environmental microbiology reports

Environmental Microbiology Reports (2012) **4**(3), 367–372 doi:10.1111/j.1758-2229.2012.00345.x

Which sequencing depth is sufficient to describe patterns in bacterial α- and β-diversity?

Daniel Lundin,[1,2] Ina Severin,[3] Jürg Brendan Logue,[3†] Örjan Östman,[4] Anders F. Andersson[1] and Eva S. Lindström[3*]**

[1]*KTH Royal Institute of Technology, Science for Life Laboratory, School of Biotechnology, Division of Gene Technology, 17165 Solna, Sweden.*

[2]*BILS, Bioinformatics Infrastructure for Life Sciences, Vetenskapsrådet, Sweden.*

Departments of [3]*Ecology and Genetics/Limnology and* [4]*Ecology and Genetics/Population Biology, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden.*

Summary

The vastness of microbial diversity implies that an almost infinite number of individuals needs to be identified to accurately describe such communities. Practical and economical constraints may therefore prevent appropriate study designs. However, for many questions in ecology it is not essential to know the actual diversity but rather the trends among samples thereof. It is, hence, important to know to what depth microbial communities need to be sampled to accurately measure trends in diversity. We used three data sets of freshwater and sediment bacteria, where diversity was explored using 454 pyrosequencing. Each data set contained 6–15 communities from which 15 000–20 000 16S rRNA gene sequences each were obtained. These data sets were subsampled repeatedly to 10 different depths down to 200 sequences per community. Diversity estimates varied with sequencing depth, yet, trends in diversity among samples were less sensitive. We found that 1000 denoised sequences per sample explained to 90% the trends in β-diversity (Bray-Curtis index) among samples observed for 15 000–20 000 sequences. Similarly, 5000 denoised sequences were sufficient to describe trends in α-diversity (Shannon index) with the same accuracy. Further, 5000 denoised sequences captured to more than 80% the trends in Chao1 richness and Pielou's evenness.

Received 19 January, 2012; accepted 25 March, 2012. For correspondence. *E-mail Eva.Lindstrom@ebc.uu.se; Tel. (+46) 18 4716497; **E-mail Anders.Andersson@scilifelab.se; Tel. (+46) 8 52481414. †Present address: Dept. of Biology/Aquatic Ecology, Lund University, Sölvegatan 37, 22362 Lund, Sweden.

Introduction

The spatial and temporal variability in diversity and composition of microbial communities has, in recent years, received considerable attention. As such, it is debated whether microbial communities follow similar ecological rules as observed for larger organisms (Martiny *et al*., 2006; Lindström and Langenheder, 2012). Bacterial β-diversity, i.e. the variation in bacterial community composition (BCC) among communities in different sites, has to date been predominantly investigated using fingerprinting techniques. Yet, these methods are less suitable for estimating α-diversity, i.e. richness and evenness within a community (Blackwood *et al*., 2007). Recent advances in sequencing technology, e.g. 454 pyrosequencing, have enabled scientists to obtain sequences with less effort and in greater numbers (Shendure and Ji, 2008) and may, hence, replace fingerprinting methods. Moreover, with more sequences being assessed and, thus, more individuals within communities being identified, improved estimates of α-diversity are possible.

It is well known that the number of individuals analysed in a sample affects β- as well as α-diversity estimates (Wolda, 1981; Shaw *et al*., 2008; Lozupone *et al*., 2011; Gihring *et al*., 2012). Yet, for most ecological questions the absolute values are not of interest but rather the trends among samples, i.e. how diversity differs among samples, and co-varies with, for instance, environmental, temporal or geographic gradients (Shaw *et al*., 2008). In this respect, it may be sufficient to adequately describe the most common taxa (Heino and Soininen, 2010; Pommier *et al*., 2010), which is fortunate, since microbial communities are often extremely diverse and greatly skewed (Quince *et al*., 2008). Keeping the number of sequences per sample low is desirable because it allows including a greater number of samples and, as such, a robust study design (Prosser, 2010; Kuczynski *et al*., 2010a). However, the question of 'How many sequences are sufficient to accurately describe patterns in α- and β-diversity?' yet remains.

To address this question, we analysed how sequencing depth affected trends in α- and β-diversity within three

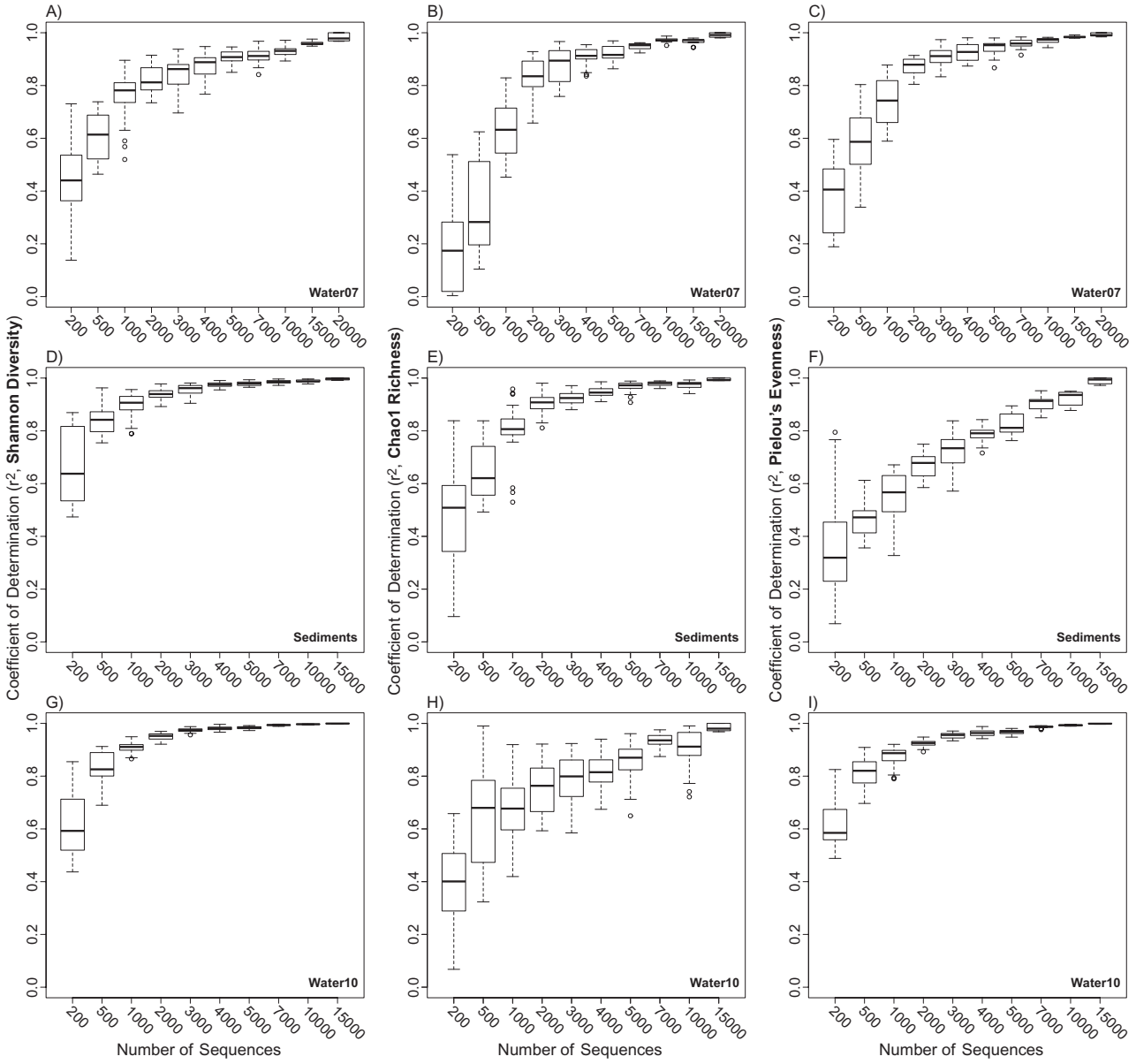

Fig. 1. Analysis of trends in α-diversity in relation to sequencing depth. Results from Pearson's correlations between data sets of different sequencing depth and the full data set of 20 000 or 15 000 sequences per sample. (A, D, G) Shannon index; (B, E, H) Chao1 richness; (C, F, I) Pielou's evenness; (A, B, C) 15 lakes ('Water07'); (D, E, F) 10 sediment samples ('Sediments'); and (G, H, I) 5 lakes and 1 stream ('Water10').

different bacterial data sets, consisting of freshwater communities from sediment ('Sediments') and water ('Water07', 'Water10').

Results

α-Diversity

In all three data sets, there was a consistent change in diversity estimates with changing sequencing depth (i.e. the number of denoised sequences per sample) independent of the α-diversity index used (Fig. S1). Pearson's correlation analyses between Shannon diversity indices obtained at the lower sequencing depths and those at the greatest sequencing depth showed that coefficients of determination (r^2) increased with increasing sequencing depth and the variation among replicates decreased for all data sets (Fig. 1A, D and G). At a sequencing depth of 3000, both the Sediments (Fig. 1D) and Water10 (Fig. 1G) data sets showed median $r^2 > 0.95$, while for the

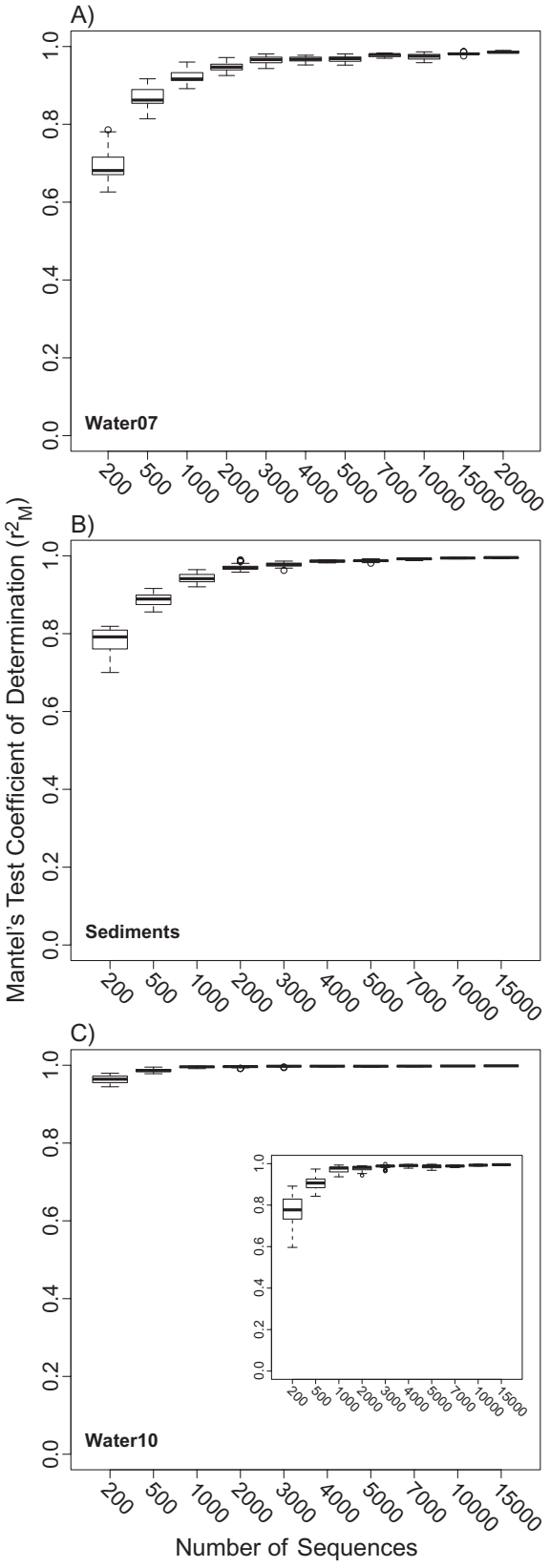

Fig. 2. Analysis of trends in β-diversity (Bray-Curtis index) in relation to sequencing depth. Results from Mantel tests between data sets of different sequencing depth and the full data set of 20 000 or 15 000 sequences per sample. (A) 15 lakes ('Water07'); (B) 10 sediment samples ('Sediments'); (C) 5 lakes and 1 stream ('Water10'); insert to (C) 5 lakes (as in C but excluding the stream).

Water07 data set this was reached at a greater sequencing depth (Fig. 1A). However, at the sequencing depth of 5000 sequences all three data sets showed $r^2 > 0.90$. Thus, applying a sequencing depth of 5000, patterns in Shannon diversity among samples were to 90%, or more, identical to the pattern in Shannon diversity using the complete depth of 15 000–20 000 sequences.

Chao1 estimated richness (Fig. 1B, E and H) required more sequences than Shannon to recapture with the same accuracy the trends obtained with the greatest sequencing depth. At a sequencing depth of 7000 sequences per sample, both the Sediments (Fig. 1E) and Water07 (Fig. 1B) data sets showed median $r^2 > 0.95$, while for Water10 this was not reached at any sequencing depth (Fig. 1H). For all three data sets, median r^2-values were > 0.90 at a sequencing depth of 7000 and > 0.80 at 5000.

Analyses of trends in Pielou's evenness showed the same tendencies as for the other two α-diversity indices (Fig. 1C, F and I). The Sediments data set required the greatest sequencing depth to obtain $r^2 > 0.95$, where the r^2-values never reached a plateau (Fig. 1F). For the two water data sets $r^2 > 0.95$ were obtained at a sequencing depth of 4000–7000. Seven thousand sequences were required to reach median r^2-values > 0.90 in all three data sets, while, 5000 for median $r^2 > 0.80$.

β-Diversity

β-Diversity decreased with increasing sequencing depth in all three data sets (Fig. S2). Similar to the trends in α-diversity indices, Mantel r^2 (r_M^2) of the Bray-Curtis distance matrices showed a positive relationship to sequencing depth (Fig. 2). However, values were generally higher and showed smaller variability. For all three data sets $r_M^2 > 0.90$ was already obtained at a sequencing depth of 1000. For the Water07 and Sediments data sets r_M^2 were > 0.95 at about 2000–3000 sequences (Fig. 2A and B), while for the Water10 data set $r_M^2 > 0.95$ were already reached at a sequencing depth of 500 sequences (Fig. 2C). Excluding the stream sample, which deviated strongly from the other samples (Fig. S2), from the Water10 data set changed r_M^2-values considerably, showing a pattern comparable to that of the other two data sets, i.e. with $r_M^2 > 0.95$ at about 2000–3000 sequences (inset in Fig. 2C).

Discussion

The number of sequences required to reach a specific level of accuracy in estimates of trends among samples of the subsampled compared with the full sequencing depth used (15 000 or 20 000 sequences) differed between diversity estimates and data sets. Least sensitive to sequencing depth were patterns in β-diversity, estimated using the Bray-Curtis index, where a sequencing depth of 3000 always provided $r_M^2 > 0.95$ and 1000 sequences was always sufficient for $r_M^2 > 0.90$. We therefore argue that 1000 denoised sequences per sample give nearly equally good results as results from 15 000 or 20 000 sequences for estimates of trends in β-diversity. Such strong correlations obtained at shallow sequencing depths indicates that the trends will be correlated to true trends in diversity, i.e. at even greater sequencing depths than 15 000–20 000 sequences per sample. This conclusion is furthermore supported by findings by Caporaso and colleagues (2011).

More sequences were generally required for patterns in α-diversity. The Shannon index was least sensitive to sequencing depth, with 5000 sequences per sample being sufficient for $r^2 > 0.90$. Hence, obtaining 5000 denoised sequences per sample may be a good compromise between precision in estimates of trends in α-diversity and economy. If the interest, however, is in Chao1 richness and Pielous evenness more sequences may be needed to obtain $r^2 > 0.90$. However, it should be noted that also for these indices major trends appear to have been captured with 5000 sequences since median r^2-values then were always > 0.80. Investigators using these indices may, thus, choose to use this number of sequences but should be aware that they thereby introduce a certain amount of noise in the data set.

Trends in diversity among samples are, thus, sensitive to the number of sequences, yet, the actual diversity values are even more so. Both α- and β-diversity values consistently changed with increasing sequencing depth, while the r^2-values, representing the accuracy in the estimate of trends among samples, generally reached a plateau more often and at lower sequencing depths. We can, therefore, only stress all the more that comparisons of α- and β-diversity estimates should not be performed between samples of differing sequencing depths but rather between samples of a similar depth or that have been trimmed to the same number of sequences (Gihring *et al*., 2012).

A common finding was that the data set that showed the greatest range in diversity estimates among samples for a certain diversity measure required the lowest number of sequences to detect robust trends among samples, and vice versa. This type of result has also been detected earlier (Kuczynski *et al*., 2010a,b), yet the range of diversity estimates within our data sets seemed to be more important than the actual richness or evenness of the samples, since otherwise the Sediments data set would have always required the greatest number of sequences, which was not the case. Therefore, to obtain a better understanding of how deeply microbial communities in general need to be sequenced, more systematic surveys in different kinds of environments need to be conducted. Further, we would like to stress that caution should be exercised if the aim of a study is to follow the distribution and abundance of rare taxa. In this case, it may be necessary to obtain a greater number of sequences.

Conclusion

We conclude from these data sets that around 1000 denoised sequences per sample are needed for an accurate and precise (> 90%) estimation of trends in β-diversity (Bray-Curtis distance). For α-diversity 5000 denoised sequences may be sufficient, although it differs between indices, where Shannon seems less sensitive to sequencing depth than Chao1 richness and Pielous' evenness. Since, on average, 17% of the sequences disappeared in the first cleaning step, approximately 6000 sequences per sample should therefore be obtained from the sequencing facility.

Experimental procedures

Study sites and sampling

Water samples were collected from 15 oligotrophic lakes in the area of Jämtland, Sweden, in August 2007 as described by Logue and colleagues (2011) (this data set is henceforth called 'Water07'). In addition, water samples were collected from five mesotrophic lakes and one mesotrophic stream in the region of Uppland, Sweden ('Water10') in June and September 2010 respectively. Sediment samples were collected from sediments of the five Uppland lakes and from another five lakes in Jämtland in June 2010 ('Sediments'). See Table S1 for an in-detail description of the sampling sites. DNA was extracted and the bacterial hypervariable regions V3 and V4 of the 16S rRNA gene were PCR-amplified and analysed by 454 pyrosequencing as described in *Supporting information*. The 454 pyrosequencing reads have been deposited in the National Center for Biotechnology Information Sequence Read Archive under Accession No. SRP005457 (Water07) and SRP011432 (Water10 and Sediments).

Sequence analyses

Sequences were, prior to analyses, quality-checked and truncated to 400 bases, which reduced the number of reads per

sample by, on average, 17%. Each data set (Water07, Water10, Sediments) was thereupon subsampled to 200, 500, 1000, 2000, 3000, 4000, 5000, 7000, 10 000, 15 000 (Water07, Water10, Sediments) and 20 000 (only Water07) sequences per sample (drawn from the 'dat' files generated by AmpliconNoise). Each subsample was drawn from the complete set of sequences without replacement, i.e. each sequence could only be included in the subsample once. For subsampling up to 5000 sequences, 10 replicates were generated from each data set, while for 7000 and 10000 sequences and for 15 000 and 20 000 sequences five and three replicate subsamplings were carried out, respectively. The resulting tables are defined as 'constructed data sets', which represent 3–10 replicates of each sequencing depth and each of the three original data sets. Each constructed data set was individually processed with AmpliconNoise to reduce the number of 454 sequencing and PCR artefacts, and PCR chimeras (Quince *et al*., 2011). Denoising was done on the subsampled data rather than on the full data set since sample size may have an effect on noise reduction, and hence diversity estimates. The removal of chimeras resulted in a decrease in the total number of sequences within a sample (Table S2). To compare results obtained within a sequencing depth we therefore resampled each sample to the minimum number of chimera-checked sequences within each sequencing depth and data set. Finally, operational taxonomic units (OTUs) were defined using complete linkage clustering at a level of 97% sequence identity. For a graphical illustration of the procedure see Fig. S3.

Analysis of trends in α- and β-diversity

Shannon diversity indices, Chao1 species richness estimates and Pielou's evenness indices were computed for all samples in all constructed data sets. The values obtained were analysed by Pearson's correlation analyses, correlating the values within each constructed data set to the values obtained for the three replicates at the greatest sequencing depth (Water07: 20 000, Water10 and Sediments: 15 000).

Bray-Curtis distances were calculated between all sample pairs within a constructed data set. Mantel tests were run between the similarity matrices of each constructed data set and the constructed data set obtained from the three replicates of the greatest sequencing depth. Mantel tests were conducted based on Pearson's correlation coefficients with 999 permutations.

All statistical data analyses were conducted in R, employing the Vegan package (Oksanen *et al*., 2011) for the computation of α- and β-diversity indices and Mantel tests.

Acknowledgements

We thank Ines Kohler, Joel Segersten and Jan Johansson for help with sampling. Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX (projects b2010008 and b2010073). Financial support was given from the Helge Ax:son Johnsons foundation to J.B. Logue, from the Olsson-Borghs foundation granted individually to E.S. Lindström, J.B. Logue and Ö. Östman, from Formas to A.F. Andersson, from the Carl Tryggers foundation to E.S. Lindström and Ö. Östman, and from the Swedish Research Council granted individually to E.S. Lindström, A.F. Andersson and Ö. Östman.

References

Blackwood, C.B., Hudleston, D., Zak, D.R., and Buyer, J.S. (2007) Interpreting ecological diversity indices applied to terminal restriction fragment length polymorphism data: insights from simulated microbial communities. *Appl Environ Microbiol* **73:** 5276–5283.

Caporaso, J.G., Lauber, C., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., *et al*. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA* **108:** 4516–4522.

Gihring, T.M., Green, S.J., and Schadt, C.W. (2012) Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environ Microbiol* **14:** 285–290.

Heino, J., and Soininen, J. (2010) Are common species sufficient in describing the turnover in aquatic metacommunities along environmental and spatial gradients? *Limnol Oceanogr* **55:** 2397–2402.

Kuczynski, J., Costello, E.K., Nemergut, D.R., Zaneveld, J., Lauber, C.L., Knights, D., *et al*. (2010a) Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol* **11:** 210.

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., and Knight, R. (2010b) Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods* **10:** 813–819.

Lindström, E.S., and Langenheder, S. (2012) Local and regional factors influencing bacterial community assembly. *Environ Microbiol Rep* **4:** 1–9.

Logue, J.B., Langenheder, S., Andersson, A.F., Bertilsson, S., Drakare, S., Lanzén, A., and Lindström, E.S. (2011) Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species-area relationships. *ISME J* doi:10.1038/ismej.2011.184.

Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J., and Knight, R. (2011) Unifrac: an effective distance metric for microbial community comparison. *ISME J* **5:** 169–172.

Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R.K., Fuhrman, J.A., Green, J.L., *et al*. (2006) Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4:** 102–112.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L., *et al*. (2011) *Vegan: community ecology package* [WWW document]. URL http://CRAN.R-project.org/package=vegan.

Pommier, T., Neal, P.R., Gasol, J.M., Coll, M., Acinas, S.G., and Pedrós-Alió, C. (2010) Spatial patterns of bacterial richness and evenness in the NW Mediterranean Sea explored by pyrosequencing of the 16S rRNA. *Aquat Microb Ecol* **61:** 221–233.

Prosser, J.I. (2010) Replicate or lie. *Environ Microbiol* **12:** 1806–1810.

Quince, C., Curtis, T.P., and Sloan, W.T. (2008) The rational exploration of microbial diversity. *ISME J* **2:** 997–1006.

Quince, C., Lanzén, A., Davenport, R.J., and Turnbaugh, P.J. (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12:** 38.

Shaw, A.K., Halpern, A.L., Beeson, K., Tran, B., Venter, J.C., and Martiny, J.B.H. (2008) It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10:** 2200–2210.

Shendure, J., and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol* **26:** 1135–1145.

Wolda, H. (1981) Similarity indices, sample size and diversity. *Oecologia* **50:** 296–302.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Supplementary experimental procedures.
Supplementary results.
Fig. S1. Analysis of α-diversity in relation to sequencing depth. (A, D, G) Shannon index; (B, E, H) Chao1 richness; (C, F, I) Pielou's evenness; (A, B, C) 15 lakes ('Water07'); (D, E, F) 10 sediment samples ('Sediments') and (G, H, I) 5 lakes and 1 stream ('Water10').
Fig. S2. Analysis of β-diversity. (A) Bray-Curtis values of 15 lakes in relation to sequencing depth ('Water07'); (B) DCA plot of the Water07 samples using 20 000 sequences per sample; (C) Bray-Curtis values of 10 sediment samples in relation to sequencing depth ('Sediments'); (D) DCA plot of the sediment samples using 15 000 sequences per sample; (E) Bray-Curtis values of five lakes and 1 stream in relation to sequencing depth ('Water10'); (F) DCA plot of the Water10 samples using 15 000 sequences per sample. Abbreviations in DCA plots are the same as in Table S1.
Fig. S3. Design of the study.
Table S1. Physicochemical and biological characteristics of the study systems.
Table S2. The loss of sequences due to the perseus Chimera check step. # sequences denotes the number of sequences in each sample before chimera check. Within each data set and sequencing depth each sample was resampled to the minimum number of sequences according to the table.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.