environmental microbiology

Environmental Microbiology (2010) **12**(7), 1806–1810 doi:10.1111/j.1462-2920.2010.02201.x

Opinion

Replicate or lie

James I. Prosser*
Institute of Biological and Environmental Sciences, University of Aberdeen, Cruickshank Building, St. Machar Drive, Aberdeen, AB24 3UU, UK.

Introduction

Andrén and colleagues (2008) recently published a paper reminding us of good scientific practice, taught and learnt during our early years of research, but frequently forgotten. This article has a similar objective. It is highly likely that the majority of microbial ecologists have taken a basic course in statistics. Unfortunately, the basic principles of statistical analysis and its significance and necessity are frequently ignored. My own exposure to this problem, and my suffering of the consequences, arise through reading, reviewing and editing articles on microbial diversity. The suffering is greatest with articles utilizing 'expensive' or cutting edge techniques, such as analysis of clone library sequences, microarray analysis and (increasingly) high-throughput sequencing. The problems, however, are more widespread and exist beyond these techniques and beyond studies of microbial diversity.

Why replicate?

This issue of the journal could be filled with articles describing and discussing applications of statistical analysis in microbial ecology, but I wish to address a simple, basic and fundamental aspect – the need for replication. To exemplify this need, imagine that an undergraduate student, wishing to compare bacterial abundance in two lakes, determines cell concentration in a single 10 ml water sample from each lake, giving values of 1.4×10^6 and 3.2×10^6 cells ml^{-1}. On the basis of these two measurements, he concludes that bacterial abundance is greater in one lake than the other.

Most supervisors, I hope, would remind the student of their basic statistics lectures and of variability arising from environmental heterogeneity and variability introduced during sampling and analysis. With a single sample from each lake, it is impossible to estimate variability. Consequently, it is impossible to assess, with any confidence, whether the difference between the two counts is due to a real difference in abundance or merely a result of variability. The supervisor would likely recommend repeating the counts on replicate samples from each lake, calculation of mean values and estimation of variability associated with mean values, for example, as the standard error. Knowledge of this 'within-lake' variability allows assessment of whether the observed difference in counts is due to chance, or is real. This would typically involve testing the null hypothesis that abundances are identical and calculating the probability that the observed difference arose through variability. If the probability is low, the null hypothesis is rejected. The Student's *t*-test might be used, but I am not concerned here with the particular statistical test to be employed, numbers of replicates, or whether both technical and biological replicates are required. I am concerned only with highlighting the need for replication and I expect that most readers will agree that replication is necessary to compare cell abundance in these two lakes.

Now, imagine that the student also wants to see if bacterial communities in the two lakes differ. He constructs a 16S rRNA gene clone library from each 10 ml sample, one library from each lake, and sequences 200 clones from each library. He then performs BLAST searches, determines relative abundances of different phylogenetic groups, constructs phylogenetic trees and estimates richness and diversity indices. Unfortunately, I suspect that many supervisors would spend valuable time analysing these results with the student, comparing data for the two lakes and discussing the many potential reasons why the communities differ.

This time would (or at least should) be wasted. The need for estimation of 'within-lake' variation in bacterial community composition is just as great as when counting bacteria. Community composition will be heterogeneous and variability will be introduced during the many steps involved in preparing and analysing clone libraries. Comparison of community composition on the basis of single, unreplicated samples, with no estimate of variability, is just as crazy and invalid as comparing my height with that

Received 3 December, 2009; accepted 22 January, 2010. *For correspondence. E-mail j.i.prosser@abdn.ac.uk; Tel. (+44) 1224 273254; Fax (+44) 1224 272703.

of a randomly chosen animal ecologist and concluding, on the basis of these two values, that microbial ecologists are smaller or taller than animal ecologists.

Do we replicate?

Unfortunately, the student and supervisor may not believe that their time is wasted. Table 1 provides information on papers published in 2009 in the five major microbial ecology journals that included molecular analysis of microbial communities. Studies are grouped according to journal and to the major molecular techniques employed and, for each, the percentage of studies in which replicate samples were analysed is presented. In some cases it was not possible to determine whether replicates were analysed, or to distinguish between technical, biological, true and pseudoreplicates, and these studies were excluded. Many studies employed fingerprinting techniques [denaturing gradient gel electrophoresis (DGGE), terminal restriction fragment length polymorphism (T-RFLP)] but a surprisingly small proportion (38%) involved replication, despite the relative cheapness of these methods. Even fewer studies comparing communities using sequence data (18%) and microarrays (22%) analysed replicates. The majority compared data from individual samples with no replication and no assessment of potential variation arising from environmental heterogeneity or other sources of variability.

These results are distressing for at least two reasons,

(i) They highlight major endemic problems in experimental design by authors and researchers and also in reviewing and editing. These may arise through ignorance or, more likely and more worryingly, through a belief that these techniques are, in some way, exempt from normal standards and that bad experimental design and bad science are acceptable.
(ii) Significant resources (time and money) are expended on these studies. It is therefore painful to have to reject such articles, knowing the hours and dollars expended obtaining data that, in all other respects, are often reliable and of high quality. This is particularly pertinent as the problem is greatest in studies using the most 'cutting-edge' and expensive techniques.

Excuses for not replicating

Authors frequently complain when articles are rejected because of lack of replication. A selection of explanations, excuses, justifications and rationalisations is given below.

(i) 'It is too expensive to do replicates.' This implies knowledge and acceptance of the need for replicates and therefore implicitly admits bad experimental design and bad science. If sufficient funds are not available to do good science, then it is better to do no science than bad science. Use of expensive techniques to do flawed science is arguably more culpable than doing poor science with inexpensive techniques. The outcome is the same – meaningless data – but more money has been expended and wasted. If sufficient funds are not available, the study or experiment should not be attempted.
(ii) 'We only analysed single samples but sequenced a very large number of clones in each sample – more than anyone else previously.' This is irrelevant. It does not matter how many clones are sequenced to characterise a community if they are taken from a single sample. The result will be a single value for, e.g. the relative abundance of proteobacteria in a lake. It is not possible to compare communities in two environments without an estimate of variability, no matter how many sequences are obtained for single samples taken from each lake.
(iii) 'Other papers have been published in which communities were compared without analysis of replicates.' This is particularly worrying when such

Table 1. Analysis of studies reported in *Applied and Environmental Microbiology, Environmental Microbiology, FEMS Microbiology Ecology, ISME Journal* and *Microbial Ecology* during 2009 in which molecular techniques were used to study microbial diversity.

	Clone library analysis and pyrosequencing		Fingerprinting methods		Microarrays		Total	
	Number of articles	% with replicates	Number of articles	% with replicates	Number of articles	% with replicates	Number of articles	% with replicates
Appl Environ Microbiol	60	23	50	38	7	29	117	30
Environ Microbiol	47	15	46	37	6	17	99	25
FEMS Microbiol Ecol	29	24	72	38	2	0	103	33
ISME J	23	13	29	41	7	29	59	29
Microbial Ecol	22	9	40	38	1	0	63	27
Total	181	18	237	38	23	22	441	29

Data are presented as the total number of articles in which diversity was characterized using the major methodological approaches and the percentage of studies in which true replicates were analysed. Articles in which it was not possible to determine whether true replicates had been analysed have been omitted.

papers are published in high profile journals, such as *Nature*, *Science* and *PNAS*, and by internationally renowned research groups that lead the way with new, cutting edge, expensive techniques. These major publications often report the first use of these techniques, where publication might be justified as 'proof of concept', for example, demonstrating that pyrosequencing of environmental samples is possible. However, use of such data for scientific interpretation and inference in the absence of good experimental design is inexcusable.

(iv) 'We agree that replication is necessary but don't know the scale at which to take replicates.' This excuse is again worrying because the authors appreciate and admit the need for replication. Temporal and spatial heterogeneity in natural environments can be considerable and must be considered within the context of the specific aims of a study. The issues surrounding the ways in which this is achieved, including determination of the scale at which sampling and replication must take place, are important. Discussion of such issues, in depth, requires a separate article but it is not acceptable to ignore them completely, by performing no replication, particularly when it is acknowledged that replication is required.

(v) 'But it's clone library data.' This implies that lack of replication is common practice and acceptable or necessary for some measurements but not others. It suggests that some (expensive) measurements are immune from experimental variability or good experimental design.

(vi) 'Measurements of other (cheaper) characteristics of the system showed low variability so we don't need to do replication.' However, the variability of different characteristics of a system may vary independently. The fact that temperature in 10 ml sea water samples measured within an area of 100 m^2 shows little variability does not mean that variability in bacterial abundance within the same area will be low.

(vii) 'But Chao values come with errors.' Richness estimates from clone libraries are quoted with variability measures because they are only estimates, based on assumptions and extrapolation. This variability is determined for single samples and is different to 'biological' variability, due to environmental heterogeneity. The latter is likely to be greater (depending on the number of clones sequenced) and 'total' variability will therefore be amplified when experimental, between-sample variability is taken into account.

(viii) 'But we took three replicate samples and pooled them.' Pooling destroys spatial variability and makes it impossible to calculate true experimental variability, which is required for comparisons.

(ix) 'But we analysed the sample on replicate microarrays.' This constitutes technical replication, i.e. replication designed to determine variability associated only with the particular technique being used. It contributes to, but gives no information on variability between measurements taken on replicate samples taken from an environment.

Will we ever replicate?

We are now entering an era in which community analysis using high-throughput sequencing and PhyloChips, which are currently expensive, will become cheaper, affordable and commonplace. There are already several high-profile papers using these techniques to 'demonstrate' differences in communities in different environments, without analysis of replicate samples. This is particularly ironic as the amount of information generated by these techniques is enormous, making replication more, rather than less feasible and affordable. Comparison of single samples from two lakes on the basis of 90 000 sequences from each lake has approximately the same cost as analysing 30 000 sequences in triplicate samples from each lake. The latter strategy will reduce (but not significantly reduce) information on community composition but will allow comparisons to be made between communities in the two lakes. The former strategy represents flawed experimental design and comparisons will be meaningless.

The pyrosequencing era also mirrors the introduction of traditional clone library sequence analysis in the early 1990s. The technique is often employed to generate large amounts of data, with no clear scientific question in mind. These studies are often justified as 'discovery-driven' science. The debate regarding the relative benefits of empirical, 'look-see' studies and experimental, hypothesis-driven science is beyond the scope of this article. However, good experimental design, appropriate sample selection and replication are essential for both approaches. Pyrosequencing is a powerful technique but becomes powerless if used in a poorly designed experiment. Our understanding of microbial diversity is not advanced by the discovery that, for example, 10% of sequences in a particular environment belong to a particular phylogenetic group, if this information arises from analysis of a single sample, because this gives no indication of the confidence we can place in this result or the extent to which it is representative of the environment.

Solutions

Table 1 illustrates the scale and endemic nature of the problem. Some of the studies employed replication for

some measurements, but not for diversity analyses, and may therefore have been published even if properly reviewed and edited. However, the majority should not have been published. Remember, also, that these are papers that were accepted for publication in the major journals and 'slipped through the net'. Many others will have been rejected through poor experimental design, but may have been published elsewhere. This is not the only problem facing scientific publishing, but it is one that we can all help solve and some suggestions are given below.

Researchers must think carefully about the question they are addressing when planning experiments. They should critically assess the techniques and experimental design to be used and determine whether they are capable of achieving the aims of the experiment. They should think, in advance, how they will present and analyse their data. They should remember their basic statistics courses, which they attended for good reason and not just to tick a box. Researchers should consult with statisticians if necessary and should do this before starting an experiment, rather than after receiving damning reviewers' comments. Much research is now collaborative and researchers should decline co-authorship of papers in which they find that co-workers have not followed best practice. Importantly, if research is found to be flawed, e.g. during the review process, researchers should be honest and abandon the work, rather than send it to other journals in the hope that the flaws will not be noticed. No journal should be considered suitable for flawed science and flawed science should never knowingly be submitted for publication.

Supervisors should ensure that researchers in their groups are adequately trained in statistics and experimental design and that they seek expert advice when necessary. They should also ensure that aims of experimental work are clearly defined and analysis methods considered before experiments begin. Training in basic statistics should be available in every educational institute but the endemic nature in characterizing microbial diversity through sequence analysis might merit specific sessions (for supervisors and researchers) at relevant international conferences.

Reviewers should ensure that experimental design is adequate to meet the aims of the study and that it is clearly explained. In many of the papers examined in preparing Table 1, it was not possible to distinguish between true replicates and pseudoreplicates. If reviewers are not sure whether statistical analysis is appropriate, they should admit ignorance and highlight the need for more expert review. Reviewers should also avoid seduction by new, high-powered techniques when judging the quality of research and ask whether studies address significant scientific questions that truly advance microbial ecology.

Editors should determine whether appropriate replication has been performed before sending a paper for review. Preparing Table 1 took some time because it involved looking at several hundred articles, but it is relatively easy to do this for a single paper and no more time-consuming than determining, for example, whether the quality of the language is sufficient. Trapping poor experimental design before the review stage saves considerable reviewer time and annoyance. Editors should also resist the temptation to accept, or give the benefit of doubt to 'borderline' papers because they adopt new technologies that might attract more citations and increase journal impact factor. Ultimately, the reputations of the journal and the editors suffer through publication of bad science.

In addition, we can all follow good example and it is possible to analyse replicates even when using the most modern and expensive techniques. By way of illustration, Sundset and colleagues compared rumen microbial communities in two sets of five reindeer using DGGE, DeAngelis and colleagues (2009) used T-RFLP and PhyloChips to characterize bulk soil and root-associated communities in triplicate microcosms and pyrosequencing of triplicate soil and rhizosphere samples by Uroz and colleagues (2010) shows that replication is possible when using high-throughput sequencing and rarefaction curves from their sequence data demonstrate why it is essential.

Should we be embarrassed?

The significance of this problem is not merely the annoyance it gives readers, reviewers and editors. It represents bad science and reflects on microbial ecology as a discipline. When I have mentioned this issue to animal or plant ecologists, their responses have ranged from amazement to disbelieving laughter. The impact of articles such as this is difficult to gauge. Three years ago, Bent and colleagues (2007) highlighted the impossibility of calculating richness from DNA fingerprinting techniques, but papers still appear in which numbers of DGGE bands or T-RFLP peaks are used as measures of richness.

At the moment, if I measured the height of one Englishmen, one American, one Australian and one African and then used the data to postulate ways in which height was determined by continent I would, correctly, be ridiculed. If I were to take a single faecal sample from each of the same individuals and performed massive, parallel, high-throughput, 454-sequencing to generate sequence lists, relative abundances of different phylogenetic groups and pie-charts, and used these data to postulate continent-associated differences in intestinal microbial communities, it is likely that the paper would be published in one of

the highest profile, general science journals. There are already several precedents for this and, as a community, we should be embarrassed.

Acknowledgements

I am very grateful for the invaluable comments of both reviewers of this article, one of whom is solely responsible for the title.

References

Andrén, O., Kirchmann, H., Kätterer, T., Magid, J., Paul, E.A., and Coleman, D.C. (2008) Visions of a more precise soil biology. *Eur J Soil Sci* **59:** 380–390.

Bent, S.J., Pierson, J.D., and Forney, L.J. (2007) Measuring species richness based on microbial community fingerprints: the emperor has no clothes. *Appl Environ Microbiol* **73:** 2399.

DeAngelis, K.M., Brodie, E.L., DeSantis, T.Z., Andersen, G.L., Lindow, S.E., and Firestone, M.K. (2009) Selective progressive response of soil microbial community to wild oat roots. *ISME J* **3:** 168–178.

Sundset, M.A., Edwards, J.E., Cheng, Y.F., Senosiain, R.S., Fraile, M.N., Northwood, K.S., *et al.* (2009) Rumen microbial diversity in Svalbard reindeer, with particular emphasis on methanogenic archaea. *FEMS Microbiol Ecol* **70:** 553–562.

Uroz, S., Buée, M., Murat, C., Frey-Klett, P., and Martin, F. (2010) Pyrosequencing reveals a contrasted bacterial diversity between oak rhizosphere and surrounding soil. *Environ Microbiol Rep* **2:** 281–288.