et al. (2006), Gotelli and Colwell (2011), Gotelli and Chao (2013) and Chao and Chiu (2016) for various applications. For two assemblages, shared species richness plays an important role in assessing assemblage overlap and forms a basis for constructing various types of beta diversity and (dis)similarity measures, such as the classic Sørensen and Jaccard indices (Colwell and Coddington 1994, Magurran 2004, Jost et al. 2011, Gotelli and Chao 2013). Compared with estimating species richness in a single assemblage, the estimation of shared species richness, taking undetected species into account, has received relatively little attention; see Chao and Chiu (2012) for a review.

In traditional measures of species diversity, all species (or taxa at some other rank) are considered to be equally distinct from one another. Species differences can be based directly on their evolutionary histories, either in the form of taxonomic classification or well-supported phylogenetic trees. A rapidly growing literature addresses phylogenetic diversity metrics and related (dis)similarity measures; see Cavender-Bares et al. (2012) for a review. A widely used phylogenetic metric is Faith's (1992) *PD* (phylogenetic diversity), which is defined as the sum of the branch lengths of a phylogenetic tree connecting all

For species diversity, we apply the Good-Turing formula to intuitively derive an estimator of the number of undetected species in an assemblage. The resulting estimator turns out to be the Chao (1984) non-parametric lower bound. The two-assemblage generalized formula yields Pan et al.'s (2009) lower bound of the number of undetected shared species when a sample of individuals is taken from each of two assemblages. For phylogenetic diversity, the unified approach yields a recently published estimator of undetected *PD* in a single assemblage

In other words, α_r should be estimated by r^*/n, where $r^* = (r + 1)f_{r+1}/f_r$. The Good-Turing frequency formula is thus contrary to most people's intuition because the estimator in (1c) depends not only on the sample frequency r of the focal species, but also on the frequency information derived from species in the next frequency class, $r + 1$.

Good (1953) used a fully Bayesian approach to theoretically justify the formula (1c), whereas Robbins (1968) derived it as an empirical Bayes estimator. Good (2000) wrote "when preparing my 1953 article, I had forgotten Turing's somewhat informal proof in 1940 or 1941, which involved cards or urn models in some way, and I worked out a separate proof [Bayes estimator]. I still don't recall Turing's proof." Nevertheless, Good (1983, p. 28) provided a very intuitive non-Bayesian justification of the Good-Turing frequency formula as follows: Given an original sample of size n, consider the probability of the event that the next individual will be a species that had appeared r times in the original sample. (Mathematically, this probability is simply $\sum_{i=1}^{S} p_i I(X_i = r) = \alpha_r f_r$, as defined in Eq. 1a.) If this event occurs, then the species to which the additional individual belongs must appear $r + 1$ times in the enlarged sample of size $n + 1$. Because the order in which individuals were sampled is assumed to be irrelevant, the total number of individuals in the enlarged sample of size $n + 1$ for those species (that appeared in the additional individual and had appeared r times in the original sample) is $(r + 1)f_{r+1}$. Thus, the probability of the aforementioned event in the enlarged sample of size $n + 1$ is $(r+1)f_{r+1}/(n+1)$, which can be approximated by $(r+1)f_{r+1}/n$ if n is large enough. Dividing this by the number of such species, f_r, we obtain the mean relative abundance of those species, which is the classic Good-Turing frequency formula as given in Eq. (1c). Chiu et al. (2014*b*) proposed an improved formula $\hat{\alpha}_r$ shown below for $r = 0, 1, 2, \ldots$,

$$\hat{\alpha}_r = \frac{(r+1)f_{r+1}}{(n-r)f_r + (r+1)f_{r+1}} \approx \frac{(r+1)f_{r+1}}{(n-r)f_r}. \tag{1d}$$

This improved estimator generally has smaller mean squared error than the original Good-Turing estimator. In our subsequent derivation, we adopt the rightmost term in Eq. (1d); a simple non-Bayesian proof is provided (in Appendix S1) to facilitate the generalization to the two-assemblage case.

Undetected species richness

Statistically, species richness (observed species plus the top32n1yesian prta,(o7.6(du)ected.7(,)-254

Notice that, in the above derivation, if $\hat{\alpha}_0 \approx \hat{\alpha}_1$ (i.e., undetected species and singletons have identical mean relative abundances), then the inequality sign in Eq. (2c) mean of the products among all such shared species (there are f_{rr} such shared species) can be expressed as $\alpha_{rr} = \sum_{i=1}^{S_{12}} p_{i1} p_{i2} I(X_{i1} = r,\ X_{i2} = r)/f_{rr},\ \ r = 0,\ 1,\ 2\ldots$ The following generalized two-assemblage Good-Turing formula provides an estimator for α_{rr} (see Appendix S1 for a proof):

$$\hat{\alpha}_{rr} = \frac{(r+1}{}$$

sample that are descended from branch i. Then we can expand the set of observed species abundances to a larger branch abundance set $\{X_i^*,\ i = 1,\ 2,\ \ldots,\ B\}$ with $(X$

Undetected shared PD between two assemblages

Following the approach to the two-assemblage model formulation and the data framework described in the section *Two-assemblage Good-Turing Formulas*, we assume that all S species of the *pooled* assemblage are indexed by 1, 2,

that proposed in Eq. (8d) can be applied to each of the three terms. The variance and confidence interval associated with this estimator follow directly from those for the Chao1-shared estimator. Under the condition that undetected species and singletons have approximately homogenous abundances, the Chao1-*FAD* estimator is nearly unbiased for any given species-pairwise distance matrix.

A summary of formulas and descriptions for estimating shared species richness and *FAD* is given in Appendix S3: Table S1, where the analogy between the estimation procedures of the two measures can be seen. Chao et al. (2014*a*) define a "functional entity" as a species pair with one unit of distance between the two species. In *FAD*, a functional entity plays the same role as a "shared species" between two assemblages. For example, a species-pair with distance $d_{ij} = 5$ is counted as 5 "shared species" (i.e., 5 functional entities). Thus the measures of total distances of species pairs, $\{F_{rv}, F_{r+}, F_{+v}, F_{++}; r, v = 0, 1, 2, \ldots\}$, play the same roles as the counts of shared species richness $\{f_{rv}, f_{r+}, f_{+v}, f_{++}; r, v = 0, 1, 2, \ldots\}$ (defined in Eqs. 4a – 4d).

Undetected shared FAD between two assemblages

Under the two-assemblage model formulation and data framework described in the section *Two-assemblage Good-T ring Form las*, we further assume that the functional distance between the *i* species (including 110 singletons and 48 doubletons) among 1794 individuals in the data from the Edge habi-

Conclusion and Discussion

We have generalized the original one-assemblage Good-Turing frequency formula (Eqs. 1c and 1d) to the case of two assemblages (Eq. 5c), and also extended it to a phylogenetic version (Eqs. 8b and 10b) as well as a functional

A similar procedure can be applied to *FAD* and other

Transactions of the Royal Society of London B - Biological Sciences 345:101–118.

Colwell, R. K., A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology 5:3–21.

Diaz, S., and M. Cabido. 2001. Vive la différence: plant functional diversity matters to ecosystem processes. Trends in Ecology and Evolution 16:646–655.

Faith, D. P. 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation 61:1–10.

Ferrier, S., G. Manion, J. Elith, and K. Richardson. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and Distributions 13:252–264.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40:237–264.

Good, I. J. 1983. Good thinking: the foundations of probability and its applications. University of Minnesota Press, Minneapolis, USA.

Good, I. J. 2000. Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval enigma. Journal of Statistical Computation and Simulation 66:101–111.

Good, I. J., and G. Toulmin. 1956. The number of new species and the increase of population coverage when a sample is increased. Biometrika 43:45–63.

Gotelli, N. J., and A. Chao. 2013. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. Pages 195–211 *in* S. A. Levin, editor. Encyclopedia of biodiversity. Second editio~~n~~. Volume 5. Academic Press, Waltham, MA, USA.

Gotelli, N. J., and R. K. Colwell. 2011. Estimating species richness. Pages 39–54 *in* A. Magurran, and B. McGill, editors. Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford.

Gotelli, N. J., and D. J. McCabe. 2002. Species co-occurrence: a meta-analysis of J.M. Diamond's assembly rules model. Ecology 83:2091–2096.

Gotelli, N. J., R. M. Dorazio, A. M. Ellison, and G. D. Grossman. 2010. Detecting temporal trends in species assemblages with bootstrapping procedures and hierarchical models. Philosophical Transactions of the Royal Society B 365:3621–3631.

Hortal, J., P. A. V. Borges, and C. Gaspar. 2006. Evaluating the performance of species richness estimators: sensitivity to sample grain size. Journal of Animal Ecology 75:274–287.

Hsieh, T. C., and A. Chao. 2017. Rarefaction and extrapolation: making fair comparison of abundance-sensitive phylogenetic diversity among multiple assemblages. Systematic Biology 66:100–111.

Jesus, R. M., and S. G. Rolim. 2005. Fitossociologia da floresta atlântica de tabuleiro em Linhares (ES). Boletim Técnico SIF 19:1–149.

Jost, L., A. Chao, and R. Chazdon. 2011. Compositional similarity and beta diversity. Pages 66–84 *in* A. Magurran, and B. McGill, editors. Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford, UK.

Lozupone, C., and R. Knight. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71:8228–8235.

Magnago, L. F. S., D. P. Edwards, F. A. Edwards, A. Magrach, S. V. Martins, and W. F. Laurance. 2014. Functional attributes change but functional richness is unchanged after fragmentation of Brazilian Atlantic forests. Journal of Ecology 102:475–485.

Magurran, A. E. 2004. Measuring biological diversity. Blackwell, Oxford, UK.

Matos, F. A. R., L. F. S. Magnago, M. Gastauer, J. M. B. Carreiras, M. Simonelli, J. A. A. Meira-Neto, and D. P. Edwards. 2017. Effects of landscape configuration and composition on phylogenetic diversity of trees in a highly fragmented tropical forest. Journal of Ecology 105:265–276.

McGrayne, S. B. 2011. The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy. Yale University Press, New Haven, USA.

Pan, H. Y., A. Chao, and W. Foissner. 2009. A non-parametric lower bound for the number of species shared by multiple communities. Journal of Agricultural, Biological and Environmental Statistics 14:452–468.

Paula, A., and J. J. Soares. 2011. Estrutura horizontal de um trecho de Floresta Ombrófila Densa das Terras Baixas na Reserva Biológica de Sooretama, Linhares, ES. Revista Floresta 41:321–334.

Rangel, T. F., R. K. Colwell, G. R. Graves, K. Fučíková, C. Rahbek, and J. A. F. Diniz-Filho. 2015. Phylogenetic uncertainty revisited: Implications for ecological analyses. Evolution 69:1301–1312.

Robbins, H. E. 1968. Estimating the total probability of the