



Supporting Online Material for

**Comment on “Computational Improvements  
Reveal Great Bacterial Diversity and High Metal Toxicity in Soil”**

John Bunge,\* Slava S. Epstein, Daniel G. Peterson

\*To whom correspondence should be addressed. E-mail: [jab18@cornell.edu](mailto:jab18@cornell.edu)

Published 18 August 2006, *Science* **313**, 918c (2006)

DOI: 10.1126/science.1126593

**This PDF file includes:**

Materials and Methods

SOM Text

Fig. S1

References

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

Supporting Online Material

John Bunge  
Department of Statistical Science  
Cornell University  
Ithaca, NY 14853  
jab18@cornell.edu

Slava S. Epstein  
Department of Biology and  
Marine Science Center  
Northeastern University  
Boston, MA 02115  
s.epstein@neu.edu

Daniel G. Peterson  
Department of Plant and Soil Sciences  
Mississippi State University  
Mississippi State, MS 39762  
dpeterson@pss.msstate.edu

June 13, 2006

# 1 Introduction

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  $8.3 \times 10^6$ . Our examination shows that cumulative uncertainties in the statistical, microbiological, and biochemical assumptions underlying this estimate render it so imprecise as to be uninformative. Here we re-derive the basic mathematical model of the Cot curve, and we re-analyze the original data, fitting it to the model via standard nonlinear regression. We find an estimate of species richness in close agreement with the authors', but with statistical standard error equal to 26 times the estimate itself. In addition we discuss several potentially crucial sources of experimental and measurement error.

## 2 Mathematical model of DNA reassociation

The following discussion broadly parallels that in Gans et al. (2), while simplifying the central mathematical argument.

### 2.1 Non-repetitive DNA

In a DNA solution in which no one sequence is more prevalent than another, the process of hybridization of complementary single-stranded to double-stranded DNA occurs according to the following process. The basic reaction is



where  $C$  denotes single-stranded, and  $D$  double-stranded, DNA. Let  $C = C(t)$  denote the concentration of single-stranded DNA as a function of time  $t \geq 0$ , and let  $C_0 = C(0)$ . A standard mathematical modeling argument (3, ch. 4.1) yields the (ordinary) differential equation

$$\dot{C} = -kC^2, \quad (2)$$

where  $k$  is a positive constant. This is an example of “second-order kinetics.” The general solution to Eq. (2) is

$$C(t) = \frac{1}{a + kt},$$

where  $a$  is an arbitrary constant. Applying the initial condition, we have  $C_0 = C(0) = 1/a$  so that  $a = 1/C_0$  and finally

$$C(t) = \frac{1}{1/C_0 + kt} = \frac{C_0}{1 + kC_0t}. \quad (3)$$

Now define new variables

$$u := C_0 t \quad \text{and} \quad y(u) := \frac{1}{C_0} C \left( \frac{u}{C_0} \right) = \frac{C(t)}{C_0}.$$

In terms of  $u$  and  $y$ , Eq. (3) becomes

$$y(u) = \frac{1}{1 + ku}, \quad (4)$$

$u \geq 0$ , where  $k > 0$  is the reassociation rate constant. The function Eq. (4) is fitted to observed ‘‘Cot points’’  $(u^{(\ell)}, y^{(\ell)})$ ,  $\ell = 1 \dots n$ .

Equation (4) gives a good fit to empirical Cot data when the reassociation reaction is measured by hydroxyapatite binding, but does not fit measurements obtained via S1 nuclease resistance. In order to fit data of the latter type, one proposal is to retain the basic functional form of Eq. (4) but to apply a ‘‘retardation’’ exponent  $\gamma > 0$  (4):

$$y(u) = \frac{1}{(1 + ku)^\gamma} \quad (5)$$

While Eq. (5) can fit empirical data in some cases, its use rests on several debatable assumptions. First, Eq. (5) is a solution to the differential equation  $\dot{C} = -(\gamma k)C^{1+1/\gamma}$ , but it is not clear what reaction would be described by such an equation when  $\gamma \neq 1$ ; this is the ‘‘inverse problem of reaction kinetics’’ (3, ch. 4.7). Second, a numerical value for  $\gamma$  must be estimated empirically, but no formal statement of the error in this estimate seems to have been reported in the literature. Third, it is not obvious that a given value of  $\gamma$  should be regarded as universal, even when restricted to S1 nuclease-based measurements. Fourth, even if a such a value were known, it is not clear that it should be used for other types of measurements such as changes in hypochromicity, which was the technique utilized in Gans et al. (2). We show below that allowing  $\gamma$  to vary as a free parameter has a dramatic effect on the statistical analysis.

## 2.2 Complex DNA samples

Now suppose that the experiment involves DNA from some number of bacterial species  $S \geq 1$ . For simplicity, we assume that the species share no sequence similarity and that the DNA of each species contains no repeat sequences (both assumptions are known to be false, but are nonetheless utilized in the literature). Let  $N_i > 0$  denote the abundance of the  $i$ th species,  $i = 1, \dots, S$ . Then a simple mixture model for the reassociation kinetics is

$$y(u) = \sum_{i=1}^S \frac{N_i}{\sum_{j=1}^S N_j} \frac{1}{(1 + k_i u)^\gamma}, \quad (6)$$

where  $k_i$  is the reassociation rate for the  $i$ th species,  $i = 1, \dots, S$ . Here we assume that  $\gamma$  is constant, which as noted above is open to criticism. To further simplify Eq. (6) we introduce a deterministic model for the  $k_i$  and a stochastic model for the  $N_i$ . Other sets of assumptions are possible: it is important to note that robustness of the results relative to varying model assumptions has not been examined, except in a limited fashion as partly reported below.

Suppose first that there is a universal constant reassociation rate for a fragment of genomic DNA,  $k_f$ , and suppose further that, in a solution containing only one species, a given fragment must perform a “linear search” for its match. That is, if the original sequence contains (say)  $M$  fragments, which are assumed to be all of the same size, then a particular fragment must search through all  $M$  to find its match, so that the actual reassociation rate will be

$$k = \frac{k_f}{M}.$$

Now in a solution containing  $S \geq 1$  species, assuming that a given fragment must now search among all fragments of all species, and assuming further that each species produces  $M$  fragments of identical size, the reassociation rate for the  $i$ th species is

$$k_i = \frac{k_f}{M} \times \frac{N_i}{\sum_{j=1}^S N_j}.$$

But given the value of a reference rate  $k_r$  (obtained from experiments on *E. coli.*), we have, under the same assumptions,

$$k_r = \frac{k_f}{M}$$

and hence

$$k_i = k_r \frac{N_i}{\sum_{j=1}^S N_j}.$$

We then have

$$y(u) = \sum_{i=1}^S \frac{N_i}{\sum_{j=1}^S N_j} \frac{1}{(1 + k_i u)^\gamma} = \sum_{i=1}^S \frac{N_i}{\sum_{j=1}^S N_j} \left( 1 + k_r \frac{N_i}{\sum_{j=1}^S N_j} u \right)^{-\gamma}, \quad (7)$$

$u \geq 0$ . Note that the value of  $k_r$  must be estimated empirically, and (as for  $\gamma$ ) the literature does not appear to provide a formal standard error for this estimate.

We further simplify Eq. (7) by introducing a stochastic abundance model for the  $N_i$ , that is, we assume that  $N_1, \dots, N_S$  are independent and identically distributed (i.i.d.) random variables from some distribution  $p(\cdot; \vec{\theta})$ , where  $\vec{\theta}$  is a low-dimensional vector of parameters; let  $r = \dim \vec{\theta}$ . Finally we assume that the observed Cot data points  $(u^{(\ell)}, y^{(\ell)})$  are generated by the underlying mixture model

perturbed by i.i.d. Gaussian errors, so that

$$\begin{aligned} y^{(\ell)} &= y(u^{(\ell)}) = \sum_{i=1}^S \frac{N_i}{\sum_{j=1}^S N_j} \frac{1}{(1 + k_i u^{(\ell)})^\gamma} + \epsilon^{(\ell)} \\ &= \sum_{i=1}^S \frac{N_i}{\sum_{j=1}^S N_j} \left( 1 + k_r \frac{N_i}{\sum_{j=1}^S N_j} u^{(\ell)} \right)^{-\gamma} + \epsilon^{(\ell)}, \end{aligned} \quad (8)$$

$\epsilon^{(1)}, \dots, \epsilon^{(n)} \sim$  i.i.d.  $N(0, \sigma_\epsilon^2)$  ( $\sigma_\epsilon^2 > 0$  fixed);  $0 \leq u^{(1)} < u^{(2)} \dots < u^{(n)}$ . Our objective, then, is to estimate the mean function

$$\begin{aligned} E(y^{(\ell)}) &= E \left( \sum_{i=1}^S \frac{N_i}{\sum_{j=1}^S N_j} \left( 1 + k_r \frac{N_i}{\sum_{j=1}^S N_j} u^{(\ell)} \right)^{-\gamma} + \epsilon^{(\ell)} \right) \\ &= E \left( \sum_{i=1}^S \frac{N_i}{\sum_{j=1}^S N_j} \left( 1 + k_r \frac{N_i}{\sum_{j=1}^S N_j} u^{(\ell)} \right)^{-\gamma} \right). \end{aligned} \quad (9)$$

### 2.3 Statistical analysis

The high-dimensional integral required by Eq. (9) presents significant numerical difficulties. We can simplify Eq. (9) for the purpose of statistical analysis as follows. By the (weak or strong) Law of Large Numbers,  $(1/S) \sum_{i=1}^S N_i \rightarrow E(N)$  as  $S \rightarrow \infty$ , where convergence is in the weak or the strong sense and  $E(N)$  denotes the expected value of  $N$ . When  $S$  is large we can therefore replace  $\sum N_j$  in Eq. (9) by  $SE(N)$ , to obtain

$$\begin{aligned} E(y^{(\ell)}) &\approx E \left( \sum_{i=1}^S \frac{N_i}{SE(N)} \left( 1 + k_r \frac{N_i}{SE(N)} u^{(\ell)} \right)^{-\gamma} \right) \\ &= \sum_{i=1}^S E \left( \frac{N_i}{SE(N)} \left( 1 + k_r \frac{N_i}{SE(N)} u^{(\ell)} \right)^{-\gamma} \right) \\ &= SE \left( \frac{N_1}{SE(N)} \left( 1 + k_r \frac{N_1}{SE(N)} u^{(\ell)} \right)^{-\gamma} \right) \\ &= E \left( \frac{N_1}{E(N)} \left( 1 + \frac{k_r}{S} \frac{N_1}{E(N)} u^{(\ell)} \right)^{-\gamma} \right), \end{aligned} \quad (10)$$

$\ell = 1 \dots n$ . Now observe that Eq. (10) depends only on  $N_1/E(N)$ . To further simplify Eq. (10), define a random variable  $W$  such that

$$W =_D \frac{N_1}{E(N)},$$

where  $=_D$  denotes equality in distribution, so that  $E(W) = 1$  and

$$E(y^{(\ell)}) \approx E\left(W\left(1 + \frac{k_r}{S}Wu^{(\ell)}\right)^{-\gamma}\right), \quad (11)$$

$\ell = 1 \dots n$ . Let  $r = \dim \vec{\theta}$ . If there exists an invertible mapping

$$\vec{\theta} \mapsto \vec{\eta} := (\eta_1(\vec{\theta}) = E(N), \eta_2(\vec{\theta}), \dots, \eta_r(\vec{\theta}))',$$

then under the  $\vec{\eta}$ -parameterization our model is restricted to  $\eta_1 = 1$ . This means that the operative dimension of the parameter space is  $r - 1$ , which significantly simplifies the numeric computations.

Our objective is thus to fit model Eq. (11) to the observed Cot data  $(u^{(\ell)}, y^{(\ell)})$ ,  $\ell = 1 \dots n$ . The model has one parameter of principal interest,  $S$ , plus  $r - 1$  parameters for the distribution of  $W$  (since  $k_r$  and  $\gamma$  are taken to be fixed *a priori*). Various statistical approaches to a model of this type are possible (5), but the most straightforward in this case is nonlinear least-squares regression (6), and we use this method. The resulting analysis yields parameter estimates and associated standard errors, goodness-of-fit measures, tests of model appropriateness, and other statistics.

In Gans et al. (2) the authors derived a model similar to Eq. (11). Their equation can be rewritten as

$$E(y^{(\ell)}) = E\left(W\left(1 + \frac{k_r}{T/\mu}Wu^{(\ell)}\right)^{-\gamma}\right), \quad (12)$$

where  $\mu$  is a positive parameter and  $T := \sum_{i=1}^S N_i$ . Here  $T$  is taken to be fixed and known, although the model is not probabilistically conditional on  $T$ , which would be appropriate if  $T$  is given. The authors fit their model by a minimum  $\chi^2$  procedure which does not yield standard errors for the parameter estimates. They obtain an error estimate by an *ex post facto* calculation which (in particular) does not take into account the shape of the objective function near the parameter estimates; we show below that the resulting error figure is unrealistically low.

### 3 Results

We fit our model Eq. (11) to the non-contaminated soil data analyzed in Gans et al. (2); the data was kindly provided to us by the original investigator, Dr. Ruth-Anne Sandaa (7). The following table shows the candidate distributions we tested, along with the corresponding number of free parameters with  $E(N) = 1$ , namely  $r - 1$  as in the previous discussion:

distribution	$r - 1$
single point mass	0
exponential	0
gamma	1
inverse Gaussian	1
lognormal	1
Pareto (power law)	1
mixture of 2 exponentials	2
mixture of 3 exponentials	4
mixture of 2 point masses	2
mixture of 3 point masses	4

(We have previously used the first eight of these to estimate species richness from count data (8).) Holding  $\gamma$  fixed at 0.45 and  $k_r$  at 5.19 (2), we were able to obtain a satisfactory fit only from the mixture of 3 point masses. This has the general form  $P(W = \lambda_i) = p_i, \lambda_i > 0, i = 1, 2, 3; p_1 + p_2 + p_3 = 1$  (such models also appear as outcomes in nonparametric maximum likelihood estimation of species richness from count data (9)). We obtained  $p_1 = 1.30 \times 10^{-4}, \lambda_1 = 2.60 \times 10^3, p_2 = 2.40 \times 10^{-6}, \lambda_2 = 7.20 \times 10^4, p_3 = 9.998676 \times 10^{-1}, \lambda_3 = 4.892648 \times 10^{-1}$ . The sum of squared errors (SSE) was  $1.05 \times 10^{-3}$  and the fit was excellent (Fig. 1; the projected complete (mixed) Cot curve is also shown). The estimate of  $S$  was  $7.4 \times 10^6$ , close to the  $8.3 \times 10^6$  of Gans et al. (2), but the corresponding standard error (SE) was  $192.1 \times 10^6$ , 26 times the size of the estimate. By exploration we found that the objective function (SSE) is very flat in the region of the optimal parameter values, and the standard error formula detects this fact (6, ch. 5).

The other models all had higher SSE (with  $\gamma = 0.45$  and  $k_r = 5.19$ ). However, even when SSE seemed not too large, the models still failed to capture certain evidently non-random local fluctuations in the data; this was confirmed by the Durbin-Watson test (6, ch. 6). For example, the mixture of 2 point masses gave SSE =  $4.98 \times 10^{-3}$  and a superficially close fit to the data, but careful visual inspection and the Durbin-Watson test clearly rejected the model. It is interesting to note that the estimate of  $S$  in this case was  $3.93 \times 10^4$  with SE  $2.66 \times 10^3$ , a plausible result that is invalidated by the lack of fit.

All of these results are sensitive to the model assumptions to a degree that has yet to be investigated. We cannot analyze sensitivity to  $k_r$  because  $k_r$  is confounded with  $S$  in Eq. (11), so error in one is transmitted to the other. Regarding  $\gamma$ , suppose we take it to be a universal but unknown constant, to be estimated from the data. Then, for example, the mixture of 2 point masses fits very well with an SSE of  $1.22 \times 10^{-3}$ , close to our best result when  $\gamma = 0.45$ , but  $\gamma$  is estimated to be 0.1095 with SE of 0.003, hence far from 0.45, and the estimate and SE of  $S$  are 629 and 120

respectively. In fact the single point mass (all abundances equal) with unknown  $\gamma$  fits better than the mixture of 2 point masses with  $\gamma = 0.45$ , but it estimates  $\gamma$  to be 0.0924 (SE 0.001) and  $S$  to be 171 (SE 7). Until these and other robustness issues are clarified, any results must be regarded as at best contingent on a collection of questionable model assumptions.

In Gans et al. (2), the authors fit Eq. (12) by a minimum  $\chi^2$  procedure, obtaining  $8.3 \times 10^6$  as the estimate of  $S$ . Their informal error calculation, which included error from all sources, gave a factor of at most 8.2 relative to the estimate; we interpret this to mean that a lower bound for  $S$  is  $\approx 10^6$ . Their paper lacked certain numerical details which prevented us from replicating their results exactly (using Eq. (11) or Eq. (12) under any abundance distribution), and hence we cannot compare the fit of our selected model to theirs, except visually. Nonetheless our selected abundance distribution (the mixture of 3 point masses) is similar to their “model-free” distribution, and gives comparable fit to the data, and our point estimate ( $7.4 \times 10^6$ ) is close to theirs ( $8.3 \times 10^6$ ). But even if all the (debatable) assumptions underlying these results were correct, the statistical SE for the point estimate is so large as to render it uninformative. Clearly an SE of this magnitude makes it inter-community comparisons (e.g., richness in pristine vs. polluted environments (2)) statistically meaningless.

## 4 Experimental measurement of DNA reassociation

Underlying the claim of Gans et al. (2) is the important assumption 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 (Fig. S1). Based on this observation, it is probable that the DNA used in the Cot analyses was contaminated with eukaryotic DNA. The presence of eukaryotic genomes in the DNA extracts would introduce substantial error into estimates of bacterial richness using reassociation kinetics data.

Another potential source of error in the Cot analysis is that DNA reassociation was estimated by measuring changes in hypochromicity ( $\Delta h$ ), a practice that can greatly underestimate reassociation of repetitive sequences in complex DNA mixtures (10, 11) (Fig. 2). While repetitive DNA may not normally be an issue when studying reassociation of DNA isolated from one bacterial strain, a population of soil bacteria may be dominated by a few species (12, 13) whose disproportionate contributions to the DNA pool would cause their sequences to effectively reassociate like eukaryotic repetitive elements. Our estimated abundance distribution for this data displays just such a structure. In such a case, normal intra- and interspe-

cific variation in homologous DNA sequences would result in formation of many duplexes with partial strand mismatch which is believed to underlie the reduced  $\Delta h$  of renatured eukaryotic repeats (10). As Figure 2 illustrates, using  $\Delta h$  as a measure of reassociation when studying DNA from complex populations could result in highly erroneous conclusions, especially if partial  $\Delta h$  Cot curves were extrapolated to “completion” (100% hypochromicity) as was done in Sandaa et al. (1).

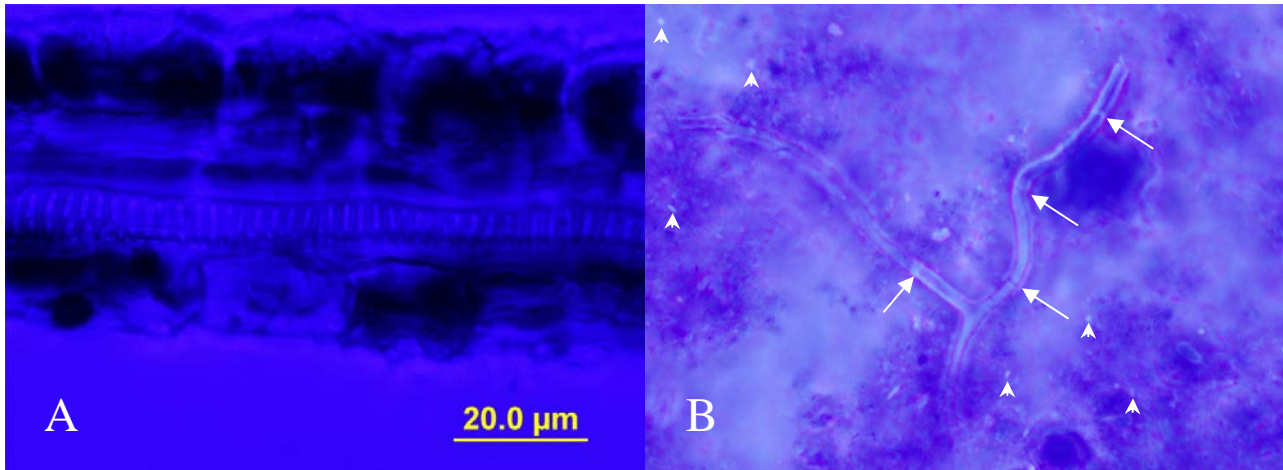
## 5 Conclusion

Current estimates of microbial richness in a soil sample span 5 orders of magnitude, from  $< 100$  phylotypes (14) to almost  $10^7$  (2). Ironically, many of the estimates constituting this extreme range may in fact be correct, though imprecise. When the standard error approaches  $2 \times 10^8$  (as above), the estimate may assume any value between unrealistic extremes, and still remain a technically correct, though practically useless, measure of the true bacterial richness. While it may well be possible to estimate species richness by analyzing DNA reassociation kinetics, such an enterprise will require a more realistic physical (chemical) model of the basic reaction (15), analysis of sensitivity to assumptions and fixed constants, comparison of competing models, and improved statistical parameter estimation and variance assessment. This is a topic for future study.

## References

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. Érdi, J. Tóth, *Mathematical Models of Chemical Reactions* (Princeton University Press, Princeton, 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. Agr. Biol. Envir. St.* **8**, 387 (2003).
6. G. A. F. Seber, C. J. Wild, *Nonlinear Regression* (Wiley, New York, 1989).
7. R.-A. Sandaa, personal communication.
8. S-H. Hong, J. Bunge, S-O. Jeon, S. S. Epstein, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 117 (2006).

9. D. Bohning, D. Schon, *J. R. Stat. Soc. Ser. C* **54**, 721 (2005).
10. D. E. Graham, B. R. Neufeld, E. H. Davidson, R. J. Britten, *Cell* **3**, 127 (1974).
11. R. J. Britten, D. E. Graham, B. R. Neufeld, *Method. Enzymol.* **29**, 363 (1974).
12. M. L. Nagy, A. Perez, F. Garcia-Pichel, *FEMS Microbiol. Ecol.* **54**, 233 (2005).
13. A. Bissett, J. Bowman, C. Burke, *FEMS Microbiol. Ecol.* **55**, 48 (2006).
14. P. F. Kemp, J. Y. Aller, *FEMS Microbiol. Ecol.* **47**, 161 (2004).
15. R. Murugan, *Biochem. Biophys. Res. Commun.* **293**, 870 (2002).
16. *Acknowledgment of support:* This work was supported in part by National Science Foundation award DBI-0421717 to D.G. Peterson.
17. *Acknowledgment of support:* This work was supported in part by National Science Foundation grants OCE-0221267, MCB-0348341 and DEB-0103599 to S. S. Epstein.



**Fig. S1** Our tests of the bacterial isolation method utilized by Sandaa et al. (1) revealed considerable contamination of the “bacterial pellet” with eukaryotic cells and tissues. For example (A) a plant tracheary element with attached cells and (B) a fungal mycelium with septa (arrows). Bright DAPI-stained bacteria are visible in the latter image (e.g., arrowheads).