Comment on "Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil"

John Bunge,[1]* Slava S. Epstein,[2] Daniel G. Peterson[3]

Gans *et al*. (Reports, 26 August 2005, p. 1387) provided an estimate of soil bacterial species richness two orders of magnitude greater than previously reported values. Using a re-derived mathematical model, we reanalyzed the data and found that the statistical error exceeds the estimate by a factor of 26. We also note two potential sources of error in the experimental data collection and measurement procedures.

Using previously published DNA reassociation kinetics (Cot curve) data (*1*), Gans *et al*. (*2*) estimated bacterial species richness (one aspect of diversity) in a soil sample to be 8.3×10^6. However, the authors' calculation of error for this estimate is unrealistically low. We re-derived the mathematical model of reassociation kinetics from first principles [arriving at a model similar to Gans *et al*. (*2*)] and applied standard nonlinear regression analysis to fit the model to the original data. We obtained a similar richness estimate (7.4×10^6), but a formal statistical error 26 times as large as the estimate itself. Furthermore, we note potential sources of error in the original experimental and measurement protocol that may contribute to the unreliability of the richness estimate.

Let us assume a DNA extract containing sequences from $S \geq 1$ bacterial species, with no interspecific sequence similarity and no intraspecific repeat sequences (both false). Applying certain simplifying assumptions (*3*, *4*), we obtain a mathematical model for the Cot (observed reassociation) data points (u^l, y^l), $l = 1 \ldots n$:

$$E\left(y^{(\ell)}\right) \approx E\left[W\left(1 + \frac{k_r}{S} W u^{(\ell)}\right)^{-\gamma}\right] \quad (1)$$

where W is a random variable representing the species' proportions and reassociation rates, and $\gamma = 0.45$ and $k_r = 5.19$ are taken to be constants (*2*). Nonlinear regression then produces parameter estimates, standard errors (SEs), and goodness-of-fit tests (*5*, *6*).

We fitted Eq. 1 to the noncontaminated soil data provided by Sandaa (*2*). We tested 10 distributions for W (*7*) and found that only one yielded a convincing fit to the observed points. This had the form $P(W = \lambda_i) = p_i$, $\lambda_i > 0$, $i = 1,2,3$; $p_1 + p_2 + p_3 = 1$, with $p_1 = 1.30 \times 10^{-4}$, $\lambda_1 = 2.60 \times 10^3$, $p_2 = 2.40 \times 10^{-6}$, $\lambda_2 = 7.20 \times 10^4$, $p_3 = 9.998676 \times 10^{-1}$, and $\lambda_3 = 4.892648 \times 10^{-1}$ (*8*). The fit was excellent (Fig. 1), with a sum of squared errors (SSE) of 1.05×10^{-3}. The estimate of S was 7.4×10^6, but with a SE of 192.1×10^6.

The model used by Gans *et al*. (*2*) can be rewritten as

$$E\left(y^{(\ell)}\right) = E\left(W\left[1 + \frac{k_r}{T/\mu} W u^{(\ell)}\right)^{-\gamma}\right] \quad (2)$$

where $\mu > 0$, and $T := \sum_{i=1}^{S} N_i$ is assumed known. The authors applied a minimum χ^2 procedure that does not yield SEs for the parameter estimates. Their report lacked certain details that prevented us from replicating their results exactly, but our model and fit are comparable. Our richness estimate is close to theirs, but the statistical SE is far higher than their informal calculation of a factor of, at most, 8.2. An SE of this magnitude makes intercommunity comparisons (e.g., richness in pristine versus polluted environments) statistically meaningless, because the range of possible values of the (unknown) richness of this community is virtually unbounded.

These results are sensitive to model assumptions to an unknown degree. For example, if γ is estimated from the data, then a simpler model fits very well with SSE 1.22×10^{-3}, but γ and S are estimated as 0.1095 (SE, 0.003) and 629 (SE, 120), respectively. Until such robustness issues are clarified, any results must be regarded as contingent on numerous questionable assumptions.

We also noted certain debatable aspects of the original experimental protocol and measurement procedure. First, Gans *et al*. (*2*) assumed that the DNA analyzed in the Cot analysis of Sandaa *et al*. (*1*) was bacterial in nature. We tested the bacterial extraction technique described (*1*) and observed considerable contamination of the bacterial pellet with eukaryotic cells/tissues. The presence of eukaryotic genomes in the DNA extract would introduce substantial error into estimates of bacterial richness using reassociation kinetics data. Second, DNA reassociation was estimated by measuring changes in hypochromicity (Δh), a practice that can greatly underestimate the reassociation of repetitive sequences in complex DNA mixtures (*9*, *10*) (Fig. 2). A population of soil bacteria may be dominated by a few species (*11*, *12*) whose sequences would effectively reassociate like eukaryotic repetitive elements; in fact, our estimated abundance distribution shows just this structure. In this case, normal variation in homologous DNA sequences would result in formation of duplexes with partial strand mismatch, which is believed to underlie the reduced Δh of renatured eukaryotic repeats (*9*).

[1]Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA. [2]Department of Biology and Marine Science Center, Northeastern University, Boston, MA 02115, USA. [3]Department of Plant and Soil Sciences, Mississippi State University, Mississippi State, MS 39762, USA.

*To whom correspondence should be addressed. E-mail: jab18@cornell.edu

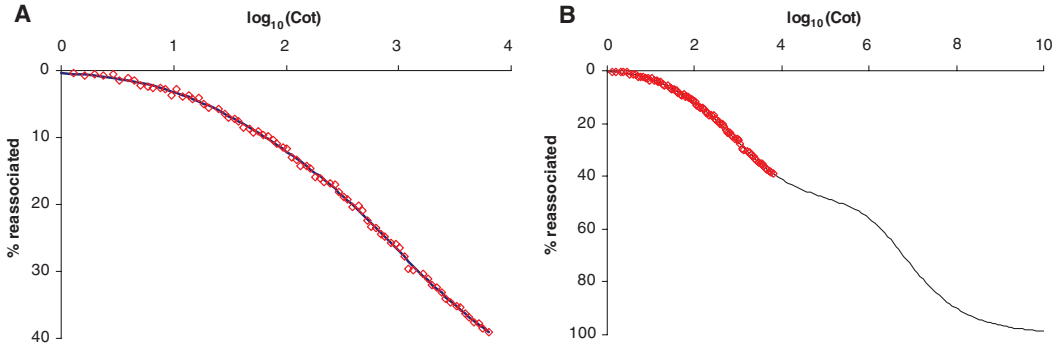

Fig. 1. Cot curves fitted to noncontaminated soil data (*2*) by nonlinear regression. (**A**) Mixture-of-three-point-masses species-abundance model, used in equation 11 in (*15*), with parameters estimated by nonlinear least-squares regression, yields function shown by solid line; data points are overlaid. (**B**) Fitted curve extended to complete (100%) reassociation. Nonconstant curvature is due to the mixture of Cot curves with varying reassociation rates. Extension of estimated curves far beyond available data is statistically inadvisable.

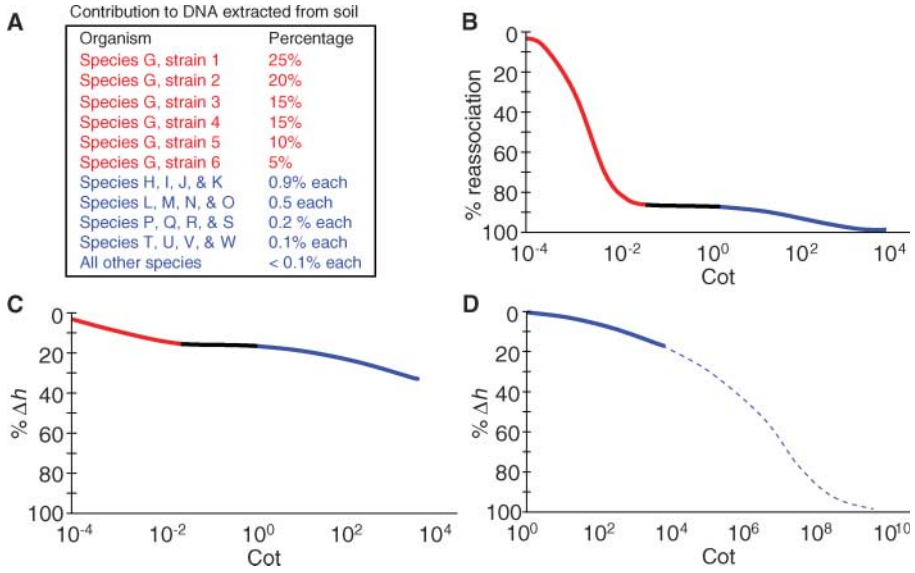

Fig. 2. Equating Δh with DNA reassociation in complex samples can produce misleading results. (**A**) DNA extracted from a soil sample represents numerous bacterial species/strains as shown, with 90% of the DNA contributed by several strains of Species G. For simplicity, assume that different species share no notable sequence homology but that DNA from strains of the same species can form duplexes during reassociation (with occasional base mismatches due to modest sequence divergence). (**B**) A hydroxyapatite chromatography–based Cot curve of the soil DNA extract would show rapid reassociation of Species G DNA (red portion of curve) compared with DNA of other species (blue portion). Although the Species G genome may contain little repetitive sequence, its relative abundance in the DNA extract would cause it to reassociate at least 100 times as fast as DNA of any other species. The gap in relative sequence redundancy between Species G and DNA sequences of other species would result in a flat region of the curve where there would be no notable DNA reassociation (black portion). (**C**) Cot curve prepared from the same soil extract, in which Δh data are used to estimate DNA reassociation. For simplicity, assume that Δh from complete native double-stranded DNA to complete denaturation accounts for a 27% change in absorbance (*9*) and that repetitive DNA (here, Species G DNA duplexes) exhibit half the Δh of native DNA, as is typical of eukaryotic repeats (*9*, *10*). As a result of its relatively low hypochromicity, reassociation of Species G DNA will occupy only 12% of the abscissa ($0.27 \times 0.5 \times 0.9 = 0.12$). At high Cot values (e.g., 10^4 M·s), reassociation of soil extract DNA will appear to be far from completion (i.e., 100% hypochromicity), when in reality it may have finished reassociating. (**D**) Reassociation of Species G DNA at relatively low Cot coupled with its reduced Δh may cause some researchers to discount its renaturation as a "collapse" hypochromicity effect; see (*16*) for definition. Consequently, they may entirely omit it from their Cot curve, as shown. Extrapolation of the curve to 100% hypochromicity (dotted blue line) would amplify the error.

Extrapolation of partial Δh Cot curves to "completion," as was done by Gans *et al.* (*2*), amplifies these errors.

Current soil bacterial species richness estimates range from < 100 (*13*) to almost 10^7 (*2*). Many of these estimates may be correct, although imprecise: When $SE \approx 2 \times 10^8$, an estimate may assume almost any value and remain correct, although uninformative. Informative estimation of species richness by DNA reassociation kinetics will require more precise parameter estimation, a more realistic physical model (*14*), and analysis of sensitivity to assumptions and constants.

References and Notes

1. R. A. Sandaa *et al.*, *FEMS Microbiol. Ecol.* **30**, 237 (1999).
2. J. Gans, M. Wolinsky, J. Dunbar, *Science* **309**, 1387 (2005).
3. P. Erdi, J. Tóth, *Mathematical Models of Chemical Reactions* (Princeton Univ. Press, Princeton, NJ, 1989).
4. M. J. Smith, R. J. Britten, E. H. Davidson, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 4805 (1975).
5. M. Davidian, D. M. Giltinan, *J. Agric. Biol. Environ. Stat.* **8**, 387 (2003).
6. G. A. F. Seber, C. J. Wild, *Nonlinear Regression* (Wiley, New York, 1989).
7. S.-H. Hong, J. Bunge, S.-O. Jeon, S. S. Epstein, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 117 (2006).
8. D. Bohning, D. Schon, *J. R. Stat. Soc. Ser. C* **54**, 721 (2005).
9. D. E. Graham, B. R. Neufeld, E. H. Davidson, R. J. Britten, *Cell* **1**, 127 (1974).
10. R. J. Britten, D. E. Graham, B. R. Neufeld, *Methods Enzymol.* **29**, 363 (1974).
11. M. L. Nagy, A. Perez, F. Garcia-Pichel, *FEMS Microbiol. Ecol.* **54**, 233 (2005).
12. A. Bissett, J. Bowman, C. Burke, *FEMS Microbiol. Ecol.* **55**, 48 (2006).
13. P. F. Kemp, J. Y. Aller, *FEMS Microbiol. Ecol.* **47**, 161 (2004).
14. R. Murugan, *Biochem. Biophys. Res. Commun.* **293**, 870 (2002).
15. Materials and methods are available as supporting material on *Science* Online.
16. A. J. Bendich, R. S. Anderson, *Biochemistry* **16**, 4655 (1977).
17. This work was supported in part by National Science Foundation award DBI-0421717 to D.G.P. and by National Science Foundation grants OCE-0221267, MCB-0348341, and DEB-0103599 to S.S.E.

Supporting Online Material

www.sciencemag.org/cgi/content/full/313/5789/918c/DC1
Materials and Methods
SOM Text
Fig. S1
References

23 February 2006; accepted 5 July 2006
10.1126/science.1126593