Supplementary Information

Robust estimation of microbial diversity in theory and in practice

B. Haegeman, J. Hamelin, J. Moriarty, P. Neal, J. Dushoff, J. S. Weitz

Supplementary Text

Text S1 Contribution of rare species to rarefaction curve
Text S2 Contribution of rare species to Hill diversities
Text S3 Hill diversities and rarefaction curve
Text S4 Estimating species abundances from sample data
Text S5 Estimating Hill diversities from sample data

Supplementary Tables

Table S1 Description of communities used in Figure 2
Table S2 Data for empirically-sampled microbial communities

Supplementary Figures

Figure S1 Sample data are insensitive to rare species tail of community
Figure S2 Hill diversity for large $\alpha$ is insensitive to rare species tail
Figure S3 Rank-abundances curve of empirical microbial community samples
Figure S4 Community-size dependence of Hill diversity estimates

Computer code

Matlab code to compute Hill diversity estimates
**Supplementary Text**

**Text S1**

**Contribution of rare species to rarefaction curve**

We define $S_m$ as the expected number of species in a sample of $m$ individuals taken from the community. The rarefaction curve of the community is the plot of the number of species $S_m$ as a function of the sample size $m$. We consider a community consisting of $S$ species with relative abundance $p_1, p_2, \ldots, p_S$. Then the expected number of sampled species $S_m$ is given by

$$S_m = \sum_{i=1}^{S} \left( 1 - (1 - p_i)^m \right). \quad (S1)$$

It is important to distinguish the community rarefaction curve (S1) from the rarefaction curve estimated from sample data. We consider a sample of size $M$ taken from the community. We denote the number of species observed in the sample by $S_{\text{obs}}$, and the number of species with abundance $k$ in the sample by $F_k$. For $m \leq M$ the rarefaction curve $S_m$ can be estimated by taking subsamples of size $m$ out of the sample. The average number of species observed in the subsample (averaged over all subsamples of size $m$) is an estimator for $S_m$,

$$\hat{S}_m = \sum_{k \geq 1} F_k \left( 1 - \frac{(M-k)}{(m)} \right), \quad m \leq M. \quad (S2)$$

This estimator is reliable in the sense that it is unbiased (that is, the expected value of $\hat{S}_m$ is equal to $S_m$). Moreover, there is no other unbiased estimator with smaller variance. For $m > M$ the estimation of the rarefaction curve is necessarily based on extrapolation, leading to less reliable estimates, especially for $m \gg M$.

We define a species to be rare if its relative abundance is much smaller than $\frac{1}{50M}$. This means that a rare species is unlikely to be present in the sample (of size $M$). For concreteness we say that

$$\text{species } i \text{ is rare if } p_i \leq \frac{1}{50M}. \quad (S3)$$
Note that our definition of rarity depends on the sample size $M$. The choice of a threshold for rarity is arbitrary, though our results are robust to changes in the constant (which in this case has been set to 50) so long as it is much greater than 1.

We consider the rarefaction curve (S1) up to sample size $M$. The contribution of species $i$ can be written as

$$1 - (1 - p_i)^m = \sum_{j=1}^{m} \left( \begin{array}{c} m \\ j \end{array} \right) p_i^j (1 - p_i)^{m-j}, \quad m \leq M.$$ 

The $j$-th term in this sum is the probability that species $i$ is represented $j$ times in a sample of size $m$. For a rare species $i$ we have $p_i \ll \frac{1}{m} \leq \frac{1}{m}$, and the first term dominates the other terms. Hence,

$$1 - (1 - p_i)^m \approx m p_i (1 - p_i)^{m-1} \approx m p_i, \quad m \leq M.$$ 

Partitioning the set of species into rare and non-rare species, we get

$$S_m \approx \sum_{i=1}^{S} \left( 1 - (1 - p_i)^m \right) + \sum_{i=1}^{S} m p_i$$

$$= \sum_{i=1}^{S} \left( 1 - (1 - p_i)^m \right) + m p_{\text{rare}}, \quad m \leq M,$$ (S4)

with $p_{\text{rare}}$ the total relative abundance of the set of rare species in the community.

From Equation (S4) it follows that the rarefaction curve does not depend on the abundance distribution of the rare species, but only on the total abundance of the rare species. This follows directly from Definition (S3): it is unlikely that a rare species will be observed twice in a sample of size $m$ (when $m < M$). Therefore, the contribution of the rare species to the sample species richness depends only on their prevalence in the sample which, in turn, depends only on their prevalence in the community. In particular, rarefaction curves obtained for different abundance distributions of the rare species are indistinguishable, see Figure S1.
Text S2

Contribution of rare species to Hill diversities

In the main text we have introduced the Hill diversities $D_\alpha$,  

$$D_\alpha = \left( \sum_{i=1}^{S} p_{i}^{\alpha} \right)^{\frac{1}{1-\alpha}}. \quad (S5)$$

The Hill diversity of order 1 is defined as the limit $D_1 = \lim_{\alpha \to 1} D_\alpha$, and is related to the Shannon diversity index $H$,  

$$D_1 = e^{H} \quad \text{with} \quad H = \sum_{i=1}^{S} -p_i \ln p_i. \quad (S6)$$

The Hill diversity of order 2 is related to the Simpson concentration index $C$,  

$$D_2 = \frac{1}{C} \quad \text{with} \quad C = \sum_{i=1}^{S} p_i^2.$$  

The Hill diversity of order $\infty$ is related to the relative abundance $p_{\text{max}}$ of the most abundant species,  

$$D_\infty = \frac{1}{p_{\text{max}}} \quad \text{with} \quad p_{\text{max}} = \max \{ p_1, p_2, \ldots, p_S \}.$$  

We consider a community in which the rare species occupy a fraction $p_{\text{rare}}$ of the total community abundance. We study the dependence of the Hill diversity on the number of rare species $S_{\text{rare}}$. Assuming that the rare species have equal abundance, we get  

$$D_\alpha = \left( \sum_{i=1}^{S_{\text{non-rare}}} p_i^{\alpha} + \sum_{i=1}^{S_{\text{rare}}} p_i^{\alpha} \right)^{\frac{1}{1-\alpha}} \quad (S7)$$

The first term inside the brackets contains the contribution of the non-rare species. The second term inside the brackets, $p_{\text{rare}}^\alpha S_{\text{rare}}^{1-\alpha}$, contains the contribution of the rare species. The contribution of the
non-rare species is independent of $S_{\text{rare}}$. For $\alpha > 1$ the contribution of the rare species decreases with $S_{\text{rare}}$ and vanishes for $S_{\text{rare}} \to \infty$. Hence, the rare species contribute only weakly to the Hill diversity $D_\alpha$ for $\alpha > 1$. For $\alpha < 1$ the contribution of the rare species increases with $S_{\text{rare}}$ and diverges for $S_{\text{rare}} \to \infty$. Hence, for sufficiently large $S_{\text{rare}}$ the rare species contribution dominates the Hill diversity $D_\alpha$ for $\alpha < 1$. Note that the relative contribution of the rare to the non-rare species has a power-law dependence on $S_{\text{rare}}$ with exponent $1 - \alpha$. For the Hill diversity $D_1$ the relative contribution of the rare to the non-rare species has a logarithmic dependence on $S_{\text{rare}}$, see (S6).
Hill diversities and rarefaction curve

We follow Mao (2007) to establish a link between the rarefaction curve \( S_m \) and the Hill diversities \( D_\alpha \).

Rewriting the sum \( \sum_i p_i^\alpha \), we get

\[
\sum_{i=1}^S p_i^\alpha = \sum_{i=1}^S (1 - (1 - p_i))^\alpha \\
= \sum_{i=1}^S \sum_{m=0}^\infty \frac{(-1)^m \Gamma(\alpha + 1)}{m! \Gamma(\alpha - m + 1)} (1 - p_i)^m \\
= S \sum_{m=0}^\infty \frac{(-1)^m \Gamma(\alpha + 1)}{m! \Gamma(\alpha - m + 1)} - \sum_{m=0}^\infty \frac{(-1)^m \Gamma(\alpha + 1)}{m! \Gamma(\alpha - m + 1)} \sum_{i=1}^S (1 - (1 - p_i)^m) \\
= \sum_{m=1}^\infty \frac{\alpha \Gamma(m - \alpha)}{m! \Gamma(1 - \alpha)} S_m \\
= \sum_{m=1}^\infty \frac{\alpha \Gamma(m - \alpha)}{m! \Gamma(1 - \alpha)} S_m
\]

where \( \Gamma \) denotes the gamma function. Hence,

\[
D_\alpha = \left( \sum_{m=1}^\infty \frac{\alpha \Gamma(m - \alpha)}{m \Gamma(1 - \alpha)} S_m \right)^{-\frac{1}{1-\alpha}}. \tag{S8}
\]

We express the link with the rarefaction curve in terms of the Tsallis entropies \( T_\alpha \) (Tsallis, 1988),

\[
T_\alpha = \frac{1}{1-\alpha} \left( \sum_{i=1}^S p_i^\alpha - 1 \right),
\]

which is closely related to the Hill diversities \( D_\alpha \),

\[
D_\alpha = (1 + (1 - \alpha) T_\alpha)^{-\frac{1}{1-\alpha}}. \tag{S9}
\]
Equation (S8) becomes

\[ T_\alpha = \frac{1}{1 - \alpha} \left( \alpha - 1 + \sum_{m=2}^{\infty} \frac{\alpha \Gamma(m - \alpha)}{m! \Gamma(1 - \alpha)} S_m \right) \]

\[ = -1 + \sum_{m=2}^{\infty} \frac{\alpha \Gamma(m - \alpha)}{m! \Gamma(2 - \alpha)} S_m. \]

We study the behavior of the coefficients \( c_m \) in this infinite sum,

\[ c_m = \frac{\alpha \Gamma(m - \alpha)}{m! \Gamma(2 - \alpha)} \]

For \( \alpha \in (0, 2) \) all coefficients \( c_m \) are positive, and

\[ c_m \sim m^{-(\alpha+1)} \quad \text{as} \quad m \to \infty. \] (S10)

This shows that different Tsallis entropies \( T_\alpha \) depend on different parts of the rarefaction curve \( S_m \). For \( \alpha \) close to 2, the Tsallis entropy \( T_\alpha \) is mainly determined by the rarefaction curve for small \( m \). For decreasing \( \alpha \), the contribution of the rarefaction curve for large \( m \) increases. For the limit cases \( \alpha \to 0 \) and \( \alpha \to 2 \) the constant of proportionality in (S10) vanishes. For \( \alpha = 2 \) we have \( T_2 = 1 - C = S_2 - 1 \): the only contribution of the rarefaction curve is at \( m = 2 \). For \( \alpha = 0 \) we have \( T_0 = S - 1 = S_\infty - 1 \): the contribution of the rarefaction curve is entirely shifted to \( m \to \infty \). This analysis also holds for the Hill diversities \( D_\alpha \) because \( D_\alpha \) is an increasing function of \( T_\alpha \), see (S9).

As an illustration, we apply (S8) to a community with a power-law tail. That is, we consider an artificial community consisting of an infinite number of species, for which the species are arranged in decreasing order of abundance, and for which

\[ p_i \sim i^{-z} \quad \text{as} \quad i \to \infty. \]

The abundances should be summable, so we have to impose that \( z > 1 \). The tail of the abundance distribution determines the asymptotic behavior of the rarefaction curve,

\[ S_m \sim m^{1/z} \quad \text{as} \quad m \to \infty. \]
From (S8) and (S10) it follows that the diversity $D_\alpha$ is finite for $\alpha > \frac{1}{z}$, and diverges for $\alpha \leq \frac{1}{z}$. This can be checked directly from Definition (S5).
Estimating species abundances from sample data

The Good-Turing estimators (Good, 1953) are a well-known family of frequency estimators. Here we present a compact derivation, given in Nádas (1985), which demonstrates that the Good-Turing estimators are non-parametric, that is, free of assumptions about the abundance distribution.

Let $\Theta$ be a random variable taking values between 0 and 1, with a distribution function $G(\theta)$ about which nothing is known. Suppose that $R$ is another random variable whose conditional distribution $p_M(r|\theta)$, when $\Theta$ has the value $\theta$, is binomial with parameters $M$ and $\theta$,

$$p_M(r|\theta) = \binom{M}{r} \theta^r (1 - \theta)^{M-r}. \quad (S11)$$

Then we have the identity

$$\theta p_M(r|\theta) = \frac{r+1}{M+1} p_{M+1}(r+1|\theta) \quad (S12)$$

Suppose now that we wish to estimate the value of $\theta$ given that $R$ is observed to take the value $r$. Taking a Bayesian approach with prior distribution $G$, the posterior mean for $\theta$ is

$$E[\theta|R = r] = \frac{r+1}{M+1} \frac{p_{M+1}(r+1)}{p_M(r)} \quad (S13)$$

where $p_M$ is the unconditional probability mass function of $R$ (that is, integrated out over $G$). This derivation is non-parametric in that $G$ is not only unknown, but no assumptions are made about $G$: the probability mass function $p_M$ must therefore be estimated directly from the sample data, so that we are in fact performing empirical Bayes estimation.

In the context of diversity estimation, we regard $G$ as the community abundance distribution, $\theta$ as the species abundance to be estimated and $r$ as the number of times that this species occurs in the sample. We use the maximum likelihood estimates for $p_M(r)$ and $p_{M+1}(r+1)$ given by $F_r/M$ and $F_{r+1}/(M+1)$, respectively. Plugging the estimates into (S13) and assuming that $M \gg 1$, we get the estimated community abundance $\hat{\theta}_r$ of a species observed $r$ times in the sample,

$$\hat{\theta}_r = \frac{r+1}{M} \frac{F_{r+1}}{F_r}, \quad (S14)$$
which are the Good-Turing frequency estimators.

As a corollary of (S14) we get the estimator for the total abundance of the observed species,

$$\sum_{r \geq 1} F_r \hat{q}_r = \frac{1}{M} \sum_{r \geq 1} (r + 1) F_{r+1} = \frac{M - F_1}{M},$$

so that the total abundance $p_{\text{unobs}}$ of the unobserved species is estimated as

$$\hat{p}_{\text{unobs}} = \frac{F_1}{M} \quad (S15)$$

In words, the total relative abundance of unobserved species in the community is estimated as the total relative abundance of singletons in the sample.
Text S5

Estimating Hill diversities from sample data

We construct estimators for the Hill diversity $D_\alpha$ based on a sample of size $M$ taken from the community. Our strategy consists in first estimating the rarefaction curve $S_m$ and then using the link (S8) between $D_\alpha$ and $S_m$.

The estimation of the rarefaction curve decomposes into two parts. For the part $m \leq M$ the rarefaction curve can be estimated unbiasedly using the estimator (S2). For the part $m > M$ the sample data have to be extrapolated, and no unbiased estimator exists. We denote the relative abundances of the unobserved species by $q_1, q_2, \ldots$ (there are $S - S_{\text{obs}}$ unobserved species). If we knew the abundances $q_i$, then we could compute the rarefaction curve using the formula,

$$\hat{S}_m = S_{\text{obs}} + \sum_{i \geq 1} \left(1 - (1 - q_i)^{m-M}\right) \quad m > M.$$  \hfill (S16)

As we have argued in the main text, the sample data contain little information about the abundances $q_i$ of unobserved species. However, the Good-Turing estimator (S15) for the total abundance $p_{\text{unobs}} = \sum_{i \geq 1} q_i$ of the unobserved species is available. It follows from (S16) that the estimation of the rarefaction curve $S_m$ for $m > M$ reduces to distributing the estimated abundance $\hat{p}_{\text{unobs}}$ over the individual unobserved species.

We work out two scenarios, see Figure 3 of the main text. In the first scenario we distribute $\hat{p}_{\text{unobs}}$ so as to obtain the lowest possible value of the diversity $D_\alpha$ consistent with the sample data. By this we mean that $\hat{p}_{\text{unobs}}$ must be distributed in a manner which remains consistent with the estimates $\hat{\theta}_r$. The lowest diversity occurs when all unobserved species have the same abundance, $q_1 = q_2 = \ldots = q^-$, and this abundance is as high as possible. However, as noted in Good (1953), the frequency estimates $\hat{\theta}_r$ must increase as $r$ increases: this implies an upper bound for $q^-$, namely $\hat{\theta}_1$ (which is the estimated community abundance of any species observed exactly once in the sample). We therefore take $q^- = \hat{\theta}_1 = \frac{2F_2}{MF_1}$ so that, from (S15), there are $\frac{F_2^2}{MF_1}$ unobserved species. Hence, the estimated rarefaction curve (S16) becomes

$$\hat{S}_m = S_{\text{obs}} + \frac{F_2^2}{2MF_1} \left(1 - \left(1 - \frac{2F_2}{MF_1}\right)^{m-M}\right) \quad m > M.$$  \hfill (S17)
where the superscript in $\hat{S}_m^-$ indicates the low-diversity scenario.

In the second scenario we distribute $\hat{\pi}_{\text{unobs}}$ so as to obtain the highest possible value of the diversity $D_\alpha$. The highest diversity is obtained when all unobserved species have the same abundance, $q_1 = q_2 = \ldots = q^+$, and this abundance is as small as possible. The smallest abundance a species can have in a community of size $N$ is equal to $\frac{1}{N}$, corresponding to a species represented by a single individual. We therefore take $q^+ = \frac{1}{N}$ so that, from (S15), there are $\frac{NF_1}{M}$ unobserved species. Hence, the estimated rarefaction curve (S16) becomes

$$\hat{S}_m^+ = S_{\text{obs}} + \frac{NF_1}{M} \left( 1 - \left( 1 - \frac{1}{N} \right)^{m-M} \right) \quad m > M,$$

(S18)

where the superscript in $\hat{S}_m^+$ indicates the high-diversity scenario. Note that the upper estimator (S18) depends on the community size $N$, in contrast to the estimator (S17).

To summarize, we have obtained two estimators for the Hill diversity $D_\alpha$, a lower estimate $\hat{D}_\alpha^-$ and an upper estimate $\hat{D}_\alpha^+$. They can be computed as follows:

**Lower estimate** First, compute the lower estimate of the rarefaction curve. From (S2) and (S17),

$$\hat{S}_m^- = \begin{cases} 
\sum_{k \geq 1} F_k \left( 1 - \left( \frac{m-k}{m} \right) \right) & \text{if } m = 1, 2, \ldots, M \\
S_{\text{obs}} + \frac{F_1^2}{2M} \left( 1 - \left( 1 - \frac{2F_1}{M} \right)^{m-M} \right) & \text{if } M + 1, M + 2, \ldots
\end{cases}$$

(S19)

Then, substitute this result into (S8) to estimate the Hill diversity,

$$\hat{D}_\alpha^- = \left( \sum_{m=1}^{\infty} \frac{\alpha \Gamma(m-\alpha)}{m! \Gamma(1-\alpha)} \hat{S}_m^- \right)^{\frac{1}{\alpha}}.$$ (S20)

**Upper estimate** First, compute the upper estimate of the rarefaction curve. From (S2) and (S18),

$$\hat{S}_m^+ = \begin{cases} 
\sum_{k \geq 1} F_k \left( 1 - \left( \frac{m-k}{m} \right) \right) & \text{if } m = 1, 2, \ldots, M \\
S_{\text{obs}} + \frac{NF_1}{M} \left( 1 - \left( 1 - \frac{1}{N} \right)^{m-M} \right) & \text{if } M + 1, M + 2, \ldots
\end{cases}$$

(S21)
Then, substitute this result into (S8) to estimate the Hill diversity,

$$
\hat{D}_+^+ = \left( \sum_{m=1}^{\infty} \frac{\alpha \Gamma(m-\alpha)}{m! \Gamma(1-\alpha)} \hat{S}_m^+ \right)^{\frac{1}{1-\alpha}}.
$$

(S22)

The Matlab code to compute the Hill diversity estimates \( \hat{D}_{-\alpha} \) and \( \hat{D}_{+\alpha} \) is part of the Supplementary Information.

We discuss three properties of the estimators \( \hat{D}_{-\alpha} \) and \( \hat{D}_{+\alpha} \) that follow directly from their definitions. First, the lower estimate \( \hat{D}_{-\alpha} \) generalizes Chao’s estimator for species richness,

$$
\hat{D}_{-0} = \hat{S}_\infty = S_{\text{obs}} + \frac{F_1^2}{2F_2}.
$$

Note that the lower estimate, like Chao’s estimator, only gives meaningful results if the number of species observed once or twice in the sample is sufficiently large, and at least \( F_2 > 0 \). These conditions are typically satisfied in practice, especially for highly diverse communities.

Second, the upper estimate \( \hat{D}_{+\alpha} \) depends on community size \( N \), which is typically several orders of magnitude larger than sample size \( M \). It is therefore instructive to consider the limit \( N \to \infty \). A computation analogous to the one in Text S2 shows that the upper estimate \( \hat{D}_{+\alpha} \) diverges as \( N^{1-\alpha} \) for \( \alpha < 1 \), and as \( \log N \) for \( \alpha = 1 \). Hence, we expect large values of the upper estimate (and therefore large estimation uncertainty) for \( \alpha < 1 \), especially for \( \alpha \) close to zero (that is, close to species richness).

Third, the estimators \( \hat{D}_{-\alpha} \) and \( \hat{D}_{+\alpha} \) coincide for the Simpson diversity. The Simpson diversity \( D_2 \) is the only Hill diversity \( D_\alpha \) that does not depend on the extrapolation of the rarefaction curve. It is a function of the rarefaction curve at \( m = 2 \): \( D_2 = \frac{1}{2-D_2} \). Because the initial part of the estimated rarefaction curve is the same for the lower and upper estimate, the Simpson diversity estimates are equal, \( \hat{D}_{-2} = \hat{D}_{+2} \). The Simpson diversity is not sensitive to the extrapolation of the rarefaction curve, and therefore easy to estimate.


## Supplementary Tables

### Table S1

Table S1: Description of communities used in Figure 2. Communities C1, C2 and C3 have a power-law abundance distribution, with parameters $S$, the number of species in the community, and $z$, the exponent of the power-law. The Hill diversity of order $\alpha = 0$ is equal to the number of species, $D_0 = S$; the Hill diversity of order $\alpha = 1$ is the Shannon diversity; the Hill diversity of order $\alpha = 2$ is the Simpson diversity. For a sample of size $2 \times 10^4$, the number of observed species is denoted by $S_{\text{obs}}$ and Chao’s estimator for species richness is denoted by $\hat{S}_{\text{Chao}}$.

<table>
<thead>
<tr>
<th>Community</th>
<th>$S$</th>
<th>$z$</th>
<th>$D_1$</th>
<th>$D_2$</th>
<th>$S_{\text{obs}}$</th>
<th>$\hat{S}_{\text{Chao}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>$5 \times 10^4$</td>
<td>1.1</td>
<td>640</td>
<td>35</td>
<td>$4.8 \times 10^3$</td>
<td>$1.5 \times 10^4$</td>
</tr>
<tr>
<td>C2</td>
<td>$2 \times 10^5$</td>
<td>1.3</td>
<td>100</td>
<td>11</td>
<td>$2.4 \times 10^3$</td>
<td>$8.3 \times 10^3$</td>
</tr>
<tr>
<td>C3</td>
<td>$10^6$</td>
<td>1.6</td>
<td>15</td>
<td>4.5</td>
<td>690</td>
<td>$1.8 \times 10^3$</td>
</tr>
</tbody>
</table>
Table S2

Table S2: Data for empirically-sampled microbial communities. We report the sample size $M$, the number of species observed in the sample $S_{\text{obs}}$, the number of singleton species $F_1$, that is, the number of species that have been sampled only once, the estimated relative abundance of the unobserved species $\hat{p}_{\text{unobs}}$, and the Chao estimate $\hat{S}_{\text{Chao}}$ for the number of species in the community. The data sets are taken from Quince et al. (2008): a seawater bacterial sample from the upper ocean (Rusch et al., 2007), soil bacterial samples at four locations: Brazil, Florida, Illinois and Canada (Roesch et al., 2007), and seawater samples from deep-sea vents at two locations: FS312 and FS396, separated into bacteria and archaea (Huber et al., 2007).

<table>
<thead>
<tr>
<th></th>
<th>$M$</th>
<th>$S_{\text{obs}}$</th>
<th>$F_1$</th>
<th>$\hat{p}_{\text{unobs}}$</th>
<th>$\hat{S}_{\text{Chao}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>upper ocean</td>
<td>7068</td>
<td>811</td>
<td>311</td>
<td>0.044</td>
<td>1038</td>
</tr>
<tr>
<td>soil, Brazil</td>
<td>26079</td>
<td>2880</td>
<td>1176</td>
<td>0.045</td>
<td>4604</td>
</tr>
<tr>
<td>soil, Florida</td>
<td>28150</td>
<td>3440</td>
<td>1541</td>
<td>0.055</td>
<td>5643</td>
</tr>
<tr>
<td>soil, Illinois</td>
<td>31621</td>
<td>3357</td>
<td>1466</td>
<td>0.046</td>
<td>5745</td>
</tr>
<tr>
<td>soil, Canada</td>
<td>52773</td>
<td>5515</td>
<td>2634</td>
<td>0.050</td>
<td>10394</td>
</tr>
<tr>
<td>FS312, bacteria</td>
<td>442062</td>
<td>12183</td>
<td>5339</td>
<td>0.012</td>
<td>19568</td>
</tr>
<tr>
<td>FS312, archaea</td>
<td>200199</td>
<td>1594</td>
<td>460</td>
<td>0.002</td>
<td>2175</td>
</tr>
<tr>
<td>FS396, bacteria</td>
<td>247826</td>
<td>5843</td>
<td>2825</td>
<td>0.011</td>
<td>10570</td>
</tr>
<tr>
<td>FS396, archaea</td>
<td>16428</td>
<td>418</td>
<td>158</td>
<td>0.010</td>
<td>630</td>
</tr>
</tbody>
</table>
Figure S1: Sample data are insensitive to rare species tail of community. We generated three community abundance distributions, shown in red, blue and green (panels a–c). The three communities have the same abundance distribution for species with relative abundance above $10^{-6}$ (the part of the rank-abundance curve to the left of the dashed black line). This common part consists of $6 \times 10^3$ species, occupying 99% of the community abundance. The communities differ in the tail of rare species: the community in panel a has $1.6 \times 10^4$ species; the community in panel b has $1.6 \times 10^5$ species; the community in panel c has $10^6$ species. Despite the marked differences, the rarefaction curves of the three communities up to sample size $2 \times 10^4$ are identical (see panel d). This observation holds generally: any set of rare species leads to the same rarefaction curve if each rare species has relative abundance below $10^{-6}$ and the total relative abundance of the set of rare species equals 0.01.
Figure S2

Figure S2: Hill diversity for large $\alpha$ is insensitive to rare species tail. Panel a: We computed the Hill diversity $D_\alpha$ for the three communities of Figure S1. The Hill diversities for $\alpha > 1$ almost coincide because the communities have the same set of non-rare species. The Hill diversities for $\alpha < 1$ differ because the communities have different rare species tails. Panel b: We computed the Hill diversity $D_\alpha$ for the three communities of Figure 2. The curves of Hill diversities intersect. For small $\alpha$, the most species-rich community (C3, green) has the largest Hill diversity, and the most species-poor community (C1, red) has the smallest Hill diversity. For larger $\alpha$, the most even community (C1, red) has the largest Hill diversity, and the most uneven community (C3, green) has the smallest Hill diversity.
Figure S3

Figure S3: Rank-abundances curve of empirical microbial community samples. Relative abundance in the sample is plotted against species rank in the sample. We used the same data sets as Quince et al. (2008): a seawater bacterial sample from the upper ocean (Rusch et al., 2007), soil bacterial samples at four locations: Brazil, Florida, Illinois and Canada (Roesch et al., 2007), and seawater samples from deep-sea vents at two locations: FS312 and FS396, separated into bacteria and archaea (Huber et al., 2007).
Figure S4: Community-size dependence of Hill diversity estimates. Same data sets as in Figure 5, but for three values of community size $N$. The lower estimate is independent of $N$; the upper estimate increases with increasing $N$ (from left to right: $N = 10^{10}$, $N = 10^{15}$, $N = 10^{20}$). We observe the same behavior as for the in silico generated data sets of Figure 4.
References


