Supporting Information

Connolly et al. 10.1073/pnas.1406664111

SI Text

SI Materials and Methods

Candidate Neutral and Nonneutral Approximations. A set of noninteracting populations undergoing pure random drift in population size (birth rate equals death rate, no immigration, emigration, or environmental stochasticity) produces a species abundance distribution in which the probability that a species has a given abundance, n, varies inversely with abundance (1). On log-log scale, this is a straight line with a slope of −1:

$$\log(f(n)) = \log(\kappa) - \log(n), \qquad \textbf{[S1]}$$

where $f(n)$ is the probability that a species has abundance n, and κ is a normalizing constant. Neutral models have two characteristics that cause them to depart from the case of pure random drift. First, because species are ecologically identical, there is a constraint on total community size that is independent of species richness. Using a maximum entropy argument, a modification to this power-law model can be derived that accounts for this constraint (1):

$$\log(f(n)) = \log(\kappa) - \log(n) - \phi n. \qquad \textbf{[S2]}$$

Eq. **S2** is equivalent to Fisher's log-series distribution (1). Second, neutral models also may have characteristics that cause individual species' dynamics to depart from the pure drift assumption, such as dispersal limitation (2), or unequal birth and death rates (3). Pueyo (1) conceptualizes small departures from pure drift as perturbations to the value of the slope of −1 in Eq. **S1**. The combination of these two extensions to Eq. **S1** yields the following:

$$\log(f(n)) = \log(\kappa) - \beta\ \log(n) - \phi n. \qquad \textbf{[S3]}$$

Note that, by setting $\beta = 1 - k$ and $\phi = 1/a$, and the normalization constant $\kappa = (\Gamma(k)a^k)^{-1}$, it becomes apparent that $f(n)$ in Eq. **S3** is a gamma distribution with shape k and scale a. Because it is well known that many neutral models can depart markedly from the log-series distribution (2, 4, 5), we take the gamma distribution as our candidate neutral approximation.

Increasingly large departures from neutrality might be poorly approximated by a perturbation to the slope of a power-law relationship, in which case a second-order perturbation may be needed, where a quadratic term is added to the first-order model:

$$\log(f(n)) = \log(\kappa) - \beta\ \log(n) + c\ [\log(n)]^2. \qquad \textbf{[S4]}$$

If we set $\beta = 1 - \mu/\sigma^2$, $c = -1/(2\sigma^2)$, and $\log(\kappa) = -(\frac{\mu^2}{2\sigma^2} + \log(\sqrt{2\pi}\sigma))$, then $f(n)$ in Eq. **S4** is a lognormal distribution where μ and σ are the mean and SD of $\log(n)$, respectively (1). We therefore take the lognormal as our candidate nonneutral approximation.

Because the gamma and lognormal distributions are continuous, whereas abundances are integer-valued, and because many species abundance data are incomplete samples from an underlying community abundance distribution, in our analyses we assess our neutral and nonneutral approximations by fitting Poisson-gamma (i.e., negative binomial) and Poisson-lognormal mixture distributions:

$$P(r) = \int_{\lambda=0}^{\infty} \frac{\lambda^r e^{-\lambda}}{r!} f(\lambda)\ d\lambda, \qquad \textbf{[S5]}$$

where $P(r)$ is the probability that a species has abundance r in the sample, λ is the mean of the Poisson distribution (and thus integrated out of the likelihood), and $f(\lambda)$ is either the lognormal or the gamma distribution. These distributions are commonly used to represent random samples of individuals from underlying gamma or lognormal community abundance distributions, respectively (6–8). More specifically, we use the zero-truncated forms of the Poisson-gamma and Poisson-lognormal distributions, because, by definition, a species is not observed in the sample if it has zero abundance (6):

$$p(r) = \frac{P(r)}{1 - P(0)}. \qquad \textbf{[S6]}$$

Assessing the Neutral Approximation. Our five candidate neutral models exhibited a broad range of auxiliary assumptions. In Hubbell's "original neutral model," local communities are partially isolated by dispersal from the broader metacommunity, and new species arise with a fixed probability from individual birth events (analogous to mutation events in population-genetic neutral models) (9). The "protracted speciation neutral model" is similar to the original neutral model, but it incorporates a time lag between the appearance of an incipient new lineage, and its recognition as a distinct species (10). In the "fission speciation model," speciation occurs by random division of existing species (e.g., via vicarance); this model can exhibit a more superficially lognormal-like species abundance pattern than point speciation models, in that its log-abundance distributions are more symmetric about a single mode than other neutral models (5). In the "independent species model" (3, 11), population dynamics are density independent, per-capita birth rate is less than per-capita death rate, and there is a constant immigration rate. Finally, in the spatially explicit neutral model (4), speciation follows a point-mutation process (as in the original neutral model), and dispersal distances follow a Gaussian kernel. The first four models have explicit mathematical expressions for the species abundance distribution at equilibrium, which facilitates formally evaluating the neutral approximation: see equations below). For the spatially explicit neutral model, we used the approximate species abundance distributions generated by simulation in the original paper and kindly provided by the authors (4).

As noted in the main text, the strict definition of neutrality that applies to these models contrasts with symmetric models that implicitly allow for niche or demographic differences among species, for instance, by having within-species competition be stronger than between species competition (12), by implicitly including temporal niche differentiation via different responses to environmental fluctuations (13), or by allowing species with different life history types to differ in their speciation rates (14).

To assess how well the Poisson-gamma distribution approximates our alternative neutral models, we considered a broad range of neutral model parameter space spanning most of the realistic range for real species abundance data (hundreds to tens of thousands of individuals, and from less than 10 to many hundreds of species). For each neutral model parameter combination, we used the Kullback–Leibler (K-L) divergence, a measure of the information lost when one distribution is used as an approximation for another (15). Specifically, we found the Poisson-gamma distribution parameters that minimized the K-L divergence. For discrete data, such as counts, K-L divergence is as follows:

$$D = \sum_{n} \pi(n) \log\left(\frac{\pi(n)}{p(n)}\right), \qquad \text{[S7]}$$

where n indexes the possible values of the random variable (in this case, abundance), $\pi(n)$ is the distribution being approximated (the relevant neutral model), and $p(n)$ is the approximating model—in this case, the zero-truncated Poisson-gamma distribution (Eq. **S6**).

Because our analysis of the empirical data is largely a comparative assessment of the Poisson-gamma and Poisson-lognormal distributions, our conclusions rely on an implicit assumption that a Poisson-gamma distribution would outperform a Poisson-lognormal if data were actually generated by neutral dynamics. Therefore, in addition to assessing the performance of the Poisson-gamma as a neutral approximation in absolute terms, we also simulated 100 species abundance distributions from each of the 126 equilibrium neutral abundance distributions used in the previous analysis (Fig. S1), and we compared the best-fit Poisson-gamma and Poisson-lognormal distributions for the 12,600 simulated abundance distributions, exactly as we did for the empirical species abundance distributions.

Criteria for Empirical Data Inclusion. Our criteria for data inclusion were as follows. First, the data needed to record counts of individual organisms for a given level of sampling effort (e.g., sample volume, or transect area). Second, data needed to be collected by experts (i.e., survey programs including data collected by amateurs were excluded), to minimize the risks of misidentification or miscounting. Third, data needed to be focused on the assemblage level, rather than on specific target species. Fourth, if sampling effort varied within species abundance samples, it had to be possible to standardize to a common level of effort. For instance, if fishes were counted on 10-m^2 and 50-m^2 transects, then 10/50 = 20% of the individuals on the larger transects were subsampled and pooled with the counts from the smaller transects (16). Three of the datasets we used required subsampling [Great Barrier Reef Fish (GBR), National Oceanic and Atmospheric Administration (NOAA) Central Pacific Reef Fish (CPF), and South East Fishery: Shelf Fish (SEF)].

Model Fitting. To assess the relative performance of the Poisson-gamma and Poisson-lognormal for both simulated neutral and real species abundance data, we found the gamma or neutral model parameters that maximized the log-likelihood for the zero-truncated forms of the Poisson-gamma and Poisson-lognormal abundance distributions:

$$\mathcal{L} = \sum_{r} n_r \log(p(r)), \qquad \text{[S8]}$$

where n_r is the number of species with abundance r in the sample, and $p(r)$ is the zero-truncated probability that a species has abundance r (Eq. **S6**). Best-fit models were obtained by finding the gamma or neutral model parameters that maximized the log-likelihood for each site.

Analysis of Variation in the Shapes of Species Abundance Distributions. To determine whether there was any systematic variation in the strength of evidence for gamma-like versus lognormal-like distributions, and whether any such variation was associated with systematic differences in the patterns of commonness and rarity in communities, we needed a sample-standardized measure of the relative strength of support for a candidate model. Specifically, the maximum log-likelihood for a species abundance model at a given site is the sum of the contributions of each species' abundance value to the log-likelihood. To control for this effect of the number of observations, we computed, for each site, a per-observation average log-likelihood: the site's maximum log-likelihood divided by the number of species abundances contributing to that log-likelihood. This approach is used in time series analysis, when models that have been fitted to different numbers of observations (e.g., models with different time lags) must be compared (17). Our standardized measure of model support was simply the difference between the standardized gamma and lognormal maximum log-likelihoods.

As our measure of the dominance of common species, we took, in the first instance, the abundance of the most abundant species, expressed as a proportion of the total number of individuals in the species abundance distribution. As our rarity measure, we took the proportion of species that were singletons (i.e., represented by a single individual in the abundance distribution). We used linear mixed-effects models to characterize the extent to which these two quantities explained variation within and among datasets in the standardized support for the lognormal over the gamma, at all scales (site, mesoscale, regional). To confirm that our results were not sensitive to the particular commonness or rarity metrics we considered, we repeated our analysis using the combined abundance of the three most abundant species, and using the proportion of species in the bottom two octaves of abundance (i.e., with proportion of species with abundance three or less).

Parametric Bootstrap Goodness of Fit. Goodness of fit to the empirical data was assessed with parametric bootstrapping, using a hypergeometric algorithm described in detail elsewhere (7). Parametric bootstrapping involves simulating datasets that conform to the assumptions of a particular fitted species abundance model. For example, to test the goodness of fit of the Poisson-lognormal, one simulates Poisson random sampling of individuals from an underlying lognormal distribution of species abundances. Then, the model is fitted to each simulated dataset, and a goodness of fit statistic calculated. The frequency distribution of this statistic across simulated datasets approximates the statistic's expected distribution, under the null hypothesis that the data conform to the model. As a goodness of fit statistic, we use a normalized measure of model deviance, which, following convention, we term $\hat{c}$ (16). Deviance is a likelihood-based measure of how far away the model is from exhibiting a perfect fit to the data. $\hat{c}$ is obtained by taking all deviances for the model's fits to the observed and simulated data, and dividing each by the average of the simulated deviances. Thus, $\hat{c}$ has an expected value of 1.0. We judged the lack of fit as statistically significant if the $\hat{c}$ of the observed data was greater than 95% of the corresponding simulated $\hat{c}$ values.

Species Pool Estimation. Using the maximum-likelihood estimates, the probability that a species is present in the species pool but has abundance zero in the sample, $P(0)$, is calculated from Eq. **S5**, by substituting 0 for r. Then, the number of species in the community that has been sampled can be estimated from the following:

$$\hat{S} = \frac{S_{obs}}{1 - P(0)}, \qquad \text{[S9]}$$

where $\hat{S}$ is the estimated number of species in the community, and S_{obs} is the number of species observed in the data. Nonparametric jackknife estimates were calculated using the frequency distribution of species occurrences across sites (i.e., presence–absence data: see ref. 16). Jackknife order was calculated separately for each dataset, using the sequential testing procedure recommended by ref. 18.

SI Results

Performance of the Neutral Approximation. Fig. S1 depicts the fit of the neutral approximation to our five alternative neutral models. For the first three models, these plots encompass three order-of-magnitude variation in local community sizes, J (10^2 to 10^4 individuals), because most species abundance distributions are on the order of hundreds to (occasionally) tens of thousands of individuals. Similarly, we show a broad range of immigration rates from $m = 0.01$ (1% of newborns are immigrants) to 1.0 (an entirely open local community). We plot a range of values of the biodiversity parameter, θ, so that the expected number of species in the community spanned a very broad range (typically from a low of about five species, for small, isolated communities with low θ, to many hundreds of species for large, high-immigration communities with large θ). Note that the range of values of θ needed to span these richness values differs between the fission speciation model and the first two models, because the parameter is defined somewhat differently in this model. The protracted speciation model includes an additional parameter, τ', which is the number of generations required for speciation to occur, relative to the metacommunity size (the special case $\tau' = 0$ corresponds to the original neutral model). The fourth (independent species) neutral model differs from the others in that it does not explicitly characterize dynamics at the metacommunity scale. Rather, it implicitly assumes that species have equal abundance in the metacommunity (and thus they all have the same rate of immigration to the local community, γ), and that species' local population dynamics are independent of one another, and thus a function of only γ and the ratio of local per-capita birth to death rates, x. Because within-species dynamics are also density independent, this is consistent with the neutrality assumption (individuals have no effect on one another's per-capita growth rates, regardless of whether they belong to the same or different species). This density-independent assumption means that the model is a probability distribution of species abundances, and not a model of overall species frequencies. Consequently, unlike the previous neutral models, it does not predict species richness. Similarly, for the spatially explicit model, the form of the species abundance distribution depends on the speciation probability (ν), and the ratio of the sampling area A (i.e., the local community size) to the squared width of the dispersal kernel, L, rather than either of the latter two variables independently (4). Thus, a given shape for the species abundance distribution can correspond to a broad range of different community species richness values, depending on whether A and L are both small or both large.

Fig. S1 shows that the Poisson-gamma neutral approximation performs very well in the overwhelming majority of cases. There are, however, some cases where the approximation performs less well. These typically correspond to parameter combinations that imply very species-poor assemblages. One class of such cases corresponds to small (~100 individuals), very low-immigration, low-diversity assemblages (~5 species: e.g., *Top Left* of Fig. S1*A*). Here, the neutral model has an elevated probability that one species is nearly monodominant (the curve bends upward at the right, for species abundances close to the total community size), which the Poisson-gamma distribution cannot capture. A second class of cases, specific to the protracted speciation model, involves a flattening of the species abundance distribution at low abundances (e.g., $m = 0.1$, $\theta = 4$, $J = 10^4$ in Fig. S1*C*). This effect is too small to see clearly for the range of parameter values shown in Fig. S1, but is somewhat more pronounced in very large communities with very low values of the biodiversity parameter ($\theta \sim 1$), for which the ratio of individuals to species is very high (e.g., a local community with 10,000 individuals but only about 10 species). The third class of cases are specific to the fission speciation model and involve an excess of rare species, relative to the Poisson-gamma distribution (e.g., $m = 0.01$, $J = 10^4$, $\theta = 40$ in Fig. S1*D*). As with the second class of cases, this effect is relatively small in Fig. S1, but can be more pronounced for very large, particularly isolated, communities with few species (e.g., 10,000 individuals and about 10 species, implying mean abundances of about 1,000). For the data analyzed in this paper, however, most sites are very far from these extreme low-diversity cases. The typical (median) site is a sample of 422 individuals containing 17 species, and very few sites contain so few species at such large sample sizes (86% of sites, for instance, have mean species abundances of 100 or less). Moreover, the individual datasets vary substantially in community size and observed species richness (e.g., mean site richness varies from 9 to 126 species across the 14 datasets, and average species abundances at the site level range from 4 to 123 across all datasets except one). Thus, the overwhelming majority of our sites could not correspond to those regions of parameter space where the Poisson-gamma distribution performs less well as an approximation for neutral dynamics.

Robustness to Ecological and Taxonomic Heterogeneity. Although neutral models have previously been applied to very heterogeneous communities (19), including benthic marine invertebrates (2) [and indeed their capacity to characterize such systems has been invoked as evidence of their robustness (2)], most neutral model communities are conceptualized as a guild of organisms competing for a shared set of resources. Some of our datasets are relatively taxonomically and ecologically homogeneous [e.g., Indo-Pacific Coral Crustaceans (IPC), which contains only crustaceans associated with dead coral heads]. However, others are more heterogeneous. Therefore, to determine whether our results were sensitive to this taxonomic and ecological heterogeneity of the assemblages, we classified our species into guilds, where information was available, and reanalyzed our species abundance data, limiting the analysis to species from the most species-rich guild for each dataset (Table S3). Such a classification is necessarily approximate for marine animals, given the high degree of omnivory in the ocean. Nevertheless, the analysis allows us to evaluate whether or not our conclusions are sensitive to the extent of heterogeneity in the data. The resolution of the groupings for this analysis depended somewhat on both the taxonomic and ecological heterogeneity of the original data, and also on the species richness in the samples. Specifically, we used as a rule of thumb that guilds should have a minimum of 10 species, necessitating use of more coarse groupings for more species-poor datasets.

By restricting the analysis to a subset of the species, the statistical power to detect differences between Poisson-lognormal and Poisson-gamma species abundances is reduced—the more heterogeneous the original dataset, the smaller the subset of species that could be included in the analysis. Nevertheless, strong support for the Poisson-lognormal remained: across site, mesoscale, and regional levels, the Poisson-lognormal was strongly (>95%) supported in 27 cases, whereas the Poisson-gamma was strongly supported in only 1 (Table S3).

Analysis of Variation in the Shapes of Species Abundance Distributions. Standardizing model support by dividing by the number of observed species abundances successfully controlled for the effects of statistical power shown in Fig. 2, at least at the site level and mesoscale: mixed-effects linear model analyses using the number of distinct species abundance values as an explanatory variable indicated that the overall effect did not differ significantly from zero, and explained about 1–10% of the variation in standardized model support across datasets (see R^2 values in Table S4). In contrast, the strength of support for the Poisson-lognormal over the Poisson-gamma increased strongly with the relative abundance of the most abundant species at site, mesoscale, and regional (whole-dataset) levels. The positive relationship was highly consistent between datasets at both the site level and mesoscale (gray lines in Fig. S4 *A* and *C*), and explained about one-half or more of the variation (Table S4 and Fig. S4 *B*, *D*, and *E*). In contrast, the proportion of singletons was a poor predictor of relative model performance: the estimated direction of the effect was not consistent across datasets at the site level (gray lines in Fig. S5*A*), and the estimated overall effect did not differ significantly from zero at any scale (Table S4 and black lines in Fig. S5 *A*, *C*, and *E*) and never explained more than 16% of the variation (Table S4).

To further assess the strength of these results, we repeated our common-species analysis using the combined relative abundance of the three most abundant species. This, too, was strongly positively related to support for the Poisson-lognormal distribution (slope: 0.40 ± 0.02, pseudo-$R^2 = 0.44$ at site level; 0.24 ± 0.05, pseudo-$R^2 = 0.53$ at mesoscale level; 0.15 ± 0.04, $R^2 = 0.55$ at regional scale). Conversely, expanding our definition of rarity to encompass the proportion of species in the bottom two octaves (species with abundance 3 or less) did not improve its effectiveness as a predictor of standardized support for the Poisson-lognormal over the Poisson-gamma: the overall relationship did not differ significantly from zero at any scale (no slopes significantly different from zero, pseudo-$R^2 < 0.02$ at site-scale and mesoscale levels, $R^2 = 0.21$ at regional level).

1. Pueyo S (2006) Diversity: Between neutrality and structure. *Oikos* 112(2):392–405.
2. Hubbell SP (2001) *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univ Press, Princeton).
3. Volkov I, Banavar JR, Hubbell SP, Maritan A (2007) Patterns of relative species abundance in rainforests and coral reefs. *Nature* 450(7166):45–49.
4. Rosindell J, Cornell SJ (2013) Universal scaling of species-abundance distributions across multiple scales. *Oikos* 122(7):1101–1111.
5. Etienne R, Haegeman B (2011) The neutral theory of biodiversity with random fission speciation. *Theor Ecol* 4(1):87–109.
6. Pielou EC (1977) *Mathematical Ecology* (Wiley, New York).
7. Connolly SR, Dornelas M, Bellwood DR, Hughes TP (2009) Testing species abundance models: A new bootstrap approach applied to Indo-Pacific coral reefs. *Ecology* 90(11): 3138–3149.
8. Sæther BE, Engen S, Grøtan V (2013) Species diversity and community similarity in fluctuating environments: Parametric approaches using species abundance distributions. *J Anim Ecol* 82(4):721–738.
9. Etienne RS, Alonso D (2005) A dispersal-limited sampling theory for species and alleles. *Ecol Lett* 8(11):1147–1156.
10. Rosindell J, Cornell SJ, Hubbell SP, Etienne RS (2010) Protracted speciation revitalizes the neutral theory of biodiversity. *Ecol Lett* 13(6):716–727.
11. He F (2005) Deriving a neutral model of species abundance from fundamental mechanisms of population dynamics. *Funct Ecol* 19(1):187–193.
12. Volkov I, Banavar JR, He FL, Hubbell SP, Maritan A (2005) Density dependence explains tree species abundance and diversity in tropical forests. *Nature* 438(7068): 658–661.
13. Allen AP, Savage VM (2007) Setting the absolute tempo of biodiversity dynamics. *Ecol Lett* 10(7):637–646.
14. Ostling A (2012) Do fitness-equalizing tradeoffs lead to neutral communities? *Theor Ecol* 5(2):181–194.
15. Burnham KP, Anderson DR (2002) *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach* (Springer, Heidelberg).
16. Connolly SR, Hughes TP, Bellwood DR, Karlson RH (2005) Community structure of corals and reef fishes at multiple scales. *Science* 309(5739):1363–1365.
17. McQuarrie ADR, Tsai C-L (1998) *Regression and Time Series Model Selection* (World Scientific, River Edge, NJ).
18. Burnham KP, Overton WS (1979) Robust estimation of population size when capture probabilities vary among animals. *Ecology* 60(5):927–936.
19. de Aguiar MAM, Baranger M, Baptestini EM, Kaufman L, Bar-Yam Y (2009) Global patterns of speciation and diversity. *Nature* 460(7253):384–387.

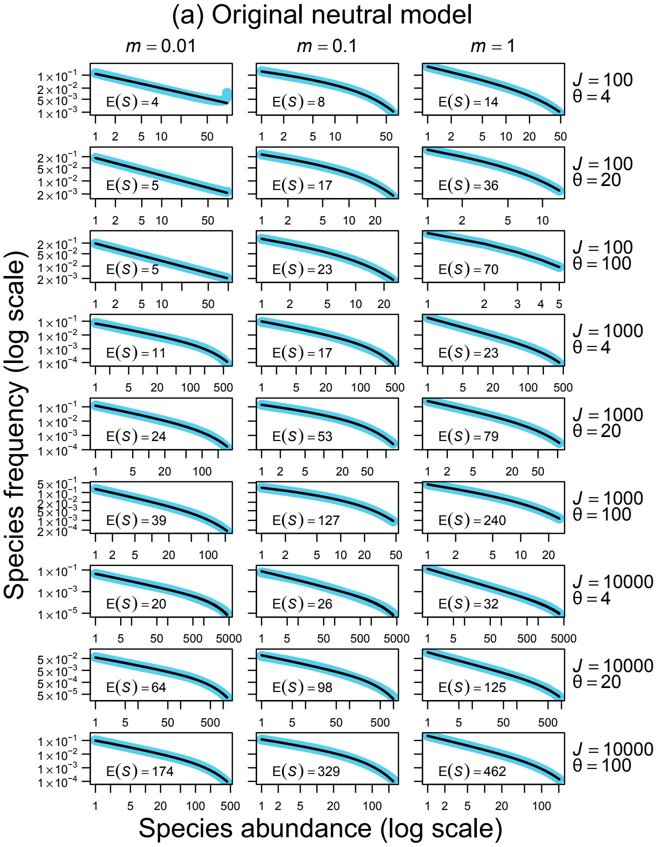

Fig. S1. (Continued)

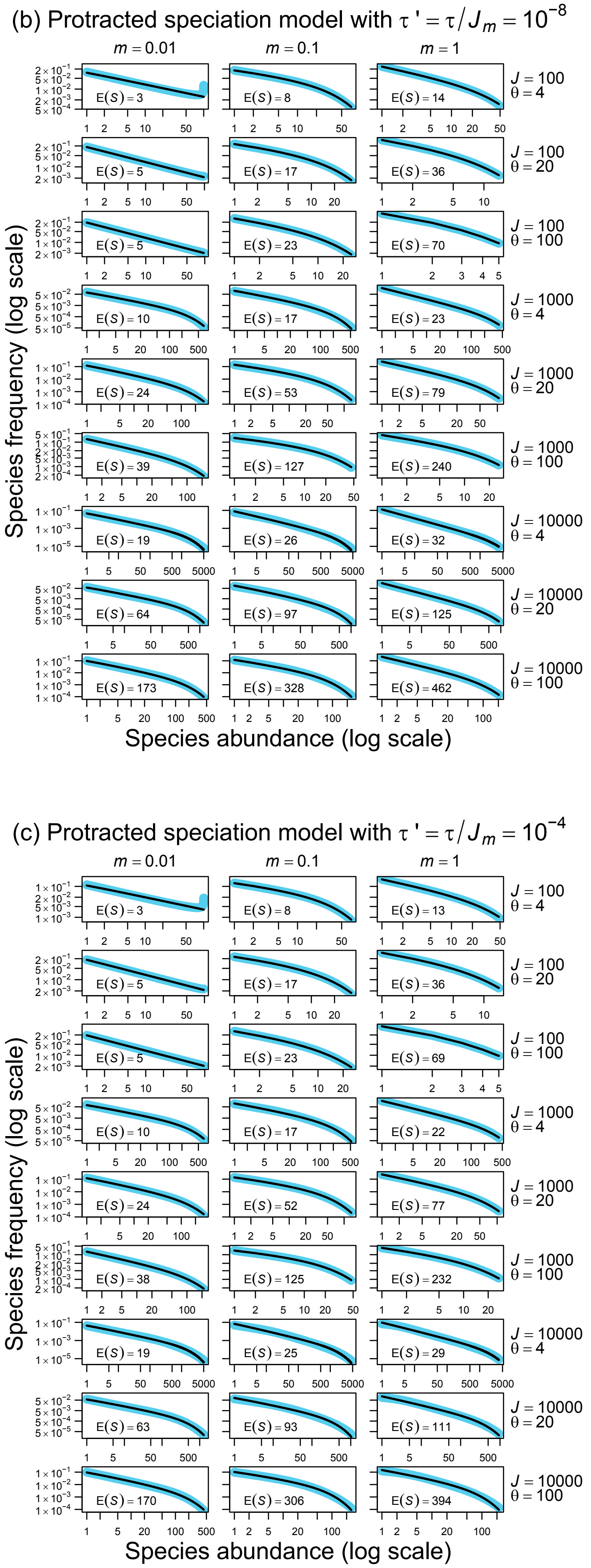

Fig. S1. (Continued)

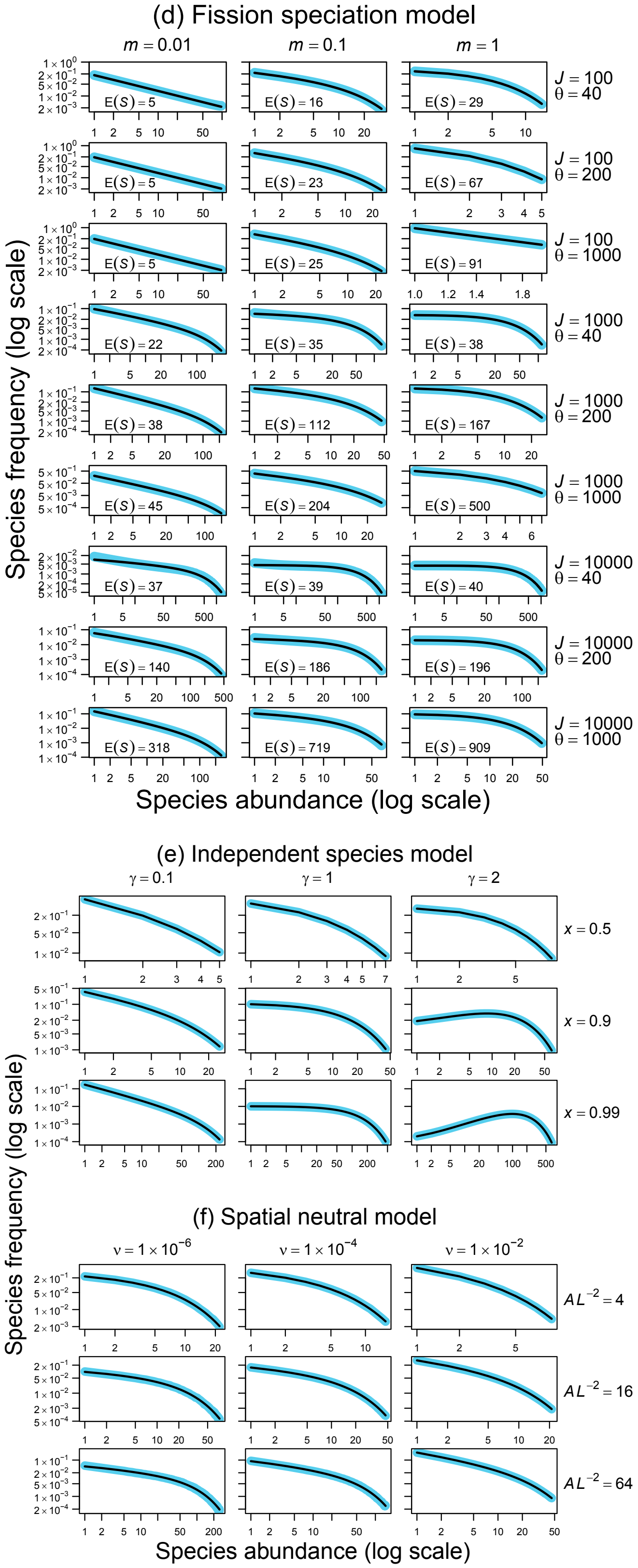

Fig. S1. Examples of the fit of the Poisson-gamma neutral approximation (black line) to the five candidate neutral approximations (blue lines). The fits are shown as "Pueyo plots": both the vertical and horizontal axes are shown on a log scale. The horizontal axis is truncated at the abundance value where the cumulative expected number of species equals 99% of the total (i.e., on average, only 1 of 100 species would be expected to have greater abundance). (*A*–*F*) The six different neutral models, as specified at the top of the corresponding group of panels. For models *A*–*D*, E(S) in each panel indicates the expected number of species for that parameter combination.

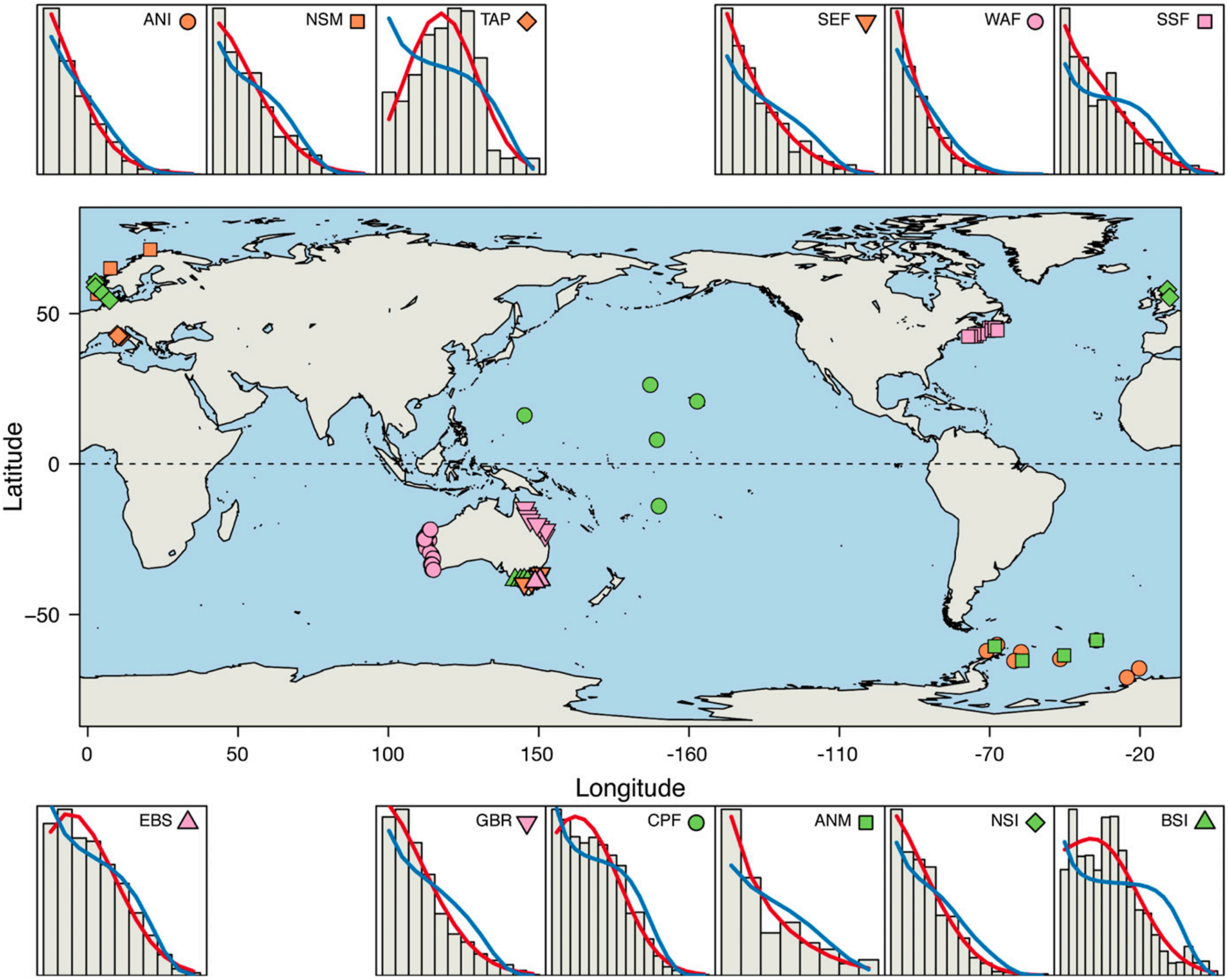

Fig. S2. Observed and best-fit mesoscale abundance distributions. On the map, different combinations of colors and symbols correspond to different datasets: these are reproduced in the corresponding figure panels. See Table S1 for metadata, including abbreviations. Each point on the map is located at the centroid of the individual sites that were pooled to generate each mesoscale abundance distribution. Panels above and below the map compare observed and fitted species abundance distributions at this scale. The bars represent the mean proportion of species in different octave classes of abundance, across all mesoscale abundance distributions from the corresponding ecosystem (these are shown as true doubling classes: the first bar represents species with abundance 1; the second, abundances 2–3; the third, abundances 4–7; etc.). The blue and red lines show the mean of fitted values from fits of the Poisson-gamma and Poisson-lognormal distributions, respectively, to each mesoscale abundance distribution.

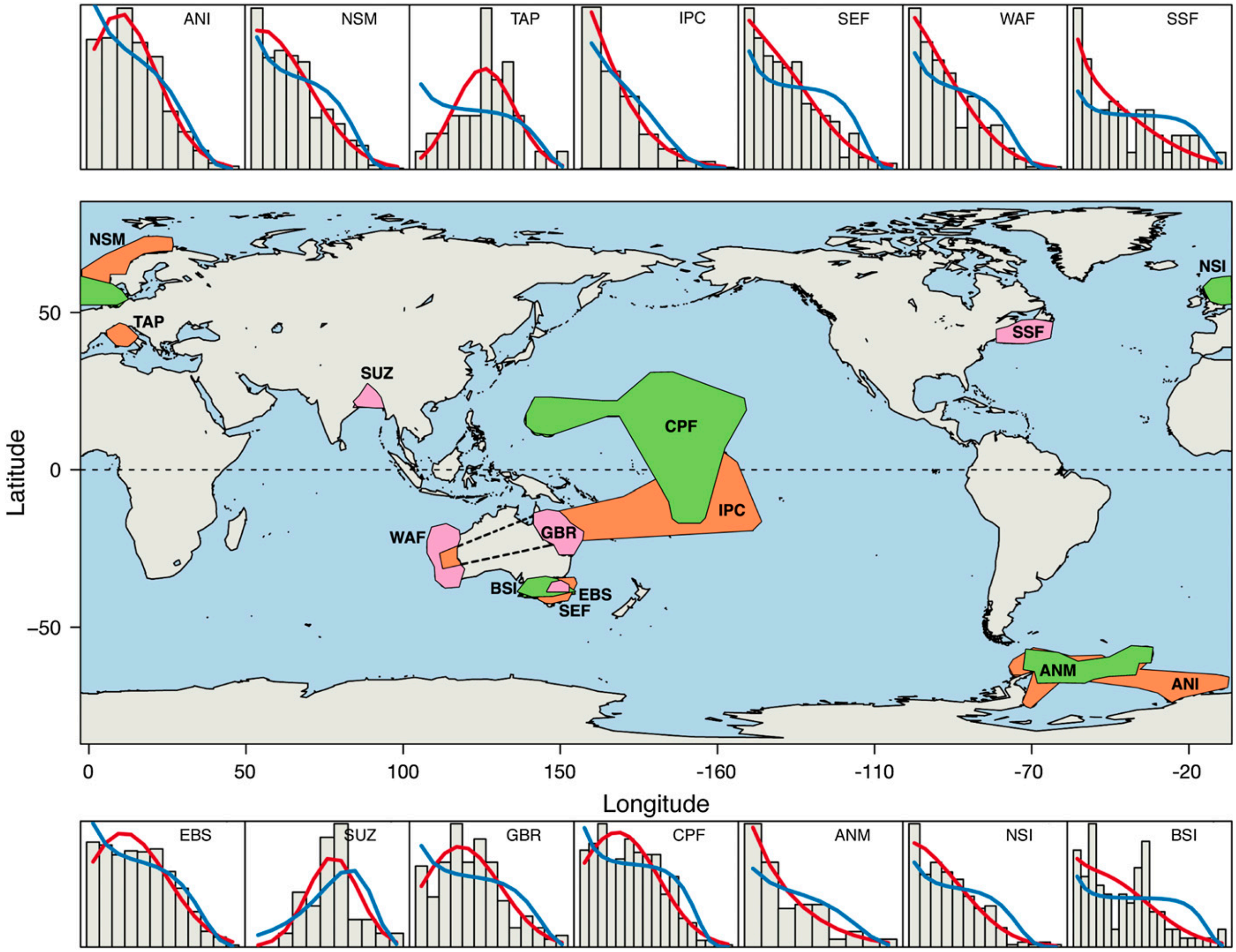

Fig. S3. Observed and best-fit regional abundance distributions. On the map, the area over which sites were pooled for each regional abundance distribution has been outlined. See Table S1 for metadata, including abbreviations. The panels above and below the map compare observed and fitted species abundance distributions at this scale. The bars represent the proportion of species in different octave classes of abundance (these are shown as true doubling classes: the first bar represents species with abundance 1; the second, abundances 2–3; the third, abundances 4–7; etc.). The blue and red lines show the fitted values from fits of the Poisson-gamma and Poisson-lognormal distributions to the data, respectively.

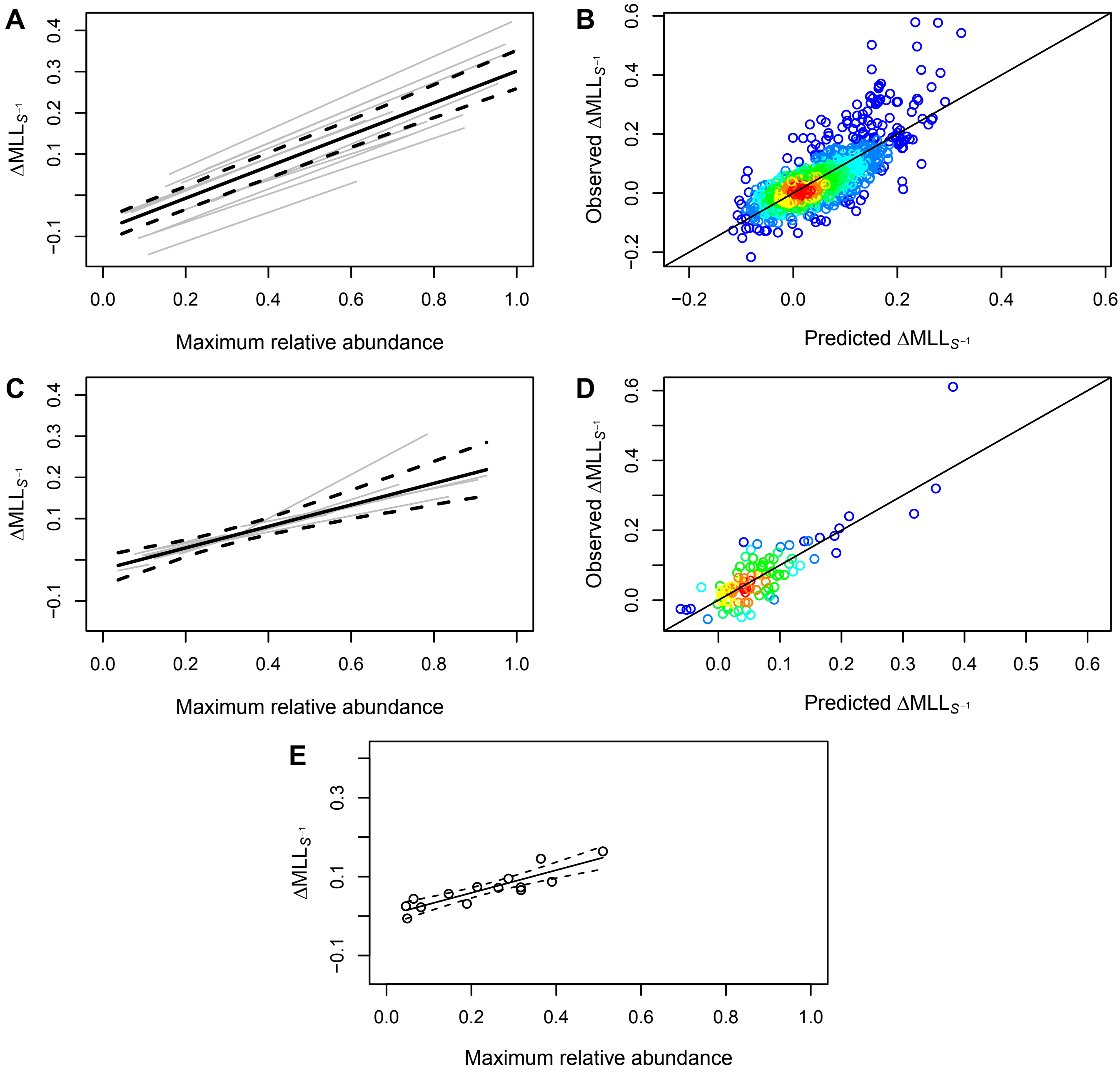

Fig. S4. Analysis of the variation in standardized support for the lognormal explained by the relative abundance of the most-abundant species (expressed as a fraction of the number of individuals sampled), at the (*A* and *B*) site scale, (*C* and *D*) mesoscale, and (*E*) regional scale. Positive relationships indicate stronger evidence against the gamma neutral approximation as the most-abundant species becomes more dominant. In *A* and *C*, the thick solid and dashed lines represent the overall (i.e., fixed effects) relationship, with 95% confidence intervals. The gray lines represent the relationships for the 14 individual datasets, based on the estimated random effects; individual lines are drawn to span only the range of horizontal axis values observed in the corresponding dataset. *B* and *D* show corresponding plots of observed versus predicted values, as estimated from the full fitted model. Because there is substantial overlap of points, the points have been color-coded according to the number of nearby observations, grading from red (high density of points) to blue. *E* is an ordinary least-squares (OLS) regression: because there is only one (pooled) regional abundance distribution per site, there is no random effect.

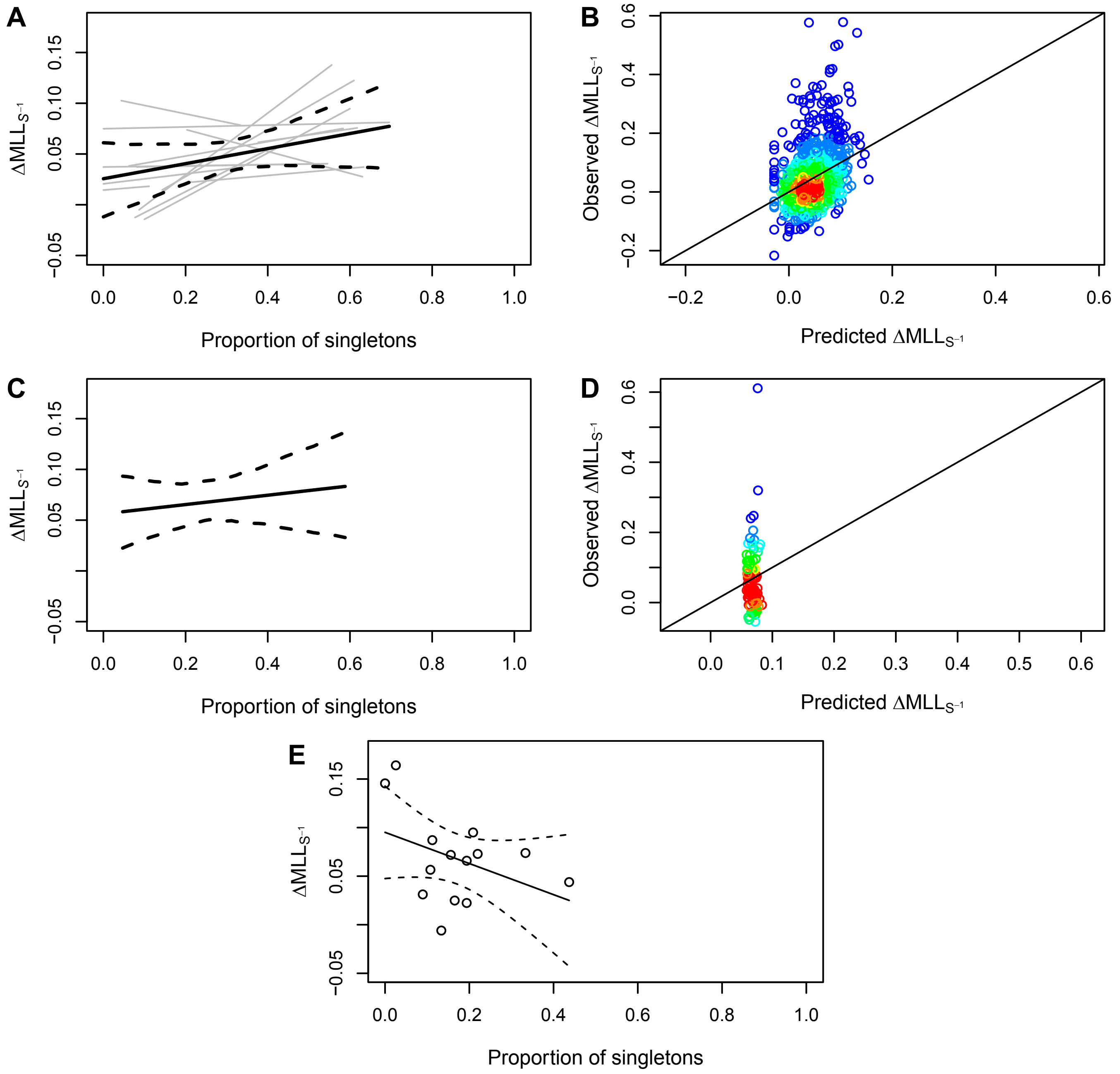

Fig. S5. Analysis of the variation in standardized support for the lognormal explained by the proportion of species that are singletons, at the (*A* and *B*) site scale, (*C* and *D*) mesoscale, and (*E*) regional scale. Positive relationships indicate stronger evidence against the gamma neutral approximation as proportion of singletons increases. In *A* and *C*, the thick solid and dashed lines represent the overall (i.e., fixed effects) relationship, with 95% confidence intervals. The gray lines represent the relationships for the 14 individual datasets, based on the estimated random effects; individual lines are drawn to span only the range of horizontal axis values observed in the corresponding dataset. *B* and *D* show corresponding plots of observed versus predicted values, as estimated from the full fitted model. Because there is substantial overlap of points, the points have been color-coded according to the number of nearby observations, grading from red (high density of points) to blue. *E* is an OLS regression: because there is only one (pooled) regional abundance distribution per site, there is no random effect.

Table S1. Metadata summary of global marine species abundance distribution samples

Dataset name	Summary	Latitudinal limits	Longitudinal limits	Depth range, m	Sampling method	Data contact
South East Fishery: Shelf Fish (SEF)	Fish from southeastern Australia; total of 173 species at 189 sites pooled into 13 mesoscale SADs	−39.0 −36.4	146.5 150.3	16 254	Fish trawl	A.W. Alan.Williams@csiro.au
Western Australia: Deep Fish (WAF)	Fish from western Australia; total of 282 species at 65 sites pooled into 23 locations	−35.1 −20.1	111.4 115.2	197 1,580	Fish trawl	A.W. Alan.Williams@csiro.au
Great Barrier Reef Fish (GBR)	Fish from underwater visual census surveys of coral reefs on the Great Barrier Reef (1); total of 195 species at 74 sites pooled into 8 mesoscale SADs	−23.9 −14.5	145.3 152.7	7	UVS	Hugh Sweatman Australian Institute of Marine Science, Townsville, Australia h.sweatman@aims.gov.au
Antarctic Molluscs (ANM)	Deep-water bivalves from the Scotia Arc, Antarctica; total of 96 species at 20 sites pooled into 4 mesoscale SADs	−58.2 −65.5	−60.0 −23.6	774 6,348	Epibenthic sledge	K.L. kl@bas.ac.uk
Indo-Pacific Coral Crustaceans (IPC)	Crustacean samples encompassing a total of 411 species from individual dead coral heads at 8 sites in the Indo-Pacific; not pooled at mesoscale	−23.4 6.4	−113.7 −149.8	10	Hand counts	L.P. PlaisanceL@si.edu
Tuscany Archipelago Fish (TAP)	Fish abundance from the Tuscany Archipelago; total of 39 species at 30 sites pooled into 4 mesoscale SADs	42.2 43.1	9.8 11.1	8 12	UVS	L.B.-C. lbenedetti@biologia.unipi.it
Eastern Bass Strait Invertebrates (EBS)	Invertebrates from the Eastern Bass Strait, Australia (2); total of 801 species at 47 sites pooled into 3 mesoscale SADs	−37.9 −37.8	148.2 148.7	17 51	Grab sample	R.S.W., G.C.B.P. rwilson@museum.vic.gov.au
NOAA Central Pacific Reef Fish (CPF)	Fish from underwater visual surveys of coral reefs throughout the Pacific; total of 491 species at 49 sites pooled into 5 mesoscale SADs	−14.6 28.5	−154.8 142.8	8.24 17.11	UVS	R.E.B. Rusty.Brainard@noaa.gov
Sunderban Zooplankton (SUZ)	Zooplankton from Sunderban mangrove wetland, India; total of 31 species at 7 sites; not pooled at mesoscale	21.6 22.3	88.0 88.9	2.0 8.9	Plankton tow	S.K.S. sarkar22@yahoo.com
Scotian Shelf Fish (SSF)	Fish from the Scotian Shelf, Northwestern Atlantic; total of 98 species at 458 sites pooled into 14 mesoscale SADs	42.1 45.6	−67.2 −57.3	16 176	Trawl	Steven E. Campana Fisheries and Oceans Canada, Bedford Institute of Oceanography, Dartmouth, Canada Steven.Campana@dfo-mpo.gc.ca
Bass Strait Intertidal Macroinvertebrates (BSI)	Invertebrates from the Bass Strait, Australia (3); total of 98 species at 53 sites pooled into 5 mesoscale SADs	−39.1 −37.6	141.4 149.8	Intertidal	UVS	T.D.O. tohara@museum.vic.gov.au
North Sea Invertebrates (NSI)*	Benthic invertebrates from the North Sea; total of 244 species at 46 sites pooled into 6 mesoscale SADs	54.3 60.4	−1.0 8.0	38 115	vanVeen grab	U.S. Ulrike.Schueckel@senckenberg.de

Table S1. Cont.

Dataset name	Summary	Latitudinal limits	Longitudinal limits	Depth range, m	Sampling method	Data contact
Norwegian Shelf Macrobenthos (NSM)†	Benthic invertebrates collected along the Norwegian Shelf (4); total of 805 species at 101 sites pooled into 4 mesoscale SADs	56.0 71.8	1.7 23.5	65 434	vanVeen grab	K.E.E. Kari.Ellingsen@nina.no
Antarctic Isopods (ANI)‡	Isopods from the Southern Ocean (5); total of 502 species at 38 sites pooled into 8 mesoscale SADs	−71.3 −58.2	0.0 −64.7	774 6,348	Epibenthic sledge	A.B. abrandt@zoologie.uni-hamburg.de

SAD, species abundance distribution; UVS, underwater visual survey (belt transects in all cases).

1. Sweatman H, et al. (2008) Long-Term Monitoring of the Great Barrier Reef (Australian Institute of Marine Science, Townsville, Australia), Status Report no. 8.
2. Gray JS, et al. (1997) Coastal and deep-sea benthic diversities compared. *Mar Ecol Prog Ser* 159:97–103.
3. O'Hara TD, Addison PFE, Gazzard R, Costa TL, Pocklington JB (2010) A rapid biodiversity assessment methodology tested on intertidal rocky shores. *Aquat Conserv* 20(4):452–463.
4. Ellingsen KE, Gray JS (2002) Spatial patterns of benthic diversity: Is there a latitudinal gradient along the Norwegian continental shelf? *J Anim Ecol* 71(3):373–389.
5. Brandt A, et al. (2007) First insights into the biodiversity and biogeography of the Southern Ocean deep sea. *Nature* 447(7142):307–311.

Table S2. Summary of the abundance distributions analyzed, at each of the three scales

	Site				Mesoscale				Regional			
Dataset	N*	S†	MRA‡	PS§	N*	S†	MRA‡	PS§	N*	S†	MRA‡	PS§
Antarctic Isopods (ANI)	310 (1–2,651)	40.2 (1–91)	0.24 (0.05–1.00)	0.49 (0.21–1.00)	1,474 (284–3,531)	141.4 (73–193)	0.13 (0.08–0.24)	0.36 (0.23–0.44)	11,788	502	0.05	0.17
Antarctic Molluscs (ANM)	109 (1–560)	14.0 (1–33)	0.39 (0.17–1.00)	0.45 (0.10–1.00)	544 (353–816)	41.0 (27–79)	0.34 (0.12–0.56)	0.37 (0.21–0.48)	2,175	96	0.21	0.33
Tuscany Archipelago Fish (TAP)	577 (223–1,118)	21.2 (16–26)	0.49 (0.29–0.74)	0.17 (0.04–0.33)	4,326 (3,276–5,867)	32.2 (30–34)	0.50 (0.48–0.54)	0.08 (0.06–0.09)	17,303	39	0.51	0.03
Indo-Pacific Coral Commensals (IPC)	336 (71–1,018)	66.9 (25–127)	0.14 (0.09–0.20)	0.46 (0.38–0.62)	NA	NA	NA	NA	644	135	0.06	0.44
SE Australia: Shelf Fish (SEF)	660 (2–10,803)	17.0 (1–36)	0.51 (0.16–1.00)	0.32 (0.00–0.71)	9,598 (112–22,978)	60.8 (20–93)	0.37 (0.19–0.71)	0.24 (0.11–0.38)	124,777	173	0.26	0.16
W Australia: Deep Fish (WAF)	373 (18–3,714)	18.7 (8–32)	0.37 (0.07–0.97)	0.37 (0.09–0.67)	1,055 (44–5,868)	43.8 (12–101)	0.37 (0.10–0.93)	0.35 (0.16–0.59)	24,272	282	0.32	0.22
Scotian Shelf Fish (SSF)	855 (6–21,436)	9.4 (1–21)	0.65 (0.14–1.00)	0.28 (0.00–0.75)	27,963 (3,493–78,019)	37.5 (19–53)	0.56 (0.33–0.90)	0.18 (0.05–0.37)	391,484	98	0.32	0.19
Eastern Bass Strait Invertebrates (EBS)	1,519 (48–8,641)	125.7 (23–384)	0.18 (0.05–0.70)	0.34 (0.00–0.70)	23,794 (16,182–28,145)	573.7 (500–661)	0.06 (0.05–0.07)	0.16 (0.14–0.18)	71,382	801	0.05	0.13
Sunderban Zooplankton (SUZ)	670 (148–1,190)	15.3 (7–28)	0.45 (0.24–0.70)	0.02 (0.00–0.11)	NA	NA	NA	NA	4,689	31	0.36	0.00
Great Barrier Reef Fish (GBR)	591 (64–2,535)	44.4 (7–66)	0.32 (0.08–0.87)	0.41 (0.14–0.61)	5,471 (686–9,315)	116.9 (78–144)	0.20 (0.14–0.30)	0.21 (0.14–0.33)	43,768	195	0.15	0.11
Central Pacific Reef Fish (CPF)	9,677 (548–85,423)	104.1 (32–183)	0.30 (0.11–0.61)	0.20 (0.08–0.39)	94,830 (35,819–223,060)	238.0 (171–293)	0.30 (0.16–0.40)	0.13 (0.08–0.16)	474,151	491	0.19	0.09
Norwegian Shelf Macrobenthos (NSM)	676 (87–2,677)	86.7 (35–148)	0.15 (0.07–0.51)	0.38 (0.20–0.63)	17,074 (5,201–42,572)	357.5 (176–552)	0.08 (0.04–0.11)	0.24 (0.19–0.29)	68,298	805	0.08	0.19
North Sea Invertebrates (NSI)	806 (92–4,608)	44.4 (28–77)	0.39 (0.09–0.87)	0.35 (0.08–0.51)	6,183 (1,224–12,361)	101.2 (54–147)	0.41 (0.10–0.83)	0.23 (0.09–0.37)	37,097	244	0.29	0.21
Bass Strait Intertidal (BSI)	31,148 (1,200–116,444)	30.2 (15–44)	0.61 (0.26–0.95)	0.14 (0.00–0.33)	330,168 (96,743–643,500)	61.6 (54–76)	0.47 (0.34–0.78)	0.07 (0.05–0.09)	1,650,842	98	0.39	0.11

Values reported are means, with ranges in parentheses. Due to the small number of sites, there are no mesoscale abundance distributions for IPC and SUZ (*Materials and Methods*). Because the regional abundance distribution pools all samples in that dataset, there is only one abundance distribution, so no range of values is reported.

*N, number of individuals sampled.

†S, number of species appearing in the sample.

‡MRA, maximum relative abundance (abundance of the most abundant species, as a proportion of the total number of individuals in the abundance distribution).

§PS, proportion of species sampled that are singletons (i.e., abundance = 1).

Table S3. Groupings used and model selection for single-guild analysis

Dataset	Guild name	No. of species in group	% of sites fitted at site level	% support for lognormal		
				Site	Mesoscale	Regional
Antarctic Isopods (ANI)	Detritus feeders	486	82	**>0.9999**	**>0.9999**	**>0.9999**
Antarctic Molluscs (ANM)	Suspension feeders	24	0	NA	0.7048	0.8639
Tuscany Archipelago Fish (TAP)	Benthic feeders	38	100	**>0.9999**	**0.9999**	**0.9973**
Indo-Pacific Coral Crustaceans (IPC)	Decapods	334	100	**>0.9999**	NA	**>0.9999**
SE Australia: Shelf Fish (SEF)	Invertivores, benthic prey	101	57	**>0.9999**	**0.9999**	0.8741
W Australia: Deep Fish (WAF)	Invertivores, benthic prey	121	38	0.0199	**>0.9999**	**0.9958**
Scotian Shelf Fish (SSF)	Invertivores	48	6	0.0727	**>0.9999**	**0.9998**
Eastern Bass Strait Invertebrates (EBS)	Deposit feeders	347	91	**>0.9999**	**>0.9999**	0.2905
Sunderban Zooplankton (SUZ)	Planktivores	27	100	**>0.9999**	NA	**0.9772**
Great Barrier Reef Fish (GBR)	Herbivores	60	53	**>0.9999**	**>0.9999**	**0.9515**
Central Pacific Reef Fish (CPF)	Invertivores	164	100	**>0.9999**	**0.9998**	0.8776
Norwegian Shelf Macrobenthos (NSM)	Deposit feeders, Malacostraca only	76	4	0.6079	**0.9902**	0.7062
North Sea Invertebrates (NSI)	Deposit feeders	74	74	0.6953	0.0902	0.0593
Bass Strait Intertidal (BSI)	Grazers	49	100	**>0.9999**	**>0.9999**	**0.9937**
Overall				**>0.9999**	**>0.9999**	**>0.9999**

Percentage support values indicate relative support for the Poisson-lognormal over Poisson-gamma model fitted to the species abundance data at three scales: site level, mesoscale, and regional. Each row represents a different dataset. For ANM, there were no sites with more than five distinct species abundance values for the most species-rich functional group, so model selection was only done at the mesoscale and regional scale. For IPC and SUZ, there were too few species abundance distributions to create mesoscale groupings. The last row is an overall test, based on summing the log-likelihoods across all datasets. Where lognormal model has at least 95% support, the model's weight is shown in bold. Where Poisson-gamma model has at least 95% support, the model's weight is shown underlined.

Table S4. Mixed-effects and OLS regression model results for analysis of standardized relative support for the Poisson-lognormal

Explanatory variable	Overall slope[†] ± SE	t[‡]	Marginal R^2 [§]	Pseudo-R^2	AIC
Site level					
Intercept only	NA	NA	0.00	0.08	−1,879.6
Log(no. of distinct abundances)	−0.009 ± 0.015	−0.58	<0.01	0.11	−1,883.2
Maximum relative abundance	0.386 ± 0.020	19.66***	0.47	0.60	−2,641.5
Proportion of singletons	0.074 ± 0.051	1.45	0.01	0.14	−1,913.6
Mesoscale					
Intercept only	NA	NA	0.00	0.00	−186.4
Log(no. of distinct abundances)	−0.009 ± 0.013	−0.74	0.01	0.01	−181.0
Maximum relative abundance	0.261 ± 0.051	5.08***	0.24	0.71	−270.8
Proportion of singletons	0.046 ± 0.074	0.63	<0.01	<0.01	−180.8
Regional	Slope ± SE	t	R^2		AIC
Intercept only	NA	NA	0.00		−43.3
Log(no. of distinct abundances)	−0.039 ± 0.014	−2.83*	0.40		−48.5
Maximum relative abundance	0.288 ± 0.043	6.75***	0.79		−63.3
Proportion of singletons	−0.160 ± 0.108	−1.48	0.16		−43.7

Except for Akaike's information criterion (AIC), which was calculated from maximum-likelihood fits, all values reported in the table were obtained using restricted maximum likelihood (REML).

[†] "Overall slope" refers to the fixed effects component of the model.

[‡] The t statistic for the slope parameter. The number of asterisks indicates the level of statistical significance: *$0.01 < P < 0.05$, **$0.001 < P < 0.01$, and ***$P < 0.001$.

[§] Percentage of variation explained by the fixed effect only.

[¶] Calculated from the residuals of the full fitted model $(1-(\sigma^2_{resid}/\sigma^2_{tot}))$.