Only Simpson diversity can be estimated accurately from microbial community fingerprints.

Bart Haegeman, Biswarup Sen, Jean-Jacques Godon, Jérôme Hamelin

Only Simpson diversity can be estimated accurately from microbial community fingerprints

Bart Haegeman · Biswarup Sen · Jean-Jacques Godon · Jérôme Hamelin

Received: date / Accepted: date

Abstract Lalande et al. (2013, Microb. Ecol. 66(3):647–658) introduced a promising approach to quantify microbial diversity from fingerprinting profiles. Their analysis is based on extrapolating the abundance of the phylotypes detectable in a fingerprint towards the rare phylotypes of the community. By considering a set of reconstructed communities Lalande et al. obtained a range of estimates for phylotype richness, Shannon diversity and Simpson diversity. They reported narrow ranges indicating accurate estimation, especially for Shannon and Simpson diversity. Here we show that a much larger set of reconstructed communities than the one considered by Lalande et al. is consistent with the fingerprint. We find that the estimates for phylotype richness and Shannon diversity vary over orders of magnitude, but that the estimates for Simpson diversity are restricted to a narrow range (around 10%). We conclude that only Simpson diversity can be estimated accurately from fingerprints.

Keywords denaturing gradient gel electrophoresis (DGGE) · microbial community fingerprinting · phylotype abundance distribution · phylotype richness · Shannon diversity · Simpson diversity · single-strand conformation polymorphism (SSCP)

B. Haegeman
Centre for Biodiversity Theory and Modelling, Centre National de la Recherche Scientifique, Moulis, France. Tel: +33 561040579. Fax: +33 561960851. E-mail: bart.haegeman@ecoex-moulis.cnrs.fr

B. Sen
Department of Environmental Engineering and Science, Feng Chia University, Taichung, Taiwan

J.-J. Godon
INRA, UR50, Laboratoire de Biotechnologie de l'Environnement, Narbonne, France

J. Hamelin
INRA, UR50, Laboratoire de Biotechnologie de l'Environnement, Narbonne, France

A fast, inexpensive and accurate method to measure microbial diversity would be a welcome addition to the toolbox of microbial ecology. Fingerprinting techniques are still considered to be good candidates, but their lack of quantifiability are generally viewed as a major obstacle. A recent paper in this journal by Lalande et al. [1] sheds new light on this problem. Building on previous simulation studies [2,3] they propose a quantitative framework to analyze the estimation properties of diversity metrics from fingerprints.

The framework of Lalande et al. [1] consists of three steps. First, the fingerprinting profile is analyzed: peaks are detected, the area under the peaks is determined and the background, that is, the area under the profile not attributed to any of the peaks, is computed. Peak areas are assumed to represent the abundance of the dominant phylotypes of the community. The background area is assumed to be equal to the total abundance of the other, rare phylotypes.

Second, the community abundance distribution is reconstructed starting from the abundance of the dominant phylotypes. This extrapolation step requires an assumption about the abundance distribution of the rare phylotypes. Different assumptions lead to different reconstructed communities, which may be very dissimilar to the true (and unknown) community. Lalande et al. [1] consider a set of reconstructed communities to quantify the effect of the assumed abundance distribution.

Third, phylotype richness, Shannon diversity and Simpson diversity are computed for each of the reconstructed communities. Each of the computed diversity values are estimates of the true community diversity. If for a diversity metric the range of estimates is wide, then the estimation depends strongly on the assumed abundance distribution and the diversity metric cannot be estimated accurately. A narrow range of estimates indicates that the diversity metric can be estimated accurately. In that case, the range of estimates can be interpreted as a measure of the accuracy with which the diversity metric can be estimated.

Lalande et al. [1] applied this framework to nine *in silico* generated fingerprints. They obtained narrow estimation ranges of the order of $\pm 10\%$ both for Shannon and Simpson diversity, and somewhat wider ranges for phylotype richness. These findings lead to the conclusion that accurate diversity estimation from fingerprints is possible, especially for Shannon and Simpson diversity.

In our opinion the framework proposed by Lalande et al. [1] is a valuable contribution for evaluating the accuracy of diversity estimation from fingerprints. We note that a similar framework was introduced recently to assess diversity estimation from metagenomic data sets [4]. However, we argue that the framework can yield stronger conclusions than those presented in Ref. [1]. In particular, by considering a larger set of reconstructed communities, we show that the estimation range for phylotype richness and for Shannon diversity becomes very wide. Only for Simpson diversity we find a narrow estima-

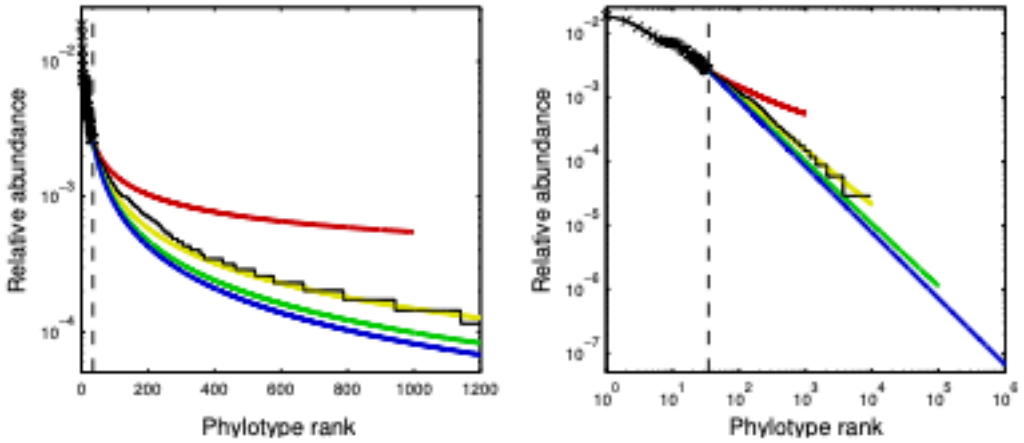

Fig. 1 A variety of reconstructed communities is consistent with a fingerprinting profile. Rank-abundance curves are shown for the dominant phylotypes obtained from the fingerprint (black ×-marks, to the left of the dashed line), for four reconstructed communities of the rare phylotypes (red, yellow, green and blue lines, to the right of the dashed line), and for the "true" community from which the fingerprint was generated (black line). The two panels are identical, except that in the left-hand panel the scale of the x-axis is linear, whereas in the right-hand panel the scale of the x-axis is logarithmic.

tion range. Hence, we are left to conclude that only Simpson diversity can be estimated accurately from fingerprints. This stands in sharp contrast to the conclusions of Ref. [1].

To make our argument, we consider in Figure 1 the data set used in Figure 1 of Ref. [1]. The left-hand panel uses the same axis scaling as Ref. [1] (linear on x-axis, logarithmic on y-axis). The right-hand panel is identical to the left-hand panel except that we use double logarithmic scaling, which is more convenient for our purpose. The fingerprint peak areas are represented as ×-marks. Recall that these areas are assumed to be equal to the abundance of the dominant phylotypes in the community. The black line represents the rank-abundance curve of the data set from which the fingerprint was generated, but which is unavailable for the diversity estimation from the fingerprinting profile.

Four community reconstructions (see Appendix for details) are shown as colored lines (red, yellow, green, blue). Each of these reconstructed communities is consistent with the fingerprinting data. That is, if one would generate a fingerprint of the reconstructed communities, one would get fingerprints that are very similar to the fingerprint we are analyzing. In other words, fingerpinting cannot tell the difference between these four communities. Nevertheless, the structure of these four communities is very different, as shown by their rank-abundance curve.

The yellow community is a realistic reconstruction because it is close to the data set from which the fingerprint was generated (compare yellow and black lines in Figure 1). The other three communities have a qualitatively similar structure, but very different phylotype richness. Are these numbers, such as 10^6 phylotypes for the blue community, realistic? In fact, there is no reason

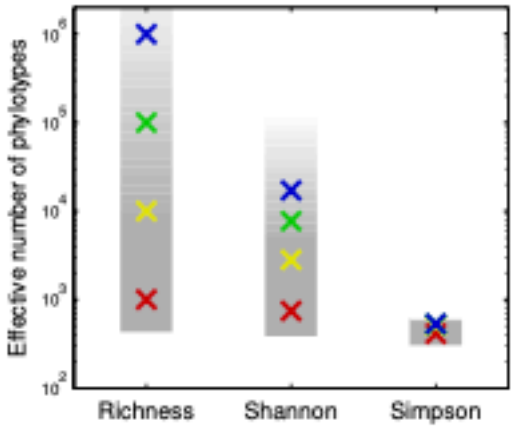

Fig. 2 Estimation ranges differ greatly between diversity metrics. For the four reconstructed communities of Figure 1 (red, yellow, green, blue) we plot phylotype richness (first column), Shannon diversity (second column) and Simpson diversity (third column). The three diversity metrics are expressed as effective numbers of phylotypes. The grey-shaded regions indicate the range of diversity estimates consistent with the fingerprint (see Appendix for details).

to consider them to be unrealistic. Even large metagenomic data sets are not sufficiently informative to rule out such large numbers of phylotypes, as argued in Ref. [4]. Therefore, when analyzing diversity estimation from fingerprints, we should also take into account more extreme reconstructions such as the blue community.

The difference between the reconstructed communities has important consequences for the diversity estimation problem, see Figure 2. For each reconstructed commmunity we plot phylotype richness, Shannon diversity and Simpson diversity. Recall that these values are interpreted as possible diversity estimates. The range of estimates for phylotype richness (from 10^3 to 10^6) and for Shannon diversity (from 700 to $2\ 10^4$) is very wide. This implies that phylotype richness and Shannon diversity cannot be estimated accurately. The range of estimates for Simpson diversity is narrow (from 410 to 530, or $470 \pm 13\%$), implying that Simpson diversity can be estimated accurately.

The above analysis is based on only four reconstructed communities. How sensitive are the results to the choice of these communities? To answer this question, we propose a general analysis in which we take into account *all* communities consistent with the fingerprint (see Appendix for details). We determine lower and upper bounds for the estimation ranges of the three diversity metrics, shown as grey-shaded regions in Figure 2. As before, we find a narrow estimation range for Simpson diversity only, confirming that Simpson diversity, but neither phylotype richness nor Shannon diversity can be estimated accurately from fingerprints. Interestingly, a similar analysis for metagenomics data sets indicated that both Shannon and Simpson diversity, but not phylotype richness can be estimated accurately [4].

To summarize, we have shown that the theoretical framework of Lalande et al. [1] can be extended to reach the following conclusions: (1) phylotype richness and Shannon diversity cannot be estimated accurately from fingerprinting profiles, and (2) Simpson diversity can be estimated with an accuracy of the order of 10%. These conclusions should be relevant for various fingerprinting techniques, such as denaturing gradient gel electrophoresis (DGGE), single-strand conformation polymorphism (SSCP), ribosomal intergenic spacer analysis (RISA) and terminal restriction fragment length polymorphism (T-RFLP).

Appendix

Here we describe the reconstructed communities of Figure 1 and the diversity estimates shown in Figure 2.

First, we extracted the fingerprint peak areas from Figure 1 of Ref. [1]. The total area of the 34 extracted peak equals 20% of the total area under the fingerprinting profile (hence, the peak-to-signal ratio PSR = 0.20 in the terminology of Ref. [1]). The remaining 80% of the area under the profile corresponds to the background (that is, the subpeak background percentage SBP = 0.80 in the terminology of Ref. [2]).

Second, we constructed four communities consistent with the fingerprint data. The 34 most abundant phylotypes correspond to the fingerprint peaks. The relative abundance of these phylotypes is equal to the peak areas divided by the total area under the profile. Hence, the total relative abudance of the most abundant phylotypes is equal to 0.20. We chose the abundance distribution of the rare phylotypes such that the following conditions are satisfied: (1) the total relative abundance of the rare phylotypes is equal to 0.80, and (2) the abundance of a rare phylotype is smaller than the abundance of each of the most abundant phylotypes.

We report the abundance distribution of the rare phylotypes as rank-abundance curves, that is, we give the relationship between relative abundance p_i and rank i for the rare phylotypes (with rank $i > 34$):

- The red community has 10^3 phylotypes. Its rank-abundance curve is quadratic on a log-log plot, $\ln p_i = -3.391 - 0.8554 \ln i + 0.03750 (\ln i)^2$ for $34 < i \leq 10^3$.
- The yellow community has 10^4 phylotypes. Its rank-abundance curve is linear on a log-log plot, $\ln p_i = -2.924 - 0.8535 \ln i$ for $34 < i \leq 10^4$.
- The green community has 10^5 phylotypes. Its rank-abundance curve is linear on a log-log plot, $\ln p_i = -2.492 - 0.9750 \ln i$ for $34 < i \leq 10^5$.
- The blue community has 10^6 phylotypes. Its rank-abundance curve is linear on a log-log plot, $\ln p_i = -2.294 - 1.0306 \ln i$ for $34 < i \leq 10^6$.

For the yellow, green and blue community the abundance distribution of the rare phylotypes is power-law. For the red community this distribution is approximately power-law (the rank-abundance curve is slightly convex, see Figure 1, right-hand panel). For a community with 10^3 phylotypes a power-law distribution for the rare phylotypes does not match smoothly the abundance of the dominant phylotypes.

Third, we computed three diversity metrics for the four reconstructed communities: phylotype richness D_0, Shannon diversity D_1,

$$D_1 = \mathrm{e}^H \qquad \text{with} \quad H = -\sum_i p_i \ln p_i,$$

and Simpson diversity D_2,

$$D_2 = \frac{1}{C} \qquad \text{with} \quad C = \sum_i p_i^2.$$

The notation D_0, D_1 and D_2 refers to Hill diversities of order 0, 1 and 2 (see Ref. [4] for details). Because Hill diversities can be interpreted as effective numbers of phylotypes, they are intercomparable. Therefore we prefer to use the transformed diversity metrics D_1 and D_2 rather than Shannon diversity index H and Simpson concentration index C. We find:

- For red community: $D_0 = 10^3$, $D_1 = 7.4\ 10^2$ and $D_2 = 4.1\ 10^2$.
- For yellow community: $D_0 = 10^4$, $D_1 = 2.8\ 10^3$ and $D_2 = 5.0\ 10^2$.
- For green community: $D_0 = 10^5$, $D_1 = 7.7\ 10^3$ and $D_2 = 5.2\ 10^2$.
- For blue community: $D_0 = 10^6$, $D_1 = 1.7\ 10^4$ and $D_2 = 5.3\ 10^2$.

Finally, we generalized the analysis to a much large set of reconstructed communities. More precisely, we considered *all* reconstructed communities satisfying conditions (1) and (2) above. This set, although it contains unrealistic communities (for example, communities with an abrupt transition from dominant to rare phylotypes), is useful to obtain lower and upper bounds for the estimation range of the diversity metrics. Indeed, it is possible to determine the community in this set yielding the lowest and highest diversity estimates. The lowest diversity estimate is obtained for a community in which the rare phylotypes all have the same abundance as the smallest abundance of the most abundant phylotypes. The highest diversity estimate is obtained for a community in which there are a large number R of rare phylotypes which all have the same relative abundance $0.20/R$.

The results of this further analysis are shown as the grey-shaded regions in Figure 2. The lower end of these regions are equal to the lowest diversity estimate. At the upper end the shade of grey becomes gradually lighter, corresponding to the higest diversity estimate with R ranging from 10^4 to 10^7. It is interesting to note the dependence of the highest diversity estimate on the number of rare phylotypes R for the three diversity metrics: when R is

large, the estimate for phylotype richness increases proportional to R, the estimate for Shannon diversity increases proportional to $\ln R$, and the estimate for Simpson diversity tends to a fixed value. This establishes another argument of why Simpson diversity can be estimated more accurately than Shannon diversity and phylotype richness.

Acknowledgements This work was supported by the SYSCOMM project DISCO (ANR-09-SYSC-003) and by the TULIP Laboratory of Excellence (ANR-10-LABX-41).

References

1. Lalande J, Villemur R, Deschênes L (2013) A new framework to accurately quantify soil bacterial community diversity from DGGE. Microb Ecol 66(3):647–658
2. Loisel P, Harmand J, Zemb O, Latrille E, Lobry C, Delgenès JP, Godon JJ (2006) Denaturing gradient electrophoresis (DGE) and single-strand conformation polymorphism (SSCP) molecular fingerprintings revisited by simulation and used as a tool to measure microbial diversity. Environ Microbiol 8(4):720–731
3. Blackwood CB, Hudleston D, Zak DR, Buyer JS (2007) Interpreting ecological diversity indices applied to terminal restriction fragment length polymorphism data: insights from simulated microbial communities. Appl Environ Microbiol 73(16):5276–5283
4. Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS (2013) Robust estimation of microbial diversity in theory and in practice. ISME J 7(6):1092–1101