

Leinster, T., Cobbold, C.A., 2012. Measuring diversity: the importance of species similarity. Ecology 93, 477–489. http://dx.doi.org/10.1890/10-2402.1.

Lewontin, R.C., 1972. The apportionment of human diversity. Evol. Biol. 6, 381–398. http://dx.doi.org/10.1007/978-1-4684-9063-3_14.

Liu, C., Whittaker, R.J., Ma, K., Malcolm, J.R., 2007. Unifying and distinguishing diversity ordering methods for comparing communities. Popul. Ecol. 49, 89–100. http://dx.doi.org/10.1007/s10144-006-0026-0.

Luypaert, T., Bueno, A.S., Masseli, G.S., Kaefer, I.L., Campos-Cerqueira, M., Peres, C.A., Haugaasen, T., 2022. A framework for quantifying soundscape diversity using Hill numbers. Methods Ecol. Evol. 13, 2262–2274. http://dx.doi.org/10.1111/2041-210X.13924.

Ma, Z., Li, L., 2018. Measuring metagenome diversity and similarity with Hill numbers. Mol. Ecol. Resour. 18, 1339–1355. http://dx.doi.org/10.1111/1755-0998.12923.

Magurran, A.E., 2004. Measuring Biological Diversity. Blackwell Science, Malden, MA.

Maruki, T., Kumar, S., Kim, Y., 2012. Purifying selection modulates the estimates of population differentiation and confounds genome-wide comparisons across single-nucleotide polymorphisms. Mol. Biol. Evol. 29, 3617–3623. http://dx.doi.org/10.1093/molbev/mss187.

Maturo, F., Di Battista, T., 2018. A functional approach to Hill's numbers for assessing changes in species variety of ecological communities over time. Ecol. Indic. 84, 70–81. http://dx.doi.org/10.1016/j.ecolind.2017.08.016.

Morrison, M.L., Rosenberg, N.A., 2023. Mathematical bounds on Shannon entropy given the abundance of the *i*th most abundant taxon. J. Math. Biol. 87, 76. http://dx.doi.org/10.1007/s00285-023-01997-3.

Novembre, J., 2022. The background and legacy of Lewontin's apportionment of human genetic diversity. Philos. Trans. R. Soc. B 377, 20200406. http://dx.doi.org/10.1098/rstb.2020.0406.

Ohlmann, M., Miele, V., Dray, S., Chalmandrier, L., O'Connor, L., Thuiller, W., 2019. Diversity indices for ecological networks: a unifying framework using Hill numbers. Ecol. Lett. 22, 737–747. http://dx.doi.org/10.1111/ele.13221.

Pielou, E.C., 1975. Ecological Diversity. Wiley, New York.

Rényi, A., 1961. On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics. University of California Press, Berkeley, pp. 547–561.

Rosenberg, N.A., Jakobsson, M., 2008. The relationship between homozygosity and the frequency of the most frequent allele. Genetics 179, 2027–2036. http://dx.doi.org/10.1534/genetics.107.084772.

Roswell, M., Dushoff, J., Winfree, R., 2021. A conceptual guide to measuring species diversity. Oikos 130, 321–338. http://dx.doi.org/10.1111/oik.07202.

Winther, R.G., 2021. Lewontin 1972. In: Lorusso, L., Winther, R.G. (Eds.), Remapping Race in a Global Context. Routledge, London, pp. 9–47.